



UNIVERSITAT DE
BARCELONA

Design of bioinformatic tools for integrative analysis of microRNA-mRNA interactome applied to digestive cancers

Maria Vila Casadesús



Aquesta tesi doctoral està subjecta a la llicència **Reconeixement- Compartigual 4.0. Espanya de Creative Commons.**

Esta tesis doctoral está sujeta a la licencia **Reconocimiento - Compartigual 4.0. España de Creative Commons.**

This doctoral thesis is licensed under the **Creative Commons Attribution-ShareAlike 4.0. Spain License.**

Design of bioinformatic tools for integrative analysis of microRNA-mRNA interactome applied to digestive cancers

Memòria presentada per
Maria Vila Casadesús

Per optar al grau de
Doctora per la Universitat de Barcelona

Tesi dirigida pel
Dr. Juan José Lozano Salvatella i la Dra. Meritxell Gironella Cos

Dr. Juan José Lozano Salvatella
Director

Dra. Meritxell Gironella Cos
Directora

Dr. Victor Raúl Moreno Aguado
Tutor

Maria Vila Casadesús
Doctoranda

La tesi s'ha realitzat al
Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas
(CIBEREHD) i
l'Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS)

Programa de Doctorat en Biomedicina
Universitat de Barcelona, 2017

Agafa la vida amb un somriure,

Agraïments

Voldria aprofitar aquestes pàgines per agrair totes aquelles persones que, d'alguna manera o altra, han estat importants per l'elaboració d'aquesta tesi:

En primer lloc, als meus directors Txell i Juanjo l'oportunitat de realitzar aquesta tesi. Al Juanjo per haver confiat en mi i haver-me permès iniciar aquest projecte just acabada la llicenciatura i haver-me deixat fer lliurement. A la Txell, li he d'agrair molt especialment tota la dedicació, per ajudar-me a enfocar la temàtica d'aquesta tesi, per les paraules d'ànim i transmetre'm aquesta passió. Moltes gràcies als dos!

Però aquesta tesi no es podria haver fet sense la col·laboració de moltes altres més persones. En concret, Elena Vila, que ha realitzat totes les validacions de miRNAs, moltes gràcies! També agrair a Jan Graffelman, que va corregir el treball fi de màster (que va ser l'embrió d'aquesta tesis). Són importants també totes les persones amb qui he tingut l'oportunitat de conèixer i/o col·laborar durant aquests anys, ja sigui directament relacionat amb temes de miRNAs o per altre tipus de dades: Sergi Castellví, Clara Esteban, Jordi Camps, Isa Quintanilla, Antoni Castells, Roser Pinyol, Pau Sancho, Delia Blaya, Mar Coll, Helena Cornellà, Xevi Bofill, Cristina Fillat, Sergio Lario, i molts altres investigadors del CEK o altres centres. Conèixer altres temes, altres camps, altres punts de vista, ha estat molt enriquidor. Moltes gràcies per les estones, part de la tesi no es podria explicar sense vosaltres.

També han sigut importants la Manuela Hummel, l'Annette Kopp-Schneider i la resta del grup de bioestadística del DKFZ a Heidelberg, a on vaig realitzar una estada la tardor de 2015. *The three months spent there showed me new ways to work, new ideas, and were a breath of fresh air. Danke schön.*

En terreny més personal, també he d'agrair a tots els bioinfos i companys de soterrani: Sebas (company de despatx durant molt temps!), Eva, Marcos, Guillaume, Keyvan, Dani. A tots vosaltres (així com també la resta de gent del grup de gastro en bloc), moltes gràcies

per les estones disteses fent el cafè o dinant, que a vegades és quan surten les millors idees (o les més esbojarrades), i especialment aquests últims mesos per tot el suport que m'heu donat: m'heu ajudat moltíssim. I també els que ja han marxat però hi van ser durant una part important: Marc, Brisa i Esther, Pau, Núria, Joan, Sessa, Guerau: moltes gràcies per les estones *freaks* i els consells.

Als antics companys de Vic, allà on va començar l'aventura. Perquè no n'esperava res d'estudiar allà i en va sortir molt. Moltes gràcies pels moments: Jenny, Sandra, Nati, Laura, Miriam, Rocio i tots els altres. Perquè no es poden resumir 4 anys (i els que han vingut després) amb una línia, però sapiguen que m'ha encantat conèixer-vos. Al grup de *Drosophila* del Parc Científic (Bàrbara en especial), per ensenyar-me a classificar mosques, però sobretot què era un doctorat. Als companys del màster, perquè van ser ells qui em van ensenyar el *cantó fosc* i l'obsessió de fer-ho tot amb \LaTeX .

A totes les persones que he anat coneguent a Barcelona: a tots els companys del pis d'Aribau que vaig anar tenint durant 5 anys, amb els que ha estat un plaer compartir tants moments i riures junts; a la gent de la plaça del rei (a la Núria Duran per introduir-m'hi), als "manelics"; per fer aquest camí molt més lleuger. A l'Enric i l'Imma per obrir les portes de casa seva. I a totes les altres persones que he tingut l'oportunitat de conèixer aquests últims anys. Perdoneu si no us puc nombrar tots, alguns ens veiem més, altres menys, però tots heu deixat una empremta única.

I no podria acabar sense agrair molt especialment a la meva família: pare, mare, Núria i des de fa molt de temps el Marc, per tot. Per haver-me donat sempre suport des de que era petita i animar-me en els moments difícils, però també compartir els bons. Per haver-me donat les eines per arribar aquí, per donar un punt de vista des de fora, perquè la família no es tria però de la meva n'estic molt contenta i orgullosa.

I finalment donar moltes gràcies al Francesc, per haver compartit i viscut l'última part d'aquest viatge. Pels consells, per les estones, pels ànims que m'has donat, per les rialles, pels descobriments i les aventures, per ser allà quan ha fet falta, per moltes coses que no em cabrien aquí, però sobretot per les que encara han de venir.

Moltes gràcies a tots!

Table of Contents

Abstract	xi
Abbreviations	xiii
1 Introduction	1
1.1 MiRNA	1
1.1.1 MiRNA biogenesis	3
1.1.2 MiRNAs in cancer	6
1.2 MiRNA-mRNA interactions	12
1.2.1 Databases of precomputed miRNA-mRNA interactions	13
1.2.1.1 MicroCosm	15
1.2.1.2 TargetScan	15
1.2.1.3 MiRSVR	18
1.2.1.4 MiRDB	19
1.2.1.5 MiRWalk and databases of validated interactions	20
1.3 Expression-based methods for detecting miRNA-mRNA interactions	23
1.3.1 Methods considering up-down pairing	24
1.3.2 Methods considering correlation	26
1.3.3 Methods considering other procedures	29
1.3.4 Other tools for working with miRNAs	31
2 Objectives	33
2.1 Main objective	33
2.2 Secondary objectives	33

3	Materials & Methods	35
3.1	Data obtention & Preprocessing	35
3.1.1	STUDY 1 – MiRComb in five digestive cancers	35
3.1.1.1	Samples	35
3.1.1.2	Preprocessing	35
3.1.2	STUDY 2 – MiRComb in pancreatic cancer	36
3.1.2.1	Samples	36
3.1.2.2	Next Generation Sequencing	37
3.1.2.3	Gene expression arrays	39
3.1.2.4	Cell culture	39
3.1.2.5	CRISPR/Cas9 targeting of miR-21 in PANC-1 cells	39
3.1.2.6	RNA extraction and Target expression analysis by qRT-PCR	40
3.2	Design of a new tool for analysing miRNA-mRNA interactions: miRComb	41
3.3	Differential Expression	42
3.3.1	T-test	44
3.3.2	Wilcoxon Test	46
3.3.3	Limma (Linear Models for Microarray Data)	46
3.3.4	RankProd	49
3.3.5	DESeq, edgeR	50
3.4	Subset selection	54
3.5	MiRNA-mRNA associations	56
3.5.1	Pearson Product-Moment Correlation Coefficient	57
3.5.2	Spearman Rank Correlation Coefficient	57
3.5.3	Kendall τ Correlation Coefficient	58
3.5.4	Generalised Linear Models	59
3.6	Database integration	62
3.6.1	P value combination: Fisher method	62
3.6.2	P value combination: Stouffer method	64
3.6.3	Intersection	65
3.7	Multiple testing correction	67
3.7.1	Bonferroni Correction	68
3.7.2	Benjamini & Hochberg correction	69
3.8	Functional analysis of miRNA targets	72

3.8.1	Enrichment Analysis	72
3.8.1.1	Proportions test	73
3.8.1.2	Hypergeometric test	74
3.9	Analysis of number of targets per miRNA	76
3.9.1	Hypergeometric Test	76
3.9.2	Logistic Regression	78
3.9.3	Gene Set Enrichment Analysis	79
3.10	Additional pipelines	82
3.10.1	Time-series analysis	82
3.10.2	Non-matched miRNA-mRNA data	84
4	Results	85
4.1	MiRComb R package	85
4.1.1	MiRComb statistics	87
4.1.2	MiRTools	89
4.1.2.1	MiRTranslator	90
4.1.2.2	MiRCircos	93
4.2	MiRComb parameters exploration	98
4.2.1	Differential expression methods	98
4.2.2	Effect of subset selection	100
4.2.3	Pearson vs Spearman vs Kendall vs Glmnet	102
4.2.3.1	Correlation methods	103
4.2.3.2	Glmnet vs Pearson estimates	103
4.2.4	Integrative approaches	105
4.2.4.1	Fisher vs Stouffer	106
4.2.4.2	Fisher vs Stouffer vs Intersection	110
4.3	STUDY 1 – MiRComb in digestive cancers	112
4.3.1	MiRComb analysis of miRNA-mRNA interactions of 5 different di- gestive cancers	112
4.3.1.1	Summary of datasets composition	112
4.3.1.2	Analysis of miRNA-mRNA interactions	115
4.3.1.3	Functional enrichment analysis of miRNAs according to their targets	118
4.3.2	Integrative analysis of the miRComb miRNA-mRNA interactions from the 5 digestive cancers	119

TABLE OF CONTENTS

4.3.2.1	Shared and specific miRNA-mRNA interactions	119
4.3.2.2	Cluster analysis of miRNA-mRNA interactions	121
4.4	STUDY 2 – MiRComb in pancreatic cancer	129
4.4.1	Data exploration	129
4.4.1.1	Top differentially expressed miRNAs or mRNAs	129
4.4.1.2	Intersection with miRNA target prediction databases	131
4.4.2	MiRComb results in the pancreatic cancer set	132
4.4.3	Confirmation of miR-21 targets in a pancreatic cancer cellular model	139
5	Discussion	143
6	Conclusions	155
	Bibliography	157
	Appendices	177
A	Reports of STUDIES 1 and 2	179
A.1	Study 1: Colon adenocarcinoma	179
A.2	Study 1: Esophageal carcinoma	188
A.3	Study 1: Liver hepatocellular carcinoma	196
A.4	Study 1: Rectum adenocarcinoma	205
A.5	Study 1: Stomach adenocarcinoma	214
A.6	Study 2: Pancreatic ductal adenocarcinoma	223
B	MiRComb vignettes and manuals	233
B.1	Main vignette	233
B.2	Additional vignette	250
B.3	Manual	256

Abstract

Introduction MicroRNAs (miRNAs) are small RNA molecules that regulate the expression of target mRNAs by specific binding on the mRNA molecule and mostly promoting mRNA degradation. It is of great interest to know the specific targets of a miRNA in order to study them in a particular disease context. Some algorithms have been designed to predict potential miRNA-mRNA interactions based on sequence hybridisation, but one of the main problems of them is that they have too many false positives and do not take into account disease-specific interactions.

Objectives The main aim of the study was to build a tool able to analyse miRNA-mRNA interactions based on the combination of biological information and theoretical information (databases of miRNA-mRNA interactions). Secondary objectives are to analyse miRNA-mRNA interactomes in the context of digestive cancers and to validate some of the results.

Methods We used the following methodology: firstly, we obtained expression data from patient samples. Secondly, we selected differentially expressed miRNAs and mRNAs and used them to compute miRNA-mRNA correlations. Then, we matched the negative correlations with preexisting target prediction databases. The final selected miRNA-mRNA interactions were those that their expression is negatively correlated, and appear as predicted in at least one of the selected databases. Functional analysis on the miRNA-mRNA pairs can also be done.

Results We built an R package –miRComb– that is able to carry out the entire analysis and allows to choose between different options in each step, as well as web-based tools aimed to deal with miRNA data. MiRComb package was tested in public available data (TCGA data from colon, esophagus, liver, rectum and stomach cancer) and a custom

set of pancreatic cancer samples. MicroRNA-mRNA interactomes of these cancers were revealed and summarised into reports using miRComb report function. In the first study, a meta-analysis of all the TCGA cancers was also performed, highlighting the similarities and differences between them. In the second study, we focused on the miRNA-mRNA interactions in the context of pancreatic cancer, and two miRNA-mRNA interactions from *hsa-miR-21* were also validated in a pancreatic cancer model.

Conclusions MiRComb package performs the entire analysis of miRNA-mRNA interactions in a single software environment and summarises the results in a useful way. A methodology of reference has been proposed, and miRNA-mRNA interactomes of colon, rectum, stomach, esophageal, liver and pancreatic cancer have been reported and are ready for further experiments in a wet lab.

Abbreviations

AUC	area under the curve
bp	base pair
cDNA	complementary DNA
COAD	colon adenocarcinoma
DNA	deoxyribonucleic acid
FC	fold change
FDR	false discovery rate
FWER	family wise error rate
FPKM	fragments per kilobase per million mapped reads
ES	enrichment score
ESCA	esophageal carcinoma
GO	gene ontology
GSEA	gene set enrichment analysis
IPMN	intraductal papillary mucinous neoplasm
IQR	inter quartile range
KEGG	Kyoto encyclopedia of genes and genomes
LASSO	least absolute shrinkage and selection operator
LIHC	liver hepatocellular carcinoma
miRNA	micro ribonucleic acid (microRNA)
mRNA	messenger RNA
NES	normalised enrichment score
nt	nucleotide
NGS	next-generation sequencing
OOP	object-oriented programming
OR	odds ratio
ORF	open reading frame
READ	rectum adenocarcinoma
RNA	ribonucleic acid

TABLE OF CONTENTS

ROC	receiver operating characteristic
STAD	stomach adenocarcinoma
SVR	supor vector regression
TSS	tissue source site
UTR	untranslated region

Chapter 1

Introduction

1.1 MiRNA

MicroRNAs (miRNAs) are a non-coding, single-stranded RNAs of 18-25nt long and constitute a novel class of gene regulators that are found in both plants and animals [1, 2]. They negatively regulate the expression of their targets (one mRNA is a *target* of a miRNA if this miRNA regulates that mRNA) in one of two ways (which are detailed in Section 1.1.1) depending on the degree of complementarity between the miRNA and the target.

MiRNAs are one of the elements regulating mRNA expression, the other known elements participating in the regulation are shown in Figure 1.1:

- **Epigenome:** the epigenome is attached DNA modifications that do not change the DNA sequence but can affect gene activity, and can be inherited. Common epigenetic marks are DNA methylation (mostly cytosine nucleotides on CpG islands), histone modifications and nucleosome positioning [3]. DNA methylation has been related to underexpressed genes, and the other factors influence DNA transcription. Abnormal methylation and other epigenetic changes have been related to cancer and other diseases [4].
- **Other RNA elements:** *Long non-coding RNA (lncRNA)* are long non-coding RNA molecules (more than 200nt long) with no clear function, but some of them are described to be able to inhibit or activate genes [5, 6], and their expression can be specific to the tissue or cell and vary across time or respond to stimulus [7]. *Circular RNA (circRNA)* are circular fragments of RNA of different length. Circular

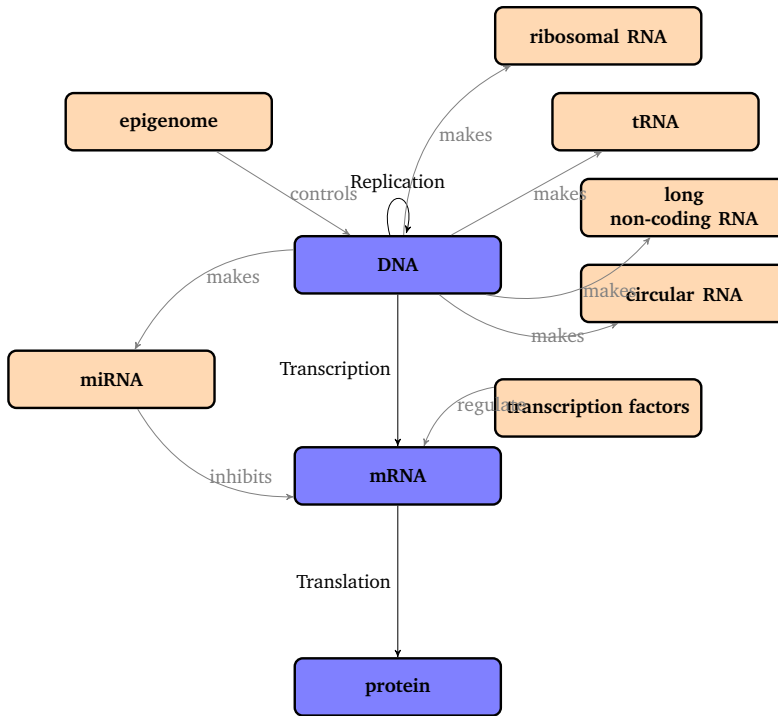


Figure 1.1: Summary of basic elements regulating DNA, mRNA and protein expression. MiRNAs are one of small RNA elements and regulate. Each of these processes is controlled by, or involve, proteins.

RNAs have the ability to rearrange the order of genomic elements of the DNA strand and influence RNA folding. Moreover, they can act as templates for viral replication, regulators of transcription or miRNA sponges [8]. *Short interfering RNA (siRNA)* are similar to miRNAs (20-25nt length) that attach to the mRNA and degrade them. However, the difference between them is that siRNAs came from other parts of the genome and are double-stranded RNA molecules [9]. *Piwi-interacting RNA (piRNA)* are small RNAs slightly larger than miRNAs (26–31nt), and more complex and not conserved between species. Functions and mechanisms of action are diverse and still being studied, but they, among others, may cause transposon silencing and interfere normal gene expression [10]. *Small nucleolar RNA (snoRNA)* are small RNAs that primarily guide chemical modifications of other RNAs, mainly ribosomal RNAs [11]. There are also other types of RNA, such as transfer RNAs (tRNA) and ribosomal RNAs (rRNA), that have structural functions but they also participate in the translation process [12, 13].

- Apart from that, **proteins** itself, as enzymes and factors, participate in all the previ-

ously described processes. *Transcription factors* are proteins that regulate the transcription of mRNA.

Among all the regulatory elements, miRNAs are especially interesting because one miRNA can regulate up to hundreds of mRNAs and are more stable than other RNA elements. MiRNAs are this thesis' subject matter.

The most complete database about miRNAs is miRBase database [14]. It is a searchable database of published miRNA sequences and annotation. The miRBase registry provides miRNA gene hunters with unique names for novel miRNA genes prior to publication of results. All sequences are available for searching and browsing, and entries can also be retrieved by name, keyword, references and annotation. Moreover, all the information is also available for download.

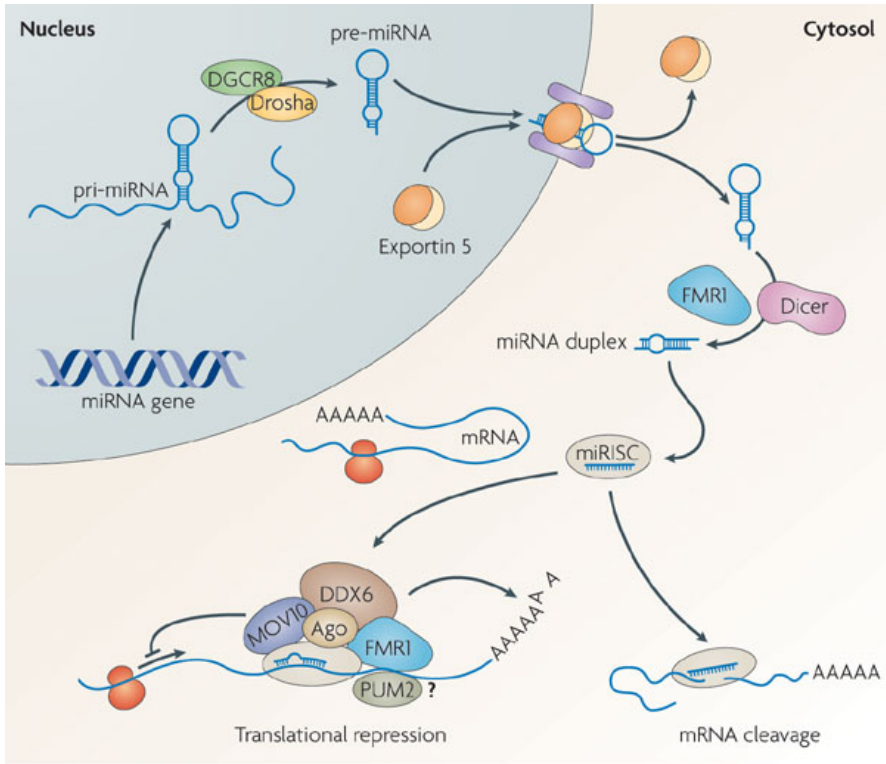
1.1.1 MiRNA biogenesis

Figure 1.2 shows miRNA biogenesis and their mechanism of action. [15, 16].

Transcription of microRNA (miRNA) genes by RNA polymerase II leads to the generation of primary miRNA transcripts (pri-miRNAs –primary-miRNA–, such as *pri-mir-21*). Drosha and Pasha constitute the microprocessor complex that cleaves the pri-miRNA to a pre-miRNA (precursor-miRNA, for example, *-pre-miR-21-*) hairpin [17].

The pre-miRNA is exported to the cytoplasm by an exportin 5-dependent pathway. There, the pre-miRNA is cut by the RNase Dicer (the activity of which can be modulated by accessory proteins, for example, FMR1) to an intermediate miRNA duplex. This duplex contains two mature miRNA fragments, for example, the '*miR-21-5p*' and the '*miR-21-3p*'. The *miR-21-5p* is transcribed in the original direction of translation, while the *miR-21-3p* is the complementary. Usually, one of them is immediately degraded (commonly the '*3p*'). In previous versions of miRBase, the degraded miRNA was labelled with an asterisk (for example, *miR-21**) and the other was not labelled (for example, *miR-21*) [18]. However, this is not a rule of thumb and sometimes both fragments remain stable and functional in the cell [19]. That is why in recent versions of miRBase the mature miRNAs are referred as '*5p*' and '*3p*'.

The mature miRNA (either the '*5p*' or the '*3p*') are integrated into a multiprotein complex called miRNA-induced silencing complex (miRISC). This complex is guided to target mRNAs where they both hybridise, preferentially in locations near the 3' untranslated



Nature Reviews | Neuroscience

Figure 1.2: RNA biogenesis and mode of action, extracted from [16].

regions of the mRNA. A complete match is required on the seed, but some mismatch or bundles are allowed on the other parts of the union [20].

There, two options are possible: translational repression or degradation of the target mRNA, where mRNA degradation occurs around 84% of the times [15]. In both cases, the protein level is always repressed. The detailed mechanism of the miRISC is still controversial but includes argonaute (Ago) [21], helicases MOV10 and DDX6 (also known as RCK and p54), plus RNA-binding proteins such as FMR1 and PUM2 [16].

It is expected that more than 60% of the total mRNAs present in a cell are possibly regulated by miRNAs [22]. Although there is no described biological limit, while a miRNA can have hundreds of targets, it is not expected that a mRNA have a very high number of miRNAs targeting it at the 3' end, as each miRNA needs its own RNA sequence to recognise and hybridise.

MiRNAs can be grouped into families or clusters. A miRNA family is formed by dif-

Nomenclature	
miRNA gene	MIR21
primary miRNA	hsa-mir-21 (not specified)
precursor miRNA	hsa-mir-21
mature miRNA	<i>hsa-miR-21-5p</i> and <i>hsa-miR-21-3p</i> (mirbase ≥ 18) <i>hsa-miR-21</i> and <i>hsa-miR-21*</i> (mirbase ≤ 17)

Table 1.1: Nomenclature of the miRNAs according to their maturity state. Sometimes the species of the miRNA (in this case, *hsa*-) can be omitted. Precursor and primary can also be labelled with the prefixes *pre*- and *pri*-, respectively.

ferent miRNAs that share the same seed region. The seed region consists of the 8 first nucleotides on the 5' and is important for miRNA-mRNA matching (see Section 1.2) [23]. Thus, the miRNAs of the same family share most of their targets due to sequence similarity.

Clusters are groups of miRNAs that are transcribed together because they are located nearby in the genome. They are not necessary members of the same family and their sequence can be very different and target different mRNAs (Figure 1.3).

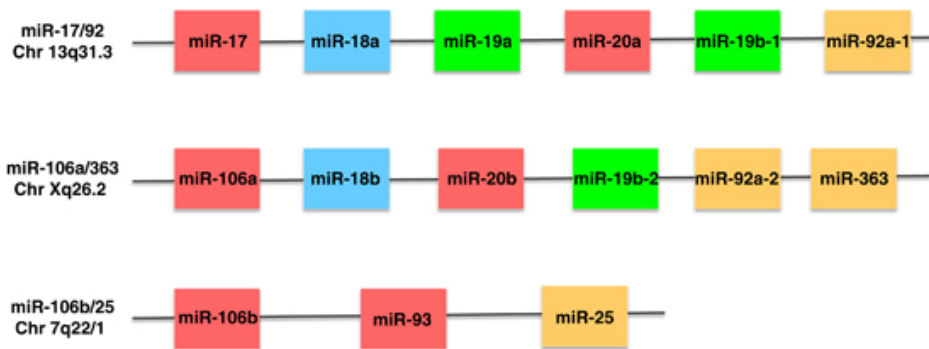


Figure 1.3: Figure extracted from [24] showing the difference between clusters (rows: miR-17/92, miR-106a/363 and miR-106b/25 clusters) and families (identified by different colours. Red: members of the miR-17 family; blue: members of the miR-18 family; green: members of the miR-19 family; orange: members of the miR-92 family).

However, despite not targeting the same mRNAs, some miRNA clusters have been shown to target mRNAs of the same pathway, meaning that the clusters might have some kind of biological sense. MiR-19/92 cluster, located on chromosome 13, encodes seven mature miRNAs and is a typical example of common action in inhibiting cell growth and proliferation pathways in cancer [25, 26, 27].

MiRNA genes can also be found in several places on the genome. For example, in Figure 1.3 we can see that miR-19b is coded either on miR-17/92 cluster located on chromosome 13 and on miR-106a/363 cluster located on chromosome X. Both genes produce the same miRNAs, which cannot be differentiated, and precursor miRNAs can be differentiated by an added number to their nomenclature (miR-19b-1 and miR-19b-2).

1.1.2 MiRNAs in cancer

Cancer is the name given to a collection of related diseases. In all types of cancer, some of the body's cells begin to divide without stopping forming masses of tissue called tumours (except some cancers such as blood cell cancers –leukaemias–). Ultimately, these cells are able to spread into surrounding tissues and form new tumours there –metastasis–.

Figure 1.4 shows a general model for cancer evolution. The change to normal tissue to invasive and metastatic cancer is driven by genetic and epigenetic changes of the cells. The chronology of this progression is specific for each cancer and tissue, ranging from several months since the initiation of the lesion or more than 20 years [28].

Cancer stage can be defined based on several criteria (TNM –Tumour, Node, Metastasis– system is one of the most used ones [29]), that can be summarised in four stages: being *I* a small tumour not spread to lymph nodes, *II* and *III* a tumour that started a spread to lymph nodes and *IV* a metastatic tumour. Stage 0 is sometimes used for precursor lesions. Precursor lesions (or dysplasias) are any alteration of the cells that cause abnormal development, observable either at macroscopic or microscopical level that may develop to a cancer tumour. Although the probability that these lesions eventually evolve to cancer cells is in overall low (Figure 1.4), they are studied in many types of cancer. A typical example is colon cancer prevention, where adenomas and other lesions are systematically removed from the patients undergoing colonoscopy [30]. The stage at diagnostics is related to survival [31], as well as the original tissue [32].

There are more than 100 distinct types of cancer, and subtypes of tumours can be found in specific organs. Despite this huge diversity, it is proposed that all these cancer cell genotypes are a manifestation of six essential alterations in cell physiology that collectively dictate malignant growth (Figure 1.5), shared in common by most and perhaps all types of human tumours:

Any cancer has acquired all of these capabilities, but its means of doing so vary significantly, both mechanistically and chronologically. These capabilities are:

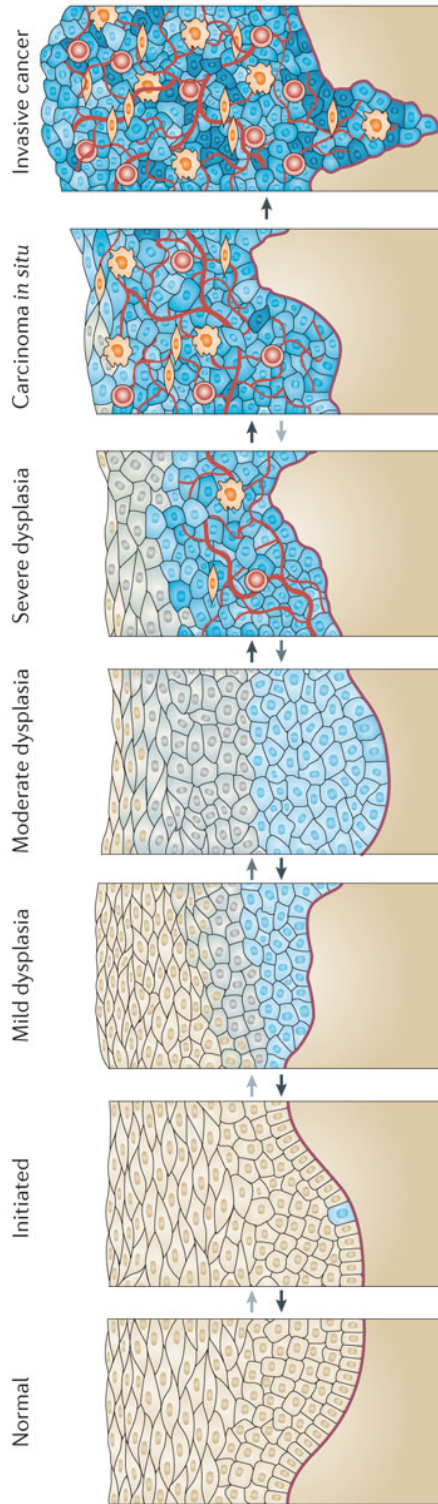


Figure 1.4: General model for cancer progression. The intensity colour of the arrow indicates the probability and direction of stage transition. Adapted from *Future directions in cancer prevention* Asad Umar, Barbara K. Dunn & Peter Greenwald [28]

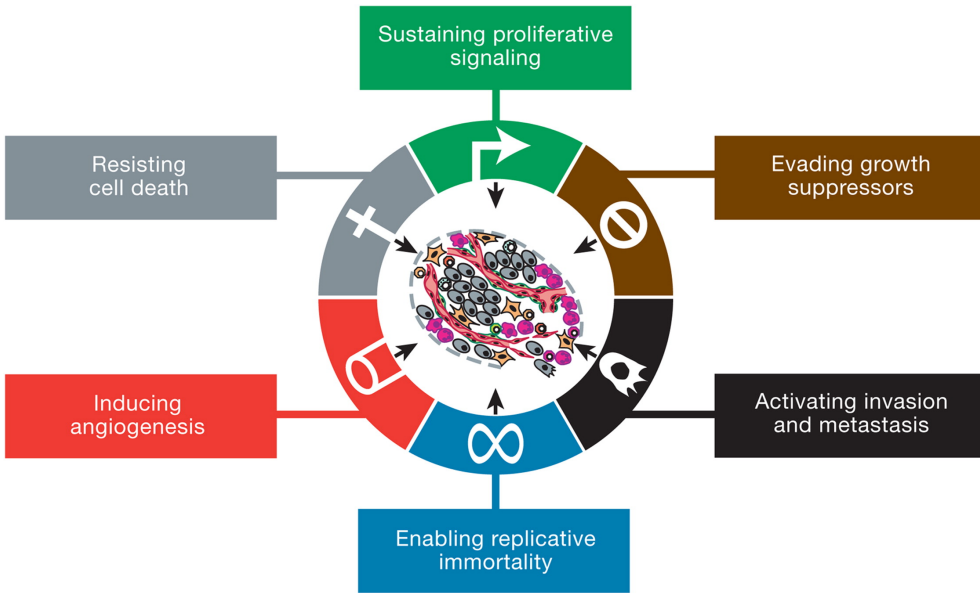


Figure 1.5: Hallmarks of cancer. Extracted from [33] and [34]. Other characteristics called enabling characteristics are included on ampliations of these hallmarks: *Genome instability and high mutation rate*, and *tumour associated inflammation response*. New hallmarks such as *Reprogramming energy metabolism* and *Evading Immune Destruction* are further discussed in [33].

- Sustaining Proliferative Signaling:** cells require mitogenic growth signals to move from a quiescent state into an active proliferative state. There are many strategies to stimulate the growth: produce growth signals (PDGF (platelet-derived growth factor) [35] or TGF α (tumour growth factor α) [36]), overexpress or change the receptors on the cell surface; or interfere on the signalling cascade (ras mutations on the SOS-Ras-Raf-MAP kinase mitogenic cascade) [37, 38].
- Evading Growth Suppressors:** normal cells can stop growing in two ways: stationary stop, where the cell is forced to a quiescent but reversible state (regulated by TGF β and associated receptors [39]) or permanent stop due to cell differentiation (regulated by the Mad–Max complexes [40]). Tumour cells are able to evade growth suppressor signals [34].
- Resisting Cell Death:** The apoptotic machinery can be broadly divided into two classes of components: sensors (FAS receptors of TNF α [41]) and effectors (caspases, cytochrome release, and pathways involving p53 [42]). Inactivation of sensors and overexpression of effector cascades are found in cancers [43, 44].
- Enabling Replicative Immortality:** telomeres are repetitive regions at the end of

the chromosomes shortened in each replication, eventually causing them to lose their ability to protect the ends of chromosomal DNA and causing cell death. Over 85%-90% of the tumour cells have increased telomerase (enzyme responsible for telomere elongation) activity [45].

- **Inducing Angiogenesis:** at the start of tumour formation, cells are not able to form new blood vessels. Only tumours able to start angiogenesis (usually via overexpression of VEGF factor) evolve to bigger and more malignant tumours [46, 47].
- **Activating Invasion and Metastasis:** in order to spread to other tissues, cells must change its configuration first by deadhering of the initial tissue epithelial–mesenchymal transition (EMT) and then adhering to a new tissue (MET) [48]. This process of is controlled mainly by extracellular proteases, cell–cell adhesion molecules (CAMs, including cadherin family) and integrins [49].

MiRNAs, as gene regulators, are participating in all of this processes [33, 50], controlling a wide range of biological functions such as cellular proliferation, differentiation and apoptosis [51]. Moreover, some pathways are interconnected, (such as cell survival, cell death and cell cycle pathways [52]), meaning that the range of functions that can regulate one single miRNA can be extensive.

MiRNAs can act as tumour suppressors or oncogenes (they are therefore referred to as "oncomirs") depending on the function of the target they are regulating. Furthermore, factors that are required for the biogenesis of miRNAs have also been associated with various cancers and might themselves function as tumour suppressors or oncogenes [53, 54].

Broadly known oncomirs are the miR-17/92 cluster or *hsa-miR-21*. For example, *hsa-miR-21* is overexpressed in a huge variety of cancers and targets BTG2 (a tumour suppressor gene [55]) and many other genes involved in critical regulation pathways. It is mainly involved in signalling pathways related to apoptosis and induction to cell survival, and it has been linked to chemotherapy resistance. This miRNA is perhaps one of the most studied miRNAs in cancer [56, 55, 57].

On the other hand, known miRNA-tumour suppressors are the miR-200 family and *hsa-miR-34a*, which are protectors of the epithelial phenotype. They repress EMT transition forming a negative loop with ZEB1 and ZEB2 transcription factors [58]. Despite these advances, however, the functional meaning of most of the deregulated miRNAs in the context of digestive cancers is still largely unknown.

Apart from that, expression profiling of miRNA has also been shown to be a more accu-

rate method of classifying cancer subtypes than using the expression profiles of protein-coding genes: the differential expression of certain miRNAs in various tumours might become a powerful tool to aid in the diagnosis and treatment of cancer [53].

Cancer impact

Cancer is the 2nd cause of death worldwide (globally, nearly 1 in 6 deaths is due to cancer) and one of the leading causes of morbidity, with 14.1 million new cases and 8.2 million deaths in 2012 [59]. Apart from that, World Health Organization has estimated that the number of new cases is expected to rise by about 70% over the next 2 decades [60].

Among all cancers, differences in incidence and survival are remarkable. Figure 1.6 shows cancer statistics divided per type of original tissue. Colorectum cancer, which comprises colon and rectum cancer, is the 3rd most incident cancer, and although it has a good prognosis compared to other cancers, is the 4th cancer ranked by mortality. When taking digestive cancers together (colon, rectum, stomach, liver, esophagus and pancreatic cancer), they account for a total of 27.6% of the incidence, and the 35.3% of the total deaths by cancer [59].

Survival rates for pancreatic and liver cancer are around 5% at 5 years [59], which makes them ones of the most lethal cancers. Esophageal and stomach cancer, with survival rates of 20% at 5 years are also on the ranking of top malignant cancers [32]. Furthermore, these rates have barely improved during the last 40 years, specially in the case of pancreatic cancer, which has not virtually increased [32]. Aberrant expression of miRNAs has been widely reported in all of these cancers.

For example, colon and rectum cancer have frequently APC, TGFBR2, TP53, SMAD4 or PTEN genes inactivated, KRAS activated and MYC overexpressed [61]. The relation between these genes, the pathways in which they are participating and miRNAs has been studied, observing that miRNAs often participate in positive or negative feedback loops [62]. Apart from that, these miRNAs (including miRNAs with still unknown function) can act as also as biomarkers or predictors of response to treatment [63].

Similarly, miRNAs have found associated with esophageal [64], liver [65, 66], stomach [67] and pancreatic [68, 69, 70, 71] cancers. Some miRNAs are specific to one type of cancer(s), while others are commonly deregulated [72].

Specifically to pancreatic cancer, several studies of miRNA expression profiling have defined miRNA signatures for PDAC that are associated with diagnosis, staging, progres-

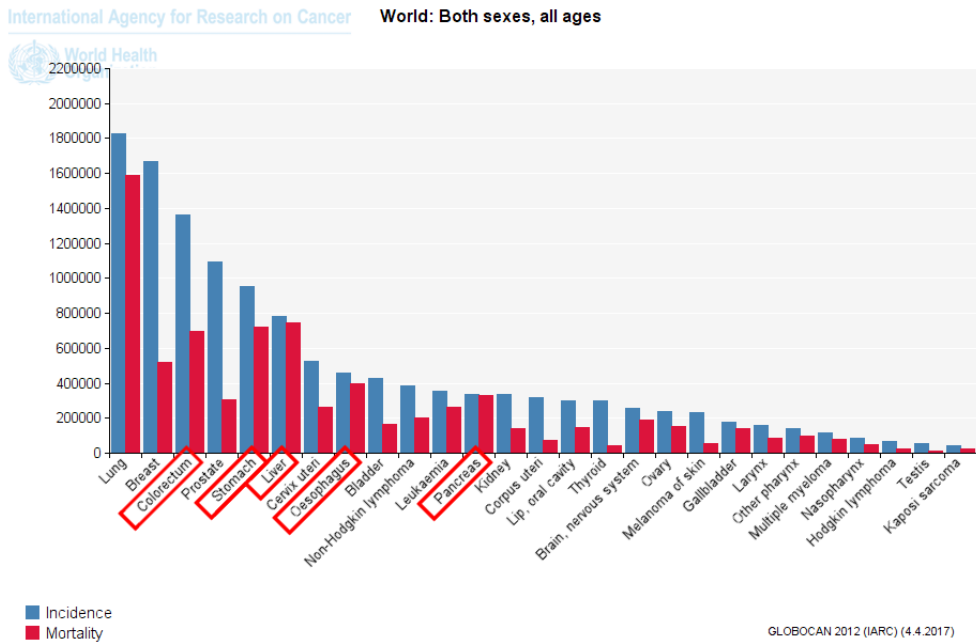


Figure 1.6: Heterogeneity of total incidence and mortality worldwide by type of cancer. Data obtained from GLOBOCAN [59]. Digestive cancers are highlighted in red.

sion, prognosis, and response to treatment [73, 74, 75, 76].

Our group has also been working with miRNAs in the context of pancreatic cancer and several miRNAs have been proposed as early detection biomarkers [77].

1.2 MiRNA-mRNA interactions

MiRNA-mRNA interactions are based on RNA hybridisation: cytosine (C) matches with guanine (G) and adenosine (A) matches with uracil (U) to form double stranded RNA molecules. Figure 1.7 shows the details and different parts of a miRNA-mRNA interaction hybridization.

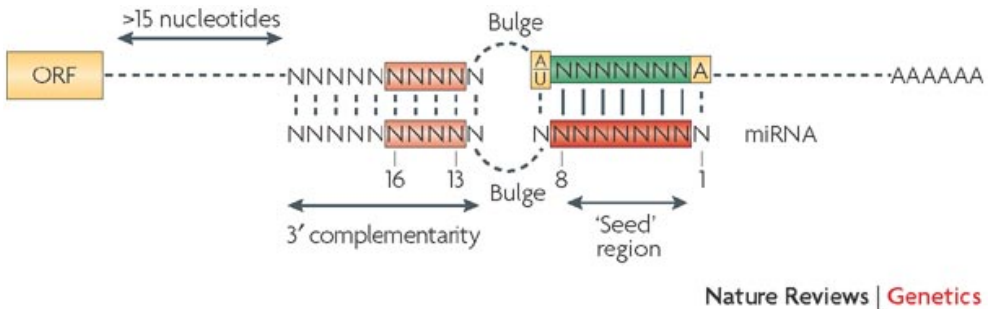


Figure 1.7: Representation of a miRNA-mRNA interaction. MiRNA-seed region determines the miRNA family. Extracted from [78]

In miRNA-mRNA hybridisation, the most critical part of the interaction is the seed region of the miRNA (8-6 base pairs at the 5' of the miRNA, starting on the second position of the miRNA) in which the hybridisation must be perfect. The seed region is used to determine miRNA families [14]. As this is the most important part to determine a miRNA-mRNA interaction, it is not rare that miRNAs of the same family regulate the same targets. A residue at position 1 of the miRNA, and an A or U at position 9 (shown in yellow) improve the site efficiency, although they do not need to be paired with miRNA nucleotides [79].

The other region of the miRNA helps to determine the energy and specificity of the union, and where the miRSC complex is located [23, 20]. More sequence similarity will produce strong miRNA-mRNA interactions, but this is not a perfect union, as bulges and/or mismatches are allowed (but the biggest ones only in the central region), and other factors help to determine a miRNA-mRNA union [80]. There must be reasonable complementarity to the miRNA 3' end to stabilise the interaction. Mismatches and small bulges are also tolerated in this region, although good base pairing, particularly to residues 13–16 of the miRNA (shown in orange on Figure 1.7), becomes important when matching in the seed region is suboptimal [81].

Other factors that can improve miRNA-mRNA hybridisation are: a location not too far

away from the poly(A) tail or the termination codon (specially for long 3' UTRs) and an AU-rich neighbourhood. These factors can make the 3' UTR regions less structured and hence more accessible to miRNA recognition [82].

Apart from that, multiple sites for the same or different miRNAs are generally required for effective repression. When they are present close to each other (< 8 nt) they tend to act competitively, when two miRNAs compete for the same target, while when they are separated by more than 8 nt (usually 8–40 nucleotides apart) they tend to act cooperatively and multiply the effect of the repression [83, 20]. This competence with other miRNAs and other regulatory mechanisms determine the miRNA-mRNA interactions taking place in each cellular state and explains, in part, why a physical hybridisation does not necessarily mean that the interaction is taking place at that specific moment.

It is known that miRNA-mRNA interactions are dependent on the situation of the cell: some interactions only occur upon certain characteristics but not in others (for example *hsa-miR-21* is known to be deregulated in several types of cancer and cardiovascular diseases [17], but has not yet described in other diseases such as Alzheimer or Parkinson). This is why, although is possible to predict these interactions bioinformatically (measuring the energy of hybridisation, among others), is not possible to rely only on these theoretical predictions: **databases are useful in the sense that tell us that the interaction can be produced, but they are not telling us if the interaction is actually happening.**

In fact, according to several studies, the estimated percentage of false positives on database predictions may range from 24% to 70% [84, 85, 86].

1.2.1 Databases of precomputed miRNA-mRNA interactions

MiRNA-mRNA interactions can be predicted using different methods. As mentioned before, the hybridisation depends on the seed section of the miRNA plus complementarity on the 3' region of the miRNA. Other factors such as sequence context also determine if an interaction is produced or not [87]. Moreover, sequence conservation between species also helps identify functional regions of the DNA, where target sites are preferentially located [22, 88].

In a whole, a lot of factors determine miRNA-mRNA interactions in a physical way. Table 1.2 summarises the main characteristics of the databases that we have used for this work.

Database	Approximate number of interactions*	# miRNA	# mRNA	Algorithm	Current version	Last update
microCosm [14]	563179	690	22107	miRanda + free energy + sequence conservation	5	October 2007
targetScan [89]	19985	329**	12445	SVR	7.1	June 2016
miRSVR [87]	598741	249	19144	miRanda + SVR	3	August 2010
miRDB [90]	827626	2588	16218	SVR	5.0	August 2014
mirWalk [91]	959552	2578	20022	Union of databases	2.0	February 2016
miRecords [92]	2115	304	1114	Manually curated	4	April 2013
miR2Disease [93]	804	181	407	Manually curated	1	March 2011
TarBase [94]	>500000(***)	304	1114	Manually curated from published experiments	7.0	2014

Table 1.2: Currently databases available for download. Only human miRNA-mRNA pairs are reported, although almost all the databases of predicted targets support mouse targets and other organisms. Validated databases, however, are more focused on human miRNA-mRNA interactions. *The number of interactions depend on the significance limit used. Here we computed the number of interactions according to the recommended significance cutoff. First group of rows: predicted databases; second group of rows: validated miRNA-mRNA interactions. **In TargetScan, there are counted as miRNA families. ***Total interactions in 24 species. SVR: support vector regression.

1.2.1.1 MicroCosm

Microcosm [14, 95, 96] is the database developed by Enright Lab at the EMBL-EBI formerly known by "MiRBase Targets" that predicts the interaction between miRNA and mRNA.

The miRanda algorithm (Figure 1.8) computes the p value of each possible miRNA-mRNA interaction and it is based on the work from Enright AJ. et al. in 2003 [95]. Other databases are based on similar algorithms, as they take into account the same features (seed complementarity, the energy of hybridisation and conservation) but including small changes to the algorithms. The algorithm takes several steps:

1. It searches for sequence complementarity between the miRNA and the mRNA using a position-weighted local alignment algorithm. The last versions of miRanda require a perfect complementarity in the seed region and the 3' of the mRNA.
2. Once a match is found, the free energies of the miRNA-mRNA are computed (Vienna folding routines) [95].
3. It computes the conservation of the target sites in related genomes [81]. It has been described that conserved sequences are correlated with functionality (a more conserved site is more likely to be functional across species). The algorithm computes the inter-species conservation. Apart from that, a site has to be conserved in at least two species in order to be included in the database (with the exception of human and chimpanzee, whose sequences are too similar).

In summary, the program finds the energetically most favourable hybridisation sites of a small RNA (miRNA) in a large RNA (mRNA) and gives it a score and p value. Only pairs with a p value ≤ 0.05 are reported in the database.

Although there are more recent databases that compute miRNA-mRNA targets, as far as of our knowledge, MicroCosm is the only one that reports p values (Table 1.2), which is useful in some options of our algorithm. Interactions can be downloaded here: <http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5/>.

1.2.1.2 TargetScan

TargetScan [89, 97] is a database of miRNA-mRNA predicted interactions. It currently supports miRNA-mRNA interactions for human (*Homo sapiens*), mouse (*Mus musculus*),

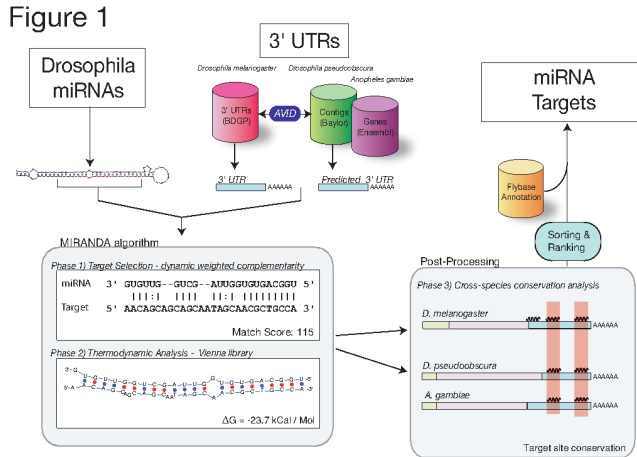


Figure 1.8: Figure extracted from www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5/info.html showing the miRanda algorithm.

worm (*Caenorhabditis elegans*), fly (*Drosophila melanogaster*) and zebrafish (*Danio rerio*) and it is updated on a regular basis.

TargetScan predicts biological targets of miRNAs by searching for the presence of conserved 8mer, 7mer, and 6mer sites that match the seed region of each miRNA [89]. Then, the rest of the miRNA and surrounding region is used to compute the strength of the interaction. For example, a miRNA-mRNA interaction with one mismatch on the seed region can be ranked before another with a perfect match if the former has best complementary features.


In the current version of Targetscan, they used 74 miRNA transfection experiments to train the model, where they considered 26 features [97]. However, they found that not all these features were relevant to predict miRNA-mRNA interactions, or some of them might be redundant. The final model is a linear regression including 14 selected features related to miRNA accessibility (Table 1.3 [97]), which gives the context++ scores of each site. Features were standardised prior to model selection to avoid biases.

TargetScan webpage (<http://www.targetscan.org>) allows to download and/or access the context++ scores for human, mouse, worm, fly and fish (zebrafish) miRNA-target predictions, and on-line queries can be also easily made. These queries can be made using a gene target, a miRNA or a miRNA family. Figure 1.9 shows an example output, that uses *hsa-miR-17-5p* to search for miRNA-mRNA interactions.

By default, TargetScan groups the targets by family. By default, the output page gives

www.targetscan.org/cgi-bin/targetscan/vert_71/targetscan.cgi?species=Human&mir_sc=mirR-17-5p/20-5p/93-5p/106-5p/519-3p

Apps Joon Lee a Twitter



Release 7.1: June 2016 Agrawal et al., 2015

Human | miR-17-5p/20-5p/93-5p/106-5p/519-3p
 1384 transcripts with conserved sites, containing a total of **1649** conserved sites and **898** poorly conserved sites.
 Genes with only poorly conserved sites are not shown. [View top predicted targets, irrespective of site conservation]
 Table sorted by cumulative weighted context++ score [Sort table by aggregate P Ct]
 The table shows at most one transcript per gene, selected for being the most prevalent, based on 3P-seq tags (for the one with the longest 3' UTR, in case of a tie). [Download table]

Target gene	Representative transcript	Gene name	Number of 3P-seq tags supporting UTR + 5'	Link to sites in UTRs	Conserved sites			Poorly conserved sites			7mer-m8	7mer-A1	7mer-m8	7mer-A1	Gene 7mer sites	Representative miRNA	Cumulative miRNA context++ score	Total context++ score	Aggregate P Ct	Previous TargetScan publication(s)
					total 7mer	7mer-m8	7mer-A1	total 7mer	7mer-m8	7mer-A1										
PDCD1LG2	ENST00000397745.2	programmed cell death 1 ligand 2	64	Sites in UTR	2	2	0	0	1	0	0	0	0	0	0	0	0	0.83	0.96	2009, 2011
GPR6	ENST00000275169.3	G protein-coupled receptor 6	5	Sites in UTR	2	2	0	0	0	0	0	0	0	0	0	0	0	0.97	0.89	2009, 2011
CTDSP1	ENST00000443903.2	CTD (carboxy-terminal domain, RNA polymerase II, polypeptide A) small phosphatase-like	372	Sites in UTR	1*	0	0	0	0	2	0	1	1	0	0	0	0	< 0.1	0.11	2011
GPR137C	ENST00000321662.6	G protein-coupled receptor 137C	110	Sites in UTR	3	1	2	0	0	1	0	0	1	1	0	0	0	> 0.99	0.84	2009, 2011
HN1	ENST00000476256.1	hematological and neurological expressed 1	15013	Sites in UTR	1	1	0	0	0	0	0	1	0	0	0	0	0	0.90	0.90	2007, 2009, 2011
NAGK	ENST00000418807.3	N-acetylglucosaminase	69	Sites in UTR	1	1	0	0	0	1	0	0	0	0	0	0	0	0.95	0.85	2005, 2007, 2009, 2011
ZBTB7A	ENST00000322357.4	zinc finger and BTB domain containing 7A	37	Sites in UTR	4	0	3	1	0	0	0	0	0	0	0	0	0	0.99	0.84	2007, 2009, 2011
ENPP5	ENST00000371383.2	ectonucleotide pyrophosphatase/phosphodiesterase 5 (putative)	15	Sites in UTR	2	2	0	0	0	4	0	1	3	1	0	0	0	> 0.99	0.77	2005, 2007, 2009, 2011
HAUS9	ENST00000233669.5	HAUS augmin-like complex, subunit 8	827	Sites in UTR	1	1	0	0	0	0	0	0	0	0	0	0	0	0.80	0.72	2005, 2007, 2009, 2011
RAB22A	ENST00000244040.3	RAB22A, member RAS oncogene family	484	Sites in UTR	2	0	2	0	1	0	1	0	0	3	0	0	0	0.93	0.71	2005, 2007, 2009, 2011
BTG3	ENST00000339775.6	BTG family, member 3	3656	Sites in UTR	1	1	0	0	0	0	0	0	0	0	0	0	0	0.91	0.70	2005, 2007, 2009, 2011
PKD2	ENST00000237596.2	polycystic kidney disease 2 (autosomal dominant)	915	Sites in UTR	2	2	0	0	1	0	1	0	0	0	0	0	0	> 0.99	0.69	2005, 2007, 2009, 2011
CCL1	ENST00000225942.3	chemokine (C-C motif) ligand 1	6	Sites in UTR	1	1	0	0	0	0	0	0	0	1	0	0	0	0.81	0.68	2009, 2011
IRF9	ENST00000396864.3	interferon regulatory factor 9	390	Sites in UTR	1	1	0	0	0	0	0	0	0	0	0	0	0	0.66	0.66	2011
CYBRD1	ENST00000375252.3	cytochrome b reductase 1	4467	Sites in UTR	1	1	0	0	3	0	0	0	0	0	0	0	0	0.14	0.71	2006, 2011
BRMS1L	ENST00000216907.7	breast cancer metastasis-suppressor 1-like	230	Sites in UTR	4	2	1	1	3	0	3	0	0	0	0	0	0	> 0.99	0.93	2007, 2009, 2011
EZH1	ENST00000428926.2	enhancer of zeste homolog 1 (Drosophila)	70	Sites in UTR	2	2	0	0	0	0	0	0	0	0	0	0	0	> 0.99	0.65	2005, 2007, 2009, 2011
PTH1H	ENST00000395872.1	parathyroid hormone-like hormone	32	Sites in UTR	1	1	0	0	0	0	0	0	0	0	0	0	0	0.91	0.64	2011
NKIRAS1	ENST00000388759.3	NFKB inhibitor interacting Ras-like 1	123	Sites in UTR	1	1	0	0	0	0	0	0	0	0	0	0	0	0.94	0.63	2005, 2007, 2009, 2011

TargetScan Release 7.1
 Questions: wibr-bonformatc@wi.mit.edu
 Whitehead Institute for Biomedical Research
 Compatibility

Figure 1.9: Targetscan output, with Target gene and Representative miRNA columns highlighted. All the information can be downloaded in .txt or .xlsx format. Predictions can also be ranked according to only PCT (site conservation) feature [22].

Abbreviation	Description	Citation(s)
TA_3UTR	Number of sites in all annotated 3' UTRs	[98, 99]
SPS	Predicted thermodynamic stability of seed pairing	[99]
sRNA1	Identity of nucleotide at position 1 of the sRNA	
sRNA8	Identity of nucleotide at position 8 of the sRNA	
site8	Identity of nucleotide at position 8 of the site	
local_AU	AU content near the site	[20, 83]
3P_score	Supplementary pairing at the miRNA 3' end	[83]
SA	Predicted structural accessibility: \log_{10} (Probability that a 14 nt segment centred on the match to sRNA positions 7 and 8 is unpaired)	
min_dist	Minimum distance: \log_{10} (Minimum distance of site from stop codon or polyadenylation site)	[83, 88, 100]
PCT	Probability of site conservation, controlling for dinucleotide evolution and site context	[22]
len_ORF	\log_{10} (Length of the ORF)	
len_3UTR	\log_{10} (Length of the 3' UTR)	[101]
off6m	Number of offset-6mer sites in the 3' UTR	[22]
ORF8m	Number of 8mer sites in the ORF	[89, 102]

Table 1.3: Table adapted from Agarwal et al. 2015 [97]. Features used to compute the context++ score for a miRNA-mRNA pair according to TargetScan database since version 7.

all the mRNA targets of the miR-17/92 family because, as miRNAs from the same family share the seed sequence (one of the most important features to determine a miRNA-mRNA interaction), they are likely to share most of their targets. However, specific miRNA-mRNA pairs can be obtained using *Target gene* and *Representative miRNA* columns.

1.2.1.3 MiRSVR

MiRSVR is based also on miRanda algorithm (Figure 1.8) but implements small modifications from 2010 and ranks the miRNA-mRNA pairs according to miRSVR scoring method [87].

The SVR (support vector regression) scoring method applied on miRSVR database is similar to context++ in TargetScan, which uses a weighted sum of features. The selected features are divided into three categories:

- **Duplex features:** base pairing at the seed region, and 3'end of the miRNA.

- **Sequence features:** A/U composition near the target sites and secondary structure accessibility.
- **Global features:** length of the UTR, relative position of the target site in the UTR and conservation score.

The scores can be interpreted as an empirical probability of downregulation, which provides a meaningful guide for selecting a score cutoff. Scoring of the genes with multiple target sites is done by simple addition of the individual target scores. The current release includes all target site predictions which have either a 6-mer or better seed site, or a mirSVR score ≤ -0.1 .

On the last release, training data consisted of 18 samples divided into 9 paired samples (transfected versus control at 12 hours, and transfected versus control at 24 hours) for a total of 9 miRNAs. Selected miRNAs were: *hsa-miRNA-7*, *hsa-miRNA-9*, *hsa-miRNA-122a*, *hsa-miRNA-128a*, *hsa-miRNA-132*, *hsa-miRNA-133a*, *hsa-miRNA-142*, *hsa-miRNA-148b* and *hsa-miRNA-181a* (GSE8501 [83]). Test data consisted on 71 samples transfected individually with one of the 16 following miRNAs in different cell conditions: *hsa-miRNA-16*, *hsa-miR-106b*, *hsa-miR-15a*, *hsa-miR-15b*, *hsa-miR-103*, *hsa-miR-200a*, *hsa-miR-141*, *hsa-miR-106b*, *hsa-miR-103*, *hsa-miR-192*, *hsa-miR-215*, *hsa-miR-17-5p*, *hsa-miR-20*, *hsa-let-7c*, *hsa-miR-195* or *hsa-miR-107*. Small interfering RNAs were also included in the study (GSE6838 [103]).

1.2.1.4 MiRDB

MiRDB [90, 104] is an online database for miRNA target prediction and functional annotations. All the targets in miRDB were predicted MirTarget algorithm [105]. Supported organisms are human, mouse, rat, dog and chicken.

MirTarget, like TargetScan and miRSVR, is another algorithm that takes advantage of support vector machine learning to make feature selection. The features enriched in the downregulated genes respect to normal genes were used as predictors of miRNA-mRNA interactions. They evaluated 50 features (Table 3 of their paper [104]), but the most important ones are:

1. **Seed site conservation:** number of species (human, mouse, rat, dog and chicken) where RNA sequences are conserved. The more conserved they are, the more likely are to be target sites.

2. **Target site location in UTR:** distance expressed in bases from the end of 3'-UTR. Target sites are likely to be found on the first 200 nucleotides, while they are rarely found more than 800 nucleotides far from the 3'-UTR.
3. **GC content** either on the target or specific regions also helps to determine if a gene is a target or not, as well as UG, AG, GC, UA and other counts.
4. **Free energy** of the seed sequence binding, and specific combinations of miRNA sequences are also ranked on the list (for example an A on the first position of the miRNA).

The used training data was GSE6838 [103], which included miRNAs from the hsa-miR-16 family in 71 samples. MiRDB score ranges from 0 to 100, being 0 not likely a target, and 100 likely a target; proposed cutoff is 50.

Test data was generated by transfecting a mimic of *hsa-miR-124* in HepG2 cell line [106]. Overexpression of *hsa-miR-124* predicted targets were evaluated at times 4, 8, 16, 24, 32, 72 and 120 hours. They found that most of the targets overexpressed before 72h. Thus, validated predicted targets have a median score higher than 50 (the proposed cutoff), while non validated targets (those that does not overexpressed during the experiment), had a median score around 30.

MiRDB has an additional feature is that lists the called "functional miRNAs" [90]. A miRNA is considered "functional" if it has two or more characteristics: being reported in PubMed, having orthologous miRNAs, having been detected in RNA-seq experiments or having been classified as High Confidence by miRBase. Users can check PubMed references and other links (miRBase, TarBase targets and genomic location) associated to each miRNA. Finally, as a recent update, users may provide their own sequences for customised target prediction.

1.2.1.5 MiRWalk and databases of validated interactions

MiRWalk [91, 107] is a widely used database that gathers the information from other databases in a customised way. The last version (updated on February 2016) includes information about 12 prediction databases, plus pathway information of associations with 597 KEGG, and 522 Wiki pathways; functional information from 18394 Gene Ontology and 456 Panther terms; and 2035 disease ontologies (DO), 6727 Human Phenotype ontologies (HPO) and 4980 OMIM disorders. It can be accessed here <http://zmf.umm.uni-heidelberg.de/apps/zmf/mirwalk2/>.

Although it is probably one of the most comprehensive databases found (it also allows to find homologies between 15 species), the drawback is that it only allows to specify queries that contain less than a hundred miRNAs each time (human miRNome have actually more than two thousand miRNAs) and databases are not instantly updated.

MiRWalk is also known for its **Validated targets** module, which offers an exhaustive list of validated miRNA-mRNA interactions. Validated interactions can be found using different criteria, but can be differentiated between automatic ones or manually curated. Lee et al. offered a good review of them [108].

MiRWalk validated miRNA-mRNA pairs are found through text mining on PubMed articles, similarly to miRTarBase [109] (another database of validated miRNA-mRNA interactions). MiRWalk output is a table which includes the name of the miRNA, the gene name, plus their respective miRNA ID and EntrezID and the PubMed ID of the article that refers this interaction. Queries can be made from a miRNA or a mRNA target. Plus, MiRWalk also offers link to miRNA homologues, gene functions, etc.

TarBase (sometimes called DIANA-TarBase or DIANA) [110, 94] is another database that offers miRNA-mRNA interactions. It collects information from PubMed articles, plus low and high-throughput experiments (in this case, all the interactions are reported, not only the ones reported in the paper), including positive and negative associations, which makes it the most exhaustive database of validated miRNA-mRNA interactions. Plus, it offers the option to filter the experimental methodology used, and check the experimental conditions of the experiment (such as cell type and treatments) and it includes a prediction score. There are a lot of experimental methodologies reported, but some of the most relevant are immunoprecipitation, luciferase reporter assays (one of the most reliable ways to validate a miRNA-mRNA interaction), RNA-seq, microarrays and immunofluorescence experiments. TarBase has more than half million miRNA-mRNA interactions from 24 organisms (including human, mouse, fly, worm, rat, *Arabidopsis thaliana*, etc). Although it is freely available on the website app, downloaded versions are only accessible after a register on their database and a positive authorisation. The current URL is: <http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=tarbase/index>.

miRecords [92] and miR2Disease [93] on the other hand, are databases of validated interactions manually curated. Each miRNA-mRNA interaction is linked to the PubMed article or articles that have found the interaction and additionally, miR2Disease also links each miRNA-mRNA interaction to a disease (such as colon cancer or heart disease).

MiR2Disease has 3273 miRNA-mRNA interactions from 349 miRNAs associated with 163 diseases, but has not been updated since its creation. MiRecords includes information from other species, and has a total of 2608 unique human miRNA-mRNA interactions (plus 857 non-human miRNA-mRNA interactions), including 304 miRNAs and 1114 mRNAs. However, they have not been updated since 2013 and although they are useful for checking known interactions of a miRNA (as they are very accurate), they do not contain any of the recently discovered interactions and thus PubMed search of the miRNAs of interest cannot be avoided. The current URLs are: <http://c1.accurascience.com/miRecords/> (miRecords), <http://www.mir2disease.org/> (miR2Disease).

These databases of **validated interactions** offer a good starting point to check validated miRNA-mRNA interactions. However, all of them have some issues that must be taken into account. First of all, they are biased to previous experimental research. Consequently, miRNA-mRNA interactions related to "hot" miRNAs (they have been known for more time or have been related to a specific disease) have been more studied than others; and for that reason, they have more validated miRNA-mRNA interactions predicted. In this case, the lack of information on the other miRNAs, specially from the newly discovered miRNAs, it is not a direct evidence that they do not have any target, it is the direct consequence of the lack of tests performed on them. Secondly, each study and experimental procedure may have its own level of significance, so validated miRNA-mRNA interactions should always be checked with the primary sources. Finally, these databases may contain false positives in the sense that miRNA-mRNA interactions that have been found on a specific condition may not always be replicated in all the other cell conditions.

1.3 Expression-based methods for detecting miRNA-mRNA interactions

There are currently several methods and tools available to analyse miRNA-mRNA interactions from miRNA and mRNA expression data [111]. Figure 1.10 summarises the problem that is analysed with these methods: elucidate the miRNA-mRNA interactions from biological data.

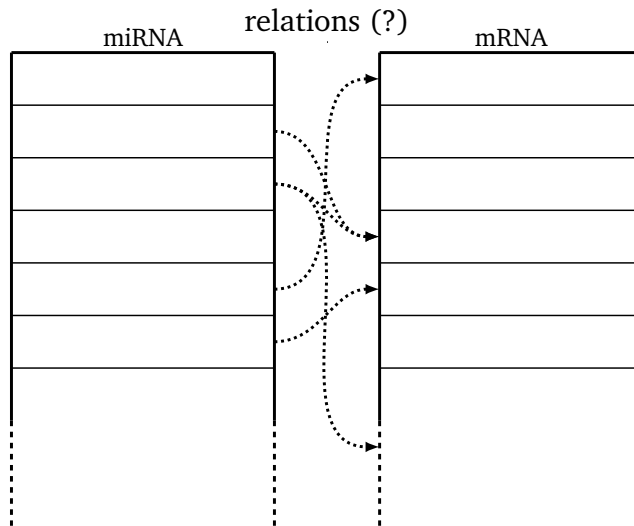


Figure 1.10: Summary of the problem: the miRNAs regulate the mRNAs, but the the exact interactions are not known (dotted lines). Each row represents the intensities for each miRNA/mRNA measured in several individuals (in most of the cases, there are the same individuals in each dataset). A miRNA can regulate many mRNAs, a mRNA can be regulated by many miRNAs, and some miRNAs or mRNAs can not interact. There are hundreds of miRNAs and thousands of mRNAs.

All the methods that will be described here are based on the same principle: in order to pick a miRNA-mRNA interaction, the miRNA and the mRNA must show some kind of negative relation between their expressions, and their sequences must be able to hybridise. However, the criteria used in each case is different, and it is possible to group them into several groups:

1.3.1 Methods considering up-down pairing

This approach can be used only when two biological conditions are compared. For example control samples vs. tumour samples. The method consists in picking the miRNA-mRNA pairs if the miRNA is upregulated and the mRNA downregulated or viceversa, and if the pair is predicted to interact in one database (or a combination of databases, depending on the case). This principle is summarised in Figure 1.11.

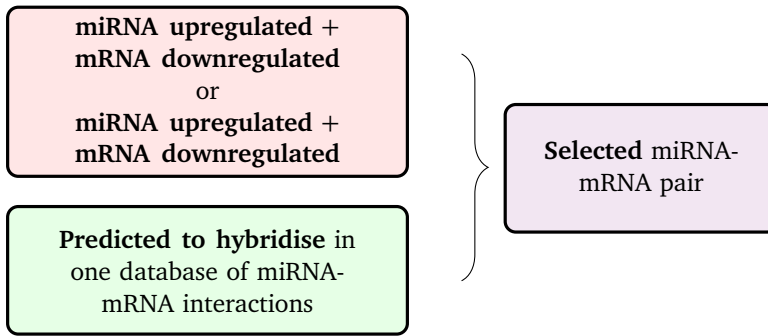


Figure 1.11: Classic up-down pairing.

The main drawback of this method is that it does not take into account the value of the expression in each sample. However, this approach can be useful when there are no common samples between the miRNA and mRNA dataset. Although the use of paired miRNA and mRNA samples is strongly recommended, sometimes is not possible to obtain paired samples in good conditions due to technical reasons or because of the nature of the study (for example, a retrospective study). Moreover, it is worth to take into account that in some cases the use of this method can lead to wrong (or at least non-optimal) solutions.

Figure 1.12 shows a case where two significant miRNA and mRNA, which would be picked according to this method, show no clear correlation. Although at first sight it seems to be a good pair to be validated, Figure 1.12(c) may be indicating us that there is no clear relation between the two variables, and their expression behave in opposite direction on the tumour due to different reasons.

Another drawback of the method is that it is difficult to use this method when there are no groups (for example only case samples) or there are more than two groups, which in this case it is difficult to assign a label such as "upregulated" or "downregulated", and the comparisons should be done one by one.

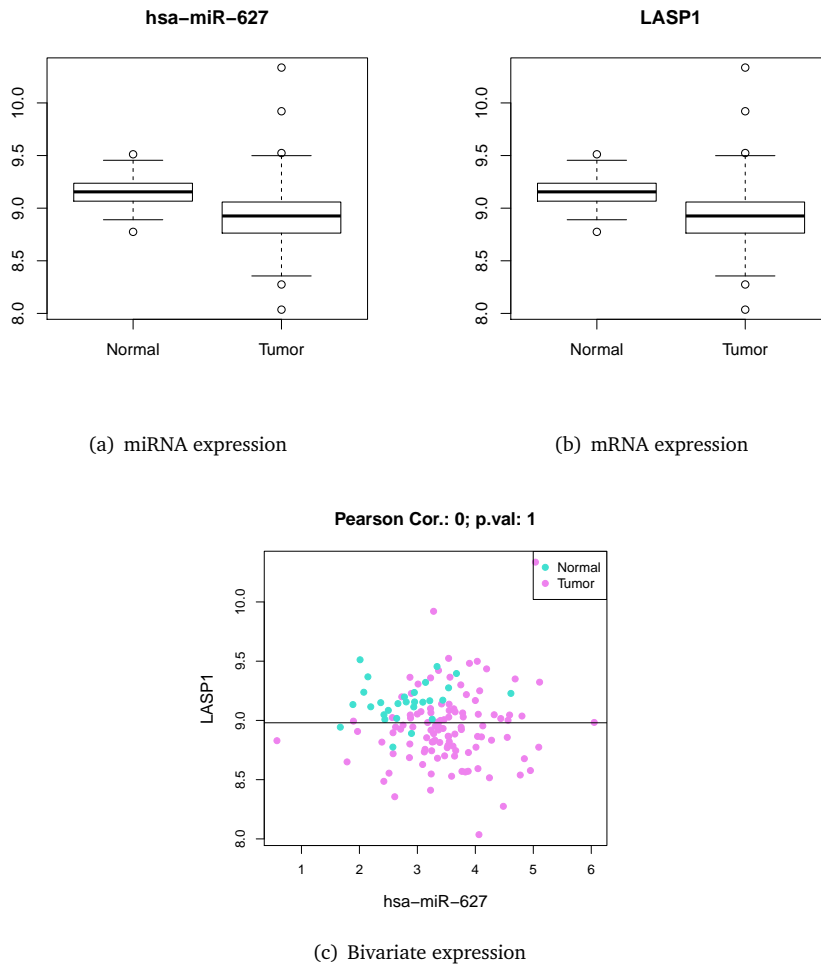


Figure 1.12: Data extracted from GSE21032 [112]. Which contains 139 samples (111 tumour taken after radical prostatectomy, 28 control) with paired miRNA and mRNA expression. We analysed 373 human miRNA and 21494 mRNA. Limma analysis using standard options showed 185 deregulated miRNAs with FDR < 0.05 (80 upregulated and 105 downregulated); and 6512 mRNAs with FDR < 0.05 (3396 upregulated and 3116 downregulated). **Subfigure (a)** shows the expression of *hsa-miR-627* between normal and tumour group. **Subfigure (b)** shows the expression of *LASP1* between normal and tumour group. **Subfigure (c)** shows the bivariate expression of *hsa-miR-627* and *LASP1* (X and Y-axis: normalised log2 expression).

Nonetheless, because of its simplicity and clarity, a lot of studies and softwares use this method to analyse miRNA-mRNA interactions. In this group programs like miRTRail [113], Ingenuity Pathways Analysis (IPA) [114] and related publications can be found.

MiRTrail [113] is a web application <http://mirtrail.bioinf.uni-sb.de/> that uses the criteria described in Figure 1.11 to give a list of miRNA-mRNA interactions. The first versions of the web allowed to use only human data and microCosm database (allowing to filter microCosm according to p value). In the last update, miRTrail offers support for mouse and zebrafish miRNA-mRNA interactions, and allows to use a custom miRNA-target database. MiRTrail also allows to plot small networks with the predicted pairs. MiRTrail analysis has been used and cited in several publications [115, 116].

IPA (Ingenuity Pathways Analysis) [114] is a commercial software aimed to deal with omics data analysis. It is able to deal with RNA-seq, small RNA-seq, microarrays including miRNA and SNP, metabolomics, proteomics, and small scale experiments. It is specialised on the interpretation of the data regarding the identification of pathway key regulators as well as the prediction of downstream effects on biological and disease processes. It also provides targeted data on genes, proteins, chemicals, and drugs; and allows to build and interpret interactive models of experimental systems.

Regarding miRNA-mRNA analysis, although it provides a very exhaustive system to interpret miRNA effects (based on their targets) and has been broadly used [117, 118, 119], miRNA-mRNA predictions are based only on FoldChange differences, and uses TargetScan and TarBase databases. Moreover, the fact that is not free makes it not accessible for the whole scientific community.

1.3.2 Methods considering correlation

Figure 1.13 shows the idea below this kind of approach. Equally to classic approach (Figure 1.11) in this case two premises should be accomplished to be predicted as good miRNA-mRNA interaction: 1) evidence of biological relation, in this case via negative correlation, and 2) evidence of physical interaction, in this case also via a database(s) of predicted miRNA-mRNA interactions.

In contrast to the previous approach, Figure 1.14 shows two miRNA-mRNA pairs from the same dataset, but in this case, the selected miRNA-mRNA pairs show deregulation and a clear negative correlation. These two pairs would also be found according to "Classical" up & dw pairing, but we can hypothesise that these pairs are more likely to be true than the shown in Figure 1.12.

One strength of this method is that it can be extrapolated to more than one group (or

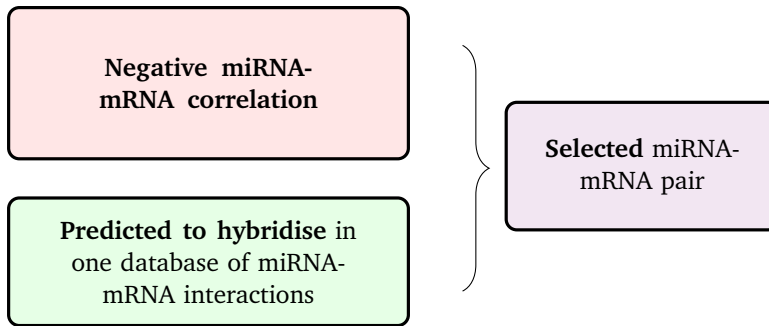


Figure 1.13: Idea below correlation.

even o groups): in both Figures 1.14(a) and 1.14(b), although in Normal samples may be not evident, a significant negative correlation is observed in Tumour samples alone. The limitation of this method is that it needs paired miRNA and mRNA samples.

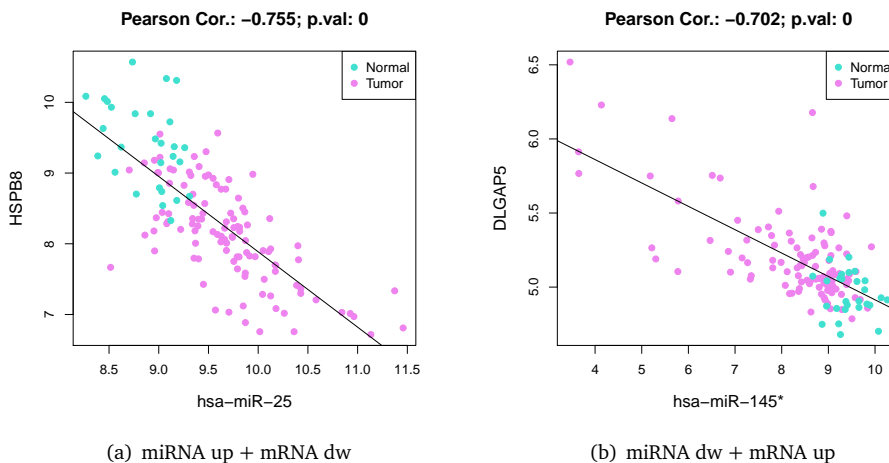


Figure 1.14: Data extracted from GSE21032 [112] (See Figure 1.12). 1.14(a): bivariate expression of *hsa-miR-25* (upregulated in cancer samples) and HSPB8 (downregulated in cancer samples). 1.14(b): bivariate expression of *hsa-miR-145** (downregulated in cancer samples) and DLGAP5 (upregulated in cancer samples). X and Y-axis: normalised log2 expression.

We were able to find several articles extremely useful for understanding the concepts used. However, they have not developed any statistical tool to reproduce systematically the procedure used, or, if they have developed one, their objectives were not the same as the described in this work. Furthermore, there are also different ways of performing the integration between the correlation and the databases. The most relevant articles

and tools that take into account the value of correlation between miRNA-mRNA pairs are commented below:

Gade et. al [120]. This article describes the basic idea of p value combination. The authors have used the data integration specifically for analysing the survival of the patients with prostate cancer and make a clinical outcome predictor. They used Pearson correlation to measure the correlation between each pair of miRNA and mRNA, and MicroCosm database for obtaining information about the physical interaction. Thus, they combine both p values into one using Stouffer correlation [121]. Using this methodology, plus a Cox model for the survival analysis, they were able to select the best features to predict the probability of biochemical relapse of prostate cancer. They showed that combined miRNA-mRNA analysis made a better risk assessment of their set of patients than the one that would be obtained using non-integrated analysis of miRNA and mRNA data [120].

Peng et. al [122]. The authors used a similar idea for combining miRNA and mRNA data, but they do not use p value combination: they use a simple intersection (final miRNA-mRNA pairs are those that fulfil both initial hypotheses), which has the advantage that can be applied for other miRNA-mRNA target prediction databases. By using this approach, other databases apart from microCosm can be selected. In their study, Peng and collaborators chose miRBase (miRanda targets, the basis of miRSVR) and TargetScan. Final miRNA-mRNA interactions were those predicted in at least one database and negatively correlated (they used permutation-based tests to estimate the false detection rate).

They focused the interpretation of the results in the description of regulatory modules (a group of miRNAs which regulate together a set of targets), that were partially interpreted using IPA software [114]. They also describe in detail some of the regulatory modules found and their relation with hepatitis C viral infection in human livers.

MAGIA [123] is a software representative of this approach. It takes into account the correlation between miRNAs and mRNAs for the data integration. It is freely accessible from <http://gencomp.bio.unipd.it/magia/start>. Figure 1.15 shows the output of the analysis, which includes:

- MiRNA-target prediction: it allows to use miRanda, PITA and TargetScan databases, and union and intersection combinations of them.
- Integrated analysis of miRNA-mRNA data: Pearson Correlation, Spearman Correlation, mutual information and Bayesian analysis when paired samples are available,

and chi-square test –searching for opposite FoldChanges– when no paired samples are available.

- Functional annotation with links to external databases: miRBase, Entrez Gene, PubFocus (either for miRNA or mRNA), miR2Disease and EBImed (either for miRNA or mRNA).
- Enrichment analysis using DAVID (an extensive database that has a gene enrichment analysis tool) [124, 125].
- Visualisation of the post-transcriptional regulatory network on the final screen and option to download all the data.

MAGIA article [123] and its update Magia² [126] have been cited more than 150 times, showing that the software has generated interest and has been a useful tool for the scientific community. However, it does not cover all the pipeline and options that we will present here. Apart from that, it does not offer the possibility of using p value combination, and the fact that it is a web-based software makes it not useful if we want to repeat the process several times or our dataset is large (it has to be uploaded each time).

1.3.3 Methods considering other procedures

TaLasso is another system for finding miRNA-mRNA interactions based on regression analysis proposed by **Muniategui et al.** [127].

Figure 1.16 shows the main idea of the procedure: a regression analysis including variable selection (LASSO –least absolute shrinkage and selection operator–, which will be further explained on methods) is only computed from the miRNA-mRNA pairs already predicted. Non-positive constraints are added to assure negative relations between the miRNAs and their targets. Intuitively, TaLasso selects the best interactions from the whole set of putative interactions.

TaLasso used microCosm, miRanda and DIANA, but other databases can be used. Functional analysis of the top-selected mRNAs showed biologically meaningful results these final interactions are enriched on validated miRNA-mRNA interactions from miRWalk or miRecords + TarBase [127].

The tool can be used either in R or Matlab, and there is also a web-based tool that can

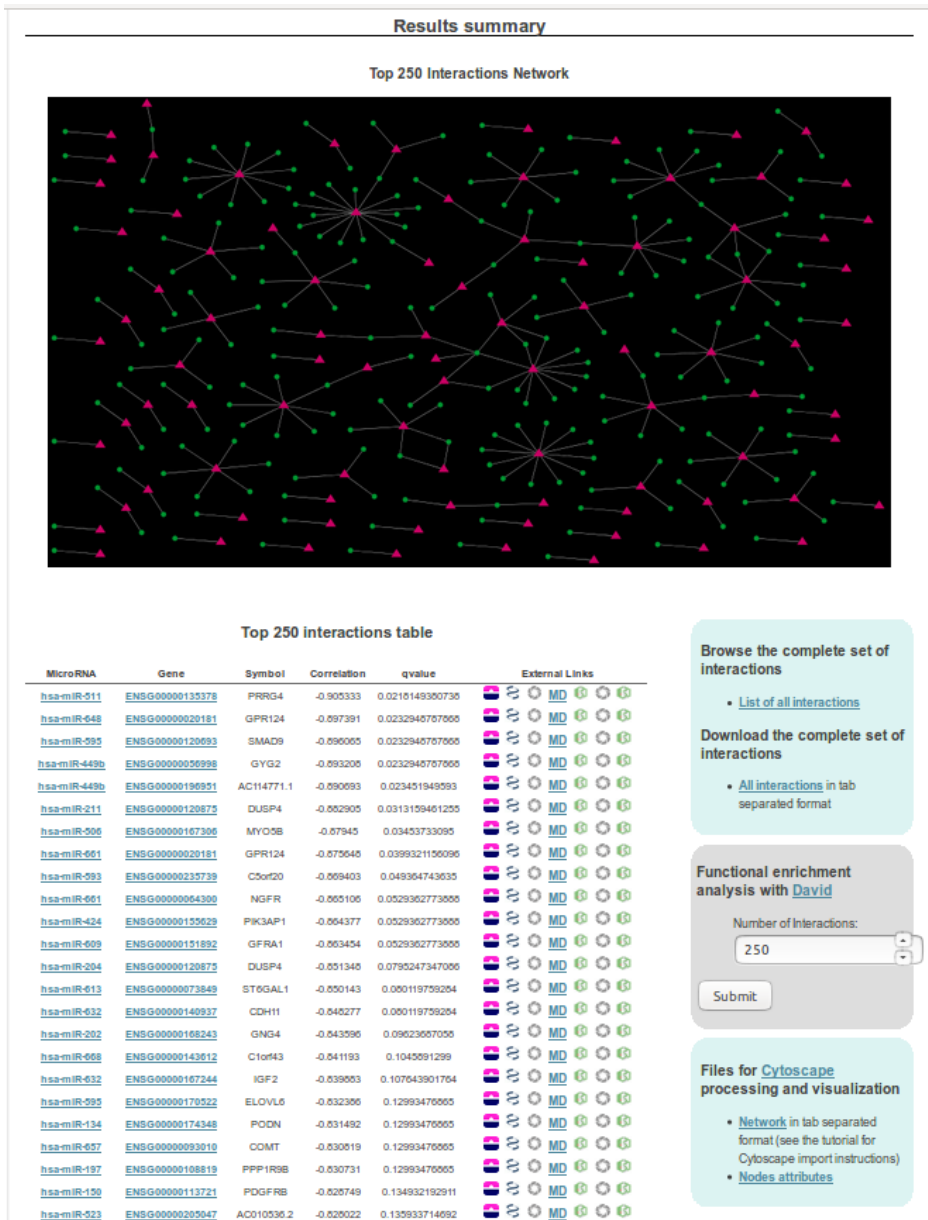


Figure 1.15: Caption of MAGIA output using provided example data by the authors.

be accessed from the following URL: <http://talasso.cnb.csic.es/>.

However, one of the drawbacks of TaLasso method is that only selects the most prominent miRNA-mRNA interactions and does not give information about the other ones.

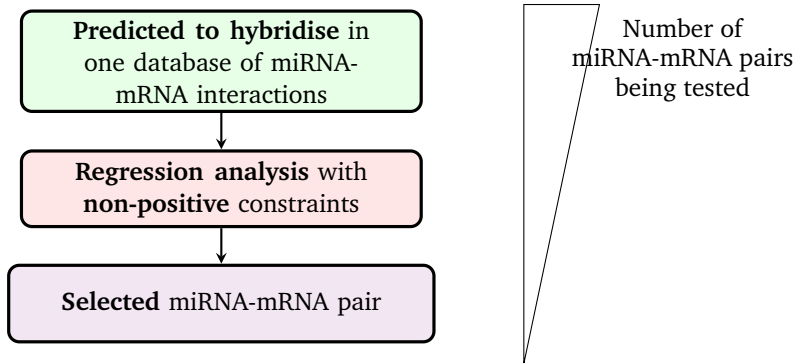


Figure 1.16: Idea below TaLasso algorithm.

Moreover, the whole output needs to be recalculated if the database is updated.

1.3.4 Other tools for working with miRNAs

R is a free software programming language and a software environment for statistical computing and graphics broadly used in data analysis. It appeared in 1993 as a derivation of the S language. It can be extended with the *packages*, which contain specific functions defined by the user. Bioconductor started in 2001 and is the project that stores thousands of R packages suitable for handling data from experiments in molecular biology [128].

At the starting point of this thesis, there were some R/Bioconductor packages [128] that dealt with miRNA data. However, none of them specifically covers the idea presented in this work:

- Rmir [129] is a package that couples data from miRNA and mRNA experiments and miRNA-mRNA prediction databases: miRSVR, TargetScan, miRanda, TarBase, miRDB and PicTar databases available on the last version. It allows to filter the miRNA-mRNA pairs using the absolute value of Pearson Correlation between the expression of the miRNA and the mRNA target. Apart from that, is focused on time-series experiments and not on case-control designs.
- CORNA [130, 131] is a package that uses publicly available information to explore miRNA-mRNA interactions and associated pathways. It uses miRanda miRNA-mRNA predictions, microarray data from GEO database, Gene Ontology terms from Biomart

and pathways from KEGG. Microarray expression information is only used for making informative plots but not for filtering miRNA-mRNA interactions. Pathways are tested using appropriated statistical tests.

- `miRNApath` [132] is expected to find enriched mRNA pathways for a given miRNAs list. It allows to find pathways from a list of miRNAs or a list of miRNA-mRNA associations, and to export the results and plot a heatmap summary of the pathways if more than two groups of samples are compared.
- `microRNA` [133] contains some functions that allow, among others, to find target regions for a given mRNA and miRNA sequences, having a function similar to miRNA-mRNA prediction databases.
- `miRNAConverter` [134] is an R/Bioconductor package that translates miRNA names between different versions of miRBase database.

Other tools that are not R packages:

- `MiRNApath` [135] is a web interface to explore metabolic pathways that are affected by miRNAs (URL: <http://lgmb.fmrp.usp.br/mirnapath/>). They use DIANA database [94] to predict miRNA-mRNA interactions. Although DIANA provides *validated targets*, these have not been validated in all the biological conditions, so they might be false positives in certain conditions. Thus, the database is biased in the sense that offers more information about the most studied miRNAs.

Chapter 2

Objectives

2.1 Main objective

- To **develop** a software suitable for the **integrative analysis of miRNA-mRNA interactions** in a specific biological context.

2.2 Secondary objectives

1. To identify the most appropriated methodology to predict potential miRNA-mRNA interactions.
2. To obtain a list of relevant miRNA targets for a given physiological situation.
3. To describe the miRNA-mRNA interactome of several digestive cancers (colorectal cancer, gastric cancer, esophageal cancer, liver cancer and pancreatic cancer)
4. To perform an integrative analysis of the above obtained miRNA-mRNA interactomes from several digestive cancers.
5. To validate some miRNA targets obtained from our tool in experimental models of pancreatic cancer.

Chapter 3

Materials & Methods

3.1 Data obtention & Preprocessing

3.1.1 STUDY 1 – MiRComb in five digestive cancers

3.1.1.1 Samples

For our first study, we have used data from The Cancer Genome Atlas (TCGA) [136]. Specifically, we downloaded RNA-seq and miRNA-seq data from 1645 samples among 5 different digestive cancers (colon adenocarcinoma (COAD); esophageal carcinoma (ESCA); liver hepatocellular carcinoma (LIHC); rectum adenocarcinoma (READ); stomach adenocarcinoma (STAD)) that had simultaneously miRNA-seq and RNA-seq data. All data was downloaded from TCGA data portal <https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.html> and have been processed with the same procedure.

3.1.1.2 Preprocessing

We selected only those samples that had paired miRNA and mRNA information and came from centers (properly identified with their corresponding Tissue Source Sites –TSS– codes) that collected more than one sample. Primary Solid Tumor and Solid Tissue Normal were used. MiRNAs and mRNAs with no id (on mirbase17) or median expression

< 10 raw counts were removed. Voom transformation [137] and quantile normalization were applied to allow using parametric methods (limma [138] and most importantly, Pearson correlation).

MD Anderson Cancer Center launched a website to explore TCGA data [139], which showed batch effects due to the Tissue Source Sites (TSS) origin of the samples of the selected cancers (Figure 3.1). In order to correct these batch effects according to TSS, we used ComBat function [140] implemented sva R package [141].

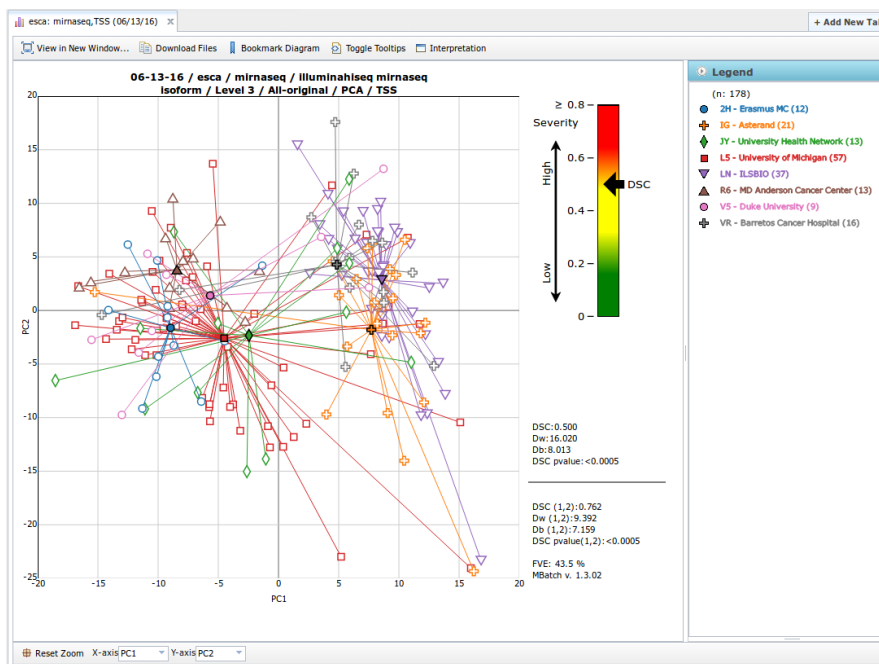


Figure 3.1: TSS batch effect in TCGA-ESCA miRNA samples.

3.1.2 STUDY 2 – MiRComb in pancreatic cancer

3.1.2.1 Samples

A set of 12 surgical pancreatic tissue samples (9 PDAC and 3 Healthy) from Hospital Clínic of Barcelona patients were included. The same samples were used for both genome-wide miRNA and mRNA profiling. Sample dissection was performed by experienced pathologists who split tissue samples in two different parts: one for gene expression analysis and

the other for diagnostic confirmation. Pancreatic tissues were kept on dry ice at all times during handling, flash frozen in liquid nitrogen and stored at -80°C until RNA isolation. Healthy pancreatic samples correspond to the healthy tissue of patients who underwent surgery for other reasons (i.e., ampulloma or neuroendocrine tumours). None of the patients with PDAC had received chemo or radiotherapy before sample collection.

This study was approved by the Institutional Ethics Committee of Hospital Clínic of Barcelona (March 27, 2008) and written informed consent was obtained from all patients in accordance with the Declaration of Helsinki. Total RNA including miRNA was isolated from frozen macrodissected tissues using the miRNeasy Mini Kit (Qiagen, Valencia, CA, USA), according to the manufacturer protocol. RNA concentrations and purity were evaluated using NanoDrop 1000 Spectrophotometer (Wilmington, DE, USA) and RNA quality was determined by Bioanalyzer 2100 (Agilent, CA, USA).

3.1.2.2 Next Generation Sequencing

The starting amount was $1\mu\text{g}$ of total RNA, and the preparation protocol was performed according to the manufacturer's recommendations. Small RNA (18-30nt long) was isolated using an polyacrylamide gel electrophoresis. Adapters were included on both 3' and 5' ends. The high-throughput sequencing of the cDNA was done in a 38 bp single-end read run on an Illumina Genome Analyzer Iix (Illumina, California, USA). Image analysis and base calling was performed with the Illumina Genome Analysis pipeline software version 1.5.1.

Data from the high-throughput sequencing were obtained in FASTQ format, 1 data file per sample. Samples contain a median of 6530636 reads (IQR: 5839068–7209569). Data quality was checked using FastQC [142], which confirmed that the quality is adequate to continue the analysis, as most of the base calls are of quality > 30 , which means that the probability of error is 10^{-3} per call (Figure 3.2).

To obtain the number of counts, the sequencing adaptors were clipped and removed using the FASTX-Toolkit [143], allowing no mismatches for adaptor identification. The remaining sequencing data were collapsed and counted into groups of identical sequences. The curated sequences were processed with miRDeep2 [144] to identify miRNAs an obtain count data from the miRBase (release 18, based on *Homo sapiens* hg18 genome reference) data repository, allowing for 1 mismatch. This system was able to identify counts for 1733 miRNAs.

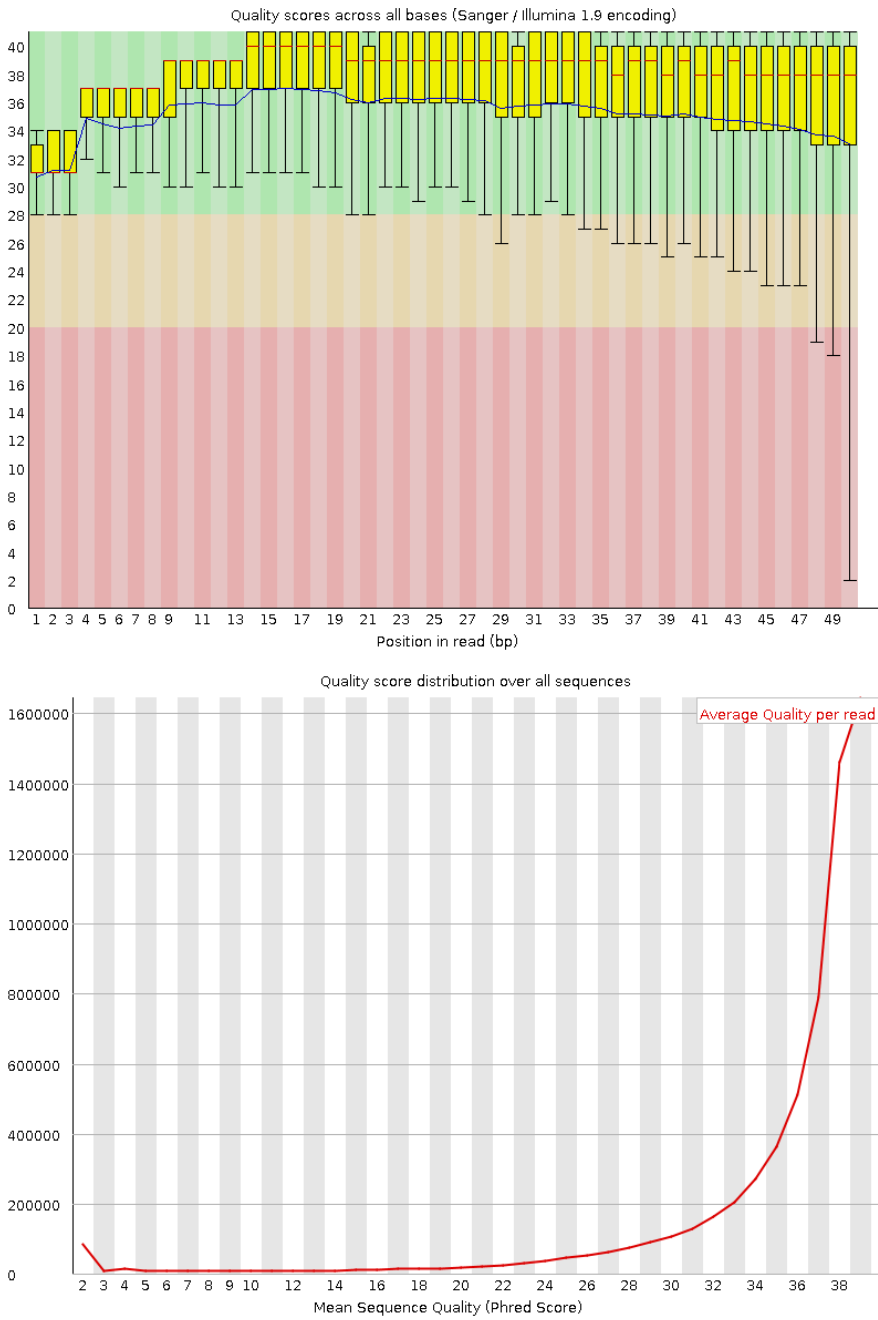


Figure 3.2: Selected images from FastQC report of one representative sample (177_L001). Image from *per base quality* (top) and *per sequence quality* (bottom).

3.1.2.3 Gene expression arrays

Matched genome-wide mRNA profiling was analyzed by microarray technology with Human Genome U219 Gene Expression Arrays (Affymetrix, Santa Clara, CA, USA) and normalized according to `limma` procedure [138].

3.1.2.4 Cell culture

Human pancreatic cancer cell line PANC-1 was obtained from European Collection of Cell Cultures (ECACC, Wiltshire, UK) and cultured in Dulbecco's modified Eagle's medium (GIBCO, Thermo Fisher Scientific, Waltham, MA, USA) supplemented with 10% fetal bovine serum (GIBCO, Thermo Fisher Scientific) and 1% penicillin/streptomycin (GIBCO, Thermo Fisher Scientific). Cells were incubated at 37°C and 5% CO₂ in a humidified chamber.

3.1.2.5 CRISPR/Cas9 targeting of miR-21 in PANC-1 cells

gRNA design The gRNA of miR-21 was designed using the "CRISPR design tool" from Feng Zhang Lab (<http://crispr.mit.edu/>). We chose a PAM sequence in the pre-miR-21 region and select a 20-bp sequence upstream as the targeting sequence (5'-TCAT-GGCAACACCAGTCGAT-3'). Oligonucleotides of the indicated sequence were purchased from IDT (Leuven, BE), annealed and cloned into the plentiCRISPRv2 vector following Lentiviral CRISPR Tool box instructions from Zhang Lab deposited to Addgene.

Verification of gRNA-mediated genome cleavage HEK293T cells were transfected with the plentiCRISPRv2 containing miR-21 gRNA by CalPhos mammalian transfection kit (Clontech, Takara Bio Company Inc., Mountain View, CA, USA). Cells were treated with 4 μ g/ml puromycin for one week. Next, genomic DNA from transfected and wild-type cells was isolated and submitted to PCR amplification of a 555 bp fragment that encompasses miR-21 region using the following primers: Fwd: 5'- CCACACTCTGTCGTATCTGTG-3', Rev: 5'- AAGTGCCACCAGACAGAAGG-3'. PCR fragments were subjected to SURVEYOR nuclease assay (Transgenomic) and resolved on 1.5% agarose gel. Mutations were confirmed by DNA sequencing.

Generation of miR-21-deleted PANC-1 cells Lentiviral particles were generated by transfection of vectors plentiCRISPRv2miR-21gRNA or plentiCRISPRv2-Control (for control cells), pVSV-G and pCMV Δ 8.91 into HEK293T by CalPhos mammalian transfection kit. At 48h the viral supernatant was collected, filtered and added to PANC-1 cells. Three days after transduction, cells were selected in 8 μ g/ml puromycin for one week. Next, limiting dilution was carried out to generate individual clones from PANC-1 infected with miR-21gRNA cells and three weeks later several clones were analyzed for DNA mutation and *hsa-miR-21* expression.

3.1.2.6 RNA extraction and Target expression analysis by qRT-PCR

Total RNA was isolated from cell cultures using the miRNeasy Mini Kit (Qiagen, Valencia, CA, USA), according to the manufacturer protocol. The final elution volume was 30 μ L. RNA concentrations and purity were evaluated using NanoDrop 1000 Spectrophotometer (Wilmington, DE, USA).

The expression of PDCD4 and BTG2 was analyzed by qRT-PCR using TaqMan High Capacity cDNA Reverse Transcription Kit (Applied Biosystems Inc., Foster City, CA, USA). A two-step protocol involves reverse transcription, followed by a real time PCR with TaqMan probes. Briefly, 1 μ g total RNA was used per reverse transcription reaction performed in final volume of 10 μ L (5 μ L RNA, 0,4 μ L of 100mM dNTPs, 0,5 μ L of Multiscribe Reverse Transcriptase (50U μ L⁻¹), 1 μ L of 10X RT buffer, 0,5 μ L of RNase inhibitor (20U μ L⁻¹), 1 μ L 10x RT random primers and 1,6 μ L Nuclease-free water) and incubated for: 10 minutes, 25°C; 120 minutes, 37°C; 5 minutes, 85°C; hold at 4°C. The 10 μ L PCR mixture included 4 μ L cDNA, 6 μ L of TaqMan 2X Universal PCR Master Mix with no AmpErase UNG and 0,5 μ L of TaqMan 20X MicroRNA Assay. PCR reactions were incubated in a 384-well optical plate and run on the Viia7 Real-Time PCR System (Applied Biosystems Inc.) as follows: 95°C for 10 min and 50 cycles of 95°C for 15 sec and 60°C for 1 min. All specimens were amplified in duplicates. Amplification data was normalized against Cyclophilin as endogenous control. Ct values were calculated from automatic threshold. No template controls showed any amplification.

3.2 Design of a new tool for analysing miRNA-mRNA interactions: miRComb

The miRComb pipeline (the work that we present here) is mainly based on the work of S. Gade (2011) and [120] W. Zhang (2012) [145]. Figure 3.3 summarises the workflow implemented in our tool: biological information (red) is combined with theoretical information (green) in order to achieve an overall conclusion (violet) and interpret the interactions using functional data analysis.

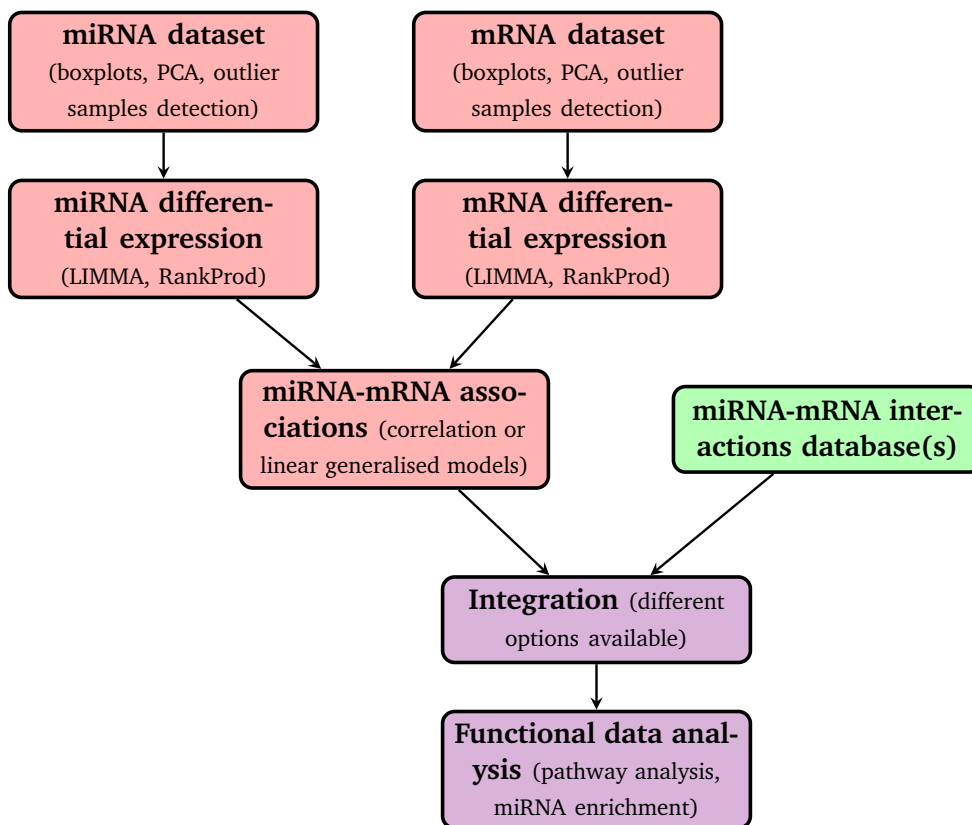


Figure 3.3: Outline of the pipeline of miRComb R package.

The exact methods that we have implemented on miRComb R package are described in the following sections.

3.3 Differential Expression

Differential expression can be assessed in different ways, from the traditional ones to the most recent ones (Table 3.1).

	Parametric	Non parametric
Simple testing	T-test	Wilcoxon-test
Microarrays	limma [138]	RankProd [146]
NGS	DESeq [147], edgeR [148], voom [137]	

Table 3.1: Methods used for testing differential expression.

The first two methods (T-test and Wilcoxon-Test) are the classic methods for testing differences on means (or medians, for non-parametric) but they are not the most suitable ones for analysing gene (or miRNA) expression. Gene expression experiments must take special consideration as usually there are far more more genes to test than samples available.

The data we obtain from microarrays are intensity measures of fluorophores attached to complementary DNA (cDNA) from the RNA sample of interest. Each spot represents a cell on the raw data matrix: these intensities are centered in one value and show variation, but they are approximately symmetrical and their values range between specific limits. So, the distribution is considered to follow a normal distribution (Figure 3.4(a) shows an example of microarray data). Based on that, some methods have been developed such as *limma* [138] and *RankProd* [146] to analyse this type of data and deal with the mentioned drawbacks (small sample size and a lot of features tested each time). These methods show better performance [149] than traditional ones as they are able to use the most of the data available.

Regarding to Next Generation Sequencing (NGS) experiments: in a typical RNA-seq experiment, a sample of RNA is converted to a library of cDNA fragments and then sequenced on a high-throughput commercially available platform (the most common is Illumina). The raw data comprise the sequences of 50-200nt long of these sequences. Each sample might have millions of reads. The reads are then mapped to the genome using softwares such as Bowtie [150], TopHat [151] or BWA [152]. Gene expression is measured by the number of reads mapped to a gene. Thus, RNA-seq results in a discrete measurement for gene expression, which is different from the fluorescence intensity

measurement from microarray, that has been treated as a continuous and normally distributed variable. Figure 3.4 shows the difference between the expression of the genes obtained from a microarray experiment or from a sequencing. We can see that NGS distribution is not symmetrical and contains a lot of zero values on the left of the distribution, corresponding to genes that have not been detected (thus the distribution is said to be zero-inflated).

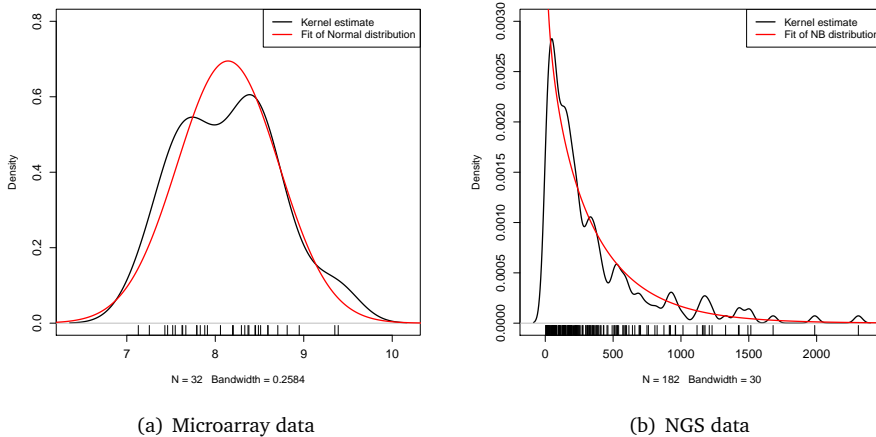


Figure 3.4: Differences between microarray ((a)) and NGS ((b)) data. Data extracted from GSE32676 dataset [153] and TCGA data from pancreatic cancer [154, 155]. GSE data has been normalised with limma, and for TCGA data FPKM (fragments per kilobase per million mapped reads) adjusted reads have been picked.

Consequently, if we want to use parametric tests, the statistical methods used to analyse microarray data (based on normal distributions) are not directly applicable for these data. Two main discrete probability distributions have been proposed to model the count data from RNA-seq studies: Poisson, and negative binomial (NB) [156]. As in microarray data, non-parametric tests should only be considered if there are outliers on the samples, for any other case is better to use parametric tests.

Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time in function of a specific rate. Each event is independent from the previous ones. R packages such as GPsseq package [157] or methods like a two-stage Poisson model proposed by Auer and Doerge (2011) [158] are based on Poisson distribution, but are not the most used. The other methods are based on Negative Binomial distribution, as it deals better with zero-inflated distributions and overdispersed data (Poisson distribution has only one parameter, so the mean

is related to the dispersion).

Apart from that, in differential gene expression analysis, one of the measures usually reported is the FoldChange (sometimes referred as simply FC). Being \bar{X} the mean expression of a group of samples X and \bar{Y} the mean expression of the control group samples Y , the FoldChange is defined as:

$$\text{FC} = \begin{cases} \frac{\bar{X}}{\bar{Y}} & \text{if } \bar{X} \geq \bar{Y} \quad (\text{upregulated}) \\ -\frac{\bar{Y}}{\bar{X}} & \text{if } \bar{X} < \bar{Y} \quad (\text{downregulated}) \end{cases}$$

However, this measure is discontinuous and has no values on the interval $[-1, 1)$. A transformation of the FC is the logratio (which is continuous) is equally used and it is defined as:

$$\text{logratio} = \begin{cases} \log_2(\text{FC}) & \text{if } \text{FC} > 0 \\ \log_2(-1/\text{FC}) & \text{if } \text{FC} < 0 \end{cases}$$

3.3.1 T-test

The simplest way to assess differential expression is by a T-test. This test is designed to test if the means of two populations differ, assuming that they come from a normal population. The hypothesis to test is if the means of both populations are equal (H_0) or not (H_1):

$$\begin{cases} H_0 : \mu_X = \mu_Y \\ H_1 : \mu_X \neq \mu_Y \end{cases}$$

Assuming equal variances and being $X = X_1, \dots, X_m$ and $Y = Y_1, \dots, Y_n$ samples from the two populations to compare, the t statistic can be computed as:

$$t = \frac{\bar{X} - \bar{Y}}{s_p \sqrt{\frac{1}{m} + \frac{1}{n}}} \quad (3.1)$$

Where

$$s_p = \sqrt{\frac{(m-1)s_X^2 + (n-1)s_Y^2}{m+n-2}} ;$$

$$\bar{X} = \frac{1}{m} \sum_{i=1}^m X_i ; \quad \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i ;$$

$$s_X^2 = \frac{1}{m-1} \sum_{i=1}^m (X_i - \bar{X})^2 ; \quad \text{and} \quad s_Y^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2 .$$

And t follows a Student's t distribution with a total of $m+n-2$ degrees of freedom.

In case that the variances are not equal the test is:

$$t = \frac{\bar{X} - \bar{Y}}{s_p} \tag{3.2}$$

Where

$$s_p = \sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}$$

In this case the distribution of the t statistic is approximated as a Student's t distribution with m degrees of freedom.

$$m = \frac{(s_X^2/m + s_Y^2/n)^2}{(s_X^2/m)^2/(m-1) + (s_Y^2/n)^2/(n-1)}$$

In both cases, unilateral or bilateral p values can be derived from the distribution of the t statistic.

This test is implemented in the function `t.test` from the `stats` package in R. Among other options, it allows to choose between equal variances or not (the default value is to consider unequal variances) and there is also a variant of t-test specific for paired samples.

This is a simple and well-known test for testing difference of means when the variables are considered to follow a normal distribution. However it is not suited for microarray data because the estimation of the variance may become unstable in this kind of data, as usually there are much more genes to test than number of samples available [138].

3.3.2 Wilcoxon Test

The Wilcoxon Rank Sum test (sometimes called also Mann-Whitney test) is the non-parametric version of the T-test when the interest is to test if both medians (M) are equal or not.

$$\begin{cases} H_0 : M_X = M_Y \\ H_1 : M_X \neq M_Y \end{cases}$$

Being the same samples $X = X_1, \dots, X_m, Y = Y_1, \dots, Y_n$ and if we want to test if $M_X = M_Y$ (M =median), we sort all the data and j is the rank.

$$T_X = \sum_{j=1}^{m+n} jI_j,$$

where

$$I_j = \begin{cases} 1 & \text{if observation with rank } j \text{ comes from } X \\ 0 & \text{if observation with rank } j \text{ comes from } Y \end{cases}$$

The exact distribution of T_X is tabulated for each value of m and n . Furthermore, the asymptotic distribution of T_X is normal with parameters:

$$E(T_X) = \frac{m(m+n+1)}{2}, \quad V(T_X) = \frac{mn(m+n+1)}{12}$$

This test is implemented in the function `wilcox.test` from the `stats` package in R. An exact calculation of the p value is also available.

3.3.3 Limma (Linear Models for Microarray Data)

Limma is an R/Bioconductor package specifically designed for the analysis of gene – mRNA– (or miRNA) expression microarray data [159, 138]. It uses a combination of linear models and Empirical Bayesian methods for analysing designed experiments and

assess differential expression. Limma provides the ability to analyse comparisons between many RNAs simultaneously from simple microarray experiments to arbitrary complicated designed microarray experiments [159, 138].

The expression data (usually expressed in log-intensities) is assumed to follow a normal distribution and the central idea is to fit a linear model to the expression data for each mRNA (or miRNA). Empirical Bayes and other shrinkage methods are used to borrow information across genes making the analyses stable even for experiments with small number of arrays (samples).

The hypotheses to test are, like in T-test, the equality between means in both groups, but for each gene of the dataset on the same test:

$$\begin{cases} H_0 : \mu_{X_i} = \mu_{Y_i} \\ H_1 : \mu_{X_i} \neq \mu_{Y_i} \end{cases} \quad \text{for } i = 1, \dots, \# \text{genes}$$

Limma will make most of all the data available to make a better estimation of the gene variance. In order to fit linear models it is assumed that the response variable is $\mathbf{y}_g = (X_1, \dots, X_m, Y_1, \dots, Y_n)$, $g = (1, \dots, m + n)$, and its expectation is:

$$E(\mathbf{y}_g) = X \boldsymbol{\alpha}_g$$

Where X is a design matrix and $\boldsymbol{\alpha}_g$ an unknown coefficient vector. It is also assumed that:

$$\text{var}(\mathbf{y}_g) = W_g \sigma_g^2$$

where σ_g^2 is the gene-variance and W_g is a known non-negative definite weight matrix. This weights can refer to the quality of the individual observations or other quality measures. \mathbf{y}_g may have missing values and W_g can also contain diagonal weights equal to zero (for example for the missing values).

The contrast estimators are defined for $\boldsymbol{\beta}_g = C^T \boldsymbol{\alpha}_g$, where C^T is a constant vector (for example in two-comparison group we can have $C^T = (0, 1)$ to pick out the coefficient relating to the difference between groups).

The linear model is fitted for each gene to obtain the $\hat{\boldsymbol{\alpha}}_g$ estimators and the s_g^2 (the estimator of σ_g^2).

$$\text{var}(\hat{\alpha}_g) = V_g s_g^2$$

where V_g is a positive definite matrix not depending on s_g^2 . So the contrast estimators are $\hat{\beta}_g = C^T \hat{\alpha}_g$, with estimated covariances:

$$\text{var}(\hat{\beta}_g) = C^T V_g C s_g^2$$

The contrast estimators $\hat{\beta}_g$, once β_g and σ_g^2 are known, are assumed to follow a normal distribution with mean β_g and covariance matrix $C^T V_g C \sigma_g^2$; and the residual variances, once σ_g^2 are known, are assumed to follow approximately a scaled chisquare distribution.

In summary:

$$\begin{aligned} \hat{\beta}_g \mid \beta_g, \sigma_g^2 &\sim N(\beta_g, v_g \sigma_g^2) \\ s_g^2 \mid \sigma_g^2 &\sim \frac{\sigma_g^2}{d_g} \chi_{d_g}^2 \end{aligned}$$

Where $v_g = C^T V_g C \sigma_g^2$ and d_g is the residual degrees of freedom for the linear model for gene g .

In this step Bayesian Methods are used to estimate the variance from the data. The prior information is described as:

$$\frac{1}{\sigma_g^2} \sim \frac{1}{d_0 s_0^2} \chi_{d_0}^2$$

That is to say that the inverse of the variance follow a χ^2 distribution divided by $d_0 s_0^2$, with d_0 degrees of freedom.

Which is updated as:

$$\tilde{s}_g^2 = E(\sigma_g^2 \mid s_g^2) = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g}$$

Then the "moderated" t statistic is an evolution of the t statistic (Equation 3.1):

$$\tilde{t}_{gj} = \frac{\hat{\beta}_{gj}}{\tilde{s}_g \sqrt{v_{gj}}} \sim t_{d_g+d_0} \quad (3.3)$$

Where $\hat{\beta}_g$ can be interpreted as the group effect $\bar{X} - \bar{Y}$. \tilde{t}_{gj} follows a t -distribution with $d_g + d_0$ degrees of freedom. The added d_0 degrees of freedom over the classical t_g ($t_g = \frac{\hat{\beta}_g}{s_g \sqrt{v_g}} \sim t_{d_g}$) reflect the information that is "borrowed" from the data. The overall estimate variation is s_0 , the per-gene deviation variation is s_g^2 , \tilde{s}_g^2 is the shrinkage variation, where $\frac{d_0}{d_0+d_g}$ is the weight coefficient associated with all probes and $\frac{d_g}{d_0+d_g}$ is the coefficient associated with gene g .

voom is a function included on `limma` package that allows to transform counts obtained from NGS sequencing and convert them to values that can be treated like microarray data [137]. Explained in a few words, it mainly does a log-2 transformation plus a count normalisation. Expression data transformed using this method should be then analysed using a variation of `limma`, which is called *limma-trend* procedure [137].

3.3.4 RankProd

RankProd [146, 160] is the non-parametric option for testing differential expression as it is based on ranks rather than quantitative measurements. Similarly to the Wilcoxon test, it is aimed to test if the medians between two groups are equal or not, but testing all the genes at the same time:

$$\begin{cases} H_0 : M_{X_i} = M_{Y_i} \\ H_1 : M_{X_i} \neq M_{Y_i} \end{cases} \quad \text{for } i = 1, \dots, \#\text{genes}$$

This test is implemented in the RankProd R/Bioconductor package [146, 160]. This function was first designed for analysing differential expression when the data came from different microarray sources, but is also useful when the original data does not follow a normal distribution.

RankProd is an heuristic method based on a series of permutation tests. It works as

follows: having m and n samples for each condition (X and Y) the algorithm is:

1. Compute the following ratios for each gene: $X_1/Y_1, X_2/Y_1, \dots, X_{m-1}/Y_n, X_m/Y_n$.
2. Rank the ratio within each comparison i ($i = 1, \dots, K$, where $K = m \times n$, and the largest is rank=1): r_{gi} .
3. The rank product for each gene is $RP_g = \left(\prod_i r_{gi}\right)^{1/K}$
4. Permute expression value and repeat steps 1-3. This will be the reference distribution $RP_g^{(l)}$, ($l = 1, \dots, L$).
5. Repeat step 3 L times and determine heuristically the p value and FDR associated with each gene.

This method is free from any prior assumption of the distribution of the data.

In summary, the `limma` method is more appropriate unless biological outliers are present. `Limma` capitalises on stability across samples, while `RankProd` not. `RankProd` is also less sensitive to outliers.

3.3.5 DESeq, edgeR

There are several methods (usually already implemented on R/Bioconductor packages) to deal with NGS data in a parametric way. Different studies compare the performance of the implemented methods on `miRComb` (`DESeq`, `edgeR`) and other methods (`BaySeq` [161], `NOISeq` [162] and others). The overall conclusion is that no single method works better in all conditions [137, 163, 164, 165], but some of them offer options of analysis than the others do not.

At this moment, `DESeq` and `edgeR` are so far the most used ones. As they are based on the same concepts, we will explain `DESeq` in detail and then spot the differences with `edgeR`.

`DESeq` [147] (and more recently its update `DESeq2` [166]) is an R/Bioconductor package aimed to deal with RNA-seq data based on negative binomial distribution.

As for `limma`, the hypotheses to test are equal means between two groups, testing all

the genes at the same time:

$$\begin{cases} H_0 : \mu_{X_i} = \mu_{Y_i} \\ H_1 : \mu_{X_i} \neq \mu_{Y_i} \end{cases} \quad \text{for } i = 1, \dots, \#\text{genes}$$

We assume that the number of reads in sample j that are assigned to gene i (k_{ij}) can be modeled by a negative binomial (NB) distribution that has two parameters (size and scale):

$$k_{ij} \sim NB(\mu_{ij}, \sigma_{ij}^2) \quad (3.4)$$

Where μ_{ij} is expectation value of the observed counts for gene i in sample j (size), and σ_{ij}^2 is the variance term (scale). μ_{ij} can be, in turn decomposed to:

$$\mu_{ij} = q_{i,\rho(j)} s_j$$

Where $q_{i,\rho(j)}$ is proportional to the expectation value of the true (but unknown) concentration of fragments from gene i under condition $\rho(j)$ ($\rho(j) \in (X, Y)$) and the size factor s_j represents the coverage, or sampling depth, of library from sample j . The purpose of the size factors s_j is to make counts from different samples, which may have been sequenced to different depths, comparable.

The variance term (σ_{ij}^2) is the sum of shot noise (proportional to mean expression) and a raw variance:

$$\sigma_{ij}^2 = \mu_{ij} + s_j^2 \nu_{i,\rho(j)}$$

$\nu_{i,\rho(j)}$ is the gene variance associated to gene i in sample j and can be estimated through a smoothing function from $q_{i,\rho(j)}$: $\nu_{i,\rho(j)} = \nu_\rho(q_{i,\rho(j)})$. That means that genes with similar expression will have similar raw variances. This assumption is needed because the number of replicates is typically too low to get a precise estimate of the variance for gene i from just the data available for this gene. This assumption allows us to pool the data from genes with similar expression strength for the purpose of variance estimation.

Our raw data is a table of counts of dimensions $N \times M$ where $i = 1, \dots, N$ indexes the genes, and $j = 1, \dots, M$ indexes the samples that will be used for estimating the different parameters.

Total number of reads has been shown to not be a good estimate to size factors s_j as a few highly and differentially expressed genes may have strong influence on the total read count, even with samples sequenced at the same depth. In this package, they take the median of the ratios of observed counts divided by a pseudo-reference sample obtained by taking the geometric mean across samples:

$$\widehat{s}_j = \text{median}_i \frac{k_{ij}}{\left(\prod_{v=1}^m k_{iv}\right)^{1/m}}$$

To estimate $q_{i\rho(j)}$ (shortened to $q_{i\rho}$), we use the average of the counts from the samples j corresponding to condition ρ , transformed to the normalised scale:

$$\widehat{q}_{i\rho} = \frac{1}{M_\rho} \sum_{j:\rho(j)=\rho} \frac{k_{ij}}{\widehat{s}_j} \quad \text{then,} \quad \widehat{\mu}_{ij} = \widehat{q}_{i\rho} \widehat{s}_j$$

where M_ρ is the number of replicates of condition ρ and the sum runs over these replicates. Variances on the normalised scale can be computed as:

$$w_{i\rho} = \frac{1}{M_\rho - 1} \sum_{j:\rho(j)=\rho} \left(\frac{k_{ij}}{\widehat{s}_j} - \widehat{q}_{i\rho} \right)^2$$

A new parameter $z_{i\rho}$ is defined as:

$$z_{i\rho} = \frac{\widehat{q}_{i\rho}}{M_\rho} \sum_{j:\rho(j)=\rho} \frac{1}{\widehat{s}_j}$$

$w_{i\rho} - z_{i\rho}$ is an unbiased estimator for the raw variance parameter $v_{i\rho}$, but in the case of small sample size, this value is highly variable so it would be better to use an estimator obtained from a smoothing function ($w_\rho(q)$):

$$\widehat{v}_\rho(\widehat{q}_{i\rho}) = w_\rho(\widehat{q}_{i\rho}) - z_{i\rho}$$

That means that:

$$\widehat{\sigma}_{ij}^2 = \widehat{\mu}_{ij} + \widehat{s}_j^2 \widehat{v}_\rho(\widehat{q}_{i\rho})$$

That allows us to rewrite the initial hypothesis to test as $q_{iX} = q_{iY}$:

$$\begin{cases} H_0 : q_{iX} = q_{iY} \\ H_1 : q_{iX} \neq q_{iY} \end{cases} \quad \text{for } i = 1, \dots, \#\text{genes}$$

where q_{iX} is the expression strength parameter for the samples of condition X , and q_{iY} for condition Y .

The total number of counts for one condition can be computed as:

$$K_{ij} \sim NB(\hat{\mu}, \hat{\sigma}^2)$$

Reparametrizing: $p = \frac{\hat{\mu}}{\hat{\sigma}^2}$ and $r = \frac{\hat{\mu}^2}{\hat{\sigma}^2} - \hat{\mu}$

Thus, the probability of obtaining exactly K_{ij} counts in

$$P(x = K_{ij}) = \binom{k+r-1}{r-1} p^r (1-p)^k \quad (3.5)$$

The total counts of gene i per condition X is:

$$K_{iX} = \sum_{j:\rho(j)=X} K_{ij}$$

Using the previous formula: we can compute the probabilities of the events $K_{iX} = x$ and $K_{iY} = y$ for any pair of numbers x and y . $P(x, y) = P(x)P(y)$. The p value of a pair of observed count sums (K_{iX}, K_{iY}) is then the sum of all probabilities less or equal to $P(K_{iX}, K_{iY})$, given that the overall sum is $K_{iS} = K_{iX} + K_{iY}$ (Equation 3.6)

$$p_i = \frac{\sum_{\substack{x+y=K_{iS} \\ P(x,y) \leq p(K_{iX}, K_{iY})}} P(x, y)}{\sum_{x+y=K_{iS}} P(x, y)} \quad (3.6)$$

What is important in this method is that it normalises according to median of ratios, estimates the variance taking into account genes with similar counts and estimates the p values using a negative binomial distribution.

edgeR is another R/Bioconductor package aimed to analyse RNA-seq data [148, 167]. Just as DESeq, it is based on the assumption that count data follows a negative binomial

distribution.

As a special feature, `edgeR` has functions to deal with paired samples (for example cancer and matched normal tissue). Regarding to implementation in R, it is more similar to `limma`.

Differently from `DESeq`, `edgeR` uses the trimmed mean of the log expression ratios to estimate the size factors s_j and perform the normalisation. The trimmed mean remove the extreme values, which is an idea similar to `DESeq` geometrical mean, or other packages that use the 75th quartile. In most of the cases, all of these strategies perform similar [156].

Regarding to dispersion estimates, `DESeq` uses two sources to estimate the dispersion: smoothed estimate of per-gene dispersion (related to mean) and noise, while `edgeR` does not take it into account as a specific term. In practice, that means that `DESeq` may be less sensitive when detecting differentially expressed genes, while `edgeR` may be more sensitive to outliers [168]. The selection of one or other will depend on the aim of the study and which kind of uncertainty can be assumed.

3.4 Subset selection

It is possible to define subsets of the data according to the differentially expressed miRNA and/or mRNA: there are summarised in Figure 3.5. The use of a miRNA and/or mRNA subset depends on the nature of the problem and the desired results, as the different subsets are aimed to find different kinds of relevant interactions. We proposed these three subsets, but other ones can be defined according to the researchers' interests:

- **All subset.** It includes all the miRNAs and mRNAs. This subset is designed for finding all **the miRNA-mRNA interactions that occur in the cell**, even if they are not relevant for the disease.
- **miRNA subset.** It includes all the mRNAs but only the significant miRNAs. This subset is designed for finding all **the miRNA-mRNA interactions of the miRNA that play a role in the disease**, even if the genes are not relevant for the disease.
- **miRNA/mRNA subset.** It includes only the significant miRNAs and significant mRNAs. This subset is designed for finding all **the miRNA-mRNA interactions relevant for the disease**.

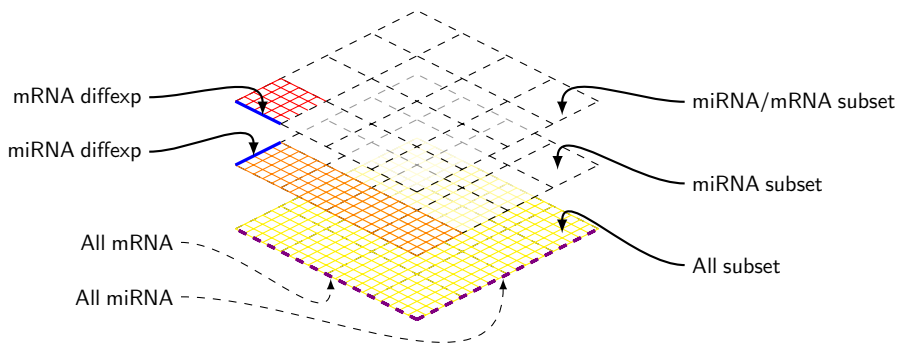


Figure 3.5: Representation of the three proposed subsets. Each little square represents one correlation. Depending on the problem, one can compute all the correlations or only a fraction of them. These fractions are defined by the lists of miRNA and mRNA selected. Note that the picture is not proportional, in most of the datasets there are many more mRNAs than miRNAs.

3.5 MiRNA-mRNA associations

Two variables X and Y are said to be associated when the value taken by one variable affects the distribution of the other variable (for example if the expression of one miRNA is related to the expression of one mRNA). On the other hand, X and Y are said to be independent if changes in one variable do not affect the other variable.

One of the most common ways to measure associations with two variables (X and Y) is with correlations. Specifically, positive correlation is said to occur when there is an increase in the values of Y as the values of X increase. Negative correlation occurs when the values of Y decrease as the values of X increase. The coefficients (r_λ , $\lambda \in$ Pearson, Kendall, Spearman) are measures of correlation, which typically reflect a monotone association (sometimes also a linear association) between the two variables.

Other systems such as linear regression models can also be used to measure association with two or more variables (for example if the expression of two miRNAs is associated with the expression of one mRNA). More detailed relationships, for example copulas in correlation, are not taken into account.

A correlation or association between two variables does not mean causality. However, a causality implies association. The first of our assumptions says that a miRNA is able to decrease the expression of their mRNA targets. That means a negative correlation or negative regression coefficients between the miRNA (for example the X) and the target(s) mRNA(s) (for example the Y). So, in all the cases the correlation test will be one sided as we want to check for the *true* targets:

$$\begin{cases} H_0 : r_\lambda \geq 0 \\ H_1 : r_\lambda < 0 \end{cases}$$

The studied correlation coefficients are:

- Pearson correlation coefficient
- Kendall τ coefficient
- Spearman correlation coefficient

Regarding to linear regression, we studied:

- Generalised Linear Models

3.5.1 Pearson Product-Moment Correlation Coefficient

The Pearson correlation is defined as:

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

So for a sample:

$$r_{X,Y} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (3.7)$$

where $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ and $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$

This coefficient is the only one of the three that reflects a linear relation between two variables: the other two reflect a monotone relation between the two variables.

The distribution of the t statistic is used for computing the p values for unilateral or bilateral hypothesis, as it follows a Student's t -distribution with $n-2$ degrees of freedom:

$$t = r_{X,Y} \sqrt{\frac{n-2}{1-r_{X,Y}^2}} \sim t_{n-2}$$

The Pearson Product-Moment Correlation is also related with the simple linear regression method. More specifically, the r_{XY} is the square root of the coefficient of determination (R^2) of a linear regression involving $Y \sim X$ (or $X \sim Y$ or more variables). A R^2 of 1 means perfect fitting of the model. This corresponds to $r_{XY} = 1$ or $r_{XY} = -1$, which also mean perfect linear correlation between the two variables. $R^2 = 0$ means, like $r_{XY} = 0$, no relation between the two variables.

Pearson correlation will be equal to the non-parametric ones (Spearman and Kendall) in the case of a linear relationship, otherwise it will take a lower value (in absolute terms).

3.5.2 Spearman Rank Correlation Coefficient

The Spearman Rank coefficient is described as follows:

Being $(X_1, Y_1), \dots, (X_n, Y_n)$, assign a Rank where $(R_1, S_1), \dots, (R_n, S_n)$

$$R_{S(X,Y)} = \frac{\sum_{i=1}^n (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^n (R_i - \bar{R})^2} \sqrt{\sum_{i=1}^n (S_i - \bar{S})^2}} \quad (3.8)$$

where $\bar{R} = \frac{1}{n} \sum_{i=1}^n R_i$ and $\bar{S} = \frac{1}{n} \sum_{i=1}^n S_i$

We can see that this is the same expression as used for the Pearson Product-Moment Correlation, but in this case it is computed from the ranks instead of the raw values of X and Y .

The exact distribution of the Spearman correlation coefficient is symmetric and can be easily obtained from precomputed tables. The asymptotic distribution of Spearman (when n tends to infinity) is:

$$\sqrt{n-1} R_{S(X,Y)} \sim N(0, 1)$$

3.5.3 Kendall τ Correlation Coefficient

The Kendall τ coefficient for a sample is defined as:

$$\tau_n = \frac{1}{\binom{n}{2}} \sum_{i < j} A_{ij} \quad (3.9)$$

where

$$A_{ij} = \begin{cases} 1 & \text{if } (X_i - X_j) \times (Y_i - Y_j) > 0 \\ -1 & \text{if } (X_i - X_j) \times (Y_i - Y_j) < 0 \end{cases}$$

The exact distribution of τ_n under the null hypotheses is symmetric and can be computed using tables. The asymptotic distribution (when n tends to infinity) is:

$$\frac{3\sqrt{n(n-1)}}{\sqrt{2(2n+5)}} \tau_n \sim N(0, 1)$$

The Kendall coefficient can be also interpreted geometrically. Note that when $A_{ij} = 1$ the two points X_i, Y_i and X_j, Y_j can be joined by a line that goes from the first to the third

quadrant; while when $A_{ij} = -1$ the two points X_i, Y_i and X_j, Y_j can be joined by a line that goes from the second to the fourth quadrant. The Kendall coefficient summarises the *mean* value of all A_{ij} 's.

3.5.4 Generalised Linear Models

Another option to measure associations between two variables is to use linear models. Linear models are used to assess if one variable can be used as a predictor of the other (if the relation is true, then the two variables are associated). Moreover, they can be used with more than two variables at the same time, for example, if several variables (for example, more than one miRNA) can be used to predict the value of one other (for example, one mRNA). These concepts fits well with our kind of data, where one miRNA is able to regulate up to hundreds of mRNAs, and one mRNA can be regulated by more than one miRNA.

To evaluate these models, we have used `glmnet` R package [169], that implements Generalised Linear Models analysis, to analyse the association between miRNAs and mRNAs. A similar method, called TaLasso has already been used for miRNA-mRNA analysis [127].

`Glmnet` is an extension of Least Absolute Shrinkage and Selector Operator –LASSO– regression model that uses elastic net regularisation (selection of variables). The elastic net is a regularised regression method that linearly combines the L1 ($\|\beta\|_1$) and L2 ($\|\beta\|_2^2$) penalties of the LASSO and ridge methods respectively. A detailed vignette can be found http://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html or in the `glmnet` paper [169] itself.

LASSO regression uses only L1 ($\|\beta\|_1$) penalties to select the significant predictor variables and is quite popular, but the use of L1 ($\|\beta\|_1$) alone has several limitations. For example, when a lot of variables are tested (in our case, thousands of mRNAs and miRNAs) with a relatively small number of samples (only a few tens), LASSO selects at most n variables before it saturates. Also if there is a group of highly correlated variables, then LASSO tends to select one variable from a group and ignore the others.

To overcome this limitation, the elastic net adds a quadratic part to the penalty ($\|\beta\|_2^2$), which when used alone is ridge regression. Ridge regression is known to shrink the coefficients of correlated predictors towards each other, allowing them to borrow strength from each other. In the extreme case of k identical predictors, they each get identical

coefficients with $1/k$ th the size that any single one would get if fitted alone. On the contrary, LASSO penalty would have picked only one coefficient and give it k size [169]. The elastic net regression is a compromise between the ridge regression ($\alpha = 0$) and LASSO penalty ($\alpha = 1$). Both elastic net or ridge regression allow to select more than n variables.

For a gaussian model assuming a normal distribution of the data, the estimates from the elastic net model are estimated by:

$$\min_{\beta_0, \beta} \frac{1}{N} \sum_{i=1}^N w_i l(y_i, \beta_0 + \beta^T x_i) + \lambda \left[(1 - \alpha) \|\beta\|_2^2 / 2 + \alpha \|\beta\|_1 \right] \quad (3.10)$$

over a grid of values of λ covering the entire range. Here $l(y, \eta)$ is the negative log-likelihood contribution for observation i , w_i is the weight of observation i . λ controls the overall strength of the penalty ($\lambda \geq 0$). N =sample size.

On the `glmnet` package, α can be selected by the user (a recommended value is $\alpha = 0.5$, which is half-way between LASSO and ridge regression), and λ are evaluated by crossvalidation, allowing to select the model with fewer errors.

As an additional feature, `glmnet` also allows to set coefficient upper and limit bounds (globally or specific for each factor), and also to add a specific penalty factor to each variable (more details about the mathematical formulation can be found on the `glmnet` package vignette [169]).

To study the relation between miRNA and mRNA expression, the following models have been implemented on the `miRComb` package:

- `Glmnet-mRNAs`: $\text{mRNAs} \sim \text{miRNAs}$

In this model, the expression of the whole set of mRNAs is defined by the expression of the whole set of miRNAs. We think that this model is the one that matches more accurately with the knowledge about the miRNA operation, but it needs to estimate a huge number of parameters. `Glmnet-miRNA` and `Glmnet-mRNA` are simplifications of this model that we think that could lead to similar conclusions and at the same time make the computations easier.

- `Glmnet-mRNA`: $\text{mRNA}_i \sim \text{miRNAs} (\forall i \in 1, \dots, \#\text{mRNAs})$

This model is focused on giving a list of miRNAs that are targeting each mRNA.

- `Glmnet-miRNA`: $\text{miRNA}_i \sim \text{mRNAs} (\forall i \in 1, \dots, \#\text{miRNAs})$

This model is focused on giving a list of the mRNA targets for each miRNA.

As a result, Glmnet-mRNAs gives a model with $\#miRNAs \times \#mRNAs$ β estimates; Glmnet-mRNA gives $\#mRNAs$ models, each of them with $\#miRNAs$ β estimates; and Glmnet-miRNA gives $\#miRNAs$ models, each of them with $\#mRNAs$ β estimates.

We set an upper bound to 0, and penalty factors were not included, as they are only available for univariate responses (Glmnet-mRNA and Glmnet-miRNA) and at this moment we wanted to compare the models to each other.

No p values are produced in Glmnet models, as non-significant items are shrunk to 0. For that reason, it is expected that all the *significant* items had already been included on the model with values different from 0, so no specific p value is needed.

3.6 Database integration

One of the main objectives of this work is to find miRNA-mRNA interactions occurring on a specific state. At this step of the work, we had two sources of information: the value of correlation or the coefficients obtained the via elastic net (**biological information**, ρ_{XY}) and the list of miRNA-mRNA pairs that are predicted to interact according to different databases –microCosm, TargetScan, etc.– (**theoretical information**).

We want that final miRNA-mRNA interactions fulfil both hypothesis:

$$\left\{ \begin{array}{l} H_0 : \rho_{XY} \geq 0 \quad \text{and/or} \quad \text{miRNA-mRNA not hybridises} \\ H_1 : \rho_{XY} < 0 \quad \text{and} \quad \text{miRNA-mRNA hybridises} \end{array} \right.$$

There are several methods to combine both sources of information. When two initial p values are available ($p_{\text{biological}}$, which can be obtained from Pearson, Spearman and Kendall correlation; and $p_{\text{theoretical}}$, which can be obtained from MicroCosm database) a p value combination test can be performed.

Several methods have been proposed to combine two or more p values [170]. There is no uniformly most powerful method for combining p values, but all of them require that the individual tests are independent in order to avoid inflating the type I error. Our data fulfil this condition, as the test for correlation is completely independent of the test of sequence hybridisation performed for the MicroCosm database team. In this work we compared two methods: the Fisher method and the Stouffer method.

3.6.1 P value combination: Fisher method

Fisher was one of the first persons to be interested in this problem and proposed this method in 1925 [171]. Initially used as a method of meta-analysis studies, it computes a statistic (t) which has been demonstrated to follow a Chi-square distribution with $2L$ degrees of freedom, where L is the total number of hypotheses to test:

$$t = -2 \sum_{i=1}^L \ln p_i \sim \chi_{2L}^2 \quad (3.11)$$

$$p = P(t \geq T) = 1 - F(t | 2L) = \int_t^\infty \frac{z^{L-1} e^{-\frac{z}{2}}}{2^L \Gamma(L)} dz \quad (3.12)$$

where,

$$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du$$

The statistic t is the sum of $\log-p$ values multiplied by minus 2. This is the same as taking the product of the square of the inverse p values and apply logarithm. Then, the statistic t is compared with the corresponding Chi-square cumulative density function (F) in order to compute the p value. $\Gamma(x)$ is the Gamma function. The result of the specific case of combining only two p values is shown in Figure 3.6.

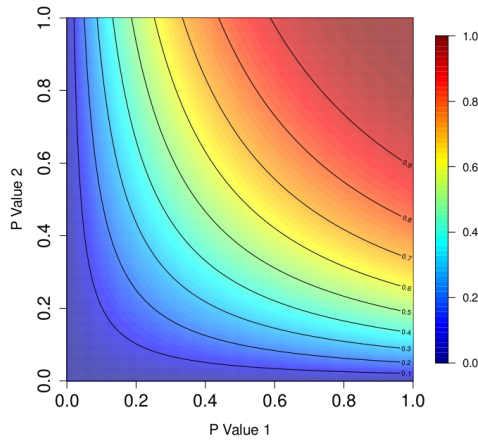


Figure 3.6: Fisher combination. This figure shows the value of the p_{Fisher} (see colour column on the right) depending on the two initial p values to combine.

This test is more sensitive than the other methods proposed [170]. However, one weakness of this statistic is that it is said to be skewed if one of the p values is small. This characteristic is related to the fact that the test does not actually assess whether the group of data sets collectively supports or refutes a common null hypothesis, it rather tests whether there is at least one significant component: so the p combined tend to be small if at least one of the p values is small (Figure 3.6). For example, if one of the p values is low (for example 0.1), the combined p value is 0.1 or less, even if the other p value is close to 1. When the p values are larger the range is wider: for example if one p value is 0.2, the overall p value can range from 0 up to 0.4, depending on the other value. If one p value is 0.6, the overall p value can be up to 0.85.

3.6.2 P value combination: Stouffer method

Stouffer proposed this method as an improvement of (among others) the Fisher method in 1949 [121]. The general equation is:

$$p = 1 - \Phi \left(\frac{1}{\sqrt{L}} \sum_{i=1}^L \Phi^{-1}(1 - p_i) \right) \quad (3.13)$$

where,

$$\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

Similarly to Fisher method, Stouffer proposes a *sum* of transformed p values. More specifically, it uses a normal transformation of the p values. The result is not a statistic which has to be contrasted, it is the p value itself.

A reformulation of the previous equation was described by Lipták in 1958 [172], which allows to weight the p values in the following way:

$$p = 1 - \Phi \left(\frac{1}{\sqrt{\sum_{i=1}^L w_i^2}} \sum_{i=1}^L w_i \Phi^{-1}(1 - p_i) \right)$$

The weights (w) are usually defined according to the sample sizes. Zaykin et al. stated in 2001 that the most powerful way to weight the p values is to assign weights according to $w_i = \sqrt{n_i}$ [173]. However, this can not be applied in our study as the sample size (or equivalent) of one of the tests is not known: the one derived from the MicroCosm database.

Figure 3.7(a) shows the behaviour of the function, which is different from Fisher combination (Figure 3.6). In that case the behaviour is similar in very small p values and very high p values. In practical terms, this means that for example for a p value of 0.5, the combination can range from 0.1 to 0.9 depending on the second p value. For a small p values (such as 0.1) the combined p value can range lower values, but only if the second p value is equal or less than 0.1 (or other small p value).

Figure 3.7(b) shows the difference of p values between the two methods. It is possible to see that the p value of Fisher method is smaller than the Stouffer except for the region enclosing the dashed diagonal, which represents when two initial p values are similar. This means that in the case of similar p values Stouffer method gives smaller p values

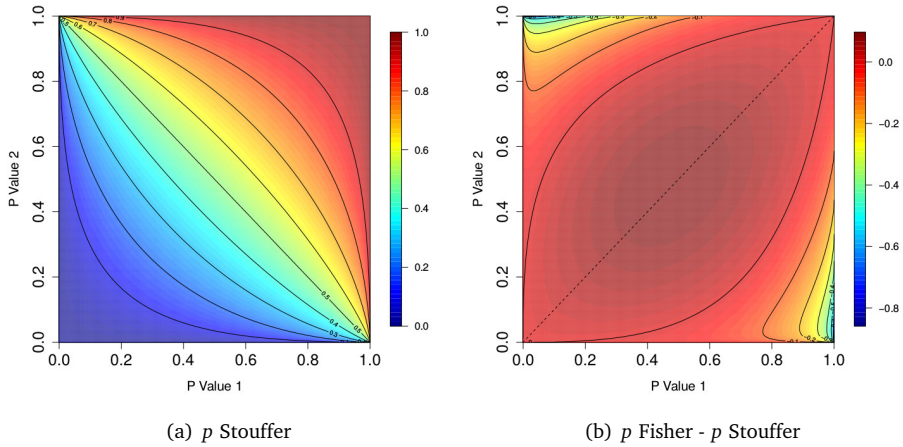


Figure 3.7: Stouffer combination, and comparison between Fisher and Stouffer combination. For (a), the figure shows the value of the p_{Stouffer} (see colour column on the right) depending on the two initial p values to combine. For (b) the figure shows the value (see colour column on the right) of $p_{\text{Fisher}} - p_{\text{Stouffer}}$.

(regions near the dotted diagonal), while in the case of one small p value the Fisher method gives smaller p values.

The Stouffer method is recommended in the majority of the cases and used by Gade S. et al. in their integrative study of miRNA and mRNA expression [120].

3.6.3 Intersection

P value combination can only be used with microCosm database and p values from correlation coefficients. This fact limits our options to study these interactions. For that reason we studied other ways of combination, such as database and correlations *intersection*. This method allows to user other databases apart from microCosm [145].

The procedure is represented in Figure 3.8. Biologically significant miRNA-mRNA pairs are labelled, then theoretically significant pairs are also labelled. Only miRNA-mRNA pairs that are labelled in both cases are labelled on the final list.

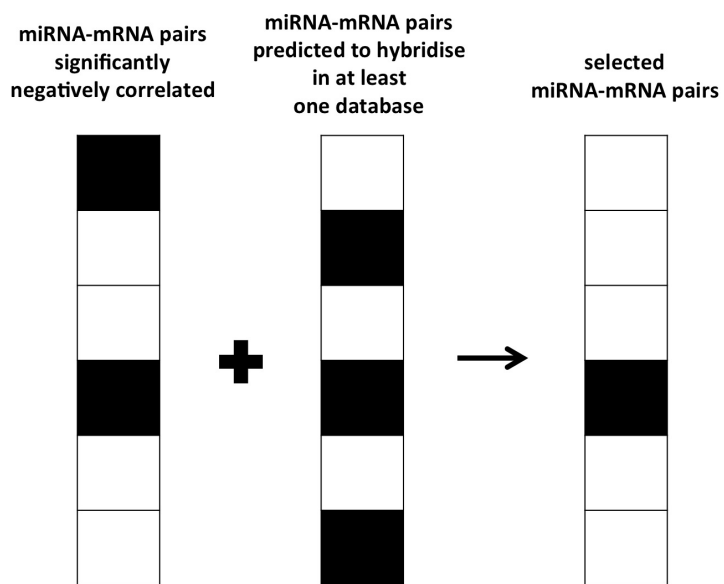


Figure 3.8: Venn showing the intersection. Black squares represent the selected miRNA-mRNA pairs in each situation.

3.7 Multiple testing correction

When performing a large number of tests, the error for false positives is no longer controlled. For example if we set an alpha value of 0.05 and we perform 100 tests we will expect to find at least 5 positives, even if the null hypothesis is true. The probability of finding at least one false positive –P(false negative)– increases in the following way (assuming that all the n tests are independent): If:

$$\alpha = P(\text{false negative})$$

$$P(\text{false positive}) = 1 - (1 - \alpha)^n$$

Meaning that for example (for $\alpha = 0.05$) when performing 10 tests the probability to find at least one positive is 40.1%, but for 100 tests the probability of finding at least a false positive is 99.4% (Figure 3.9).

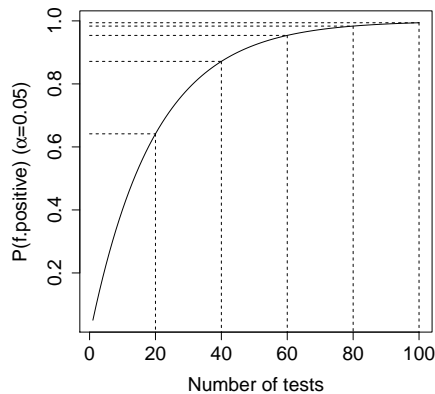


Figure 3.9: Probability of at least one false positive depending on the number of tests performed. α is set to 0.05.

In this work, the number of tests performed is specially large (up to hundreds of thousands of simultaneous tests: one per each miRNA-mRNA pair), so the probability of finding at least one false positive is almost 1. Therefore we need to find some way to control this error.

In order to avoid the increment of false positives, the reference alpha value (α , called

α_{ref} from now on) must be lowered to some α_{ind} to be used in each comparison that guarantees the control of type II errors. This α_{ind} value has different interpretations according to the correction method used. α_{ind} is unknown, but there are some methods for approximating it that will be described in the following sections.

3.7.1 Bonferroni Correction

The Bonferroni correction controls the Familywise Error Rate (FWER). That is, the probability of selecting one or more false positives among all the hypotheses tested.

Bonferroni described the theory in 1936 [174], but modern usage in biomedicine dates from 1965 [175]. Although their use is not always recommended [176] and new forms of control have been implemented (see next section) the simplicity and usefulness of the Bonferroni correction make it still used.

The proposition of the Bonferroni correction is to compute α_{ind} as:

$$\alpha_{\text{ind}} = \frac{\alpha_{\text{ref}}}{n} \tag{3.14}$$

where α_{ref} is the desired probability of at most one false positive, n is the number of tests performed and α_{ind} is the individual α to be tested to each p value in order to guarantee that the probability to find at most one false positive in all the tests is α_{ref} .

As an example, assume that we have the following p values:

$$0.02, 0.32, 0.01, 0.12, 0.78$$

And we want to know which of them are significant for $\alpha_{\text{ref}} = 0.05$. We have a total of 5 tests, so α_{ind} is $0.05/5 = 0.01$. In this case only the third p value is significant after p value correction.

This method controls the FWER at $\text{FWER} \leq \alpha_{\text{ref}}$, derived from Boole's Inequalities¹ (being n_0 the number of the true null hypotheses):

¹Boole's Inequality:

$$P\left(\bigcup_{i=1}^n E_i\right) \leq \sum_{i=1}^n P(E_i)$$

If E_i are disjoint sets, then the inequality becomes an equality.

$$\text{FWER} = P \left\{ \bigcup_{i_o} \left(p_i \leq \frac{\alpha_{\text{ref}}}{n} \right) \right\} \leq \sum_{i_o} \left\{ P \left(p_i \leq \frac{\alpha_{\text{ref}}}{n} \right) \right\} \leq n_0 \frac{\alpha_{\text{ref}}}{n} \leq n \frac{\alpha_{\text{ref}}}{n} \leq \alpha_{\text{ref}}$$

This method is free of:

- Any test distribution assumption.
- Dependence testing (the tests to be corrected can be dependent). In our case, we can not assume that all the tests are independent of each other, because the expression of one mRNA can be related to other mRNAs, so the correlation test results for these mRNAs can be related.

The disadvantage is that this correction is too conservative. It controls the FWER, but at the expense of increasing the Type II error (false negative), decreasing the power of the test for finding true positives. For this reason this method is not desirable if the number of tests is large and/or we want to increase the power of the test, which is often the case of mRNA/miRNA expression [176, 177].

3.7.2 Benjamini & Hochberg correction

The control of the FWER is sometimes not needed and it has some problems. Benjamini & Hochberg described these problems in their article in 1995 and proposed a new method for handling the multiple testing problem: the False Discovery Rate (FDR) [177].

More specifically, the problems described from the FWER/Bonferroni approach are:

- Test statistics must be multivariate normal. It is not a problem in this study but it must be considered for further research.
- The methods that control for FWER tend to have substantially less power than the Per Comparison Error Rate (true α_{ind}).
- Sometimes the control of the FWER is not even needed. For example, if testing some differentially expressed genes we want to select those really differentially expressed despite some of the other genes are falsely rejected.

The proposal from Benjamini & Hochberg is to use the False Discovery Rate, where the proportion of false positives is controlled, instead of controlling the probability of one single false positive.

To understand the FDR, we must take into account all the possible errors and right choices when doing a hypothesis test. Table 3.2 shows what possibilities we have. V correspond to the type I error: rejecting the null hypotheses when it is true; and the T represents the type II error: do not reject the null hypotheses when it is false. U and S are the correct choices: when we accept or reject the null hypotheses according to its true value.

	Non-significant	Significant	Total
True H_0	U	V	n_0
True H_1	T	S	$n - n_0$
Total	$n - R$	R	n

Table 3.2: Number of errors and correct decisions committed when testing n null hypotheses.

According to this table, the proportion of false positives (FDR) among all the rejected tests is:

$$\text{FDR} = \frac{V}{V + S}$$

Similarly, the probability of finding at least one false positive (FWER) is:

$$\text{FWER} = P(V \geq 1)$$

The Q value is the expectation of the FDR on a given sample:

$$Q = E[\text{FDR}] = E\left[\frac{V}{V + S}\right] \tag{3.15}$$

Q is an unobserved random variable, because we do not know the value of V or S , we only know the total of tests rejected: $V + S = R$.

The properties of the FDR are:

- If all null hypotheses are true, FDR is equivalent to FWER.
- If $n_0 < n$ (almost always) the FDR is smaller than or equal to FWER. So any procedure that controls for FWER also controls the FDR. However if controlling only for

FDR the procedure is less stringent and a gain of statistical power may be expected. The potential of increasing the statistical power is proportional to the difference between n_0 and n .

- FDR controls the FWER in the weak sense.

To control the Q value for each test Benjamini & Hochberg described the following procedure:

Let $p_1 \leq p_2 \leq \dots \leq p_n$ the ordered p values. And H_i the null hypothesis corresponding to p_i . Being k the largest i for which:

$$p_i \leq \frac{i}{n} q^*$$

Then reject all H_i ($i = 1, 2, \dots, k$). In that case, the FDR is controlled at q^* level (Appendix A from [177]). Individual q values can be computed as:

$$q_{value} = \min \{ p_{ref} \cdot n/k, 1 \}$$

So this means that for $q_{value} = \alpha$, the group of tests that have $q_{value} \leq \alpha$ have a proportion of α false positives, being the FWER also controlled.

As an example, if we take the values from the previous example

$$0.02, 0.32, 0.01, 0.12, 0.78$$

We get the following q values:

$$0.05, 0.40, 0.05, 0.20, 0.78$$

Note that in this case, the first but also the third are significant after p value correction. This means that if we take the first and third p values the FDR is 0.05, if we take the first, third and fourth p values the overall FDR is 0.20, and so on. In this example, we can see that this method controls also the FWER (it includes at least the p values from the Bonferroni correction, plus the others which also are FDR < 0.05).

3.8 Functional analysis of miRNA targets

Once the miRNA-mRNA interaction list is obtained, it is possible to do a functional analysis. In our package, the functional analysis is aimed to see which specific functions are regulating the miRNAs. Gene ontology (GO) terms are used to annotate the function of the genes in standardised terms [178], and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways are a collection of manually curated pathways that represent cell metabolism [179]. Both classifications can be used to infer the function of a miRNA(s) according to the functions of its(their) targets.

We are interested in the overrepresented functions of the mRNAs of the miRNA-mRNA interaction list, as they are the functions that are regulated by the miRNAs. For that reason, we test if the proportion of the gene ontology (GO) terms in our miRNA-mRNA list (p_{subset}) is larger than a given reference (p_{overall}). In other words, we test if the counts of GO terms in groups of genes are likely to be found by chance or not.

3.8.1 Enrichment Analysis

More specifically, the miRComb package allows testing for overrepresented GO and KEGG terms. The specific hypotheses are:

$$\begin{cases} H_0 : p_{\text{subset}} \leq p_{\text{overall}} \\ H_1 : p_{\text{subset}} > p_{\text{overall}} \end{cases}$$

Where p_{subset} is the proportion of a given GO term respect the other GO terms in the subset of mRNA tested and p_{overall} is the proportion of a given GO term respect the other GO terms in the all the possible mRNAs.

This analysis will help us to answer the following questions:

- Which functions are deregulated in the disease?
- Does the miRNA regulate a set of genes with specific functions?

There are many ways to test enrichment, but we focused on testing for overrepresentation because we were interested in the functions that *are* on the list, not those that *are*

not on the list.

3.8.1.1 Proportions test

One way to test the overrepresentation of a given set of terms is the Proportions test (also known as Chi-squared Test). This option is used by the DAVID program (Functional Annotation Bioinformatics Microarray Analysis [124]), a program widely used in the scientific community to test for overrepresentation of gene categories. The χ^2 value is computed as follows, and follows a χ^2 distribution with $n - p$ degrees of freedom:

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i} \simeq \chi_{n-p}^2 \quad (3.16)$$

where χ^2 is Pearson's cumulative test statistic, which asymptotically approaches a χ_{n-p}^2 distribution; O_i is the observed frequency; E_i is the expected (theoretical) frequency, asserted by the null hypothesis; n is the number of cells in the table; p are the degrees of freedom ($p = (\#rows - 1) \times (\#columns - 1)$). When the contingency table is 2×2 this test is equivalent to the Z-test of proportions.

For the specific case of our problem the contingency table is computed as follows (Table 3.3):

(a) Observed frequencies		
	Selected gene list	Genome
In the pathway	30	400
Not in the pathway	270	29600
(b) Expected frequencies		
	Selected gene list	Genome
In the pathway	4.26	425.74
Not in the pathway	295.74	29574.26

Table 3.3: Example of a contingency table. Example extracted from DAVID program. A specific term is found 400 times in all the human genome (30000) genes. On list of 300 genes, 30 of them are related to its specific pathway. We want to test if 30 in 300 is different from 400 in 30000. For example 4.26 is computed as: $(30 + 400) \times (30 + 270) / (30 + 270 + 400 + 29600)$. The chi-square statistic is 159.4724. The p-value is $< .00001$. The OR is 8.22.

However, the approximation for the chi-squared distribution has problems if at least one cell has an expected frequency below 5. In this case a continuity problem appears because we are approximating a discrete variable (χ^2) with a continuous one. If this happens, the use Yates' correction for continuity [180] is proposed:

$$\chi_{\text{Yates}}^2 = \sum_{i=1}^N \frac{(|O_i - E_i| - 0.5)^2}{E_i}$$

This correction lowers the Chi-square statistic so it increases the p value. It reduces the overall error but it tends to be overconservative and increase the type II error (fail to reject H_0 when H_1 is true).

Another option is to compute directly an exact test, it is computationally slower but the results are more accurate.

3.8.1.2 Hypergeometric test

Another test aimed to see overrepresented terms is the hypergeometric test. In this test, we assume that a random variable X (in our case, the number of GO terms) follows the hypergeometric distribution if its probability mass function (pmf) is given by:

$$P(X = k) = \frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}} \quad (3.17)$$

Where N is the population size; K is the number of success states in the population; n is the number of draws; k is the number of successes; and $\binom{a}{b}$ is a binomial coefficient.

Then, the unilateral p value to test for overrepresentation will be computed as:

$$P(X \geq k) = \sum_{i=k}^n \frac{\binom{K}{i} \binom{N-K}{n-i}}{\binom{N}{n}}$$

In which we evaluate if the number of terms is higher than the expected one or not. In the previous example, $P(X \geq 30) = 1.25^{-17}$.

In both cases (proportions and hypergeometric test) a multiple test correction is also needed because all the GO or KEGG terms are tested at the same time.

The advantage of the Chi-square test is that is computationally fast, but the results may be inaccurate for extreme distributions. For that reason, we decided to preferentially use

the hypergeometric test, which has been already implemented in the `G0stats` package from R/Bioconductor.

3.9 Analysis of number of targets per miRNA

Similarly to the analysis of overexpressed GO terms on the mRNA targets, we can test if one miRNA has more targets than the expected by random. In this context, we defined miRNA-mRNA interactions expected by random as those predicted by the miRNA-mRNA databases, but not taking place in our disease (referred also as false positives of the database). Figure 3.10 shows how we hypothesise that miRNA-mRNA interactions are divided: among all the predicted miRNA-mRNA interactions, some of them are taking place in our disease while others not.

3.9.1 Hypergeometric Test

Similarly to pathway enrichment, we can imagine a situation similar that the one showed in Table 3.4. In this example, the databases predict 430 miRNA-mRNA interactions. 30 of them are negatively correlated with their targets (predicted by miRComb), while 400 are not. The Odds Ratio (OR) is 8.22 and it is significant (Table 3.3).

(a) Observed frequencies		
	Significant negative correlations	Other correlations
Predicted miRNA-mRNA interaction	30	400
Not predicted miRNA-mRNA interaction	270	29600
Proportion of predicted miRNA-mRNA interactions	0.1	0.01333

(b) Proposed underlying distribution of the miRComb predicted miRNA-mRNA interactions		
	Significant correlation	Other correlations
MiRNA-mRNA interactions taking place on the disease	26	∅
MiRNA-mRNA interactions not taking place on the disease	4	400

Table 3.4: Example of an underlying distribution from Figure 3.10. Subtable (b) extracted from observed frequencies of Subtable (a).

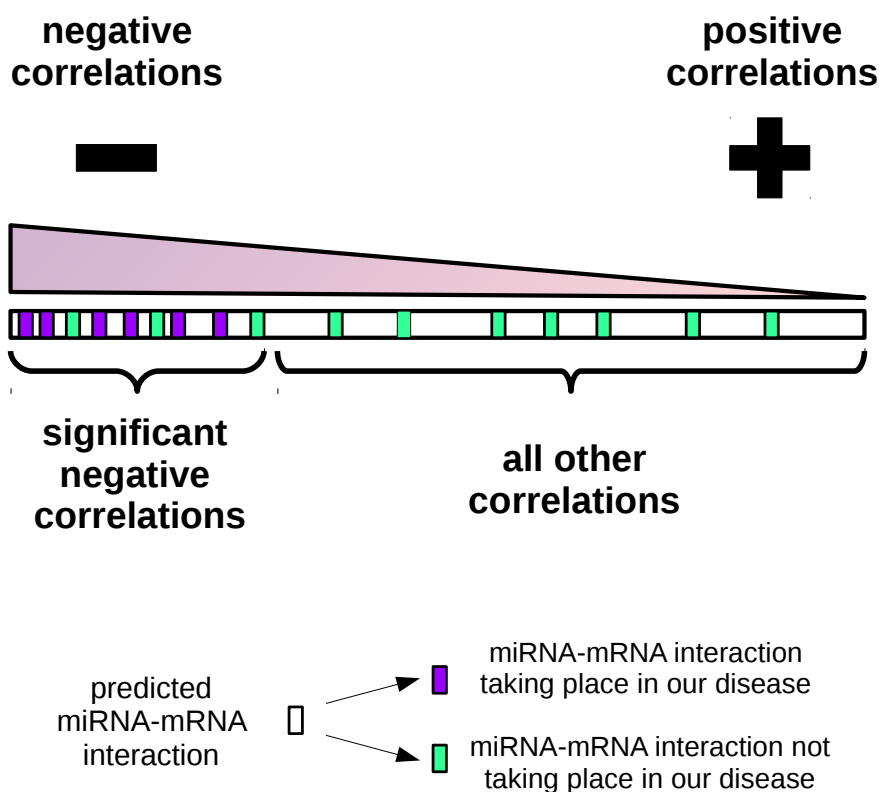


Figure 3.10: Distribution of the two types of miRNA-mRNA interactions. MiRNA-mRNA interactions not taking place in our disease (turquoise) are randomly distributed around all the correlations. MiRNA-mRNA interactions taking place in our disease (violet) are located in negative correlations. In an ideal situation, they are all found on significant negative correlations ($FDR < 0.05$).

According to our proposed underlying distribution (Figure 3.10), the 400 targets found on *other correlations* are not taking place on the studied disease. They correspond to turquoise interactions of Figure 3.10. Regarding the 30 interactions found on negative correlations, we expect a mix of interactions taking place on the disease and interactions not taking place on the disease. Using the expected frequency of miRNA-mRNA interactions not taking place on the disease (0.0133 per miRNA-mRNA interaction tested), 4 ($0.0133 \times (30 + 270)$) of the original 30 can be also miRNA-mRNA interactions not taking place on the disease. The 26 ($30 - 4$) resting ones can be interactions taking place on the disease. We do not know which interactions are taking place on the disease or not,

but we hypothesise that we can find 13.3% of false positives (4/30) on the original 30 miRNA-mRNA miRComb miRNA-mRNA pairs output.

The drawback of this method is that, as it is based on a specific cutoff (usually $p = 0.05$), the computations can vary dependent on this chosen cutoff. Other alternatives can be proposed that are not dependent on any cutoff.

3.9.2 Logistic Regression

Logistic regression is another way to measure relations between a binary response and a predictor, which can be numerical, categorical or ordinal. In this case, the response is if a miRNA-mRNA pair is a correlation predicted by the database and the predictor is the value of the correlation. Although using this method we are not able to estimate the number of false positives, we will be able to estimate if there is any relation between "being predicted by the database" and "show a negative correlation".

The logistic regression can be expressed as:

$$g(F(x)) = \ln\left(\frac{F(x)}{1-F(x)}\right) = \beta_0 + \beta_1 x \quad (3.18)$$

Where $F(x)$ is the probability that a subject is a target, β_0 and β_1 are constants and x is the value of the correlation. If the value of the correlation is related to the probability of being a target, then β_1 will be different from 0. Specifically, β_1 must be negative, as less correlation equals more probability of being a target. The area under the curve (AUC) can be used to assess the predictive value of these models.

As an example, Figure 3.11 shows simulated data where 200 miRNA-mRNA pairs have a basal probability of 20% of being miRNA-mRNA interactions (correlations ranging from -1 to 1), and we 50 added significant negative correlations (from -1 to -0.75 in this case) that have a probability of 80% of being miRNA-mRNA interactions.

β_0 and β_1 are 0.253 ($p = 7.26e-16$) and -0.225 ($p = 9.81e-07$) respectively. ROC curve (Figure 3.11(b)) is also able to detect this difference of probabilities.

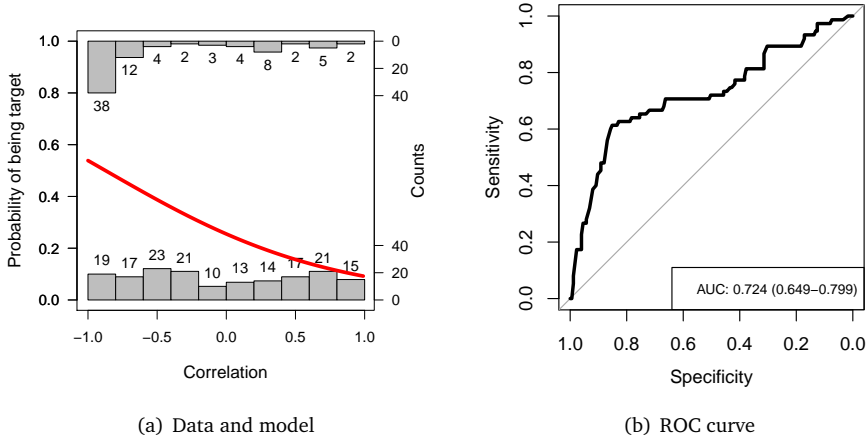


Figure 3.11: (a): visual representation of data. (b): receiver operating characteristic (ROC) curve and AUC including 95% confidence interval.

3.9.3 Gene Set Enrichment Analysis

Gene Set Enrichment Analysis (GSEA) is a computational method originally developed to determine whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states [181].

GSEA system sorts a list of genes according to different criteria (FoldChange between two characteristics, correlation with a biomarker, etc.) and checks if the genes belonging to a specific pathway (S) are systematically located on the top or bottom of this list (L). In this case, similarly to logistic regression, this method allows us to check if the miRNA-mRNA interactions predicted by the databases are preferentially located in negative correlations.

The score is calculated by walking down the list, increasing a running-sum statistic when we encounter a gene in the pathway and decreasing it when we encounter genes not present on it. The magnitude of the increment depends on the correlation of the gene with the phenotype if it is available (other options are available if no sorting score is reported –Supporting information of [181]–).

The walking sum is computed as follows: for i in $1, \dots, N$ where N is the length of the list L and N_H the length of the pathway S .

$$P_{\text{hit}}(S, i) = \sum_{g_j \in S, j \leq i} \frac{|r_j|^p}{N_R}, \text{ where } N_R = \sum_{g_j \in S} |r_j|^p$$
$$P_{\text{miss}}(S, i) = \sum_{g_j \notin S, j \leq i} \frac{1}{(N - N_H)} \quad (3.19)$$

The ran is $P(S, i) = P_{\text{hit}}(S, i) - P_{\text{miss}}(S, i)$. The enrichment score (*ES*) is the maximum deviation from zero encountered in the ran. *P* values are computed using permutation tests.

When more than one pathway is evaluated (*S*) additional coefficients are also computed. In this cases, a Normalised Enrichment Score (*NES*) is computed as follows $NES = \frac{ES}{\text{mean (all permutations of } ES)}$. FDR-like values are also computed but not recommended for general use. This value is based on median null distribution and for example when it is zero indicates that observed scores are larger than the values obtained by at least half of the random permutations, similarly to the cases when $FDR < 0.25$. Thus, when required, a cutoff of $FDR < 0.25$ offers reasonable results [181].

Using the same example of the logistic regression, Figure 3.12 shows the ran $P(S, i)$ and Enrichment Score of the same data using `fgsea` R/Bioconductor package [182], which implements GSEA analysis. Maximum positive ran is 0.052, and the minimum is -0.625. The greatest deviation from zero is -0.625, so this value is the Enrichment Score. *P* value is 0.011. This means that this method also detects the concentration of targets preferentially located on the negative correlations (highest ranks, as they are decreasingly sorted).

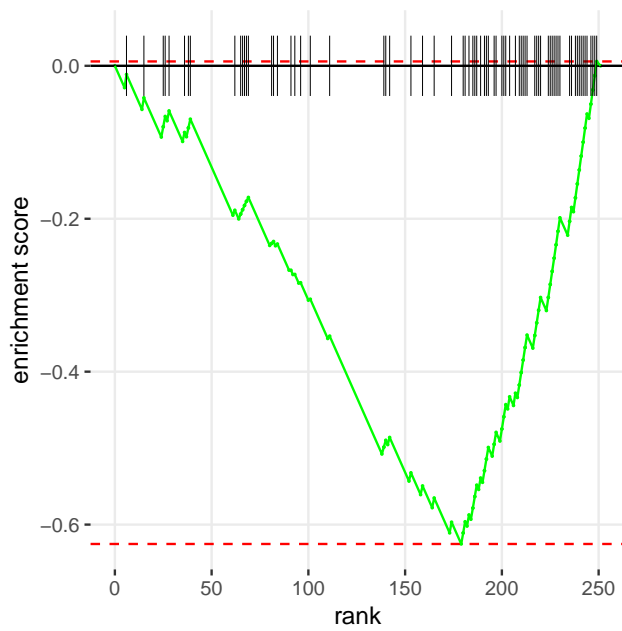


Figure 3.12: GSEA plot using `fgsea` R/Bioconductor package [182] showing the *ES* of simulated data from Figure 3.11(a).

3.10 Additional pipelines

Other types of data can be analysed with miRComb. In this cases, small variations of the pipeline should be done, which are described below:

3.10.1 Time-series analysis

Sometimes, instead of two groups, the data can consist of the expression of miRNAs and mRNAs measured in a group of samples across different times.

In this case, the step of filtering for deregulated miRNAs and mRNAs is different. Subfigure 3.13(a) shows a custom example of a miRNA-mRNA pair in which the miRNA expression increases over time and the mRNA expression decreases over time. Then, in the correlation step, time-course samples are treated as the other cases, where each dot represents a sample in a specific time (Subfigure 3.13(b)). We presented two ways of filtering time-course experiments:

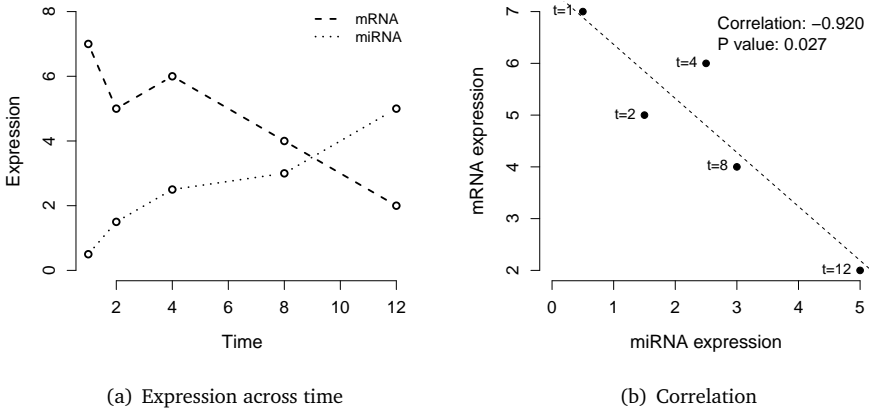


Figure 3.13: Example of a time analysis data. Analysis of one sample measured at times 1, 2, 4, 8, 12.

- **Final point.** Consists on analysing t_{fin} versus t_{init} . In this case we compare the expression of the miRNA or mRNA on the final time (t_{fin}) versus the expression of the miRNA or mRNA on the initial time (t_{init}).

$$\begin{cases} H_0 : \mu_{t_{\text{init}}} = \mu_{t_{\text{fin}}} \\ H_1 : \mu_{t_{\text{init}}} \neq \mu_{t_{\text{fin}}} \end{cases}$$

We need to have more than one sample analysed if we want to obtain a p value. This option of analysis is useful when: 1) the evolution of the miRNA or the mRNA is not linear across time, 2) we are only interested in the difference between initial time and final time, or 3) we have only analysed two points. The FoldChange equivalent is the value of $t_{\text{fin}} - t_{\text{init}}$. In the example exposed on Figure 3.13, the FoldChange for the miRNA is 4.5, and the FoldChange for the mRNA is -5. No p values can be associated to these "FoldChanges" as there is only one sample represented, but typical differential expression tests can be considered if more than one sample is measured.

- **Linear regression.** In this case we perform a linear regression of the expression of the miRNA or mRNAs versus time. The slope of this regression is recorded as the equivalent of the FoldChange, and the null and alternative hypothesis to test is if the slope is 0.

$$\begin{cases} H_0 : \text{slope (miRNA(or mRNA) } \sim \text{ time)} = 0 \\ H_1 : \text{slope (miRNA(or mRNA) } \sim \text{ time)} \neq 0 \end{cases}$$

For this version of the analysis, one sample measured in different time-points is enough to perform the analysis. If there are more samples available they are also be included in the computation. In the example exposed on Figure 3.13, the slope for the miRNA is 0.360 with an associated p value of 0.00625. This means that the miRNA increases its expression in 0.36 units per unit of time, and this value is significantly different from 0. The slope of the mRNA is -0.39 with a p value of 0.0225.

In both cases, the phenotypical file that will be used for miRComb package should be designed according to the times when the samples are measured.

3.10.2 Non-matched miRNA-mRNA data

When only **non-matched miRNA-mRNA data** is available, we suggest using FoldChanges instead of correlation values. Although it is less powerful than correlation, in most of the cases it will also lead to the same conclusions. Slope function summarises the relation between both FoldChanges:

$$\text{score} = -(\text{logratio}_{\text{miRNA}} \cdot \text{logratio}_{\text{mRNA}})$$

Figure 3.14 shows the value of the Score according to both logratios. If the logratios have opposite sign then the score is positive. The higher both logratios are in absolute terms, the higher the score value is. The score is available in miRComb since its first version, but we have also included the possibility to perform the integrative analysis based on selecting one score cutoff instead of the correlation value and its associated p value.

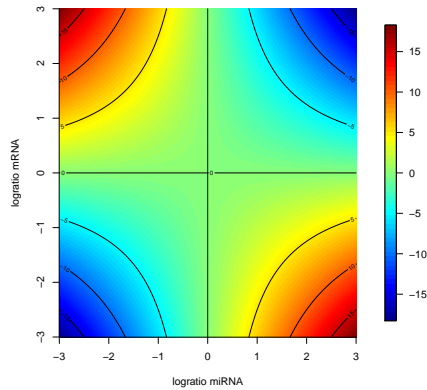


Figure 3.14: Score function.

Chapter 4

Results

4.1 MiRComb R package

MiRComb package is built in R, which is probably one of the softwares more used in biological data analysis. Apart from R code, the package uses C++ code used for the `addCorrelation` function and \LaTeX [183] and Sweave [184] for the `mkReport` function. We released an initial package which has been loaded to Sourceforge, which offered free hosting for the package source files and successive releases, plus the option to store a corporative webpage of the project. Moreover, in order to increase its visibility and make it easier to share the code with other bionformaticians, `miRComb` and `miRData` have been recently added to **GitHub**: <https://github.com/mariavica/mircomb> and <https://github.com/mariavica/mirdata>.

The package is designed using **object-oriented programming** (OOP): it creates an object and adds all the features (attributes) needed. In programming language, the difference between an object and a list (which can also contain "attributes") is that we can define methods to the object. Methods are generic functions that change their behaviour according to the object. For example, the `plot` function in R makes different plots depending on the type of data/object that we are plotting: a scatter plot for an object containing numeric data, a boxplot for an object containing numeric and categorical data, a residuals plot for the object resulted from a linear regression, etc. Having designed the package using OOP allows us to use generic functions if desired. In `miRComb` we have designed a new class called `corObject`. This class creates an object that contains all the information needed for miRNA-mRNA analysis, and successive layers of information

are added to it during the analysis pipeline.

Figure 4.1 shows how the `miRComb` package is organised: in fact, there are two packages: `miRData helper` package: which contains databases and has a size of 134.7 MB; and `miRComb` itself, which contains the functions and some example data and has a size of only 5.3 MB (version 0.8.5). They are organised in this way because main updates will be on `miRComb` package, while `miRData` is supposed to be more stable. `miRData` version 0.6 contains information from `microCosm`, `TargetScan`, `miRSVR` and `miRDB` databases, genome coordinates for human miRNAs and mRNAs and HUGO Gene Symbol to Entrez Gene Human translator.

`miRComb` functions can be divided into several categories:

- `CorObject` is a new **class** that stores all the information needed for the analysis. Minimum requirements of the class are miRNA expression matrix, mRNA expression matrix, sample information of miRNA samples, and sample information of mRNA samples.
- **Analysing functions** are those that add information to the `corObject` class or filter the data. They can add slots to the `corObject` class. A slot is a new layer of information and can be: the output of differential expression analysis of the miRNAs, the output of differential expression analysis of the mRNAs, the matrix of correlation values, the list of miRNA-mRNA interactions (which can also include features), or more complex structures containing functional information or summary of the analysis performed.
- **Plotting functions** are those aimed to explore the data on a visual way. They are options for plotting miRNA or mRNA datasets, individual miRNA-mRNA interactions or networks of miRNA-mRNA interactions.
- Finally, there are other functions aimed to **summarise and export** the information. `MkReport` function is the most representative, as it makes pdf report summarising the main findings, but the list of miRNA-mRNA interactions can also be exported to excel or csv files.

Regarding **time-series analysis**, as well as **non-matched miRNA-mRNA data**, we created an alternative vignette called *Additional Examples* (Appendix B.2), which explains how these analysis can be done. For the time-analysis, `addLong` function, was included in a later release of the package, and for non-matched miRNA-mRNA data, the use of a score cutoff is suggested.

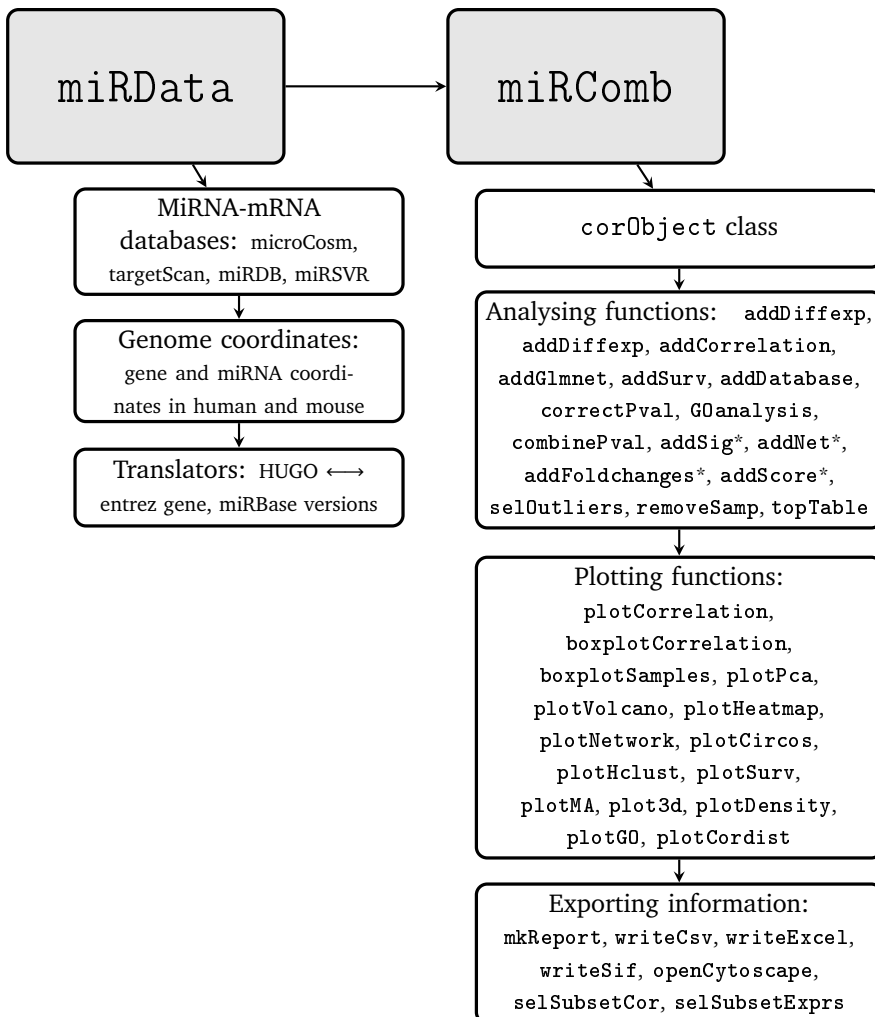


Figure 4.1: Organization of miRComb functions. *Complementary functions: they are used for preparing the data for other functions and are needed for the analysis.

4.1.1 MiRComb statistics

Since miRComb was made public in June 2015, there have been constant downloads of files located on the miRComb project webpage (Table 4.1). The files available on the project are: miRComb and miRData R packages (both in Windows and Linux/Mac versions), two Vignettes, an example data for longitudinal analysis and a README file.

During this period we have collaborated in three studies that used miRComb. In over-

all, the papers that involve miRComb or miRComb analysis are:

1. Coll M, Taghdouini AE, Perea L, Mannaerts I, Vila-Casadesús M, Blaya D, Rodrigo-Torres D, Affò S, Morales-Ibanez O, Graupera I, Lozano JJ, Najimi M, Sokal E, Lambrecht J, Ginès P, van Grunsven LA, Sancho-Bru P. *Integrative miRNA and Gene Expression Profiling Analysis of Human Quiescent Hepatic Stellate Cells*. Scientific Reports 2015 Jun; 5:11549 [185].
2. Bofill-De Ros X, Santos M, Vila-Casadesús M, Villanueva E, Andreu N, Dierssen M, Fillat C. *Genome-wide miR-155 and miR-802 target gene identification in the hippocampus of Ts65Dn Down syndrome mouse model by miRNA sponges*. BMC Genomics. 2015;16:907 [186].
3. Vila-Casadesús M, Gironella M*, Lozano JJ*. *MiRComb: An R Package to Analyse miRNA-mRNA Interactions. Examples across Five Digestive Cancers*. PLoS ONE 11(3): e0151127. 2016. doi:10.1371/journal.pone.0151127 [187].
4. Blaya D, Coll M, Rodrigo-Torres D, Vila-Casadesús M, Altamirano J, Llopis M, Graupera I, Perea L, Aguilar-Bravo B, Díaz A, Banales JM, Clària J, Lozano JJ, Bataller R, Caballería J, Ginès P, Sancho-Bru P. *Integrative microRNA profiling in alcoholic hepatitis reveals a role for microRNA-182 in liver injury and inflammation*. Gut. 2016 May; p.gutjnl-2015-311314 [188].
5. Vila-Casadesús M, Vila-Navarro E, Raimondi G, Fillat C, Castells A, Lozano JJ, Gironella M. *Deciphering microRNA targets in pancreatic cancer using miRComb R package. (manuscript in preparation)*

The geographical origin of the people downloading miRComb is varied, being United States the country with more downloads: 49% of them. The following countries are Spain (7.8%), Germany (6.4%) and China (6.0%). There are a total of 40 countries with at least one download, and 14 with more than 30 downloads. The maximum number of downloads per month was in March 2016, coinciding with the publication of miRComb's article in PLoS ONE [187].

Sourceforge allowed also to upload a webpage (a snapshot is shown in Figure 4.3), that we used for advertising the package and the links to miRTools (that were stored in our server). Regarding webpage visitors, Figure 4.4 shows the number of visitors of <http://mircomb.sourceforge.net> per day during 1-month period. Visitors came mostly referred from PLoS ONE publication [187], but also sporadically from the other publications: Coll et al., 2015 [185], Bofill-De Ros et al., 2015 [186] and Blaya et al., 2016 [188], or from Google direct search.

Month	Total downloads	Related articles
2015-06	13	Study that uses miRComb [185]
2015-07	99	
2015-08	55	
2015-09	50	
2015-10	19	Study that uses miRComb [186]
2015-11	54	
2015-12	42	
2016-01	54	
2016-02	28	
2016-03	783	MiRComb release article [187]
2016-04	34	
2016-05	129	
2016-06	69	Study that uses miRComb [188]
2016-07	117	
2016-08	122	
2016-09	182	
2016-10	197	
2016-11	32	
2016-12	93	First miRComb citation [189]
2017-01	83	
2017-02	108	
2017-03	118	
2017-04	107	

Table 4.1: Number of miRComb’s project files download since the publication of the first article mentioning miRComb R package.

4.1.2 MiRTools

During this thesis we also developed other tools apart from miRComb. These functions have also been integrated into the package: `translate` and `checkmiRNAs` for miR-Translator and `plotCircos` for miRCircos, but we also considered that they can work separately so we have built two on-line tools that are aimed to deal with these specific parts of miRNA analysis.

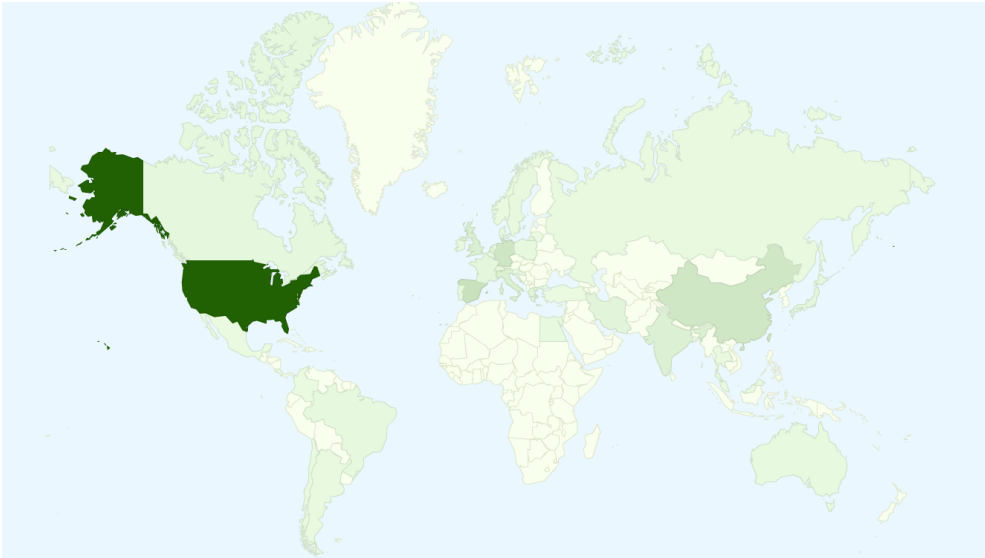


Figure 4.2: Map showing the precedence of the miRComb and associated files downloads. Countries green-filled have at least one download. Dark green means more downloads. There were a total of 2588 downloads during this period.

4.1.2.1 MiRTranslator

We saw that miRBase [14] has been changing the names of the miRNAs across its successive releases, apart from including (and sometimes removing) new miRNAs. The most relevant change was from version 17 to 18, when almost all miRNA names were changed from the style miR-X and miR-X* to miR-X-5p and miR-X-3p. However, not all the miRNAs were changed according to this criteria (sometimes -3p and -5p was inverted, or not included) and other miRNAs showed different changes.

Some of these changes are spotted on Table 4.2. In this table, we selected miRNAs that did not follow the traditional rules, that shows why identifying the version of a set of miRNAs may not be always entirely straightforward. For example *hsa-miR-101* and *hsa-miR-27a* took the "-3p" ending in miRBase version 18, when in most of the cases, is the form with the asterisk that takes the "-3p" ending. *Hsa-miR-28-3p* or *hsa-miR-193a-5p* were not found on early versions of miRBase, and used the "-3p" ending earlier than others, while *hsa-miR-216* and *hsa-miR-301* were renamed to *hsa-miR-216a* and *hsa-miR-301a* and added the "-5p" and "-3p" endings later than the others. Moreover, although it is not usual, some miRNAs may be removed for miRBase, such as *hsa-miR-220*. A lot more of examples can be found on the full miRBase records.

MIMAT	miRBase_8.2	miRBase_12.0	miRBase_17	miRBase_18	miRBase_21
MIMAT0000084	hsa-miR-27a	hsa-miR-27a	hsa-miR-27a	hsa-miR-27a-3p	hsa-miR-27a-3p
MIMAT0000085	hsa-miR-28	hsa-miR-28-5p	hsa-miR-28-5p	hsa-miR-28-5p	hsa-miR-28-5p
MIMAT0004502		hsa-miR-28-3p	hsa-miR-28-3p	hsa-miR-28-3p	hsa-miR-28-3p
MIMAT0000087	hsa-miR-30a-5p	hsa-miR-30a	hsa-miR-30a	hsa-miR-30a-5p	hsa-miR-30a-5p
MIMAT0000099	hsa-miR-101	hsa-miR-101	hsa-miR-101	hsa-miR-101-3p	hsa-miR-101-3p
MIMAT0000231	hsa-miR-199a	hsa-miR-199a-5p	hsa-miR-199a-5p	hsa-miR-199a-5p	hsa-miR-199a-5p
MIMAT0000232	hsa-miR-199a*	hsa-miR-199a-3p	hsa-miR-199a-3p	hsa-miR-199a-3p	hsa-miR-199a-3p
MIMAT0000273	hsa-miR-216	hsa-miR-216a	hsa-miR-216a	hsa-miR-216a	hsa-miR-216a-5p
MIMAT0000277	hsa-miR-220	hsa-miR-220a			
MIMAT0000458	hsa-miR-190	hsa-miR-190	hsa-miR-190	hsa-miR-190a	hsa-miR-190a-5p
MIMAT0004614		hsa-miR-193a-5p	hsa-miR-193a-5p	hsa-miR-193a-5p	hsa-miR-193a-5p
MIMAT0000459	hsa-miR-193a	hsa-miR-193a-3p	hsa-miR-193a-3p	hsa-miR-193a-3p	hsa-miR-193a-3p
MIMAT0000688	hsa-miR-301	hsa-miR-301a	hsa-miR-301a	hsa-miR-301a-3p	hsa-miR-301a-3p

Table 4.2: Portion of the database which contains the name of the miRNAs across versions. Non consecutive miRBase versions have been picked in order to highlight the changes.

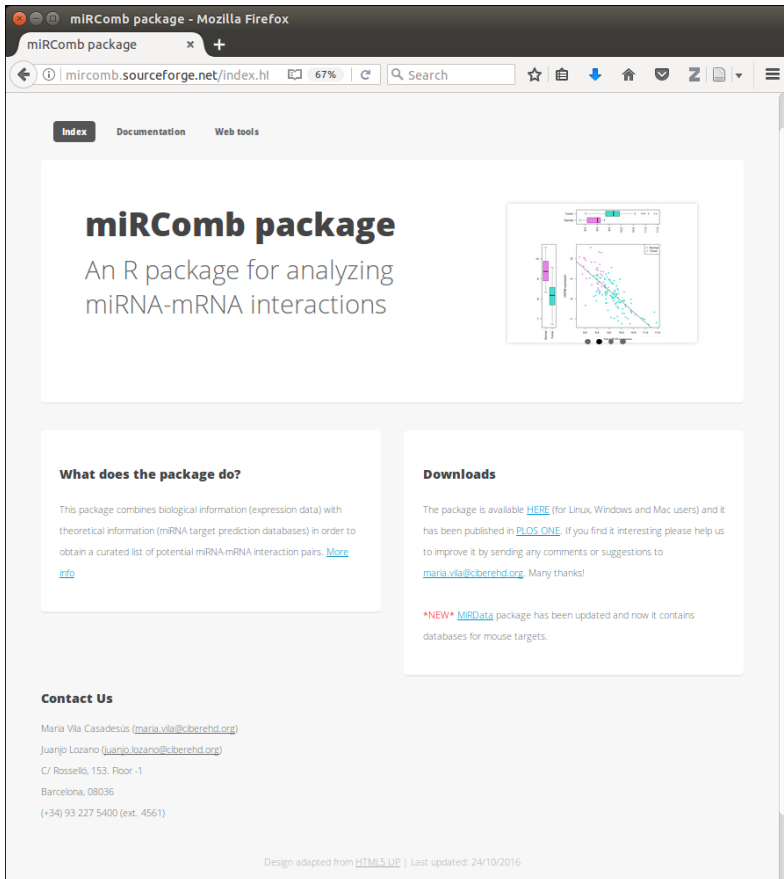


Figure 4.3: Caption of miRComb webpage <http://mircomb.sourceforge.net>.

These non-standard changes make difficult to compare miRNAs from different versions, even to look for miRNAs targets in different databases. For that reason, we considered interesting to build a small tool to help on that process.

The interface of the tool is simple (Figure 4.5) and allows to put the miRNA names in your version, and then output the names in the desired version. It is similar to `miRNAme-Converter` package [134] but web-based, which makes it better for the standard user and includes the option "*I don't know*" when selecting the initial miRBase version.

The page is built using PHP language. Once the name(s) of the miRNA(s) is/are taken, they are sent to a MySQL [191] database on the server (which contains the table with the mirna names). A portion of this table is shown in Table 4.2. Using the information of Table 4.2, miRTranslator is able to obtain the MIMAT of a miRNA (once given its original



Figure 4.4: Graph of visitors per day of the miRComb webpage (<http://mircomb.sourceforge.net>). 152 visitors and 325 actions in this 1-month period. Extracted from Clicky Real Time Web Analytics [190].

version) and return the name of the miRNA in any version. The tool returns the output on the screen, plus a csv file including both initial and translated name. Optionally, if provided, an e-mail with a copy of the link of the results is also sent to the person making the request.

When the user selects the option "*I don't know*", the tool does a previous step to identify the miRBase version of the input miRNAs. It counts how many miRNAs match with each of the miRNA version and the picks the one with more matches. In the case of a tie, it selects the most recent one. Then, the tool uses this predicted version as input version and uses it to make the appropriated translations.

4.1.2.2 MiRCircos

A chord diagram is a graphical method of displaying the inter-relationships between data in a matrix. The data is arranged radially around a circle and the relationships between the points typically drawn as arcs connecting both points. In our case, the circle represents genomic locations, and the arcs connect a miRNA and a mRNA from a miRNA-mRNA interaction. Circos [192] is a visualisation tool that allows to draw and customise chord diagrams. Circos has also been implemented in R circlize package [193], which is the

Index **Documentation** **Webtools**

MiRNA conversor

Paste your list of miRNAs (For example: hsa-miR-21, version 17)

```
hsa-miR-21
hsa-miR-21*
hsa-miR-148a
```

MiRBase version original name

MiRBase version desired name

Your email:

Successful send! An e-mail with the results has also been sented to you.
This is your conversion table:

mirbase version 17	mirbase version 21
hsa-miR-21	hsa-miR-21-5p
hsa-miR-148a	hsa-miR-148a-3p
hsa-miR-21*	hsa-miR-21-3p

The csv file with the conversion pairs is available [here](#)

Figure 4.5: Web interface of the miRNA versions translator. <http://bioinfo.ciberehd.org/mircomb/conversion.html>

package that we use to create the plot in this application.

Figure 4.6 shows the steps of the tool we implemented in our web server. The input query is the name(s) of the miRNA(s) that will be analysed. This name is used for querying the MySQL database, which outputs two tables (**I** and **II** from Figure 4.6): the first contains the miRNA-mRNA pairs (first column for the miRNA and second for the mRNA); and the second the genome coordinates for each miRNA and mRNA (first column the name of the miRNA or mRNA and the second the coordinates –chromosome and base number–). Then, an R script processes all this information and gives the following outputs:

1. A circos plot (plot **a** from Figure 4.6) where lines represent miRNA-mRNA interactions and a track which highlights the regions significantly enriched in target mRNAs (hypergeometric test with $\alpha = 0.05$ used).
2. A table with all the miRNA-mRNA interactions and their corresponding genome coordinates (table **III** from Figure 4.6).
3. An excel file containing the track analysis (table **IV** from Figure 4.6). Each row represents a region (has its chromosome, starting and ending) and has its associated p value indicating if it has more targets than expected or not. The excel file contains one table per sheet, where each table represents one grouping (for example, one sheet per chromosome divisions, and other sheet is a custom bed is included, etc.).
4. (optional) The user receives an email with the links that allow him to download all the tables and plot.

Figure 4.7 shows the output of *hsa-miR-21-5p*. In this version, mRNA targets are predicted using TargetScan. The regions to make the comparison can be defined by the user. In this plot, several layers (entire chromosomes, arms or cytobands) are simultaneously plotted. For example, when we checked chromosomes we saw that chromosomes 2 and 4 were enriched. But other regions can be also highlighted using other divisions: for example, the plot highlights that the p arm of chromosome 12 is enriched, as well as the q arm of chromosome 4 (but not the p arm). And we also observed that the enrichment of chromosome 2 is due by targets located in different cytobands. Apart from that, and e-mail is sent to the user with a copy of the link to download the results.

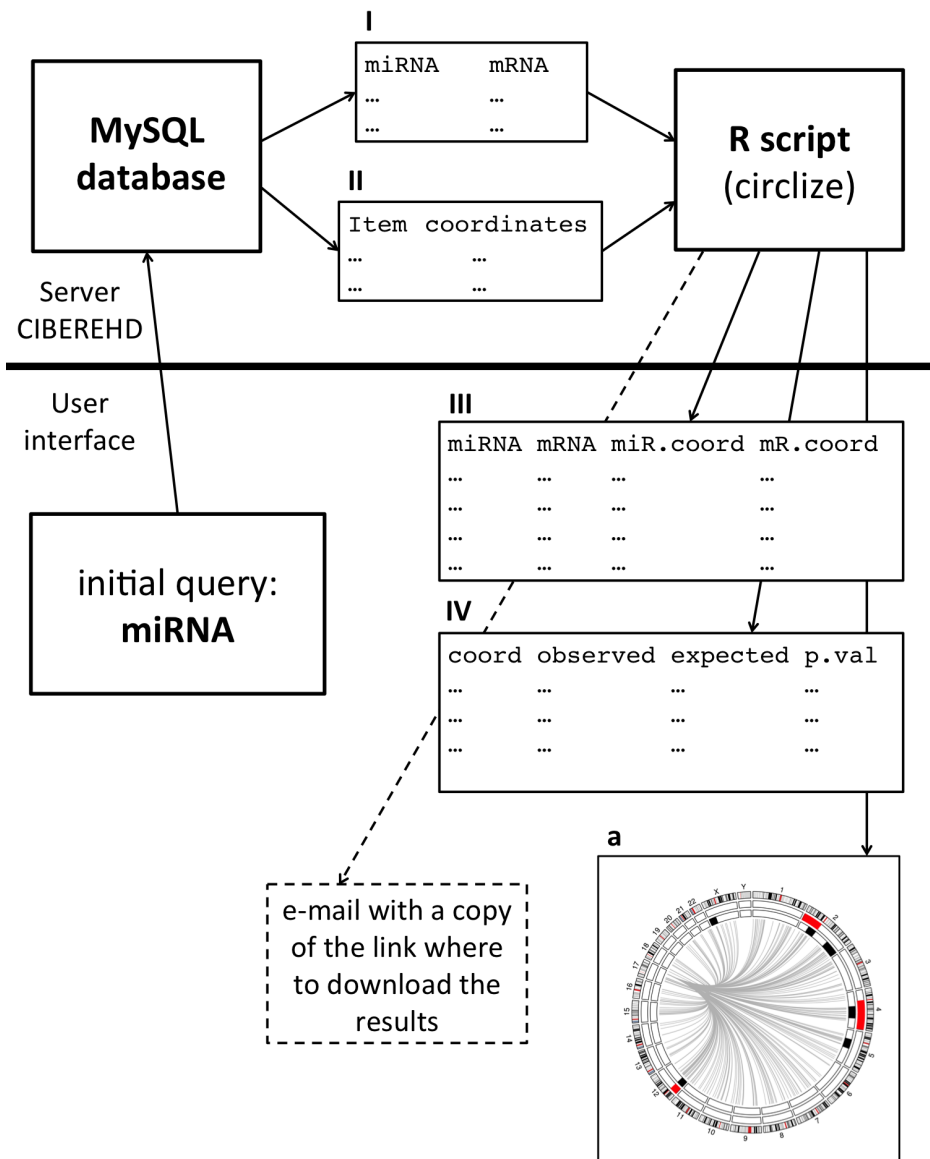


Figure 4.6: Underlying schema of the miRCircos tool. Coordinates include chromosome, start and end.

Index Documentation Webtools

Circos plot - enriched region finder

[Still in test. If it doesn't work, you can report any bug to maria.vila@ciberehd.org indicating the error message. Many thanks!]

Paste your miRNA in plain text (miRBase version 18!* (only human miRNAs) For example: hsa-miR-21-5p or hsa-miR-200c-3p)

hsa-miR-21-5p

Which regions do you want to look at? (select at least 1)

- Cytobands (f.ex p11.1)
- Cytobands grouped level 1 (f.ex p11.x)
- Cytobands grouped level 2 (f.ex p1x.x)
- P and Q arms
- Chromosomes
- Custom bed No file selected.

Your email (optional):
(if desired, a link with the results will be sented to you)

* If you need to convert your miRNA names from another version of miRBase to miRBase version you can use this converter: [[Tab-separated](#) | [Excel \(.xls\)](#)]

Successful send! An e-mail with the results has also been sented to you.
The plot is available [here](#)
The csv file with the miRNA-mRNA pairs is available [here](#)
The excel file with the track analysis is available [here](#)

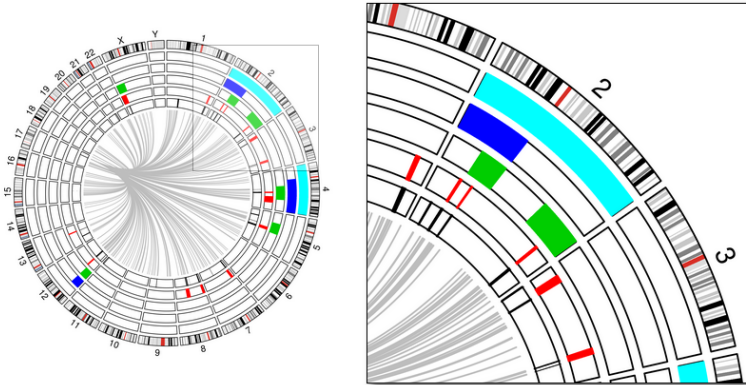


Figure 4.7: Web interface of the miRCircos tool. MiRCircos can be accessed here: http://bioinfo.ciberehd.org/mircomb/index_circos.html

4.2 MiRComb parameters exploration

For studying the implications of the different choices that can be done using miRComb, we used liver cancer mRNA-sequencing data from TCGA [154]. We selected a balanced and medium-small subset (49 cases and 49 controls) and we explored the different methods implemented on the miRComb package.

4.2.1 Differential expression methods

Raw sequencing data was downloaded from TCGA portal. Matched normal-tumour samples were selected and different procedures were applied:

- **Voom+T-test** (or T-test): voom data transformation plus a T-test considering equal variances among the two groups. Then, multiple testing correction according to Benjamini & Hochberg method (FDR).
- **Voom+Wilcoxon** (or Wilcoxon): voom data transformation plus a Wilcoxon test. Then, multiple testing correction according to Benjamini & Hochberg method (FDR).
- **Voom+limma-trend** (or limma, limma-trend): voom data transformation and limma-trend procedure.
- **Voom+RankProd** (or RankProd): voom data transformation and RankProd.
- **DESeq**: DESeq standard procedure: estimate size factors using the median, and local fit for the dispersion parameter.
- **EdgeR**: edgeR standard procedure: estimate common dispersion plus tagwise dispersion and then fit a negative binomial method.

Figure 4.8 shows the number of differentially expressed genes ($FDR < 0.05$) according to each method. Limma, T-test and Wilcoxon were the methods that found more differentially expressed mRNAs. EdgeR found a large number of mRNAs too, while RankProd and DESeq detected half of the mRNAs compared to limma.

As we do not know the true differentially expressed mRNAs, it is difficult to estimate the number of false positives/specificity, but previous studies suggest a similar performance between the different methods [137, 164, 163], although results may vary depending on the used dataset and its characteristics.

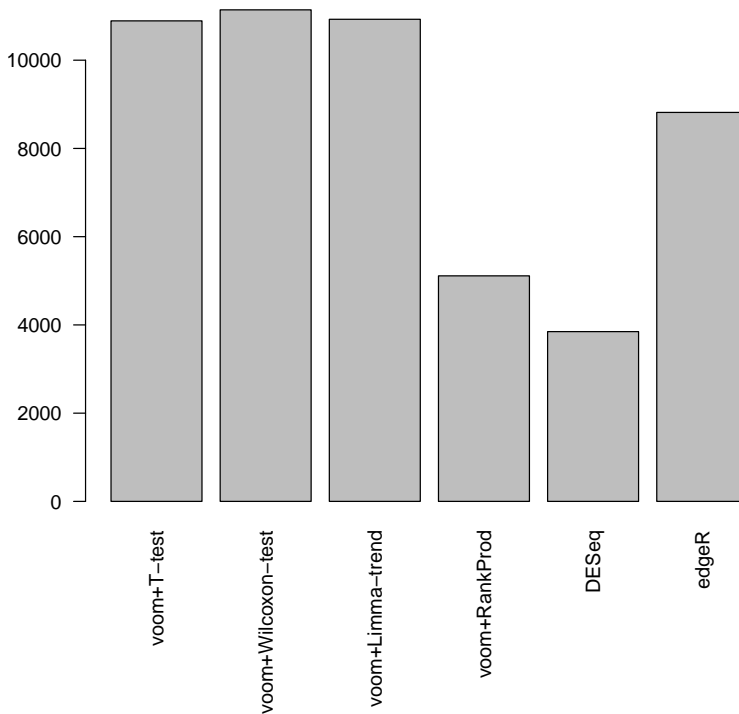


Figure 4.8: Total number of differentially expressed mRNAs (FDR < 0.05) using voom+T-test, voom+Wilcoxon, voom+limma-trend, voom+RankProd, DESeq and edgeR procedures.

Studying the reproducibility across the different methods (Figure 4.9), all the mRNAs detected by T-test were also detected for limma-trend procedure (except for 3 mRNAs). There were 290 mRNAs detected by limma that were not found on Wilcoxon, and 504 mRNAs that were detected on Wilcoxon but not limma. A vast majority of the differentially expressed mRNAs (10618) mRNAs were shared across these 3 methods. Based on that, t-test and Wilcoxon were not selected for further analysis as they do not show performances significantly better than limma, and limma implementation offers more features.

Figure 4.9 also shows the differences between the 4 methods specific for gene expression: limma, RankProd, DESeq and edgeR.

The DESeq and edgeR methods seem to be a little less specific detecting low expressed mRNAs than limma. In fact, 5.0% of the deregulated mRNAs by DESeq have a mean

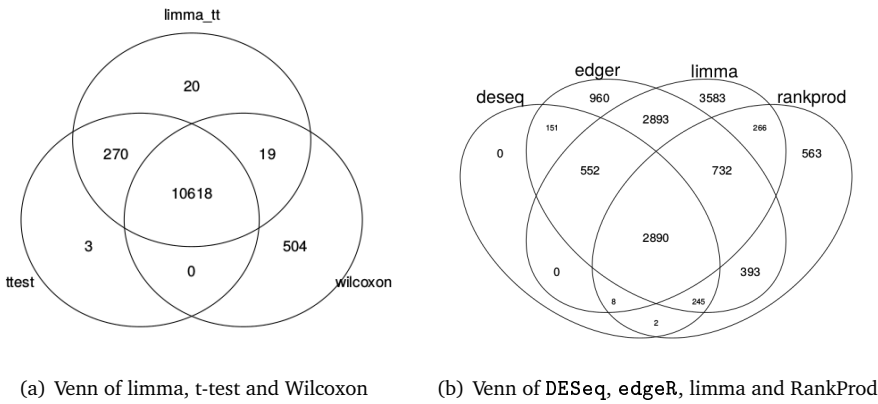


Figure 4.9: Venn diagram showing the number of differentially expressed mRNAs (FDR<0.05) shared between DESeq, edgeR, limma and RankProd procedures.

count of less than 40 counts (5.1% for edgeR), while 11.5% of the deregulated mRNAs by limma have a mean of less than 40 counts.

Figure 4.10 shows the difference between library size estimations. DESeq library size estimation, which is based on the median of counts, seems to be more related to total sample counts (correlation=0.66, p value=7.3e-14). EdgeR is based on the trimmed mean and is almost no related to original library size (correlation = 0.12, p = 0.238). Relative count sum after voom transformation (shown in orange) seems also to not be related to the original count sum (correlation = 0.15, p = 0.136).

4.2.2 Effect of subset selection

The next step on miRComb analysis is to filter for differentially expressed miRNAs and mRNAs. We explored which is the effect of subset selection.

Figure 4.11 shows the number of negative significant correlations (FDR < 0.05) depending on the filtering used. We can see that stringent filters lead to less significant negative correlations at the end. However, this is in part due to the fact that we are increasing the number of miRNAs and mRNAs to test.

When we compare the number of significant miRNA-mRNA correlations respect to the total number of computed miRNA-mRNA correlations we can observe that despite a slight decrease in the total number of final significant correlations, the proportion of

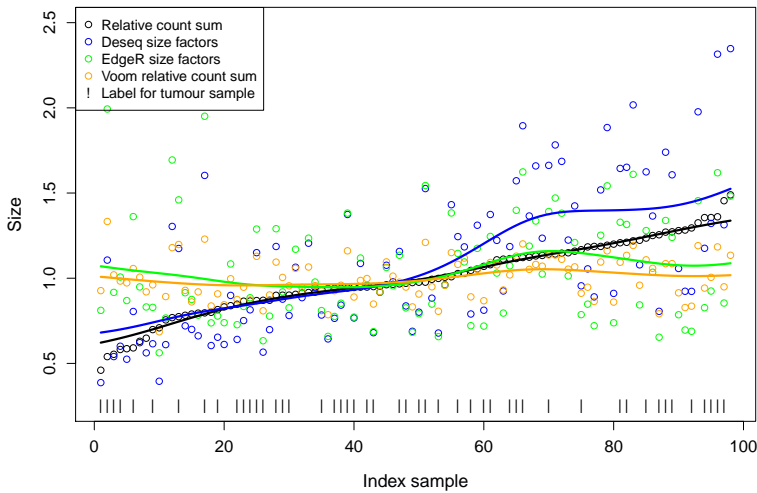
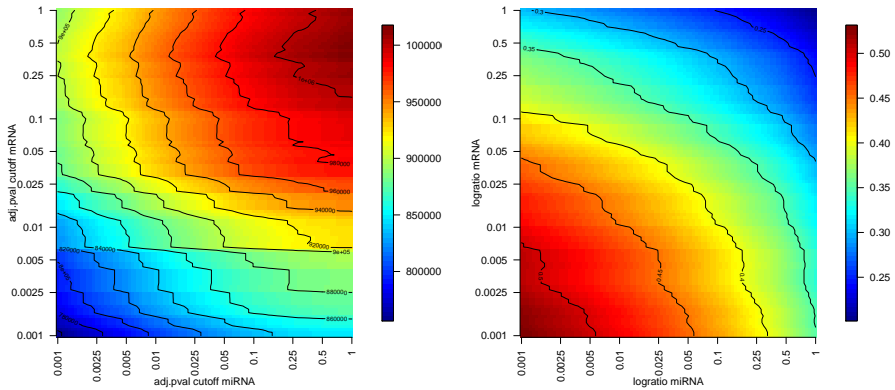


Figure 4.10: Plot showing the differences between library size estimations. Samples are sorted according to their relative count sum. Lines summarise the main trend by using kernel smoothed regression.



(a) Total number of significant miRNA-mRNA correlations (b) Proportion of significant miRNA-mRNA correlations

Figure 4.11: Contour plots showing the total number ((a)) or proportion ((b)) of significant ($FDR < 0.05$) miRNA-mRNA correlations depending on the filters used for selecting the miRNAs and mRNAs to correlate.

significant correlations increases. Figure 4.12 shows also this proportion: the proportion of negative (and thus significant) correlations increases when the sets have been strongly

filtered (lower FDRs).

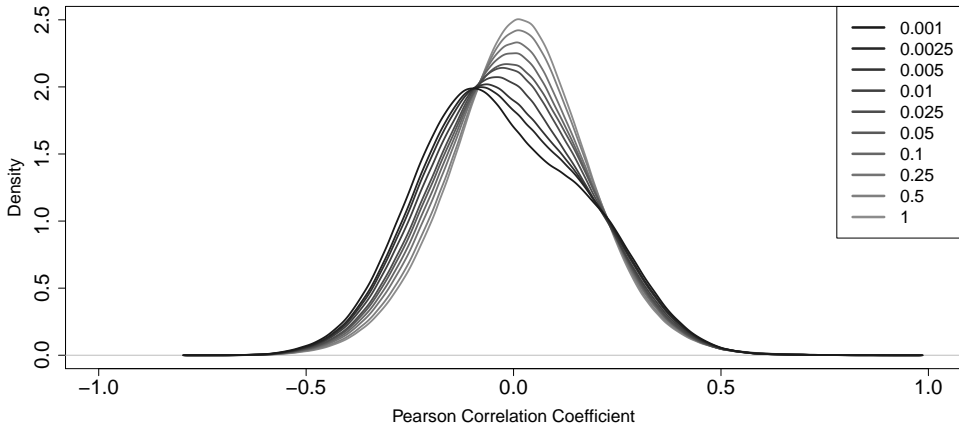


Figure 4.12: Density of the Pearson correlation coefficient of miRNA-mRNA pairs depending on the FDR filter applied to miRNA and mRNA dataset.

The results also replicate for Spearman and Kendall correlation estimates. It was not possible to perform a complete Glmnet-mRNAs estimation with all the miRNAs and mRNAs.

4.2.3 Pearson vs Spearman vs Kendall vs Glmnet

We first shown that voom [137] transformation gave us reasonable results. Voom transformation is also needed for estimating parametric measures of correlation (Pearson) and linear models (Glmnet-mRNAs and Glmnet-mRNA and Glmnet-miRNA) and (III)).

The filtering step is also mandatory to compare all these methods, as Glmnet-mRNAs cannot be performed using all the miRNAs and mRNAs. For these reasons, we used voom estimates plus we filtered the differentially expressed miRNAs and mRNAs ($FDR < 0.05$) in order to compare the different methods.

There are 142 significant miRNAs and 10927 significant mRNAs, which leads to a total of 1551634 miRNA-mRNA pairs that can be evaluated.

4.2.3.1 Correlation methods

First of all, we compared the different methods for measuring correlation: Pearson, Spearman and Kendall correlation coefficients. Figure 4.13 shows the distribution of the correlation coefficients and the number of significant negative correlations ($FDR < 0.05$).

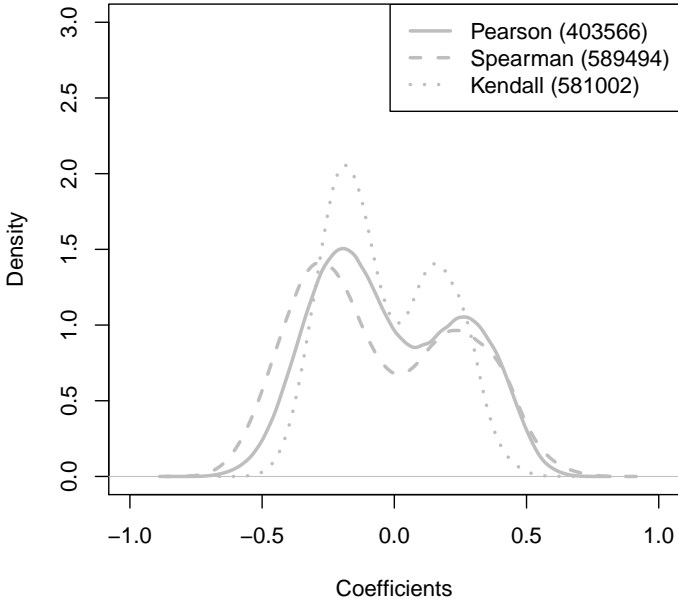
We can see that Kendall estimates are lower than Pearson or Spearman ones. Pearson and Spearman give similar coefficients. Figure 4.14 shows randomly selected correlations. Although Spearman estimates are slightly lower, the plots show that in general, a linear assumption is also reasonable to hold. Correlation between Pearson and Spearman estimates is 0.95. Different subsets (Subfigures 4.13(b) and 4.13(c)) show that Pearson and Spearman behave almost similar in small subsets, and Kendall estimates are systematically lower than the other two.

4.2.3.2 Glmnet vs Pearson estimates

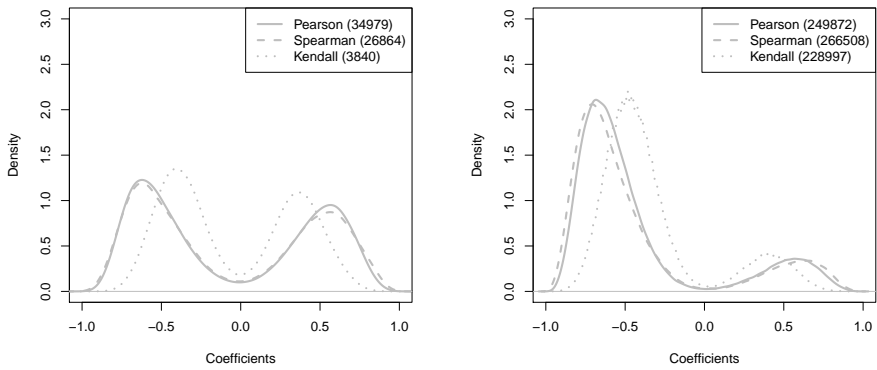
We also compared Glmnet versus the most similar correlation estimate, which is Pearson correlation, as it holds a linear assumption of the relation between the two variables (the miRNAs and the mRNAs). Figure 4.15 shows the relation between the coefficients of the different models of Glmnet and Pearson Correlation.

Glmnet-miRNA, apart from giving high coefficients (which can be corrected using feature scaling), gives coefficients not correlated either with Glmnet-mRNA or Glmnet-mRNAs. Glmnet-miRNA model assigns coefficient values equal to zero to 99.7% of the tested miRNA-mRNA pairs. In fact, Glmnet-miRNA model is "only" able to estimate 4044 miRNA-mRNA relations.

On the contrary, Glmnet-mRNA is able to estimate 68257 miRNA-mRNA coefficients (still a low number –4.4% of the total–, but greater than Glmnet-miRNA), and Glmnet-mRNAs estimates 363516 miRNA-mRNA coefficients –23.4% of the total–. However, Glmnet-mRNAs takes a long time and, in overall, gives similar results (correlation: 0.7) to Glmnet-mRNA (Figure 4.15(d)).



(a) All dataset: 98 samples, 142 miRNAs and 10927 mRNAs



(b) Subset 1: 12 samples, 66 miRNAs and 5534 mRNAs (c) Subset 2: 12 samples, 53 miRNAs and 8970 mRNAs

Figure 4.13: Density of the correlation coefficients.

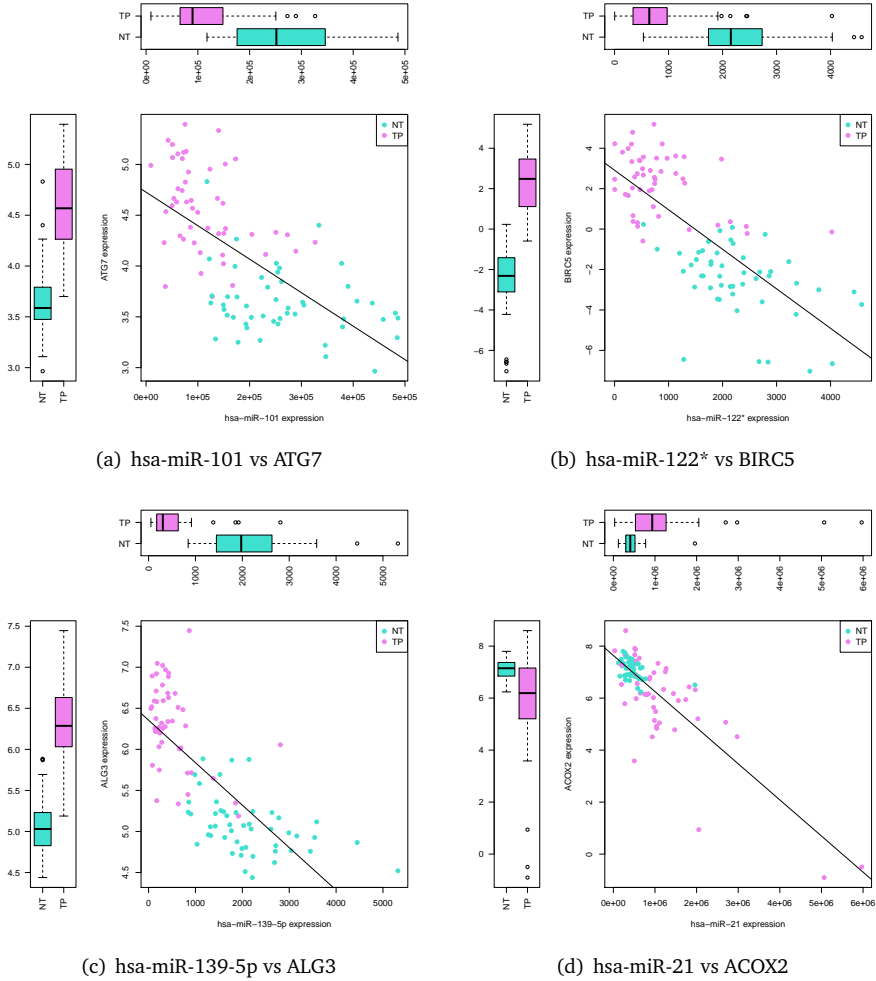


Figure 4.14: 4 randomly selected miRNA-mRNA correlations.

4.2.4 Integrative approaches

Finally, we compared the different methods used for the integration of external databases. Fisher and Stouffer p value combination can only be used with databases that report p values, while the intersection procedure can be used with any database.

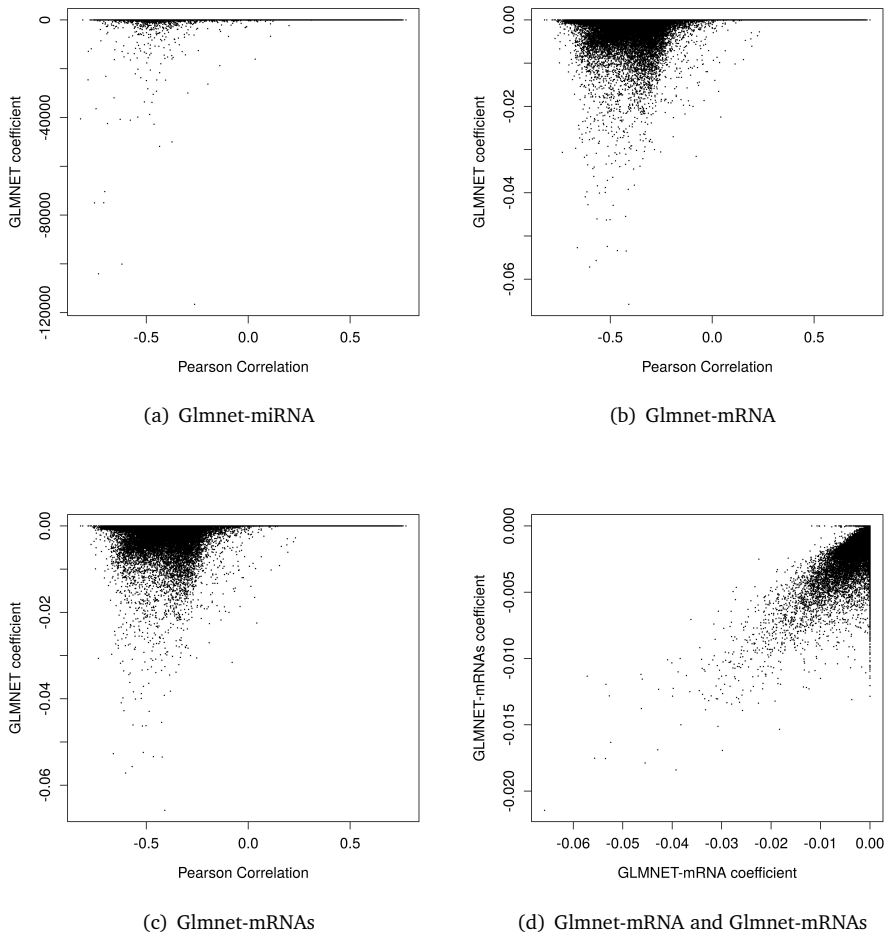


Figure 4.15: Comparison of different Glmnet models implemented on miRComb package.

4.2.4.1 Fisher vs Stouffer

First, we explored Fisher and Stouffer combination using MicroCosm database. Figure 4.16 gives an overall idea of the behaviour of the different p values. We have started the plot For example we can see that the lower p vaule of correlation ($p_{\text{Correlation}}$) is around $1e-16$ and increases slightly in these values (half order of magnitude), while the lower p value from MicroCosm ($p_{\text{MicroCosm}}$) is around $1e-7$, increasing only a little (0.5 orders of magnitude also). This means that in general $p_{\text{Correlation}}$ are several orders of magnitude

less than $p_{\text{MicroCosm}}$. However, we have to take into account that while p_{cor} range from 0 to 1, the p_{pred} has a small percentage of values below 0.05 but the rest (not reported by microCosm) are set to 1 as suggested by Gade et al. 2011 [120], which is not reflected in this figure.

Regarding the methods of combination, Subfigure 4.16(a) shows that combined p values fall between both initial p values to combine. Exceptions to this rule can occur when both original p values are small (some points at Figure 4.16(b)), and then the combined p value is smaller than both of the original ones.

Due the mathematical formulation, Fisher and Stouffer combination differ in one aspect that is worth to mention: when $p_{\text{Microcosm}}$ is 1 p_{Stouffer} is also 1, but p_{Fisher} takes a value several orders of magnitude greater than $p_{\text{Correlation}}$, but giving still significant ($p_{\text{Fisher}} < 0.05$).

In summary, from Figure 4.16 we can see that:

- If both p values are small (below 0.05 for example), both combined p values are smaller than any of the initial p values.
- If one of the initial p values is non-significant, bot combined p values take a value between both original p values, but:
 - p_{Fisher} values are usually smaller than p_{Stouffer} values (Fisher method is skewed to the smaller p value).
 - p_{Stouffer} is equal to 1 if one of the p values is 1, but p_{Fisher} can lead to *significant* results (for example $<1e-13$) even if one p value is equal to 1.

In practice, we do not want to base our decision on one large p value, we want that *both* premises are true. For that reason, even if Stouffer coefficients are higher than Fisher's, they are more reasonable and they perform better in our dataset.

Figure 4.17 shows also the same idea. In this figure we can differentiate the different behaviours described before:

- There is a group of p values that correlate (lower diagonal, starting at around $10e-5$). This group corresponds to low combined p values. We can say that if the p_{Stouffer} is lower than $10e-5$ is probably similar to p_{Fisher} (although p_{Fisher} is systematically slightly lower). This corresponds to low $p_{\text{Correlation}}$ and $p_{\text{MicroCosm}}$. There are 2031 miRNA-mRNA pairs in this group, which represent the 0.13% of the total.

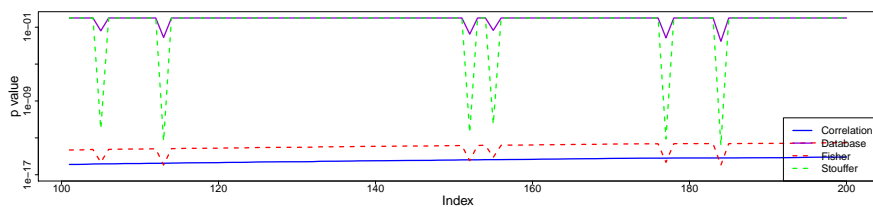
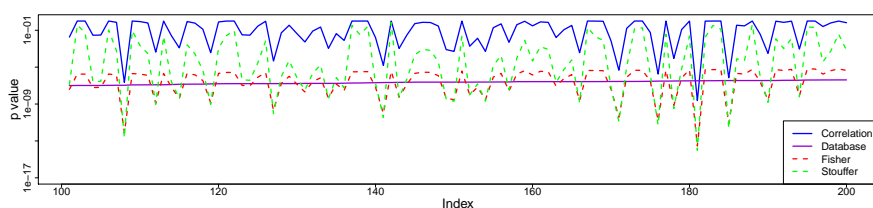
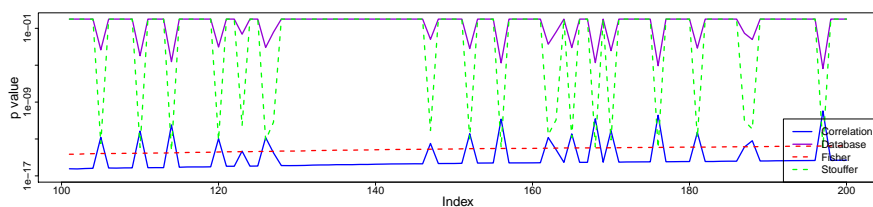
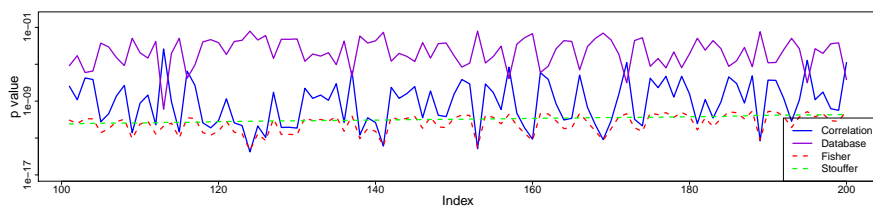
(a) Sort by $p_{\text{Correlation}}$ (b) Sort by $p_{\text{MicroCosm}}$ (c) Sort by p_{Fisher} (d) Sort by p_{Stouffer}

Figure 4.16: Sorting of the p values. Each plot shows the first 100 lower p values sorted by $p_{\text{Correlation}}$ (a), $p_{\text{MicroCosm}}$ (b), p_{Fisher} (c) or p_{Stouffer} (d). First 100 values were omitted because some of them contain p values equal to 0 due to floating point rounding in R, and could not be displayed on a y-log scale.

- On the top of the plot, there is another group of p values where p_{Stouffer} are larger than p_{Fisher} . At the same time, we can differentiate:
 - The upper line corresponds to p_{pred} equal to 1. In this case, the p_{Stouffer} is 1, but as we saw in Figure 4.16, p_{Fisher} can reach lower values. It is the case of the major part of the pairs represented in Figure 4.16(a). There are 1527080 pairs in this group, which represent the 98.4% of the total.
 - Near this line, there is the other subgroup, more dispersed than the others. These interactions represent pairs with one original p value (usually $p_{\text{MicroCosm}}$) several orders of magnitude greater than the other one. This group reflects better the skewing of p_{Fisher} to small p values. There are 22449 pairs in this group, which represent the 1.45% of the total.

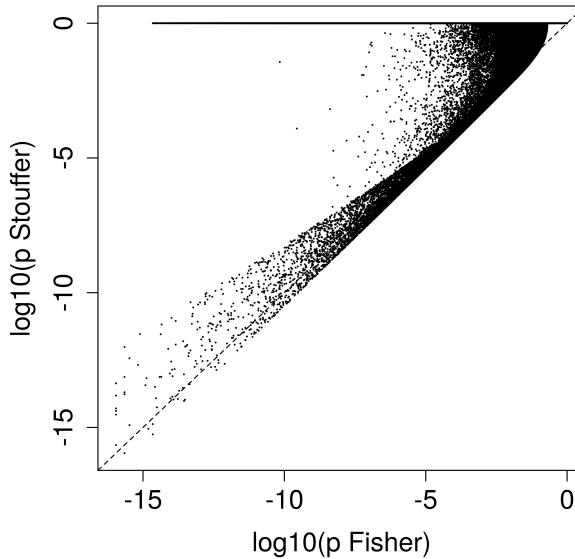
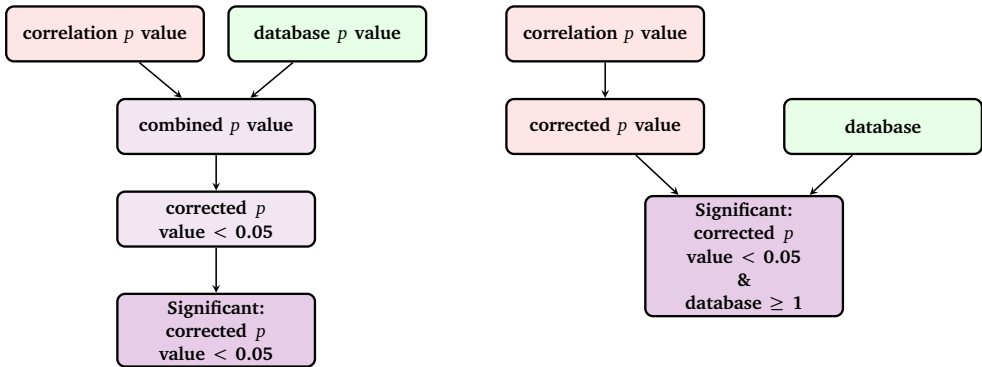


Figure 4.17: Bivariate distribution of p_{Fisher} and p_{Stouffer} .

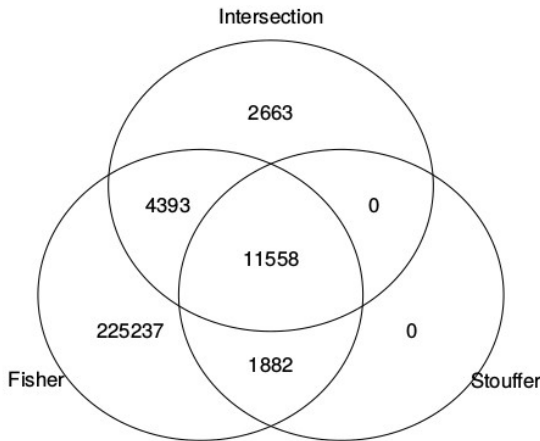
So the major part of the discrepancies between Fisher and Stouffer methods are located in the group with $p_{\text{MicroCosm}} = 1$, and there are a few number of pairs where both p values are similar (lower diagonal) or showing a similar trend. In practical terms, this figure means that when both hypotheses are true, both methods give similar results, but in the case when one is not significant (which is the main part of the cases) the methods disagree and lead to different conclusions.

4.2.4.2 Fisher vs Stouffer vs Intersection

Finally, we compared different strategies (Figure 4.18(a)) available. The most remarkable difference is when p value correction is performed. In the Fisher or Stouffer combination, the p value correction is performed after p value combination; while in the Intersection procedure, as there is no combined p value, the correction is made directly to correlation p value.



(a) Strategies for selecting miRNA-mRNA intersections. **Left:** Fisher or Stouffer combination. **Right:** Integration procedure.



(b) Number of miRNA-mRNA interactions

Figure 4.18: 4.18(a): strategies for selecting miRNA-mRNA intersections. 4.18(b): number of miRNA-mRNA intersections found with each criteria.

Figure 4.18(b) shows the number of miRNA-mRNA pairs according to the strategies shown in Figure 4.18(a). Fisher's combination method is the one that finds more miRNA-mRNA interactions due to two factors: first of all, it is skewed to small p values, so their p values are systematically lower than Stouffer's, plus by the distribution of p values (not concentrated in 1), it is less affected by p-value correction.

Intersection adds 2663 miRNA-mRNA interactions to both Fisher and Stouffer criteria due to the inclusion of TargetScan database. There are also 4393 interactions not considered by Stouffer that were detected using Intersection (these interactions had strong p_{Stouffer} , but were not included due to FDR correction). Only 1882 miRNA-mRNA interactions detected with Stouffer method are no longer significant after correcting the p values directly from correlation p values. These are pairs with a strong $p_{\text{MicroCosm}}$ but a regular $p_{\text{Correlation}}$.

4.3 STUDY 1 – MiRComb in digestive cancers

In our first study we have used publicly available data from The Cancer Genome Atlas (TCGA) [136] for different digestive cancers. The results of this section highlight potential miRNA-mRNA interactomes of five digestive cancers and offer an unbiased view of miRComb functions. As far as we know, there is still no global analysis of this kind in digestive cancers. We obtained the results that are detailed in the next sections, and were also published in the following article:

Citation

Vila-Casadesús M, Gironella M*, Lozano JJ* (2016) MiRComb: An R Package to Analyse miRNA-mRNA Interactions. Examples across Five Digestive Cancers. PLoS ONE 11(3): e0151127. doi:10.1371/journal.pone.0151127

4.3.1 MiRComb analysis of miRNA-mRNA interactions of 5 different digestive cancers

The following MiRComb parameters were used to compute miRNA-mRNA interactions: no filter for differentially expressed miRNAs and mRNAs (unfortunately, they not were enough controls to produce correct estimates of the mean expression per group), Pearson correlation was used to compute the associations between miRNA and mRNA pairs, and then intersection with microCosm and TargetScan databases were used to select the final miRNA-mRNA pairs. Final miRNA-mRNA interactions were those with negative correlations ($FDR < 0.05$) and predicted in at least one of the selected databases.

Standard `mkReport` function was used to produce one PDF report for each sample. As an example, Figure 4.19 shows the main figures from the LIHC report. Full reports of this analysis can be found in the Appendix A of this thesis.

4.3.1.1 Summary of datasets composition

Table 4.3 shows the number of samples available for each cancer and the total number of significant correlations. COAD, LIHC and STAD cancer had more than 400 samples

4.3. STUDY 1 – MIRCOMB IN DIGESTIVE CANCERS

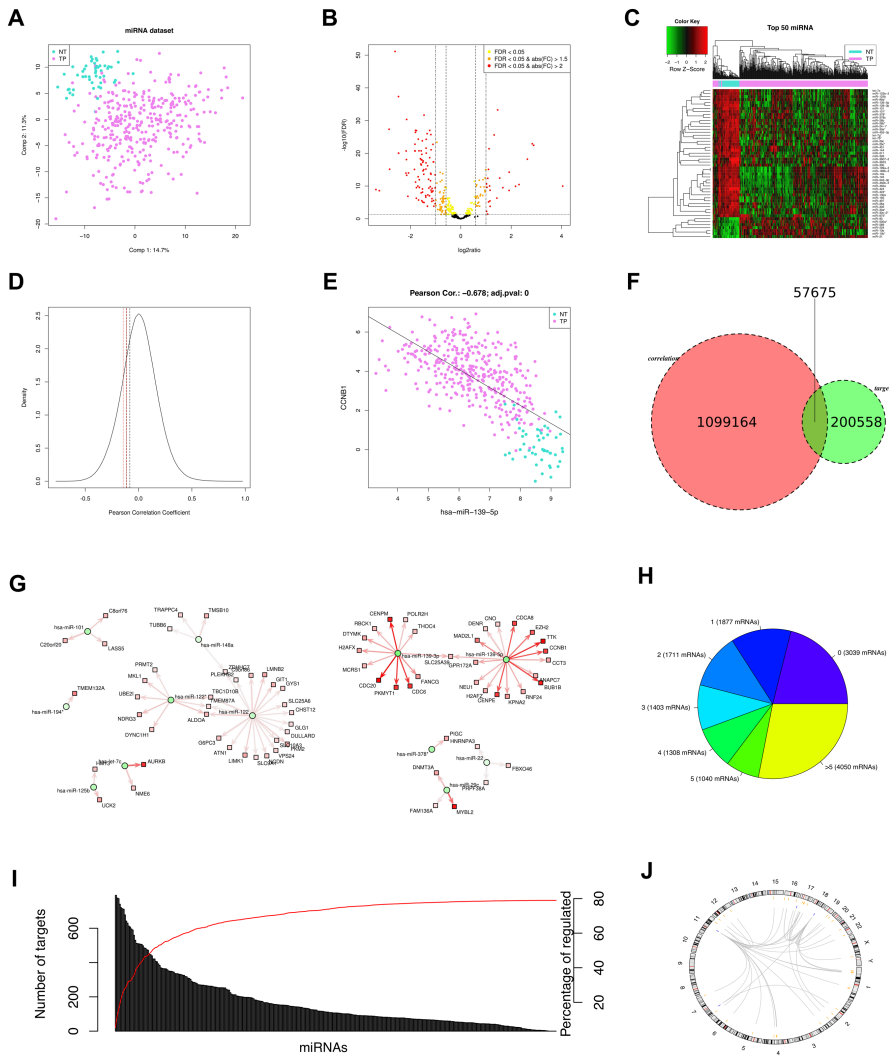


Figure 4.19: (continues next page)

available for analysis, while ESCA cancer and READ cancer datasets had 191 and 160 samples, respectively. Moreover, the ratio between cases and controls is also a term to take into account. While ESCA, LIHC and STAD disposed of a *reasonable* amount of controls (approximately 1:13 for ESCA, 1:7 for LIHC and 1:10 for STAD), in and LIHC we disposed of only 8 and 3 controls respectively (a ratio of approximately 1:50).

The number of available samples influences the number of correlations with $FDR < 0.05$ found: the more samples we have, the higher is the power for detecting correlations different from 0. The number of significant correlations found is higher than 15% (even

Figure 4.19: Summary of main `mkReport` figures. **A)** Principal Components Analysis (PCA) (based on correlation matrix) of miRNA samples. **B)** Volcano plot showing the miRNAs according to its logratio between cancer and control. **C)** Heatmap of the top 50 most deregulated miRNAs according to its FDR. **D)** Density plot of the Pearson correlation coefficients of all possible miRNA-mRNA interactions. Lines show different cutoff: $p\text{-value} < 0.05$, $p\text{-value} < 0.01$, $FDR < 0.05$ and $FDR < 0.01$. **E)** Correlation of miR-139-5p and CCNB1 as an example. **F)** Venn diagram showing the total number of significant correlations ($FDR < 0.05$), the total number of predicted interactions in at least one database (TargetScan or MicroCosm), and the intersection of both. **G)** Network of selected interactions. Each miRNA-mRNA interaction is negatively correlated ($FDR < 10^{-33}$) and predicted at least in one database (Targetscan or MicroCosm). Circles represent miRNAs and squares mRNAs; red fill means upregulated miRNA/mRNA, while green fill means down-regulated miRNA/mRNA; lines indicate the miRNA-mRNA pairs; red line means positive score and green line means negative score; arrow width is proportional to the number of appearances on the databases (TargetScan or MicroCosm). **H)** Pie chart showing the number of mRNAs regulated by 0, 1, 2, 3, 4, 5, and >5 miRNAs. **I)** Barplot showing the number of targets per miRNA and the percentage of mRNAs that are cumulatively regulated by the miRNAs. **J)** Circos plot of the top 45 miRNA-mRNA interactions sorted by FDR, a line means a miRNA-mRNA pair. Blue lines are the position of the miRNAs and orange lines are the position of the mRNAs.

	COAD	ESCA	LIHC	READ	STAD
Number of samples (cases, controls)	444 (436, 8)	191 (178, 13)	407 (357, 50)	160 (157, 3)	443 (406, 37)
Expressed miRNAs	325	338	343	325	330
Expressed mRNAs	14860	18807	14428	14973	18565
Total correlations computed	4829500	6356766	4948804	4866225	6126450
Significant correlations (%respect total correlations computed)	823121 (17.04%)	568914 (8.95%)	1156839 (24.38%)	423296 (8.70%)	1390596 (22.70%)
Significant correlations + targets	47134	30061	57675	24941	71464

Table 4.3: Summary of the main `miRComb` computations on the five digestive cancer datasets analysed.

after FDR correction) in the data sets with more than 400 samples (STAD, LIHC, COAD), while this percentage does not reach 10% in the cases of READ and ESCA (less than 200 samples available). In short, it seems that a dataset with a bigger sample size and a balanced design should provide a greater number of correlations that one that is smaller and not balanced.

Although 20,531 mRNAs and 1025 miRNAs were sequenced, only around 32–34% of the miRNAs were considered expressed (median counts > 10 across all samples) in each cancer dataset. In contrast, 70–90% of the mRNAs were detected with a median > 10 counts. In general, PCA analysis (pages 1 and 2 of the reports made by mkReport that can be found on the Appendix A) of samples revealed a really slight control clusterization (except for miRNA dataset in COAD, READ and in both data sets in LIHC). Overall, this leads to the idea that the main drawback of the data set is the lack of a reasonable number controls, reinforcing the thoughts that differential expression between both groups can be computed and used as an informative item, but not as a filtering step (although is preferred to do that (Section 4.2.2), that could lead to failures in the sense of false negatives).

Volcano plots (pages 3 and 4 of the reports) highlight in red the selected miRNAs and mRNAs. Figure 4.19B has been adapted to highlight different possible selections. First of all the horizontal line sets the statistical evidence, in this case FDR < 0.05 or not. The vertical lines set different levels of biological evidence: 1.5 and 2. MiRNAs or mRNAs with FoldChange less than 1.5 are likely little biologically relevant (yellow). MiRNAs or mRNAs with FoldChange greater than 1.5 but smaller than 2 show medium biological relevance (orange). MiRNAs or mRNAs with FoldChange greater than 2 are highly biologically relevant (red). Heatmaps are also plotted (pages 3 and 4 of the reports). Heatmap of LIHC as an example is also shown in 4.19C.

4.3.1.2 Analysis of miRNA-mRNA interactions

Page 5 of the pdf reports shows the summary of the computed correlations. The next step is to intersect the significant correlations with predicted miRNA-mRNA potential interactions from Microcosm [14] or TargetScan [89] prediction databases (pages 6 and 7 of the reports). For the case of LIHC (4.19F), we observed that the predicted number of miRNA-mRNA interactions was reduced from 258233 to 57675, therefore, we could estimate that around 80% of the initial miRNA-mRNA predicted interactions from databases were false positives for this disease because they did not show a negative correlation between the specific miRNA and the specific mRNA expression in vivo in the tissue.

Furthermore, Figure 4.20 shows that we can also depict the proportion of false positive predicted targets of each miRNA from databases in a given situation. Concerning LIHC, the number of false positives ranges from 22% to 99%. In the case of *hsa-miR-122*, *hsa-miR-122** or *hsa-miR-378c* these percentages are quite low compared to the others (22%, 27% and 24% respectively), therefore these miRNAs show a high ratio of predicted targets

confirmed by miRComb. Interestingly, *hsa-miR-122* is the most frequent miRNA in the adult liver, and plays a central role in liver biology and hepatocarcinoma disease [194].

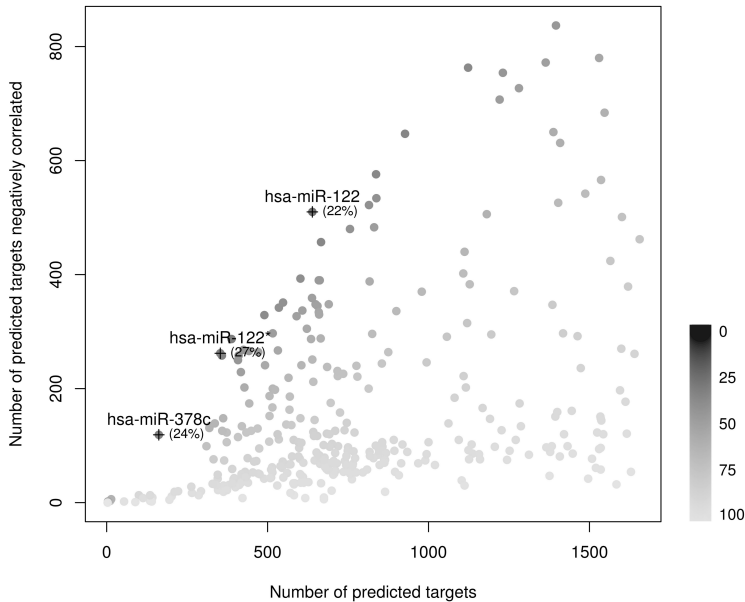


Figure 4.20: Percentage of false positive miRNA-mRNA predicted interactions in LIHC. Plot showing the ratios of negatively correlated predicted targets respect to all predicted targets according to the databases for each miRNA. The intensity of the grey colour dot is related to the percentage of false positive miRNA-mRNA predicted interactions. In brackets, the exact percentages of false positives from selected miRNAs (*hsa-miR-122*; *hsa-miR-122**; *hsa-miR-378c*).

Page 6 of the pdf report shows the top 15 miRNA-mRNA interactions (sorted by adjusted p value, taking into account that they must have been predicted in at least one database) in each cancer. Page 9 of the pdf report shows the network of all the miRNA-mRNA interactions. All the interactions are plotted by default and this could result in a very dense figure difficult to interpret, as it is the case in our examples. For the case of all the interactions of LIHC (Appendix A.3, page 9) we can see two main patterns: on the left we can find mostly downregulated miRNAs in LIHC (plotted as green circles) together with their corresponding mRNA targets (plotted as red squares). On the right the roles are inverted, and the predominant miRNA-mRNA interactions shown consist on upregulated miRNAs with downregulated mRNAs. This general pattern is reproduced in all the studied cancers (Appendix A). To solve this problem, we suggest adapting in every case the number of interactions to be plotted depending on the goal of the figure. In Figure 4.19G we have plotted a reduced amount of interactions and we can see some of the details. For example, two targets (DNMT3A and MYBL2) of *hsa-miR-29c* (bottom

right) are predicted by two databases, while the target FAM136A is predicted by only one database (the arrow is thinner). Moreover, regarding the targets of *hsa-let-7c*, the AURKB is more deregulated in LIHC than the NME6, and the interactions of *hsa-miR-122* or *hsa-miR-122** (top left) have lower scores (lower intensity of arrow colour) than the interactions of *hsa-miR-139-3p* and *hsa-miR-139-5p* (higher intensity of arrow colour; top right).

In LIHC more than 75% of the expressed mRNAs are being targeted by at least one miRNA (Figure 4.19H and 4.19I and page 10 of the pdf report), in COAD and STAD that number is between 70% and 60%, while in READ and ESCA is less than 50%. However, we have to take into account that these percentages are partially affected by the total number of miRNA-mRNA predicted interactions: the higher number of interactions, the higher number of miRNAs per mRNA (and viceversa). For example, more than 25% of the miRNAs in LIHC are predicted to be targeted by more than five miRNAs. This percentage is lower in the other cancers, but it is still 8% in READ. It is worth to mention that this is a first approach that will require interactions to be experimentally confirmed in a wet lab. This unusual number of miRNAs targeting the same mRNA could be attributed to the fact that miRComb does not take into account competitiveness between different miRNAs hybridising to the same target, which is unlikely to happen in real organisms.

Page 11 of the pdf report shows the first 20 miRNAs sorted by the number of targets. As an example, *hsa-miR-106a* has 766 interactions predicted in COAD, *hsa-miR-27a* has 450 interactions in ESCA, *hsa-miR-27b* has 792 interactions in LIHC, *hsa-miR-106a* has 582 interactions in READ, and *hsa-miR-29a* has 798 interactions in STAD. Although miRNAs are expected to regulate up to hundreds of genes, these interactions should be experimentally validated in order to discard false positives or indirect relations, as mentioned above. Colours in these pages show the direction of miRNA deregulation (red: up-regulated; green: down-regulated). While in COAD, READ and ESCA the top miRNAs are in general upregulated, in LIHC and STAD they are mostly downregulated. MRNAs can also be sorted according to the number of miRNAs that are targeting them (page 12 of the report) and are also coloured according to the direction of deregulation. Overall, mRNAs do not have more than 50 miRNAs regulating them. Exceptionally, in STAD there are some mRNAs with more than 60 miRNAs (eg. 74 for FOXP2). However, it is worth to take into account that the vast majority of mRNAs that are regulated by at least 1 miRNA, are simultaneously regulated by up to 4 miRNAs.

In general terms, the main direction of the top mRNAs (sorted by the number of miRNA targeting them, report page 12) is the inverse of the main direction of the top miRNAs (sorted by number of targets, report page 11).

4.3.1.3 Functional enrichment analysis of miRNAs according to their targets

In pages 13–15 of the report, we can find the Gene Ontology (GO) and KEGG functional analysis of the results. As an example, we tested if the mRNAs that are regulated by miRNAs are enriched in any of the GO and KEGG categories.

Results of this section are quite similar between all digestive cancer datasets because they include all mRNAs that are targeted by at least one miRNA and it includes more than 50% of the expressed mRNAs on average. Depending on the goal of the study different filters could be applied (differential expressed miRNAs and/or mRNAs, targets from one specific miRNA, etc.) and, then, results would be different.

In this case, BP (Biological Process) overrepresented terms include cellular process and other regulating and signalling processes. CC (Cellular Component) overrepresented terms are mostly related to intracellular-cytoplasm compartments. MF (Molecular Function) overrepresented terms are centred in protein binding and other binding (enzyme, anion binding) actions.

KEGG pathways are more concise and all of them include the term *Pathways in cancer*. COAD also included prostate cancer and chronic myeloid leukaemia and glioma, ESCA also small cell lung cancer, LIHC included prostate cancer, colorectal cancer, pancreatic cancer, chronic myeloid leukaemia and renal cell carcinoma; READ included renal cell carcinoma, STAD also included small cell lung cancer and prostate cancer. This suggests that, as known, many cancers share similar patterns. Other pathways that are shared across the different studied data sets are: Focal adhesion, Fc-gamma R-mediated phagocytosis (COAD, ESCA, STAD), or TGF-beta signalling pathway (COAD, READ).

More targeted results can be obtained by testing for enrichment the targets of a specific miRNA. For example, the targets of *hsa-miR-148a* in liver cancer are enriched in antigen processing and presentation KEGG Pathway (FDR = 0.006) (Figure 4.21). In a practical sense, this means that this pathway is involved in liver cancer through a deregulation of *hsa-miR-148a*, and that this pathway could be, at least partially, modulated by modifying *hsa-miR-148a* expression. Other pathways involved in liver cancer that could be modulated by altering miRNA expression are RNA transport (FDR = 0.030), Cell cycle (FDR = 0.031) and Ubiquitin mediated proteolysis (FDR = 0.031) for *hsa-miR-424*, or Lysine degradation (FDR = 0.006) for *hsa-miR-29c*.

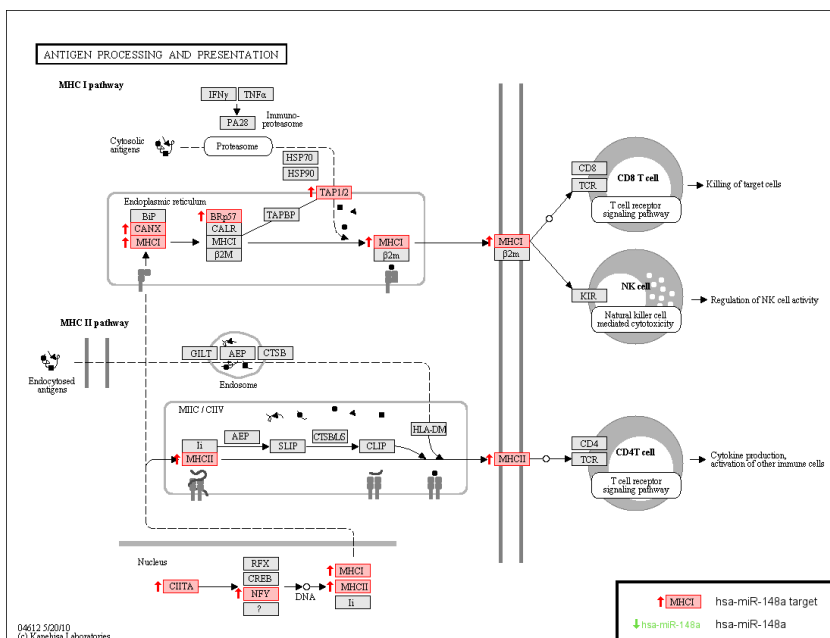


Figure 4.21: KEGG Pathway for *hsa-miR-148a* targets in liver cancer: Antigen Processing and Presentation. *Hsa-miR-148a* targets (negative correlation with *hsa-miR-148a* (FDR < 0.05) and predicted in at least one database) are highlighted in red.

4.3.2 Integrative analysis of the miRComb miRNA-mRNA interactions from the 5 digestive cancers

4.3.2.1 Shared and specific miRNA-mRNA interactions

Figure 4.22 shows the number of shared miRComb miRNA-mRNA pairs among the 5 studied digestive cancer datasets. 1570 miRNA-mRNA interactions are shared for all 5 sets, but a more relevant number is shared in at least 2 or more of them, being only less than 40% of miRNA-mRNA pairs specific of each cancer dataset. STAD is the one with more miRNA-mRNA interactions found.

In Figure 4.23, a network represents the 1570 common miRNA-mRNA interactions among the five studied mentioned data sets. We can see two networks: the big network on the left contains mostly downregulated miRNAs with their upregulated mRNA targets (780 miRNAs + mRNAs, and 1305 miRNA-mRNA pairs), while the smaller network on the right contains mostly upregulated miRNAs and their downregulated mRNA targets

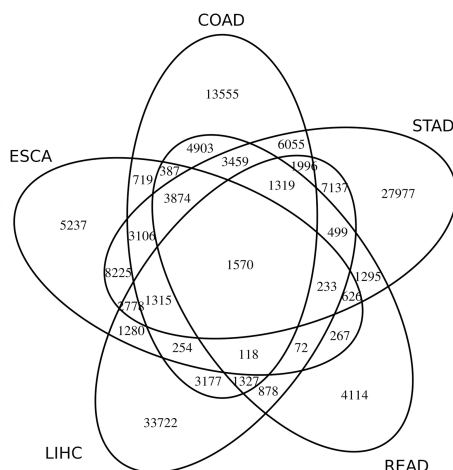


Figure 4.22: Venn diagram showing miRComb miRNA-mRNA interactions (FDR < 0.05 and predicted in at least one database) that are present in at least one cancer. 1570 miRNA-mRNA interactions appear in the 5 studied digestive cancers.

(173 miRNAs + mRNAs, and 187 miRNA-mRNA pairs). We have spotted the mRNAs that have KEGG terms related to cancer, such as *Cell Cycle* (red), *Pathways in Cancer* (yellow) and *MAPK Signalling Pathway* (blue). Combinations of these terms are also displayed in different colours. The network on the right contains some mRNAs related to Cell Cycle, while the big on the left is mostly related to MAPK Signalling Pathway, Pathways in Cancer, or both terms (green).

The common interactions can be related to pathways that are shared by all the studied digestive cancers. However, it is also interesting to study the interactions that can be specific of each one. Tables 4.4 4.5 4.6 4.7 4.8 show the top 10 miRNAs with more miRComb miRNA-mRNA specific interactions for each cancer (a specific interaction is the one that has been found significantly negatively correlated in one data set but not in the others). Full tables containing all the miRNA-mRNA specific interactions can be found on the Supplementary information of the PLoS ONE article [187].

Figure 4.24 also shows the number of specific interactions depending on the miRNAs involved in LIHC. MiRNAs on the line corresponding to ratio 1:1 are those that are only expressed in liver cancer. The others are expressed in at least another cancer, but they have some specific interactions in LIHC, the closer to the ratio 1:1 line are, the higher the specificity is.

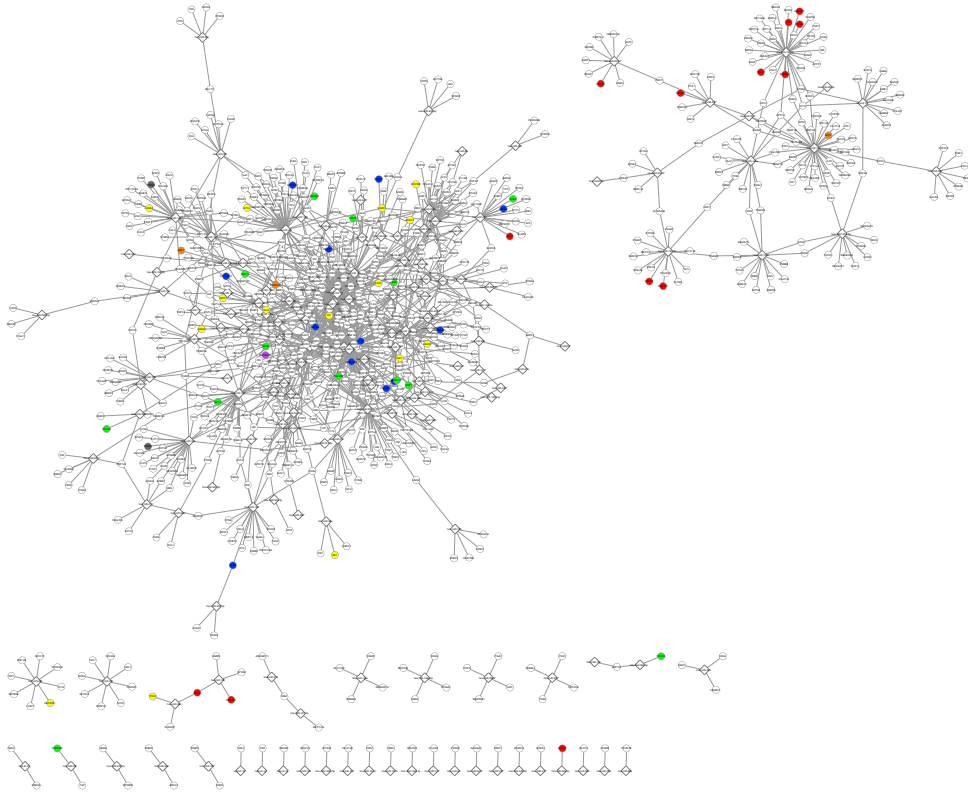


Figure 4.23: Network for common 1573 miRComb miRNA-mRNA interactions across five digestive cancers. Circles are mRNAs, while diamonds are miRNAs. Fill colour represents in which pathways the resulting protein of the mRNA is involved. Yellow: *Pathways in cancer*; Blue: *MAPK Signalling Pathway*; Red: *Cell Cycle*; Green: *Pathways in Cancer+MAPK Signalling Pathway*; Orange: *Pathways in cancer+Cell cycle*; Magenta: *MAPK Signalling Pathway+Cell cycle*; Grey: *MAPK Signalling pathway+Pathways in Cancer+Cell cycle*.

4.3.2.2 Cluster analysis of miRNA-mRNA interactions

Globally, there are 106426 miRNA-mRNA interactions measured in all cancer datasets, and significantly negatively correlated in at least one of them. In order to classify them into similar patterns, we applied clustering methods in order to summarise the main trends. We used the K-means method with 4 clusters as it gave a reasonable interpretation of the results (Figure 4.25).

Interestingly, hierarchical clustering of cancers according to the mean correlation co-

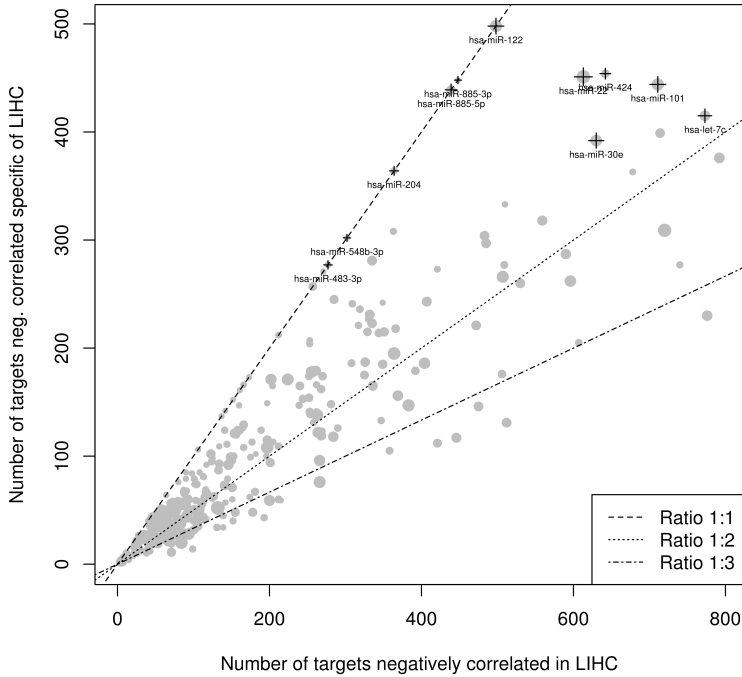


Figure 4.24: Number of total miRNA targets in LIHC versus number of miRNA targets present only in LIHC but not in COAD, ESCA, READ or STAD. The size of the dots is proportional to the mean miRNA expression on the LIHC samples included.

efficients of the clusters gives the following result: STAD and ESCA are first grouped, as well as READ and COAD. Next, these four cancers are grouped, and finally LIHC is added to the tree. Principal Components Analysis shows the same pattern. It is an expected result and is reasonable with biological similarities of these tumors, what is relevant is that this classification can also be reproduced using miRNA-mRNA interactions. Successive increase of the number of clusters allow to differentiate other trends according to the correlations (data not shown), but the tree structure described before (COAD+READ, ESCA+STAD and then LIHC) is always maintained.

Clusters can be interpreted as follows: Cluster 1: miRNA-mRNA interactions slightly negatively regulated across all cancers and interactions that do not fit other clusters; Cluster 2: miRNA-mRNA interactions negatively correlated in COAD and READ, but not in the other cancers; Cluster 3: miRNA-mRNA interactions negatively regulated in ESCA and STAD, but not in the other cancers; Cluster 4: miRNA-mRNA interactions negatively regulated in LIHC, but not in the other cancers. For example, *hsa-miR-106a* and its targets are quite specific of Cluster 2 –COAD and READ– (although they are also present in some

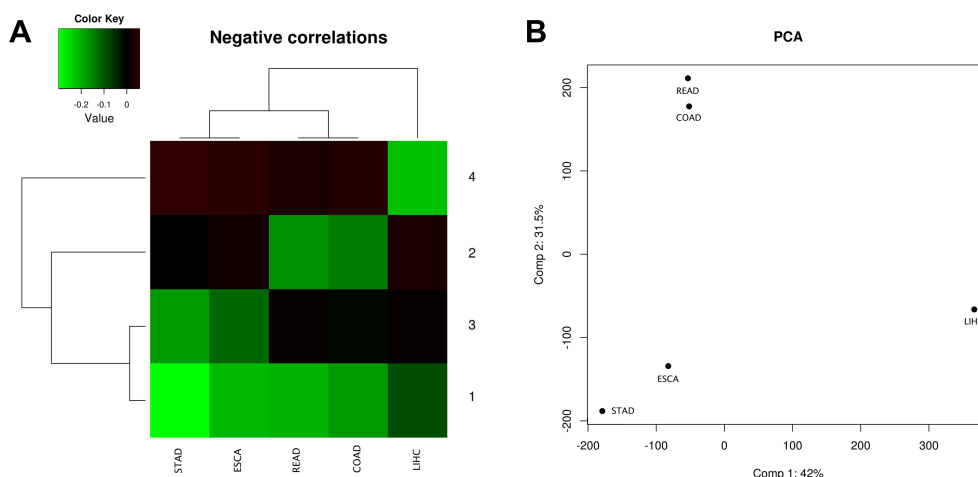


Figure 4.25: Clustering and Principal Components Analysis of the five digestive cancers. Computations are based on the correlation coefficients of the 106.426 miRNA-mRNA pairs that are expressed across all five cancer data sets. **A)** Heatmap showing the centers of the different clusters. Values represent the mean of the Pearson correlation coefficient of the miRNA-mRNA pairs that fall into the cluster. **B)** Principal Components Analysis (based on correlation matrix) of the Pearson correlation coefficient of the miRNA-mRNA pairs from the five digestive cancer data sets.

extent in Cluster 1 –all cancers–). Another example, *hsa-miR-29c* targets are specific from Clusters 3 –LIHC and ESCA– and 4 –LIHC –, and have almost no presence in Cluster 2. Furthermore, *hsa-miR-22* targets are specific from Cluster 4 –LIHC–, and others such as *hsa-miR-30b* or *hsa-let-7b* targets seem to not show any clear specificity (Figure 4.26, figure in full resolution can be found at Supplementary Figure 3 of PLoS ONE article [187]: <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0151127#sec018>).

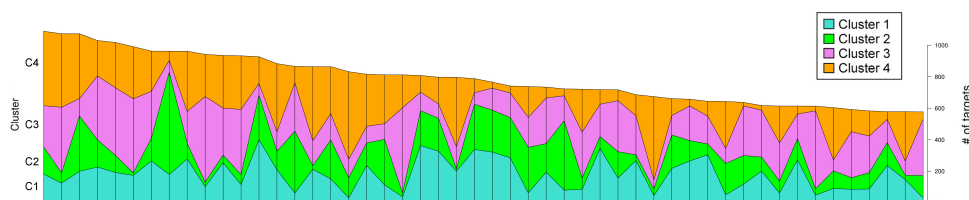


Figure 4.26: Number of miRNA targets in each cluster. The represented miRNAs are the top 50 sorted by the total number of targets. Blue: Cluster 1; Green: Cluster 2; Magenta: Cluster 3; Orange: Cluster 4.

miRNA	n.targets	mRNAs
miR-30c	264	MRAS, ZEB2, FCER1G, CFL2, SMNDC1, KIAA1949, CLIP4, SIGLEC5, C10orf119, CHORDC1, DZIP1L, LRCH2, SAP30, INTS2, ATF1, RPRD1A, PLXNC1, HCFC2, EIF5A2, TMEFF1
miR-16	213	TRANK1, KHNYN, PHLPP2, KIAA1370, C5orf41, ACOX1, CCDC88C, LCOR, CCNJL, SYNRG, CHD2, ZBTB34, SESN2, PDCD6IP, GCC2, MLL2, WNT5B, KIAA0317, NBR1, TNIP1
miR-17	178	TAOK3, GBF1, ZFYVE26, PPARD, TEP1, CYP26B1, KDM6B, BTBD7, CD68, NRBP1, NCOR1, KIAA1671, GOLGA2, ARHGAP21, MINK1, ALOX15B, KIAA1522, PSEN1, ARHGEF18, SEMA7A
miR-454	174	KIAA1211, HADHB, CSF1, FAM107B, PANK3, BTG1, ADAM28, FAM78A, MIER3, MXD1, BTD, RNASEL, MOBKL2B, GZMK, B2M, HADHA, TP53INP1, TCF7L2, DCTN2, TAGAP
miR-106a	173	SLC36A1, LASP1, TANC2, FGFR2, ANKFY1, BAHD1, KDM6B, SLC22A23, STAT3, CRK, C15orf17, TADA2B, ABHD5, MAP3K9, IQSEC1, ARHGEF11, NDEL1, CNNM2, KIAA1522, RCOR1
miR-301a	170	FAM107B, KIAA1211, ZDHHC7, MAML3, MXD1, MTF1, BTG1, RAB5B, PANK3, KLF3, LMTK2, MOBKL2B, NDEL1, ABHD5, C8orf4, FAM78A, LAMA3, KLHL20, HOXD1, TSPAN3
miR-181c	159	MAP3K6, ERI1, UGT8, KPNB1, SAP30, MORC4, SLC25A37, RNF125, PAX9, E2F7, ZIC2, WASF1, TUBB, PKNOX1, XKR9, MAP2K1, KITLG, XPO7, SLC25A4, C18orf55
miR-539	157	ZNF609, WNK1, YLPM1, HIVEP2, PHF3, MED1, DDX24, SPEN, INO80D, LRP6, SP1, SH3PXD2A, C10orf118, AHR, IWS1, SETD5, HNRNPK, RNMT, KIAA1244, LCOR
miR-181a	146	TM4SF19, POM121C, WDR45L, ACAN, SPP1, PROCR, ZNF207, PITPNB, EME1, STC1, RAN, MELK, EEF1E1, MRPS23, SCD, SAP30, TUBB, RNF8, CDCA4, SLC35B1
miR-106b	140	SMAD7, FAM102A, TEP1, NCOR1, SESN2, KIAA1522, PANK3, PPARD, GBF1, CRK, SLC22A23, WDR37, TRIM36, CYP26B1, ANKFY1, MYO1F, TMEM156, KIAA1671, MBNL3, MAP3K12

Table 4.4: Top 10 miRNAs sorted by the number of specific targets in COAD. Target mRNAs are sorted according to its negative correlation value (top 20 are displayed).

4.3. STUDY 1 – MIRCOMB IN DIGESTIVE CANCERS

miRNA	n.targets	mRNAs
miR-944	313	SLC41A2, HNMT, LONRF3, GATA6, ARHGAP18, MGAT4A, ICA1, LPIN2, VPS13C, SLC12A2, NR3C2, HSD17B11, FOXP1, THRA, C2orf88, PTPRB, TMEM50B, C20orf112, C11orf54, SEPSECS
miR-205	261	ENPP4, SERPINA5, PTPRJ, SPATA13, BTNL8, SPRY1, ACACB, SLC4A4, PHF17, MGRN1, PTP4A2, MGAT4A, MAGI1, DOK1, EXOC6, PRDM16, NEK6, CASC4, HSD17B11, FBP1
let-7b	177	TTL6, TFPI, TMEM135, SFMBT1, GALC, SLC46A3, CCL23, YPEL2, MUC3A, ITGB3, C7orf58, ATP8A1, SEC14L1, INSR, GXYL1, BHMT2, KLF9, HGF, MLXIP, MAP4K3
miR-338-5p	141	SAFB, VRK2, NEDD1, ABHD12B, LIMK2, AEBP2, TANC2, QSER1, RAB38, CERS3, ROBO1, MBD2, SP3, SYPL1, SCPEP1, ATP2C1, UNK, CCNA1, FKBP3, NCK1
miR-27a	128	RALGPS1, NFATC2, EEPD1, PLEKHA6, MAN2A2, SPATA13, PPM1H, KIAA1958, FOXA3, PRDM16, KBTBD11, SEMA3B, PTPRJ, RALGAPA2, TRIM2, PPFIA2, KIAA1147, GPD1, CAPN9, NKTR
miR-23a	115	KLHDC7A, PTPRB, REPS2, LONRF3, ZC3H12B, ZNF420, C11orf75, FUCA1, TTC6, TBC1D12, RAB17, ZNF518A, MLPH, ZNF238, GPRC5B, C10orf68, CRBN, ZNF780B, ZNF506, ZNF253
miR-34c-5p	97	MLPH, TM9SF3, AHCYL2, CAPN5, LONRF3, CREB3L1, MYO7B, LGR4, C10orf81, BACE2, PARP4, MGAT4A, TGFB2, IYD, MICAL2, LRCH1, FUT8, GOLPH3L, UBR1, TM9SF2
miR-24	87	MEGF11, NDST3, SNTB1, HNF1B, ATP6V0A2, AHI1, EPB41L1, SNED1, SLC12A3, C9orf96, ARHGDIG, C20orf112, FCGRT, TCAI, NLK, ARHGAP26, IDUA, SLC37A1, UBN2, SMPDL3B
miR-27b	87	ACAA2, PEAK1, ZFP36L2, JMJD1C, ARL14, GPD1, PLCL2, CTH, PDK4, PLEKHA6, ZC3H12B, PTPRB, GPR126, FOXA3, OXER1, NR2F2, KBTBD11, SLC46A3, PAPSS2, GORASP1
miR-149	81	PLEKHA6, GJB1, CREB3L3, ACHE, GAB2, GRK5, FZD5, GPR114, RILP, MIA, MMP15, RPH3AL, MUC5B, DENND3, MUC5AC, SEMA3B, C11orf86, BIN1, ANPEP, IGJ

Table 4.5: Top 10 miRNAs sorted by the number of specific targets in ESCA. Target mRNAs are sorted according to its negative correlation value (top 20 are displayed).

miRNA	n.targets	mRNAs
miR-122	498	SLC9A1, G6PC3, PKM2, VPS24, TBC1D10B, NCDN, ZDHHC7, C9orf86, GYS1, CHST12, GIT1, DULLARD, ALDOA, PLEKHB2, ATN1, SLC10A3, SLC25A6, TMEM87A, LMNB2, GLG1
miR-424	454	APLN, AMIGO3, RECQL5, FAM189B, UBE2Q1, MXD3, SNRPC, BAT4, ZNHIT3, NSMCE2, TOMM20, MTX1, BCAP31, PUF60, E4F1, CDKN2A, DUS1L, NFKBIL1, TARBP1, DEDD
miR-22	451	FBXO46, RCC2, UTP18, NAT9, H2AFX, COPS7B, UBE2Z, PHF5A, MCM6, KIF18A, C17orf53, OLA1, POGK, WDR62, HNRNPH1, FAM49B, FBXL19, TPM3, ENTPD2, RFXANK
miR-885-3p	448	BMP1, KIAA0174, ACCN2, C9orf116, CCDC103, E2F4, CDK6, RARG, SP5, OTUD5, OSR1, RALY, EIF2B4, CLDN2, PRMT2, PLSCR3, CDYL2, GTF3C5, CCDC40, PPP1R12C
miR-101	444	LASS5, DNMT3A, NAP1L1, EZH2, RIT1, UCK2, SMARCA4, SUB1, C1orf77, KIAA1841, SMARCD1, RASD2, STK19, DSTYK, ATP6V1E1, ATP5G2, UBE2D2, MFSD6, C12orf34, EED
miR-885-5p	439	NKD1, ADAMTS9, C20orf196, CMIP, VLDLR, DNAL1, RPRG1P1L, AP2M1, CDYL2, HSPB8, MFSD5, AAK1, HIF1AN, LAMA5, WWTR1, LUZP6, TTC30A, RNASEL, CFLAR, CHMP5
let-7c	415	ARID3A, IGF2BP1, NAP1L1, PCBP4, NPEPL1, C7orf49, ABCC5, DLGAP4, ABCC10, BAX, SLC12A9, C15orf39, IRGQ, CYB561D1, IGF2BP3, FBXL19, GGA3, DUSP9, MMP11, AARSD1
miR-125b	399	SLC26A6, RBCK1, NUP210, NEU1, THOC5, P2RX4, ARID3A, ATP5G2, STK11IP, GLT ϵ , LIMK1, MAZ, RIT1, PLXNA1, MAN1B1, CD2BP2, C15orf39, MSI1, RFXANK, TAZ
miR-30e	392	C8orf76, FKBP1A, MICAL1, DTX2, C19orf50, NME6, STK39, STOML1, DGKZ, TMC7, TTC39A, USF1, VOPP1, SEMA7A, TTC35, GNPDA1, FZD2, LENG9, AURKB, RPS19BP1
miR-27b	376	PSMD7, KIAA0513, HM13, EFNA3, WDR45, ACCN2, SLC7A11, WDR8, ATP6AP1, ELOVL1, SCAMP3, PIGT, MRPL33, BRSK1, KIAA0226, FAM21B, UNC45A, MEPCE, TSEN54, RRP12

Table 4.6: Top 10 miRNAs sorted by the number of specific targets in LIHC. Target mRNAs are sorted according to its negative correlation value (top 20 are displayed).

4.3. STUDY 1 – MIRCOMB IN DIGESTIVE CANCERS

miRNA	n.targets	mRNAs
miR-323-3p	262	KIAA0907, MYLIP, MACC1, RBM41, EFNA3, RBBP6, ABI1, TPR, TMEM106B, MLL5, PHF14, MKLN1, SLC25A36, AFTPH, NCBP2, ZNF292, RBM39, RSBN1, ZNF485, NCOA3
miR-23a	179	WASL, UBE2D1, MTM1, PLEKHM3, TMEM87B, PPP1R12A, CBL1, WAC, MLLT4, CDC40, PTP4A2, AEBP2, RPRD2, RBBP6, CPEB2, TSR2, BMPR2, BACH2, PURB, ZYG11B
miR-369-3p	120	BMPR2, STYX, STON1, ZEB1, GOPC, RC3H1, RAB3GAP2, TMEM87B, PHACTR2, IQSEC1, GABPA, ZNF350, SEC63, TNRC6A, RAB11FIP2, UBE2J1, JHDM1D, VPS36, SMG1, OSTM1
miR-382	109	BTBD7, PRKAA1, NT5C2, FBXO28, DHX32, MBNL1, HIPK1, ZMYM2, MIER1, PLEKHA1, ZNF638, C3orf63, DDX3X, RSBN1L, ZNF197, FOXN2, CCDC132, PDE5A, C9orf68, CASP3
miR-409-3p	105	GPBP1L1, C9orf68, CCNT2, TCF7L2, CREB1, FANCL, ZNF14, ARHGAP5, CLK4, C5orf28, NSUN6, DPY19L4, PPHLN1, EBAG9, NDFIP2, ATXN3, TBL1XR1, SLC35F5, ZNF540, SAV1
miR-23b	101	NCOA6, EEA1, ADNP, TSR2, PAPD5, TAB3, TXLNG, FAM123B, IYD, ZNF81, FMR1, UBN2, WASL, GCC1, WIPF2, XIAP, ZBTB44, PICALM, KLHL15, SIAH1
miR-381	99	AKAP6, SORBS1, CACNB2, PBX1, ANK3, LMO3, MBNL1, ZFYVE21, BTBD7, SPPL3, TES, NBEA, MYST4, CHMP1B, ARHGAP5, CACNA1C, CASD1, KIAA1143, ADAMTSL3, RABGAP1
miR-106a	93	ZBTB6, GMCL1, CDC40, FAM3C, PHTF2, ZNF800, TBC1D15, HOOK3, PTP4A2, SLC4A7, LMBRD1, ZBTB41, CNOT6L, ITGB8, DEGS1, CMPK1, SNX16, SGTB, TMEM168, SNTB2
miR-27a	68	EGFR, STON1, CSF1, SERTAD2, MARCKS, HGSNAT, ATP2B1, SGMS1, C5orf41, SMCR8, SMCHD1, GPD2, SSH1, SEPNI, ARHGAP21, TICAM2, WIPF2, PLS1, DIRC2, C16orf54
miR-409-5p	62	ANKRD13C, MON2, TLK1, DYRK1A, PDE4D, FRS2, FAM129A, PDIK1L, RAB3GAP1, C9orf45, NBEA, ZBTB34, PRKAA2, USP15, ARID4B, SFRS11, ENSA, KIAA1598, BRAP, MKL2

Table 4.7: Top 10 miRNAs sorted by the number of specific targets in READ. Target mRNAs are sorted according to its negative correlation value (top 20 are displayed).

miRNA	n.targets	mRNAs
miR-330-3p	390	PRUNE2, NFIA, LMOD3, PARVA, TMEM35, KANK2, ZNF25, HCFC2, FOXP2, ATP2B4, PDE5A, TEAD1, HOXA3, DPYSL3, RNF180, NRP2, TSHZ3, SMAD9, DDR2, SHISA9
miR-26a	357	KIAA1737, UBR3, RANBP9, TMEM106B, G3BP2, KPNA6, ZNF148, STXBP4, ZYG11B, FAM8A1, HEATR5A, UBE2H, UBE2G1, RLF, PEX13, UBR1, SCAMP1, AHI1, LIMS1, FBXW2
miR-1	326	PIGW, UHMK1, CAPRIN1, MTHFD1, NXT2, POLA1, PHF6, CMTM8, AZIN1, SMG7, HOOK1, TMED5, SLC39A9, FAF2, NUP54, IPO9, SMCR7L, PASK, SF3B3, SPTLC1
miR-340	319	LPP, VEZF1, ETV1, RBFOX2, NEK7, SLC25A12, SLC20A2, VAMP4, SGMS2, FBXO8, ZCWPW2, TEAD1, VCL, FAT3, DIXDC1, NCAM2, SGCD, CALD1, MACF1, FBXO3
let-7g	315	RBFOX2, SLC8A2, DMD, CPEB1, GHR, KLHL4, NEFM, HLF, WNK3, DOCK3, FGF5, LEPR, NFASC, TGFB3, KLF8, KIAA2022, EZH1, NOVA1, PBX1, FOXN3
miR-129-5p	314	TMEM62, COL11A1, NXT2, C6orf223, WDFY1, FCGR1A, DTL, NOX1, TRIAP1, PRPF40A, WDR12, TGIF2, CACYBP, SLBP, ALG6, MRPL13, TPM3, RPIA, NDUFA10, E2F7
let-7f	313	ACTR10, FGF5, MAP4K3, BACH1, PPAPDC1A, SNX6, RBFOX2, CALM1, DPH3, CALU, SESTD1, SLMAP, BAG2, CRBN, ELOVL4, SGCD, COPS4, FBXO32, PRKAB2, KPNA4
miR-29a	310	DIP2C, IL17RD, DNAL1, RMND5A, TGFB2, BACE1, FBXL20, PRICKLE2, ATP2B4, ILDR2, OXTR, SBF2, RYBP, PCYT1B, CALU, CACNA1C, C16orf72, CDKL2, KIF5A, JAZF1
miR-15b	290	FOXP2, NOS1, GRPR, KATNAL1, TEAD1, ANKRD53, GPR135, PENK, KY, WNK3, PRTG, CHIC1, TLE4, BAI1, AASS, KCNQ5, BCL2, SYDE2, PID1, BMPR1A
miR-30b	286	VSTM4, TEAD1, AFF4, ABCC9, BCL6, KLF11, ZYG11B, PRKAR1A, UBE2G1, EPN2, C3orf58, ZCCHC24, CCDC6, PCDH10, SETD7, AMOTL2, YPEL2, SAMD4A, ZNF264, PHACTR2

Table 4.8: Top 10 miRNAs sorted by the number of specific targets in STAD. Target mRNAs are sorted according to its negative correlation value (top 20 are displayed).

4.4 STUDY 2 – MiRComb in pancreatic cancer

We also applied miRComb analysis to a dataset of pancreatic cancer samples.

Citation

Vila-Casadesús M, Vila-Navarro E, Raimondi G, Fillat C, Castells A, Lozano JJ, Gironella M. Deciphering microRNA targets in pancreatic cancer using miRComb R package. (*manuscript in preparation*)

4.4.1 Data exploration

The dataset consists of 3 controls and 9 cases with paired miRNA-mRNA data, including the expression of 1733 miRNAs and 18570 mRNAs. Figure 4.27A shows Principal Components Analysis of the dataset. We can see that PDAC samples can be easily differentiated from healthy (H) ones in both miRNA and mRNA dataset.

4.4.1.1 Top differentially expressed miRNAs or mRNAs

Figure 4.27B shows the differentially expressed miRNAs and mRNAs between pancreatic cancer and healthy tissue. There are 201 significantly upregulated and 342 significantly downregulated miRNAs in our pancreatic cancer set. Those represent 31.1% of the total expressed miRNAs. 30 of the upregulated miRNAs were validated by RT-qPCR in two larger cohorts in our previous article [77].

There are also 1613 significantly upregulated and 2030 significantly downregulated mRNAs. Those represent 19.6% of the total expressed mRNAs. Figure 4.27C shows their respective volcano plots colouring the miRNAs and mRNAs according to their Fold-Change. The miRNAs and mRNAs that were selected for further exploration were those with $FDR < 0.05$ regardless of their FoldChange (highlighted in yellow, orange and red).

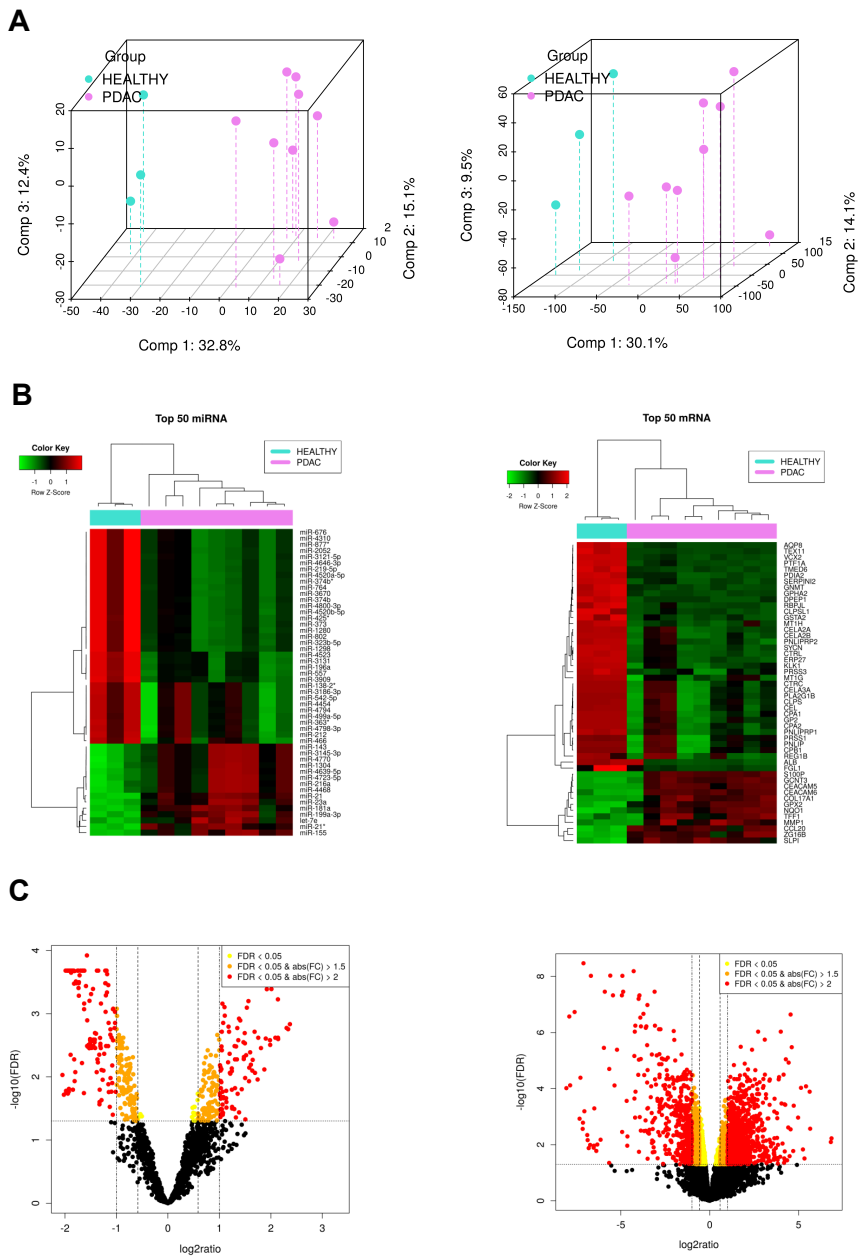


Figure 4.27: A) 3d-PCA (based on correlation matrix) plots for miRNA and mRNA dataset. B) Heatmaps of the top 50 miRNAs and 50 mRNAs sorted by FoldChange. C) Volcano plot of the miRNAs and mRNAs.

4.4.1.2 Intersection with miRNA target prediction databases

We then selected the 543 and 3643 significantly deregulated miRNAs and mRNAs, respectively, and computed all possible correlations using miRComb package. Multiple testing corrections using Benjamini & Hochberg procedure (FDR) was applied. Among 1978149 possible miRNA-mRNA pairs, there were 959775 that correlated negatively and, among them, we found 443100 miRNA-mRNA pairs where this correlation had $FDR < 0.05$. This number which represented 22.4% of the total miRNA-mRNA possible combinations.

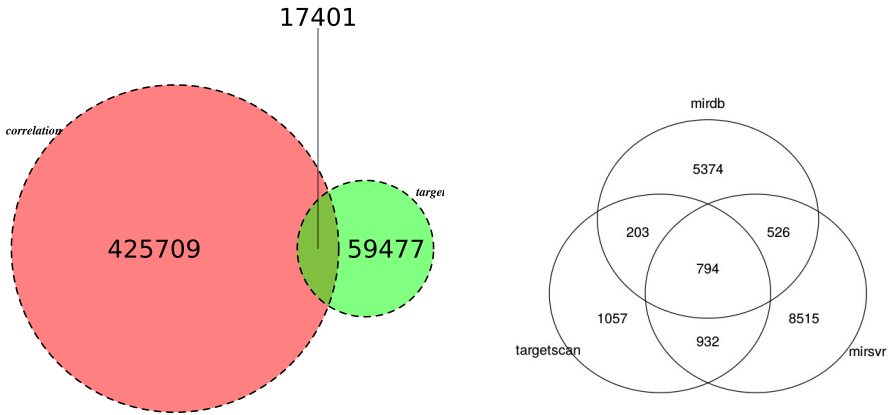
Furthermore, we used the information given by 3 miRNA target prediction databases (TargetScan, miRDB, miRSVR) to intersect with the above mentioned correlations. If we only took into account the 3 miRNA target prediction databases, we would have found a total number of 76878 potential miRNA targets present in at least one of them, and using the interaction calculated by miRComb we reduce this number nearly five times, as we found 17401 miRNA-mRNA pairs that were present in at least one database. That means that only 22.6% of the miRNA-mRNA interactions appearing in these databases were reported as negatively correlated pairs in our dataset. Figure 4.28(a) shows the number of negatively correlated miRNA-mRNA pairs, the number of predicted miRNA-mRNA pairs, and the pairs that fulfil both conditions.

Figure 4.28(b) shows also, among the 17401 miRNA-mRNA interactions, how many are predicted by each database. We can see that miRSVR provided more of the miRComb predicted miRNA-mRNA pairs than the other databases (10767 in total, while TargetScan and miRDB predicted 2986 and 6897 respectively), probably due to the fact that this database has globally more miRNA-mRNA interactions described.

Only 794 of the negatively correlated miRNA-mRNA pairs were simultaneously present in the 3 databases, confirming the little overlap that exists between them. That number corresponds to a 0.88% of the total miRNA-mRNA possible combinations existing from the tissue expression analysis. That means that this step considerably reduces the number of miRNA target interactions that are likely to occur in pancreatic oncogenesis. Moreover, due to the fact that they have been predicted in the three databases, these interactions can be prioritised ahead of the others in terms of confidence.

Figure 4.29 shows the network of all these 704 high-confident interactions. MiRNA-mRNA interactions are divided into downregulated miRNAs and their upregulated target mRNAs (left), and upregulated miRNA with downregulated target mRNAs (right), although the mRNAs can later interact between them.

Interestingly, in this plot is possible to see that miRNAs from the same family share



(a) Negatively correlated miRNA-mRNA pairs and pairs predicted by at least one database

(b) Overlap between the 17401 miRNA-mRNA pairs negatively correlated and predicted at least in one database (TargetScan, miRDB and miRSVR) used

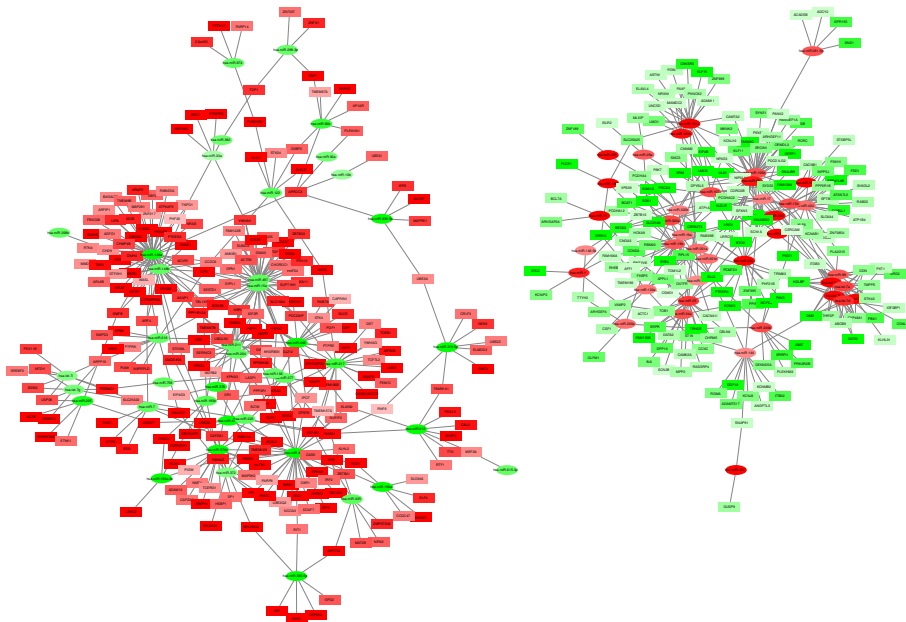
Figure 4.28: Venn diagrams showing the number of miRNA-mRNA interactions predicted by miRComb and intersections between databases.

most of their target mRNAs, which is why they are represented close to each other. For example, *hsa-miR-148a/b* family on the left of the network (Figure 4.29(b)). Both miRNAs are close, which means that share most of their targets, as it can be observed. *Hsa-miR-15a* and *hsa-miR-497* are officially not members of the same family, but they also share most of their targets.

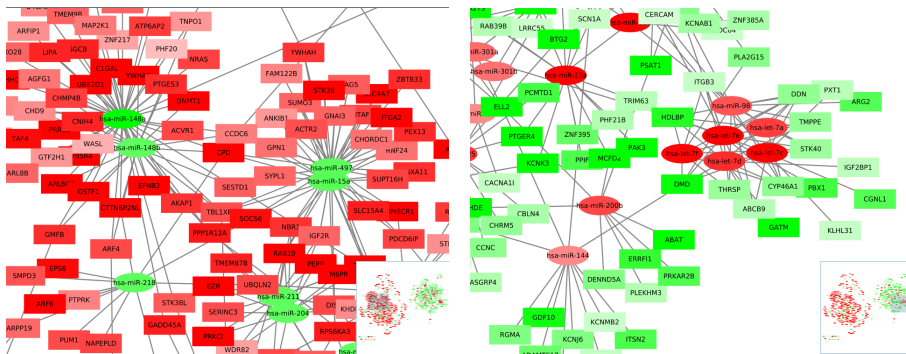
Similarly, members of the *hsa-let-7* family (*hsa-let-7a/c/d/e/f* and *hsa-miR-28*, which is also a member of the *let-7* family) are clustered on the right. *Hsa-let-7i/g* do not cluster with the others as they are downregulated in pancreatic cancer and appear on the network of the left, and *hsa-let-7b* was not differentially expressed so it does not appear here.

4.4.2 MiRComb results in the pancreatic cancer set

Figure 4.30 shows a plot with the top 12 miRNA-mRNA interactions with more significant negative correlations obtained from miRComb that appear simultaneously in the three mentioned databases (Table 4.9). These are the miRNA-mRNA interactions that are more likely to be true. *Hsa-miR-106b*, *hsa-miR-21*, *hsa-miR-148a*, *hsa-miR-93* are responsible of the first 4 miRNA-mRNA interactions. All of them are miRNAs widely studied in a variety



(a) Network of the 704 high-confident interactions



(b) zoom

(c) zoom

Figure 4.29: Full network (a), and zooms of specific parts of it (b), (c). All miRNA-mRNA high-confident interactions have correlation FDR < 0.05 and are simultaneously predicted in the 3 databases.

of contexts, and appear several times on the list [195, 57, 196, 197].

The first plot (*hsa-miR-106b* and *LRRC55*) shows a clear linear regression, but in the others, it is difficult to see the complete trend because the difference in expression between two groups monopolises the distance, and gives results similar to a situation like classical up & dw pairing. We can find miRNAs upregulated with mRNAs downregulated and miRNAs downregulated with mRNAs upregulated in this list. Even in this situation,

miRNA	mRNA	cor	FDR	FC.miRNA	FC.mRNA	dat.sum
miR-106b	LRRC55	-0.97	8.80e-03	2.07	-1.23	3
miR-21	PDCD4	-0.93	9.55e-03	9.91	-7.90	3
miR-148a	YWHA8	-0.93	9.55e-03	-2.98	3.11	3
miR-93	FAM129A	-0.92	1.00e-02	2.60	-11.48	3
miR-330-5p	GPI	-0.91	1.05e-02	-3.64	3.38	3
miR-330-5p	BHLHE40	-0.91	1.06e-02	-3.64	7.97	3
miR-93	LRIG1	-0.91	1.10e-02	2.60	-4.13	3
miR-23a	LRIG1	-0.91	1.10e-02	4.40	-4.13	3
miR-148a	ARF4	-0.91	1.11e-02	-2.98	2.11	3
miR-106b	FAM129A	-0.90	1.13e-02	2.07	-11.48	3
miR-148a	ACVR1	-0.90	1.17e-02	-2.98	2.11	3
miR-148a	CTTNBP2NL	-0.90	1.19e-02	-2.98	2.76	3
miR-107	PDK4	-0.90	1.19e-02	2.08	-12.85	3
miR-106b	LMO3	-0.89	1.24e-02	2.07	-4.07	3
miR-148a	C1GALT1	-0.89	1.24e-02	-2.98	6.38	3
miR-330-5p	CAPN12	-0.89	1.25e-02	-3.64	3.99	3
miR-148a	TBL1XR1	-0.89	1.27e-02	-2.98	2.06	3
miR-320b	KIAA1324	-0.89	1.28e-02	1.66	-12.22	3
miR-320a	LMO3	-0.88	1.31e-02	2.14	-4.07	3
miR-93	SCN1A	-0.88	1.36e-02	2.60	-1.25	3
miR-148a	CNIH4	-0.87	1.37e-02	-2.98	2.46	3
miR-148a	DNMT1	-0.87	1.38e-02	-2.98	3.09	3
miR-320b	RPL15	-0.87	1.38e-02	1.66	-2.09	3
miR-193b	TNFRSF21	-0.87	1.38e-02	-2.05	8.10	3
miR-148a	UBE2D1	-0.87	1.38e-02	-2.98	3.74	3
miR-181a	LMO3	-0.87	1.39e-02	5.17	-4.07	3
miR-193b	YWHAZ	-0.87	1.42e-02	-2.05	2.59	3
miR-424	LRIG1	-0.86	1.45e-02	1.82	-4.13	3
miR-106b	PDCD11LG2	-0.86	1.45e-02	2.07	-1.30	3
miR-130a	LRIG1	-0.86	1.46e-02	1.74	-4.13	3
miR-497	ITGA2	-0.86	1.46e-02	-1.94	23.44	3
miR-15a	ITGA2	-0.86	1.46e-02	-1.96	23.44	3
miR-34a	VAMP2	-0.86	1.47e-02	2.05	-1.50	3
miR-155	SCN1A	-0.86	1.47e-02	4.03	-1.25	3
miR-299-3p	TOP1	-0.86	1.47e-02	-1.87	2.35	3
miR-367	TOB1	-0.86	1.48e-02	1.61	-1.60	3
miR-330-5p	ARPC5L	-0.86	1.48e-02	-3.64	3.17	3
miR-19b	RBM20	-0.86	1.49e-02	2.00	-1.80	3
miR-34a	INA	-0.86	1.49e-02	2.05	-1.72	3
miR-148a	CPD	-0.86	1.50e-02	-2.98	3.44	3

Table 4.9: Top 40 miRNA-mRNA pairs (sorted by adjusted p-value) that have: $p_{\text{val-corrected}} < 0.05$ and appear predicted simultaneously the three databases.

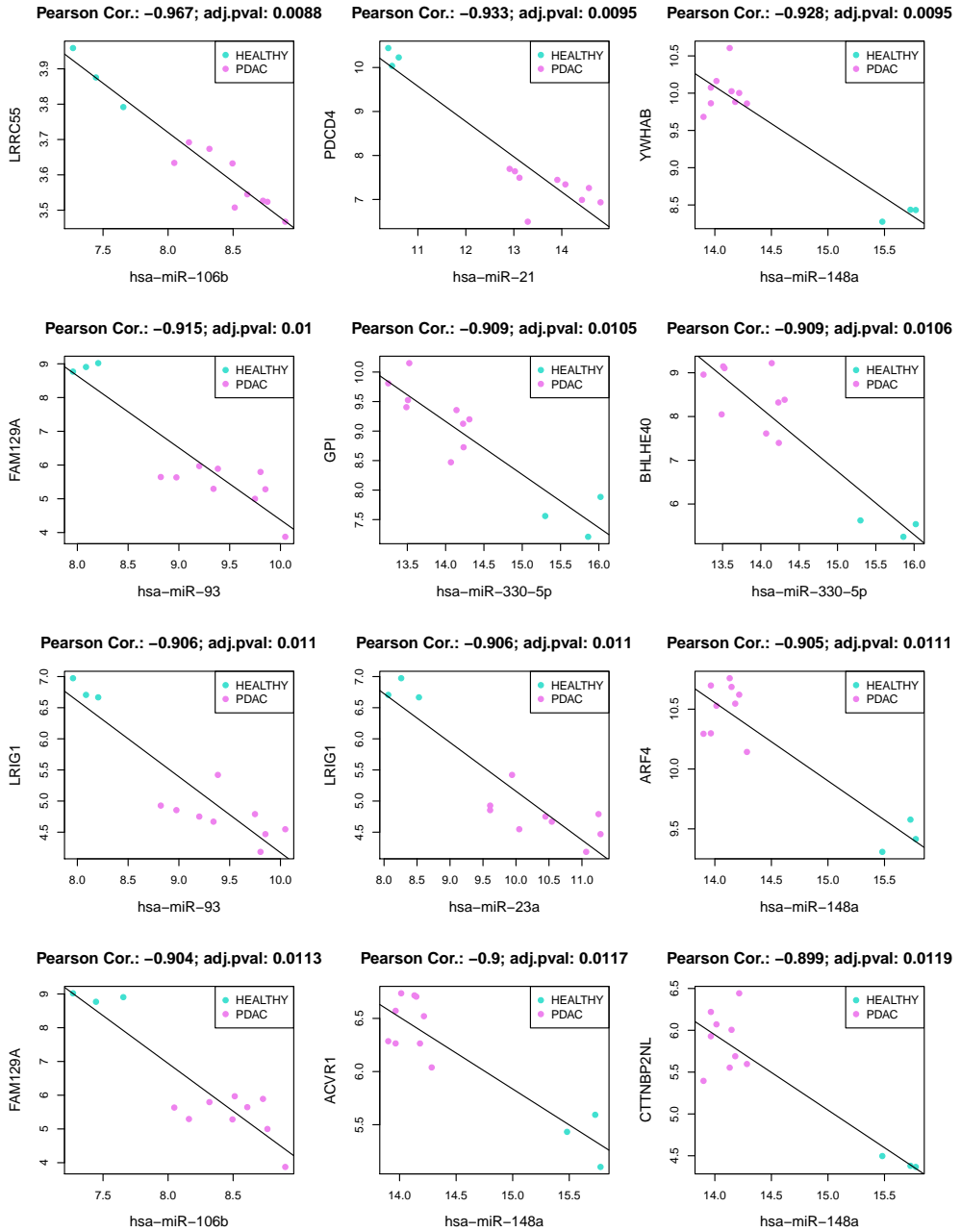


Figure 4.30: First 12 miRNA-mRNA interactions sorted by FDR. All interactions are FDR < 0.05 and predicted in the three used databases: TargetScan, miRDB and miRSVR.

Pearson Correlation is able to correctly select negative relations.

The miRComb report helps to highlight some characteristics of the interactions. Table 4.10 shows the miRNAs with more targets predicted by miRComb. Plus, we added a column showing the number of targets initially predicted by the database, highlighting the reduction of the initial number of miRNA-mRNA interactions.

Each miRNA shows different levels of reduction, the average is 77% but this number is different for each miRNA. Other examples not shown on the table are *hsa-miR-21* (67%) and *hsa-miR-122* (73%). Regarding the miRNAs on the table, for example, *hsa-miR-148a* shows less reduction than the others. Further exploration showed that miRNA-mRNA interactions are preferentially located on negative miRNA-mRNA correlations evaluated by different methods. According to the model proposed in the methods section (Section 3.9), we could check that:

- Specifically, significantly negative correlations showed 2.12 times more targets than non-significant correlations ($p = 1.68e-23$). This number is related to the database false positives from *hsa-miR-148a* and according to our model, that makes that among the 363 miRNA-mRNA interactions predicted by miRComb, 171 might be spurious interactions not taking place or relevant for the disease, but 192 (53%) are expected to be playing a role in the disease.
- With the logarithmic model, we saw that β_1 is -0.55 ($p = 1.26e-21$). This means that, for example decreasing 0.1 units of correlation increments the odds of being predicted by the databases in 5%, 0.2 units 11.6%, etc. AUC is 0.602 and significant ($p = 1.63e-15$), which means that the value of correlation is a predictor of a miRNA-mRNA pair is present in the database or not.
- GSEA Enrichment Score is -0.215 ($p = 3.4e-4$, NES = -4.64), which implies that miRNA-mRNA interactions are preferentially (and significant) located on negatively correlated pairs, using a non-parametric test (Figure 4.31)

This behaviour, however, was not a general trend observed in all the miRNAs and both characteristics (less reduction and concentration of miRNA-mRNA targets on the negative correlations) are not necessarily related.

Moreover, this percentage varies across databases. For example, if we only considered TargetScan database (which predicted 105 miRNA-mRNA interactions), the OddsRatio would have been 2.60 ($p = 2.04e-09$); for miRSVR, 2.19 ($p = 1.4e-23$, predicting 336 miRNA-mRNA interactions); and for miRDB, 3.85 ($p = 4.5e-10$, predicting 71 miRNA-mRNA interactions).

miRNA	#tgs	orig.	cum. %	targets (top 20)
miR-374b	381	(866, 56%)	10.46	PMEPAL, CD58, TMSB10, CCL20, CT5B, HSPH1, DNMT1, DIS3, ELF1, UBAC2, FAT1, CCDC47, PTPN12, COPB1, FAM122B, IL8, CTTNBP2NL, FAM96A, H2AFY, ACVR1
miR-148a	363	(595, 39%)	16.85	HLA-A, KLF5, CT5B, TNFRSF21, TMSB10, BID, TMMEM123, KCNK1, B2M, PGRMC1, YWHAB, TAGLN2, ENDOD1, PTPN12, UBE2A, ACSL3, MYO1D, AMMEGR1, PLEKHB2, ACTG1
miR-181a	259	(828, 69%)	23.96	PDCD4, IFRD1, DFFB, EPB41L4B, ANGPT1, LRIG1, KCNN1, NUCB2, DMGDH, FKBP11, EPB41, TMED6, LMO3, VCX2, MYO15A, RPL15, SLC25A53, PSAT1, ITSN2, SPATA20
miR-373	258	(647, 60%)	26.52	HLA-A, ENDOD1, B2M, PGRMC1, BID, DIS3, TAGLN2, CCDC47, PTPN12, MDK, PON2, MYO1D, SKAP2, CTTNBP2NL, FAM96A, IL8, H2AFY, PSMA2, ACVR1, C1D
miR-320a	252	(751, 66%)	31.62	WNT9B, PDCD4, TMED6, PAIP2B, SFTPC, ADRA1B, MS4A10, HHIPL1, CACNB1, AOX1, IFRD1, SND1, CECR2, GPHA2, KCNAB1, OSBP2, ERO1LB, EPB41L4B, LMO3, BACE1
miR-448	245	(608, 60%)	33.24	ENDOD1, GBP2, LITAF, LIMS1, DNMT1, ELF1, PTPN12, IL8, FAM96A, VPS13C, SEPT10, SKAP2, CTTNBP2NL, FAM122B, CALM2, RBM41, PPFIA1, IVNS1ABP, NEK6, PFKP
miR-93	238	(813, 71%)	36.62	IFRD1, FAM129A, LRIG1, ATXN7L2, MLC1, EPB41L4B, SH2D5, ANGPT1, ISM2, MS4A10, SYBU, SCN1A, MYO15A, PCMTD1, FBXO24, SLC46A2, EPB41, ITSN2, PAIP2B, WNT9B
miR-106b	234	(766, 69%)	37.47	LRRC55, FNDC5, ZNF385A, SH2D5, FAM129A, MYT1, MLC1, LMO3, IFRD1, C17orf67, KPNA7, APOBEC3H, SLC41A1, TIMM8A, ATOH8, PAIP2B, ARHGAP18, ERO1LB, PRND, MUMIL1
miR-217	230	(533, 57%)	39.28	TNFRSF21, CTNNA1, ARPC2, CLINT1, RAB11A, YWHAH, KLF5, PFKP, MAP4K4, YWHAB, CAP1, PTTG1IP, RAC1, SPTLC2, ADAM9, PRKCI, ISG20, TES, DDX60, TMEM87B
miR-539	225	(868, 74%)	41.23	CCDC109B, NQO1, SULF2, KCNKI, MARCKSL1, ITGA2, PSMB8, ARPC2, DENND2D, HSBP1, SLC44A1, MRPL50, B2M, ENCI, FAM108C1, MAT2B, GCC2, HLA-A, DYNLT1, PNP

Table 4-10: Top 10 miRNA with more targets (each miRNA-mRNA pair has pval-corrected<0.05 and appears at least 1 times in the following databases: targetScan v7.1_17, miRSVR_aug10_17, miRDB v5.0_17). MiRNAs in red are upregulated in PDAC, miRNAs in green are downregulated in PDAC. #tgs : Number of target mRNAs; orig.: number of miRNA-mRNA pairs predicted in at least one database (considering positive and negatively correlated miRNA-mRNA pairs) and percentage of these original pairs that are removed after considering only negatively correlated miRNA-mRNA pairs.

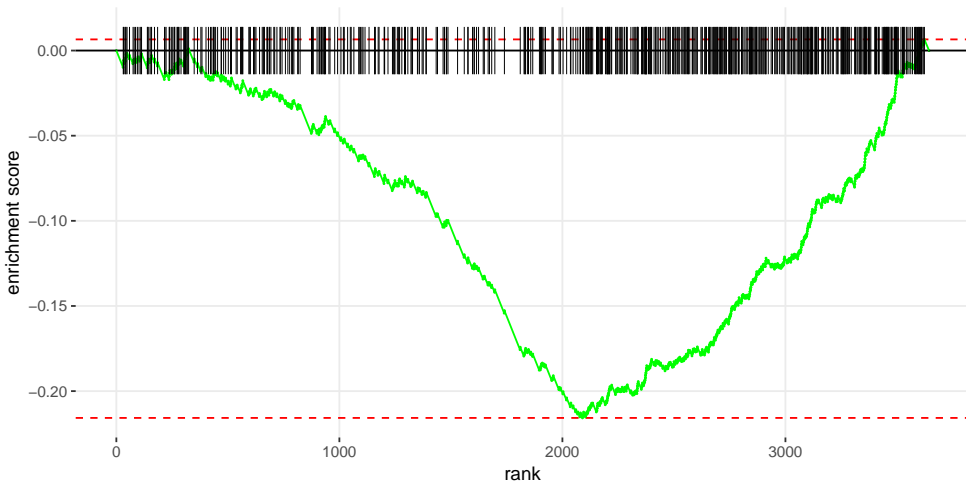


Figure 4.31: GSEA plot of *hsa-miR-148a*. Positive correlations are on the left and negative on the right. Bars show miRNA-mRNA pairs that are predicted in at least 1 database of TargetScan, miRSVR or miRDB. Significant negative correlations start at rank 2098.

Despite these variations, in overall terms, the OddsRatio showed more variation across miRNAs (inter-quartile range of the logratios 1.31 for the combination of the three databases, and 1.16, 1.13 and 0.95 for TargetScan, miRSVR and miRDB respectively) than across databases (mean inter-quartile range of the logratios is 0.31, in the miRNAs that were possible to evaluate).

Target enrichment on revealed target enrichment of *hsa-miR-148a* targets in the Notch Signaling Pathway (FDR = 0.01). Figure 4.32 shows the targets involved in the pathway. This pathway participates in cell fate control and signal integration development [198], and has also been related to cancer development [199, 200].

Other observations can be obtained from the miRComb report. Figure 4.33 shows a summary of the number of targets per miRNA, and a summary of number of miRNAs per mRNA target. Figure 4.33(a) shows the number of targets per miRNA and the cumulative number of mRNAs that are regulating (among the deregulated mRNAs). 50% of the mRNAs are regulated by the top 17 miRNAs, and almost no mRNAs are added by the last others. Both figures show that 75% of the deregulated mRNAs are targeted by at least one miRNA.

Specifically, Figure 4.33(b) shows the number of miRNAs that are targeting each mRNA. As said before, 75% of the mRNAs are targeted by at least one miRNA, and interestingly, 1149 mRNAs (41.7% of the deregulated ones) are targeted by more than 5

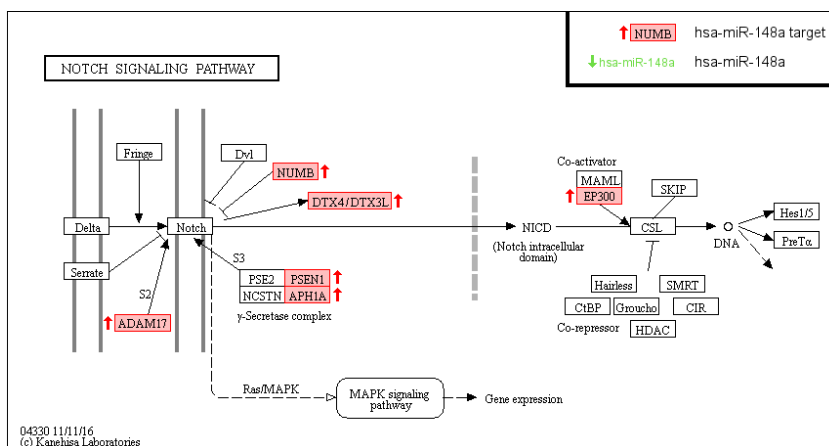


Figure 4.32: KEGG Pathway for *hsa-miR-148a* targets in the context of pancreatic cancer: *Notch Signaling Pathway*. *Hsa-miR-148a* targets (mRNAs that have a negative correlation with *hsa-miR-148a* (FDR < 0.05) and predicted in at least one database) are highlighted in red.

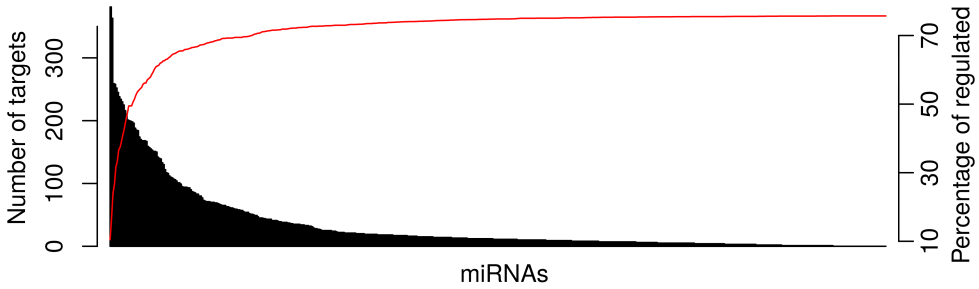
miRNAs.

4.4.3 Confirmation of miR-21 targets in a pancreatic cancer cellular model

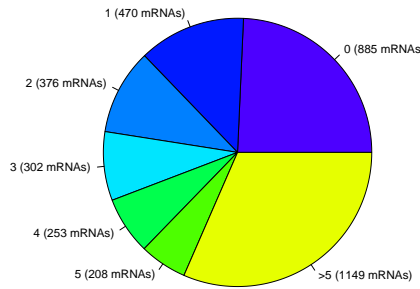
To go one step further, we focused on the list of miRComb predicted targets for *hsa-miR-21* (Table 4.11). *Hsa-miR-21* is currently one of the best studied miRNA that plays relevant roles in cancer as it is named oncomiR-21 [57, 201]. It has also been described to have important roles in pancreatic cancer [75, 77, 202, 203]

In order to experimentally confirm if some of these predicted targets act as *hsa-miR-21* targets in the context of pancreatic cancer, we selected 2 targets (PDCD4, BTG2) from the list (Table 4.11). These targets have good correlation values and have been predicted in more than one database. Furthermore, their interactions with *hsa-miR-21* have been described in other cancer lines (PDCD4: [204, 205, 206, 207, 208], BTG2: [209, 210, 211]), and thus this makes them good candidates to be validated in our model. Selecting mRNA targets predicted in more than one database also tries to avoid the number of false positives.

We generated a pancreatic cancer cellular model (PANC-1) lacking *hsa-miR-21* expres-



(a) Barplot showing the number of mRNA targets per each miRNA (each bar represents a miRNA and they are sorted by the number of targets). MiRNA-mRNA interactions have p -value < 0.05 and predicted at least 1 time on the following databases: targetScan, miRSVR miRDB. Red line (and right axis) represents the percentage of deregulated mRNAs that are cumulatively targeted by the miRNAs.



(b) Pie chart representing the number of miRNAs targeting the mRNAs, p -value < 0.05 and predicted at least in one of the following databases: TargetScan, miRSVR, miRDB

Figure 4.33: Plots summarising the number of interactions per miRNA and mRNA.

sion by using CRISPR/Cas9 methodology. After confirming which clones did not express *hsa-miR-21* (Figure 4.34A), we measured the basal expression of PDCD4 and BTG2 in three PANC-1 KO miR-21 clones, and compared it with the control pancreatic cancer cell line PANC-1 expressing high levels of *hsa-miR-21*. As expected, both, PDC4 and BTG2, showed significantly increased expression in the absence of *hsa-miR-21* (Figure 4.34B, 4.34C), confirming that the expression of these genes is, in part, regulated by *hsa-miR-21* in a pancreatic cancer context.

miRNA	mRNA	cor	adj.pval	TargetScan	miRSVR	miRDB	dat.sum
miR-21	PDCD4	-0.93	0.01	1	1	1	3
miR-21	PAIP2B	-0.90	0.01	1	1	0	2
miR-21	SMARCD1	-0.88	0.01	1	1	0	2
miR-21	SERP1	-0.85	0.02	1	1	0	2
miR-21	B3GAT2	-0.84	0.02	0	1	1	2
miR-21	BTG2	-0.84	0.02	1	1	0	2
miR-21	BCL7A	-0.83	0.02	1	1	1	3
miR-21	ALX4	-0.83	0.02	1	0	1	2
miR-21	SEC63	-0.81	0.02	0	1	1	2
miR-21	RNF182	-0.79	0.02	0	1	1	2
miR-21	ARHGAP24	-0.79	0.02	1	1	1	3
miR-21	STK40	-0.79	0.02	1	0	1	2
miR-21	CNTFR	-0.78	0.02	1	1	0	2
miR-21	NPAS3	-0.77	0.02	1	1	0	2
miR-21	ABAT	-0.77	0.02	0	1	1	2
miR-21	KLF9	-0.76	0.03	1	1	0	2
miR-21	EPM2A	-0.74	0.03	0	1	1	2
miR-21	ADCY2	-0.73	0.03	0	1	1	2
miR-21	PIKFYVE	-0.70	0.04	1	1	1	3
miR-21	SLC16A10	-0.70	0.04	1	1	0	2

Table 4.11: Targets of *hsa-miR-21* that have FDR < 0.05 and are predicted in at least two databases. TS: TargetScan database; SVR: miRSVR database; mDB: miRDB database.

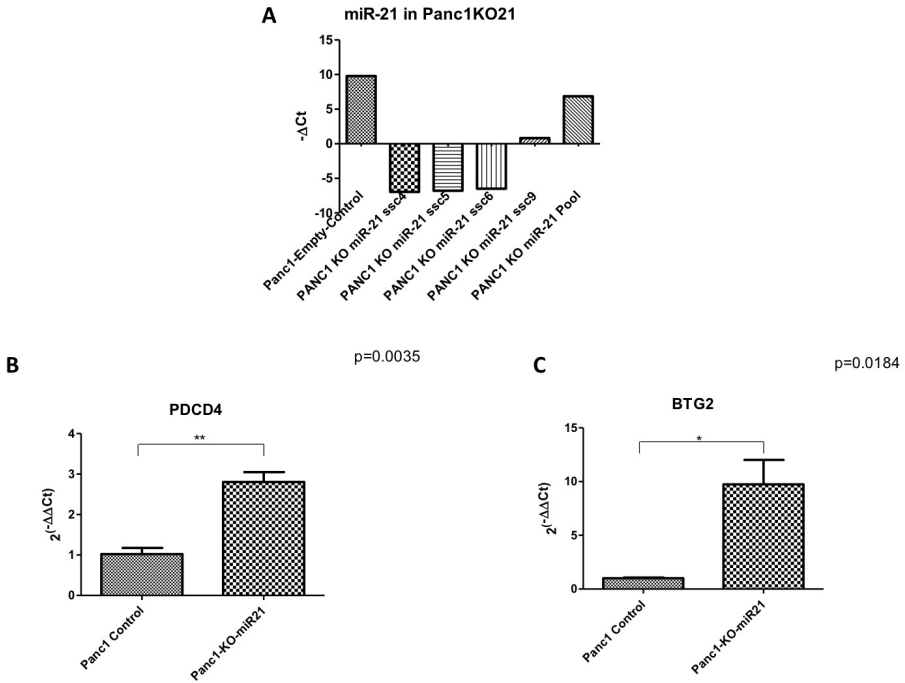


Figure 4.34: Figure showing the validation of *hsa-miR-21* targets. A) expression of *hsa-miR-21* in the generated PANC-1 KO miR-21 clones and empty control. B) expression of PDCD4 in both Panc1 Control and PANC-1 KO miR-21 clone. C) expression of BTG2 in both Panc1 Control and PANC-1 KO miR-21 clone. Relative expression levels of mRNAs were calculated as $2^{(-\Delta\Delta Ct)}$. Statistical differences between groups ($n = 3$) were computed by using non-parametric tests (Mann-Whitney test).

Chapter 5

Discussion

If we had to summarise all the work in one sentence we would say that in this work it is possible to see how to go from a huge amount of data to a few biologically relevant results. Moving from the raw data to the final results is a process that involves many steps. Main results of this thesis have been already commented in the previous chapter, but in this chapter we will summarise the main findings of each study, as well as aspects that have been remained unmentioned, and options than can help to improve further research in the future.

MiRComb package

The first step was **to think about a procedure** that can be used to answer our question: *which are the most probable relations between miRNAs and mRNAs that are biologically occurring in this dataset?*. There are some articles describing integrative procedures for miRNA and mRNA data (Gade et al., 2011 [120] was the more influential article for its clarity), but this field is still growing and several articles describing integrative analysis of miRNA and mRNA expression data have been published during the development of this thesis [117, 212, 213, 214, 215, 216, 217]. However, having no clear example of how to do this kind of analysis, we present here a method which connects differential expression analysis, correlation analysis, p value combination or database interaction and p value correction in order to obtain a list of miRNA-mRNA pairs with a reliable p value associated, and some tools for interpreting this list, all in one single software environment

(Figure 3.3).

We chose **R** for implementing our pipeline due to several reasons. First of all, jointly with Bioconductor, it is probably one of the most used software for analysing big data in biology [128]. Secondly, it allows to build packages (your own functions) and easily share them with other researchers and/or make it public. Thirdly, it is free and works with any platform (Unix, Windows, Mac), which makes accessible for anyone. And finally, it is well documented and kept updated. There is a huge community of R-users that share programming tips or help to fix any possible reported bug. For that reason, miRComb has been also uploaded to GitHub [218].

Once the general ideas of the pipeline were postulated, we built the miRComb **R package** and published it [187]. Apart from storing the most widely used methods, it allows us to choose between different options in each step. For example, in the correlation analysis step, we can choose between Pearson, Spearman and Kendall correlation (and recent versions include Glmnet procedure), in differential expression we can use parametric or non-parametric methods, and the filters are easily modulated. It is very important that the package allows to choose between these options because, although in this work we concluded that `limma`, a filter of $FDR < 0.05$, Pearson correlation and database intersection were the best options to analyse our datasets, it is possible to use the same tool using other parameters.

Regarding the **design of the package**, the modular structure of the package allows us to include other methods and improving the package itself whenever it is necessary. For example, time-series analysis has been also included as a new method for selecting differentially expressed miRNAs. It is also very easy to include new miRNA-mRNA databases and updates of the ones that you are already using. Until now, that has allowed us to integrate more miRNA-mRNA prediction databases in successive releases (for example TargetScan, miRDB and miRSVR, that were not included on the first release). Methods for testing differential expression in NGS data such as DESeq2 [147] or edgeR [148] have been also included on the last version of the package and are fully integrated into the pipeline, as the `addCorrelation` function is able to use DESeq and edgeR weights to normalise the data before making the correlations.

It is also worth mentioning the **report function**, which was included after several updates of the package. The sections of the report were designed using our own criteria but also after receiving feedback from the scientists that used the package. It has been also one of the most appreciated functions by the users, and it gives an added value to the package.

MiRTools, despite not being part of the "typical" miRComb analysis, are web-based tools that we have developed in the context of this thesis, aimed to make easy to deal with miRNA data. MiRTranslator is similar to other R packages such as `miRNANameConverter` [134], but it is on-line and includes the "*I don't know*" option, which add advantages to the preexisting ones. MiRCircos is a tool to plot miRNA-mRNA interactions in a circos-like plot [192, 193] and explore if there are any overtargeted regions on the genome specified by the user using an hypergeometric test.

The **downloads** of the miRComb package and related documentation around the world have achieved a satisfactory rate. Downloads increased with the publication of the PLoS ONE article and at this moment are stable at around a mean of 100 downloads (including miRComb, miRData and vignettes) per month. It is hard to evaluate the real impact because miRComb is aimed to help to design new hypotheses, that still need to be tested. Nevertheless, as shown in the last section, there have been scientists interested on the package, they have used it, and in fact, some studies have been already published [185, 186, 188, 189]. Moreover, we have received more than 20 consultations and queries via e-mail, mostly related to small installation problems that were easily solved but others suggesting more functionalities, that some of them have been included on the different miRComb releases, such as time-series analysis or non-matched miRNA-mRNA data.

MiRComb parameter settings

In the Section 4.2, we compared different miRNA-mRNA settings. We evaluated the strengths and weaknesses of all the options that miRComb offers.

Despite having removed miRNAs and mRNAs with a mean of less than 10 counts, NGS specific **differential expression methods** seem to be a little less sensitive with low expressed miRNAs and mRNAs, suggesting that the limit of 10 counts may be still too low. On the contrary, voom transformation + limma is able to detect differentially expressed miRNAs and mRNAs in all ranges of expression. T-test and Wilcoxon test found a similar number of differentially expressed items, but Wilcoxon is most dependent on sample size than the others (for example you will not be able to have a p value less than 0.1 in a sample size of 3 per group, while you will be able to achieve significantly lower p values on the t-test if the sample groups show enough differences). Thus, specific methods designed for analysing gene expression such as limma or RankProd are expected to be more accurate and should be used [137, 138, 146]. RankProd should only be used if a normal distribution of the data cannot be assumed, otherwise limma is more sensible.

Moreover, these results show that NGS data can be transformed with voom and analysed with limma, obtaining reliable results, as suggested by Law et al. [137]

Regarding the **filtering for differentially expressed miRNAs/mRNAs**, we analysed if the previous filtering of differentially expressed miRNAs and mRNAs helps the detection of significant negatively correlated miRNA-mRNA pairs. The dataset of Section 4.2 shows that previous filtering (and consequent reduction of the number of miRNA-mRNA interactions being tested) decreases the total number of miRNA-mRNA significantly negatively correlated pairs (Figure 4.11(a)). However, the proportion of significant negatively correlated miRNA-mRNA pairs increases with previous filtering (Figures 4.11(b) and 4.12), meaning that these negative correlations are more likely to be found within differentially expressed miRNAs and mRNAs.

Previous datasets analysed with miRComb showed also this tendency [219]. Moreover, in that analysis, apart from the proportion of significant miRNA-mRNA pairs, the number of miRNA-mRNA interactions increased when significant miRNAs and mRNAs are selected. That means that although filtering allows to focus on specific miRNA-mRNA pairs (which is usually the aim of the study), it is always suggested because it reduces the time of the analysis and no relevant miRNA-mRNA pairs are usually lost.

Exploring the **correlation methods**, we observed that Pearson and Spearman coefficients give similar results, and Kendall differs more of the former ones. Spearman detects more negative correlations when the number of samples is high (Figure 4.13(a)), but has less sensitivity in smaller datasets (Figures 4.13(b) and 4.13(c), and study [219]). Top miRNA-mRNA interactions of Studies 1 and 2 (Sections 4.3 and 4.4 respectively) plotted on the Appendix A also point that miRNA-mRNA relations may not be completely linear, but Pearson coefficient is able to summarise the main trend and keep this essence. This data suggests that miRNA-mRNA relations might not be completely linear, but Pearson coefficient is still able to detect the significant ones. Non-parametric or more sophisticated relations (such as polynomial regression, or logarithmic trends) may be suitable when the number of samples is high, while Pearson coefficient is suggested for smaller datasets, in which non-parametric methods are less powerful. In summary, Pearson is suggested for smaller datasets and for making the study comparable between others (as Pearson correlation has been so far the most used one in literature [120, 122, 123, 216, 220, 221, 222]).

Regarding **Glmnet methodology** [169], several models have been tested: Glmnet-miRNA, Glmnet-mRNA and Glmnet-mRNAs. The later is the one that represents better the biology known about the topic (a miRNA can simultaneously be regulating several mRNAs, and a mRNA can be regulated by more than one miRNA, including competence between miRNAs [83, 20]), but it takes a considerable amount of time to estimate all the

coefficients in the same model and also can be unstable with normal or small datasets.

Alternative models such as Glmnet-miRNA and Glmnet-mRNA were formulated to avoid these problems. Glmnet-mRNA coefficients are similar to Glmnet-mRNAs although is only able to estimate 10% of them (resulting the selection of much more less miRNA-mRNA pairs than Pearson correlation). On the contrary, coefficients estimated by Glmnet-miRNA are not related either to Glmnet-mRNAs or Glmnet-mRNA and may correspond to the mRNAs that negatively regulate miRNA expression, and not viceversa. We still recommend Pearson Correlation for comparative purposes, but we think that Glmnet methods can be further explored. More specifically, Glmnet has an option to weight predictors. For example, miRNA-mRNA pairs predicted by the databases can be forced to take part in the model, or can be prioritised ahead of the miRNA-mRNA pairs that have not been predicted. This option is not available for Glmnet-mRNAs formulation, but it is for Glmnet-mRNA and Glmnet-miRNA. As we have seen that Glmnet-mRNA seems to be a valid model, it will be explored on further releases of the package.

Finally, the different methods of **database integration** were compared. *P* value combination is the formula that fits better the real model. Despite Fisher combination may give significant results with only one initial hypothesis being enough significant, Stouffer combination reflects the fact that both initial hypothesis (negative correlation and predicted on the database) must be true. However, *p* value combination has a big problem of implementation, which is that not all the miRNA-mRNA prediction databases report *p* values. MicroCosm is the only database that reports *p* values but it is not the most recent one, and newly discovered miRNAs are not taken into account, making a bias against these miRNAs. Moreover, microCosm only reports *p* values lower than 0.05. In our analysis, we set all unreported *p* values to 1 as suggested by Gade et al. 2011 [120]. This is the most conservative option and also reflects well the reality, as we know that unreported values correspond to non-significant miRNA-mRNA predictions; but it may also induce a bias on the FDR computation. Figure 4.18(b) shows that including one more database compensate longer than required the miRNA-mRNA interactions that might be lost using the Stouffer method.

MiRComb analysis in five digestive cancers

Having explored all the options of miRComb, we proceeded to analyse five different digestive cancers: COAD, ESCA, LIHC, READ and STAD using freely available data from The Cancer Genome Atlas [154]. This was the first test of miRComb with high-throughput

data.

We decided to use the following criteria to analyse the miRNA-mRNA interactions for the following reasons:

- **No filter** for differentially expressed miRNAs and mRNAs. Although we recommended filtering whenever it is possible, most of the studied cancers in the TCGA repository were highly imbalanced (the less imbalanced was LIHC but still at a ratio of 1 control per 7 cases), and in some others there were no enough control samples to make a clear assumption, READ for example had only 3 controls available. For this reason, we decided to not use the FoldChanges for filtering, but we kept them for informative purposes.
- We chose **Pearson correlation coefficient** as a measure of correlation because, 1) after checking non-parametric coefficients, linear assumption were reasonable and 2) as our aim was to proportion data to the scientific community, it was important that our measure was comparable, and Pearson correlation coefficient is the most used in this kind of studies.
- **MicroCosm and TargetScan databases** as the most consolidated ones, which implies to use database intersection method and define a miRNA-mRNA interaction as: $FDR < 0.05$ (corrected p value from the original unilateral test, where $H_0 = r_\lambda \geq 0$) and predicted in MicroCosm or TargetScan (or both).

On the Section 4.3.1 we explored individually each cancer according to the selected parameters, chosen to enhance comparability among datasets. Main results are showed on the reports. Not surprisingly, we saw that cancers with more samples and a more balanced dataset were more sensible to detect miRNA-mRNA interactions. Control samples are important to emphasise the differences in expression between healthy and disease samples, but we were still able to detect miRNA-mRNA interactions on specific cancers even with low controls because of the expression range of miRNAs and mRNAs in the cancer samples, probably due to different stages of the patients' cancer.

Reports were shown to be a good tool for comparing the different analysed cancers. We obtained **thousands of miRNA-mRNA interactions** that were occurring on each cancer, plus tens of pathways that can be controlled, at least in part, by miRNAs.

We observed that some miRNAs have less **database false positives** (miRNA-mRNA interactions predicted by the databases that are not negatively correlated) than others, and the range was quite huge (99% to 22%). For example, Figure 4.20 highlights that some of the miRNAs with less false positives have been already described as key regulators

of the liver machinery in cancer [194], suggesting that miRNAs that are playing a role in the disease may have also less percentage of false positive miRNA-mRNA. That means that the ratio of database false positives can be related to the function of the miRNA in this specific context. However, the difficulty of defining what is "related to a disease" makes hard to define a global test for checking this effect.

Some **functional enrichment** was described on targets of *hsa-miR-148a* in liver cancer. Specifically, 7 genes participating in the *Antigen Processing and Presentation* KEGG pathway are overexpressed in liver cancer because they are targets of *miR-148a*, which is underexpressed in it (Figure 4.21). This pathway participates in tumour immunogenicity and has been described to be altered in cancer [223, 224]. Although this result should be validated in experimental models, this may be indicating that our miRNA-mRNA predictions are biologically meaningful.

Regarding the **integrative analysis**, we found 1570 miRNA-mRNA interactions present in all five digestive cancers from Study 1. In this subset of miRNA-mRNA interactions, we identified mRNAs that were present on already known pathways related to cancer such as *Mapk signaling pathway* (which is involved in cell growth [225]). The genes on this pathway are already known potential target drug candidates [226] and these results show that they can also be considered to miRNA-based therapies in the context of these digestive cancers. Pathways and target mRNAs specific for each cancer still have to be explored among the miRNA-mRNA specific interactions, and new pathways may still need to be elucidated.

Moreover, integrative analysis of the miRNA-mRNA interactions allowed us to **classify** original cancers according to the values of correlation of their miRNA-mRNA interactions. Cancers with similar pathophysiology share more miRNA-mRNA interactions than cancers that are more different. Furthermore, the number of interactions of each miRNA is related to the type of cancer (Figure 4.25). These results suggest that miRNA-mRNA interactome may be associated with cancer type and tissue of origin.

MiRComb analysis in pancreatic cancer

In the second study, we analysed the miRNA-mRNA interactome in the context of **pancreatic cancer** with an own dataset. This study allowed us to test miRComb performance in a small set of samples. Small datasets are usually the norm in diseases in which samples are hard to obtain, either because they are rare diseases or because they are not easily

diagnosed. Pancreatic cancer is not one of the most frequent ones (Figure 1.6, [59]) and it is usually diagnosed on late phases of the disease. We wanted to check if the miRComb methodology was also useful in these cases.

Exploratory analysis of the data revealed different expression patterns on both miRNA and mRNA dataset between healthy and tumour tissue (Figure 4.27). In contrast to the previous study, these differences (and the fact that the ratio between cases and controls was considerably lower than on the case of The Cancer Genome Atlas Data) allowed us to select differentially expressed miRNAs and mRNAs, focusing on the study of miRNA-mRNA interactions taking place in the specific context of pancreatic cancer.

In this study, we also updated the **miRNA-mRNA predictions databases** used. For that reason, we decided to use the three most recent databases that included support vector regression methods to predict miRNA-mRNA interactions, which are TargetScan [89], miRSVR [87] and miRDB [90]. The little overlap on the predictions of these three databases (Figure 4.28(b)) reinforces the idea that is better to use more than one database and take advantage of other sources of information such as miRNA and mRNA expression to filter out the results. Although it was not observed in our data, previous studies suggest that combinations of miRNA-mRNA databases have less false positives [86] and this should be considered for further studies. In our study, we used the number of coincidences across database to prioritise the miRNA-mRNA interactions that would be selected for validation.

Final miRComb number of potential miRNA-mRNA interactions in pancreatic cancer is 17401, that corresponds to a 0.88% of the total miRNA-mRNA possible combinations existing from the tissue expression analysis. Although the experimental confirmation of these interactions have not been done, and therefore there may be some false positives among them, this reduction means that this analysis considerably reduces the number of miRNA target interactions that are likely to occur in pancreatic oncogenesis and helps to interpret the data.

Comparing the number of miRComb selected miRNA-mRNA pairs with the initial miRNA-mRNA interactions predicted by the databases, we saw that around 77% of predicted miRNA-mRNA interactions do not show negative correlations and can be considered as **database false positives** (Figure 4.28(a), current estimations of database false positives range from 24% to 70% [86]). However, we observed that this number is **different for each miRNA**, for example *hsa-miR-122* is the miRNA with less database false positives in LIHC (22%) [187], but not in pancreatic cancer, where it has 73% of database false positives, a percentage much higher than other miRNAs such as *hsa-miR-148a* (39%). These percentages also vary depending on the databases used, but are higher across miRNAs,

suggesting that this value should better not be used to evaluate the number of database false positives, but may be useful for evaluating the false positives per miRNA.

The model of miRNA-mRNA distributions presented in the methods section (Section 3.9) is similar to the evaluation of the number of false positives, but it also takes into account the number of negative miRNA-mRNA interactions. Like in the case of false positives, it shows more variation between miRNAs than between databases.

Apart from that, we observed that miRNAs that are described to be **related to a disease** (such as *hsa-miR-148a* in pancreatic cancer [227, 228] and *hsa-miR-122* in liver cancer [194]) seem to have less database false positives, suggesting also that this number may be related to its importance on the disease. Although we were not able to compare this trend in all the miRNAs, based on these findings, we hypothesise that perhaps most of the "database false positives" may be circumstantial false positives, as these miRNA-mRNA interactions might be observed in other contexts.

The **percentage of mRNAs regulated by miRNAs** is 75% in pancreatic cancer (Figure 4.33(b)). This percentage is computed among deregulated mRNAs (and therefore relevant in pancreatic cancer), but in the first study, similar percentages were found on the total amount of mRNAs (from 46% (ESCA) to 78.9% (LICH), Appendix A) [187]. These estimations are in concordance from the percentage of 60% described in the literature [22, 229], but, apart from the total number of mRNAs regulated, what is relevant is the large number of targets that have some miRNA hubs, as well as the large number of mRNAs that are regulated by more than 5 miRNAs.

Strikingly, the top 20 miRNAs with more targets are able to regulate 41.23% of the deregulated mRNAs in PDAC. Moreover, 31.5% of the mRNAs are regulated by more than 5 miRNAs. Usually, mRNAs are not regulated by one single miRNA, and competence and cooperativity between miRNAs have also been described [20, 83]. Altogether, these data confirms that miRNAs are acting as **fine-tuning regulators** in a wide range of diseases [230, 231, 232].

Table 4.9 shows the **top miRNA-mRNA interactions** that should be more likely to occur in a pancreatic cancer context, given that they are the most negative correlated that appear simultaneously in the three target prediction databases. Among them, it is noteworthy that there are miRNAs that have been previously described as important in pancreatic cancer for being significantly upregulated or downregulated in tumour tissue in comparison to healthy pancreas. For example, *hsa-miR-106b*, *hsa-miR-107*, *hsa-miR-130a*, *hsa-miR-34* [73], *hsa-miR-93*, *hsa-miR-155*, *hsa-miR-181a*, *hsa-miR-21*, *hsa-miR-23a*, *hsa-miR-320a* [77], *hsa-miR-193b*, *hsa-miR-320b* [233] are significantly upregulated

and *hsa-miR-148a* [234, 235], *hsa-miR-330-5p* [236], *hsa-miR-373* [237] significantly down-regulated. It is important to highlight the high number of *hsa-miR-148a* interactions that appear among the most significant (12/50), suggesting it may have a central role in pancreatic tumorigenesis. It is likely that *hsa-miR-148a* is involved in more pancreatic cancer pathways than those reported so far for apoptosis and cell survival [238, 227].

Table 4.10 highlights the **miRNAs with more targets**. These miRNAs are probably those playing more central roles in PDAC because they are the ones with more targets and they seem to be regulating a huge number of mRNAs simultaneously. Most of them appear also on Table 4.9 and, as mentioned above, most of them have already been reported to be significantly deregulated in PDAC. These data suggests that these miRNAs may constitute central players of pancreatic tumorigenesis and could be new therapeutic target candidates.

Returning to *hsa-miR-148a*, it is important to note that **functional enrichment** analysis according to its targets by KEGG revealed significant target enrichment in the *Notch signaling pathway*. Figure 4.32 shows proteins involved in that pathway highlighting those that appeared as miRComb predicted targets for *miR-148a* as NUMB, DTX4, DTX3L, PSEN1, APH1A, ADAM17 and EP300. In recent years, accumulated evidence has demonstrated that *Notch signaling pathway* plays critical roles in the development and progression of PDAC [199]. It has been well documented that the *Notch signaling pathway* is critical for cell proliferation, differentiation, development and homeostasis [239]. Reactivation of Notch signaling is observed in early PDAC pathogenesis and persists throughout the progression of the disease [240, 241, 242, 243, 244]. However, no relationships between *hsa-miR-148a* and *Notch signaling pathway* have been described so far in pancreatic cancer and more studies would be needed to confirm and explore this relationship. In this sense, evidences about *hsa-miR-148a* regulation of Notch pathway members have been recently reported in hepatocellular carcinoma [245].

Another important miRNA that seems to play important roles in PDAC is *hsa-miR-21*, as we can see it appears in Table 4.9 with its interaction with PDCD4. *Hsa-miR-21* is currently one of the best studied miRNAs that plays relevant roles in cancer as it is named as **oncomiR-21** [57, 201]. It has also been described to have important roles in pancreatic cancer [75, 77, 202, 203]. In order to experimentally confirm if some of these predicted targets act as *hsa-miR-21* targets in the context of pancreatic cancer, we selected 2 targets (PDCD4, BTG2) from the list of prioritised miRNA-mRNA interactions (Table 4.11). Both PDCD4 and BTG2 are described to play a tumor suppressor role in several cancers and are downregulated in PDAC [246]. PDCD4 is also a known target of *hsa-miR-21* in several types of cancer (colon cancer [204, 205] or diffuse large B-cell lymphoma

[206]), including PDAC [207, 208]. BTG2 has been related to pancreatic cancer [247], and the relation between *hsa-miR-21* and BTG2 interaction has been observed in other cancers (multiple myeloma, [209], liver cancer [210], prostate cancer [211]), but they still have not been directly linked in pancreatic cancer.

In this study we have confirmed the involvement of **miR-21-PDCD4 and miR-21-BTG2 interactions** in the pancreatic cancer cell with the help of a *hsa-miR-21* depleted CRISPR/Cas9 generated model. However, we cannot affirm that all the interactions proposed by miRComb really exist because they should be experimentally validated one by one. Nevertheless, the aim of this study was to unveil a list of high confident miRNA-mRNA interactions for pancreatic cancer that can be the seed for a high number of studies aiming to understand the molecular pathogenesis of PDAC more deeply.

Chapter 6

Conclusions

- MiRComb is a modifiable R package that analyses miRNA-mRNA interactions and summarises the output into a user-friendly pdf report. This tool helps scientists to filter the huge amount of potential interactions given by preexisting databases.
- There is no unique approach for all miRNA-mRNA interactome analysis and settings have to be adjusted for every dataset. However:
 - Filtering for differentially expressed miRNAs and mRNAs between two or more conditions is recommended whenever is possible.
 - Among the tested methods, Pearson correlation between miRNA and mRNA data is the most efficient method to compute miRNA-mRNA associations.
 - Using more than one target prediction database is recommended to increase the number of predicted miRNA-mRNA interactions. But a minimum of coincidences between databases should be used in order to prioritise the most high-confident miRNA-mRNA interactions.
- Despite using combinations of preexisting miRNA-mRNA predictions databases, this theoretical information presents a high number of false positive miRNA targets. This number is highly variable across miRNAs and types of cancer, ranging from 20% to 99%.
- The miRNA-mRNA interactomes of colon, rectum, esophagus, stomach, liver and pancreatic cancer have been reported and are ready for further experiments in a wet lab.

- Some miRNA-mRNA interactions are specific of a type of cancer, while others can be shared across different digestive cancers. Common miRNA-mRNA interactions on colon, rectum, stomach, esophagus and liver cancers are related to basic cancer processes. Cancers from similar tissue origin share more miRNA-mRNA interactions.
- MiRNAs are key fine-tuning regulators of gene expression as they affect the expression of 50-75% of the total mRNAs, and a high proportion of them are regulated by more than 5 miRNAs.
- PDCD4 and BTG2 have been confirmed as *hsa-miR-21* targets in a pancreatic cancer model.

Bibliography

- [1] Lim LP, Lau NC, Weinstein EG, Abdelhakim A, Yekta S, Rhoades MW, Burge CB, Bartel DP. The microRNAs of *Caenorhabditis elegans*. *Genes & Development*. 2003 Apr;17(8):991–1008.
- [2] Bentwich I, Avniel A, Karov Y, Aharonov R, Gilad S, Barad O, Barzilai A, Einat P, Einav U, Meiri E, Sharon E, Spector Y, Bentwich Z. Identification of hundreds of conserved and non-conserved human microRNAs. *Nature Genetics*. 2005 Jul;37(7):766–770.
- [3] Esteller M. Epigenetics in evolution and disease. *The Lancet*. 2008 Dec;372:S90–S96.
- [4] Esteller M, Herman JG. Cancer as an epigenetic disease: DNA methylation and chromatin alterations in human tumours. *The Journal of Pathology*. 2002 Jan;196(1):1–7.
- [5] Peschansky VJ, Wahlestedt C. Non-coding RNAs as direct and indirect modulators of epigenetic regulation. *Epigenetics*. 2014 Jan;9(1):3–12.
- [6] Mattick JS, Rinn JL. Discovery and annotation of long noncoding RNAs. *Nature Structural & Molecular Biology*. 2015 Jan;22(1):5–7.
- [7] Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, Lagarde J, Veeravalli L, Ruan X, Ruan Y, Lassmann T, Carninci P, Brown JB, Lipovich L, Gonzalez JM, Thomas M, Davis CA, Shiekhhattar R, Gingeras TR, Hubbard TJ, Notredame C, Harrow J, Guigó R. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Research*. 2012 Jan;22(9):1775–1789.
- [8] Lasda E, Parker R. Circular RNAs: diversity of form and function. *RNA*. 2014 Jan;20(12):1829–1842.
- [9] Elbashir SM, Harborth J, Lendeckel W, Yalcin A, Weber K, Tuschl T. Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature*. 2001 May;411(6836):494–498.
- [10] Iwasaki YW, Siomi MC, Siomi H. PIWI-Interacting RNA: Its Biogenesis and Functions. *Annual Review of Biochemistry*. 2015;84(1):405–433.
- [11] Tollervy D, Kiss T. Function and synthesis of small nucleolar RNAs. *Current Opinion in Cell Biology*. 1997 Jun;9(3):337–342.

- [12] Crick FHC. The origin of the genetic code. *Journal of Molecular Biology*. 1968 Dec;38(3):367–379.
- [13] Brimacombe R, Stiege W. Structure and function of ribosomal RNA. *Biochemical Journal*. 1985 Jul;229(1):1–17.
- [14] Griffiths-Jones S, Grocock RJ, Dongen Sv, Bateman A, Enright AJ. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research*. 2006 Jan;34(suppl 1):D140–D144.
- [15] Guo H, Ingolia NT, Weissman JS, Bartel DP. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*. 2010 Aug;466(7308):835–840.
- [16] Schratz G. microRNAs at the synapse. *Nature Reviews Neuroscience*. 2009 Dec;10(12):842–849.
- [17] Esquela-Kerscher A, Slack FJ. Oncomirs — microRNAs with a role in cancer. *Nature Reviews Cancer*. 2006 Apr;6(4):259–269.
- [18] Meister G. Argonaute proteins: functional insights and emerging roles. *Nature Reviews Genetics*. 2013 Jul;14(7):447–459.
- [19] Bartel DP. MicroRNAs: target recognition and regulatory functions. *Cell*. 2009 Jan;136(2):215–233.
- [20] Nielsen CB, Shomron N, Sandberg R, Hornstein E, Kitzman J, Burge CB. Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. *RNA*. 2007 Jan;13(11):1894–1910.
- [21] Hutvagner G, Simard MJ. Argonaute proteins: key players in RNA silencing. *Nature Reviews Molecular Cell Biology*. 2008 Jan;9(1):22–32.
- [22] Friedman RC, Farh KKH, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*. 2009 Jan;19(1):92–105.
- [23] Bartel DP. MicroRNA Target Recognition and Regulatory Functions. *Cell*. 2009 Jan;136(2):215–233.
- [24] Mogilyansky E, Rigoutsos I. The miR-17/92 cluster: a comprehensive update on its genomics, genetics, functions and increasingly important and numerous roles in health and disease. *Cell Death and Differentiation*. 2013 Dec;20(12):1603–1614.
- [25] Hausser J, Zavolan M. Identification and consequences of miRNA-target interactions — beyond repression of gene expression. *Nature Reviews Genetics*. 2014 Sep;15(9):599–612.
- [26] Concepcion CP, Bonetti C, Ventura A. The miR-17-92 family of microRNA clusters in development and disease. *Cancer journal (Sudbury, Mass)*. 2012;18(3):262–267.
- [27] Mestdagh P, Boström AK, Impens F, Fredlund E, Peer GV, Antonellis PD, Stedingk Kv, Ghesquière B, Schulte S, Dewes M, Thomas-Tikhonenko A, Schulte JH, Zollo M, Schramm A, Gevaert K, Axelson H, Speleman F, Vandesompele J. The miR-17-92 MicroRNA Cluster Regulates Multiple Components of the TGF-beta Pathway in Neuroblastoma. *Molecular Cell*. 2010 Dec;40(5):762–773.

- [28] Umar A, Dunn BK, Greenwald P. Future directions in cancer prevention. *Nature Reviews Cancer*. 2012 Dec;12(12):835–848.
- [29] Sellers AH. The clinical classification of malignant tumours: the TNM system. *Canadian Medical Association Journal*. 1971 Oct;105(8):836–passim.
- [30] Renteln Dv, Bouin M, Barkun AN. Current standards and new developments of colorectal polyp management and resection techniques. *Expert Review of Gastroenterology & Hepatology*. 2017 Mar;0(ja):null.
- [31] Liang W, Shao W, Jiang G, Wang Q, Liu L, Liu D, Wang Z, Zhu Z, He J. Chinese multi-institutional registry (CMIR) for resected non-small cell lung cancer: survival analysis of 5,853 cases. *Journal of Thoracic Disease*. 2013 Dec;5(6):726–729.
- [32] Quaresma M, Coleman MP, Rachet B. 40-year trends in an index of survival for all cancers combined and survival adjusted for age and sex for each cancer in England and Wales, 1971–2011: a population-based study. *The Lancet*. 2015 Mar;385(9974):1206–1218.
- [33] Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation. *Cell*. 2011 Mar;144(5):646–674.
- [34] Hanahan D, Weinberg RA. The Hallmarks of Cancer. *Cell*. 2000 Jan;100(1):57–70.
- [35] Heldin CH. Targeting the PDGF signaling pathway in tumor treatment. *Cell Communication and Signaling : CCS*. 2013 Dec;11:97.
- [36] Goustin AS, Leof EB, Shipley GD, Moses HL. Growth Factors and Cancer. *Cancer Research*. 1986 Mar;46(3):1015–1029.
- [37] Witsch E, Sela M, Yarden Y. Roles for growth factors in cancer progression. *Physiology (Bethesda, Md)*. 2010 Apr;25(2):85–101.
- [38] Martinelli E, Morgillo F, Troiani T, Ciardiello F. Cancer resistance to therapies against the EGFR-RAS-RAF pathway: The role of MEK. *Cancer Treatment Reviews*. 2017 Feb;53:61–69.
- [39] Massagué J. TGFbeta in Cancer. *Cell*. 2008 Jul;134(2):215–230.
- [40] Grandori C, Cowley SM, James LP, Eisenman RN. The Myc/Max/Mad network and the transcriptional control of cell behavior. *Annual Review of Cell and Developmental Biology*. 2000;16:653–699.
- [41] Balkwill F. Tumour necrosis factor and cancer. *Nature Reviews Cancer*. 2009 May;9(5):361–371.
- [42] Vousden KH, Lu X. Live or let die: the cell's response to p53. *Nature Reviews Cancer*. 2002 Aug;2(8):594–604.
- [43] Liedtke C, Trautwein C. The role of TNF and Fas dependent signaling in animal models of inflammatory liver injury and liver cancer. *European Journal of Cell Biology*. 2012 Jul;91(6-7):582–589.
- [44] Philchenkov A, Zavelevich M, Krocak TJ, Los M. Caspases and cancer: mechanisms of inactivation and new treatment modalities. *Experimental Oncology*. 2004 Jun;26(2):82–97.

- [45] Bryan TM, Cech TR. Telomerase and the maintenance of chromosome ends. *Current Opinion in Cell Biology*. 1999 Jun;11(3):318–324.
- [46] Nishida N, Yano H, Nishida T, Kamura T, Kojiro M. Angiogenesis in Cancer. *Vascular Health and Risk Management*. 2006 Sep;2(3):213–219.
- [47] Carmeliet P, Jain RK. Angiogenesis in cancer and other diseases. *Nature*. 2000 Sep;407(6801):249–257.
- [48] Thiery JP. Epithelial–mesenchymal transitions in tumour progression. *Nature Reviews Cancer*. 2002 Jun;2(6):442–454.
- [49] Desgrosellier JS, Cheresch DA. Integrins in cancer: biological implications and therapeutic opportunities. *Nature Reviews Cancer*. 2010 Jan;10(1):9–22.
- [50] Peng Y, Croce CM. The role of MicroRNAs in human cancer. *Signal Transduction and Targeted Therapy*. 2016 Jan;1:15004.
- [51] Chen K, Rajewsky N. The evolution of gene regulation by transcription factors and microRNAs. *Nature Reviews Genetics*. 2007 Feb;8(2):93–103.
- [52] Maddika S, Ande SR, Panigrahi S, Paranjothy T, Weglarczyk K, Zuse A, Eshraghi M, Manda KD, Wiechec E, Los M. Cell survival, cell death and cell cycle pathways are interconnected: implications for cancer therapy. *Drug Resistance Updates: Reviews and Commentaries in Antimicrobial and Anticancer Chemotherapy*. 2007 Apr;10(1-2):13–29.
- [53] Sassen S, Miska EA, Caldas C. MicroRNA—implications for cancer. *Virchows Archiv*. 2008 Jan;452(1):1–10.
- [54] Croce CM. Causes and consequences of microRNA dysregulation in cancer. *Nature Reviews Genetics*. 2009 Oct;10(10):704–714.
- [55] Mao B, Zhang Z, Wang G. BTG2: A rising star of tumor suppressors (Review). *International Journal of Oncology*. 2015 Feb;46(2):459–464.
- [56] Feng YH, Tsao CJ. Emerging role of microRNA-21 in cancer. *Biomedical Reports*. 2016 Oct;5(4):395–402.
- [57] Pfeffer SR, Yang CH, Pfeffer LM. The Role of miR-21 in Cancer. *Drug Development Research*. 2015 Sep;76(6):270–277.
- [58] Hao J, Zhang Y, Deng M, Ye R, Zhao S, Wang Y, Li J, Zhao Z. MicroRNA control of epithelial-mesenchymal transition in cancer stem cells. *International Journal of Cancer*. 2014 Sep;135(5):1019–1027.
- [59] Ferlay J, Soerjomataram I, Ervik M, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin D, Forman D, Bray F. GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11.; 2013. <http://globocan.iarc.fr>.
- [60] IARC Publications IP. World Cancer Report 2014;. <https://shop.iarc.fr/products/wcr2014>.
- [61] Network TCGA. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*. 2012 Jul;487(7407):330–337.

- [62] Strubberg AM, Madison BB. MicroRNAs in the etiology of colorectal cancer: pathways and clinical implications. *Disease Models & Mechanisms*. 2017 Mar;10(3):197–214.
- [63] Jafri MA, Zaidi SK, Ansari SA, Al-Qahtani MH, Shay JW. MicroRNAs as potential drug targets for therapeutic intervention in colorectal cancer. *Expert Opinion on Therapeutic Targets*. 2015 Dec;19(12):1705–1723.
- [64] Sharma P, Sharma R. miRNA–mRNA crosstalk in esophageal cancer: From diagnosis to therapy. *Critical Reviews in Oncology / Hematology*. 2015 Dec;96(3):449–462.
- [65] WANG L, YUE Y, WANG X, JIN H. Function and clinical potential of microRNAs in hepatocellular carcinoma. *Oncology Letters*. 2015 Dec;10(6):3345–3353.
- [66] Hayes CN, Chayama K. MicroRNAs as Biomarkers for Liver Disease and Hepatocellular Carcinoma. *International Journal of Molecular Sciences*. 2016 Feb;17(3):280.
- [67] da Silva Oliveira KC, Thomaz Araújo TM, Albuquerque CI, Barata GA, Gígeç CO, Leal MF, Wisnieski F, Rodrigues Mello Junior FA, Khayat AS, de Assumpção PP, Rodriguez Burbano RM, Smith MC, Calcagno DQ. Role of miRNAs and their potential to be useful as diagnostic and prognostic biomarkers in gastric cancer. *World Journal of Gastroenterology*. 2016;22(35):7951.
- [68] Zhang Y, Li M, Wang H, Fisher WE, Lin PH, Yao Q, Chen C. Profiling of 95 MicroRNAs in Pancreatic Cancer Cell Lines and Surgical Specimens by Real-Time PCR Analysis. *World Journal of Surgery*. 2009 Apr;33(4):698.
- [69] Yu J, Li A, Hong SM, Hruban RH, Goggins M. MicroRNA Alterations of Pancreatic Intraepithelial Neoplasias. *Clinical Cancer Research*. 2012 Feb;18(4):981–992.
- [70] Hruban RH, Goggins M, Parsons J, Kern SE. Progression Model for Pancreatic Cancer. *Clinical Cancer Research*. 2000 Aug;6(8):2969–2972.
- [71] Huang J, Liu J, Chen-Xiao K, Zhang X, Lee WNP, Go VLW, Xiao GG. Advance in microRNA as a potential biomarker for early detection of pancreatic cancer. *Biomarker Research*. 2016;4:20.
- [72] Matsuzaki J, Suzuki H. Role of MicroRNAs-221/222 in Digestive Systems. *Journal of Clinical Medicine*. 2015 Aug;4(8):1566–1577.
- [73] Jamieson NB, Morran DC, Morton JP, Ali A, Dickson EJ, Carter CR, Sansom OJ, Evans TRJ, McKay CJ, Oien KA. MicroRNA Molecular Profiles Associated with Diagnosis, Clinicopathologic Criteria, and Overall Survival in Patients with Resectable Pancreatic Ductal Adenocarcinoma. *Clinical Cancer Research*. 2012 Jan;18(2):534–545.
- [74] Papaconstantinou IG, Manta A, Gazouli M, Lyberopoulou A, Lykoudis PM, Polymeneas G, Voros D. Expression of microRNAs in patients with pancreatic cancer and its prognostic significance. *Pancreas*. 2013 Jan;42(1):67–71.
- [75] Giovannetti E, Funel N, Peters GJ, Chiaro MD, Erozenski LA, Vasile E, Leon LG, Pollina LE, Groen A, Falcone A, Danesi R, Campani D, Verheul HM, Boggi U. MicroRNA-21 in Pancreatic Cancer: Correlation with Clinical Outcome and Pharmacologic Aspects Underlying Its Role in the Modulation of Gemcitabine Activity. *Cancer Research*. 2010 Jun;70(11):4528–4538.

- [76] Giovannetti E, Velde Avd, Funel N, Vasile E, Perrone V, Leon LG, Lio ND, Avan A, Caponi S, Pollina LE, Gallá V, Sudo H, Falcone A, Campani D, Boggi U, Peters GJ. High-Throughput MicroRNA (miRNAs) Arrays Unravel the Prognostic Role of MiR-211 in Pancreatic Cancer. *PLOS ONE*. 2012 Nov;7(11):e49145.
- [77] Vila-Navarro E, Vila-Casadesús M, Moreira L, Duran-Sancho S, Sinha R, Ginés A, Fernández-Esparrach G, Miquel R, Cuatrecasas M, Castells A, Lozano JJ, Gironella M. MicroRNAs for Detection of Pancreatic Neoplasia: Biomarker Discovery by Next-generation Sequencing and Validation in 2 Independent Cohorts. *Annals of Surgery*. 2016 May;p. 1.
- [78] Filipowicz W, Bhattacharyya SN, Sonenberg N. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nature Reviews Genetics*. 2008 Feb;9(2):102–114.
- [79] Brennecke J, Stark A, Russell RB, Cohen SM. Principles of MicroRNA–Target Recognition. *PLOS Biology*. 2005 Feb;3(3):e85.
- [80] Ellwanger DC, Büttner FA, Mewes HW, Stümpflen V. The sufficient minimal set of miRNA seed types. *Bioinformatics*. 2011 May;27(10):1346–1350.
- [81] Rehmsmeier M, Steffen P, Höchsmann M, Giegerich R. Fast and effective prediction of microRNA/target duplexes. *RNA*. 2004 Jan;10(10):1507–1517.
- [82] Walters RW, Bradrick SS, Gromeier M. Poly(A)-binding protein modulates mRNA susceptibility to cap-dependent miRNA-mediated repression. *RNA*. 2010 Jan;16(1):239–250.
- [83] Grimson A, Farh KKH, Johnston WK, Garrett-Engle P, Lim LP, Bartel DP. MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing. *Molecular Cell*. 2007 Jul;27(1):91–105.
- [84] Sethupathy P, Megraw M, Hatzigeorgiou AG. A guide through present computational approaches for the identification of mammalian microRNA targets. *Nature Methods*. 2006 Nov;3(11):881–886.
- [85] Bentwich I. Prediction and validation of microRNAs and their targets. *FEBS Letters*. 2005 Oct;579(26):5904–5910.
- [86] Thomson DW, Bracken CP, Goodall GJ. Experimental strategies for microRNA target identification. *Nucleic Acids Research*. 2011 Sep;39(16):6845–6853.
- [87] Betel D, Koppal A, Agius P, Sander C, Leslie C. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biology*. 2010 Aug;11(8):R90.
- [88] Gaidatzis D, van Nimwegen E, Hausser J, Zavolan M. Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC bioinformatics*. 2007 Mar;8:69.
- [89] Lewis BP, Burge CB, Bartel DP. Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets. *Cell*. 2005 Jan;120(1):15–20.
- [90] Wong N, Wang X. miRDB: an online resource for microRNA target prediction and functional annotations. *Nucleic Acids Research*. 2015 Jan;43(D1):D146–D152.

- [91] Dweep H, Sticht C, Pandey P, Gretz N. miRWalk – Database: Prediction of possible miRNA binding sites by “walking” the genes of three genomes. *Journal of Biomedical Informatics*. 2011 Oct;44(5):839–847.
- [92] Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T. miRecords: an integrated resource for microRNA–target interactions. *Nucleic Acids Research*. 2009 Jan;37(suppl 1):D105–D110.
- [93] Jiang Q, Wang Y, Hao Y, Juan L, Teng M, Zhang X, Li M, Wang G, Liu Y. miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Research*. 2009 Jan;37(suppl 1):D98–D104.
- [94] Vlachos IS, Paraskevopoulou MD, Karagkouni D, Georgakilas G, Vergoulis T, Kanellos I, Anastasopoulos IL, Maniou S, Karathanou K, Kalfakakou D, Fevgas A, Dalamagas T, Hatzigeorgiou AG. DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. *Nucleic Acids Research*. 2014 Nov;p. gku1215.
- [95] Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS. MicroRNA targets in *Drosophila*. *Genome Biology*. 2003;5:R1.
- [96] Griffiths-Jones S, Saini HK, Dongen Sv, Enright AJ. miRBase: tools for microRNA genomics. *Nucleic Acids Research*. 2008 Jan;36(suppl 1):D154–D158.
- [97] Agarwal V, Bell GW, Nam JW, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. *eLife*. 2015 Aug;4:e05005.
- [98] Arvey A, Larsson E, Sander C, Leslie CS, Marks DS. Target mRNA abundance dilutes microRNA and siRNA activity. *Molecular Systems Biology*. 2010 Apr;6:363.
- [99] Garcia DM, Baek D, Shin C, Bell GW, Grimson A, Bartel DP. Weak seed-pairing stability and high target-site abundance decrease the proficiency of *lscy-6* and other microRNAs. *Nature Structural & Molecular Biology*. 2011 Oct;18(10):1139–1146.
- [100] Majoros WH, Ohler U. Spatial preferences of microRNA targets in 3' untranslated regions. *BMC genomics*. 2007 Jun;8:152.
- [101] Hausser J, Landthaler M, Jaskiewicz L, Gaidatzis D, Zavolan M. Relative contribution of sequence and structure features to the mRNA binding of Argonaute/EIF2C-miRNA complexes and the degradation of miRNA targets. *Genome Research*. 2009 Nov;19(11):2009–2020.
- [102] Reczko M, Maragkakis M, Alexiou P, Grosse I, Hatzigeorgiou AG. Functional microRNA targets in protein coding sequences. *Bioinformatics (Oxford, England)*. 2012 Mar;28(6):771–776.
- [103] Linsley PS, Schelter J, Burchard J, Kibukawa M, Martin MM, Bartz SR, Johnson JM, Cummins JM, Raymond CK, Dai H, Chau N, Cleary M, Jackson AL, Carleton M, Lim L. Transcripts Targeted by the MicroRNA-16 Family Cooperatively Regulate Cell Cycle Progression. *Molecular and Cellular Biology*. 2007 Mar;27(6):2240–2252.
- [104] Wang X. Improving microRNA target prediction by modeling with unambiguously identified microRNA-target pairs from CLIP-ligation studies. 2016;

BIBLIOGRAPHY

- [105] Wang X, Naqa IME. Prediction of both conserved and nonconserved microRNA targets in animals. *Bioinformatics*. 2008 Jan;24(3):325–332.
- [106] Wang X, Wang X. Systematic identification of microRNA functions by combining target prediction and expression profiling. *Nucleic Acids Research*. 2006 Jan;34(5):1646–1652.
- [107] Dweep H, Gretz N. miRWalk2.0: a comprehensive atlas of microRNA-target interactions. *Nature Methods*. 2015 Aug;12(8):697–697.
- [108] Lee YJD, Kim V, Muth DC, Witwer KW. Validated MicroRNA Target Databases: An Evaluation. *Drug Development Research*. 2015 Nov;76(7):389–396.
- [109] Chou CH, Chang NW, Shrestha S, Hsu SD, Lin YL, Lee WH, Yang CD, Hong HC, Wei TY, Tu SJ, Tsai TR, Ho SY, Jian TY, Wu HY, Chen PR, Lin NC, Huang HT, Yang TL, Pai CY, Tai CS, Chen WL, Huang CY, Liu CC, Weng SL, Liao KW, Hsu WL, Huang HD. miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Research*. 2016 Jan;44(D1):D239–247.
- [110] Vergoulis T, Vlachos IS, Alexiou P, Georgakilas G, Maragkakis M, Reczko M, Gerangelos S, Koziris N, Dalamagas T, Hatzigeorgiou AG. TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. *Nucleic Acids Research*. 2012 Jan;40(D1):D222–D229.
- [111] Muniategui A, Pey J, Planes FJ, Rubio A. Joint analysis of miRNA and mRNA expression data. *Briefings in Bioinformatics*. 2013 May;14(3):263–278.
- [112] Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao Y, Carver BS, Arora VK, Kaushik P, Cerami E, Reva B, Antipin Y, Mitsiades N, Landers T, Dolgalev I, Major JE, Wilson M, Socci ND, Lash AE, Heguy A, Eastham JA, Scher HI, Reuter VE, Scardino PT, Sander C, Sawyers CL, Gerald WL. Integrative Genomic Profiling of Human Prostate Cancer. *Cancer Cell*. 2010 Jul;18(1):11–22.
- [113] Laczny C, Leidinger P, Haas J, Ludwig N, Backes C, Gerasch A, Kaufmann M, Vogel B, Katus HA, Meder B, Stähler C, Meese E, Lenhof HP, Keller A. miRTrail - a comprehensive web-server for analyzing gene and miRNA patterns to enhance the understanding of regulatory mechanisms in diseases. *BMC Bioinformatics*. 2012 Feb;13(1):36.
- [114] Ingenuity I. Ingenuity IPA - Integrate and understand complex 'omics data;. <http://www.ingenuity.com/products/ipa>.
- [115] Li X, Jiang C, Wu X, Sun Y, Bu J, Li J, Xiao M, Zheng Y, Zhang J. A systems biology approach to study the biology characteristics of esophageal squamous cell carcinoma by integrating microRNA and messenger RNA expression profiling. *Cell Biochemistry and Biophysics*. 2014 Nov;70(2):1369–1376.
- [116] Hamed M, Spaniol C, Zapp A, Helms V. Integrative network-based approach identifies key genetic elements in breast invasive carcinoma. *BMC genomics*. 2015;16 Suppl 5:S2.
- [117] Freitas RCCd, Bortolin RH, Lopes MB, Hirata MH, Hirata RDC, Silbiger VN, Luchessi AD. Integrated analysis of miRNA and mRNA gene expression microarrays: Influence on platelet reactivity, clopidogrel response and drug-induced toxicity. *Gene*. 2016 Nov;593(1):172–178.

- [118] Shen L, Lin Y, Sun Z, Yuan X, Chen L, Shen B. Knowledge-Guided Bioinformatics Model for Identifying Autism Spectrum Disorder Diagnostic MicroRNA Biomarkers. *Scientific Reports*. 2016 Dec;6:39663.
- [119] Wang F, Lu J, Peng X, Wang J, Liu X, Chen X, Jiang Y, Li X, Zhang B. Integrated analysis of microRNA regulatory network in nasopharyngeal carcinoma with deep sequencing. *Journal of Experimental & Clinical Cancer Research*. 2016;35:17.
- [120] Gade S, Porzelius C, Fälth M, Brase JC, Wuttig D, Kuner R, Binder H, Sülthmann H, Beißbarth T. Graph based fusion of miRNA and mRNA expression data improves clinical outcome prediction in prostate cancer. *BMC Bioinformatics*. 2011 Dec;12(1):488.
- [121] Stouffer SA, Suchman EA, DeVinney LC, Star SA, Williams RM. The american soldier: Adjustment during army life. volume i. *Journal of the American Medical Association*. 1949 Aug;140(14):1189–1189.
- [122] Peng X, Li Y, Walters KA, Rosenzweig ER, Lederer SL, Aicher LD, Proll S, Katze MG. Computational identification of hepatitis C virus associated microRNA-mRNA regulatory modules in human livers. *BMC Genomics*. 2009 Aug;10(1):373.
- [123] Sales G, Coppe A, Bisognin A, Biasiolo M, Bortoluzzi S, Romualdi C. MAGIA, a web-based tool for miRNA and Genes Integrated Analysis. *Nucleic Acids Research*. 2010 Jan;38(suppl 2):W352–W359.
- [124] Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*. 2008 Dec;4(1):44–57.
- [125] Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*. 2009 Jan;37(1):1–13.
- [126] Bisognin A, Sales G, Coppe A, Bortoluzzi S, Romualdi C. MAGIA2: from miRNA and genes expression data integrative analysis to microRNA-transcription factor mixed regulatory circuits (2012 update). *Nucleic Acids Research*. 2012 Jul;40(Web Server issue):W13–21.
- [127] Muniategui A, Nogales-Cadenas R, Vázquez M, L Aranguren X, Agirre X, Luttun A, Prosper F, Pascual-Montano A, Rubio A. Quantification of miRNA-mRNA Interactions. *PLoS ONE*. 2012 Feb;7(2):e30766.
- [128] Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*. 2004 Sep;5(10):R80.
- [129] Favero F. RmiR: Package to work with miRNAs and miRNA targets with R. R package version 1.30.0; 2016. <https://www.bioconductor.org/packages/release/bioc/html/RmiR.html>.

- [130] Zhao Y, Ransom JF, Li A, Vedantham V, Drehle Mv, Muth AN, Tsuchihashi T, McManus MT, Schwartz RJ, Srivastava D. Dysregulation of Cardiogenesis, Cardiac Conduction, and Cell Cycle in Mice Lacking miRNA-1-2. *Cell*. 2007 Apr;129(2):303–317.
- [131] Wu X, Watson M. CORNA: testing gene lists for regulation by microRNAs. *Bioinformatics*. 2009 Mar;25(6):832–833.
- [132] Cogswell JP, Ward J, Taylor IA, Waters M, Shi Y, Cannon B, Kelnar K, Kemppainen J, Brown D, Chen C, Prinjha RK, Richardson JC, Saunders AM, Roses AD, Richards CA. Identification of miRNA Changes in Alzheimer’s Disease Brain and CSF Yields Putative Biomarkers and Insights into Disease Pathways. *Journal of Alzheimer’s Disease*. 2008 Jan;14(1):27–41.
- [133] Gentleman R, Falcon S. microRNA: Data and functions for dealing with microRNAs. R package version 1.32.0; 2016. <https://www.bioconductor.org/packages/release/bioc/html/microRNA.html>.
- [134] Haunsberger SJ, Connolly NMC, Prehn JHM. miRNANameConverter: an R/Bioconductor package for translating mature miRNA names to different miRBase versions. *Bioinformatics*. 2016 Oct;p. btw660.
- [135] Chiromatzo AO, Oliveira TYK, Pereira G, Costa AY, Montesco CaE, Gras DE, Yosetake F, Vilar JB, Cervato M, Prado PRR, Cardenas RGCCL, Cerri R, Borges RL, Lemos RN, Alvarenga SM, Perallis VRC, Pinheiro DG, Silva IT, Brandão RM, Cunha MaV, Giuliatti S, Silva WA. miRNA-path: a database of miRNAs, target genes and metabolic pathways. *Genetics and molecular research: GMR*. 2007 Oct;6(4):859–865.
- [136] Hudson TJ, Anderson W, Aretz A, Barker AD, Bell C, Bernabé RR, Bhan MK, Calvo F, Eerola I, Gerhard DS, Guttmacher A, Guyer M, Hemsley FM, Jennings JL, Kerr D, Klatt P, Kolar P, Kusuda J, Lane DP, Laplace F, Lu Y, Nettekoven G, Ozenberger B, Peterson J, Rao TS, Remacle J, Schafer AJ, Shibata T, Stratton MR, Vockley JG, Watanabe K, Yang H, Yuen MME, (Leader) BMK, Bobrow M, Cambon-Thomsen A, Dressler LG, Dyke SOM, Joly Y, Kato K, Kennedy KL, Nicolás P, Parker MJ, Rial-Sebbag E, Romeo-Casabona CM, Shaw KM, Wallace S, Wiesner GL, Zeps N, (Leader) PL, et al. International network of cancer genome projects. *Nature*. 2010 Apr;464(7291):993–998.
- [137] Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*. 2014 Feb;15(2):R29.
- [138] Smyth GK. limma: Linear Models for Microarray Data. In: Gentleman R, Carey VJ, Huber W, Irizarry RA, Dudoit S, editors. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Statistics for Biology and Health. Springer New York; 2005. p. 397–420. http://link.springer.com/chapter/10.1007/0-387-29362-0_23.
- [139] Center MAC. TCGA Batch Effects Tool; <http://bioinformatics.mdanderson.org/tcgambatch/>.
- [140] Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007 Jan;8(1):118–127.

- [141] Leek JT, Johnson WE, Parker HS, Fertig EJ, Jaffe AE, Storey JD. *sva*: Surrogate Variable Analysis. R package version 3.22.0. Bioconductor;.
- [142] Bioinformatics B. FastQC: a quality control tool for high throughput sequence data. Version 0.11.5. Babraham Institute; 2016. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
- [143] Lab H. FASTX Toolkit. Hannon Lab; http://hannonlab.cshl.edu/fastx_toolkit/index.html.
- [144] Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Research*. 2012 Jan;40(1):37–52.
- [145] Zhang W, Edwards A, Fan W, Flemington EK, Zhang K. miRNA-mRNA Correlation-Network Modules in Human Prostate Cancer and the Differences between Primary and Metastatic Tumor Subtypes. *PLoS ONE*. 2012 Jun;7(6):e40130.
- [146] Hong F, Breitling R, McEntee CW, Wittner BS, Nemhauser JL, Chory J. RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*. 2006 Nov;22(22):2825–2827.
- [147] Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biology*. 2010 Oct;11(10):R106.
- [148] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010 Jan;26(1):139–140.
- [149] Jeanmougin M, Reynies Ad, Marisa L, Paccard C, Nuel G, Guedj M. Should We Abandon the t-Test in the Analysis of Gene Expression Microarray Data: A Comparison of Variance Modeling Strategies. *PLOS ONE*. 2010 Sep;5(9):e12336.
- [150] Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*. 2009;10:R25.
- [151] Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*. 2012 Mar;7(3):562–578.
- [152] Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*. 2009 Jul;25(14):1754–1760.
- [153] Donahue TR, Tran LM, Hill R, Li Y, Kovochich A, Calvopina JH, Patel SG, Wu N, Hindoyan A, Farrell JJ, Li X, Dawson DW, Wu H. Integrative Survival-Based Molecular Profiling of Human Pancreatic Cancer. *Clinical Cancer Research*. 2012 Jan;18(5):1352–1363.
- [154] The Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*. 2013 Oct;45(10):1113–1120.

- [155] Bailey P, Chang DK, Nones K, Johns AL, Patch AM, Gingras MC, Miller DK, Christ AN, Bruxner TJC, Quinn MC, Nourse C, Murtaugh LC, Harliwong I, Idrisoglu S, Manning S, Nourbakhsh E, Wani S, Fink L, Holmes O, Chin V, Anderson MJ, Kazakoff S, Leonard C, Newell F, Waddell N, Wood S, Xu Q, Wilson PJ, Cloonan N, Kassahn KS, Taylor D, Quek K, Robertson A, Pantano L, Mincarelli L, Sanchez LN, Evers L, Wu J, Pinese M, Cowley MJ, Jones MD, Colvin EK, Nagrial AM, Humphrey ES, Chantrill LA, Mawson A, Humphris J, Chou A, Pajic M, Scarlett CJ, et al. Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature*. 2016 Mar;531(7592):47–52.
- [156] Kvam VM, Liu P, Si Y. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *American Journal of Botany*. 2012 Feb;99(2):248–256.
- [157] Srivastava S, Chen L. A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Research*. 2010 Jan;38(17):e170–e170.
- [158] Auer PL, Doerge RW. A Two-Stage Poisson Model for Testing RNA-Seq Data. *Statistical Applications in Genetics and Molecular Biology*. 2011;10(1).
- [159] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*. 2015 Jan;p. gkv007.
- [160] Del Carratore F, Janckevis A, Hong F, Wittner BS, Breitling R, Battke F. Rank Product method for identifying differentially expressed genes with application in meta-analysis. 2016;.
- [161] Hardcastle T. baySeq: Empirical Bayesian analysis of patterns of differential expression in count data. R package version 2.8.0. R/Bioconductor; 2012. <https://www.bioconductor.org/packages/release/bioc/html/baySeq.html>.
- [162] Tarazona S, Furió-Tarí P, Turrà D, Pietro AD, Nueda MJ, Ferrer A, Conesa A. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Research*. 2015 Dec;43(21):e140–e140.
- [163] Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, Mason CE, Socci ND, Betel D. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biology*. 2013;14(9):R95.
- [164] Seyednasrollah F, Laiho A, Elo LL. Comparison of software packages for detecting differential expression in RNA-seq studies. *Briefings in Bioinformatics*. 2015 Jan;16(1):59–70.
- [165] Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szczeniński MW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A. A survey of best practices for RNA-seq data analysis. *Genome Biology*. 2016;17:13.
- [166] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*. 2014;15:550.
- [167] McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*. 2012 May;40(10):4288–4297.

- [168] Anders S, McCarthy DJ, Chen Y, Okoniewski M, Smyth GK, Huber W, Robinson MD. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nature Protocols*. 2013 Sep;8(9):1765–1786.
- [169] Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software*. 2010;33(1):1–22.
- [170] Zaykin DV, Zhivotovsky LA, Westfall PH, Weir BS. Truncated product method for combining P-values. *Genetic Epidemiology*. 2002 Feb;22(2):170–185.
- [171] Fisher R. *Statistical methods for research workers*; 1925. <http://trove.nla.gov.au/version/15607207>.
- [172] Lipták P. On the combination of independent tests. *Magyar Tud Akad Mat Kutato Int Kozl*. 1958;3:171–197.
- [173] Zaykin DV. Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. *Journal of Evolutionary Biology*. 2011 Aug;24(8):1836–1841.
- [174] Bonferroni CE. *Teoria statistica delle classi e calcolo delle probabilità*. Libreria internazionale Seeber; 1936.
- [175] Dunn OJ, Clark VA. *Basic Statistics: A Primer for the Biomedical Sciences*. Wiley; 1965. <http://www.wiley.com/WileyCDA/WileyTitle/productCd-0470248793.html>.
- [176] Perneger TV. What's wrong with Bonferroni adjustments. *BMJ*. 1998 Apr;316(7139):1236–1238.
- [177] Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B (Methodological)*. 1995 Jan;57(1):289–300.
- [178] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*. 2000 May;25(1):25–29.
- [179] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*. 2000 Jan;28(1):27–30.
- [180] Yates F. Contingency Tables Involving Small Numbers and the X² Test. Supplement to the *Journal of the Royal Statistical Society*. 1934;1(2):217–235.
- [181] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*. 2005 Oct;102(43):15545–15550.
- [182] Sergushichev A. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *bioRxiv*. 2016 Jun;p. 060012.

- [183] Project L. LaTeX - A document preparation system;. LaTeX is based on Donald E. Knuth's TeX typesetting language or certain extensions. LaTeX was first developed in 1985 by Leslie Lamport, and is now being maintained and developed by the LaTeX3 Project. www.latex-project.org.
- [184] Leisch F. Sweave: Dynamic Generation of Statistical Reports Using Literate Data Analysis. In: Härdle PDW, Rönz PDB, editors. *Compstat*. Physica-Verlag HD; 2002. p. 575–580. DOI: 10.1007/978-3-642-57489-4_89. http://link.springer.com/chapter/10.1007/978-3-642-57489-4_89.
- [185] Coll M, Taghdouini AE, Perea L, Mannaerts I, Vila-Casadesús M, Blaya D, Rodrigo-Torres D, Affò S, Morales-Ibanez O, Graupera I, Lozano JJ, Najimi M, Sokal E, Lambrecht J, Ginès P, van Grunsven LA, Sancho-Bru P Integrative miRNA and Gene Expression Profiling Analysis of Human Quiescent Hepatic Stellate Cells. *Scientific Reports*. 2015 Jun;5:11549.
- [186] Bofill-De Ros X, Santos M, Vila-Casadesús M, Villanueva E, Andreu N, Dierssen M, Fillat C. Genome-wide miR-155 and miR-802 target gene identification in the hippocampus of Ts65Dn Down syndrome mouse model by miRNA sponges. *BMC Genomics*. 2015;16:907.
- [187] Vila-Casadesús M, Gironella M, Lozano JJ. MiRComb: An R Package to Analyse miRNA-mRNA Interactions. Examples across Five Digestive Cancers. *PLOS ONE*. 2016 Mar;11(3):e0151127.
- [188] Blaya D, Coll M, Rodrigo-Torres D, Vila-Casadesús M, Altamirano J, Llopis M, Graupera I, Perea L, Aguilar-Bravo B, Díaz A, Banales JM, Clària J, Lozano JJ, Bataller R, Caballería J, Ginès P, Sancho-Bru P Integrative microRNA profiling in alcoholic hepatitis reveals a role for microRNA-182 in liver injury and inflammation. *Gut*. 2016 May;p. [gutjnl-2015-311314](https://doi.org/10.1136/gutjnl-2015-311314).
- [189] Xiong D, Pan J, Zhang Q, Szabo E, Miller MS, Lubet RA, You M, Wang Y, Xiong D, Pan J, Zhang Q, Szabo E, Steven Miller M, Lubet RA, You M, Wang Y. Bronchial airway gene expression signatures in mouse lung squamous cell carcinoma and their modulation by cancer chemopreventive agents. *Oncotarget*. 2016 Dec;5(0).
- [190] Software R. *Web Analytics in Real Time | Clicky*; 2016. <https://clicky.com/>.
- [191] Widenius M, Axmark D. *MySQL Reference Manual*. 1st ed. DuBois P editor. Sebastopol, CA, USA: O'Reilly & Associates, Inc.; 2002.
- [192] Krzywinski MI, Schein JE, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. Circos: An information aesthetic for comparative genomics. *Genome Research*. 2009 Jun;.
- [193] Gu Z, Gu L, Eils R, Schlesner M, Brors B. circlize implements and enhances circular visualization in R. *Bioinformatics*. 2014 Oct;30(19):2811–2812.
- [194] Bandiera S, Pfeffer S, Baumert TF, Zeisel MB. miR-122 – A key factor and therapeutic target in liver disease. *Journal of Hepatology*. 2015 Feb;62(2):448–457.
- [195] Petrocca F, Vecchione A, Croce CM. Emerging role of miR-106b-25/miR-17-92 clusters in the control of transforming growth factor beta signaling. *Cancer Research*. 2008 Oct;68(20):8191–8194.

- [196] Li Y, Deng X, Zeng X, Peng X. The Role of Mir-148a in Cancer. *Journal of Cancer*. 2016;7(10):1233–1241.
- [197] Li C, Lyu J, Meng QH. MiR-93 Promotes Tumorigenesis and Metastasis of Non-Small Cell Lung Cancer Cells by Activating the PI3K/Akt Pathway via Inhibition of LKB1/PTEN/CDKN1A. *Journal of Cancer*. 2017;8(5):870–879.
- [198] Artavanis-Tsakonas S, Rand MD, Lake RJ. Notch signaling: cell fate control and signal integration in development. *Science (New York, NY)*. 1999 Apr;284(5415):770–776.
- [199] Gao J, Long B, Wang Z. Role of Notch signaling pathway in pancreatic cancer. *American Journal of Cancer Research*. 2017;7(2):173–186.
- [200] Aster JC, Pear WS, Blacklow SC. The Varied Roles of Notch in Cancer. *Annual Review of Pathology: Mechanisms of Disease*. 2017;12(1):245–275.
- [201] Zhu W, Xu B. MicroRNA-21 identified as predictor of cancer outcome: a meta-analysis. *PloS One*. 2014;9(8):e103373.
- [202] Dillhoff M, Liu J, Frankel W, Croce C, Bloomston M. MicroRNA-21 is overexpressed in pancreatic cancer and a potential predictor of survival. *Journal of Gastrointestinal Surgery: Official Journal of the Society for Surgery of the Alimentary Tract*. 2008 Dec;12(12):2171–2176.
- [203] Moriyama T, Ohuchida K, Mizumoto K, Yu J, Sato N, Nabae T, Takahata S, Toma H, Nagai E, Tanaka M. MicroRNA-21 modulates biological functions of pancreatic cancer cells including their proliferation, invasion, and chemoresistance. *Molecular Cancer Therapeutics*. 2009 May;8(5):1067–1074.
- [204] Asangani IA, Rasheed SaK, Nikolova DA, Leupold JH, Colburn NH, Post S, Allgayer H. MicroRNA-21 (miR-21) post-transcriptionally downregulates tumor suppressor Pcd4 and stimulates invasion, intravasation and metastasis in colorectal cancer. *Oncogene*. 2008 Apr;27(15):2128–2136.
- [205] Peacock O, Lee AC, Cameron F, Tarbox R, Vafadar-Isfahani N, Tufarelli C, Lund JN. Inflammation and MiR-21 Pathways Functionally Interact to Downregulate PDCD4 in Colorectal Cancer. *PLOS ONE*. 2014 Oct;9(10):e110267.
- [206] Gu L, Song G, Chen L, Nie Z, He B, Pan Y, Xu Y, Li R, Gao T, Cho WC, Wang S. Inhibition of miR-21 induces biological and behavioral alterations in diffuse large B-cell lymphoma. *Acta Haematologica*. 2013;130(2):87–94.
- [207] Nagao Y, Hisaoka M, Matsuyama A, Kanemitsu S, Hamada T, Fukuyama T, Nakano R, Uchiyama A, Kawamoto M, Yamaguchi K, Hashimoto H. Association of microRNA-21 expression with its targets, PDCD4 and TIMP3, in pancreatic ductal adenocarcinoma. *Modern Pathology: An Official Journal of the United States and Canadian Academy of Pathology, Inc*. 2012 Jan;25(1):112–121.
- [208] Roldo C, Missiaglia E, Hagan JP, Falconi M, Capelli P, Bersani S, Calin GA, Volinia S, Liu CG, Scarpa A, Croce CM. MicroRNA expression abnormalities in pancreatic endocrine and acinar tumors are associated with distinctive pathologic features and clinical behavior. *Journal*

- of Clinical Oncology: Official Journal of the American Society of Clinical Oncology. 2006 Oct;24(29):4677–4684.
- [209] Leone E, Morelli E, Di Martino MT, Amodio N, Foresta U, Gullà A, Rossi M, Neri A, Giordano A, Munshi NC, Anderson KC, Tagliaferri P, Tassone P. Targeting miR-21 inhibits in vitro and in vivo multiple myeloma cell growth. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*. 2013 Apr;19(8):2096–2106.
- [210] Mao B, Xiao H, Zhang Z, Wang D, Wang G. MicroRNA-21 regulates the expression of BTG2 in HepG2 liver cancer cells. *Molecular Medicine Reports*. 2015 Oct;12(4):4917–4924.
- [211] Coppola V, Musumeci M, Patrizii M, Cannistraci A, Addario A, Maugeri-Saccà M, Biffoni M, Francescangeli F, Cordenonsi M, Piccolo S, Memeo L, Pagliuca A, Muto G, Zeuner A, De Maria R, Bonci D. Btg2 loss and miR-21 upregulation contribute to prostate cell transformation by inducing luminal markers expression and epithelial-mesenchymal transition. *Oncogene*. 2013 Apr;32(14):1843–1853.
- [212] Itersen Mv, Bervoets S, Meijer EJD, Buermans HP, Hoen PAC, Menezes RX, Boer JM. Integrated analysis of microRNA and mRNA expression: adding biological significance to microRNA target predictions. *Nucleic Acids Research*. 2013 Jun;p. gkt525.
- [213] Kwon MS, Kim Y, Lee S, Namkung J, Yun T, Yi SG, Han S, Kang M, Kim SW, Jang JY, Park T. Integrative analysis of multi-omics data for identifying multi-markers for diagnosing pancreatic cancer. *BMC Genomics*. 2015;16(9):1–10.
- [214] Tang Z, Yang Y, Wang Z, Zhao S, Mu Y, Li K. Integrated analysis of miRNA and mRNA paired expression profiling of prenatal skeletal muscle development in three genotype pigs. *Scientific Reports*. 2015 Oct;5.
- [215] Quitadamo A, Tian L, Hall B, Shi X. An integrated network of microRNA and gene expression in ovarian cancer. *BMC Bioinformatics*. 2015;16(5):S5.
- [216] Zhuang X, Li Z, Lin H, Gu L, Lin Q, Lu Z, Tzeng CM. Integrated miRNA and mRNA expression profiling to identify mRNA targets of dysregulated miRNAs in non-obstructive azoospermia. *Scientific Reports*. 2015 Jan;5.
- [217] Seo J, Jin D, Choi CH, Lee H. Integration of MicroRNA, mRNA, and Protein Expression Data for the Identification of Cancer-Related MicroRNAs. *PLOS ONE*. 2017 Jan;12(1):e0168412.
- [218] GitHub. GitHub Guide; 2017. <https://guides.github.com/activities/hello-world/>.
- [219] Vila-Casadesús M. Analysis of miRNA-mRNA interactions in alcoholic hepatitis. *UPC Commons*. 2014 Jan;.
- [220] Laxman N, Rubin CJ, Mallmin H, Nilsson O, Pastinen T, Grundberg E, Kindmark A. Global miRNA expression and correlation with mRNA levels in primary human bone cells. *RNA (New York, NY)*. 2015 Aug;21(8):1433–1443.
- [221] Wang YP, Li KB. Correlation of expression profiles between microRNAs and mRNA targets using NCI-60 data. *BMC Genomics*. 2009 May;10(1):218.

- [222] Giles CB, Girija-Devi R, Dozmorov MG, Wren JD. mirCoX: a database of miRNA-mRNA expression correlations derived from RNA-seq meta-analysis. *BMC Bioinformatics*. 2013 Oct;14(Suppl 14):S17.
- [223] Leone P, Shin EC, Perosa F, Vacca A, Dammacco F, Racanelli V. MHC class I antigen processing and presenting machinery: organization, function, and defects in tumor cells. *Journal of the National Cancer Institute*. 2013 Aug;105(16):1172–1187.
- [224] Thibodeau J, Bourgeois-Daigneault MC, Lapointe R. Targeting the MHC Class II antigen presentation pathway in cancer immunotherapy. *Oncoimmunology*. 2012 Sep;1(6):908–916.
- [225] Zhang W, Liu HT. MAPK signal pathways in the regulation of cell proliferation in mammalian cells. *Cell Research*. 2002;12(1):9–18.
- [226] Santarpia L, Lippman SM, El-Naggar AK. Targeting the MAPK-RAS-RAF signaling pathway in cancer therapy. *Expert Opinion on Therapeutic Targets*. 2012 Jan;16(1):103–119.
- [227] Zhang R, Li M, Zang W, Chen X, Wang Y, Li P, Du Y, Zhao G, Li L. MiR-148a regulates the growth and apoptosis in pancreatic cancer by targeting CCKBR and Bcl-2. *Tumour Biology: The Journal of the International Society for Oncodevelopmental Biology and Medicine*. 2014 Jan;35(1):837–844.
- [228] Feng H, Wang Y, Su J, Liang H, Zhang CY, Chen X, Yao W. MicroRNA-148a Suppresses the Proliferation and Migration of Pancreatic Cancer Cells by Down-regulating ErbB3. *Pancreas*. 2016 Oct;45(9):1263–1271.
- [229] Huntzinger E, Izaurralde E. Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nature Reviews Genetics*. 2011 Feb;12(2):99–110.
- [230] Sevignani C, Calin GA, Siracusa LD, Croce CM. Mammalian microRNAs: a small world for fine-tuning gene expression. *Mammalian Genome*. 2006 Mar;17(3):189–202.
- [231] Wu CT, Chiou CY, Chiu HC, Yang UC. Fine-tuning of microRNA-mediated repression of mRNA by splicing-regulated and highly repressive microRNA recognition element. *BMC Genomics*. 2013;14:438.
- [232] Lai X, Wolkenhauer O, Vera J. Understanding microRNA-mediated gene regulatory networks through mathematical modelling. *Nucleic Acids Research*. 2016 Jul;44(13):6019–6035.
- [233] Liu R, Chen X, Du Y, Yao W, Shen L, Wang C, Hu Z, Zhuang R, Ning G, Zhang C, Yuan Y, Li Z, Zen K, Ba Y, Zhang CY. Serum MicroRNA Expression Profile as a Biomarker in the Diagnosis and Prognosis of Pancreatic Cancer. *Clinical Chemistry*. 2012 Mar;58(3):610–618.
- [234] Hanoun N, Delpu Y, Suriawinata AA, Bournet B, Bureau C, Selves J, Tsongalis GJ, Dufresne M, Buscail L, Cordelier P, Torrisani J. The Silencing of MicroRNA 148a Production by DNA Hypermethylation Is an Early Event in Pancreatic Carcinogenesis. *Clinical Chemistry*. 2010 Jul;56(7):1107–1118.
- [235] Bofill-De Ros X, Gironella M, Fillat C. miR-148a- and miR-216a-regulated oncolytic adenoviruses targeting pancreatic tumors attenuate tissue damage without perturbation of

- miRNA activity. *Molecular Therapy: The Journal of the American Society of Gene Therapy*. 2014 Sep;22(9):1665–1677.
- [236] Tréhoux S, Lahdaoui F, Delpu Y, Renaud F, Leteurtre E, Torrisani J, Jonckheere N, Van Seuning I. Micro-RNAs miR-29a and miR-330-5p function as tumor suppressors by targeting the MUC1 mucin in pancreatic cancer cells. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*. 2015 Oct;1853(10, Part A):2392–2403.
- [237] Nakata K, Ohuchida K, Mizumoto K, Aishima S, Oda Y, Nagai E, Tanaka M. Micro RNA-373 is down-regulated in pancreatic cancer and inhibits cancer cell invasion. *Annals of Surgical Oncology*. 2014 Dec;21 Suppl 4:S564–574.
- [238] Liffers ST, Munding JB, Vogt M, Kuhlmann JD, Verdoodt B, Nambiar S, Maghnouj A, Mir-mohammadsadegh A, Hahn SA, Tannapfel A. MicroRNA-148a is down-regulated in human pancreatic ductal adenocarcinomas and regulates cell survival by targeting CDC25B. *Laboratory Investigation*. 2011 Oct;91(10):1472–1479.
- [239] Ranganathan P, Weaver KL, Capobianco AJ. Notch signalling in solid tumours: a little bit of everything but not all the time. *Nature Reviews Cancer*. 2011 May;11(5):338–351.
- [240] Tremblay I, Paré E, Arseneault D, Douziech M, Boucher MJ. The MEK/ERK Pathway Promotes NOTCH Signalling in Pancreatic Cancer Cells. *PLOS ONE*. 2013 Dec;8(12):e85502.
- [241] Hingorani SR, Petricoin EF, Maitra A, Rajapakse V, King C, Jacobetz MA, Ross S, Conrads TP, Veenstra TD, Hitt BA, Kawaguchi Y, Johann D, Liotta LA, Crawford HC, Putt ME, Jacks T, Wright CVE, Hruban RH, Lowy AM, Tuveson DA. Preinvasive and invasive ductal pancreatic cancer and its early detection in the mouse. *Cancer Cell*. 2003 Dec;4(6):437–450.
- [242] Miyamoto Y, Maitra A, Ghosh B, Zechner U, Argani P, Iacobuzio-Donahue CA, Sriuranpong V, Iso T, Meszoely IM, Wolfe MS, Hruban RH, Ball DW, Schmid RM, Leach SD. Notch mediates TGF α -induced changes in epithelial differentiation during pancreatic tumorigenesis. *Cancer Cell*. 2003 Jun;3(6):565–576.
- [243] Magliano MPd, Sekine S, Ermilov A, Ferris J, Dlugosz AA, Hebrok M. Hedgehog/Ras interactions regulate early stages of pancreatic cancer. *Genes & Development*. 2006 Nov;20(22):3161–3173.
- [244] Stanger BZ, Stiles B, Lauwers GY, Bardeesy N, Mendoza M, Wang Y, Greenwood A, Cheng Kh, McLaughlin M, Brown D, DePinho RA, Wu H, Melton DA, Dor Y. Pten constrains centroacinar cell expansion and malignant transformation in the pancreas. *Cancer Cell*. 2005 Sep;8(3):185–195.
- [245] Jung KH, Zhang J, Zhou C, Shen H, Gagea M, Rodriguez-Aguayo C, Lopez-Berestein G, Sood AK, Beretta L. Differentiation therapy for hepatocellular carcinoma: Multifaceted effects of miR-148a on tumor growth and phenotype and liver fibrosis. *Hepatology (Baltimore, Md)*. 2016 Mar;63(3):864–879.
- [246] Tan ZJ, Hu XG, Cao GS, Tang Y. Analysis of gene expression profile of pancreatic carcinoma using cDNA microarray. *World Journal of Gastroenterology : WJG*. 2003 Apr;9(4):818–823.

- [247] Jin G, Hu XG, Ying K, Tang Y, Liu R, Zhang YJ, Jing ZP, Xie Y, Mao YM. Discovery and analysis of pancreatic adenocarcinoma genes using cDNA microarrays. *World Journal of Gastroenterology*. 2005 Nov;11(41):6543–6548.

Appendices

Appendix A

Reports of STUDIES 1 and 2

A.1 Study 1: Colon adenocarcinoma

Default miRComb output

/home/nvilia/Baixades/TCGA/colon

May 13, 2015

1 Exploratory analysis of miRNA dataset

Number of miRNAs analysed 325
 Number of samples 444

Table 1: Basic information of the miRNA dataset.

group.n	CvH	center	sample	batch
1	NT: 8	Min: -0.000	AA :160	TCGA-5M-AAAT4-01: 1
		1st Qu.:1.000	A6 : 57	TCGA-5M-AAAT5-01: 1
		Median :1.000	CM : 37	TCGA-5M-AAAT6-01: 1
		Mean :0.982	D5 : 31	TCGA-5M-AAAT7-01: 1
		3rd Qu.:1.000	G4 : 27	TCGA-5M-AAAT8-01: 1
		Max. :1.000	DM : 25	TCGA-A6-2670-01: 1
			(Other):107	(Other) :438
				(Other) :189

Table 2: Summary of the phenotypical information of the miRNA dataset.

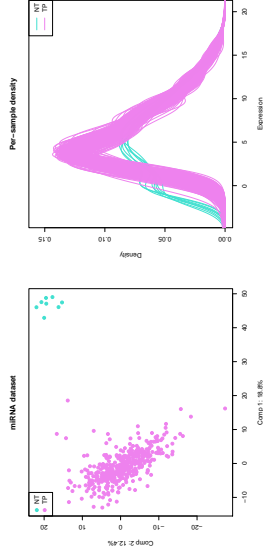


Figure 1: PCA and density plot for miRNAs.

2 Exploratory analysis of mRNA dataset

Number of mRNAs analysed 14860
 Number of samples 444

Table 3: Basic information of the mRNA dataset.

group.n	CvH	center	sample	batch
1	NT: 8	Min: -0.000	AA :160	TCGA-5M-AAAT4-01: 1
		1st Qu.:1.000	A6 : 57	TCGA-5M-AAAT5-01: 1
		Median :1.000	CM : 37	TCGA-5M-AAAT6-01: 1
		Mean :0.982	D5 : 31	TCGA-5M-AAAT7-01: 1
		3rd Qu.:1.000	G4 : 27	TCGA-5M-AAAT8-01: 1
		Max. :1.000	DM : 25	TCGA-A6-2670-01: 1
			(Other):107	(Other) :438
				(Other) :189

Table 4: Summary of the phenotypical information of the mRNA dataset.

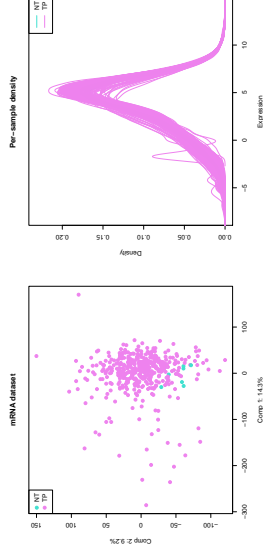


Figure 2: PCA and density plot for mRNAs.

3 Differentially expressed miRNAs

Analysis performed	Comparative used: CvH; method used: limma.
Number of differentially expressed miRNAs	325 (187 upregulated, 138 downregulated)
Number of samples	444
Criteria for selecting miRNAs	adj.pval < 1

Table 5: Basic statistics

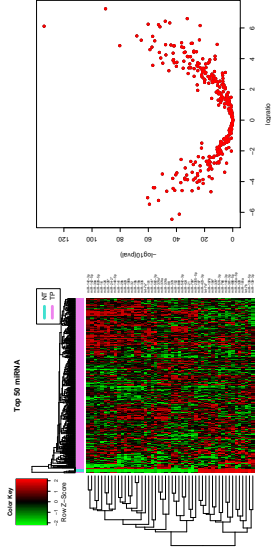


Figure 3: A) Heatmap with the top 50 most significant miRNAs (sorted by adjusted p-value). B) Volcano plot showing the selected miRNAs.

4 Differentially expressed mRNAs

Analysis performed	Comparative used: CvH; method used: limma.
Number of differentially expressed mRNAs	14860 (8526 upregulated, 6334 downregulated)
Number of samples	444
Criteria for selecting mRNAs	adj.pval < 1

Table 6: Basic statistics

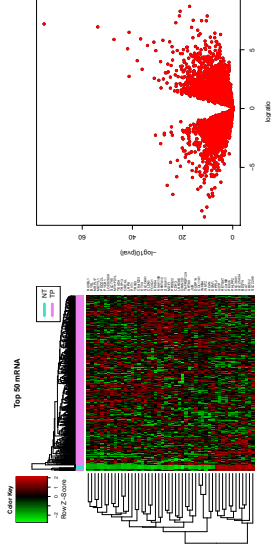


Figure 4: A) Heatmap with the top 50 most significant mRNAs (sorted by adjusted p-value). B) Volcano plot showing the selected mRNAs.

5 Correlation & intersection with databases

Number of miRNAs	325
Number of mRNAs	14860
Total miRNA-mRNA combinations	4829500
Number of samples	444

Table 7: Number of miRNAs, mRNAs and samples used for correlation.

	Number	%
Total correlations	4829500	100
Total negative correlations	2363105	48.93
Total correlations $p < 0.05$	1205347	24.96
Total correlations $p < 0.01$	849917	17.6
Total correlations $\text{adj. } p < 0.05$	823121	17.04
Total correlations $\text{adj. } p < 0.01$	568570	11.77

Table 8: Basic statistics for correlation results. Correlation hypothesis: two-sided.

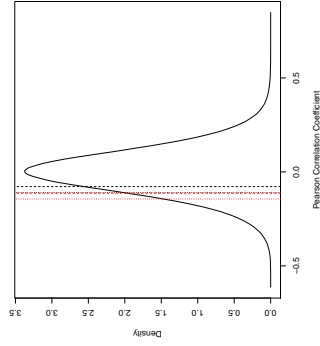


Figure 5: Density of a total of 4829500 miRNA-mRNA pairs. Dashed lines distinguish correlations whose p-value is lower than 0.05, dotted lines for 0.01. Black is for raw p-value and red for adjusted p-value.

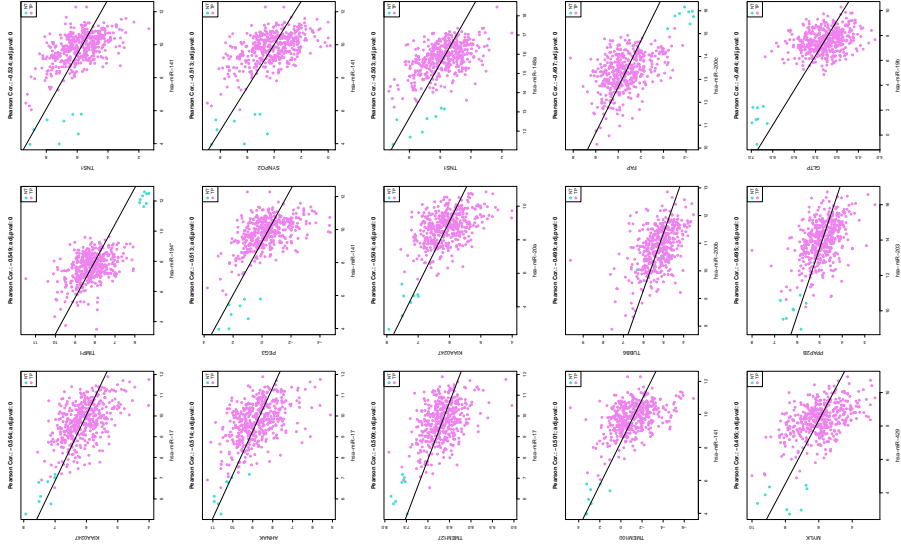


Figure 6: Plot of 15 top correlations, sorted by adjusted p-value. Databases used: microCosm_v3_L8, targetScan_v0.2.18 (each miRNA-mRNA pair has to appear at least 1 times).

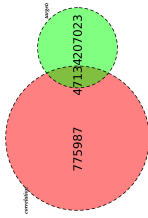


Figure 7: Venn Diagram. Left (red): number of miRNA-miRNA pairs with adjusted p-value < 0.05. Right (green): number of all the theoretical miRNA-miRNA pairs reported at least 1 times in the following databases: microCosm_v5.18, targetScan_v6.2.18. Intersection: miRNA-miRNA pairs that fulfil both conditions.

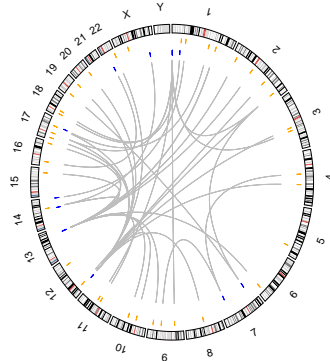


Figure 8: Circos plot for the first 45 miRNA-miRNA pairs (sorted by adjusted p-value) that have: pval-corrected < 0.05 and appear at least 1 times in the following databases: microCosm_v5.18, targetScan_v6.2.18. Blue: miRNAs, Orange: target mRNAs

miRNA	miRNA	cor	adj_pval	FC:miRNA	FC:miRNA	dat.sum
hsa-miR-17	KIAA0247	-0.56	8.86e-33	10.38	-2.39	1
hsa-miR-194*	TIMP1	-0.55	5.52e-31	-23.96	6.31	1
hsa-miR-141	TNS1	-0.52	5.03e-28	28.51	-2.68	1
hsa-miR-17	AHNAK	-0.51	7.18e-27	10.38	-3.01	1
hsa-miR-141	PEG3	-0.51	1.07e-26	28.51	-4.02	1
hsa-miR-141	SYNP02	-0.51	1.13e-26	28.51	-5.63	1
hsa-miR-17	TMEM127	-0.51	2.80e-26	10.38	-1.98	1
hsa-miR-20a	KIAA0247	-0.50	1.17e-25	21.55	-2.89	1
hsa-miR-148a	TNS1	-0.50	1.33e-25	10.39	-2.68	1
hsa-miR-141	TMEM100	-0.50	2.30e-25	28.51	-14.32	1
hsa-miR-200b	TUBB6	-0.50	3.87e-25	1.48	-1.03	1
hsa-miR-200c	FAP	-0.50	5.81e-25	-5.41	36.99	1
hsa-miR-429	MYLK	-0.50	9.52e-25	27.25	-4.06	1
hsa-miR-203	PPAP2B	-0.49	1.10e-24	15.89	-2.47	1
hsa-miR-196	GLTP	-0.49	1.15e-24	75.71	-2.95	1
hsa-miR-592	PRDM8	-0.49	1.50e-24	30.41	-1.27	1
hsa-miR-592	DAPK1	-0.49	2.05e-24	30.41	-2.11	1
hsa-miR-200a	TNS1	-0.49	3.15e-24	9.14	-2.68	1
hsa-miR-17	FAM129A	-0.49	3.60e-24	10.38	-4.18	1
hsa-miR-141	LMO3	-0.49	4.47e-24	28.51	-6.79	1
hsa-miR-17	ZBTB4	-0.49	5.14e-24	10.38	-1.29	1
hsa-miR-17	GSN	-0.49	5.14e-24	10.38	-2.98	1
hsa-miR-106a	KIAA0247	-0.49	6.09e-24	8.97	-2.39	1
hsa-miR-200c	FAM19A5	-0.49	6.70e-24	-5.41	4.10	1
hsa-miR-200b	DNAJB5	-0.49	8.22e-24	1.48	-1.86	1
hsa-miR-200b*	P4HA3	-0.48	1.58e-23	-3.84	6.41	1
hsa-miR-17	KCNMA1	-0.48	1.90e-23	10.38	-5.31	1
hsa-miR-148a	CNN1	-0.48	2.98e-23	10.59	-3.92	1
hsa-miR-141	CDC20	-0.48	3.02e-23	28.51	-2.29	2
hsa-miR-194*	FHL3	-0.48	3.62e-23	-23.96	2.31	1
hsa-miR-200c	RAB34	-0.48	4.16e-23	-5.41	-1.05	1
hsa-miR-692*	C20orf43	-0.48	4.50e-23	-2.55	1.33	1
hsa-miR-17	KIAA0513	-0.48	4.50e-23	10.38	-4.91	1
hsa-miR-200c	DOK5	-0.48	7.01e-23	-5.41	1.72	1
hsa-miR-200b	RAB34	-0.48	8.66e-23	1.48	-1.05	1
hsa-miR-16	SLC39A1	-0.47	4.55e-22	15.32	-4.79	1
hsa-miR-141	CXCL12	-0.47	4.68e-22	28.51	-5.30	1
hsa-miR-21*	PEXIP1	-0.47	5.59e-22	5.48	-1.83	1
hsa-miR-17	AKAP13	-0.47	5.99e-22	10.38	-1.81	1
hsa-miR-20a	FAM129A	-0.47	6.62e-22	21.55	-4.18	1
hsa-miR-552	KCTD1	-0.47	8.67e-22	8.58	-1.32	1
hsa-miR-130b	MYH11	-0.47	8.85e-22	3.28	-11.11	1
hsa-miR-200c	SULF1	-0.47	8.87e-22	-5.41	5.32	1
hsa-miR-17	OSR1	-0.47	9.24e-22	10.38	-5.28	1
hsa-miR-17	SYNE1	-0.47	9.83e-22	10.38	-2.59	1

Table 9: Top 45 miRNA-miRNA pairs (sorted by adjusted p-value) that have: pval-corrected < 0.05 and appear at least 1 times in the following databases: microCosm_v5.18, targetScan_v6.2.18.

6 Functional analysis

6.1 Network analysis

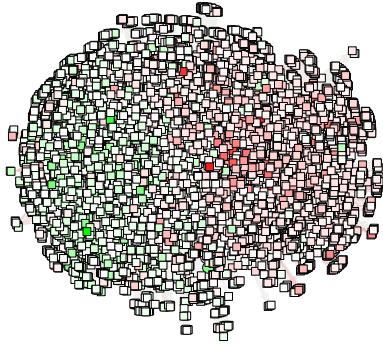


Figure 9: Network for all the miRNA-mRNA pairs that have: $p\text{-val-corrected} < 0.05$ and appear at least 1 times in the following databases: `microCosm_v5.18`, `targetScan_v6.2.18`. Circles represent the miRNAs, and squares the mRNA. Red fill means upregulated miRNAs/mRNAs, while green fill means downregulated miRNA/mRNAs in comparative CVH; lines indicate the miRNA-mRNA pairs, red line means positive score and green line means negative score.

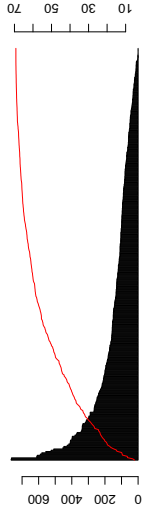


Figure 10: Barplot for miRNAs, $p\text{-val-corrected} < 0.05$ and `Targets=microCosm_v5.18`, `targetScan_v6.2.18`(minimum coincidences between databases:1). Red line (and right axis) represents the percentage of deregulated miRNAs that are targeted by the mRNAs.

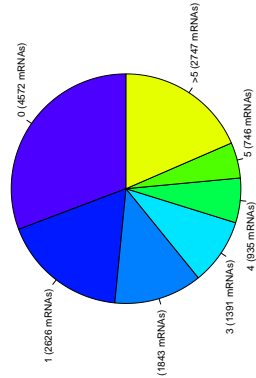


Figure 11: Pie chart representing the number of miRNAs targeting the mRNAs, $p\text{-val-corrected} < 0.05$ and `Targets=microCosm_v5.18`, `targetScan_v6.2.18`(minimum coincidences between databases:1).

miRNA	#targets	cum. % targets (top 20)	miRNAs (top 20)	mRNAs (top 20)
hsa-miR-106a	766	5.15	KIAA0247, AKAP13, AHNAK, SH3PXD2A, PTPN21, TMEM127, GOLGA2, LUZP1, ZBTB4, MTMR3, BCL2L2, FZD4, FRMD4B, FYCO1, TRIP11, VCL, ANKRD12, SYNE1, SNRK, NCOR1, KIAA0247, AHNAK, TMEM127, FAMI129A, ZBTB4, GSN, KCNMA1, KIAA0513, AKAP13, OSRI, SYNE1, FGFR2, PSD, TNS1, PTPN21, ZBTB47, NTN1, TIMP2, STG6GALNAC6, TXNIP	QKI 52 hsa-miR-141, hsa-miR-577, hsa-miR-200a, hsa-miR-200b, hsa-miR-429, hsa-miR-96, hsa-miR-194, hsa-miR-130b, hsa-miR-93, hsa-miR-362-5p, hsa-miR-183, hsa-miR-93*, hsa-miR-200c, hsa-miR-17, hsa-miR-576-5p, hsa-miR-19a, hsa-miR-106b, hsa-miR-30c, hsa-miR-375, hsa-miR-148a
hsa-miR-17	764	6.1	TRIP11, VCL, ANKRD12, SYNE1, SNRK, NCOR1, KIAA0247, AHNAK, TMEM127, FAMI129A, ZBTB4, GSN, KCNMA1, KIAA0513, AKAP13, OSRI, SYNE1, FGFR2, PSD, TNS1, PTPN21, ZBTB47, NTN1, TIMP2, STG6GALNAC6, TXNIP	FOXP2 49 hsa-miR-7-1*, hsa-miR-128, hsa-miR-7, hsa-miR-203, hsa-miR-186, hsa-miR-200a, hsa-miR-141, hsa-miR-576-5p, hsa-miR-577, hsa-miR-222, hsa-miR-16, hsa-miR-590-3p, hsa-miR-20a*, hsa-miR-20a*, hsa-miR-15a, hsa-miR-500-5p, hsa-miR-660, hsa-miR-10a*, hsa-miR-29a*, hsa-miR-15a, hsa-miR-1300*
hsa-miR-19a	612	8.99	CLTP, PTPN21, SCN9A, MYH11, SYNP02, C10orf26, PDE5A, CBX7, CILP, PLCL2, FBX032, TNS1, ZERI, SLC9A1, LPP, PDE7B, SCN4B, BCAR3, NPTN, SCN1B	RORA 48 hsa-miR-577, hsa-miR-93, hsa-miR-17, hsa-miR-141, hsa-miR-106a, hsa-miR-200a, hsa-miR-345, hsa-miR-166, hsa-miR-835*, hsa-miR-183, hsa-miR-20a*, hsa-miR-19a, hsa-miR-18a, hsa-miR-148a, hsa-miR-19b, hsa-miR-3013-3p, hsa-miR-576-3p, hsa-miR-20a, hsa-miR-148a, hsa-miR-19b
hsa-miR-20a	609	9.12	KIAA0247, FAMI129A, KIAA0513, KCNMA1, OSRI, TMEM127, PSD, FGFR2, NTN1, AHNAK, TNS1, SYNE1, PTPN21, FGL2, CNN1, KIAA1683, RELL1, LMO3, STG6GALNAC6, CYBRD1	IGF1 45 hsa-miR-130b, hsa-miR-577, hsa-miR-19a, hsa-miR-576-5p, hsa-miR-19b, hsa-miR-18a, hsa-miR-425, hsa-miR-192, hsa-miR-625, hsa-miR-186, hsa-miR-454, hsa-miR-15a, hsa-miR-335*, hsa-miR-942, hsa-miR-29b, hsa-miR-148a, hsa-miR-301a, hsa-miR-590-3p, hsa-miR-196a
hsa-miR-16	590	11.74	SLC36A1, LAOD1, MYLK, TMEM100, SCN4B, SMAD7, PPA2A, CNN1, DCLK1, KIAA0247, PPAP2B, ADAMTSL3, MAP1A, PRELP, BCL2L2, CNNM2, SMPD1, IGF1, NFE2L1, DDXDC1	BNC2 44 hsa-miR-141, hsa-miR-17, hsa-miR-577, hsa-miR-93, hsa-miR-429, hsa-miR-200a, hsa-miR-200b, hsa-miR-106b, hsa-miR-191, hsa-miR-20a, hsa-miR-183, hsa-miR-200c*, hsa-miR-125, hsa-miR-576-5p, hsa-miR-183, hsa-miR-200c*, hsa-miR-142, hsa-miR-19b
hsa-miR-106b	581	11.9	KIAA0247, TMEM127, TMEM100, KCNMA1, SLC36A1, FAMI129A, FBLM1, TNS1, LMO3, KIAA0513, CNN1, CYBRD1, PSD, RGMA, AHNAK, PTPN21, NTN1, SGCA, SYNM, SV2B	TNRC6B 41 hsa-miR-17, hsa-miR-106a, hsa-miR-203, hsa-miR-362-5p, hsa-miR-590-5p, hsa-miR-335, hsa-miR-19a, hsa-miR-2255-5p, hsa-miR-19b, hsa-miR-18a, hsa-miR-16, hsa-miR-301a, hsa-miR-106b, hsa-miR-135b, hsa-miR-503, hsa-miR-148a, hsa-miR-424, hsa-miR-29b, hsa-miR-369-3p
hsa-miR-19b	557	12.17	CLTP, MYH11, MAF, PLCL2, SLC9A1, NPTN, PTPN21, PDZD4, PDE5A, ABHD5, SYNP02, PSD, CILP, C10orf26, BCAR3, SGK1, TNS1, CSF1, FBX032, CLIP4	LPP 40 hsa-miR-19a, hsa-miR-18a, hsa-miR-19b, hsa-miR-16, hsa-miR-96, hsa-miR-141, hsa-miR-596-5p, hsa-miR-577, hsa-miR-32, hsa-miR-19a, hsa-miR-425, hsa-miR-424, hsa-miR-203, hsa-miR-428-5p, hsa-miR-576-5p, hsa-miR-30c, hsa-miR-490-5p, hsa-miR-142-3p, hsa-miR-2355-5p, hsa-miR-26b
hsa-miR-30c	525	14.68	LOX, CALU, ADAM12, FAP, LIMS1, TPM4, ITGA5, MEX3B, SNA2, PRR16, CTHRC1, STC1, FRMD6, GJA1, LPPR4, ZNF281, MRAS, ADAMT53, DCBLD1, CALD1	RUNX1IT1 40 hsa-miR-130b, hsa-miR-429, hsa-miR-25, hsa-miR-16, hsa-miR-200c, hsa-miR-15a, hsa-miR-203, hsa-miR-186, hsa-miR-500a, hsa-miR-192, hsa-miR-19a, hsa-miR-501-3p, hsa-miR-584, hsa-miR-19b, hsa-miR-148a, hsa-miR-29b, hsa-miR-15b, hsa-miR-148b, hsa-miR-215
hsa-miR-93	509	15.03	TIMP2, BNC2, CRYAB, CLIP4, SGCA, GPR137B, PSD, RGMA, LMO3, TNS1, TGFB1I1, KCNMA1, JAZF1, CYBRD1, ATP8B2, ZFPM2, PTGER3, FRMD6, GUCY1A3, ZBTB47	TRPS1 39 hsa-miR-93, hsa-miR-362-5p, hsa-miR-17, hsa-miR-194, hsa-miR-130b, hsa-miR-203, hsa-miR-148a, hsa-miR-19a, hsa-miR-106b, hsa-miR-200c, hsa-miR-429, hsa-miR-200b, hsa-miR-19b, hsa-miR-186, hsa-miR-106a, hsa-miR-33a, hsa-miR-500b, hsa-miR-3613-3p, hsa-miR-345
hsa-miR-96	454	16.05	LMOD1, TNS1, ASB2, MYL9, C20orf194, GSTM5, CRYAB, ADCY5, ITPRI1, MAF, CCDC80, BNC2, SCN9A, FILIP1, C10orf54, CD36, NLGN4X, FAMI10B, LDB3, NAALADL1	BACH2 38 hsa-miR-141, hsa-miR-130b, hsa-miR-186, hsa-miR-425, hsa-miR-429, hsa-miR-200a, hsa-miR-200b, hsa-miR-148a, hsa-miR-33a, hsa-miR-454, hsa-miR-301a, hsa-miR-29b, hsa-miR-200c, hsa-miR-16, hsa-miR-183, hsa-miR-15a, hsa-miR-96, hsa-miR-148b, hsa-miR-552, hsa-miR-33b

Table 10: Top 10 miRNA with more targets (each miRNA-mRNA pair has pval-corrected<0.05 and appears at least 1 times in the following databases: microCosm.v5.18, targetScan.v6.2.18). miRNAs in red are upregulated in CVH. miRNAs in green are downregulated in CVH.

Table 11: Top 10 mRNA with more miRNAs targeting them (each miRNA-mRNA pair has pval-corrected<0.05 and appears at least 1 times in the following databases: microCosm.v5.18, targetScan.v6.2.18). MRNAs in red are upregulated in CVH. MRNAs in green are downregulated in CVH.

6.2 GO analysis

GOBPID	Term	Count	Size	ExpCount	OddsRatio	fdr	Pvalue
GO:0035556	intracellular transduction signal	1401	2009	1137.75	1.92	1.58e-34	1.39e-38
GO:0048518	positive regulation of biological process	2490	3817	2161.66	1.62	1.18e-32	2.08e-36
GO:0044267	cellular metabolic process	2216	3370	1908.51	1.64	2.62e-31	6.95e-35
GO:0044237	cellular metabolic process	5373	8855	5014.80	1.52	3.88e-31	1.37e-34
GO:0048522	positive regulation of cellular process	2224	3397	1923.80	1.62	6.77e-30	3.83e-33
GO:0006464	cellular protein modification process	1733	2583	1462.82	1.71	6.77e-30	4.18e-33
GO:0038211	protein modification process	1733	2583	1462.82	1.71	6.77e-30	4.18e-33
GO:0048583	regulation of response to stimulus	1827	2741	1552.30	1.68	1.57e-29	1.17e-32
GO:0006793	phosphorus metabolic process	1868	2809	1590.81	1.67	1.57e-29	1.28e-32
GO:0006796	phosphate-containing compound metabolic process	1840	2763	1564.75	1.68	1.57e-29	1.39e-32

Table 12: Biological Process . Options used: mRNAs that are present in a mRNA-mRNA pair that has adjusted-pval cutoff <0.05; that also appears at least 1 times (databases: microCosm_v5.18, targetScan_v6.2.18); organism: human.

GOCCID	Term	Count	Size	ExpCount	OddsRatio	fdr	Pvalue
GO:0044424	intracellular part	7302	12258	6561.96	2.53	4.27e-147	3.27e-150
GO:0005622	intracellular	7360	12396	6635.83	2.53	4.63e-144	7.09e-147
GO:0005737	cytoplasm	5814	9342	5000.96	2.23	7.82e-141	1.80e-143
GO:0043227	membrane-bounded organelle	6297	10399	5566.79	2.13	4.86e-119	1.49e-121
GO:0043226	organelle	6690	11215	6003.62	2.13	4.30e-112	1.64e-114
GO:0044444	cytoplasmic part	4380	6864	3674.44	2.06	1.17e-111	5.38e-114
GO:0043231	intracellular nonmembrane-bounded organelle	5829	9551	5112.84	2.04	4.10e-110	2.20e-112
GO:0043229	intracellular organelle	6360	10617	5683.49	2.03	7.45e-104	4.56e-106
GO:0004446	intracellular organelle part	3934	6274	3358.60	1.82	1.53e-74	1.06e-76
GO:0044422	organelle part	4008	6449	3452.28	1.77	8.90e-69	6.82e-71

Table 13: Cellular Component . Options used: mRNAs that are present in a mRNA-mRNA pair that has adjusted-pval cutoff <0.05; that also appears at least 1 times (databases: microCosm_v5.18, targetScan_v6.2.18); organism: human.

GOMFID	Term	Count	Size	ExpCount	OddsRatio	fdr	Pvalue
GO:0005515	protein binding	4960	7904	4344.09	1.94	2.97e-87	8.95e-91
GO:0005488	binding	7046	12075	6636.50	1.87	4.06e-59	2.44e-62
GO:0003824	catalytic activity	3195	5197	2856.31	1.50	9.52e-29	8.60e-32
GO:0019899	enzyme binding	821	1175	645.79	2.00	8.66e-25	1.04e-27
GO:0043168	anion binding	1585	2493	1370.17	1.53	6.65e-19	1.00e-21
GO:0043167	ion binding	3457	5793	3183.87	1.36	1.18e-17	2.03e-20
GO:0032403	protein complex binding	464	656	360.54	2.04	9.22e-15	1.94e-17
GO:0019901	protein kinase binding	288	383	210.50	2.54	3.60e-14	8.67e-17
GO:0019904	protein domain specific binding	383	582	292.39	2.16	5.48e-14	1.61e-16
GO:0007367	carbohydrate derivative binding	1335	2115	1102.42	1.48	5.48e-14	1.65e-16

Table 14: Molecular Function . Options used: mRNAs that are present in a mRNA-mRNA pair that has adjusted-pval cutoff <0.05; that also appears at least 1 times (databases: microCosm_v5.18, targetScan_v6.2.18); organism: human.

KEGGID	Term	Count	Size	ExpCount	OddsRatio	fdr	P value
05200	Pathways in cancer	236	314	183.44	2.24	3.16e-08	1.40e-10
04510	Focal adhesion	150	192	112.17	2.61	4.49e-07	3.98e-09
04142	Lysosome	95	117	68.35	3.13	7.93e-06	1.05e-07
04350	TGF-beta signaling pathway	70	83	48.49	3.89	1.70e-05	3.02e-07
05215	Prostate cancer	72	87	50.83	3.47	4.19e-05	9.28e-07
05220	Chronic myeloid leukemia	61	72	42.06	4.00	4.87e-05	1.29e-06
04380	Osteoclast differentiation	93	118	68.94	2.70	6.18e-05	1.91e-06
05142	Chagas disease (American trypanosomiasis)	81	102	59.59	2.79	1.38e-04	5.15e-06
04066	Fc gamma R-mediated phagocytosis	74	92	53.75	2.97	1.38e-04	5.50e-06
05214	Glioma	53	63	36.80	3.82	1.98e-04	1.02e-05

Table 15: Kegg Pathways . Oprions used: mRNAs that are present in a mRNA-mRNA pair that has adjusted-pval cutoff <0.05; that also appears at least 1 times (databases: microCosm_v5_18, targetScan_v6.2.18); organism: human.

A.2 Study 1: Esophageal carcinoma

Default miRComb output

/home/mvilla/Baixades/TCGA/isophagous

May 13, 2015

1 Exploratory analysis of miRNA dataset

Number of miRNAs analysed 338
 Number of samples 191

Table 1: Basic information of the miRNA dataset.

group.n	CvH	center	sample	batch
1	NT: 13	Min.: 0.0000	TCGA-2H-A9GF-01: 1	Batch 382 :39
2	TP: 178	1st Qu.: 1.0000	TCGA-2H-A9GG-01: 1	Batch 272 :37
3		Median: 1.0000	TCGA-2H-A9GH-01: 1	Batch 374 :22
4		Mean: 0.9319	VR: 16	TCGA-2H-A9GI-01: 1
5		3rd Qu.: 1.0000	JY: 13	TCGA-2H-A9GJ-01: 1
6		Max.: 1.0000	R6: 13	TCGA-2H-A9GK-01: 1
7			(Other): 36	(Other) :185

Table 2: Summary of the phenotypical information of the miRNA dataset.

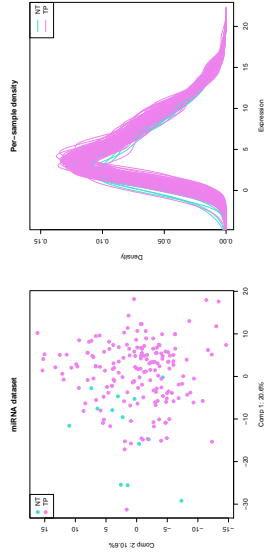


Figure 1: PCA and density plot for miRNAs.

2 Exploratory analysis of mRNA dataset

Number of mRNAs analysed 18807
 Number of samples 191

Table 3: Basic information of the mRNA dataset.

group.n	CvH	center	sample	batch
1	NT: 13	Min.: 0.0000	TCGA-2H-A9GF-01: 1	Batch 382 :39
2	TP: 178	1st Qu.: 1.0000	TCGA-2H-A9GG-01: 1	Batch 272 :37
3		Median: 1.0000	IG: 20	TCGA-2H-A9GH-01: 1
4		Mean: 0.9319	VR: 16	TCGA-2H-A9GI-01: 1
5		3rd Qu.: 1.0000	JY: 13	TCGA-2H-A9GJ-01: 1
6		Max.: 1.0000	R6: 13	TCGA-2H-A9GK-01: 1
7			(Other): 36	(Other) :185

Table 4: Summary of the phenotypical information of the mRNA dataset.

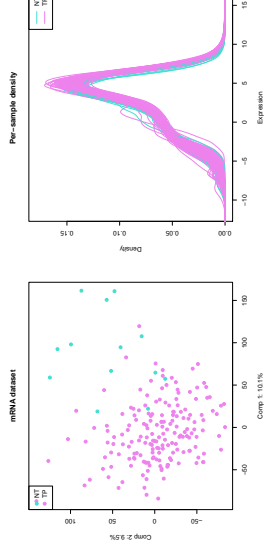


Figure 2: PCA and density plot for mRNAs.

3 Differentially expressed miRNAs

Analysis performed	Comparative used: CvH; method used: limma.
Number of differentially expressed miRNAs	338 (247 upregulated, 91 downregulated)
Number of samples	191
Criteria for selecting miRNAs	adj.pval < 1

Table 5: Basic statistics

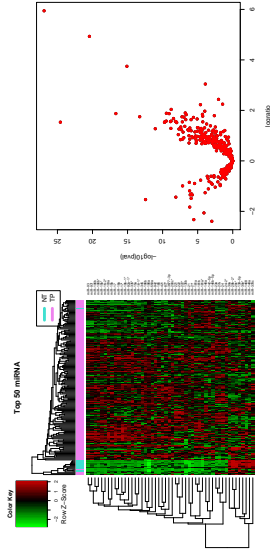


Figure 3: A) Heatmap with the top 50 most significant miRNAs (sorted by adjusted p-value). B) Volcano plot showing the selected miRNAs.

4 Differentially expressed mRNAs

Analysis performed	Comparative used: CvH; method used: limma.
Number of differentially expressed mRNAs	18807 (13153 upregulated, 5654 downregulated)
Number of samples	191
Criteria for selecting mRNAs	adj.pval < 1

Table 6: Basic statistics

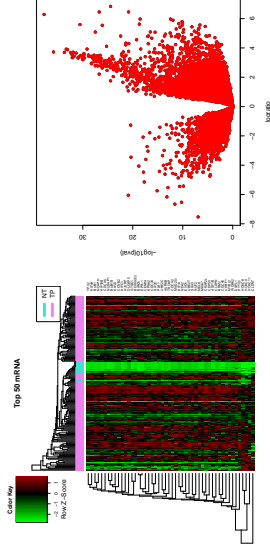


Figure 4: A) Heatmap with the top 50 most significant mRNAs (sorted by adjusted p-value). B) Volcano plot showing the selected mRNAs.

5 Correlation & intersection with databases

Number of miRNAs	338
Number of mRNAs	18807
Total miRNA-mRNA combinations	6350766
Number of samples	191

Table 7: Number of miRNAs, mRNAs and samples used for correlation.

	Number	%
Total correlations	6356766	100
Total negative correlations	2792875	43.94
Total correlations $p < 0.05$	1138006	17.9
Total correlations $p < 0.01$	705933	11.11
Total correlations adj $p < 0.05$	568914	8.95
Total correlations adj $p < 0.01$	336673	5.3

Table 8: Basic statistics for correlation results. Correlation hypothesis: two-sided.

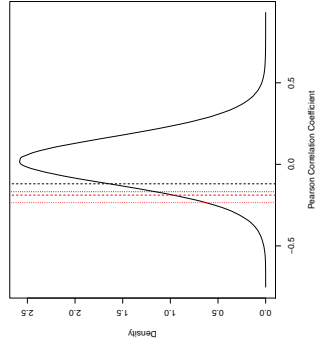


Figure 5: Density of a total of 6356766 miRNA-mRNA pairs. Dashed lines distinguish correlations whose p-value is lower than 0.05, dotted lines for 0.01. Black is for raw p-value and red for adjusted p-value.

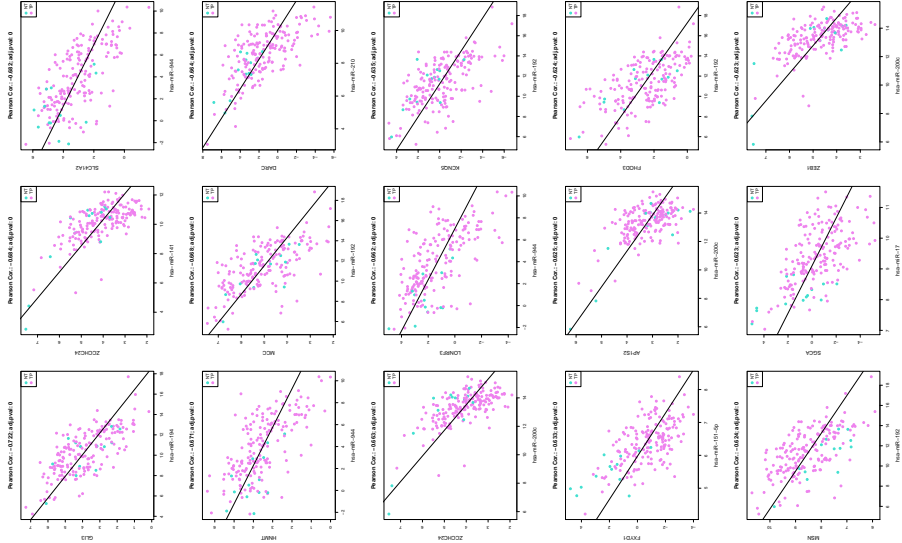


Figure 6: Plot of 15 top correlations, sorted by adjusted p-value. Databases used: microCosm_v3_L8, targetScan_v0.2.18 (each miRNA-mRNA pair has to appear at least 1 times).

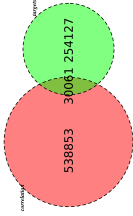


Figure 7: Venn Diagram. Left (red): number of miRNA-miRNA pairs with adjusted p-value<0.05. Right (green): number of all the theoretical miRNA-miRNA pairs reported at least 1 times in the following databases: microCosm_v5.18, targetScan_v6.2.18. Intersection: miRNA-miRNA pairs that fulfil both conditions.

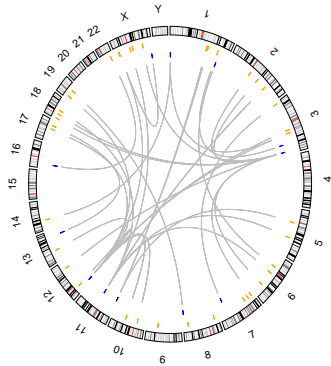


Figure 8: Circos plot for the first 45 miRNA-miRNA pairs (sorted by adjusted p-value) that have: pval-corrected<0.05 and appear at least 1 times in the following databases: microCosm_v5.18, targetScan_v6.2.18. Blue: miRNAs, Orange: target mRNAs

miRNA	miRNA	cor	adj.pval	FC:miRNA	FC:miRNA	dat.sum
hsa-miR-194	GLI3	-0.72	4.02e-26	1.15	1.06	2
hsa-miR-141	ZCCHC24	-0.68	1.71e-22	1.99	-2.01	1
hsa-miR-944	SLC11A2	-0.68	2.40e-22	8.25	-2.31	2
hsa-miR-944	HNMT	-0.67	2.80e-21	8.25	-1.38	2
hsa-miR-192	MCC	-0.67	5.45e-21	1.30	1.01	1
hsa-miR-210	DARC	-0.66	1.08e-20	2.25	-4.51	1
hsa-miR-200c	ZCCHC24	-0.66	1.17e-20	1.75	-2.01	1
hsa-miR-944	LONRF3	-0.66	1.33e-20	8.25	-1.66	1
hsa-miR-192	KCNQ3	-0.64	1.53e-18	1.30	-2.52	1
hsa-miR-151-5p	FXYD1	-0.63	2.22e-18	1.89	-6.73	1
hsa-miR-200c	API52	-0.62	8.54e-18	1.75	-1.03	2
hsa-miR-192	FHOD3	-0.62	9.08e-18	1.30	-1.29	1
hsa-miR-192	MSN	-0.62	9.36e-18	1.30	1.65	1
hsa-miR-17	SGCA	-0.62	1.09e-17	2.58	-3.67	1
hsa-miR-200c	ZEB1	-0.62	1.11e-17	1.75	-1.30	2
hsa-miR-141	ZEB1	-0.62	2.49e-17	1.99	-1.30	1
hsa-miR-192	OSBPL6	-0.62	2.76e-17	1.30	-1.31	1
hsa-miR-200c	LHFP	-0.62	3.02e-17	1.75	-1.01	1
hsa-miR-200c	MYLK	-0.62	3.26e-17	1.75	-2.03	1
hsa-miR-16	LMOD1	-0.62	3.74e-17	1.75	-5.10	1
hsa-miR-944	GATA6	-0.61	4.60e-17	8.25	-2.24	2
hsa-miR-194	BICD2	-0.60	1.95e-16	1.15	1.61	1
hsa-miR-944	ARHGAP18	-0.60	2.65e-16	8.25	-1.48	1
hsa-miR-141	CDC80	-0.60	2.76e-16	1.99	-1.98	2
hsa-miR-194	DFNA5	-0.60	4.16e-16	1.15	1.39	1
hsa-miR-193b	MGAT3	-0.60	6.13e-16	-1.03	-1.03	1
hsa-miR-194	MAGEE1	-0.60	6.36e-16	1.15	-2.40	1
hsa-miR-452	REPS2	-0.59	8.80e-16	3.41	-2.62	1
hsa-miR-205	ENPP4	-0.59	1.04e-15	5.41	-2.42	1
hsa-miR-96	ITPRI	-0.59	1.06e-15	2.75	-2.56	2
hsa-miR-944	MGAT4A	-0.59	1.53e-15	8.25	-1.12	1
hsa-miR-944	ICA1	-0.59	1.68e-15	8.25	-1.15	1
hsa-miR-15b	LMOD1	-0.59	2.16e-15	2.40	-5.10	1
hsa-miR-200c	ABCC9	-0.59	2.54e-15	1.75	-2.08	2
hsa-miR-429	MYLK	-0.59	2.59e-15	2.05	-2.03	1
hsa-miR-16	CNN1	-0.58	3.59e-15	1.75	-5.60	1
hsa-miR-200c	KANK2	-0.58	4.46e-15	1.75	-2.16	1
hsa-miR-141	MAP3K3	-0.58	4.48e-15	1.99	-1.11	1
hsa-miR-210	PPAP2A	-0.58	5.52e-15	2.25	-1.28	1
hsa-miR-944	LPIN2	-0.58	6.20e-15	8.25	-1.29	1
hsa-miR-194	TUSC3	-0.58	7.14e-15	1.15	1.37	1
hsa-miR-429	CFL2	-0.58	7.51e-15	2.05	-1.95	1
hsa-miR-194	MTS1	-0.58	7.87e-15	1.15	-1.27	1
hsa-miR-210	NPRI	-0.58	7.87e-15	2.25	-2.41	1
hsa-miR-192*	OSBPL6	-0.58	7.98e-15	1.32	-1.31	1

Table 9: Top 45 miRNA-miRNA pairs (sorted by adjusted p-value) that have: pval-corrected<0.05 and appear at least 1 times in the following databases: microCosm_v5.18, targetScan_v6.2.18.

6 Functional analysis

6.1 Network analysis

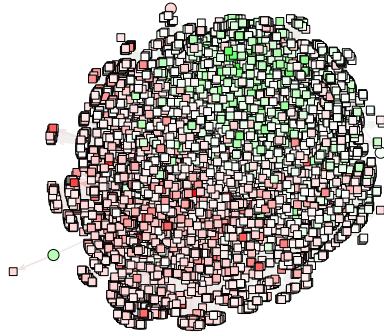


Figure 9: Network for all the miRNA-mRNA pairs that have: $p\text{-val-corrected} < 0.05$ and appear at least 1 times in the following databases: microCosm_v5.18, targetScan_v6.2.18. Circles represent the miRNAs, and squares the mRNA. Red fill means upregulated miRNAs/mRNAs, while green fill means downregulated miRNA/mRNAs in comparative CVH; lines indicate the miRNA-mRNA pairs, red line means positive score and green line means negative score.

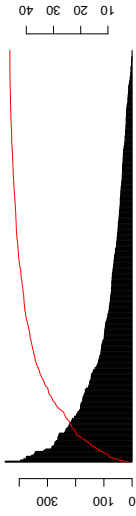


Figure 10: Barplot for miRNAs, $p\text{-val-corrected} < 0.05$ and Targets=microCosm_v5.18, targetScan_v6.2.18 (minimum coincidences between databases:1). Red line (and right axis) represents the percentage of deregulated miRNAs that are targeted by the miRNAs.

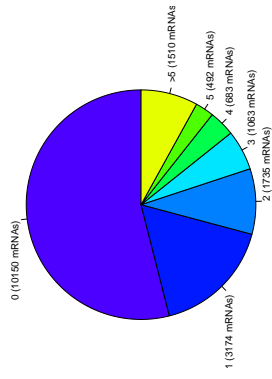


Figure 11: Pie chart representing the number of miRNAs targeting the miRNAs, $p\text{-val-corrected} < 0.05$ and Targets=microCosm_v5.18, targetScan_v6.2.18 (minimum coincidences between databases:1).

miRNA	#targets	cum. %	targets (top 20)	miRNAs	#miRNAs	miRNAs (top 20)
hsa-miR-27a	450	2.39	ADCY6, PDK4, SELENBP1, RALGAP5, CCC2, DET1, NFA1C2, SIEBGR2, GPR133, SORBS1, PTF5KB, EEPD1, PLEKHG6, REFS2, RAB1FIP2, PBXIP1, PTPRB, MAN2A2, NR2F2, SPATA13	45	hsa-miR-141, hsa-miR-200c, hsa-miR-590-5p, hsa-miR-200a*, hsa-miR-96, hsa-miR-429, hsa-miR-101, hsa-let-7g*, hsa-miR-10a, hsa-miR-183, hsa-miR-592, hsa-miR-17, hsa-miR-3127-5p, hsa-miR-182, hsa-miR-130b, hsa-miR-19, hsa-miR-196a, hsa-miR-200b, hsa-miR-425, hsa-miR-532-5p	
hsa-miR-29c	395	4.49	EIF2S1, FBOXO5, CSE1L, ODF2, REXO4, RC2, TUBA1B, CDK2, CCNA2, TMEM201, MTHFD1L, YY1, YWHAE, COL11A1, VTA1, C9orf89, GCCT, ULBP2, MAP4K4, Clorf135	42	hsa-miR-590-3p, hsa-miR-18a, hsa-miR-1976, hsa-miR-141, hsa-miR-96, hsa-miR-31, hsa-miR-19f, hsa-miR-92a, hsa-miR-203, hsa-miR-183, hsa-miR-425, hsa-miR-19a, hsa-miR-16, hsa-miR-424, hsa-miR-32, hsa-miR-942, hsa-miR-182, hsa-miR-142-3p, hsa-miR-390-5p, hsa-miR-200a	
hsa-miR-203	384	6.23	ANXA6, ZEB1, RBPM5, SLC1A7, GNC2, MEF2C, GPRC5B, Clg6orf45, FKBP7, PAP2B, PAR3B, RUNX1T1, FOXP1, DDXDC1, INSR, RGS5, ITGA9, PRKCB, FGL2, COL4A4	40	hsa-miR-141, hsa-miR-429, hsa-miR-200c, hsa-miR-16, hsa-miR-200a, hsa-miR-532-5p, hsa-let-7g*, hsa-miR-30a, hsa-miR-10b, hsa-miR-425*, hsa-miR-19a, hsa-miR-577, hsa-miR-30b, hsa-miR-339-5p, hsa-miR-100a, hsa-miR-130b, hsa-miR-30d, hsa-miR-592, hsa-miR-17	
hsa-miR-23a	375	7.62	ESRRG, TSPAN12, UBL3, PLCXD3, SELENBP1, MYOCD, RAB1FIP2, FAM66C, MAGI1, DAPK1, LRR31, KLHD7A, SLC4A4, LRIG1, PTPRB, GPR64, ZC4H2, TOX3, REFS2, ARIF3	40	hsa-miR-200c, hsa-miR-33a, hsa-miR-203, hsa-miR-455-3p, hsa-miR-16, hsa-miR-92a, hsa-miR-15b, hsa-miR-429, hsa-miR-130b, hsa-miR-19a, hsa-miR-19b, hsa-miR-19f, hsa-miR-584, hsa-miR-15a, hsa-miR-320b, hsa-miR-27a, hsa-miR-103a, hsa-miR-148b, hsa-let-7d, hsa-miR-23a	
hsa-miR-429	375	9.26	MYLK, CFL2, NFASC, DNABP5, API52, ABCG9, CLIC4, ZCCHC24, FHL1, LHFP, ZFPM2, TLN1, SYDEL1, NPTX1, QKI, OSTM1, KANK2, PEX3, NCS1, FSTL1	37	hsa-miR-192*, hsa-miR-29a*, hsa-miR-877, hsa-miR-660, hsa-miR-222, hsa-miR-19a, hsa-miR-7, hsa-miR-501-3p, hsa-miR-21, hsa-miR-196a, hsa-miR-15a, hsa-miR-34a, hsa-miR-92a, hsa-miR-577, hsa-miR-171*, hsa-miR-671-3p, hsa-miR-180, hsa-miR-503, hsa-miR-502-3p, hsa-miR-19b	
hsa-miR-200c	356	9.62	ZCCHC24, API52, ZEB1, LHFP, MYLK, ABCG9, KANK2, FHL1, TLN1, CFL2, ZNF423, RECK, RUNX1T1, FSTL1, ZEB2, TIMP2, GPRASP1, ZFPM2, NFASC, DLG1	36	hsa-miR-19a, hsa-miR-18a, hsa-miR-942, hsa-miR-196b, hsa-miR-19f, hsa-miR-15b, hsa-miR-16, hsa-miR-222, hsa-miR-130b, hsa-miR-15a, hsa-miR-196a, hsa-miR-576-5p, hsa-miR-1976, hsa-miR-378, hsa-miR-503, hsa-miR-425, hsa-miR-192, hsa-miR-335*, hsa-miR-92b, hsa-miR-590-3p	
hsa-miR-200b	347	9.76	CFL2, DNABP5, MYLK, NFASC, API52, SYDEL1, LHFP, ZCCHC24, NCS1, QKI, ZFPM2, ABCG9, CLIC4, NPTX1, FHL1, FSTL1, CITED2, OSTM1, TLN1, KANK2	35	hsa-miR-194, hsa-miR-141, hsa-miR-200c, hsa-miR-429, hsa-miR-200b, hsa-miR-577, hsa-miR-200a, hsa-miR-502-3p, hsa-let-7g*, hsa-miR-17*, hsa-miR-19b, hsa-miR-501-5p, hsa-miR-19a, hsa-miR-590-3p, hsa-miR-20a*, hsa-miR-18a, hsa-miR-100b, hsa-miR-106a, hsa-miR-335, hsa-miR-148a*	
hsa-miR-96	345	10.89	ITPRI, FILIP1, LDB3, TNS1, DDXDC1, ZEB1, CACNA1C, POPDC2, CSRNP1, FYCO1, LMOD1, ITM2A, LRCH2, PPP1R12C, MAP3K3, ACVHL1, PDE7B, MYOCD, IAZF1, CELF2	34	hsa-miR-17, hsa-miR-20a, hsa-miR-20a*, hsa-miR-135b, hsa-miR-335*, hsa-miR-16-2*, hsa-miR-33a, hsa-miR-942, hsa-miR-192*, hsa-miR-100b, hsa-miR-425, hsa-miR-532-5p, hsa-miR-677-5p, hsa-miR-106a*, hsa-miR-21*, hsa-miR-29a*, hsa-miR-339-3p, hsa-miR-34a, hsa-miR-584	
hsa-miR-27b	342	11.62	CCG2, FNDCA, NR5A2, DMXL2, DENND5B, LPN2, NKTR, RAPGEF2, ACA2, GFPT1, NLK, PEAK1, SLC30A7, TMED5, CDH2, RAB20, PLEKHH1, ZC3H12D, GCA, GOLI1	34	hsa-miR-20a*, hsa-miR-18a, hsa-miR-92a, hsa-miR-17, hsa-miR-19a, hsa-miR-335*, hsa-miR-16, hsa-miR-20a, hsa-miR-19b, hsa-miR-100b, hsa-miR-107, hsa-miR-3615-5p, hsa-miR-141, hsa-miR-183, hsa-miR-652, hsa-miR-15a, hsa-miR-21*, hsa-miR-92b, hsa-miR-550a, hsa-miR-148b	
hsa-miR-944	313	12.6	SLC41A2, HNMT, LONRF3, GATA6, ARHGAP18, MGAT4A, ICAM1, LPN2, VPS13C, SLC12A2, NR3C2, HSD17B11, FOXP1, THRA, C2orf88, PTPRB, TMEM50B, C20orf112, C11orf54, SEPSecs	32	hsa-miR-141, hsa-miR-200c, hsa-miR-141*, hsa-let-7a*, hsa-miR-200b*, hsa-miR-200a*, hsa-miR-17, hsa-miR-429, hsa-miR-16, hsa-miR-24, hsa-miR-19b-1*, hsa-miR-130b, hsa-miR-151-5p, hsa-miR-106b, hsa-miR-20a, hsa-miR-200b, hsa-miR-19b, hsa-let-7b*, hsa-miR-19a, hsa-miR-93	

Table 10: Top 10 miRNA with more targets (each miRNA-mRNA pair has pval-corrected<0.05 and appears at least 1 times in the following databases: microCosm.v5.18, targetScan.v6.2.18). miRNAs in red are upregulated in CVH. miRNAs in green are downregulated in CVH.

miRNA	#targets	cum. %	targets (top 20)	miRNAs	#miRNAs	miRNAs (top 20)
hsa-miR-27a	450	2.39	ADCY6, PDK4, SELENBP1, RALGAP5, CCC2, DET1, NFA1C2, SIEBGR2, GPR133, SORBS1, PTF5KB, EEPD1, PLEKHG6, REFS2, RAB1FIP2, PBXIP1, PTPRB, MAN2A2, NR2F2, SPATA13	45	hsa-miR-141, hsa-miR-200c, hsa-miR-590-5p, hsa-miR-200a*, hsa-miR-96, hsa-miR-429, hsa-miR-101, hsa-let-7g*, hsa-miR-10a, hsa-miR-183, hsa-miR-592, hsa-miR-17, hsa-miR-3127-5p, hsa-miR-182, hsa-miR-130b, hsa-miR-19, hsa-miR-196a, hsa-miR-200b, hsa-miR-425, hsa-miR-532-5p	
hsa-miR-29c	395	4.49	EIF2S1, FBOXO5, CSE1L, ODF2, REXO4, RC2, TUBA1B, CDK2, CCNA2, TMEM201, MTHFD1L, YY1, YWHAE, COL11A1, VTA1, C9orf89, GCCT, ULBP2, MAP4K4, Clorf135	42	hsa-miR-590-3p, hsa-miR-18a, hsa-miR-1976, hsa-miR-141, hsa-miR-96, hsa-miR-31, hsa-miR-19f, hsa-miR-92a, hsa-miR-203, hsa-miR-183, hsa-miR-425, hsa-miR-19a, hsa-miR-16, hsa-miR-424, hsa-miR-32, hsa-miR-942, hsa-miR-182, hsa-miR-142-3p, hsa-miR-390-5p, hsa-miR-200a	
hsa-miR-203	384	6.23	ANXA6, ZEB1, RBPM5, SLC1A7, GNC2, MEF2C, GPRC5B, Clg6orf45, FKBP7, PAP2B, PAR3B, RUNX1T1, FOXP1, DDXDC1, INSR, RGS5, ITGA9, PRKCB, FGL2, COL4A4	40	hsa-miR-141, hsa-miR-429, hsa-miR-200c, hsa-miR-16, hsa-miR-200a, hsa-miR-532-5p, hsa-let-7g*, hsa-miR-30a, hsa-miR-10b, hsa-miR-425*, hsa-miR-19a, hsa-miR-577, hsa-miR-30b, hsa-miR-339-5p, hsa-miR-100a, hsa-miR-130b, hsa-miR-30d, hsa-miR-592, hsa-miR-17	
hsa-miR-23a	375	7.62	ESRRG, TSPAN12, UBL3, PLCXD3, SELENBP1, MYOCD, RAB1FIP2, FAM66C, MAGI1, DAPK1, LRR31, KLHD7A, SLC4A4, LRIG1, PTPRB, GPR64, ZC4H2, TOX3, REFS2, ARIF3	40	hsa-miR-200c, hsa-miR-33a, hsa-miR-203, hsa-miR-455-3p, hsa-miR-16, hsa-miR-92a, hsa-miR-15b, hsa-miR-429, hsa-miR-130b, hsa-miR-19a, hsa-miR-19b, hsa-miR-19f, hsa-miR-584, hsa-miR-15a, hsa-miR-320b, hsa-miR-27a, hsa-miR-103a, hsa-miR-148b, hsa-let-7d, hsa-miR-23a	
hsa-miR-429	375	9.26	MYLK, CFL2, NFASC, DNABP5, API52, ABCG9, CLIC4, ZCCHC24, FHL1, LHFP, ZFPM2, TLN1, SYDEL1, NPTX1, QKI, OSTM1, KANK2, PEX3, NCS1, FSTL1	37	hsa-miR-192*, hsa-miR-29a*, hsa-miR-877, hsa-miR-660, hsa-miR-222, hsa-miR-19a, hsa-miR-7, hsa-miR-501-3p, hsa-miR-21, hsa-miR-196a, hsa-miR-15a, hsa-miR-34a, hsa-miR-92a, hsa-miR-577, hsa-miR-171*, hsa-miR-671-3p, hsa-miR-180, hsa-miR-503, hsa-miR-502-3p, hsa-miR-19b	
hsa-miR-200c	356	9.62	ZCCHC24, API52, ZEB1, LHFP, MYLK, ABCG9, KANK2, FHL1, TLN1, CFL2, ZNF423, RECK, RUNX1T1, FSTL1, ZEB2, TIMP2, GPRASP1, ZFPM2, NFASC, DLG1	36	hsa-miR-19a, hsa-miR-18a, hsa-miR-942, hsa-miR-196b, hsa-miR-19f, hsa-miR-15b, hsa-miR-16, hsa-miR-222, hsa-miR-130b, hsa-miR-15a, hsa-miR-196a, hsa-miR-576-5p, hsa-miR-1976, hsa-miR-378, hsa-miR-503, hsa-miR-425, hsa-miR-192, hsa-miR-335*, hsa-miR-92b, hsa-miR-590-3p	
hsa-miR-200b	347	9.76	CFL2, DNABP5, MYLK, NFASC, API52, SYDEL1, LHFP, ZCCHC24, NCS1, QKI, ZFPM2, ABCG9, CLIC4, NPTX1, FHL1, FSTL1, CITED2, OSTM1, TLN1, KANK2	35	hsa-miR-194, hsa-miR-141, hsa-miR-200c, hsa-miR-429, hsa-miR-200b, hsa-miR-577, hsa-miR-200a, hsa-miR-502-3p, hsa-let-7g*, hsa-miR-17*, hsa-miR-19b, hsa-miR-501-5p, hsa-miR-19a, hsa-miR-590-3p, hsa-miR-20a*, hsa-miR-18a, hsa-miR-100b, hsa-miR-106a, hsa-miR-335, hsa-miR-148a*	
hsa-miR-96	345	10.89	ITPRI, FILIP1, LDB3, TNS1, DDXDC1, ZEB1, CACNA1C, POPDC2, CSRNP1, FYCO1, LMOD1, ITM2A, LRCH2, PPP1R12C, MAP3K3, ACVHL1, PDE7B, MYOCD, IAZF1, CELF2	34	hsa-miR-17, hsa-miR-20a, hsa-miR-20a*, hsa-miR-135b, hsa-miR-335*, hsa-miR-16-2*, hsa-miR-33a, hsa-miR-942, hsa-miR-192*, hsa-miR-100b, hsa-miR-425, hsa-miR-532-5p, hsa-miR-677-5p, hsa-miR-106a*, hsa-miR-21*, hsa-miR-29a*, hsa-miR-339-3p, hsa-miR-34a, hsa-miR-584	
hsa-miR-27b	342	11.62	CCG2, FNDCA, NR5A2, DMXL2, DENND5B, LPN2, NKTR, RAPGEF2, ACA2, GFPT1, NLK, PEAK1, SLC30A7, TMED5, CDH2, RAB20, PLEKHH1, ZC3H12D, GCA, GOLI1	34	hsa-miR-20a*, hsa-miR-18a, hsa-miR-92a, hsa-miR-17, hsa-miR-19a, hsa-miR-335*, hsa-miR-16, hsa-miR-20a, hsa-miR-19b, hsa-miR-100b, hsa-miR-107, hsa-miR-3615-5p, hsa-miR-141, hsa-miR-183, hsa-miR-652, hsa-miR-15a, hsa-miR-21*, hsa-miR-92b, hsa-miR-550a, hsa-miR-148b	
hsa-miR-944	313	12.6	SLC41A2, HNMT, LONRF3, GATA6, ARHGAP18, MGAT4A, ICAM1, LPN2, VPS13C, SLC12A2, NR3C2, HSD17B11, FOXP1, THRA, C2orf88, PTPRB, TMEM50B, C20orf112, C11orf54, SEPSecs	32	hsa-miR-141, hsa-miR-200c, hsa-miR-141*, hsa-let-7a*, hsa-miR-200b*, hsa-miR-200a*, hsa-miR-17, hsa-miR-429, hsa-miR-16, hsa-miR-24, hsa-miR-19b-1*, hsa-miR-130b, hsa-miR-151-5p, hsa-miR-106b, hsa-miR-20a, hsa-miR-200b, hsa-miR-19b, hsa-let-7b*, hsa-miR-19a, hsa-miR-93	

Table 11: Top 10 miRNA with more miRNAs targeting them (each miRNA-mRNA pair has pval-corrected<0.05 and appears at least 1 times in the following databases: microCosm.v5.18, targetScan.v6.2.18). miRNAs in red are upregulated in CVH. miRNAs in green are downregulated in CVH.

6.2 GO analysis

GOBPID	Term	Count	Size	ExpCount	OddsRatio	fdr	Pvalue
GO:0045731	system development	1968	3416	1664.51	1.59	1.03e-28	9.45e-33
GO:0009653	anatomical structure morphogenesis	1280	2131	1038.37	1.71	2.51e-26	4.00e-30
GO:0007339	nervous system development	1108	1818	885.86	1.76	1.40e-25	3.86e-29
GO:0016043	cellular component organization	2662	4619	2250.70	1.49	1.83e-25	8.31e-29
GO:0023052	signaling	2652	4798	2337.92	1.48	1.83e-25	1.01e-28
GO:0044700	single organism signaling	2652	4798	2337.92	1.48	1.83e-25	1.01e-28
GO:0007154	cell communication	2687	4871	2373.49	1.47	3.14e-25	2.02e-28
GO:0065007	biological regulation	4653	8895	4334.26	1.45	2.20e-24	1.61e-27
GO:0071840	cellular component organization or biogenesis	2539	4715	2287.47	1.46	9.27e-24	7.64e-27
GO:0007165	signal transduction	2404	4333	2111.34	1.47	1.76e-23	1.61e-26

Table 12: Biological Process . Options used: mRNAs that are present in a mRNA-mRNA pair that has adjusted-pval cutoff <0.05; that also appears at least 1 times (databases: microCosm_v5.18, targetScan_v6.2.18); organism: human.

GOCCID	Term	Count	Size	ExpCount	OddsRatio	fdr	Pvalue
GO:0005737	cytoplasm	4858	9342	4302.38	1.73	3.70e-65	2.97e-68
GO:0044224	intracellular part	6105	12258	5645.33	1.79	5.00e-57	8.01e-60
GO:0005622	intracellular cytoplasmic part	6162	12396	5708.88	1.80	9.20e-57	2.21e-59
GO:0044444	cytoplasmic part	3608	6864	3161.16	1.56	6.63e-43	2.12e-45
GO:0043226	organelle	5573	11215	5164.98	1.57	1.16e-39	4.66e-42
GO:0043227	membrane-bounded organelle	5188	10399	4780.18	1.51	1.63e-35	7.82e-38
GO:0043229	intracellular organelle	5208	10617	4889.58	1.49	1.75e-32	9.80e-35
GO:0043231	intracellular membrane-bounded organelle	4765	9551	4398.64	1.44	9.52e-29	6.10e-31
GO:0005829	cytosol	1436	2545	1172.08	1.64	4.37e-28	3.15e-30
GO:0044464	cell part	6915	14496	6676.02	1.69	9.18e-27	7.35e-29

Table 13: Cellular Component . Options used: mRNAs that are present in a mRNA-mRNA pair that has adjusted-pval cutoff <0.05; that also appears at least 1 times (databases: microCosm_v5.18, targetScan_v6.2.18); organism: human.

GOMFID	Term	Count	Size	ExpCount	OddsRatio	fdr	Pvalue
GO:000515	protein binding	4229	7904	3713.51	1.76	9.13e-63	2.92e-66
GO:0005688	binding	6009	12075	5673.15	1.76	7.76e-40	4.98e-43
GO:0008092	cytoskeletal binding	453	704	330.76	2.11	1.88e-18	1.81e-21
GO:0043168	anion binding	1378	2483	1171.28	1.49	5.65e-17	7.25e-20
GO:0019899	enzyme binding	697	1175	552.05	1.72	4.54e-16	7.27e-19
GO:0019904	protein domain specific binding	347	582	249.95	2.17	2.68e-15	5.16e-18
GO:0004672	protein kinase activity	365	565	265.45	2.12	3.08e-15	6.31e-17
GO:0016773	phosphotransferase activity, alcohol group as acceptor	426	679	319.01	1.86	9.17e-15	2.35e-17
GO:0016301	kinase activity	453	732	343.91	1.89	2.64e-14	7.62e-17
GO:0043167	ion binding	2966	5793	2721.70	1.32	4.87e-14	1.56e-16

Table 14: Molecular Function . Options used: mRNAs that are present in a mRNA-mRNA pair that has adjusted-pval cutoff <0.05; that also appears at least 1 times (databases: microCosm_v5.18, targetScan_v6.2.18); organism: human.

KEGGID	Term	Count	Size	ExpCount	OddsRatio	fdr	Pvalue
04510	Focal adhesion	142	192	92.71	3.15	3.62e-11	1.59e-13
05200	Pathways in cancer	209	314	151.61	2.23	1.63e-09	1.44e-11
04512	ECM-receptor interaction	65	81	39.11	4.43	1.89e-07	2.50e-09
04360	Axon guidance	87	125	60.36	2.50	5.03e-05	8.86e-07
04070	Phosphatidylinositol signaling system	58	78	37.66	3.15	8.53e-05	2.12e-06
05222	Small cell lung cancer	61	83	40.08	3.01	8.53e-05	2.25e-06
04916	Melanogenesis	71	100	48.28	2.67	9.28e-05	2.86e-06
04910	Insulin signaling pathway	88	133	64.22	2.13	5.56e-04	1.96e-05
04666	Fc gamma R-mediated phagocytosis	64	92	44.42	2.48	6.44e-04	2.55e-05
04270	Vascular smooth muscle contraction	74	111	53.60	2.17	1.38e-03	6.10e-05

Table 15: Kegg Pathways . Options used: mRNAs that are present in a mRNA-mRNA pair that has adjusted-pval cutoff <0.05; that also appears at least 1 times (databases: microCosm_v5.18, targetScan_v6.2.18); organism: human.

A.3 Study 1: Liver hepatocellular carcinoma

Default miRComb output

/home/mvila/Baixades/TCGA/liver

June 4, 2015

1 Exploratory analysis of miRNA dataset

Number of miRNAs analysed 343
 Number of samples 407

Table 1: Basic information of the miRNA dataset.

group.n	CvH	center	sample	batch
1	NT: 50	Min.: -0.0000	DD :172	TCGA-2Y-A9GS-01: 1 Batch 425: 83
2	TP:357	1st Qu.:1.0000	G3 : 33	TCGA-2Y-A9GT-01: 1 Batch 100: 62
3		Median:1.0000	CC : 32	TCGA-2Y-A9GU-01: 1 Batch 203: 33
4		Mean :0.8771	BC : 31	TCGA-2Y-A9GV-01: 1 Batch 399: 32
5		3rd Qu.:1.0000	2Y : 20	TCGA-2Y-A9GW-01: 1 Batch 231: 24
6		Max. :1.0000	ED : 15	TCGA-2Y-A9GX-01: 1 Batch 384: 21
7			(Other):104 (Other):401	(Other):152

Table 2: Summary of the phenotypical information of the miRNA dataset.

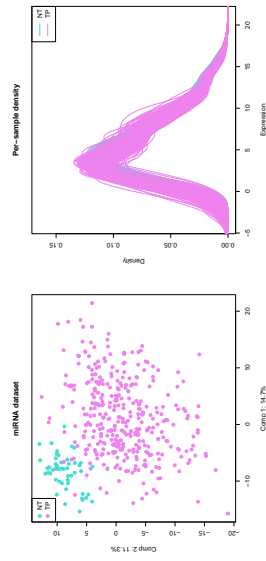


Figure 1: PCA and density plot for miRNAs.

2 Exploratory analysis of mRNA dataset

Number of mRNAs analysed 14428
 Number of samples 407

Table 3: Basic information of the mRNA dataset.

group.n	CvH	center	sample	batch
1	NT: 50	Min.: -0.0000	DD :172	TCGA-2Y-A9GS-01: 1 Batch 425: 83
2	TP:357	1st Qu.:1.0000	G3 : 33	TCGA-2Y-A9GT-01: 1 Batch 100: 62
3		Median :1.0000	CC : 32	TCGA-2Y-A9GU-01: 1 Batch 203: 33
4		Mean :0.8771	BC : 31	TCGA-2Y-A9GV-01: 1 Batch 399: 32
5		3rd Qu.:1.0000	2Y : 20	TCGA-2Y-A9GW-01: 1 Batch 231: 24
6		Max. :1.0000	ED : 15	TCGA-2Y-A9GX-01: 1 Batch 384: 21
7			(Other):104 (Other):401	(Other):152

Table 4: Summary of the phenotypical information of the mRNA dataset.

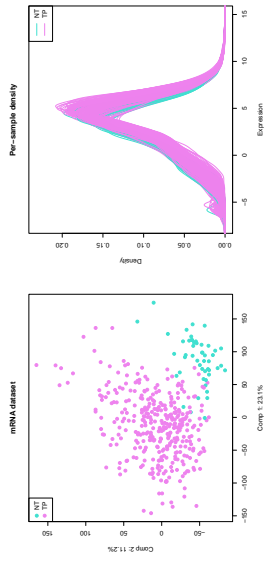


Figure 2: PCA and density plot for mRNAs.

3 Differentially expressed miRNAs

Analysis performed	Comparative used: CvH; method used: limma.
Number of differentially expressed miRNAs	343 (111 upregulated, 232 downregulated)
Number of samples	407
Criteria for selecting miRNAs	adj.pval < 1

Table 5: Basic statistics

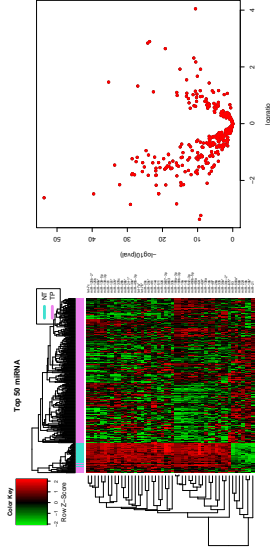


Figure 3: A) Heatmap with the top 50 most significant miRNAs (sorted by adjusted p-value). B) Volcano plot showing the selected miRNAs.

4 Differentially expressed mRNAs

Analysis performed	Comparative used: CvH; method used: limma.
Number of differentially expressed mRNAs	14428 (11134 upregulated, 3294 downregulated)
Number of samples	407
Criteria for selecting mRNAs	adj.pval < 1

Table 6: Basic statistics

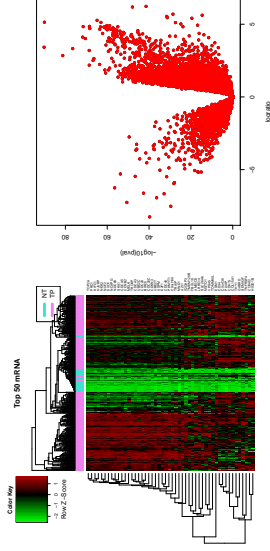


Figure 4: A) Heatmap with the top 50 most significant mRNAs (sorted by adjusted p-value). B) Volcano plot showing the selected mRNAs.

5 Correlation & intersection with databases

Number of miRNAs	343
Number of mRNAs	14428
Total miRNA-mRNA combinations	494804
Number of samples	407

Table 7: Number of miRNAs, mRNAs and samples used for correlation.

	Number	%
Total correlations	494804	100
Total negative correlations	2424205	48.99
Total correlations $p < 0.05$	1460936	29.52
Total correlations $p < 0.01$	1130070	22.84
Total correlations adj. $p < 0.05$	1156839	23.38
Total correlations adj. $p < 0.01$	889513	17.97

Table 8: Basic statistics for correlation results. Correlation hypothesis: two-sided.

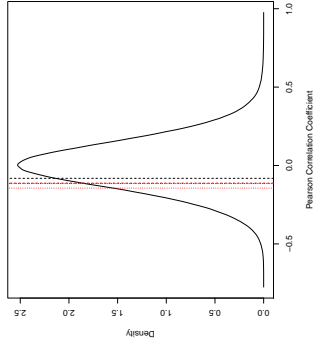


Figure 5: Density of a total of 494804 miRNA-mRNA pairs. Dashed lines distinguish correlations whose p -value is lower than 0.05, dotted lines for 0.01. Black is for raw p -value and red for adjusted p -value.

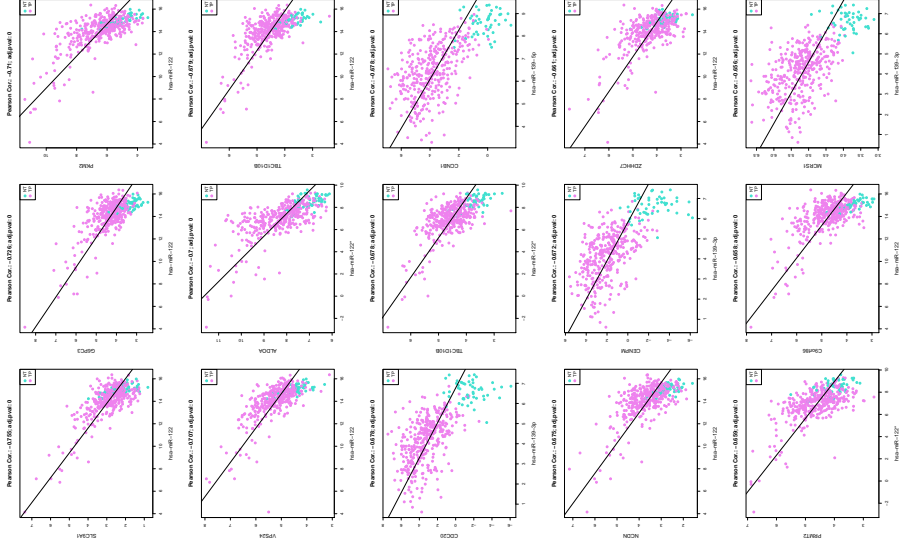


Figure 6: Plot of 15 top correlations, sorted by adjusted p -value. Databases used: microCosm_v3_L8, targetScan_v0.2.18 (each miRNA-mRNA pair has to appear at least 1 times).

miRNA	miRNA	cor	adj.pval	FC:miRNA	FC:mRNA	dat.sum
hsa-miR-122	SLC9A1	-0.75	5.07e-70	-2.12	1.59	1
hsa-miR-122	G6PC3	-0.73	3.65e-62	-2.12	2.12	2
hsa-miR-122	PKM2	-0.71	4.30e-58	-2.12	2.94	1
hsa-miR-122	VPS24	-0.71	2.03e-57	-2.12	1.84	1
hsa-miR-122*	ALDOA	-0.70	5.94e-56	-3.23	2.52	1
hsa-miR-122	TBC1D10B	-0.68	2.71e-51	-2.12	1.92	2
hsa-miR-139-3p	CDC20	-0.68	3.66e-51	-5.56	29.84	1
hsa-miR-122*	TBC1D10B	-0.68	4.09e-51	-3.23	1.92	1
hsa-miR-139-5p	CCNB1	-0.68	4.09e-51	-4.44	12.46	1
hsa-miR-122	NCDN	-0.67	1.52e-50	-2.12	2.05	1
hsa-miR-139-3p	CENPM	-0.67	5.07e-50	-5.56	20.25	1
hsa-miR-122	ZDHHC7	-0.66	5.75e-48	-2.12	1.52	1
hsa-miR-122*	PRMT2	-0.66	1.44e-47	-3.23	1.55	1
hsa-miR-122	C9orf86	-0.66	1.90e-47	-2.12	2.19	2
hsa-miR-139-3p	MCRS1	-0.66	6.67e-47	-5.56	2.54	1
hsa-miR-122	GRS1	-0.65	1.76e-46	-2.12	1.86	1
hsa-miR-122	CHST12	-0.65	4.10e-46	-2.12	1.64	1
hsa-miR-139-5p	CDC48	-0.65	1.13e-45	-4.44	12.19	1
hsa-miR-122	GTT1	-0.65	2.87e-45	-2.12	2.75	1
hsa-miR-139-3p	RBCK1	-0.64	5.54e-45	-5.56	2.42	1
hsa-miR-139-3p	CDC6	-0.64	2.52e-44	-5.56	14.70	1
hsa-miR-122	DULLARD	-0.64	6.06e-44	-2.12	1.52	1
hsa-miR-139-3p	DTYMK	-0.64	1.41e-43	-5.56	3.30	1
hsa-miR-139-5p	H2AFZ	-0.63	6.18e-43	-4.44	2.93	1
hsa-miR-122	ALDOA	-0.63	9.34e-43	-2.12	2.52	2
hsa-miR-148a	TMSB10	-0.63	2.82e-42	-1.61	2.61	2
hsa-miR-139-5p	TTK	-0.63	3.11e-42	-4.44	28.38	1
hsa-miR-139-5p	BUB1B	-0.63	8.04e-42	-4.44	19.62	1
hsa-miR-139-5p	MAD2L1	-0.62	1.91e-41	-4.44	5.89	1
hsa-miR-148a	TRAPPPC4	-0.62	4.16e-41	-1.61	1.86	1
hsa-miR-139-3p	PKMYT1	-0.62	1.31e-40	-5.56	15.09	1
hsa-miR-139-5p	KPNA2	-0.62	1.64e-40	-4.44	3.79	1
hsa-miR-101	C20orf20	-0.62	2.02e-40	-2.75	2.64	2
hsa-miR-139-5p	EZH2	-0.62	4.06e-40	-4.44	8.44	1
hsa-miR-122	PLEKHB2	-0.62	5.51e-40	-2.12	1.44	1
hsa-miR-122	ATN1	-0.61	1.52e-39	-2.12	1.92	1
hsa-miR-125b	UCK2	-0.61	2.59e-39	-2.59	3.36	1
hsa-miR-139-5p	CENPE	-0.61	2.98e-39	-4.44	13.96	1
hsa-miR-122*	NDRG3	-0.61	4.91e-39	-3.23	2.71	1
hsa-miR-139-5p	CCT3	-0.61	6.23e-39	-4.44	3.09	1
hsa-miR-139-3p	H2AFX	-0.61	6.39e-39	-5.56	3.40	1
hsa-miR-122	SLC10A3	-0.61	1.02e-38	-2.12	1.79	1
hsa-miR-139-3p	SLC25A39	-0.61	2.13e-38	-5.56	2.42	1
hsa-miR-139-5p	ANAPC7	-0.60	7.30e-38	-4.44	2.53	1
hsa-miR-22	HNRNPA3	-0.60	8.53e-38	-1.50	1.90	1

Table 9: Top 45 miRNA-mRNA pairs (sorted by adjusted p-value) that have: pval-corrected<0.05 and appear at least 1 times in the following databases: microCosm_v5.18, targetScan_v6.2.18.

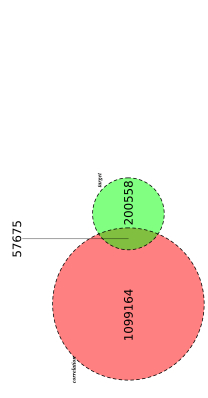


Figure 7: Venn Diagram. Left (red): number of miRNA-mRNA pairs with adjusted p-value<0.05. Right (green): number of all the theoretical miRNA-mRNA pairs reported at least 1 times in the following databases: microCosm_v5.18, targetScan_v6.2.18. Intersection: miRNA-mRNA pairs that fulfil both conditions.

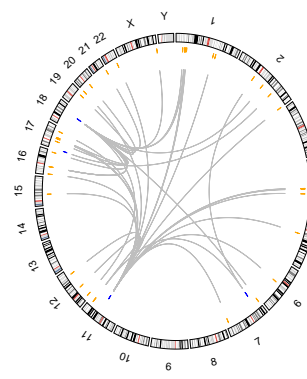


Figure 8: Circos plot for the first 45 miRNA-mRNA pairs (sorted by adjusted p-value) that have: pval-corrected<0.05 and appear at least 1 times in the following databases: microCosm_v5.18, targetScan_v6.2.18. Blue: miRNAs, Orange: target mRNAs

6 Functional analysis

6.1 Network analysis

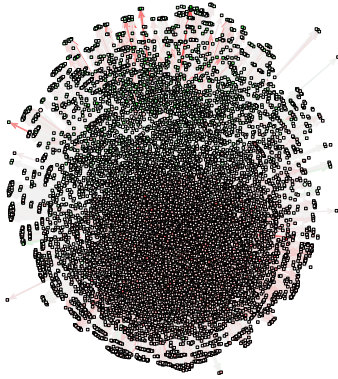


Figure 9: Network for all the miRNA-mRNA pairs that have: $p\text{-val-corrected} < 0.05$ and appear at least 1 times in the following databases: `microCosm_v5.18`, `targetScan_v6.2.18`. Circles represent the miRNAs, and squares the mRNA. Red fill means upregulated miRNAs/mRNAs, while green fill means downregulated miRNA/mRNAs in comparative CVH; lines indicate the miRNA-mRNA pairs, red line means positive score and green line means negative score.

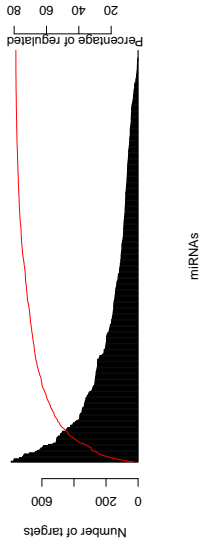


Figure 10: Barplot for miRNAs, $p\text{-val-corrected} < 0.05$ and `Targets=microCosm_v5.18`, `targetScan_v6.2.18`(minimum coincidences between databases:1). Red line (and right axis) represents the percentage of deregulated miRNAs that are targeted by the miRNAs.

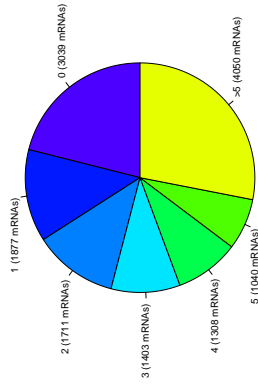


Figure 11: Pie chart representing the number of miRNAs targeting the mRNAs, $p\text{-val-corrected} < 0.05$ and `Targets=microCosm_v5.18`, `targetScan_v6.2.18`(minimum coincidences between databases:1).

miRNA	#miRNAs	miRNAs (top 20)
ROSA	41	hsa-miR-18a, hsa-miR-324-3p, hsa-miR-17, hsa-miR-181b, hsa-miR-93, hsa-miR-92b, hsa-miR-20a, hsa-miR-183, hsa-miR-106b, hsa-miR-652, hsa-miR-421, hsa-miR-92a, hsa-miR-501-3p, hsa-miR-181a, hsa-miR-19a, hsa-miR-132, hsa-miR-155, hsa-miR-671-5p, hsa-miR-19b, hsa-miR-224
SAP30L	38	hsa-miR-511, hsa-miR-336, hsa-miR-450a, hsa-miR-142-3p, hsa-miR-142-5p, hsa-miR-582-5p, hsa-miR-145, hsa-miR-450b-5p, hsa-miR-144, hsa-miR-203, hsa-miR-29c, hsa-miR-337-3p, hsa-miR-27a*, hsa-miR-136*, hsa-miR-29b, hsa-miR-26b, hsa-miR-378*, hsa-miR-493, hsa-let-7c, hsa-let-7c-1, hsa-miR-101, hsa-miR-378c, hsa-let-7c, hsa-miR-29c, hsa-miR-22, hsa-miR-26b, hsa-miR-144, hsa-miR-378, hsa-miR-3607-3p, hsa-miR-339, hsa-miR-148a, hsa-miR-130a, hsa-miR-328, hsa-miR-50a*, hsa-miR-29a, hsa-miR-20a, hsa-miR-542-3p, hsa-miR-145, hsa-miR-865-5p, hsa-miR-326
CASK	36	hsa-miR-19b-1*, hsa-miR-885-5p, hsa-miR-92a, hsa-miR-101, hsa-miR-375, hsa-miR-497, hsa-miR-203, hsa-miR-483-3p, hsa-miR-424, hsa-miR-241*, hsa-miR-144, hsa-miR-758, hsa-miR-19b, hsa-miR-424*, hsa-miR-142-3p, hsa-miR-125b-2*, hsa-miR-411, hsa-miR-195, hsa-miR-142-5p, hsa-miR-194
NF1A	36	hsa-miR-21, hsa-miR-410, hsa-miR-301a, hsa-miR-889, hsa-miR-485-3p, hsa-miR-92b, hsa-miR-92a, hsa-miR-382, hsa-miR-200b, hsa-miR-19a, hsa-miR-155, hsa-miR-338-5p, hsa-miR-200a, hsa-miR-369-5p, hsa-miR-142b, hsa-miR-19b, hsa-miR-421, hsa-miR-370, hsa-miR-501-3p, hsa-miR-134
MIRPL43	34	hsa-let-7a*, hsa-miR-378*, hsa-miR-26a, hsa-miR-10a, hsa-miR-26b, hsa-miR-542-3p, hsa-miR-378, hsa-miR-22, hsa-miR-194*, hsa-miR-451, hsa-miR-342-5p, hsa-miR-22*, hsa-miR-142-3p, hsa-miR-885-5p, hsa-miR-214*, hsa-miR-99a, hsa-miR-152, hsa-miR-126, hsa-miR-215, hsa-miR-122
ATXN7L1	33	hsa-miR-375, hsa-miR-18a, hsa-miR-17, hsa-miR-92a, hsa-miR-20a*, hsa-miR-20b, hsa-miR-19b, hsa-miR-39b, hsa-miR-19b, hsa-miR-106a, hsa-miR-889, hsa-miR-148b, hsa-miR-495, hsa-miR-538-5p, hsa-miR-537-3p, hsa-miR-744*, hsa-miR-130a, hsa-miR-379, hsa-miR-338-3p, hsa-miR-199b-5p
HLF	33	hsa-miR-483, hsa-miR-181b, hsa-miR-331-5p, hsa-miR-181d, hsa-miR-18a, hsa-miR-181a, hsa-miR-199b-5p, hsa-miR-181c, hsa-miR-181e, hsa-miR-200c, hsa-miR-338-5p, hsa-miR-652, hsa-let-7i, hsa-miR-141, hsa-miR-148b, hsa-miR-17, hsa-miR-425, hsa-miR-106b, hsa-miR-223, hsa-miR-200a
NFIB	33	hsa-miR-224-3p, hsa-miR-17, hsa-miR-192, hsa-miR-103a, hsa-miR-421, hsa-miR-20a, hsa-miR-92a, hsa-miR-539, hsa-miR-185, hsa-miR-19a, hsa-miR-410, hsa-miR-19b, hsa-miR-582-5p, hsa-miR-194, hsa-miR-382, hsa-miR-136*, hsa-miR-409-3p, hsa-miR-136, hsa-miR-301a, hsa-miR-495, hsa-miR-139-5p, hsa-miR-375, hsa-miR-483-3p, hsa-miR-199a-3p, hsa-miR-10a, hsa-let-7a*, hsa-miR-214*, hsa-miR-654-3p, hsa-miR-30e*, hsa-miR-125b, hsa-miR-20a*, hsa-miR-378*, hsa-miR-22, hsa-miR-39c, hsa-miR-326, hsa-miR-570c, hsa-let-7a*, hsa-miR-381, hsa-let-7e*, hsa-miR-21*
HDAC8	32	hsa-miR-109, hsa-let-7b*, hsa-miR-214*, hsa-miR-654-3p, hsa-miR-30e*, hsa-miR-125b, hsa-miR-20a*, hsa-miR-378*, hsa-miR-22, hsa-miR-39c, hsa-miR-326, hsa-miR-570c, hsa-let-7a*, hsa-miR-381, hsa-let-7e*, hsa-miR-21*

Table 11: Top 10 miRNA with more miRNAs targeting them (each miRNA-miRNA pair has pval-corrected<0.05 and appears at least 1 times in the following databases: mi-cosmic.v5.18, targetScan.v6.2.18). miRNAs in red are upregulated in CVH, miRNAs in green are downregulated in CVH.

miRNA	#targets	cum. % targets (top 20)	targets (top 20)
hsa-miR-27b	792	5.49	LPCAT1, PSMD7, B3GNT1L, AAK1, KIAA0513, MIR523, Chord4, MSTO1, TPM3, HML3, RPN2, EFN3, WDR45, KP21B, LMK1, GNS, DAP3, ACCN2, PSMA1, SLC7A11
hsa-miR-29c	776	10.24	FAM136A, MYBL2, DNMT3A, CCNA2, SUV420H2, CDC39, ADSL, DDX49, C5orf13, XP05, SOX12, TAF11, CENPE, TET1, PPP1CC, CSE1L, GSG2, LSM1, RFXN2, C20orf11
hsa-let-7c	773	14.24	AURKB, NME6, HMGAI, EIF2S2, CCT3, TMM17B, EFSL3, UFC1, ARID3A, UBE2T, SNRPD1, CENPA, ATP6V1F, EME1, SLC33B1, CCNF, C16orf21, RRM2, NME1, TIT1L
hsa-miR-497	740	17.54	APLN, SLC26A6, SRPK1, ACACA, GGA3, ZNFHIT3, SNRPC, NRRP1, PPL1, EHMT2, PDCD11, HMGAI, XRCC6, ZBTB9, ZSCAN2, FAM189B, TOMM20, BTF3, SERPIN1, RECQL5
hsa-miR-148a	720	20.83	TMSB10, TRAPP4, ZDHHC7, TUBB6, PCNXL2, VPS24, STX3, TAGLN2, CD151, CFL1, C9orf66, DNAJC18, MTHFD1L, ITGB4, FMNL3, SULF1, CTNBP2NL, CHMP4B, JAG2, TMSL3
hsa-miR-125b	714	23.54	UCK2, HML3, SLC26A6, COPZ1, C9orf100, ELOVL1, SMG5, SRPRB, TMEM120B, TOMM40, FIGU, KIAA1522, SNRNP, BUB1B, E2F2, TMEM201, STX6, BAK1, MCB1, UBE2I
hsa-miR-101	711	26.03	C20orf20, LASS5, C5orf76, C9orf29, DENR, TRIM11, PLXNA1, ANKRD32, DNMT3A, C5orf33, CCT4, IQGAP3, TUBA1C, NAP1L1, EZH2, ENAH, RIT1, UCK2, TCEB1, MELK
hsa-miR-144	678	27.71	SNRPE, FIGC, Clorf77, RAB11FIP4, SF3B4, DSTYK, RUSC1, ANKZF1, NOL12, GGPS1, ORMDL2, MRPS23, SNRNP, BCL6, LASS5, RASD2, DCTN2, TAF6, B3GALNT2
hsa-miR-424	642	28.4	APLN, AMIGO3, RECQL5, FAM189B, UBE2Q1, MXD3, SNRPC, BAT4, ZNHT3, NSMCE2, TOMM20, MTX1, BCAP31, PUF06, E4F1, CDKN2A, DUSL1, PSKHI, NFKB1L1, TARBP1
hsa-miR-30e	630	30.12	C5orf76, YWHAZ, MTHFD1L, ZNF706, RASL12, NPC2, FKBP1A, MICALL1, CTRHC1, FBXO32, DTX2, P4HA2, C19orf50, NME6, FAP, PSMB5, STK39, STOML1, DGKZ, TMC7

Table 10: Top 10 miRNA with more targets (each miRNA-miRNA pair has pval-corrected<0.05 and appears at least 1 times in the following databases: mi-cosmic.v5.18, targetScan.v6.2.18). miRNAs in red are upregulated in CVH, miRNAs in green are downregulated in CVH.

6.2 GO analysis

GOBPID	Term	Count	Size	ExpCount	OddsRatio	fdr	Pvalue
GO:0044237	cellular metabolic process	6098	8855	5531.80	1.99	7.67e-83	6.67e-87
GO:0008152	metabolic process	6025	9835	6144.01	1.87	2.48e-64	4.32e-68
GO:0071704	organic substance metabolic process	6270	9266	5788.55	1.82	1.85e-61	4.82e-65
GO:0044238	primary metabolic process	6092	8977	5608.01	1.81	4.81e-61	1.67e-64
GO:0044267	cellular protein metabolic process	2503	3370	2105.27	2.01	1.21e-57	5.24e-61
GO:0044280	cellular macromolecule metabolic process	4680	6758	4221.78	1.72	7.61e-53	3.97e-56
GO:0044248	cellular catabolic process	1485	1936	1209.44	2.17	2.26e-43	1.38e-46
GO:0044710	single-organism metabolic process	3289	4649	2904.27	1.71	3.76e-43	2.61e-46
GO:0071840	cellular component organization or biogenesis	3325	4715	2945.50	1.69	1.21e-41	9.44e-45
GO:0009056	catabolic process	1724	2296	1434.33	2.00	2.97e-41	2.59e-44

Table 12: Biological Process . Options used: mRNAs that are present in a mRNA-mRNA pair that has adjusted-pval cutoff <0.05; that also appears at least 1 times (databases: microCosm_v5_18, targetScan_v6_2_18); organism: human.

GOCCID	Term	Count	Size	ExpCount	OddsRatio	fdr	Pvalue
GO:0044424	intracellular part	8224	12258	7224.25	3.58	2.84e-273	2.17e-276
GO:0005622	intracellular	8285	12396	7305.58	3.57	7.34e-268	1.12e-270
GO:0005737	cytoplasm	6561	9342	5505.71	2.94	2.44e-244	5.58e-247
GO:0043227	membrane-bounded organelle	7129	10399	6128.65	2.89	5.62e-229	1.72e-231
GO:0043226	organelle	7565	11215	6609.56	2.83	9.15e-222	3.49e-224
GO:0043231	intracellular membrane-bounded organelle	6621	9551	5628.88	2.76	3.47e-217	1.59e-219
GO:0043229	intracellular organelle	7207	10617	6257.13	2.77	2.31e-209	1.28e-211
GO:0044444	cytoplasmic part	4974	6864	4045.30	2.69	4.49e-196	2.71e-198
GO:0044446	intracellular organelle part	4542	6274	3697.58	2.54	8.61e-168	5.91e-170
GO:0044422	organelle part	4627	6449	3800.72	2.45	1.29e-158	9.83e-161

Table 13: Cellular Component . Options used: mRNAs that are present in a mRNA-mRNA pair that has adjusted-pval cutoff <0.05; that also appears at least 1 times (databases: microCosm_v5_18, targetScan_v6_2_18); organism: human.

GOMFID	Term	Count	Size	ExpCount	OddsRatio	fdr	Pvalue
GO:0005515	protein binding	5539	7904	4816.90	2.25	1.84e-125	5.30e-129
GO:0005488	binding	7861	12075	7358.81	2.33	2.12e-91	1.22e-94
GO:0003824	catalytic activity	3646	5197	3167.18	1.84	1.20e-61	1.04e-64
GO:0044822	poly(A) RNA binding	886	1107	674.63	2.74	1.87e-42	2.16e-45
GO:0019899	enzyme binding	923	1175	716.08	2.50	3.92e-38	5.66e-41
GO:0003723	RNA binding	1121	1477	900.12	2.16	9.63e-35	1.67e-37
GO:1901363	heterocyclic compound binding	3758	5568	3393.28	1.56	2.60e-34	5.24e-37
GO:0097159	organic cyclic compound binding	3800	5641	3437.77	1.55	1.08e-33	2.50e-36
GO:0043168	anion binding	1785	2493	1519.30	1.76	1.61e-31	4.17e-34
GO:1901265	nucleoside phosphate binding	1618	2255	1374.25	1.76	8.25e-29	2.38e-31

Table 14: Molecular Function . Options used: mRNAs that are present in a mRNA-mRNA pair that has adjusted-pval cutoff <0.05; that also appears at least 1 times (databases: microCosm_v5_18, targetScan_v6_2_18); organism: human.

KEGGID	Term	Count	Size	ExpCount	OddsRatio	fdr	Pvalue
01100	Metabolic pathways	831	1116	729.73	1.70	6.50e-11	2.86e-13
05215	Prostate cancer	82	87	56.89	8.85	1.36e-08	1.20e-10
05200	Pathways in cancer	254	314	205.32	2.33	2.90e-08	3.90e-10
04120	Ubiquitin mediated proteolysis	114	128	83.70	4.42	2.90e-08	5.12e-10
04141	Protein processing in endoplasmic reticulum	140	163	106.58	3.31	8.94e-08	1.97e-09
05210	Colorectal cancer	59	61	39.89	15.85	9.98e-08	2.64e-09
05212	Pancreatic cancer	66	70	45.77	8.87	2.72e-07	8.38e-09
04110	Cell cycle	108	124	81.08	3.65	7.55e-07	2.81e-08
05220	Chronic myeloid leukemia	67	72	47.08	7.20	7.55e-07	2.99e-08
05211	Renal cell carcinoma	62	66	43.16	8.33	8.42e-07	3.71e-08

Table 15: Kegg Pathways . Options used: mRNAs that are present in a mRNA-mRNA pair that has adjusted-pval cutoff <0.05; that also appears at least 1 times (databases: microCosm_v5.18, targetScan_v6.2.18); organism: human.

A.4 Study 1: Rectum adenocarcinoma

Default miRComb output

/home/mvilla/Baixades/TCGA/rectum

May 13, 2015

1 Exploratory analysis of miRNA dataset

Number of miRNAs analysed 325
 Number of samples 160

Table 1: Basic information of the miRNA dataset.

group.n	CvH	center	sample	batch	platform
1	NT: 3	Min.: -0.0000	AG :71	TCGA-AF-2687-01: 1	Batch 42 :36
2	TP:157	1st Qu.:1.0000	AF :18	TCGA-AF-2689-11: 1	Batch 122:27
3		Median :1.0000	EI :17	TCGA-AF-2690-01: 1	Batch 139:25
4		Mean :0.9812	DC :13	TCGA-AF-2691-01: 1	Batch 46 :17
5		3rd Qu.:1.0000	F5 :12	TCGA-AF-2691-11: 1	Batch 158:15
6		Max. :1.0000	AH : 7	TCGA-AF-2692-01: 1	Batch 102:14
7			(Other):22	(Other) :154	(Other) :26

Table 2: Summary of the phenotypical information of the miRNA dataset.

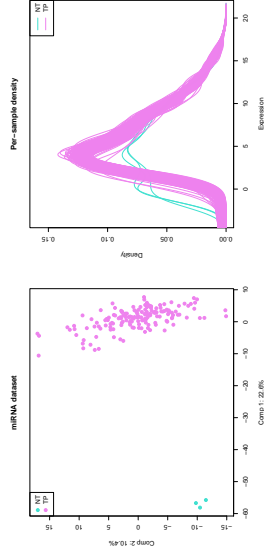


Figure 1: PCA and density plot for miRNAs.

2 Exploratory analysis of mRNA dataset

Number of mRNAs analysed 14973
 Number of samples 160

Table 3: Basic information of the mRNA dataset.

group.n	CvH	center	sample	batch	platform
1	NT: 3	Min.: -0.0000	AG :71	TCGA-AF-2687-01: 1	Batch 42 :36
2	TP:157	1st Qu.:1.0000	AF :18	TCGA-AF-2689-11: 1	Batch 122:27
3		Median :1.0000	EI :17	TCGA-AF-2690-01: 1	Batch 139:25
4		Mean :0.9812	DC :13	TCGA-AF-2691-01: 1	Batch 46 :17
5		3rd Qu.:1.0000	F5 :12	TCGA-AF-2691-11: 1	Batch 158:15
6		Max. :1.0000	AH : 7	TCGA-AF-2692-01: 1	Batch 102:14
7			(Other):22	(Other) :154	(Other) :26

Table 4: Summary of the phenotypical information of the mRNA dataset.

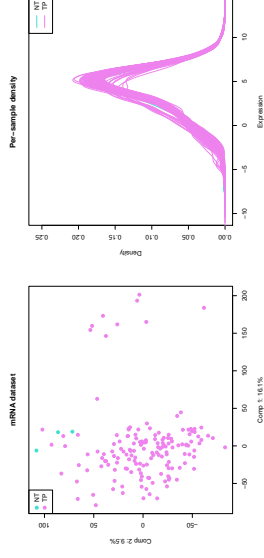


Figure 2: PCA and density plot for mRNAs.

3 Differentially expressed miRNAs

Analysis performed	Comparative used: CvH; method used: limma.
Number of differentially expressed miRNAs	225 (202 upregulated, 123 downregulated)
Number of samples	160
Criteria for selecting miRNAs	adj.pval < 1

Table 5: Basic statistics

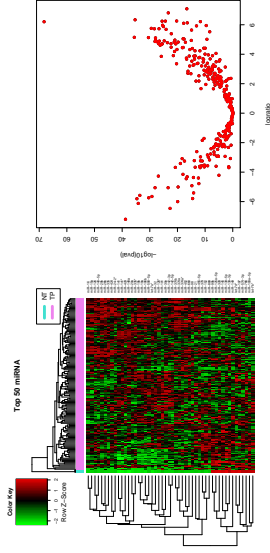


Figure 3: A) Heatmap with the top 50 most significant miRNAs (sorted by adjusted p-value). B) Volcano plot showing the selected miRNAs.

4 Differentially expressed mRNAs

Analysis performed	Comparative used: CvH; method used: limma.
Number of differentially expressed mRNAs	14973 (8595 upregulated, 6378 downregulated)
Number of samples	160
Criteria for selecting mRNAs	adj.pval < 1

Table 6: Basic statistics

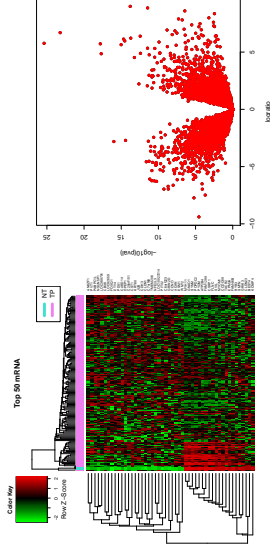


Figure 4: A) Heatmap with the top 50 most significant mRNAs (sorted by adjusted p-value). B) Volcano plot showing the selected mRNAs.

5 Correlation & intersection with databases

Number of miRNAs	325
Number of mRNAs	14973
Total miRNA-mRNA combinations	4866225
Number of samples	160

Table 7: Number of miRNAs, mRNAs and samples used for correlation.

	Number	%
Total correlations	4866225	100
Total negative correlations	2543118	52.26
Total correlations $p < 0.05$	979376	20.13
Total correlations $p < 0.01$	557475	11.46
Total correlations $\text{adj. } p < 0.05$	423296	8.7
Total correlations $\text{adj. } p < 0.01$	204266	4.2

Table 8: Basic statistics for correlation results. Correlation hypothesis: two-sided.

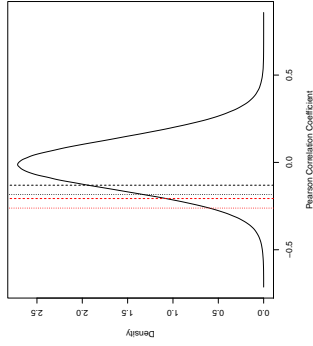


Figure 5: Density of a total of 4866225 miRNA-mRNA pairs. Dashed lines distinguish correlations whose p-value is lower than 0.05, dotted lines for 0.01. Black is for raw p-value and red for adjusted p-value.

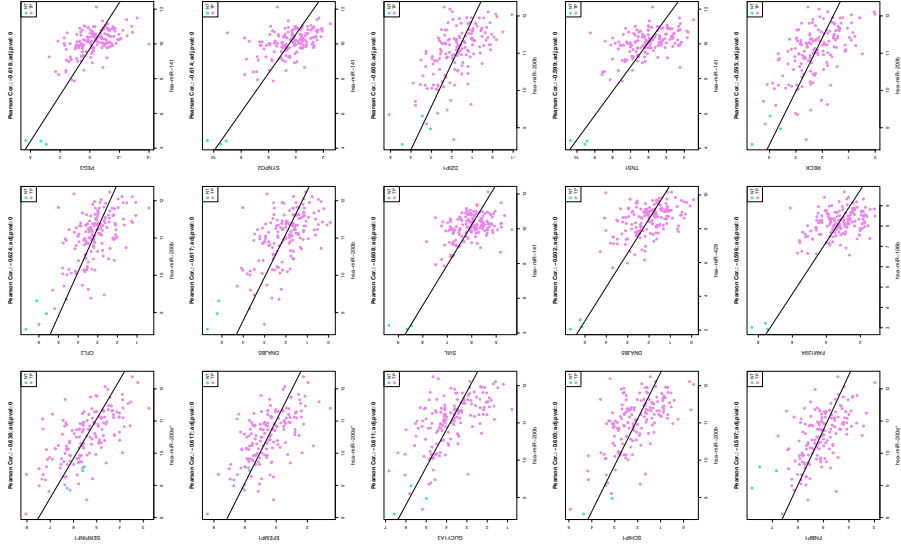


Figure 6: Plot of 15 top correlations, sorted by adjusted p-value. Databases used: microCosm_v3_L18, targetScan_v0.2.18 (each miRNA-mRNA pair has to appear at least 1 times).

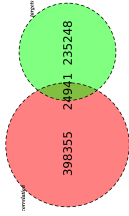


Figure 7: Venn Diagram. Left (red): number of miRNA-mRNA pairs with adjusted p-value<0.05. Right (green): number of all the theoretical miRNA-mRNA pairs reported at least 1 times in the following databases: microCosm_v5.18, targetScan_v6.2.18. Intersection: miRNA-mRNA pairs that fulfil both conditions.

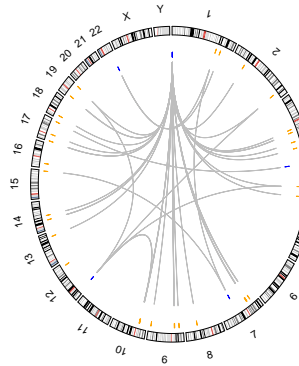


Figure 8: Circos plot for the first 45 miRNA-mRNA pairs (sorted by adjusted p-value) that have: pval-corrected<0.05 and appear at least 1 times in the following databases: microCosm_v5.18, targetScan_v6.2.18. Blue: miRNAs, Orange: target mRNAs

miRNA	mRNA	cor	adj.pval	FC.miRNA	FC.miRNA	dat.sum
hsa-miR-200a*	SERPINF1	-0.64	5.58e-15	2.32	-1.28	1
hsa-miR-200b	CFL2	-0.62	2.93e-14	4.31	-7.81	1
hsa-miR-141	PEG3	-0.62	6.11e-14	53.91	-13.95	1
hsa-miR-200a*	EFEMP1	-0.62	7.11e-14	2.32	-2.72	1
hsa-miR-200b	DNAJB5	-0.62	7.11e-14	4.31	-10.05	1
hsa-miR-141	SYNPO2	-0.61	1.03e-13	53.91	-38.72	1
hsa-miR-200b	GUCY1A3	-0.61	1.36e-13	4.31	-4.01	2
hsa-miR-141	SVIL	-0.61	2.19e-13	53.91	-6.91	1
hsa-miR-200b	DZIP1	-0.61	2.71e-13	4.31	-3.80	1
hsa-miR-200b	SCHP1	-0.61	2.88e-13	4.31	-2.99	1
hsa-miR-429	DNAJB5	-0.60	4.43e-13	75.95	-10.05	1
hsa-miR-141	TNS1	-0.60	5.68e-13	53.91	-13.80	1
hsa-miR-200a*	FNBP1	-0.60	7.15e-13	2.32	-4.91	1
hsa-miR-106b	FAM128A	-0.60	8.53e-13	35.05	-23.87	1
hsa-miR-200b	RECK	-0.59	9.47e-13	4.31	-3.56	1
hsa-miR-141	ZEB1	-0.59	1.24e-12	53.91	-5.27	1
hsa-miR-200b	RAB34	-0.59	1.25e-12	4.31	-2.05	1
hsa-miR-200a*	AEBP1	-0.59	1.31e-12	2.32	1.03	1
hsa-miR-200b*	SERPINF1	-0.59	1.33e-12	-1.55	-1.28	1
hsa-miR-429	MYLK	-0.59	1.54e-12	75.95	-16.10	1
hsa-miR-200b	BNC2	-0.59	2.62e-12	4.31	-6.84	1
hsa-miR-200a	ZEB1	-0.58	2.91e-12	21.99	-5.27	1
hsa-miR-571-3p	CBX8	-0.58	2.91e-12	-50.33	4.66	1
hsa-miR-200b	ZEB1	-0.58	3.22e-12	4.31	-5.27	2
hsa-miR-200a	IAZF1	-0.58	3.23e-12	21.99	-4.46	1
hsa-miR-200a	PEG3	-0.58	3.48e-12	21.99	-13.95	1
hsa-miR-429	KANK2	-0.58	3.71e-12	75.95	-5.55	1
hsa-miR-200b	KIAA1462	-0.58	4.07e-12	4.31	-6.87	1
hsa-miR-200b	MYLK	-0.58	4.31e-12	4.31	-16.10	1
hsa-miR-200b	IAZF1	-0.58	4.48e-12	4.31	-4.46	1
hsa-let-7a*	MYH11	-0.58	4.89e-12	18.21	-78.36	1
hsa-miR-33a	FAM120A	-0.58	4.93e-12	68.60	-23.87	1
hsa-miR-200b	TUBB6	-0.58	5.01e-12	4.31	-3.01	1
hsa-miR-200b	FRMD6	-0.58	5.78e-12	4.31	-1.96	1
hsa-miR-200a	CCDC80	-0.58	6.46e-12	21.99	-6.02	1
hsa-miR-200a	TNS1	-0.58	7.24e-12	21.99	-13.80	1
hsa-miR-200a*	CHRD	-0.58	7.73e-12	2.32	-1.69	1
hsa-let-7a*	HSPB8	-0.58	7.77e-12	18.21	-26.84	1
hsa-miR-200b	SDC2	-0.58	8.20e-12	4.31	-2.20	1
hsa-miR-200b	KANK2	-0.57	8.77e-12	4.31	-5.55	1
hsa-miR-33a	ATP2B4	-0.57	9.98e-12	68.60	-9.40	1
hsa-miR-429	CFL2	-0.57	1.18e-11	75.95	-7.81	1
hsa-miR-200b	GLI3	-0.57	1.33e-11	4.31	-7.82	2
hsa-miR-200a	SYNPO2	-0.57	1.56e-11	21.99	-38.72	1
hsa-miR-106b	SYNM	-0.57	1.67e-11	35.05	-68.22	1

Table 9: Top 45 miRNA-mRNA pairs (sorted by adjusted p-value) that have: pval-corrected<0.05 and appear at least 1 times in the following databases: microCosm_v5.18, targetScan_v6.2.18.

6 Functional analysis

6.1 Network analysis

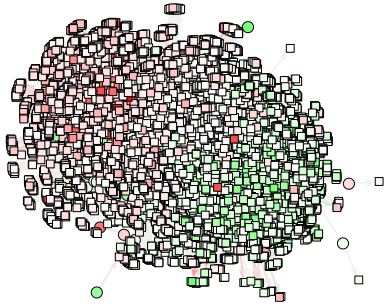


Figure 9: Network for all the miRNA-mRNA pairs that have: $p\text{-val-corrected} < 0.05$ and appear at least 1 times in the following databases: `microCosm_v5_L8`, `targetScan_v6_2_L8`. Circles represent the miRNAs, and squares the mRNA. Red fill means upregulated miRNAs/mRNAs, while green fill means downregulated miRNA/mRNAs in comparative CVH; lines indicate the miRNA-mRNA pairs, red line means positive score and green line means negative score.

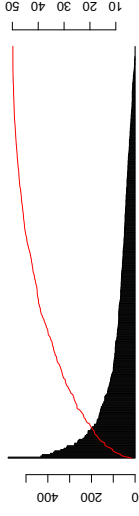


Figure 10: Barplot for miRNAs, $p\text{-val-corrected} < 0.05$ and `Targets=microCosm_v5_L8`, `targetScan_v6_2_L8`(minimum coincidences between databases:1). Red line (and right axis) represents the percentage of deregulated miRNAs that are targeted by the miRNAs.

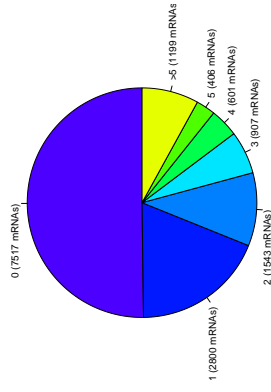


Figure 11: Pie chart representing the number of miRNAs targeting the mRNAs, $p\text{-val-corrected} < 0.05$ and `Targets=microCosm_v5_L8`, `targetScan_v6_2_L8`(minimum coincidences between databases:1).

miRNA	# targets	cum. %	targets (top 20)
hsc-miR-106a	582	3.80	AHNAK, BMP22, CYLD, PLSCR4, OSTML, XYL11, PDLIM6, ANKRD12, PGM2L1, TCF4, HEG1, ZBTB4, FCHO2, ATXN1, RAIB8B, KAT2B, AFF1, TAOK3, SNRK, C1orf84
hsc-miR-19a	432	6	SYNPO2, RNF11, ITPRI, PDE5A, NPTN, SYNM, CLIP4, MYH11, ITPKB, JAZF1, CXCL12, PRICKLE2, TNS1, SLC8A1, KIAA1462, MBNL1, FAM46B, GLTP, IAM2, IL6ST
hsc-miR-23a	425	7.80	WASL, NCOA1, MARCKS, MYO6, UBE2D1, RAPA1, SOS1, MTM1, ATP2B4, EIF4E3, CAB39, SMC5, NEGR1, WIPF2, JAZF1, BBX, RSNL1L, PEX1, CASD1, STX12
hsc-miR-19b	388	8.13	SYNPO2, NPTN, SYNM, FAM46B, CLIP4, MYH11, ITPRI, PDE5A, RNF11, CXCL12, GLTP, PRICKLE2, PDE7B, ITPKB, JAZF1, SLC8A1, MEF2C, KIAA1462, C20orf194, TNS1
hsc-miR-369-3p	374	9.54	NCOA1, AKAP6, MBNL1, DDXDC1, PLEKHM8, ITPRI, ARHGAP5, HLF, MKL2, ANKRD12, PRKAB2, ABBBP, Csof41, TMEM131, AFF1, CREB1, BMP12, SASH1, RALGPS2, SORBS1
hsc-miR-30e	368	10.55	AHNAK, AMOTL1, LPP, LIFR, NEGR1, NECAB1, MYH11, CALD1, TEAD1, GNAO1, KLF9, MBNL1, CFL2, SAMD4A, GNG2, NRP2, PDE5A, NCS1, FYCOL, ITGA5
hsc-miR-17	341	11.18	FAM129A, KCNMA1, MYLK, CLIP4, GSN, TMEM100, BAALC, ITPKB, C1orf94, STGALNAC6, SYNM, CNN1, PSD, SYNE1, CFL2, ATP1A2, FBXL22, FGL2, BNC2, SPG20
hsc-miR-106b	326	11.29	FAM129A, SYNM, ATP1A2, TNS1, CNN1, LMO3, AHNAK, ITPKB, KCNMA1, TMEM100, CLIP4, PSD, BAALC, CFL2, JAZF1, BVES, KIAA1462, CALD1, HSD17B6, STGALNAC6
hsc-miR-27a	314	12.15	GRK3, SDPR, LIFR, ADAMTS1, AFF3, ABCA8, RBPMS2, FMBP1, NCOA1, ATP1A2, SFRP1, PRICKLE2, SPARCL1, SORBS1, PHLPP2, RCAN2, MYOCD, MEF2C, EGFR, VAV1L
hsc-miR-200b	311	13.24	CFL2, DNAJB5, GUCY1A3, DZIP1, SCHIP1, RECK, RAB34, BNC2, ZEB1, KIAA1462, MYLK, JAZF1, TUBB6, FRMD6, SDCC2, KANK2, GIL3, TIMP2, SGCE, NR3C1

Table 10: Top 10 miRNA with more targets (each miRNA-mRNA pair has pval-corrected<0.05 and appears at least 1 times in the following databases: microCosm.v5.18, targetScan.v6.2.18). miRNAs in red are upregulated in CvH. miRNAs in green are downregulated in CvH.

miRNA	# miRNAs	miRNAs (top 20)
QKI	43	hsc-miR-200b, hsc-miR-200a, hsc-miR-141, hsc-miR-33b, hsc-miR-33a, hsc-miR-106a, hsc-miR-577, hsc-miR-493, hsc-miR-19a, hsc-miR-106b, hsc-miR-362-5p, hsc-miR-191, hsc-miR-27a, hsc-miR-130b, hsc-miR-135b, hsc-miR-17, hsc-miR-576-5p, hsc-miR-345
BNC2	38	hsc-miR-200b, hsc-miR-200a, hsc-miR-141, hsc-miR-429, hsc-miR-200c, hsc-miR-577, hsc-miR-106b, hsc-miR-19a, hsc-miR-17, hsc-miR-19b, hsc-miR-596-5p, hsc-miR-130b, hsc-miR-20a, hsc-miR-576-5p, hsc-miR-30e, hsc-miR-532-5p, hsc-miR-96, hsc-miR-552, hsc-miR-106a
KCNMA1	37	hsc-miR-17, hsc-miR-20a, hsc-miR-335*, hsc-miR-106b, hsc-miR-29a*, hsc-miR-20a*, hsc-miR-29b, hsc-miR-450b-5p, hsc-miR-135b, hsc-miR-224, hsc-miR-33a, hsc-miR-93, hsc-miR-576-5p, hsc-miR-16-2*, hsc-miR-584, hsc-miR-942, hsc-miR-29a, hsc-miR-629, hsc-miR-382, hsc-miR-21*
NRP2	32	hsc-miR-200b, hsc-miR-200a, hsc-miR-141, hsc-miR-429, hsc-miR-30e, hsc-miR-19a, hsc-miR-106b, hsc-miR-577, hsc-miR-19b, hsc-miR-17, hsc-miR-532-5p, hsc-miR-130b, hsc-miR-942, hsc-miR-140a, hsc-miR-106a, hsc-miR-188-5p, hsc-let-7g*, hsc-miR-20a, hsc-miR-16, hsc-miR-425*
MBNL1	31	hsc-miR-369-3p, hsc-miR-19a, hsc-miR-500-5p, hsc-miR-30c, hsc-miR-19b, hsc-miR-18a, hsc-miR-199b-3p, hsc-miR-199a-3p, hsc-miR-130b, hsc-miR-889, hsc-miR-590-3p, hsc-miR-203, hsc-miR-141, hsc-miR-21, hsc-miR-301a, hsc-miR-450b-5p, hsc-miR-135b, hsc-miR-382, hsc-miR-223, hsc-miR-552
PDE4D	31	hsc-miR-18a, hsc-miR-552, hsc-miR-130b, hsc-miR-369-3p, hsc-miR-203, hsc-miR-50e, hsc-miR-692, hsc-miR-53a, hsc-miR-193, hsc-miR-371b, hsc-miR-194, hsc-miR-432, hsc-miR-374a, hsc-miR-25, hsc-miR-409-5p, hsc-miR-454, hsc-miR-339-3p, hsc-miR-130a, hsc-miR-30b, hsc-miR-32
TSHZ3	30	hsc-miR-200b, hsc-miR-200a, hsc-miR-576-5p, hsc-miR-429, hsc-miR-141, hsc-miR-19b, hsc-miR-17, hsc-miR-106b, hsc-let-7a*, hsc-miR-19b, hsc-miR-20a, hsc-miR-194, hsc-miR-148a*, hsc-miR-20a*, hsc-miR-325, hsc-miR-590-3p, hsc-miR-106a, hsc-miR-93
HLF	29	hsc-miR-369-3p, hsc-miR-106b, hsc-miR-19a, hsc-miR-429, hsc-miR-141, hsc-miR-19b, hsc-miR-20a, hsc-miR-590-5p, hsc-miR-30e, hsc-miR-374a, hsc-miR-337-3p, hsc-miR-130b, hsc-miR-20a, hsc-miR-152, hsc-miR-200b, hsc-miR-18a, hsc-miR-29b, hsc-miR-17, hsc-miR-183, hsc-miR-148a
IGF1	29	hsc-miR-18a, hsc-miR-625, hsc-miR-130b, hsc-miR-19a, hsc-miR-26a-2*, hsc-miR-27a, hsc-miR-19b, hsc-miR-577, hsc-miR-16, hsc-miR-192, hsc-miR-90b-3p, hsc-miR-629, hsc-miR-24-2*, hsc-miR-30b, hsc-miR-30e, hsc-miR-15a, hsc-miR-222, hsc-miR-215, hsc-miR-362-5p, hsc-miR-148a
ITPRI	29	hsc-miR-19a, hsc-miR-19b, hsc-let-7a*, hsc-miR-19b, hsc-miR-203, hsc-miR-429, hsc-miR-369-3p, hsc-miR-200b, hsc-miR-32, hsc-miR-22*, hsc-miR-576-5p, hsc-miR-96, hsc-miR-126, hsc-miR-301a, hsc-miR-124, hsc-miR-136*, hsc-miR-192*, hsc-miR-628-5p, hsc-miR-26b, hsc-miR-450a

Table 11: Top 10 miRNA with more miRNAs targeting them (each miRNA-mRNA pair has pval-corrected<0.05 and appears at least 1 times in the following databases: microCosm.v5.18, targetScan.v6.2.18). miRNAs in red are upregulated in CvH. miRNAs in green are downregulated in CvH.

6.2 GO analysis

GOBPID	Term	Count	Size	ExpCount	OddsRatio	fdr	Pvalue
GO:0050794	regulation of cellular process	3015	7941	3259.02	1.50	1.33e-29	1.28e-33
GO:0050789	regulation of biological process	3784	8398	3446.57	1.48	3.72e-27	7.15e-31
GO:0050007	biological regulation	3982	8895	3650.54	1.49	4.24e-27	1.22e-30
GO:0048518	positive regulation of biological process	1860	3817	1566.51	1.53	7.68e-26	2.95e-29
GO:0035556	intracellular signal transduction	1055	2009	824.50	1.71	7.83e-26	3.76e-29
GO:0048522	positive regulation of cellular process	1673	3397	1394.14	1.55	2.47e-25	1.43e-28
GO:0016043	cellular component organization	2200	4619	1805.66	1.48	4.56e-25	3.11e-28
GO:0051716	cellular response to stimulus	2487	5295	2173.09	1.47	4.56e-25	3.02e-28
GO:0006464	cellular protein modification process	1310	2583	1060.07	1.61	4.56e-25	4.38e-28
GO:0038211	protein modification process	1310	2583	1060.07	1.61	4.56e-25	4.38e-28

Table 12: Biological Process . Options used: mRNAs that are present in a mRNA-mRNA pair that has adjusted-pval cutoff ≤ 0.05 ; that also appears at least 1 times (databases: microCosm_v5.1.8, targetScan_v6.2.18); organism: human.

GOCCID	Term	Count	Size	ExpCount	OddsRatio	fdr	Pvalue
GO:0044424	intracellular part	5256	12258	4750.75	2.00	1.98e-73	1.61e-76
GO:0005622	intracellular	5296	12396	4804.23	1.99	2.50e-71	4.05e-74
GO:0005737	cytoplasm	4174	9342	3620.61	1.78	1.83e-68	4.47e-71
GO:0043227	membrane-bounded organelle	4524	10309	4030.27	1.72	1.64e-57	5.34e-60
GO:0043226	organelle	4815	11215	4346.52	1.74	1.58e-55	6.42e-58
GO:0043231	intracellular membrane-bounded organelle	4183	9551	3701.61	1.66	2.68e-52	1.31e-54
GO:0043229	intracellular organelle	4573	10617	4114.76	1.67	1.85e-50	1.05e-52
GO:0044444	cytoplasmic part	3109	6864	2660.23	1.59	1.83e-45	1.19e-47
GO:0005529	cytosol	1262	2545	986.35	1.69	1.52e-31	1.11e-33
GO:0044446	intracellular organelle part	2795	6274	2431.57	1.47	9.55e-31	7.76e-33

Table 13: Cellular Component . Options used: mRNAs that are present in a mRNA-mRNA pair that has adjusted-pval cutoff ≤ 0.05 ; that also appears at least 1 times (databases: microCosm_v5.1.8, targetScan_v6.2.18); organism: human.

GOMFID	Term	Count	Size	ExpCount	OddsRatio	fdr	Pvalue
GO:0005515	protein binding	3683	7904	3143.00	1.83	2.95e-69	1.03e-72
GO:0005488	binding	5151	12075	4801.59	1.88	5.87e-46	4.10e-49
GO:0019899	enzyme binding	616	1175	467.24	1.75	4.24e-17	4.44e-20
GO:0043167	ion binding	2538	5793	2303.57	1.31	5.20e-13	7.25e-16
GO:0004672	protein kinase activity	315	565	224.67	1.96	2.25e-12	3.92e-15
GO:0008092	cytoskeletal protein binding	376	704	279.94	1.79	2.27e-11	4.75e-14
GO:0019904	protein domain specific binding	295	532	211.55	1.93	3.09e-11	7.54e-14
GO:0016773	phosphotransferase activity, alcohol group as acceptor	361	679	270.00	1.77	1.28e-10	3.58e-13
GO:0043168	antion binding	1149	2493	991.33	1.36	4.17e-10	1.31e-12
GO:0016301	kinase activity	380	732	291.08	1.68	1.84e-09	6.42e-12

Table 14: Molecular Function . Options used: mRNAs that are present in a mRNA-mRNA pair that has adjusted-pval cutoff ≤ 0.05 ; that also appears at least 1 times (databases: microCosm_v5.1.8, targetScan_v6.2.18); organism: human.

KEGGID	Term	Count	Size	ExpCount	OddsRatio	fdr	Pvalue
05200	Pathways in cancer	182	314	129.57	2.04	1.47e-07	6.50e-10
04510	Focal adhesion	118	192	79.23	2.34	8.84e-07	7.83e-09
04360	Axon guidance	78	125	51.58	2.41	8.66e-05	1.19e-06
04144	Endocytosis	113	196	80.88	1.99	8.66e-05	1.87e-06
04514	Cell adhesion molecules (CAMs)	79	128	52.82	2.34	8.06e-05	1.92e-06
04142	Lysosome	70	117	48.28	2.15	1.25e-03	3.32e-05
04810	Regulation of actin cytoskeleton	111	203	83.76	1.75	1.74e-03	5.96e-05
04350	TGF-beta signaling pathway	52	83	34.25	2.42	1.74e-03	6.17e-05
05211	Renal cell carcinoma	43	66	27.23	2.69	1.76e-03	7.04e-05
04520	Adherens junction	45	70	28.88	2.59	1.76e-03	7.80e-05

Table 15: Kegg Pathways . Options used: mRNAs that are present in a mRNA-mRNA pair that has adjusted-pval cutoff <0.05; that also appears at least 1 times (databases: microCosm_v5.18, targetScan_v6.2.18); organism: human.

A.5 Study 1: Stomach adenocarcinoma

Default miRComb output

/home/invilla/Baixades/TCGA/stomach

May 13, 2015

1 Exploratory analysis of miRNA dataset

Number of miRNAs analysed 330
 Number of samples 443

Table 1: Basic information of the miRNA dataset.

group.n	CvH	center	sample	batch
1	NT: 37	Min: -0.0000	BR :138	TCGA-3M-AB16-01: 1 Batch 220: 58
2	TP:406	1st Qu.:1.0000	VQ : 66	TCGA-3M-AB47-01: 1 Batch 427: 58
3		Median :1.0000	CG : 43	TCGA-B7-5816-01: 1 Batch 269: 52
4		Mean :0.9165	HU : 41	TCGA-B7-5818-01: 1 Batch 57 : 31
5		3rd Qu.:1.0000	D7 : 40	TCGA-B7-A5TL-01: 1 Batch 95 : 29
6		Max. :1.0000	CD : 27	TCGA-B7-A5TL-01: 1 Batch 242: 28
7			(Other): 88	(Other) :437

Table 2: Summary of the phenotypical information of the miRNA dataset.

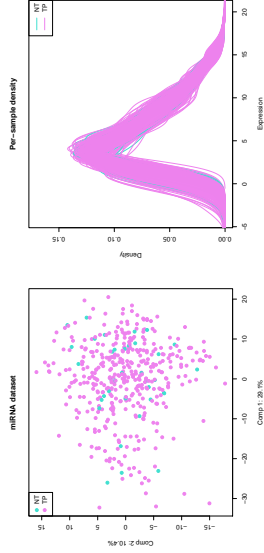


Figure 1: PCA and density plot for miRNAs.

2 Exploratory analysis of mRNA dataset

Number of mRNAs analysed 18565
 Number of samples 443

Table 3: Basic information of the mRNA dataset.

group.n	CvH	center	sample	batch
1	NT: 37	Min: -0.0000	BR :138	TCGA-3M-AB16-01: 1 Batch 220: 58
2	TP:406	1st Qu.:1.0000	VQ : 66	TCGA-3M-AB47-01: 1 Batch 427: 58
3		Median :1.0000	CG : 43	TCGA-B7-5816-01: 1 Batch 269: 52
4		Mean :0.9165	HU : 41	TCGA-B7-5818-01: 1 Batch 57 : 31
5		3rd Qu.:1.0000	D7 : 40	TCGA-B7-A5TL-01: 1 Batch 95 : 29
6		Max. :1.0000	CD : 27	TCGA-B7-A5TL-01: 1 Batch 242: 28
7			(Other): 88	(Other) :437

Table 4: Summary of the phenotypical information of the mRNA dataset.

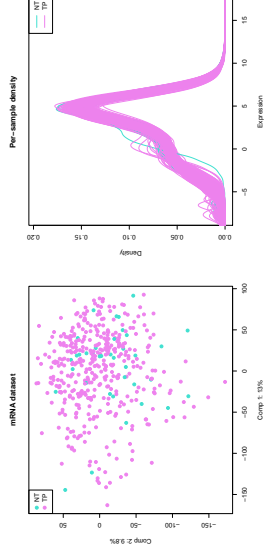


Figure 2: PCA and density plot for mRNAs.

3 Differentially expressed miRNAs

Analysis performed	Comparative used: CvH; method used: limma.
Number of differentially expressed miRNAs	330 (150 upregulated, 180 downregulated)
Number of samples	443
Criteria for selecting miRNAs	adj.pval < 1

Table 5: Basic statistics

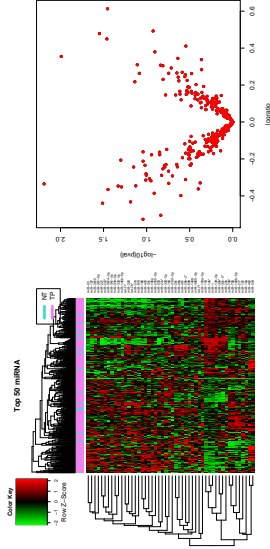


Figure 3: A) Heatmap with the top 50 most significant miRNAs (sorted by adjusted p-value). B) Volcano plot showing the selected miRNAs.

4 Differentially expressed mRNAs

Analysis performed	Comparative used: CvH; method used: limma.
Number of differentially expressed mRNAs	18565 (10588 upregulated, 7977 downregulated)
Number of samples	443
Criteria for selecting mRNAs	adj.pval < 1

Table 6: Basic statistics

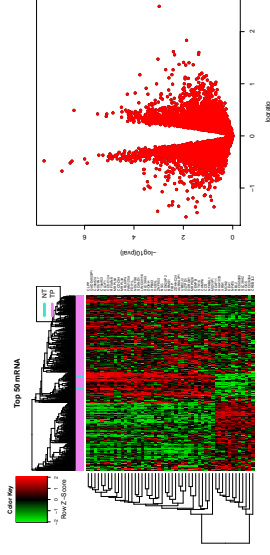


Figure 4: A) Heatmap with the top 50 most significant mRNAs (sorted by adjusted p-value). B) Volcano plot showing the selected mRNAs.

5 Correlation & intersection with databases

Number of miRNAs	330
Number of mRNAs	18565
Total miRNA-mRNA combinations	6126450
Number of samples	443

Table 7: Number of miRNAs, mRNAs and samples used for correlation.

	Number	%
Total correlations	6126450	100
Total negative correlations	2828605	46.17
Total correlations $p < 0.05$	1736533	28.25
Total correlations $p < 0.01$	1367300	22.32
Total correlations adj. $p < 0.05$	1390596	22.7
Total correlations adj. $p < 0.01$	1108323	18.09

Table 8: Basic statistics for correlation results. Correlation hypothesis: two-sided.

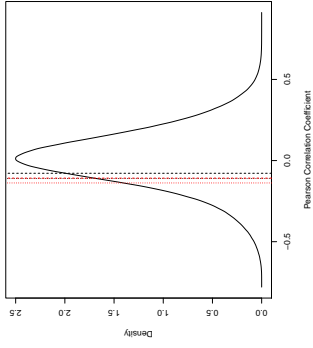


Figure 5: Density of a total of 6126450 miRNA-mRNA pairs. Dashed lines distinguish correlations whose p-value is lower than 0.05, dotted lines for 0.01. Black is for raw p-value and red for adjusted p-value.

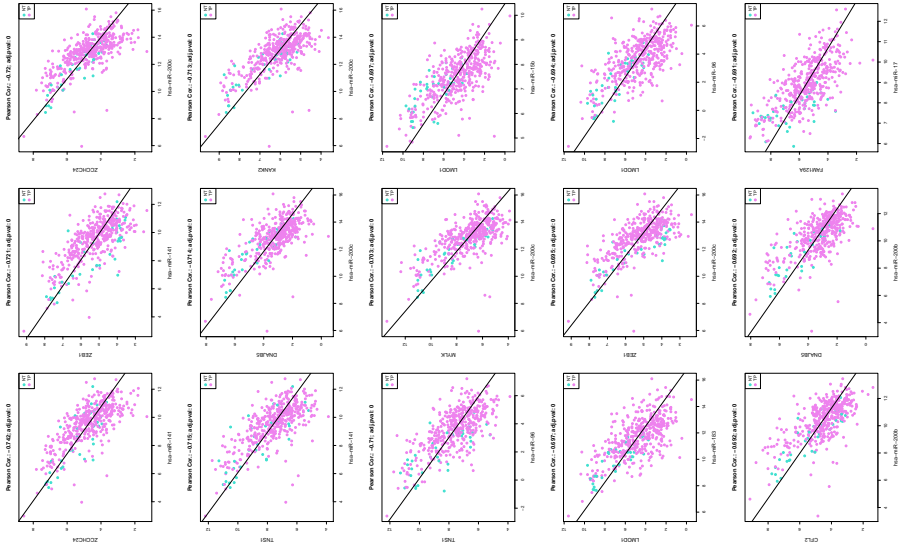


Figure 6: Plot of 15 top correlations, sorted by adjusted p-value. Databases used: microCosm_v3_L8, targetScan_v0.2_L18 (each miRNA-mRNA pair has to appear at least 1 times).

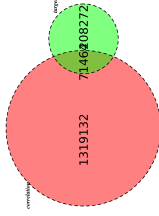


Figure 7: Venn Diagram. Left (red): number of miRNA-mRNA pairs with adjusted p-value < 0.05. Right (green): number of all the theoretical miRNA-mRNA pairs reported at least 1 times in the following databases: microCosm_v5.18, targetScan_v6.2.18. Intersection: miRNA-mRNA pairs that fulfil both conditions.

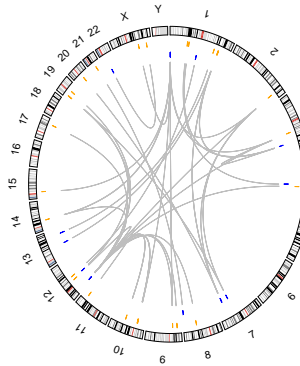


Figure 8: Circos plot for the first 45 miRNA-mRNA pairs (sorted by adjusted p-value) that have: pval-corrected < 0.05 and appear at least 1 times in the following databases: microCosm_v5.18, targetScan_v6.2.18. Blue: miRNAs, Orange: target mRNAs

miRNA	mRNA	cor	adj_pval	FC_miRNA	FC_mRNA	dist.sum
hsa-miR-141	ZCCHC24	-0.74	1.01e-72	1.02	1.03	1
hsa-miR-141	ZEB1	-0.72	5.00e-67	1.02	1.24	1
hsa-miR-200c	ZCCHC24	-0.72	1.46e-66	-1.01	1.03	1
hsa-miR-141	TNSI	-0.72	2.39e-65	1.02	1.10	1
hsa-miR-200c	DNAJB5	-0.71	3.50e-65	-1.01	1.03	1
hsa-miR-200c	KANK2	-0.71	8.71e-65	-1.01	1.00	1
hsa-miR-96	TNSI	-0.71	4.03e-64	-1.05	1.10	1
hsa-miR-200c	MYLK	-0.70	1.49e-62	-1.01	1.15	1
hsa-miR-15b	LMOD1	-0.70	5.93e-61	-1.12	-1.08	1
hsa-miR-183	LMOD1	-0.70	5.93e-61	1.03	-1.08	1
hsa-miR-200c	ZEB1	-0.69	1.32e-60	-1.01	1.24	2
hsa-miR-36	LMOD1	-0.69	2.02e-60	-1.05	-1.08	1
hsa-miR-200b	CFL2	-0.69	5.90e-60	-1.01	1.15	1
hsa-miR-200b	DNAJB5	-0.69	7.89e-60	-1.01	1.03	1
hsa-miR-17	FAM129A	-0.69	1.38e-59	1.04	1.00	1
hsa-miR-15b	MYLK	-0.69	7.07e-59	-1.12	1.15	1
hsa-miR-151-5p	NEXN	-0.69	1.45e-58	-1.10	1.02	1
hsa-miR-429	CFL2	-0.69	1.80e-58	-1.06	1.15	1
hsa-miR-141	DIXDC1	-0.69	1.94e-58	1.02	1.15	1
hsa-miR-130b*	FHL1	-0.68	1.45e-57	-1.04	1.14	1
hsa-miR-942	MSRB3	-0.68	2.22e-57	-1.13	1.25	1
hsa-miR-151-5p	JAM2	-0.68	6.62e-57	1.04	-1.00	1
hsa-miR-96	MYL9	-0.68	9.62e-57	-1.05	-1.08	1
hsa-miR-141	TSHZ3	-0.68	2.25e-56	1.02	1.13	2
hsa-miR-16	LMOD1	-0.68	2.42e-56	1.03	-1.08	1
hsa-miR-141	SVIL	-0.67	4.26e-56	1.02	1.23	1
hsa-miR-200a	TNSI	-0.67	4.96e-56	-1.01	1.10	1
hsa-miR-200c	CFL2	-0.67	5.07e-56	-1.01	1.15	1
hsa-miR-141	MAP3K3	-0.67	5.13e-56	1.02	1.06	1
hsa-miR-141	SYNP2	-0.67	5.34e-56	1.02	1.16	1
hsa-miR-141	BNC2	-0.67	5.72e-56	1.02	1.31	1
hsa-miR-15a	LMOD1	-0.67	1.24e-55	1.06	-1.08	1
hsa-miR-106b	TNSI	-0.67	1.34e-55	-1.04	1.10	1
hsa-miR-429	MYLK	-0.67	1.57e-55	-1.06	1.15	1
hsa-miR-141	NECAB1	-0.67	1.87e-55	1.02	1.10	1
hsa-miR-200b*	GPRASP1	-0.67	3.08e-55	-1.11	1.13	1
hsa-miR-429	DNAJB5	-0.67	3.59e-55	-1.06	1.03	1
hsa-miR-200c	POU6F1	-0.67	1.09e-54	-1.01	1.07	1
hsa-miR-200a	SVIL	-0.67	1.14e-54	-1.01	1.23	1
hsa-miR-148b	CNN1	-0.67	1.73e-54	-1.05	-1.15	1
hsa-miR-200b	NDN	-0.67	2.37e-54	-1.01	-1.04	2
hsa-miR-15b	PRELP	-0.67	3.45e-54	-1.12	1.11	1
hsa-miR-576-5p	NEGR1	-0.67	3.77e-54	1.04	1.06	1
hsa-miR-106b	CNN1	-0.66	5.20e-54	-1.04	-1.15	2
hsa-miR-16	MYLK	-0.66	5.89e-54	1.03	1.15	1

Table 9: Top 45 miRNA-mRNA pairs (sorted by adjusted p-value) that have: pval-corrected < 0.05 and appear at least 1 times in the following databases: microCosm_v5.18, targetScan_v6.2.18.

6 Functional analysis

6.1 Network analysis



Figure 9: Network for all the miRNA-mRNA pairs that have: $p\text{-val-corrected} < 0.05$ and appear at least 1 times in the following databases: `microCosm_v5_L8`, `targetScan_v6_2_L8`. Circles represent the miRNAs, and squares the mRNA. Red fill means upregulated miRNAs/mRNAs, while green fill means downregulated miRNA/mRNAs in comparative CVH; lines indicate the miRNA-mRNA pairs, red line means positive score and green line means negative score.

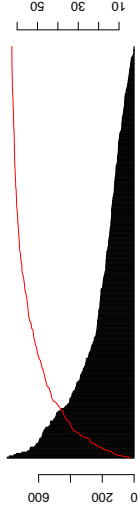


Figure 10: Barplot for miRNAs, $p\text{-val-corrected} < 0.05$ and `Targets=microCosm_v5_L8`, `targetScan_v6_2_L8` (minimum coincidences between databases:1). Red line (and right axis) represents the percentage of deregulated miRNAs that are targeted by the miRNAs.

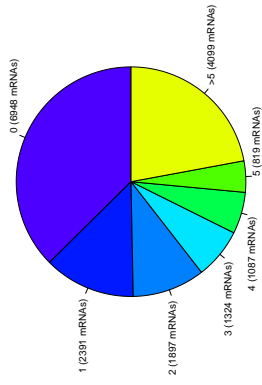


Figure 11: Pie chart representing the number of miRNAs targeting the miRNAs, $p\text{-val-corrected} < 0.05$ and `Targets=microCosm_v5_L8`, `targetScan_v6_2_L8` (minimum coincidences between databases:1).

miRNA	#miRNAs	miRNAs (top 20)
FOXP2	74	hsa-miR-15b, hsa-miR-625, hsa-miR-141, hsa-miR-16, hsa-miR-130b*, hsa-miR-590-3p, hsa-miR-7-1*, hsa-miR-576-5p, hsa-miR-15a, hsa-miR-222, hsa-miR-502-3p, hsa-miR-96, hsa-miR-200a, hsa-miR-590-5p, hsa-let-7d, hsa-miR-671-5p, hsa-miR-7, hsa-miR-34a, hsa-miR-197, hsa-miR-182
TNRC6B	67	hsa-miR-130b, hsa-miR-26b, hsa-miR-144, hsa-miR-32, hsa-miR-590-3p, hsa-miR-106b, hsa-miR-16, hsa-miR-7, hsa-miR-363-5p, hsa-miR-141, hsa-miR-124, hsa-miR-183, hsa-miR-29b, hsa-miR-29a, hsa-miR-590-5p, hsa-miR-455-3p, hsa-miR-3614-5p, hsa-miR-15b, hsa-miR-330-3p, hsa-miR-18a
QKI	65	hsa-miR-129, hsa-miR-141, hsa-miR-200a, hsa-miR-200b, hsa-miR-19b, hsa-miR-96, hsa-miR-322-3p, hsa-miR-194, hsa-miR-130b, hsa-miR-106b, hsa-miR-38a, hsa-miR-92a, hsa-miR-106b, hsa-miR-345, hsa-miR-17, hsa-miR-93*, hsa-miR-500b, hsa-miR-148a, hsa-miR-577, hsa-miR-19a
ROSA	65	hsa-miR-141, hsa-miR-106b, hsa-miR-18a, hsa-miR-17, hsa-miR-200a, hsa-miR-15b, hsa-miR-38a, hsa-miR-363-5p, hsa-miR-19b, hsa-miR-92a, hsa-miR-576-5p, hsa-miR-590-3p, hsa-miR-502-3p, hsa-miR-671-5p, hsa-miR-25, hsa-miR-15a, hsa-miR-183, hsa-miR-19a, hsa-miR-93, hsa-miR-501-3p, hsa-miR-141, hsa-miR-200c, hsa-miR-130b, hsa-miR-106b, hsa-miR-429, hsa-miR-96, hsa-miR-200a, hsa-miR-592, hsa-miR-200b, hsa-miR-7, hsa-miR-17, hsa-miR-19b, hsa-miR-183, hsa-miR-200c*, hsa-miR-191, hsa-miR-577, hsa-miR-590-5p, hsa-miR-7-1*, hsa-miR-182, hsa-miR-19a
AFP4	57	hsa-miR-30b, hsa-miR-374b, hsa-miR-106b, hsa-miR-92a, hsa-miR-29a, hsa-miR-129, hsa-miR-32, hsa-miR-197, hsa-miR-590-3p, hsa-miR-455-5p, hsa-miR-502-3p, hsa-miR-15b, hsa-miR-200c, hsa-miR-17, hsa-miR-25, hsa-miR-505, hsa-miR-374a, hsa-miR-660, hsa-miR-15a, hsa-miR-331-5p, hsa-miR-590-3p, hsa-miR-16, hsa-miR-141, hsa-miR-96, hsa-miR-942, hsa-miR-32, hsa-miR-183, hsa-miR-7, hsa-miR-1266, hsa-miR-15b, hsa-miR-182, hsa-miR-124, hsa-miR-197b, hsa-miR-576-5p, hsa-miR-19b, hsa-miR-18a, hsa-miR-425, hsa-miR-330-5p, hsa-miR-577, hsa-miR-200a
LPP	57	hsa-miR-15b, hsa-miR-130b, hsa-miR-16, hsa-miR-200c, hsa-miR-38a, hsa-miR-429, hsa-miR-148b, hsa-miR-455-3p, hsa-miR-32, hsa-miR-200b, hsa-miR-19b, hsa-miR-15a, hsa-miR-92a, hsa-miR-19a, hsa-let-7d, hsa-miR-192, hsa-miR-25, hsa-miR-107, hsa-miR-29b, hsa-miR-186
RUNX1T1	57	hsa-miR-92a, hsa-miR-942, hsa-miR-19b, hsa-miR-130b, hsa-miR-197, hsa-miR-19a, hsa-miR-500b, hsa-miR-590-3p, hsa-miR-27a, hsa-miR-362-5p, hsa-miR-25, hsa-miR-96, hsa-miR-500b, hsa-miR-590-3p, hsa-miR-3617-3p, hsa-miR-128, hsa-let-7d, hsa-miR-26b, hsa-miR-141, hsa-miR-374b, hsa-miR-192
CPBE4	55	hsa-miR-130b, hsa-miR-590-5p, hsa-miR-141, hsa-miR-32, hsa-miR-942, hsa-miR-93*, hsa-miR-191, hsa-miR-7, hsa-miR-38a, hsa-miR-494, hsa-miR-25, hsa-miR-200c, hsa-miR-429, hsa-miR-19a, hsa-miR-19b, hsa-miR-301a, hsa-miR-330-3p, hsa-miR-577, hsa-miR-92a

Table 11: Top 10 miRNA with more miRNAs targeting them (each miRNA-mRNA pair has pval-corrected<0.05 and appears at least 1 times in the following databases: miRCoCosm.v5.18, targetScan.v6.2.18). miRNAs in red are upregulated in CvH. miRNAs in green are downregulated in CvH.

miRNA	#targets	cum. % targets (top 20)
hsa-miR-29a	798	4.3
hsa-miR-29c	784	5.93
hsa-miR-195	755	9.28
hsa-miR-30b	740	12.21
hsa-miR-29b	705	12.78
hsa-miR-497	671	13.36
hsa-miR-429	656	15.55
hsa-miR-26a	654	17.44
hsa-miR-26b	652	18.24
hsa-miR-144	624	19.69

Table 10: Top 10 miRNA with more targets (each miRNA-mRNA pair has pval-corrected<0.05 and appears at least 1 times in the following databases: miRCoCosm.v5.18, targetScan.v6.2.18). miRNAs in red are upregulated in CvH. miRNAs in green are downregulated in CvH.

6.2 GO analysis

GOBPID	Term	Count	Size	ExpCount	OddsRatio	fdr	Pvalue
GO:000987	cellular process	8586	12932	8354.90	1.91	6.41e-31	5.53e-35
GO:0006464	cellular protein modifi- cation process	1931	2583	1668.78	1.78	1.51e-30	3.90e-34
GO:0030211	protein modification process	1931	2583	1668.78	1.78	1.51e-30	3.90e-34
GO:0016043	cellular component or- ganization	3304	4619	2984.17	1.58	5.28e-30	1.82e-33
GO:0071840	cellular component or- ganization or biogenesis	3365	4715	3046.19	1.57	1.57e-29	6.78e-33
GO:0043412	macromolecule modifi- cation	1994	2685	1734.68	1.74	4.11e-29	1.13e-32
GO:0044267	cellular protein metabolic process	2456	3370	2177.24	1.64	3.37e-28	2.03e-31
GO:0006793	phosphorus metabolic process	2061	2809	1814.79	1.65	3.34e-25	2.51e-28
GO:0050794	regulation of cellular process	5447	7941	5130.40	1.46	3.34e-25	2.50e-28
GO:0044237	cellular metabolic pro- cess	6031	8855	5720.90	1.47	4.77e-25	4.12e-28

Table 12: Biological Process . Options used: mRNAs that are present in a mRNA-mRNA pair that has adjusted-pval cutoff <0.05; that also appears at least 1 times (databases: microCosm_v5.18, targetScan_v6.2.18); organism: human.

GOCCID	Term	Count	Size	ExpCount	OddsRatio	fdr	Pvalue
GO:0044421	intracellular part	8246	12258	7540.33	2.46	2.62e-137	1.95e-140
GO:0005622	intracellular	8322	12396	7625.22	2.47	7.87e-137	1.17e-139
GO:0005737	cytoplasm	6465	9342	5746.60	2.10	7.01e-115	1.56e-117
GO:0043226	organelle	7566	11215	6898.74	2.13	1.04e-109	3.10e-112
GO:0043227	membrane-bounded or- ganelle	7078	10389	6396.79	2.08	6.74e-108	2.50e-110
GO:0043229	intracellular organelle	7191	10617	6530.89	2.05	9.60e-103	4.28e-105
GO:0043231	intracellular membrane- bounded organelle	6546	9551	5875.16	2.00	5.33e-101	2.78e-103
GO:0044444	cytoplasmic part	4827	6864	4222.29	1.91	2.79e-84	1.66e-86
GO:0044446	intracellular organelle part	4392	6274	3839.36	1.80	1.26e-67	8.44e-70
GO:0044422	organelle part	4502	6449	3907.01	1.80	1.78e-67	1.33e-69

Table 13: Cellular Component . Options used: mRNAs that are present in a mRNA-mRNA pair that has adjusted-pval cutoff <0.05; that also appears at least 1 times (databases: microCosm_v5.18, targetScan_v6.2.18); organism: human.

GOMFID	Term	Count	Size	ExpCount	OddsRatio	fdr	Pvalue
GO:0005515	protein binding	5593	7904	4974.68	2.03	2.25e-93	6.53e-97
GO:0005488	binding	8058	12075	7599.85	2.18	5.07e-77	2.95e-80
GO:0043167	ion binding	3986	5793	3646.04	1.51	1.80e-29	1.57e-32
GO:0043168	anion binding	1821	2493	1569.06	1.74	1.28e-28	1.49e-31
GO:1901265	nucleoside phosphate binding	1600	2255	1419.27	1.78	1.52e-28	2.24e-31
GO:0000166	nucleotide binding	1659	2254	1418.64	1.78	1.53e-28	2.07e-31
GO:0017076	purine nucleotide bind- ing	1348	1819	1144.86	1.80	1.77e-24	3.61e-27
GO:0097367	carbohydrate derivative binding	1548	2115	1331.15	1.73	1.87e-24	4.34e-27
GO:0003824	catalytic activity	3571	5197	3270.93	1.47	2.28e-24	5.96e-27
GO:0032555	purine ribonucleotide binding	1334	1800	1132.90	1.80	2.36e-24	6.84e-27

Table 14: Molecular Function . Options used: mRNAs that are present in a mRNA-mRNA pair that has adjusted-pval cutoff <0.05; that also appears at least 1 times (databases: microCosm_v5.18, targetScan_v6.2.18); organism: human.

KEGGID	Term	Count	Size	ExpCount	OddsRatio	fdr	P value
05200	Pathways in cancer	253	314	201.06	2.43	6.42e-09	3.88e-11
04510	Focal adhesion	163	192	122.94	3.26	6.42e-09	5.65e-11
04722	Neurotrophin signaling pathway	105	125	80.04	3.01	3.89e-05	5.14e-07
04810	Regulation of actin cytoskeleton	161	203	129.99	2.21	5.85e-05	1.14e-06
04141	Protein processing in endoplasmic reticulum	132	163	104.37	2.44	5.85e-05	1.29e-06
05100	Bacterial invasion of epithelial cells	61	69	44.18	4.34	1.58e-04	4.18e-06
04666	Fc gamma R-mediated phagocytosis	78	92	58.91	3.18	2.45e-04	7.57e-06
05222	Small cell lung cancer	71	83	53.15	3.37	2.51e-04	1.01e-05
05215	Prostate cancer	74	87	55.71	3.24	2.51e-04	1.02e-05
04360	Axon guidance	102	125	80.04	2.53	2.51e-04	1.11e-05

Table 15: Kegg Pathways . Optrions used: mRNAs that are present in a mRNA-mRNA pair that has adjusted-pval cutoff <0.05; that also appears at least 1 times (databases: microCosm_v5_18, targetScan_v6.2.18); organism: human.

A.6 Study 2: Pancreatic ductal adenocarcinoma

Default miRCom output

</media/mvila/1842f0bf-523b-42e1-8670-c887a7e801b7/Article2>

April 26, 2017

1 Exploratory analysis of miRNA dataset

Number of miRNAs analysed	1733
Number of samples	12
Samples	102, 589, 71, 106, 272, 792, 253, 61, 748, 795, 829, 839

Table 1: Basic information of the miRNA dataset.

Group	PDACvsH
102 HEALTHY	0
589 HEALTHY	0
71 HEALTHY	0
106 PDAC	1
272 PDAC	1
792 PDAC	1
253 PDAC	1
61 PDAC	1
748 PDAC	1
795 PDAC	1
829 PDAC	1
839 PDAC	1

Table 2: Phenotypical information of the miRNA dataset.

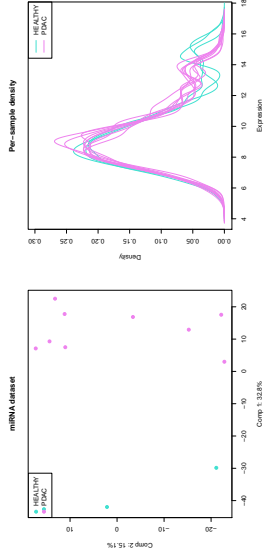


Figure 1: PCA and density plot for miRNAs.

2 Exploratory analysis of mRNA dataset

Number of miRNAs analysed	18570
Number of samples	12
Samples	102, 589, 71, 106, 272, 792, 253, 61, 748, 795, 829, 839

Table 3: Basic information of the mRNA dataset.

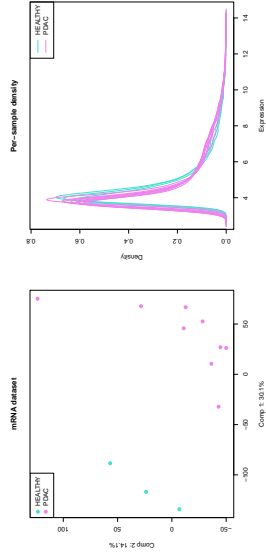


Figure 2: PCA and density plot for mRNAs.

Group	PDACvsH
102 HEALTHY	0
589 HEALTHY	0
71 HEALTHY	0
106 PDAC	1
272 PDAC	1
792 PDAC	1
253 PDAC	1
61 PDAC	1
748 PDAC	1
795 PDAC	1
829 PDAC	1
839 PDAC	1

Table 4: Phenotypical information of the mRNA dataset.

3 Differentially expressed miRNAs

Analysis performed	Comparative used: PDACvsH; method used: limma.
Number of differentially expressed miRNAs	543 (201 upregulated, 342 downregulated)
Number of samples	12
Samples	102, 589, 71, 106, 272, 792, 253, 61, 748, 795, 829, 839
Criteria for selecting miRNAs	adj.pval < 0.05

Table 5: Basic statistics

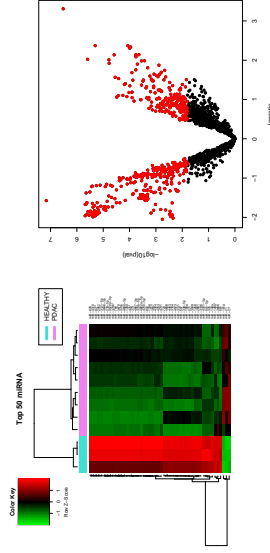


Figure 3: A) Heatmap with the top 50 most significant miRNAs (sorted by adjusted p-value). B) Volcano plot showing the selected miRNAs.

4 Differentially expressed mRNAs

Analysis performed	Comparative used: PDACvsH; method used: limma.
Number of differentially expressed mRNAs	3643 (1613 upregulated, 2030 downregulated)
Number of samples	12
Samples	102, 589, 71, 106, 272, 792, 253, 61, 748, 795, 829, 839
Criteria for selecting mRNAs	adj.pval < 0.05

Table 6: Basic statistics

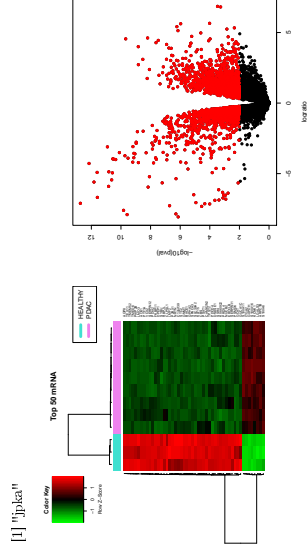


Figure 4: A) Heatmap with the top 50 most significant mRNAs (sorted by adjusted p-value). B) Volcano plot showing the selected mRNAs.

5 Correlation & intersection with databases

Number of miRNAs	543
Number of miRNAs	3643
Total miRNA-mRNA combinations	1978149
Number of samples	12
Samples	102, 589, 71, 106, 272, 792, 253, 61, 748, 795, 829, 839

Table 7: Number of miRNAs, mRNAs and samples used for correlation.

	Number	%
Total correlations	1978149	100
Total negative correlations	959775	48.52
Total negative correlations $p < 0.05$	786018	39.74
Total negative correlations $p < 0.01$	417928	21.13
Total negative correlations $\text{adj}p < 0.05$	443110	22.4
Total negative correlations $\text{adj}p < 0.01$	2929	0.15

Table 8: Basic statistics for correlation results. Correlation hypothesis: less.

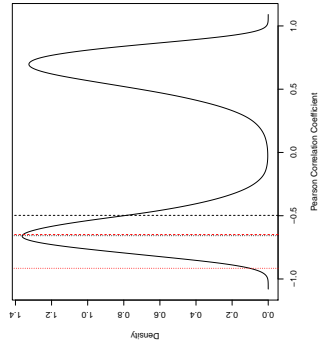


Figure 5: Density of a total of 1978149 miRNA-mRNA pairs. Dashed lines distinguish correlations whose p-value is lower than 0.05, dotted lines for 0.01. Black is for raw p-value and red for adjusted p-value.

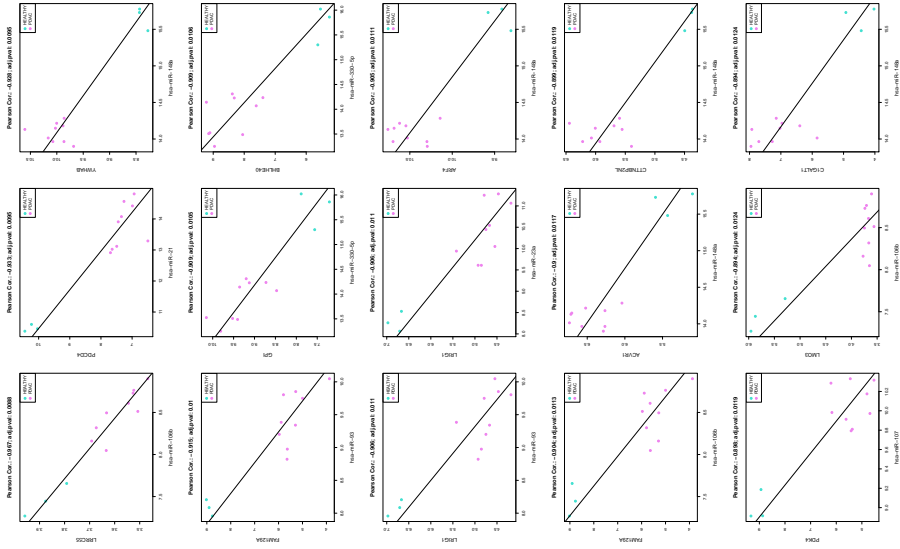


Figure 6: Plot of 15 top correlations, sorted by adjusted p-value. Databases used: fat-gsScan_v7.1_17, miRSVR_aug0_17, miRDB_v5.0_17 (each miRNA-mRNA pair has to appear at least 3 times).

miRNA	miRNA	cor	adj.pval	FC:miRNA	FC:miRNA	dat.sum
hsa-miR-106b	LRRC55	-0.97	8.80e-03	2.07	-1.23	3
hsa-miR-21	PDCD4	-0.93	9.55e-03	9.91	-7.90	3
hsa-miR-148a	YWHA8	-0.93	9.55e-03	-2.98	3.11	3
hsa-miR-83	FAM129A	-0.92	1.00e-02	2.60	-11.48	3
hsa-miR-330-3p	GPI	-0.91	1.05e-02	-3.64	3.38	3
hsa-miR-330-5p	BHLHE40	-0.91	1.06e-02	-3.64	7.97	3
hsa-miR-83	LRIG1	-0.91	1.10e-02	2.60	-4.13	3
hsa-miR-28a	LRIG1	-0.91	1.10e-02	4.40	-4.13	3
hsa-miR-148a	ARF4	-0.91	1.11e-02	-2.98	2.11	3
hsa-miR-106b	FAM129A	-0.90	1.13e-02	2.07	-11.48	3
hsa-miR-148a	ACVR1	-0.90	1.17e-02	-2.98	2.11	3
hsa-miR-148a	CITTNBP2NL	-0.90	1.19e-02	-2.98	2.76	3
hsa-miR-107	PDK4	-0.90	1.19e-02	2.08	-12.85	3
hsa-miR-106b	LMO3	-0.89	1.24e-02	2.07	-4.07	3
hsa-miR-148a	CIGAL1	-0.89	1.24e-02	-2.98	6.38	3
hsa-miR-330-3p	CAPN1	-0.89	1.25e-02	-3.64	3.99	3
hsa-miR-148a	TBL1XR1	-0.89	1.27e-02	-2.98	2.06	3
hsa-miR-320b	KIAA1824	-0.89	1.28e-02	1.66	-12.22	3
hsa-miR-320a	LMO3	-0.88	1.31e-02	2.14	-4.07	3
hsa-miR-43	SCN1A	-0.88	1.36e-02	2.60	-1.25	3
hsa-miR-148a	CNIH4	-0.87	1.37e-02	-2.98	2.46	3
hsa-miR-148a	DNNMT1	-0.87	1.38e-02	-2.98	3.09	3
hsa-miR-320b	RPL15	-0.87	1.38e-02	1.66	-2.09	3
hsa-miR-193b	TNFRSF21	-0.87	1.38e-02	-2.05	8.10	3
hsa-miR-148a	UBE2D1	-0.87	1.38e-02	-2.98	3.74	3
hsa-miR-181a	LMO3	-0.87	1.39e-02	5.17	-4.07	3
hsa-miR-193b	YWHAZ	-0.87	1.42e-02	-2.05	2.59	3
hsa-miR-424	LRIG1	-0.86	1.45e-02	1.82	-4.13	3
hsa-miR-106b	PDCD1IG2	-0.86	1.45e-02	2.07	-1.30	3
hsa-miR-130a	LRIG1	-0.86	1.46e-02	1.74	-4.13	3
hsa-miR-497	ITGA2	-0.86	1.46e-02	-1.94	23.44	3
hsa-miR-15a	ITGA2	-0.86	1.46e-02	-1.96	23.44	3
hsa-miR-34a	VAMP2	-0.86	1.47e-02	2.05	-1.50	3
hsa-miR-155	SCN1A	-0.86	1.47e-02	4.03	-1.25	3
hsa-miR-209-3p	TOP1	-0.86	1.47e-02	-1.87	2.35	3
hsa-miR-367	TOB1	-0.86	1.48e-02	1.61	-1.60	3
hsa-miR-330-5p	ARPC3L	-0.86	1.48e-02	-3.64	3.17	3
hsa-miR-19b	REN20	-0.86	1.49e-02	2.00	-1.80	3
hsa-miR-34a	INA	-0.86	1.49e-02	2.05	-1.72	3
hsa-miR-148a	CPD	-0.86	1.50e-02	-2.98	3.44	3
hsa-miR-148a	GMFB	-0.86	1.50e-02	-2.98	2.37	3
hsa-miR-374b	NMT1	-0.86	1.50e-02	-3.79	1.71	3
hsa-miR-373	RAB11A	-0.86	1.50e-02	-3.76	3.29	3
hsa-miR-374b	TCERG1	-0.85	1.51e-02	-3.79	1.59	3
hsa-miR-373	CAPZAI	-0.85	1.51e-02	-3.76	2.30	3

Table 9: Top 45 miRNA-mRNA pairs (sorted by adjusted p-value) that have: pval-corrected<0.05 and appear at least 3 times in the following databases: tar-geScan_v7.1_17, miRSVR_aug10_17, miRDB_v5.0_17.

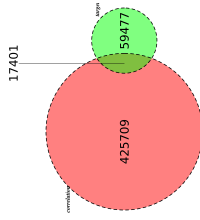


Figure 7: Venn Diagram. Left (red): number of miRNA-mRNA pairs with adjusted p-value<0.05. Right (green): number of all the theoretical miRNA-mRNA pairs reported at least 1 times in the following databases: tar-geScan_v7.1_17, miRSVR_aug10_17, miRDB_v5.0_17. Intersection: miRNA-mRNA pairs that fulfill both conditions.

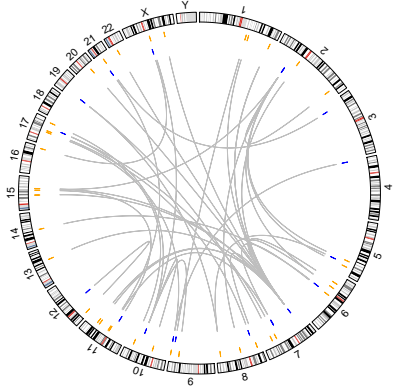


Figure 8: Circos plot for the first 45 miRNA-mRNA pairs (sorted by adjusted p-value) that have: pval-corrected<0.05 and appear at least 1 times in the following databases: tar-geScan_v7.1_17, miRSVR_aug10_17, miRDB_v5.0_17. Blue: miRNAs. Orange: target mRNAs

6 Functional analysis

6.1 Network analysis

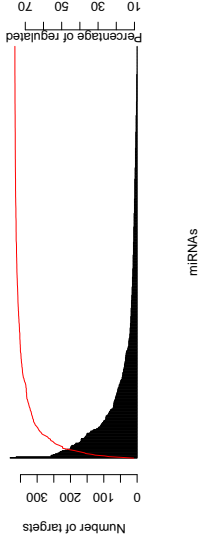


Figure 10: Barplot showing the number of mRNA targets per each miRNA (each bar represents a miRNA and they are sorted by number of targets). MiRNA-mRNA interactions have p -val-corrected <0.05 and predicted at least 1 time on the following databases: targetScan_v7.1.17; miRSVR_aug10.17; miRDE_v5.0.17. Red line (and right axis) represents the percentage of deregulated miRNAs that are cumulatively targeted by the miRNAs.

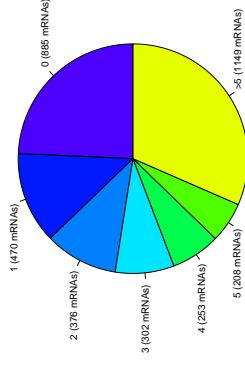


Figure 11: Pie chart representing the number of miRNAs targeting the miRNAs. p -val-corrected <0.05 and Targets=targetScan_v7.1.17; miRSVR_aug10.17; miRDE_v5.0.17(minimum coincidences between databases:1).

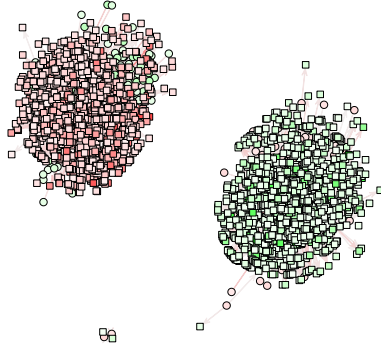


Figure 9: Network for all the miRNA-mRNA pairs that have: p -val-corrected <0.05 and appear at least 1 times in the following databases: targetScan_v7.1.17; miRSVR_aug10.17; miRDE_v5.0.17. Circles represent the miRNAs, and squares the miRNAs. Red fill means upregulated miRNAs/miRNAs, while green fill means downregulated miRNAs/miRNAs in comparative PDACvsH; lines indicate the miRNA-miRNA pairs, red line means positive score and green line means negative score.

miRNA	#miRNAs	miRNAs (top 20)
TBL1XR1	39	hsa-miR-148a*, hsa-miR-472-3p, hsa-miR-3666, hsa-miR-217, hsa-miR-4608-3p, hsa-miR-429, hsa-miR-15a, hsa-miR-497, hsa-miR-619, hsa-miR-377, hsa-miR-548l, hsa-miR-211, hsa-miR-876-5p, hsa-miR-338-3p, hsa-miR-148b, hsa-miR-548b, hsa-miR-548f, hsa-miR-548g, hsa-miR-4474-3p
CTTNBP2NL	36	hsa-miR-148a, hsa-miR-2082, hsa-miR-3167, hsa-miR-573, hsa-miR-374b, hsa-miR-448, hsa-miR-330-5p, hsa-miR-4603, hsa-miR-196a, hsa-miR-302c*, hsa-miR-367, hsa-miR-3168, hsa-miR-323-3p, hsa-miR-891b, hsa-miR-193b, hsa-miR-372, hsa-miR-377, hsa-miR-876-5p, hsa-miR-122, hsa-miR-136
YWHAZ	36	hsa-miR-193b, hsa-miR-217, hsa-miR-4429, hsa-miR-375, hsa-miR-339-5p, hsa-miR-636, hsa-miR-122, hsa-miR-758, hsa-miR-4474-3p, hsa-miR-492b, hsa-miR-204, hsa-miR-876-5p, hsa-miR-136, hsa-miR-548am, hsa-miR-211, hsa-miR-802, hsa-miR-641, hsa-miR-448, hsa-miR-7, hsa-miR-373
AMMECR1	33	hsa-miR-148a, hsa-miR-196a, hsa-miR-4310, hsa-miR-6253, hsa-miR-4700-5p, hsa-miR-448, hsa-miR-618, hsa-miR-4460-3p, hsa-miR-1236, hsa-miR-428, hsa-miR-4608-5p, hsa-miR-497, hsa-miR-15a, hsa-miR-876-5p, hsa-miR-148b, hsa-miR-548g, hsa-miR-548b, hsa-miR-4679, hsa-miR-548am, hsa-miR-548m
TNPO1	33	hsa-miR-290-3p, hsa-miR-154, hsa-miR-211, hsa-miR-4608-5p, hsa-miR-325, hsa-miR-4418, hsa-miR-218, hsa-miR-548b, hsa-miR-208b, hsa-miR-148a, hsa-miR-4469, hsa-miR-15a, hsa-miR-548f, hsa-miR-548g, hsa-miR-497, hsa-miR-4504, hsa-miR-548m, hsa-miR-548l, hsa-miR-548am, hsa-miR-4775
CCDC6	32	hsa-miR-302c*, hsa-miR-567, hsa-miR-148a, hsa-miR-39a*, hsa-miR-472-3p, hsa-miR-148a*, hsa-miR-3666, hsa-miR-641, hsa-miR-4910, hsa-miR-373, hsa-miR-802, hsa-miR-374b*, hsa-miR-3685, hsa-miR-875-5p, hsa-miR-330-5p, hsa-miR-557, hsa-miR-497, hsa-miR-15a, hsa-miR-122, hsa-miR-211
CPD	32	hsa-miR-148a, hsa-miR-306*, hsa-miR-635, hsa-miR-373, hsa-miR-641, hsa-miR-196a, hsa-miR-448, hsa-miR-497, hsa-miR-15a, hsa-miR-211, hsa-miR-377, hsa-miR-4255, hsa-miR-578c, hsa-miR-548f, hsa-miR-576-5p, hsa-miR-148b, hsa-miR-548m, hsa-miR-204, hsa-miR-338-3p, hsa-miR-4477b
CPSF6	32	hsa-miR-148a, hsa-miR-4761-3p, hsa-miR-474, hsa-miR-497, hsa-miR-15a, hsa-miR-802, hsa-miR-377, hsa-miR-204, hsa-miR-548f, hsa-miR-548am, hsa-miR-548b, hsa-miR-670, hsa-miR-136, hsa-miR-4762-3p, hsa-miR-548m, hsa-miR-448, hsa-miR-548l, hsa-miR-618, hsa-miR-4775, hsa-miR-2255-3p
G3BP2	32	hsa-miR-148a, hsa-miR-374b, hsa-miR-802, hsa-miR-6253, hsa-miR-448, hsa-miR-219-3p, hsa-miR-4603, hsa-miR-217, hsa-miR-6683, hsa-miR-5184, hsa-miR-148a*, hsa-miR-323-3p, hsa-miR-485-3p, hsa-miR-497, hsa-miR-4608-5p, hsa-miR-15a, hsa-miR-122, hsa-miR-212, hsa-miR-335, hsa-miR-148b
PDCD6IP	32	hsa-miR-148a, hsa-miR-217, hsa-miR-261*, hsa-miR-3144-3p, hsa-miR-323-3p, hsa-miR-548l, hsa-miR-148b, hsa-miR-211, hsa-miR-15a, hsa-miR-497, hsa-miR-876-5p, hsa-miR-4477b, hsa-miR-548g, hsa-miR-2914, hsa-miR-875-5p, hsa-miR-3140-5p, hsa-miR-4262, hsa-miR-4775, hsa-miR-671, hsa-miR-485-3p

Table 11: Top 10 mRNA with more miRNAs targeting them (each miRNA-mRNA pair has pval-corrected<0.05 and appears at least 1 times in the following databases: targetScan_v7.1.17, miRSVR_aug10_17, miRDB_v5.0_17). miRNAs in red are up-regulated in PDACvsH, miRNAs in green are downregulated in PDACvsH.

miRNA	#targets	cum. %	targets (top 20)
hsa-miR-374b	381	10.46	PAEPAL, CD58, TMSB10, CCL20, CTSEB, HSPH1, DNMT1, DSS, ELF1, UBAC2, FAT1, CCDC47, PTPN12, COPB1, FAM122B, ILS, CTTNBP2NL, FAM66A, H2AFY, ACVR1
hsa-miR-148a	363	16.85	HLA-A, KLF5, CTSEB, TNFRSF21, TMSB10, BID, TMEM123, KCNK1, B2M, PGRMC1, YWHAB, TAGLN2, ENDO1, PTPN12, UBRE2A, ACSL3, MYO1D, AMMECR1, PLEKH2, ACTG1
hsa-miR-181a	259	23.96	PCDC4, IFRD1, DFER, EPB414B, ANGPT1, LRIG1, KCNN1, NUCB2, DAGDB, FKBP1, EPB41, TMED6, LMO3, VCN2, MYO15A, RPL5, SLC25A33, PSAT1, ITSN2, SPATA20
hsa-miR-373	258	26.32	HLA-A, ENDO1, B2M, PGRMC1, BID, DSS, TAGLN2, CCDC47, PTPN12, MDK, PON2, MYO1D, SKAP2, CTTNBP2NL, FAM66A, ILS, H2AFY, PSMA2, ACVR1, CID
hsa-miR-320a	252	31.62	WNT9B, PDCD4, TMED6, PAIP2B, SFTTC, ADRAB, MS140, HHP11, CACNB1, AOX1, IFRD1, SND1, CECR2, GPH2, KCNAB1, OSBP2, ERO1L, EPB414B, LMO3, BACE1
hsa-miR-148	245	33.24	ENDOD1, GBP2, LITAF, LIMS1, DNMT1, ELF1, PTPN12, ILS, FAM66A, VPS13C, SEPT10, SKAP2, CTTNBP2NL, FAM122B, CALM2, RBM41, PPTAL, IUNSLABP, NEK6, PFKP
hsa-miR-93	238	36.62	IFRD1, FAM129A, LRIG1, ATXN2L2, MLC1, EPB414B, SH2D5, ANGPT1, ISM2, MS140, SYBU, SCNA1, MYO15A, PCMTD1, FBXO24, SLC46A2, EPB41, ITSN2, PAIP2B, WNT9B
hsa-miR-106b	234	37.47	LRRC55, FNDC5, ZNF385A, SH2D5, FAM129A, MYT1, MLC1, LMO3, IFRD1, C17orf67, KPNAT, APOBEC3H, SLC44A1, TM68A, ATOX8, PAIP2B, ARHGAP18, ERO1L, PRND, MUM1L1
hsa-miR-217	230	39.28	TNFRSF21, CTNNA1, ARPC2, GINT1, RAB11A, YWHAB, KLF5, PFKP, MAP4K4, YWHAB, CAPI, PTTGHP, RAC1, SPTLC2, ADAM9, PRKCI, ISG20, TES, DDN90, TMEM87B
hsa-miR-589	225	41.23	CCDC109B, NQO1, SULF2, KCNK1, MARCKSL1, ITGA2, PSM8, ARPC2, ENDO2D, HSBP1, SLC44A1, MIRP150, B2M, ENCI, FAM108CI, MAT2B, GCC2, HLA-A, DYNLL1, PNP

Table 10: Top 10 miRNA with more targets (each miRNA-mRNA pair has pval-corrected<0.05 and appears at least 1 times in the following databases: targetScan_v7.1.17, miRSVR_aug10_17, miRDB_v5.0_17). miRNAs in red are up-regulated in PDACvsH, miRNAs in green are downregulated in PDACvsH.

6.2 GO analysis

GOBPID	Term	Count	Size	ExpCount	OddsRatio	fdr	Pvalue
GO:0010033	response to organic substance	567	2895	427.34	1.53	2.29e-11	2.90e-15
GO:0051234	establishment of localization	822	4487	662.33	1.45	2.20e-11	4.83e-15
GO:0006810	transport	799	4356	643.00	1.45	2.82e-11	1.16e-14
GO:0051179	localization	974	5479	808.77	1.42	2.82e-11	1.19e-14
GO:0070857	cellular response to chemical stimulus	541	2768	408.59	1.52	4.79e-11	2.51e-14
GO:0007167	enzyme linked receptor protein signaling pathway	271	1201	177.28	1.77	6.68e-11	4.24e-14
GO:0071669	transmembrane receptor protein tyrosine kinase signaling pathway	212	893	131.82	1.88	2.57e-10	1.90e-13
GO:0006950	response to stress	721	3914	577.75	1.44	2.63e-10	2.22e-13
GO:1902578	single-organism localization	731	3986	588.38	1.43	3.95e-10	3.76e-13
GO:0071310	cellular response to organic substance	455	2294	338.62	1.53	5.18e-10	5.47e-13

Table 12: Biological Process . Options used: mRNAs that are present in a mRNA-mRNA pair that has adjusted-pval cutoff <0.05; that also appears at least 1 times (databases: targetScan_v7.1_17; miRSVR_aug10_17; miRDB_v3.0_17); organism: human.

GOCCID	Term	Count	Size	ExpCount	OddsRatio	fdr	Pvalue
GO:005237	cytoplasm	1714	10202	1476.67	1.61	2.92e-23	2.51e-26
GO:001988	membrane-bounded vesicle	696	3463	501.24	1.67	1.02e-21	1.77e-24
GO:0051982	vesicle	715	3584	518.76	1.66	1.02e-21	2.63e-24
GO:0044444	cytoplasmic part	1347	7699	1114.38	1.55	1.11e-21	3.81e-24
GO:0065010	extracellular membrane-bounded organelle	564	2708	391.96	1.72	1.21e-20	6.21e-23
GO:0070062	extracellular exosome	564	2708	391.96	1.72	1.21e-20	6.21e-23
GO:004320	extracellular organelle	565	2720	393.70	1.71	1.57e-20	1.08e-22
GO:1903561	extracellular vesicle	565	2720	393.70	1.71	1.57e-20	1.08e-22
GO:0043227	membrane-bounded organelle	1859	11533	1069.32	1.53	5.82e-17	4.46e-19
GO:0044421	extracellular region part	700	3680	532.65	1.54	5.76e-16	4.94e-18

Table 13: Cellular Component . Options used: mRNAs that are present in a mRNA-mRNA pair that has adjusted-pval cutoff <0.05; that also appears at least 1 times (databases: targetScan_v7.1_17; miRSVR_aug10_17; miRDB_v5.0_17); organism: human.

GO MFID	Term	Count	Size	ExpCount	OddsRatio	fdr	Pvalue
GO:0065015	protein binding	1714	10364	1523.69	1.53	6.52e-16	3.08e-19
GO:0005488	binding	2096	13675	2010.46	1.41	5.93e-05	5.60e-08
GO:0019859	enzyme binding	313	1644	241.70	1.42	1.61e-04	2.29e-07
GO:0050839	cell adhesion molecule binding	53	181	26.61	2.43	1.76e-04	3.33e-07
GO:0008092	cytoskeletal protein binding	167	800	117.61	1.57	3.61e-04	8.53e-07
GO:0042777	peptide binding	60	224	32.93	2.15	5.74e-04	1.63e-06
GO:0032218	amide binding	63	248	36.46	2.00	1.85e-03	6.11e-06
GO:0097367	carbohydrate derivative binding	387	2177	320.06	1.30	3.07e-03	1.16e-05
GO:0032553	ribonucleotide binding	329	1822	267.87	1.32	3.37e-03	1.60e-05
GO:0003779	actin binding	87	384	56.45	1.73	3.37e-03	1.73e-05

Table 14: Molecular Function . Options used: mRNAs that are present in a mRNA-mRNA pair that has adjusted-pval cutoff <0.05; that also appears at least 1 times (databases: targetScan_v7.1_17; miRSVR_aug10_17; miRDB_v5.0_17); organism: human.

KEGGID	Term	Count	Size	ExpCount	OddsRatio	lfr	Pvalue
05130	Pathogenic Escherichia coli infection	21	56	9.48	2.99	3.78e-02	1.74e-04
03050	Proteasome	17	44	7.45	3.13	5.16e-02	4.75e-04
03040	Spliceosome	35	123	20.82	1.99	6.39e-02	8.83e-04
04530	Tight junction	36	130	22.01	1.91	6.99e-02	1.29e-03
04141	Protein processing in endoplasmic reticulum	42	163	27.00	1.74	9.22e-02	2.48e-03
04145	Phagosome	39	149	25.23	1.77	9.32e-02	2.58e-03
05100	Bacterial invasion of epithelial cells	21	69	11.68	2.17	1.19e-01	3.84e-03
04810	Regulation of actin cytoskeleton	49	203	34.37	1.59	1.30e-01	4.78e-03
00100	Steroid biosynthesis	8	18	3.05	3.95	1.40e-01	5.81e-03
04514	Cell adhesion molecules (CAMs)	33	128	21.67	1.73	1.48e-01	6.81e-03

Table 15: Kegg Pathways . Oryctons used; mRNAs that are present in a mRNA-miRNA pair that has adjusted-pval cutoff <0.05; that also appears at least 1 times (databases: targetScan_v7.1_17, miRSVR_aug10_17, miRDE_v5.0_17); organism: human.

Appendix B

MiRComb vignettes and manuals

B.1 Main vignette

miRComb - An R package for analyzing miRNA-mRNA interactions

Maria Vila-Casadesús, Juanjo Lozano

May 4, 2017

Contents

1	Workflow	1
2	Data format	2
3	Creating the corObject	3
4	Analysis	3
4.1	Exploratory analysis	3
4.2	Differential expression	4
4.3	Correlation	6
4.3.1	Diagnostic plots	6
4.4	Organize the pairs in rows	6
4.5	Foldchanges	8
4.6	Adding targets information	9
4.7	<i>P</i> value correction	9
4.8	Interaction score	10
4.9	Save the results	11
5	Functional analysis	11
5.1	Most targeted miRNAs or mRNAs	11
5.2	Network	11
5.3	Gene Ontology analysis	11
6	Summary	12
7	Available databases	14
7.1	microCosm	14
7.2	targetScan	14
8	Session Info	15
	Bibliography	15

1 Workflow

This package provides a workflow for miRNA target analysis. Data about the miRNA databases is stored in a separate package `-miRData-`, which is automatically loaded with `miRComb`.

The main workflow of the package is represented on the following figure. We start from two datasets, where correlations are computed. Then they are combined with a database `-microCosm` or others- and a functional analysis of the results can be performed.

See full article: `miRComb` [1]

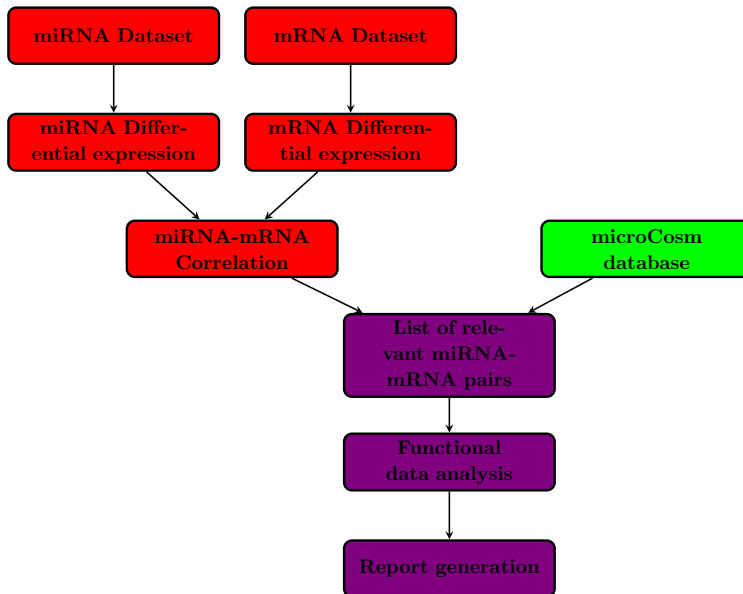


Figure 1: Outline of the pipeline.

2 Data format

We need the expression matrix for the miRNA and mRNA. The file format must be as follows:

- Expression matrices: a **matrix** with normalized data (preferably normalized log2 expression values). Columns should correspond to samples and Rows to probesets. Column names and Row names will be used as sample names and probe names respectively.
- Phenotypical information: a **data.frame**. Rows corresponding to sample names (must match with the Column names from the expression matrices). Columns with the desired combinations to test must be filled with 0 and 1. For example:

```

  group DvH
H_1     H  0
H_2     H  0
H_3     H  0
D_1     D  1
D_2     D  1
D_3     D  1
D_4     D  1
D_5     D  1
D_6     D  1
D_7     D  1
D_8     D  1
D_9     D  1

```


3 Creating the corObject

A `corObject` contains the following slots:

- `dat.miRNA`: miRNA matrix expression
- `dat.mRNA`: mRNA matrix expression
- `pheno.miRNA`: phenotypical miRNA information
- `pheno.mRNA`: phenotypical mRNA information
- `cor`: correlation matrix
- `pval`: correlation p value matrix
- `net`: a dataframe that can be used for cytoscape
- `diffexp.miRNA`: differential expression analysis from miRNA data
- `diffexp.mRNA`: differential expression analysis from mRNA data
- `sig.miRNA`: significant miRNAs
- `sig.mRNA`: significant mRNAs
- `info`: information of the tests performed

However, not all slots are mandatory for creating a simple `corObject`. A `corObject` can be created from the matrix expressions and phenotypical information. Further slots can be filled with specific functions. We can begin with the data provided as example (the data has been adapted from [2]):

```
> library(miRComb)
> data(miRNA)
> data(mRNA)
> data(pheno.miRNA)
> data(pheno.mRNA)
```

To create the `corObject`:

```
> data.obj<-new("corObject",dat.miRNA=as.matrix(miRNA),dat.mRNA=as.matrix(mRNA),
+             pheno.miRNA=pheno.miRNA,pheno.mRNA=pheno.mRNA)
```

4 Analysis

4.1 Exploratory analysis

Some plots are allowed to explore the data. For example we can plot the distances between samples of the mRNA dataset (Figure 2).

```
> plotCordist(data.obj,subset="mRNA",type="dist")
```

After this exploratory analysis, it is also possible to remove some samples and/or miRNAs/mRNAs. In this case we must indicate which sample and in which dataset we want to remove. The sample will be removed from the corresponding expression matrix and phenotypical dataframe. It is also possible to remove all the samples except for the selected ones. The procedures would be:

```
> #data.obj<-removeSamp(data.obj,"mRNA",c("D_4"))           #remove D_4 from the mRNA dataset
> #data.obj<-removeSamp(data.obj,"miRNA",genes="hsa-miR-21",keep=TRUE)       #keep only hsa-miR-21 :
```

Boxplots of the expression can also be plotted (Figure 3):

```
> boxplotSamples(data.obj,subset="mRNA")
```

PCA plots are also available (Figure 4, and `plot3d` function plots a PCA in 3 dimensions):

```
> plotPca(data.obj,subset="mRNA")
```

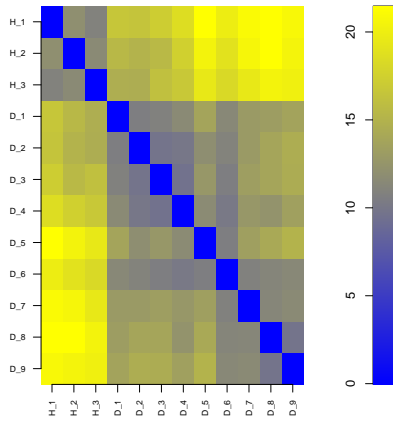


Figure 2: Plot of the distance between the samples of the mRNA dataset.

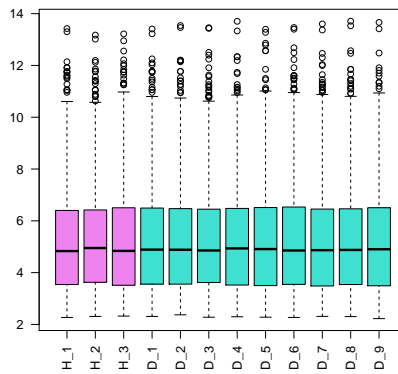


Figure 3: Boxplot of the mRNA samples.

4.2 Differential expression

We can add FoldChange information to the net from the differential expression slot. If this slot is not available, we can create it (indicating the column with the desired combination, in this case *Disease (D)* versus *Healthy (H)*, column DvH (see Section 2 to see the format of the column)):

```
> data.obj<-addDiffexp(data.obj,"miRNA",classes="DvH",method.dif="limma")
> data.obj<-addDiffexp(data.obj,"mRNA",classes="DvH",method.dif="limma")
```

Plot a heatmap of the top miRNA or mRNA (sorted by p value) (Figure 5):

```
> plotHeatmap(data.obj,"mRNA")
```

Moreover, we can obtain specific subsets, for example those genes with FoldChange greater than 10, a corrected p value less than 0.05 and specifically upregulated:

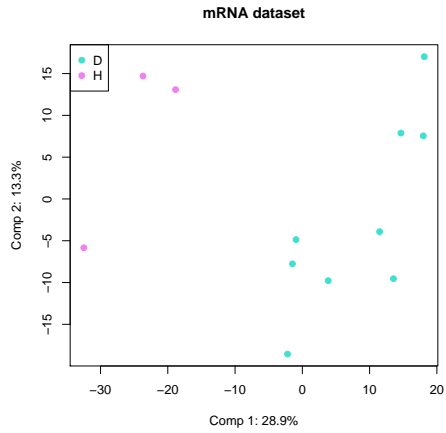


Figure 4: Principal Components Analysis (based on the correlation matrix) of the mRNA samples.

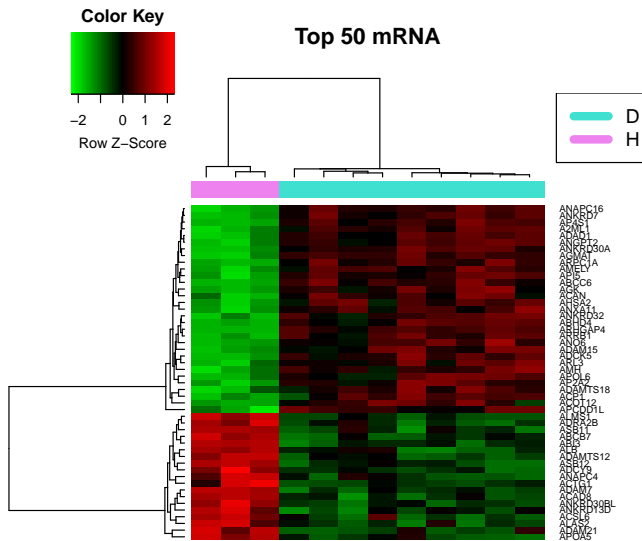


Figure 5: Heatmap of top 50 mRNAs, sorted by p value.

```
> selSubsetExprs(data.obj, "mRNA", FC=10, up=TRUE, adj.pval=0.05)
      FC logratio meanExp      pval      adj.pval
ANAPC16 35.75054 5.159893 9.454859 9.669116e-08 4.834558e-05
```

The same options can be used to add this information to the `corObject`. **The significant miRNAs and mRNAs added in this step will be used in correlation step.**

```
> data.obj<-addSig(data.obj,"mRNA",adj.pval=0.05,FC=1.5)
> data.obj<-addSig(data.obj,"miRNA",adj.pval=0.05)
```

If you have a specific list of miRNAs and/or mRNAs that you want to test, you should add them there in this step, for example:

```
> #data.obj<-addSig(data.obj,"miRNA",manual=c("hsa-miR-21","hsa-miR-21*",hsa-miR-200c"))
```

4.3 Correlation

The next step is to compute the correlation between the two matrices, the alternative hypothesis is "less" because we are interested only on negative correlations:

```
> data.obj<-addCorrelation(data.obj,alternative="less")
```

Correlating miRNA and mRNA

At this moment, the slots `cor` and `pval` have been filled. The column names are the mRNAs selected by `add.sig`, and the row names are the miRNAs selected by `add.sig` also:

```
> data.obj@cor[1:3,1:3]
                A2ML1      ABCC6      ABCD3
hsa-miR-107  -0.8616698 -0.7886117  0.5100627
hsa-miR-1208 -0.9231799 -0.8446258  0.7978511
hsa-miR-1231  0.9492489  0.8967782 -0.7377228

> data.obj@pval[1:3,1:3]
                A2ML1      ABCC6      ABCD3
hsa-miR-107  1.573829e-04 0.0011512207 0.954885976
hsa-miR-1208 9.250455e-06 0.0002730272 0.999063814
hsa-miR-1231 9.999988e-01 0.9999612867 0.003084055
```

If `add.sig` was set to `NULL`, all the miRNAs and/or mRNAs are used, respectively.

4.3.1 Diagnostic plots

It is also possible to plot the correlation for each pair (Figure 8) and some diagnostic plots for the linear correlation (Figure 7):

```
> plotCorrelation(data.obj,miRNA="hsa-miR-107",mRNA="A2ML1",type="cor",
+   col.color="group",sample.names=TRUE)
> plotCorrelation(data.obj,miRNA="hsa-miR-107",mRNA="A2ML1",type="residuals")
```

4.4 Organize the pairs in rows

These slots can be used to create another slot, which is called `net`. This slot contains a `data.frame` where each row represents a specific miRNA-mRNA pair, and each column contains information relevant to the pair, the name of the table refers to Cytoscape software, as this format can be easily imported to it [3]. In this step of the analysis the columns are: miRNA, mRNA, correlation coefficient and *p* value; other columns will be added in further steps.

```
> data.obj<-addNet(data.obj)
```

Converting to net

```
> head(data.obj@net)
```

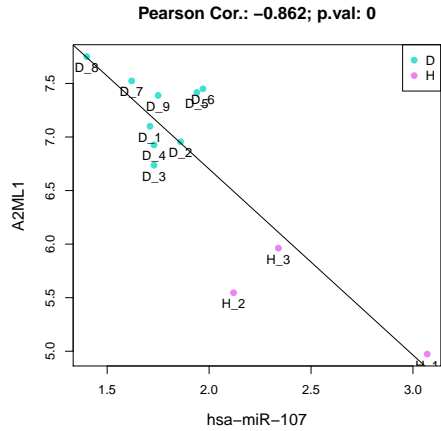


Figure 6: Plot of the correlation of one miRNA (*hsa-miR-107*) and one mRNA (*A2ML1*). Horizontal and vertical axis represent the (\log_2)-expression values (see Section 2) of the miRNA and mRNA, respectively.

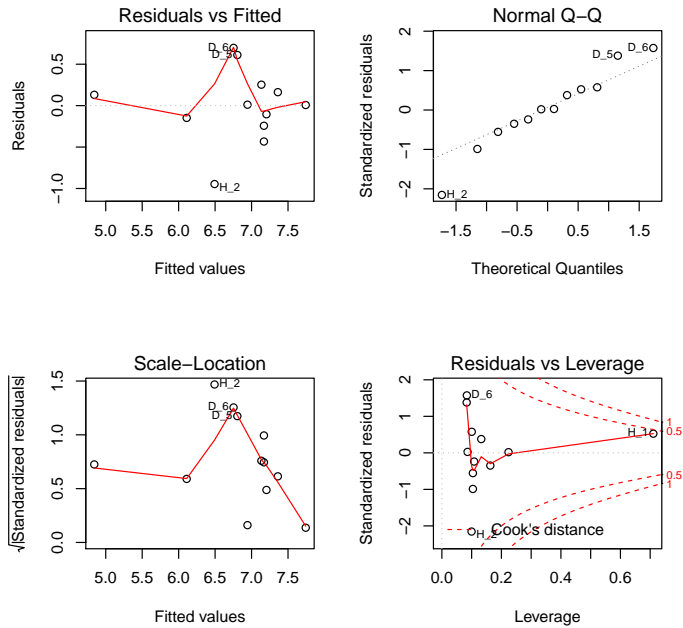


Figure 7: Diagnostic plot for the linear regression between *hsa-miR-107* and *A2ML1* (see Figure 8).

	miRNA	mRNA	cor	pval
hsa-miR-107:A2ML1	hsa-miR-107	A2ML1	-0.8616698	0.0001573829

```
[1] "violet" "violet" "violet" "turquoise" "turquoise" "turquoise"
[7] "turquoise" "turquoise" "turquoise" "turquoise" "turquoise" "turquoise"
```

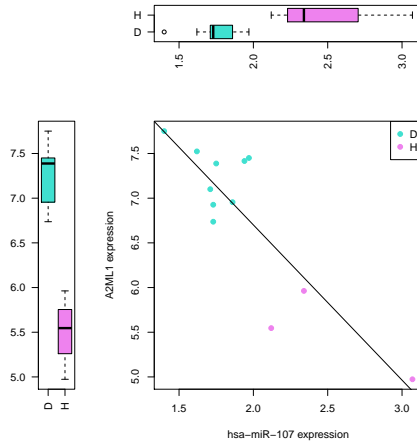


Figure 8: Plot of the correlation of one miRNA (*hsa-miR-107*) and one mRNA (*A2ML1*). Horizontal and vertical axis represent the (log₂)-expression values (see Section 2) of the miRNA and mRNA, respectively.

```
hsa-miR-107:ABCC6 hsa-miR-107 ABCC6 -0.7886117 0.0011512207
hsa-miR-107:ABCD3 hsa-miR-107 ABCD3 0.5100627 0.9548859759
hsa-miR-107:ABHD12 hsa-miR-107 ABHD12 0.9178376 0.9999871718
hsa-miR-107:ABHD4 hsa-miR-107 ABHD4 -0.7378162 0.0030790960
hsa-miR-107:ABI1 hsa-miR-107 ABI1 -0.5795687 0.0241310206
```

4.5 Foldchanges

As optional, we add the FoldChange information of the `diffexp.miRNA` and `diffexp.mRNA` slots to the `net` slot:

```
> data.obj<-addFoldchanges(data.obj)
> head(data.obj@net)
```

	miRNA	mRNA	cor	pval	logratio.miRNA
hsa-miR-107:A2ML1	hsa-miR-107	A2ML1	-0.8616698	0.0001573829	-0.7644444
hsa-miR-107:ABCC6	hsa-miR-107	ABCC6	-0.7886117	0.0011512207	-0.7644444
hsa-miR-107:ABCD3	hsa-miR-107	ABCD3	0.5100627	0.9548859759	-0.7644444
hsa-miR-107:ABHD12	hsa-miR-107	ABHD12	0.9178376	0.9999871718	-0.7644444
hsa-miR-107:ABHD4	hsa-miR-107	ABHD4	-0.7378162	0.0030790960	-0.7644444
hsa-miR-107:ABI1	hsa-miR-107	ABI1	-0.5795687	0.0241310206	-0.7644444
	meanExp.miRNA	logratio.mRNA	meanExp.mRNA		
hsa-miR-107:A2ML1	1.936667	1.7556491	6.810779		
hsa-miR-107:ABCC6	1.936667	0.8767317	5.806870		
hsa-miR-107:ABCD3	1.936667	-0.8775035	5.329774		
hsa-miR-107:ABHD12	1.936667	-1.1404364	5.685076		
hsa-miR-107:ABHD4	1.936667	1.3967358	4.653627		
hsa-miR-107:ABI1	1.936667	1.8646091	7.268884		

4.6 Adding targets information

At this moment, two databases are provided, but if it is necessary we can add more (see `?add.database`):

```
> data(microCosm_v5_18)
> data(targetScan_v6.2_18)
```

The function to add the database(s) information is (this step can take a while):

```
> data.obj<-addDatabase(data.obj,database=c("microCosm_v5_18","targetScan_v6.2_18"))
```

Intersecting with database

```
microCosm_v5_18 database chosen
targetScan_v6.2_18 database chosen
```

```
> head(data.obj@net)
```

	miRNA	mRNA	cor	pval	logratio.miRNA
hsa-miR-107:A2ML1	hsa-miR-107	A2ML1	-0.8616698	0.0001573829	-0.7644444
hsa-miR-107:ABCC6	hsa-miR-107	ABCC6	-0.7886117	0.0011512207	-0.7644444
hsa-miR-107:ABCD3	hsa-miR-107	ABCD3	0.5100627	0.9548859759	-0.7644444
hsa-miR-107:ABHD12	hsa-miR-107	ABHD12	0.9178376	0.9999871718	-0.7644444
hsa-miR-107:ABHD4	hsa-miR-107	ABHD4	-0.7378162	0.0030790960	-0.7644444
hsa-miR-107:ABI1	hsa-miR-107	ABI1	-0.5795687	0.0241310206	-0.7644444
	meanExp.miRNA	logratio.mRNA	meanExp.mRNA	dat.microCosm_v5_18	
hsa-miR-107:A2ML1	1.936667	1.7556491	6.810779		0
hsa-miR-107:ABCC6	1.936667	0.8767317	5.806870		0
hsa-miR-107:ABCD3	1.936667	-0.8775035	5.329774		0
hsa-miR-107:ABHD12	1.936667	-1.1404364	5.685076		1
hsa-miR-107:ABHD4	1.936667	1.3967358	4.653627		0
hsa-miR-107:ABI1	1.936667	1.8646091	7.268884		0
	dat.targetScan_v6.2_18	dat.sum			
hsa-miR-107:A2ML1		0			0
hsa-miR-107:ABCC6		0			0
hsa-miR-107:ABCD3		0			0
hsa-miR-107:ABHD12		0	1		1
hsa-miR-107:ABHD4		0	0		0
hsa-miR-107:ABI1		0	0		0

And we can see that some columns have been added:

- One column for each database, with the name: `dat.database_name`. 1 means that the miRNA-mRNA pair has been found as predicted in that database, 0 that the miRNA-mRNA pair is not predicted.
- The column `dat.sum`, it reports how many times that miRNA-mRNA pair has been found in the used databases.

4.7 P value correction

We can add a column with the corrected p value as follows. This step is important for controlling the Type I error of the correlations:

```
> data.obj<-correctPval(data.obj, pval="pval",method.adj="BH")
```

Correcting p.values

```
> head(data.obj@net)
```

	miRNA	mRNA	cor	pval	logratio.miRNA
hsa-miR-107:A2ML1	hsa-miR-107	A2ML1	-0.8616698	0.0001573829	-0.7644444
hsa-miR-107:ABCC6	hsa-miR-107	ABCC6	-0.7886117	0.0011512207	-0.7644444
hsa-miR-107:ABCD3	hsa-miR-107	ABCD3	0.5100627	0.9548859759	-0.7644444
hsa-miR-107:ABHD12	hsa-miR-107	ABHD12	0.9178376	0.9999871718	-0.7644444
hsa-miR-107:ABHD4	hsa-miR-107	ABHD4	-0.7378162	0.0030790960	-0.7644444
hsa-miR-107:ABI1	hsa-miR-107	ABI1	-0.5795687	0.0241310206	-0.7644444
	meanExp.miRNA	logratio.mRNA	meanExp.mRNA	dat.microCosm_v5_18	
hsa-miR-107:A2ML1	1.936667	1.7556491	6.810779		0
hsa-miR-107:ABCC6	1.936667	0.8767317	5.806870		0
hsa-miR-107:ABCD3	1.936667	-0.8775035	5.329774		0
hsa-miR-107:ABHD12	1.936667	-1.1404364	5.685076		1
hsa-miR-107:ABHD4	1.936667	1.3967358	4.653627		0
hsa-miR-107:ABI1	1.936667	1.8646091	7.268884		0
	dat.targetScan_v6.2_18	dat.sum	adj.pval		
hsa-miR-107:A2ML1	0	0	0.002656169		
hsa-miR-107:ABCC6	0	0	0.007102815		
hsa-miR-107:ABCD3	0	0	0.999999128		
hsa-miR-107:ABHD12	0	1	0.999999128		
hsa-miR-107:ABHD4	0	0	0.012746900		
hsa-miR-107:ABI1	0	0	0.054419612		

4.8 Interaction score

Finally, a score can be added to each interaction. This score is related to both logratios and it is aimed to reflect the possible *biological relevance* of the miRNA (higher score means that both miRNA and mRNA are highly deregulated in that disease).

$$\text{score} = -2(\text{logratio}_{\text{miRNA}} \cdot \text{logratio}_{\text{mRNA}})$$

```
> data.obj<-addScore(data.obj)
> head(data.obj@net)
```

	miRNA	mRNA	cor	pval	logratio.miRNA
hsa-miR-107:A2ML1	hsa-miR-107	A2ML1	-0.8616698	0.0001573829	-0.7644444
hsa-miR-107:ABCC6	hsa-miR-107	ABCC6	-0.7886117	0.0011512207	-0.7644444
hsa-miR-107:ABCD3	hsa-miR-107	ABCD3	0.5100627	0.9548859759	-0.7644444
hsa-miR-107:ABHD12	hsa-miR-107	ABHD12	0.9178376	0.9999871718	-0.7644444
hsa-miR-107:ABHD4	hsa-miR-107	ABHD4	-0.7378162	0.0030790960	-0.7644444
hsa-miR-107:ABI1	hsa-miR-107	ABI1	-0.5795687	0.0241310206	-0.7644444
	meanExp.miRNA	logratio.mRNA	meanExp.mRNA	dat.microCosm_v5_18	
hsa-miR-107:A2ML1	1.936667	1.7556491	6.810779		0
hsa-miR-107:ABCC6	1.936667	0.8767317	5.806870		0
hsa-miR-107:ABCD3	1.936667	-0.8775035	5.329774		0
hsa-miR-107:ABHD12	1.936667	-1.1404364	5.685076		1
hsa-miR-107:ABHD4	1.936667	1.3967358	4.653627		0
hsa-miR-107:ABI1	1.936667	1.8646091	7.268884		0
	dat.targetScan_v6.2_18	dat.sum	adj.pval	score	
hsa-miR-107:A2ML1	0	0	0.002656169	2.684192	
hsa-miR-107:ABCC6	0	0	0.007102815	1.340425	
hsa-miR-107:ABCD3	0	0	0.999999128	-1.341605	
hsa-miR-107:ABHD12	0	1	0.999999128	-1.743601	
hsa-miR-107:ABHD4	0	0	0.012746900	2.135454	
hsa-miR-107:ABI1	0	0	0.054419612	2.850780	

4.9 Save the results

It is possible to output the results to text files and explore them with other tools if desired (for example *excel* or *libreoffice*).

```
> writeCsv(data.obj, "results_today.csv", pval.corrected=1)
```

These networks can also be opened by cytoscape (v2.x), by indicating the pathway of the folder which contains the `cytoscape.jar` file:

```
> #openCytoscape(data.obj, p.cutoff=0.0001, cytoscape.folder="/home/mvila/Cytoscape_v2.8.3")
```

5 Functional analysis

5.1 Most targeted miRNAs or mRNAs

A table showing the number of targets for each miRNA (or a table showing the number of miRNAs that are targeting a specific mRNA) can be obtained (with the option `names=TRUE`, the names of the targets are also reported). All the miRNA/mRNA are plotted and displayed here, even if they have no targets (Figure 9).

```
> topTable(data.obj, "miRNA", names=FALSE, pval.cutoff=0.05)[1:20]
```

hsa-miR-1231	hsa-miR-1273f	hsa-miR-107	hsa-miR-1208	hsa-miR-127-5p
4	3	2	1	1
hsa-miR-1273e	hsa-miR-1258	hsa-miR-1291	hsa-miR-1296	<NA>
1	0	0	0	
<NA>	<NA>	<NA>	<NA>	<NA>
<NA>	<NA>	<NA>	<NA>	<NA>

This information can also be represented with a barplot (Figure 9):

```
> topTable(data.obj, "miRNA", names=TRUE, pval.cutoff=0.05, plot=TRUE)
```

5.2 Network

We can draw a network (Figure 10) with the following procedure. We need to give a p value cutoff (this p value refers to the corrected p value) and a minimum number of occurrences on the theoretical databases (`dat.sum`, see Section 4.6):

```
> plotNetwork(data.obj, pval.cutoff=0.01, dat.sum=1)
```

A bigger picture (Figure 11):

```
> plotNetwork(data.obj, pval.cutoff=0.05, names=FALSE)
```

A picture of the miRNA with more targets (Figure 12):

```
> hub<-names(topTable(data.obj, "miRNA"))[1]
```

```
> plotNetwork(data.obj, pval.cutoff=0.05, names=TRUE, sub.miRNA=hub, vertex.cex="interact.table")
```

***Any of these networks can be opened with Cytoscape using the function `openCytoscape`.**

5.3 Gene Ontology analysis

It is possible to select the mRNA of the pairs according to the combined p value and perform a GO enrichment analysis (reference genes are the whole human genome).

```
> GO.results<-GOanalysis(data.obj, type="GO", ontology="BP")
```

We can also compute the GO of a specific miRNA:

```
> #data.obj<-GOanalysis(data.obj, type="GO", ontology="BP", sub.miRNA="hsa-miR-516a-3p")
```

```
> #GO.results<-data.obj@GO.results[["GO:BP"]]
```

```
> #GO.results[which(GO.results$Pvalue<0.1), "Term"]
```

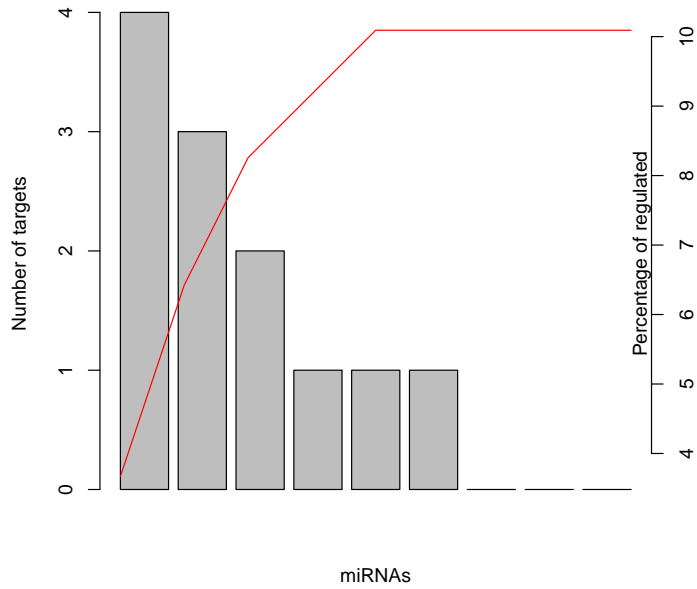


Figure 9: Barplot showing the number of targets per miRNA. The red line represents the cumulative percentage of mRNAs –respect to the total number of deregulated mRNAs– that the miRNAs are targeted by at least one miRNA.

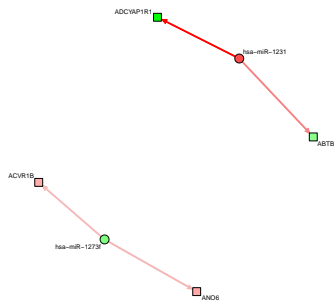


Figure 10: Network

6 Summary

Finally, a summary of the methods used can be obtained:

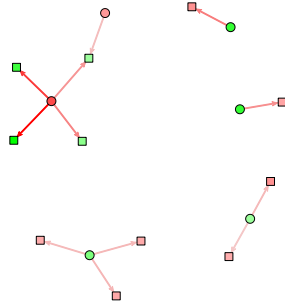


Figure 11: Network, bigger picture (without names)

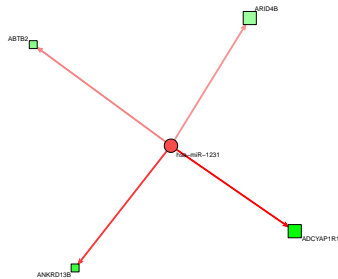


Figure 12: Targets of the miRNA hub. The size of the mRNAs reflects the number of protein-protein interactions they have (provided by `interact.table`).

```
> summary(data.obj)
```

```
corObject with:
  miRNA slot with 12 samples and 200 probesets
  mRNA slot with 12 samples and 1000 probesets
Computations done:
- Differential expression mRNA:  limma method used
                               DvH comparison used
- Differential expression miRNA: limma method used
                               DvH comparison used
- Correlation: "pearson" method used
               "correlation" function used
               12 samples used
```

```

          9 miRNAs used
          adj.pval < 0.05
        109 mRNAs used
          abs(logratio) > 0.58; abs(FC) > 1.5; adj.pval < 0.05
- Database: "microCosm_v5_18" database used
- Database: "targetScan_v6.2_18" database used
- P.value adjustment: "BH" method used

```

A pdf report can also be generated with the following function:

```
> mkReport(data.obj, "NameOfTheReport")
```

And then the file NameOfTheReport.pdf will be created.

7 Available databases

All names are from miRBase version 17. See miRData package for more information.

7.1 microCosm

<http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5/>

```
> data(microCosm_v5_18)
> head(microCosm_v5_18)
```

	mir18	mir17	mir_acc	mirmicrocosm
hsa-miR-598:A1L4H1_HUMAN	hsa-miR-598	hsa-miR-598	MIMAT0003266	hsa-miR-598
hsa-miR-181a:NR6A1	hsa-miR-181a-5p	hsa-miR-181a	MIMAT0000256	hsa-miR-181a
hsa-miR-181c:NR6A1	hsa-miR-181c-5p	hsa-miR-181c	MIMAT0000258	hsa-miR-181c
hsa-miR-181b:NR6A1	hsa-miR-181b-5p	hsa-miR-181b	MIMAT0000257	hsa-miR-181b
hsa-miR-181d:NR6A1	hsa-miR-181d	hsa-miR-181d	MIMAT0002821	hsa-miR-181d
hsa-miR-212:NP_055530.2	hsa-miR-212-3p	hsa-miR-212	MIMAT0000269	hsa-miR-212
	target_name	target_entrezid	pval	score
hsa-miR-598:A1L4H1_HUMAN	A1L4H1_HUMAN	ENST00000389623	1.07251e-16	18.1954
hsa-miR-181a:NR6A1	NR6A1	ENST00000373584	1.99162e-13	20.0243
hsa-miR-181c:NR6A1	NR6A1	ENST00000373584	1.99162e-13	19.1021
hsa-miR-181b:NR6A1	NR6A1	ENST00000373584	1.99162e-13	18.7798
hsa-miR-181d:NR6A1	NR6A1	ENST00000373584	1.99162e-13	18.4404
hsa-miR-212:NP_055530.2	NP_055530.2	ENST00000310343	5.45163e-13	16.3193
	method	names		
hsa-miR-598:A1L4H1_HUMAN	microcosm	hsa-miR-598:A1L4H1_HUMAN		
hsa-miR-181a:NR6A1	microcosm	hsa-miR-181a:NR6A1		
hsa-miR-181c:NR6A1	microcosm	hsa-miR-181c:NR6A1		
hsa-miR-181b:NR6A1	microcosm	hsa-miR-181b:NR6A1		
hsa-miR-181d:NR6A1	microcosm	hsa-miR-181d:NR6A1		
hsa-miR-212:NP_055530.2	microcosm	hsa-miR-212:NP_055530.2		

7.2 targetScan

<http://www.targetscan.org/>

```
> data(targetScan_v6.2_18)
> head(targetScan_v6.2_18)
```

	mir18	mir17	mir_acc	mirtargetscan
hsa-let-7a:DZIP1	hsa-let-7a-5p	hsa-let-7a	MIMAT0000062	hsa-let-7a
hsa-let-7a:TEX261	hsa-let-7a-5p	hsa-let-7a	MIMAT0000062	hsa-let-7a
hsa-let-7a:CAP1	hsa-let-7a-5p	hsa-let-7a	MIMAT0000062	hsa-let-7a

hsa-let-7a:GIPC1	hsa-let-7a-5p	hsa-let-7a	MIMAT0000062	hsa-let-7a
hsa-let-7a:SUCLG2	hsa-let-7a-5p	hsa-let-7a	MIMAT0000062	hsa-let-7a
hsa-let-7a:CANT1	hsa-let-7a-5p	hsa-let-7a	MIMAT0000062	hsa-let-7a
	target_name	target_entrezid	method	names
hsa-let-7a:DZIP1	DZIP1	22873	targetscan	hsa-let-7a:DZIP1
hsa-let-7a:TEX261	TEX261	113419	targetscan	hsa-let-7a:TEX261
hsa-let-7a:CAP1	CAP1	10487	targetscan	hsa-let-7a:CAP1
hsa-let-7a:GIPC1	GIPC1	10755	targetscan	hsa-let-7a:GIPC1
hsa-let-7a:SUCLG2	SUCLG2	8801	targetscan	hsa-let-7a:SUCLG2
hsa-let-7a:CANT1	CANT1	124583	targetscan	hsa-let-7a:CANT1

8 Session Info

- R version 3.3.3 (2017-03-06), x86_64-pc-linux-gnu
- Locale: LC_CTYPE=en_US.UTF-8, LC_NUMERIC=C, LC_TIME=es_ES.UTF-8, LC_COLLATE=C, LC_MONETARY=es_ES.UTF-8, LC_MESSAGES=en_US.UTF-8, LC_PAPER=es_ES.UTF-8, LC_NAME=C, LC_ADDRESS=C, LC_TELEPHONE=C, LC_MEASUREMENT=es_ES.UTF-8, LC_IDENTIFICATION=C
- Base packages: base, datasets, grDevices, graphics, grid, methods, parallel, stats, stats4, utils
- Other packages: AnnotationDbi 1.34.4, Biobase 2.32.0, BiocGenerics 0.18.0, Category 2.38.0, CircStats 0.2-4, DESeq 1.24.0, DOSE 2.10.7, Formula 1.2-1, GO.db 3.3.0, GOstats 2.38.1, Hmisc 4.0-3, IRanges 2.6.1, KEGG.db 3.2.3, MASS 7.3-45, Matrix 1.2-8, RamiGO 1.18.0, RankProd 2.44.0, Rcpp 0.12.10, ReactomePA 1.16.2, S4Vectors 0.10.3, VennDiagram 1.6.17, WriteXLS 4.0.0, boot 1.3-18, circlize 0.3.10, dtw 1.18-1, fgsea 0.99.7, fields 8.10, foreach 1.4.3, futile.logger 1.4.3, ggplot2 2.2.1, glmnet 2.0-9, gplots 3.0.1, graph 1.50.0, gsubfn 0.6-6, gtools 3.5.0, lattice 0.20-34, limma 3.28.21, locfit 1.5-9.1, maps 3.1.1, mclust 5.2.3, miRComb 0.8.9, miRData 0.6.1, mvoutlier 2.0.8, network 1.13.0, org.Hs.eg.db 3.3.0, pROC 1.9.1, pheatmap 1.0.8, proto 1.0.0, proxy 0.4-17, scatterplot3d 0.3-40, sgeostat 1.0-27, spam 1.4-0, survival 2.40-1, verification 1.42, xtable 1.8-2
- Loaded via a namespace (and not attached): AnnotationForge 1.14.2, BiocParallel 1.6.6, DBI 0.5-1, DEoptimR 1.0-8, DO.db 2.9, Gally 1.3.0, GOsemSim 1.30.3, GSEABase 1.34.1, GlobalOptions 0.0.11, KernSmooth 2.23-15, MatrixModels 0.4-1, RBGL 1.48.1, RColorBrewer 1.1-2, RCurl 1.95-4.8, RCytoscape 1.21.1, RSQLite 1.1-2, SparseM 1.74, VIM 4.6.0, XML 3.98-1.5, XMLRPC 0.3-0, acepack 1.4.1, annotate 1.50.1, assertthat 0.1, backports 1.0.5, base64enc 0.1-3, bitops 1.0-6, caTools 1.17.1, car 2.1-4, checkmate 1.8.2, class 7.3-14, cluster 2.0.5, codetools 0.2-15, colorspace 1.3-2, cvTools 0.3.2, data.table 1.10.4, digest 0.6.12, diptest 0.75-7, e1071 1.6-8, fastmatch 1.1-0, flexmix 2.3-13, foreign 0.8-67, fpc 2.1-10, futile.options 1.0.0, gdata 2.17.0, genefilter 1.54.2, geneplotter 1.50.0, graphite 1.18.1, gridExtra 2.2.1, gtable 0.2.0, htmlTable 1.9, htmltools 0.3.6, htmlwidgets 0.8, igraph 1.0.1, iterators 1.0.8, kernlab 0.9-25, knitr 1.15.1, laeken 0.4.6, lambda.r 1.1.9, latticeExtra 0.6-28, lazyeval 0.2.0, lme4 1.1-12, lmtest 0.9-35, magrittr 1.5, memoise 1.0.0, mgcv 1.8-17, minqa 1.2.4, modeltools 0.2-21, munsell 0.4.3, mvtnorm 1.0-5, nlme 3.1-131, nloptr 1.0.4, nnet 7.3-12, pbkrtest 0.4-6, pcaPP 1.9-61, pls 2.6-0, plyr 1.8.4, png 0.1-7, prabclus 2.2-6, quantreg 5.29, qvalue 2.4.2, rappdirs 0.3.1, reactome.db 1.55.0, reshape 0.8.6, reshape2 1.4.2, robCompositions 2.0.3, robustbase 0.92-7, rpart 4.1-10, rrcov 1.4-3, sROC 0.1-2, scales 0.4.1, shape 1.4.2, sp 1.2-4, splines 3.3.3, stringi 1.1.2, stringr 1.2.0, tcltk 3.3.3, tibble 1.2, tools 3.3.3, trimcluster 0.1-2, vcd 1.4-3, zoo 1.7-14

References

- [1] Maria Vila-Casadesús, Meritxell Gironella, and Juan José Lozano. MiRComb: An R Package to Analyse miRNA-mRNA Interactions. Examples across Five Digestive Cancers. *PLoS ONE*, 11(3):e0151127, March 2016.

- [2] Silvia Affò, Marlene Dominguez, Juan José Lozano, Pau Sancho-Bru, Daniel Rodrigo-Torres, Oriol Morales-Ibanez, Montserrat Moreno, Cristina Millán, Aurora Loeza-del Castillo, José Altamirano, Juan Carlos García-Pagán, Vicente Arroyo, Pere Ginès, Juan Caballería, Robert F. Schwabe, and Ramon Bataller. Transcriptome analysis identifies TNF superfamily receptors as potential therapeutic targets in alcoholic hepatitis. *Gut*, 62(3):452–460, March 2013.
- [3] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, 13(11):2498–2504, November 2003.

B.2 Additional vignette

miRComb - An R package for analyzing miRNA-mRNA interactions. Additional Examples.

Maria Vila-Casadesús

May 4, 2017

Contents

1	Time analysis	1
1.1	Time analysis	1
1.1.1	Final time	2
1.1.2	Linear regression	2
1.2	Select desired miRNAs and mRNAs (optional)	2
1.3	Rest of the analysis	2
2	Non-matched miRNA and mRNA data	3
3	Session Info	4
	Bibliography	5

Brief comment

The main analysis is described in the `miRComb` main vignette, only variations of that description are included in this file.

The article describing the package with example data has been published in PLoS ONE [1].

1 Time analysis

We can begin with the data provided as example¹ (the data has been adapted from [2]):

```
> library(miRComb)
> load("longdata.RData")
```

To create the `corObject`:

```
> data.obj<-new("corObject",dat.mirna=as.matrix(miRNA),dat.mrna=as.matrix(mRNA),
+             pheno.mirna=pheno.mirna,pheno.mrna=pheno.mrna)
```

Exploratory analysis can be done as usual.

1.1 Time analysis

We can add time information to the net from the "differential expression" slot (in this case, time analysis slot). If this slot is not available, we can create it. We have two options:

¹Data can be downloaded from here: <http://sourceforge.net/projects/mircomb/files/>.

1.1.1 Final time

Analysing $t_{fin} - t_{init}$:

In this case we compare the final time expression versus the initial expression of the miRNA or mRNA

```
> data.obj<-addLong(data.obj,"miRNA","time_alt","time.point")
> data.obj<-addLong(data.obj,"mRNA","time_alt","time.point")
> head(data.obj@diffexp.miRNA)
```

	FC	logratio	meanExp	pval	adj.pval
hsa-let-7a	1.383877	0.4687154	4.009944	0.424006073	0.9927973
hsa-let-7a*	1.396125	0.4814278	5.117326	0.360749796	0.9927973
hsa-let-7a-2*	-1.552633	-0.6347171	4.213117	0.178340836	0.9927973
hsa-let-7b	-1.922197	-0.9427565	4.328827	0.006425473	0.9927973
hsa-let-7b*	1.326903	0.4080632	4.373637	0.179096825	0.9927973
hsa-let-7c	2.082935	1.0586176	11.875354	0.054944083	0.9927973

1.1.2 Linear regression

In this case we perform a linear regression of the expression of the miRNA or mRNAs across time. The slope of this regression is recorded: *miRNA time* (and *mRNA time*)

```
> data.obj<-addLong(data.obj,"miRNA","time_cont","linear.regression")
> data.obj<-addLong(data.obj,"mRNA","time_cont","linear.regression")
> head(data.obj@diffexp.miRNA)
```

	slope	meanExp	pval	adj.pval
hsa-let-7a	0.1364423	4.076286	0.44821523	0.9856735
hsa-let-7a*	0.1697982	5.047076	0.29594318	0.9856735
hsa-let-7a-2*	-0.1899524	4.147486	0.24275037	0.9856735
hsa-let-7b	-0.2453290	4.059561	0.12562776	0.9821777
hsa-let-7b*	0.1527002	4.273131	0.27577319	0.9856735
hsa-let-7c	0.3306132	11.920419	0.07511731	0.9157318

1.2 Select desired miRNAs and mRNAs (optional)

Specific miRNAs or mRNAs can be selected for the correlation step. In this case we select a minimum absolute slope of 0.3. For the case of miRNAs, we fix also that the slope must be positive. **The selected miRNAs and mRNAs added in this step will be used in correlation step.**

```
> data.obj<-addSig(data.obj,"mRNA",slope=0.1)
> data.obj<-addSig(data.obj,"miRNA",slope=0.1)
```

If you have a specific list of miRNAs and/or mRNAs that you want to test, you should add them there in this step, for example:

```
> #data.obj<-addSig(data.obj,"miRNA",manual=c("hsa-miR-21","hsa-miR-21*",hsa-miR-200c"))
```

1.3 Rest of the analysis

The rest of the analysis can be done like this:

```
> data.obj<-addCorrelation(data.obj,alternative="less")
```

Correlating miRNA and mRNA

```
> data.obj<-addNet(data.obj)
```

Converting to net

```
> data(microCosm_v5_18)
> data(targetScan_v6.2_18)
> data.obj<-addDatabase(data.obj,database=c("microCosm_v5_18","targetScan_v6.2_18"))
```

```

Intersecting with database
microCosm_v5_18 database chosen
targetScan_v6.2_18 database chosen

> data.obj<-correctPval(data.obj, pval="pval",method.adj="BH")

```

Correcting p.values

2 Non-matched miRNA and mRNA data

In the case of non-matched data, no correlation can be computed, but individual p-values from differential expression analysis can be combined to only one p-value. This combined p-value highlights the miRNA-mRNA pairs more deregulated. In ideal conditions, these pairs should be similar to the ones computed by the correlation method.

```

> data(miRNA)
> data(mRNA)
> data(pheno.miRNA)
> data(pheno.mRNA)
> minimal<-new("corObject",dat.miRNA=miRNA,dat.mRNA=mRNA,
+             pheno.miRNA=pheno.miRNA,pheno.mRNA=pheno.mRNA)
> minimal.diffexp<-addDiffexp(minimal, "miRNA", classes="DvH",
+ method.dif="limma")
> head(minimal.diffexp@diffexp.miRNA)

             FC      logratio meanExp      pval adj.pval
hsa-let-7a   -1.051335 -0.072222222 1.729167 0.4750159 0.8119929
hsa-let-7a*  -1.000770 -0.001111111 2.212500 0.9964553 1.0000000
hsa-let-7a-2* 1.075080  0.104444444 1.381667 0.3319392 0.7295366
hsa-let-7b   1.038459  0.054444444 1.377500 0.5484407 0.8485038
hsa-let-7b*  -1.052145 -0.073333333 1.608333 0.6016110 0.8559239
hsa-let-7c   -1.213260 -0.278888889 9.370833 0.3188399 0.7164941

> minimal.diffexp<-addDiffexp(minimal.diffexp, "mRNA", classes="DvH",
+ method.dif="limma")
> minimal.diffexp<-addSig(minimal.diffexp,"miRNA",pval=1)
> minimal.diffexp<-addSig(minimal.diffexp,"mRNA",pval=1)
> minimal<-addNet(minimal.diffexp)

```

Converting to net

```

> minimal<-addFoldchanges(minimal, add.pvals=TRUE)
> head(minimal@net)

```

	miRNA	mRNA	logratio.miRNA	meanExp.miRNA
hsa-let-7a:1/2-SBSRNA4	hsa-let-7a 1/2-SBSRNA4		-0.07222222	1.729167
hsa-let-7a:A1BG	hsa-let-7a A1BG		-0.07222222	1.729167
hsa-let-7a:A1BG-AS1	hsa-let-7a A1BG-AS1		-0.07222222	1.729167
hsa-let-7a:A1CF	hsa-let-7a A1CF		-0.07222222	1.729167
hsa-let-7a:A2LD1	hsa-let-7a A2LD1		-0.07222222	1.729167
hsa-let-7a:A2M	hsa-let-7a A2M		-0.07222222	1.729167
	adj.pval.miRNA	pval.miRNA	logratio.mRNA	meanExp.mRNA
hsa-let-7a:1/2-SBSRNA4	0.8119929	0.4750159	0.510166495	7.519257
hsa-let-7a:A1BG	0.8119929	0.4750159	0.462108547	3.506738
hsa-let-7a:A1BG-AS1	0.8119929	0.4750159	-0.005313107	3.367482
hsa-let-7a:A1CF	0.8119929	0.4750159	-0.264825689	8.193657
hsa-let-7a:A2LD1	0.8119929	0.4750159	0.555888174	5.274122
hsa-let-7a:A2M	0.8119929	0.4750159	0.393856788	4.497770
	adj.pval.mRNA	pval.mRNA		
hsa-let-7a:1/2-SBSRNA4	0.015193262	0.0019295443		
hsa-let-7a:A1BG	0.293763899	0.1324875186		

```

hsa-let-7a:A1BG-AS1      0.982192601 0.9635309419
hsa-let-7a:A1CF         0.330361769 0.1565914786
hsa-let-7a:A2LD1       0.006971274 0.0006066992
hsa-let-7a:A2M         0.263706925 0.1147125122

```

```
> minimal<-combinePval(minimal, pval.1="pval.miRNA", pval.2="pval.mRNA", method="fisher")
```

Combining p.values

```
> minimal<-correctPval(minimal, pval="p.comb")
```

Correcting p.values

```
> head(minimal@net)
```

	miRNA	mRNA	logratio.miRNA	meanExp.miRNA
hsa-let-7a:1/2-SBSRNA4	hsa-let-7a 1/2-SBSRNA4		-0.07222222	1.729167
hsa-let-7a:A1BG	hsa-let-7a A1BG		-0.07222222	1.729167
hsa-let-7a:A1BG-AS1	hsa-let-7a A1BG-AS1		-0.07222222	1.729167
hsa-let-7a:A1CF	hsa-let-7a A1CF		-0.07222222	1.729167
hsa-let-7a:A2LD1	hsa-let-7a A2LD1		-0.07222222	1.729167
hsa-let-7a:A2M	hsa-let-7a A2M		-0.07222222	1.729167
	adj.pval.miRNA	pval.miRNA	logratio.mRNA	meanExp.mRNA
hsa-let-7a:1/2-SBSRNA4	0.8119929	0.4750159	0.510166495	7.519257
hsa-let-7a:A1BG	0.8119929	0.4750159	0.462108547	3.506738
hsa-let-7a:A1BG-AS1	0.8119929	0.4750159	-0.005313107	3.367482
hsa-let-7a:A1CF	0.8119929	0.4750159	-0.264825689	8.193657
hsa-let-7a:A2LD1	0.8119929	0.4750159	0.555888174	5.274122
hsa-let-7a:A2M	0.8119929	0.4750159	0.393856788	4.497770
	adj.pval.mRNA	pval.mRNA	p.comb	adj.pval
hsa-let-7a:1/2-SBSRNA4	0.015193262	0.0019295443	0.007327819	0.03578823
hsa-let-7a:A1BG	0.293763899	0.1324875186	0.236987693	0.39897621
hsa-let-7a:A1BG-AS1	0.982192601	0.9635309419	0.815405596	0.88576645
hsa-let-7a:A1CF	0.330361769	0.1565914786	0.267670434	0.43346386
hsa-let-7a:A2LD1	0.006971274	0.0006066992	0.002637498	0.01770014
hsa-let-7a:A2M	0.263706925	0.1147125122	0.213042395	0.37089231

3 Session Info

- R version 3.3.3 (2017-03-06), x86_64-pc-linux-gnu
- Locale: LC_CTYPE=en_US.UTF-8, LC_NUMERIC=C, LC_TIME=es_ES.UTF-8, LC_COLLATE=C, LC_MONETARY=es_ES.UTF-8, LC_MESSAGES=en_US.UTF-8, LC_PAPER=es_ES.UTF-8, LC_NAME=C, LC_ADDRESS=C, LC_TELEPHONE=C, LC_MEASUREMENT=es_ES.UTF-8, LC_IDENTIFICATION=C
- Base packages: base, datasets, grDevices, graphics, grid, methods, parallel, stats, stats4, utils
- Other packages: AnnotationDbi 1.34.4, Biobase 2.32.0, BiocGenerics 0.18.0, Category 2.38.0, CircStats 0.2-4, DESeq 1.24.0, DOSE 2.10.7, Formula 1.2-1, GO.db 3.3.0, GOSTats 2.38.1, Hmisc 4.0-3, IRanges 2.6.1, KEGG.db 3.2.3, MASS 7.3-45, Matrix 1.2-8, RamiGO 1.18.0, RankProd 2.44.0, Rcpp 0.12.10, ReactomePA 1.16.2, S4Vectors 0.10.3, VennDiagram 1.6.17, WriteXLS 4.0.0, boot 1.3-18, circlize 0.3.10, dtw 1.18-1, fgsea 0.99.7, fields 8.10, foreach 1.4.3, futile.logger 1.4.3, ggplot2 2.2.1, glmnet 2.0-9, gplots 3.0.1, graph 1.50.0, gsubfn 0.6-6, gtools 3.5.0, lattice 0.20-34, limma 3.28.21, locfit 1.5-9.1, maps 3.1.1, mclust 5.2.3, miRComb 0.8.9, miRData 0.6.1, mvoutlier 2.0.8, network 1.13.0, pROC 1.9.1, pheatmap 1.0.8, proto 1.0.0, proxy 0.4-17, scatterplot3d 0.3-40, sgeostat 1.0-27, spam 1.4-0, survival 2.40-1, verification 1.42, xtable 1.8-2
- Loaded via a namespace (and not attached): AnnotationForge 1.14.2, BiocParallel 1.6.6, DBI 0.5-1, DEoptimR 1.0-8, DO.db 2.9, GGally 1.3.0, GOSemSim 1.30.3, GSEABase 1.34.1, GlobalOptions 0.0.11, KernSmooth 2.23-15, MatrixModels 0.4-1, RBGL 1.48.1, RColorBrewer 1.1-2, RCurl 1.95-4.8, RCytoscape 1.21.1, RSQLite 1.1-2, SparseM 1.74, VIM 4.6.0, XML 3.98-1.5, XMLRPC 0.3-0, acepack 1.4.1, annotate 1.50.1, assertthat 0.1, backports 1.0.5, base64enc 0.1-3,

bitops 1.0-6, caTools 1.17.1, car 2.1-4, checkmate 1.8.2, class 7.3-14, cluster 2.0.5, codetools 0.2-15, colorspace 1.3-2, cvTools 0.3.2, data.table 1.10.4, digest 0.6.12, diptest 0.75-7, e1071 1.6-8, fastmatch 1.1-0, flexmix 2.3-13, foreign 0.8-67, fpc 2.1-10, futile.options 1.0.0, gdata 2.17.0, genefilter 1.54.2, geneplotter 1.50.0, graphite 1.18.1, gridExtra 2.2.1, gtable 0.2.0, htmlTable 1.9, htmltools 0.3.6, htmlwidgets 0.8, igraph 1.0.1, iterators 1.0.8, kernlab 0.9-25, knitr 1.15.1, laeken 0.4.6, lambda.r 1.1.9, latticeExtra 0.6-28, lazyeval 0.2.0, lme4 1.1-12, lmtest 0.9-35, magrittr 1.5, memoise 1.0.0, mgcv 1.8-17, minqa 1.2.4, modeltools 0.2-21, munsell 0.4.3, mvtnorm 1.0-5, nlme 3.1-131, nloptr 1.0.4, nnet 7.3-12, pbkrtest 0.4-6, pcaPP 1.9-61, pls 2.6-0, plyr 1.8.4, png 0.1-7, prabclus 2.2-6, quantreg 5.29, qvalue 2.4.2, rappdirs 0.3.1, reactome.db 1.55.0, reshape 0.8.6, reshape2 1.4.2, robCompositions 2.0.3, robustbase 0.92-7, rpart 4.1-10, rrcov 1.4-3, sROC 0.1-2, scales 0.4.1, shape 1.4.2, sp 1.2-4, splines 3.3.3, stringi 1.1.2, stringr 1.2.0, tcltk 3.3.3, tibble 1.2, tools 3.3.3, trimcluster 0.1-2, vcd 1.4-3, zoo 1.7-14

References

- [1] Maria Vila-Casadesús, Meritxell Gironella, and Juan José Lozano. MiRComb: An R Package to Analyse miRNA-mRNA Interactions. Examples across Five Digestive Cancers. *PLOS ONE*, 11(3):e0151127, March 2016.
- [2] Silvia Affò, Marlene Dominguez, Juan José Lozano, Pau Sancho-Bru, Daniel Rodrigo-Torres, Oriol Morales-Ibanez, Montserrat Moreno, Cristina Millán, Aurora Loeza-del Castillo, José Altamirano, Juan Carlos García-Pagán, Vicente Arroyo, Pere Ginès, Juan Caballería, Robert F. Schwabe, and Ramon Bataller. Transcriptome analysis identifies TNF superfamily receptors as potential therapeutic targets in alcoholic hepatitis. *Gut*, 62(3):452–460, March 2013.

B.3 Manual

R topics documented:

information-package	3
addCorrelation	4
addCorrelation.R	5
addDatabase	7
addDrifexp	8
addFoldchanges	10
addGlmnet	11
addLong	12
addNet	14
addScore	15
addSig	16
addSurv	17
boxplotCorrelation	18
boxplotSamples	19
checkmiRNAs	20
combinePval	21
corObject-class	22
correct-pval	24
data.obj	24
evaluate	25
GOanalysis	26
miRNA	28
miReport	28
miRNA	29
openCytoscape	30
pearson	31
pheno.miRNA	31
pheno.miRNA	31
plotGd	32
plotGd	32
plotCircos	33
plotCordist	34
plotCorrelation	35
plotDensity	36
plotGO	37
plotHlust	38
plotHeatmap	38
plotMA	39
plotNetwork	40
plotPca	41
plotSurv	42
plotVolcano	43
removeStamp	44
setOutliers	45
setSubsetCor	46
setSubsetExprs	47
summary.corObject	48

Package ‘miRComb’

May 4, 2017

Type Package

Title An R package for analyzing miRNA-mRNA interactions

Version 0.8.9

Date 2017-04-25

Author Maria Vila-Casadesús, Juanjo Lozano

Maintainer Maria Vila-Casadesús <mar.ia.vila@ci.berhd.org>

Description Find miRNA targets combining both biological and theoretical information.

License GPL-3

Depends R (>= 3.0), gplots, gtools, network, pheatmap, WriteXLS, Hmisc, mvoutlier, pROC, verification, glmnet, survival, mclust, RankProd, GOSTats, limma, scatterplot3d, RamiGO, circelize, VennDiagram, xtable, ReactomePA, figsea, DESeq, GO.db, KEGG.db, miRData (>= 0.4)

Imports AnnotationDbi

bioViews miRNA, mRNA, correlation

Encoding UTF-8

URL <http://mircomb.sourceforge.net>

RemoteType github

RemoteHost <https://api.github.com>

RemoteRepo mircomb

RemoteUsername mariavica

RemoteRef patch-devel

RemoteSha e9227e2bae2302489fb4493d27f132a6f68cf18

GitHubRepo mircomb

GitHubUsername mariavica

GitHubRef patch-devel

GitHubSHA1 e9227e2bae2302489fb4493d27f132a6f68cf18

topTable	48
translatemiRNAs	49
writeCSV	50
writeExcel	51
writeSif	52
Index	54

information-package *Find miRNA targets combining both biological and theoretical information.*

Description

Set of functions and databases useful for computing the miRNA targets from miRNA and miRNA expression data. It is based on the principle that miRNA targets need to be correlated and also be predicted on a database to be true target.

Details

```
Package: information
Type: Package
Version: 0.8.8
Date: 2017-04-18
License: GPL-3
```

Author(s)

Maria Vila-Casadesús, Juanjo Lozano
 Maintainer: Maria Vila-Casadesús <maria.vila@ciberhd.org>
 Contributor: Pau Erola (C++ code)

References

- The package has been published in PLoS ONE:
- Vila-Casadesús M, Gironella M*, Lozano JJ*, MIRComb: an R package to analyze miRNA-miRNA interactions. *Examples across five digestive cancers.* PLoS ONE 11(3): e0151127. (doi: 10.1371/journal.pone.0151127).
- Files can be downloaded here: <https://sourceforge.net/projects/mircomb/files/?source=navbar> and via GitHub using the following command: `devtools::install_github("maria.vila.ca/mircomb")`.
- The use of the package has been reported in the following publications:

- Coll M, Taghdouini AE, Perea L, Mannaerts I, Vila-Casadesús M, Blaya D, Rodrigo-Torres D, Afío S, Morales-hamez O, Graupera I, Lozano JJ, Najimi M, Sokal E, Lambrecht J, Ginés P, van Grunsven LA, Sanchez-Bru P. *Integrative miRNA and gene expression profiling analysis of human quiescent hepatic stellate cells.* Scientific Reports. Advanced online publication June 22, 2015, doi: 10.1038/srep11549.
- Bofill-De Ros X, Santos M, Vila-Casadesús M, Villanueva E, Andreu N, Dierssen M, Fillat C. *Genome-wide miR-155 and miR-802 target gene identification in the hippocampus of Tsk65Dn Down syndrome mouse model by miRNA sponges.* BMC Genomics 2015; 16:907 (advanced online publication November 6, 2015, doi: 10.1186/s12864-015-2160-6).
- Blaya D, Coll M, Rodrigo-Torres D, Vila-Casadesús M, Altamirano J, Llopis M, Graupera I, Perea L, Aguilár B, Díaz A, Banales JM, Claria J, Lozano JJ, Batañer R, Caballera J, Ginés P, Sanchez-Bru P. *Integrative MicroRNA Profiling in Alcoholic Hepatitis Reveals a Role for microRNA-182 in Liver Injury and Inflammation.* Gut (advanced online publication 10 May 2016; doi:10.1136/gutjnl-2015-311314).

The main page of miRComb package is <http://mircomb.sourceforge.net> (it includes some additional tools).

See Also

miRData, methods, gplots, gtools, network, pheatmap, writeXLS, hmisc, mvoutlier, pROC, verification, glmnet, survival, mclust, RankProd, Gostats, limma, scatterplot3d, Ram160, circlize, VennDiagram, xtable, ReactomePA, fgsea, DESeq, GO.db, KEGG.db

addCorrelation *Correlate miRNA and mRNA expression*

Description

The function correlates miRNA and mRNA expression from a corObject and fills the cor and pval slots.

Usage

```
addCorrelation(obj, method = "pearson", subset.miRNA = obj$sig.miRNA,
subset.miRNA = obj$sig.mRNA, common = NULL, alternative = "less")
```

Arguments

obj	a corObject
method	method used for computing correlation: "pearson" or "spearman".
subset.miRNA	Optional, character vector with the names of the miRNAs to correlate. It is recommended that miRNAs are added using addSig function.
subset.mRNA	Optional, character vector with the names of the mRNAs to correlate. It is recommended that mRNAs are added using addSig function.

common Optional, character vector with the names of the samples to correlate (the samples must appear in both miRNA and mRNA datasets).

alternative specification of the alternative hypothesis: "less" (default), "two-sided" or "greater".

Value

corObject with the slots cor and pval filled.

Note

addCorrelation.R is the slow version of this function, but has the option to compute if there are any influential samples affecting the correlation values.

See Also

corObject-class, cor, addCorrelation.R, addSig

Examples

```
data(data.obj)
data.obj.correlated@cor
subset.miRNA=c("hsa-let-7e", "hsa-miR-122"), subset.miRNA=c("A1BG", "A1CF")
data.obj.correlated@cor
data.obj.correlated@pval
```

addCorrelation.R *Correlation, old version*

Description

Correlation, old version

Usage

```
addCorrelation.R(obj, method = "pearson", subset.miRNA = obj@sig.miRNA,
subset.miRNA = obj@sig.miRNA, common = NULL, d.influences = FALSE,
alternative = "two-sided")
```

Arguments

obj a corObject.

method method used for computing correlation: "pearson" or "spearman".

subset.miRNA Optional, character vector with the names of the miRNAs to correlate. It is recommended that miRNAs are added using addSig function.

subset.miRNA Optional, character vector with the names of the mRNAs to correlate. It is recommended that mRNAs are added using addSig function.

common Optional, character vector with the names of the samples to correlate (the samples must appear in both miRNA and mRNA datasets.)

d.influences compute a matrix with the Cook's Distance of each sample in each miRNA-mRNA correlation.

alternative specification of the alternative hypothesis: "less" (default), "two-sided" or "greater".

Details

A more complete version of the addCorrelation function, but significantly slower (specially kendall correlation). Use always addCorrelation function whenever it is possible.

If d.influences = TRUE, a 3-dimension matrix is added to the info slot, labeled "influencing.sample". First dimension: miRNA names; second dimension: mRNA names; third dimension: sample names; fill: Cook's Distance for a specific sample in a specific miRNA-mRNA linear regression (defined by the dimension label-names).

Value

A corObject with the slots "cor" and "pval" filled. Optionally, a matrix named "influencing.sample" is added to the info slot.

Note

This function can take a long time to complete when is applied to large datasets.

See Also

corObject-class, cor, addCorrelation, addSig, cooks.distance

Examples

```
data(data.obj)
data.obj.correlated@cor
data.obj.correlated@pval
data.obj.correlated@info[["influencing.sample"]]
data.obj.correlated@cor
data.obj.correlated@pval
data.obj.correlated@info[["influencing.sample"]]
data.obj.correlated@cor
data.obj.correlated@pval
data.obj.correlated@info[["influencing.sample"]]
```

addDatabase *Intersect correlations with an external database.*

Description

For each miRNA-mRNA pair, add if this pair has been predicted as miRNA-mRNA interaction according to the desired external databases.

Usage

```
addDatabase(obj, database, pval.ref=1, dat.sum=1)
```

Arguments

- obj a corObject with a cytofile slot already defined.
- database "microCosm_v5_18_numeric"; or a character vector including: "microCosm_v5_18", "targetScan_v6_2_18", "PITA_v6_11_Top", "miRDB_v5_0_21", "miRSVR_aug10_17", "targetScan_v7_0_21", "miRDB_v5_0_21_mouse", "miRSVR_aug10_15_mouse"; or a character with the name of the data.frame containing the database.
- pval.ref only for "microCosm_v5_18_numeric": *p* value to set if no information is given. Default: 1.
- dat.sum only if "microCosm_v5_18_numeric" is not selected. For future purposes, the minimum number of concurrences across databases that determine that a miRNA-mRNA pair is bioinformatically predicted. By default: 1.

Value

A corObject with new columns added to the net slot. The name and content of the column is dependent on the database selected:

- `microCosm_v5_18_numeric`: `pval.database`: *p* value of the microCosm database. If there is no *p* value in the database, then the value `pval.ref` is assigned
- A character vector including the names of the databases: the following columns are created:
 - For each database: column called "dat.xxx", where "xxx" is the name of the database. The values are "0" (no target) or "1" (target).
 - A column called "dat.sum" that is, for each row, the sum of the values of all "dat.xxx" columns.

Note

If the database is a customised data.frame, the row names must follow the format "miRNA_name:mRNA_name" (check `head(microCosm_v5_18)` or `head(targetScan_v6_2_18)`). miRNA names of the corObject and the database must be of the same miRBase version. See `\link{TransLatiemRNAs}` for more details

References

TargetScan: <http://www.targetscan.org/MicroCosm>; <http://www.ebi.ac.uk/enright-srv/microcosm/html/docs/targets/v5/MIRDB>; <http://mirdb.org/>; MIRSVR: <http://www.microna.org/microna/home.do> MiRBase: <http://www.mirbase.org/>

See Also

`data(microCosm_v5_18)`, `data(targetScan_v6_2_18)`, `addNet`, `transLatiemRNAs`

Examples

```
## Load databases
data(microCosm_v5_18)
data(targetScan_v6_2_18)

## Load corObject example
data(data.obj)

## numeric example
data.obj.dat.added.numeric <- addDatabase(data.obj, "microCosm_v5_18_numeric")
head(data.obj.dat.added.numeric@net)

## non-numeric example, multiple databases
data.obj.dat.added.multiple <- addDatabase(data.obj,
c("microCosm_v5_18", "targetScan_v6_2_18"))
head(data.obj.dat.added.multiple@net)
```

addDiffExp

Calculate differential expression

Description

Calculate miRNA or mRNA differential expression from a corObject

Usage

```
addDiffExp(obj, dataset, classes, method.diff = "t.test", method.adj = "BH",
var.t.test = FALSE, trend = FALSE, norm = NULL)
```

Arguments

- obj a corObject
- dataset "miRNA" or "mRNA"
- classes column name of the pheno.miRNA or pheno.mRNA encoding the contrast to make. The column must contain "0" (reference) and "1" (case). Missing values are also allowed. More levels are accepted for "anova" method.
- method.diff method used for differential expression: "t.test", "wilcoxon", "limma", "rankprod", "anova" or "only.fc"

method.adjust multiple testing correction method used (only for method.diff = "t.test", "wilcoxon", "limma" or "anova". One of "holm", "hochberg", "hommel", "bonferroni", "BH", "BY", "FDR", "none" (see \link{p.adjust}). TRUE or FALSE (default). Only for method.diff = "t.test". If TRUE, considers equality of variances in a T-test.

trend if TRUE use "limma-trend" method (only for method.diff = "t.test"). Recommended for log-normalised counts

norm NULL (default) (in implementation, this feature will help to integrate and normalise NGS data for the correlation)

Value

a corObject with a diffExp.mRNA or diffExp.mRNA slot added, where the rownames of the data.frame are the names of the miRNAs or mRNAs. This slot is a data.frame with the following columns:

- FC: foldchange
- logratio: logratio
- meanExp: mean value log2-expression for the probe across all samples.
- pval: p values. In RankProd the minimum of both p values is reported.
- adj.pval: p values corrected for multiple testing. In the RankProd the pfp (estimated percentage of false positives, which are, in theory, equivalent to False Discovery Rate) are added.

* If method "anova" is selected, then the proportion of SS related to factor is showed instead of FC and logratio. ** If method "only.fc" is selected, then only FC and logratio are computed (useful if there is no enough samples to perform a statistical test).

See Also

`corObject-class`, `limma`, `package:RankProd`, `voom`, `t.test`, `wilcoxon`, `aov`, `p.adjust`

Examples

```
data(miRNA)
data(mRNA)
data(pheno.miRNA)
data(pheno.mRNA)

minimal<-new("corObject",dat.miRNA=miRNA,dat.mRNA=mRNA,
pheno.miRNA=pheno.miRNA,pheno.mRNA=pheno.mRNA)

minimal.diffExp<-addDiffExp(minimal, "miRNA", classes="DVH",
method.diff="limma")
head(minimal.diffExp@diffExp.miRNA)

minimal.diffExp<-addDiffExp(minimal.diffExp, "mRNA", classes="DVH",
method.diff="limma")
head(minimal.diffExp@diffExp.mRNA)
```

addFoldChanges *Add foldchanges to the net slot*

Description

Adds information regarding to miRNA and mRNA differential expression on the net slot of a corObject

Usage

```
addFoldChanges(obj, add.pvals = FALSE)
```

Arguments

obj a corObject with a net, diffExp.miRNA and diffExp.mRNA slots already defined.

add.pvals TRUE or FALSE (default). If TRUE, p values are added to the net slot.

Value

a corObject with the following columns added in the net slot:

- logratio.miRNA
- logratio.mRNA
- meanExp.miRNA
- meanExp.mRNA

plus, if add.pvals=TRUE:

- pval.miRNA
- pval.mRNA
- adj.pval.miRNA
- adj.pval.mRNA

See Also

`corObject-class`, `addDiffExp`, `addDiffExp`.

Examples

```
## obtain minimal net slot
data(data.obj)
data.obj@net <- data.obj@net[, -c(5:ncol(data.obj@net))]
head(data.obj@net)

## add the foldchanges from diffExp.miRNA and diffExp.mRNA slots
data.obj<-addFoldChanges(data.obj)
head(data.obj@net)
```

addGlmnet	<i>Add elastic net estimates</i>
-----------	----------------------------------

Description

Add elastic net estimates to a corObject

Usage

```
addGlmnet(obj, response = "mRNAs", alpha = 0.5, upper.limit = 0,
          cluster = "manual", plot = TRUE)
```

Arguments

obj a corObject
 response miRNA, mRNA or mRNAs
 alpha
 upper.limit
 cluster
 plot

References

<https://cran.r-project.org/web/packages/glmnet/index.html>

See Also

[corobject-class, package:glmnet](#)

Examples

```
##---- Should be DIRECTLY executable !! ----
##-- ==> Define data, use random,
##--or do help(data=index) for the standard data sets.
```

addLlong	<i>Time analysis</i>
----------	----------------------

Description

Time analysis of a corObject

Usage

```
addLlong(obj, dataset, classes, method.dif = "time.point", method.adj = "BH",
         var.t.test = FALSE, trend = FALSE)
```

Arguments

obj corObject
 dataset "miRNA" or "mRNA"
 classes column name of the pheno. miRNA or pheno. mRNA encoding codification of time variable. The column must contain "0" (L_init) and "1" (L_fin), or the specific times. Missing values are also allowed.
 method.dif "time.point" or "linear_regression"
 method.adj Multiple testing correction method used. One of "holm", "hochberg", "hommel", "bonferroni", "BH", "BY", "fdr", "none" (see [p.adjust](#)).
 var.t.test TRUE or FALSE (default). If TRUE, considers equality of variances in a T-test.

Details

If method.dif="time.point", both times are compared using a t-test, time variable should be encoded as kbd"0" (L_init) and "1" (L_fin).

Value

a corObject with a diffexp.miRNA or diffexp.mRNA slot added, where the rownames of the data.frame are the names of the miRNAs or mRNAs. Depending on the selected method, the diffexp is a data.frame with the following columns:

If method.dif="time.point":

- FC: foldchange between the two selected times.
- logratio: logratio between the two selected times.
- meanExp: mean value log2-expression for the probe across all samples.
- pval: p values. In RankProd the minimum of both p values is reported.
- adj.pval: p values corrected for multiple testing. In the RankProd the p values (estimated percentage of false positives, which are, in theory, equivalent to False Discovery Rate) are added.

If method.dif="linear_regression":

- slope: slope of the linear regression.
- meanExp: mean value log₂-expression for the probe across all samples.
- pval: *p* values associated to the slope of the linear regression.
- adj.pval: *p* values corrected for multiple testing.

See Also

[addDiffexp](#)

Examples

```
#####
### data special preparation ##
#####

data(miRNA)
data(mRNA)

### simulated phenotypical: 3 samples in 4 time-points, one treatment.

pheno.miRNA<-data.frame(sample=rep(1:3, each=4),
  time=rep(c(1,2,5,9),3),
  time.alt=rep(c(0,NA,NA,1),3))
rownames(pheno.miRNA)<-paste(pheno.miRNA$sample,pheno.miRNA$time,sep="_")

pheno.mRNA<-pheno.miRNA #same sample distribution

## modify sample names from miRNA and mRNA dataset
colnames(mRNA)<-rownames(pheno.miRNA)
colnames(mRNA)<-rownames(pheno.mRNA)

#####
### start of the example ##
#####

minimal<-new("corObject",dat.miRNA=miRNA,dat.mRNA=mRNA,
  pheno.miRNA=pheno.miRNA,pheno.mRNA=pheno.mRNA)

## comparing times 1 (t.init) and 9 (t.fin):
minimal.diffexp<-addLong(minimal, "miRNA", classes="time.alt",
  method.diff="time.point")
head(minimal.diffexp@diffExp.miRNA)

## treated as linear regression:
minimal.diffexp<-addLong(minimal, "miRNA", classes="time",
  method.diff="linear.regression")
head(minimal.diffexp@diffExp.miRNA)
```

addNet *Create a net slot*

Description

Creates and fills the net slot from a corObject.

Usage

```
addNet(obj)
```

Arguments

obj a corObject with the slots pval and cor already defined

Details

The net slot is a data.frame where each row represents a potential miRNA-mRNA interaction and the columns contain all the available information for each pair. The data.frame is sorted by miRNA name and then by mRNA name. See [addFOLDchanges](#), [addDatabase](#), [addScore](#) and [correctPval](#) functions that help to add more information to the net slot.

Value

A data.frame with the cytofile slot filled.

See Also

[corObject-class](#), [addFOLDchanges](#), [addDatabase](#), [code\(addScore\)](#) and [correctPval](#)

Examples

```
data(data.obj)
data.obj@net <- data.frame()
## create the minimal net slot
data.obj<-addNet(data.obj)
head(data.obj@net)
```

<i>Create a score</i>

Description

Create a score for each miRNA:miRNA pair associated to their relative miRNA and mRNA Fold-Change

Usage

```
addScore(obj)
```

Arguments

obj a corObject, with a net slot containing logratio.miRNA and logratio.mRNA columns.

Value

a corObject in which a column containing the score values (score) has been added to the net slot.

Note

This score is needed by [plotNetwork](#) function.

See Also

[corObject-class](#), [addFoldChanges](#), [addDiffExp](#), [plotNetwork](#)

Examples

```
data(data.obj)
data.obj@net$score<-NULL
head(data.obj@net)

data.obj<-addScore(data.obj)
head(data.obj@net)
```

<i>Select significant miRNAs or mRNAs</i>

Description

Select significantly deregulated miRNAs or mRNAs that will be used for correlating their expression.

Usage

```
addSig(obj, dataset, FC=NA, logratio=foldchange2logratio(FC), slope=NA, pval=NA,
adj.pval=NA, min.meanExp=NA, up=FALSE, dw=FALSE, manual=NULL)
```

Arguments

obj a corObject with the slots diffExp.miRNA and diffExp.mRNA already defined
 "miRNA" or "mRNA"
 dataset minimum FoldChange (in absolute value)
 FC minimum logratio (in absolute value)
 logratio minimum logratio (in absolute value)
 slope for longitudinal-regression analysis, minimum absolute slope
 pval maximum uncorrected p value.
 adj.pval maximum corrected p value.
 min.meanExp minimum mean expression.
 up Select only upregulated items (TRUE or FALSE).
 dw Select only downregulated items (TRUE or FALSE).
 manual character vector with miRNA or mRNA names.

Value

The same corObject with the slots sig.miRNA or sig.mRNA including the names of the miRNAs or mRNAs that will be used for correlation.

See Also

[corObject-class](#), [addDiffExp](#), [addCorrelation](#)

Examples

```
data(data.obj)

## select the significant miRNAs and mRNAs
data.obj<-addSig(data.obj, "miRNA", adj.pval=0.05)
data.obj<-addSig(data.obj, "miRNA", adj.pval=0.05, FC=1.5)
```

addSurv *Calculate hazard ratio*

Description

Calculate hazard ratio of miRNAs or mRNAs from a corObject using a Cox proportional hazards regression.

Usage

```
addSurv(obj, dataset, time, event, adjusting = NULL)
```

Arguments

obj a corObject
 dataset "miRNA" or "miRNA"
 time colname of the time column in the pheno. miRNA or pheno. mRNA slot
 event colname of the event column in the pheno. miRNA or pheno. mRNA slot
 adjusting (optional) colname of the adjusting variables that will be used

Value

a corObject with a diffExp. miRNA or diffExp. mRNA slot added. The slot is a data.frame in which row names are the names of the miRNAs or mRNAs and has the following columns:

- coef: risk associated to the expression of the miRNA/miRNA
- pval: p values
- adj.pval: p values corrected for multiple testing (FDR)

See Also

package: [survival](#), [coxph](#), [plotSurv](#)

Examples

```
data(miRNA)
data(mRNA)
data(pheno.miRNA)
data(pheno.mRNA)

minimal<-new("corObject", dat.miRNA=miRNA, dat.mRNA=mRNA,
             pheno.miRNA=pheno.miRNA, pheno.mRNA=pheno.mRNA)

minimal@pheno.miRNA$time<-runif(nrow(minimal@pheno.miRNA), 1, 20)
minimal@pheno.miRNA$event<-rbinom(nrow(minimal@pheno.miRNA), 1, 0.5)
minimal@pheno.miRNA$ai<-rnorm(nrow(minimal@pheno.miRNA), 1, 0.5)
minimal@pheno.miRNA$a2<-rnorm(nrow(minimal@pheno.miRNA), 1, 0.5)
```

```
#plotSurv(minimal, "miRNA", "hsa-let-7c", "time", "event")
minimal.diffExp<-addSurv(minimal, "miRNA", "time", "event", c("a1", "a2"))
head(minimal.diffExp@diffExp.miRNA)
minimal.diffExp<-addSurv(minimal, "miRNA", "time", "event", c("a1"))
head(minimal.diffExp@diffExp.miRNA)
minimal.diffExp<-addSurv(minimal, "miRNA", "time", "event")
head(minimal.diffExp@diffExp.miRNA)
```

boxplotCorrelation *Plot boxplot and correlation of a miRNA-mRNA pair*

Description

Given a miRNA and a mRNA, plots in the same frame a boxplot of the miRNA expression, a boxplot of the mRNA expression and a scatterplot of showing the correlation between the two.

Usage

```
boxplotCorrelation(obj, miRNA, mRNA, col.color = 1, pos.legend = "topright",
                  colors = c("turquoise", "violet"), ...)
```

Arguments

obj a corObject
 miRNA character with the name of the miRNA to plot
 mRNA character with the name of the mRNA to plot
 col.color number or name of the column in the pheno. miRNA or pheno. mRNA slot which define the grouping variables
 colors character vector indicating the colors that will be used for each variable from col.color
 pos.legend legend position: "topright", "bottomright", "topleft", or "bottomleft"
 ... other parameters

Value

a plot that includes the following figures:

- Top-right: boxplot of the miRNA
- Bottom-left: boxplot of the mRNA
- Bottom-right: scatter plot of the correlation (see [plotCorrelation](#)).

See Also[plotCorrelation](#)**Examples**

```
data(data.obj)
boxplotCorrelation(data.obj, "hsa-miR-107", "ACPF", pos.legend="topleft")
```

```
boxplotSamples          Boxplots of samples expression
```

Description

Plot boxplots of the miRNA or mRNA expression, for each sample. It is possible to colour the samples according to a phenotypical description.

Usage

```
boxplotSamples(obj, subset, col.color = 1, las = 1, colors = c("turquoise", "violet"))
```

Arguments

```
obj          a corObject
subset       "miRNA" or "mRNA"
col.color    number or name of the column in the pheno.miRNA or pheno.mRNA slot which
              define the grouping variables
las          las parameter
colors       character vector indicating the colors that will be used for each variable from
              col.color
```

See Also[plotCordist](#), [boxplotCorrelation](#)**Examples**

```
data(data.obj)
boxplotSamples(data.obj, "miRNA", col.color=1)
boxplotSamples(data.obj, "mRNA", col.color=1)
```

```
checkmiRNAs          Compare a list of miRNAs with different miRBase versions
```

Description

Compare a list of miRNAs with different miRBase versions and plot the percentage of coincidences across them.

Usage

```
checkmiRNAs(v.miRNAs, to.dataframe = FALSE)
```

Arguments

```
v.miRNAs      character vector with the miRNA names to test.
to.dataframe  FALSE (default) or TRUE. If TRUE, give a dataframe with the name of all miRBase
              versions and percentage of coincidences. If FALSE, only plots the result.
```

Details

This function needs miRdata v.0.5.0 or greater to work. Update miRdata if needed.

Value

If to.dataframe=TRUE a dataframe with the name of all miRBase versions and percentage of coincidences.

If to.dataframe=FALSE a bar plot showing the name of all miRBase versions (x-axis) and percentage of coincidences (y-axis).

See Also[translatemiRNAs](#)**Examples**

```
comp<-c("hsa-miR-20a", "hsa-miR-21", "hsa-miR-22",
        "hsa-miR-23a", "hsa-miR-24", "hsa-miR-25", "hsa-miR-26a",
        "hsa-miR-26b", "hsa-miR-27a", "hsa-miR-28-5p", "hsa-miR-28-3p",
        "hsa-miR-29a", "hsa-miR-30a")
result_in_table<-checkmiRNAs(comp, to.dataframe=TRUE)
# plot the result
result_in_table
```

combinePval *Combine p values*

Description

Combine two p values into one.

Usage

```
combinePval(obj, pval.1 = "pval", pval.2 = "pval.database", method="stouffer",
w=c(1,1))
```

Arguments

obj corObject with a net slot defined. It must have at least two columns with p values to combine.

pval.1 column name (from the net slot) of the first p value. By default: the p value of the correlation.

pval.2 column name (from the net slot) of the second p value. By default: the p value from MicroCosm database.

method "stouffer" (default) or "fisher"

w numeric vector of length two indicating the respective weights that will be applied to Stouffer combination. By default: no weighting.

Details

Stouffer and Weighted Stouffer (Lipták) combination is computed according to:

$$P_{comb} = 1 - \Phi \left(\frac{1}{\sqrt{w_1^2 + w_2^2}} (w_1 (\Phi^{-1}(1 - p_1)) + w_2 (\Phi^{-1}(1 - p_2))) \right)$$

where,

$$\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

Fisher combination is computed according to:

$$t = -2 (\ln p_1 + \ln p_2) \sim \chi^2_4$$

Value

a corObject in which a column containing the combined p values has been added to the net slot.

References

For more information about the combination methods, see:

Zaykin D.Y. Optimally weighted Z-test is a powerful method for combining probabilities in meta analysis. *Journal of Evolutionary Biology*, 2011.

Gade G., Porzelius C., Fjällh M., Bråse J.C., Wuttig D., Kuner R., Binder H., Sillmann H., and Beisbarth T. Graph based fusion of miRNA and mRNA expression data improves clinical outcome prediction of prostate cancer. *BMC Bioinformatics*, 12(488), 2011.

See Also

`combinePval`, `p.adjust`, `addNet`

Examples

```
data(data.obj)

## add column pval.database
data(enteroCosm_v5_18)
data.obj<-addDatabase(data.obj, "microCosm_v5_18_numeric")

## combine the two p-values
data.obj<-combinePval(data.obj, pval.1="pval", pval.2="pval.database")
head(data.obj@net)
```

corObject-class *Class "corObject"*

Description

Class object for storing all the information for a miRNA-mRNA correlation analysis.

Details

In this version, miRNAs should be preferentially named according to miRBase 17, and mRNAs should be named according to HUGO gene symbol nomenclature.

Objects from the Class

Objects can be created by calls of the form `new("corObject", ...)`.

Slots

dat.miRNA: Object of class "matrix". Contains the miRNA expression (rows for miRNAs and columns for samples).

dat.mRNA: Object of class "matrix". Contains the mRNA expression (rows for mRNAs and columns for samples).

pheno.miRNA: Object of class "data.frame". Rows for samples and columns for phenotypical information.

pheno.mRNA: Object of class "data.frame". Rows for samples and columns for phenotypical information.

cor: Object of class "matrix". Rows for miRNAs and columns for mRNAs.

pval: Object of class "matrix". Rows for miRNAs and columns for mRNAs.

net: Object of class "data.frame". Rows for unique miRNA:mRNA pairs and columns for their corresponding information (at least: miRNA name, mRNA name, coefficient of correlation and p value).

diffexp.miRNA: Object of class "data.frame". Rows for miRNAs and columns for their corresponding information (usually FC, logratio, mean expression, p value and corrected p value).

diffexp.mRNA: Object of class "data.frame". Rows for mRNAs and columns for their corresponding information (usually FC, logratio, mean expression, p value and corrected p value).

sig.miRNA: Object of class "vector". Vector specifying the miRNAs that are used for correlation.

sig.mRNA: Object of class "vector". Vector specifying the mRNAs that are used for correlation.
GO.results: Object of class "list". It contains the results of a GO analysis.

info: Object of class "list". It contains the information of the tests and functions used.

Methods

No methods defined with class "corObject" in the signature.

Examples

```
## minimal corObject:
data(miRNA)
data(mRNA)
data(pheno.miRNA)
data(pheno.mRNA)

minimal<-new("corObject",dat.miRNA=miRNA,dat.mRNA=mRNA,
pheno.miRNA=pheno.miRNA,pheno.mRNA=pheno.mRNA)

str(minimal)

## corObject with more slots:
data(data.obj)
str(data.obj)
```

correctPval *Correct p values*

Description

Correct p values from one column of a net slot.

Usage

```
correctPval(obj, method.adj = "BH", pval= "pval")
```

Arguments

obj a corObject with a net slot already defined.

method.adj one of "holm", "hochberg", "hommel", "bonferroni", "BH", "BY", "fdr", "none".

pval name of the column with the p values to correct.

Value

a corObject in which a column named adj.pval, which contains the corrected p values, has been added to the net slot.

See Also

[combinePval.p.adjust.adjust](#)

Examples

```
data(data.obj)
data.obj <- correctPval (data.obj, method.adj="BH", pval="pval")
head(data.obj$net)
```

data.obj *Example of a corObject*

Description

Example of a corObject that was used for the main vignette.

Usage

```
data("data.obj")
```

Format

The format is: Formal class 'corObject' [package 'miRComb'] with 13 slots: '@ dat.miRNA : numeric matrix .@ dat.miRNA : numeric matrix .@ pheno.miRNA : 'data.frame' .@ pheno.miRNA : 'data.frame' .@ cor : numeric matrix .@ pval : numeric matrix .@ net : 'data.frame' .@ diff-exp.miRNA : 'data.frame' .@ diffexp.miRNA : 'data.frame' .@ sig.miRNA : character .@ sig.miRNA : character .@ GO.results : list .@ info : list

Source

Modified from: Aflo S, Dominguez M, Lozano JJ, Sancho-Bru P et al. *Transcriptome analysis identifies TNF superfamily receptors as potential therapeutic targets in alcoholic hepatitis*. Gut 2013 Mar;62(3):452-60. PMID: 22637703

References

Aflo S, Dominguez M, Lozano JJ, Sancho-Bru P et al. *Transcriptome analysis identifies TNF superfamily receptors as potential therapeutic targets in alcoholic hepatitis*. Gut 2013 Mar;62(3):452-60. PMID: 22637703

Examples

```
data(data.obj)
str(data.obj)
```

evaluate

Evaluate function

Description

Evaluate in detail (STILL IN TEST!)

Usage

```
evaluate(obj, method = c("hypergeometric", "logistic", "GSEA"), databases = "all",
adj.pval = 0.05, plot = TRUE, miRNAs = "all", nperm = nperm)
```

Arguments

- obj
- method
- databases
- adj.pval
- plot
- miRNAs
- nperm

Examples

```
##---- Should be DIRECTLY executable !! ----
##-- ==> Define data, use random,
##--or do help(Data=index) for the standard data sets.

## The function is currently defined as
```

GOanalysis

GO and KEGG enrichment analysis

Description

GO and KEGG enrichment analysis of miRNAs from selected miRNA-mRNA interactions.

Usage

```
GOanalysis(obj, type, ontology, pval.cutoff = 0.05,
dat.sum = obj$info["dat.sum"]], score.cutoff = NULL, sub.miRNA = NULL,
exclude.miRNA = NULL, sub.miRNA = NULL, organism = "human", FC = NULL,
up = FALSE, dw = FALSE, add.miRNA = FALSE)
```

Arguments

- obj a corObject with a net slot already defined
- type "GO", "KEGG" or "REACTOME"
- ontology If type is "GO", the ontology to be analysed: "bp" (Biological Process), "CC" (Cellular Component) or "MF" (Molecular Function). If the type is "KEGG", write "KEGG"; if it is "REACTOME", write "REACTOME".
- pval.cutoff maximum p value of selected miRNA-mRNA pairs.
- dat.sum minimum concurrences of the miRNA-mRNA pairs across the database(s) used
- score.cutoff maximum score allowed
- sub.miRNA (optional) character vector, names of the miRNAs to limit the targets.
- exclude.miRNA (optional) character, use only these targets.
- sub.miRNA (optional) character, use only these targets.
- FC (optional) minimum FC for the miRNAs.
- up if TRUE, select only upregulated miRNAs.
- dw if TRUE, select only upregulated miRNAs.
- organism "human" or "mouse".
- add.miRNA if TRUE, add the miRNAs that are regulating the selected miRNAs (only if correlation p-values are computed)

Value

a corObject, with an item of the GO_results slot added. The item is a data.frame with the named "type:ontology" and with the following columns:

- Ontology: "BP" (Biological Process), "CC" (Cellular Component), "MF" (Molecular Function), "KEGG" or "REACTOME"
- ID: term ID
- Pvalue: p value
- OddsRatio: number of mRNAs found/number of expected mRNAs
- ExpCount: expected number of mRNAs
- Count: number of mRNAs in the selected category
- Size: total number of mRNAs in the selected category
- Term: term name
- fdr: corrected p value with BH method
- genescat: mRNAs in the category
- (optional) miRNAs: miRNAs regulating these mRNAs

References

Falcon S and Gentleman R. Using GOstats to test gene list for GO term association. *Bioinformatics*, 23(7):257-8, 2007.

Yu G. ReactomePA: Reactome Pathway Analysis. R package version 1.10.1.

See Also

`package:GOstats`, `package:ReactomePA`

Examples

```
data(data.obj)
data.obj<-GOanalysis(data.obj,"GO","MF",pval.cutoff=0.05,dat.sum=1)
head(data.obj@GO.results[["GO:MF"]])

data.obj<-GOanalysis(data.obj,"KEGG","KEGG",pval.cutoff=0.05,dat.sum=1)
head(data.obj@GO.results[["KEGG:KEGG"]])
```

miRNA *miRNA data expression*

Description

miRNA data expression that can be used for example, in log2-intensity units

Usage

```
data(miRNA)
```

Format

The format is: num [1:1733, 1:12] 1.86 2.4 1.35 1.25 1.76 ... - attr(*, "dimnames")=List of 2 ..\$: chr [1:1733] "hsa-let-7a" "hsa-let-7a*" "hsa-let-7a-2*" "hsa-let-7b"\$: chr [1:12] "Control_1" "Control_2" "Control_3" "Case_1" ...

Source

Modified from Sancho-Bru P group data.

Examples

```
data(miRNA)
head(miRNA)
```

mkReport *Creates a pdf report*

Description

Creates a pdf report summarizing the contents of the corObject

Usage

```
mkReport(obj, file, title = "Default \\texttt{miRComb} output", dat.sum.table = NULL)
```

Arguments

```
obj          a corObject
file         name of the file, for example "myExampleReport"
title        Title of the report
dat.sum.table Minimum dat.sum that will be applied to Table 9 and Figure 6.
```

Details

Documents myExampleReport.tex and myExampleReport.pdf will be created on the working directory.

See PLoS ONE publication for more details: Vila-Casadesús et al., "MIRComb: an R package to analyse miRNA-miRNA interactions. Examples across five digestive cancers". PLoS ONE, 2016.

Note

This function only works in Linux computers, with LaTeX and texlive already configured.

Some known problems and solutions:

- If this happens: ! => Fatal error occurred, no output PDF file produced!
Try: sudo apt-get install texlive-recommended-fonts or just sudo apt-get install texlive-full
- If you have problems with xcolor package: xcolor.sty not found sudo apt-get install latex-xcolor
- If you have problems with tikz package: tikz.sty not found sudo apt-get install pgf

Examples

```
## do not run
#data(data.obj)
#mkReport(data.obj, "myExampleReport")
## documents myExampleReport.tex and myExampleReport.pdf will be created
```

miRNA

miRNA data expression

Description

miRNA data expression that can be used for example, in log2-intensity units

Usage

```
data(miRNA)
```

Format

```
The format is: num [1:18900, 1:12] 7.06 3.38 3.23 8.41 4.63 ... - attr(*, "dimnames")=List of 2 ..$
: chr [1:18900] "1J2-SBSRNA4" "A1BG" "A1BG-AS1" "A1CF" ... ..$
: chr [1:12] "Control_1"
"Control_2" "Control_3" "Case_1" ...
```

Source

Modified from: Afío S. Dominguez M, Lozano JJ, Sancho-Bru P et al. *Transcriptome analysis identifies TNF superfamily receptors as potential therapeutic targets in alcoholic hepatitis. Gut* 2013 Mar;62(3):452-60. PMID: 22637703

References

Afío S, Dominguez M, Lozano JJ, Sancho-Bru P et al. Transcriptome analysis identifies TNF superfamily receptors as potential therapeutic targets in alcoholic hepatitis. *Gut* 2013 Mar;62(3):452-60. PMID: 22637703

Examples

```
data(miRNA)
head(miRNA)
```

openCytoscape *Open cytoscape session with the network of miRNA-miRNA interactions*

Description

Open cytoscape session and automatically load the network of miRNA-miRNA interactions.

Usage

```
openCytoscape(obj = NULL, pval.cutoff = 0.05, dat.sum =
obj$info["dat.sum"], file = NULL, cytoscape.folder =
"/home/mvila/Cytoscape_v2.8.3", sub.miRNA = NULL,
sub.miRNA = NULL, add.other = NULL, expand = FALSE)
```

Arguments

```
obj a corObject
pval.cutoff minimum corrected p value of selected miRNA-miRNA interactions
dat.sum minimum occurrences across databases of selected miRNA-miRNA interactions
file name of the ".sif" network file that will be written. If NULL, file "network_default.sif"
will be created
cytoscape.folder path where "cytoscape.jar" file is located
sub.miRNA character vector with the restricted miRNA
sub.miRNA character vector with the restricted mRNA
add.other other
expand expand the network
```

References

<http://cytoscape.org/>

See Also

[writeSif](#)

pearson 31

Examples

```
##openCytoscape(data.obj)
```

```
pearson
```

```
Pearson correlation with C++ code
```

Description

Function written in C++ that computes pearson correlation and p values. It is used by `addCorrelation`.

See Also

[addCorrelation](#)

```
pheno.mRNA
```

```
Phenotypical miRNA information
```

Description

Phenotypical miRNA information

Usage

```
data(pheno.mRNA)
```

Format

A data frame with 12 observations on the following 2 variables.

group a factor with levels H (Healthy) D (Disease)

DVH a numeric vector

Source

Modified from: *Affò S, Dominguez M, Lozano JJ, Sancho-Bru P et al. Transcriptome analysis identifies TNF superfamily receptors as potential therapeutic targets in alcoholic hepatitis. Gut 2013 Mar;62(3):452-60. PMID: 22637703*

References

Affò S, Dominguez M, Lozano JJ, Sancho-Bru P et al. Transcriptome analysis identifies TNF superfamily receptors as potential therapeutic targets in alcoholic hepatitis. *Gut* 2013 Mar;62(3):452-60. PMID: 22637703

Examples

```
data(pheno.mRNA)  
pheno.mRNA
```

32

plot3d

```
pheno.mRNA
```

```
Phenotypical mRNA information
```

Description

Phenotypical mRNA information

Usage

```
data(pheno.mRNA)
```

Format

A data frame with 12 observations on the following 2 variables.

group a factor with levels H (Healthy) D (Disease)

DVH a numeric vector

Source

Modified from: *Affò S, Dominguez M, Lozano JJ, Sancho-Bru P et al. Transcriptome analysis identifies TNF superfamily receptors as potential therapeutic targets in alcoholic hepatitis. Gut 2013 Mar;62(3):452-60. PMID: 22637703*

References

Affò S, Dominguez M, Lozano JJ, Sancho-Bru P et al. Transcriptome analysis identifies TNF superfamily receptors as potential therapeutic targets in alcoholic hepatitis. *Gut* 2013 Mar;62(3):452-60. PMID: 22637703

Examples

```
data(pheno.mRNA)
```

```
pheno.mRNA
```

```
plot3d
```

```
PCA plot in 3D
```

Description

```
PCA plot in 3D
```

Usage

```
plot3d(obj, subset, col.color = 1, angle = 45, colors = c("violet", "turquoise"), lty = 0,  
cex.points = 1, ...)
```

Arguments

obj a corObject
subset "miRNA" or "mRNA"
col.color number or name of the column in the pheno slot that will be used to color the samples
angle angle orientation
colors character vector with the colors that will be used for each group in col.color.
lty lty of lines that are plotted parallel to the z-axis for each sample. 0 means no line.
cex.points cex scaling of the dots.
 ... further arguments to be passed

Value

A 3d pca plot.

Note

This error:

```
Error in factor(as.numeric(as.factor(obj@pheno.miRNA[, col.color])) + : invalid 'labels'; length 2
should be 1 or 12
Is due to invalid length of colors option. Use character vectors with the same length of the number
of groups in col.color.
```

See Also

[plotPca](#)

Examples

```
data(data.obj)
plot3d(data.obj, "mRNA")
```

plotCircos

Circos plot

Description

Plot a circos plot showing the miRNA-mRNA interactions

Usage

```
plotCircos(obj, pval.cutoff = 0.05, dat.sum = obj@info[["dat.sum"]],
n = NULL, sub.miRNA = NULL, sub.mRNA = NULL)
```

Arguments

obj a corObject with a net slot already defined
pval.cutoff maximum corrected p value of the selected miRNA-mRNA pairs
dat.sum numeric; minimum concordance across databases of the selected miRNA-mRNA pairs
n (optional) numeric. If specified, limit to the first "n" pairs (sorted by corrected p value)
sub.miRNA (optional) character vector, include only pairs containing these miRNAs.
sub.mRNA (optional) character vector, include only pairs containing these mRNAs.

Value

a plot

References

www.circos.ca : Krzywinski MI, Schein JE, Birol J, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. Circos: An information aesthetic for comparative genomics. Genome Research, 2009.
<http://cran.r-project.org/web/packages/circlize/index.html>

See Also

[data\(genes_human_h37\), mirnas_human_17_h37](#)

Examples

```
data(data.obj)
plotCircos(data.obj, n="50")
```

plotCordist

Plot distances/correlation between miRNA or mRNA samples

Description

Plot distances/correlation between miRNA or mRNA samples

Usage

```
plotCordist(obj, subset, type = "cor", method.cor = "pearson",
method.dist = "euclidean", hierarchical = FALSE, ...)
```

Arguments

obj a corObject.
 subset "miRNA" or "mRNA"
 type "cor" (correlation) or "dist" (distance).
 method.cor method used for computing correlation: "pearson" or "spearman".
 method.dist method used for computing distance from dist function: This must be one of "euclidean", "maximum", "manhattan", "canberra", "binary" or "minkowski".
 Any unambiguous substring can be given.
 hierarchical TRUE or FALSE. Sort the rows using hierarchical clustering.
 ... further arguments.

Value

A plot of the matrix of distances (or correlations) between samples. A hierarchical clustering of the samples can also be performed to sort the rows and columns.

References

http://www.phageta4.org/R/image_matrix.html

See Also

`plot3d.plotPca.dist`

Examples

```
data(data.obj)
plotCorDist(data.obj, "miRNA", type="cor")
plotCorDist(data.obj, "mRNA", type="dist")
```

plotCorrelation

Plot correlations

Description

Plot the correlation of a miRNA and mRNA with their linear regression

Usage

```
plotCorrelation(obj, miRNA, mRNA, type = "cor", samples = "all",
  col.color = 1, i.legend = list(col.color), pos.legend = "topright",
  sample.names = FALSE, pos.sample.names = 1, cex.main = 1.35,
  alternative = "two.sided", colors = c("turquoise", "violet"))
```

Arguments

obj corObject
 miRNA character, miRNA selected
 mRNA character, mRNA selected
 type "cor" or "residuals"
 samples column name of the pheno.miRNA where to look for the grouping factor. The column must contain character names.
 col.color name of the column in the pheno slot used for coloring the samples.
 i.legend TRUE or FALSE. Plot legend
 pos.legend set the legend position from Legend: "bottomright", "bottom", "bottomleft", "left", "topleft", "top", "topright", "right" and "center".
 sample.names TRUE or FALSE. Plot text of sample names.
 pos.sample.names Position of the text for sample names from text. Values of '1', '2', '3' and '4', respectively indicate positions below, to the left of, above and to the right of the specified coordinates.
 cex.main cex parameter for the main title.
 alternative alternative hypotheses used for computing the *p* value of the correlation. One of "two.sided", "less", "greater".
 colors In case of a two-level factor grouping, colors to be used.

Value

A scatter plot showing the correlation of the miRNA and mRNA expression of a miRNA-mRNA pair. A line representing the linear regression between both variables is also plotted.

Examples

```
data(data.obj)
plotCorrelation(data.obj, miRNA="hsa-miR-107", mRNA="ACPP",
  type="cor", col.color="group", sample.names=TRUE)
plotCorrelation(data.obj, miRNA="hsa-miR-107", mRNA="ACPP",
  type="residuals")
```

plotDensity

Plot miRNA or mRNA density

Description

Plot miRNA or mRNA density per sample

Usage

```
plotDensity(obj, subset, col.color=1, colors=c("turquoise", "violet"))
```

Arguments

obj a corObject.
 subset "miRNA" or "mRNA".
 col.color name of the column in the pheno slot used for coloring the samples.
 colors In case of a two-level factor grouping, colors to be used.

Value

A density plot showing the densities per sample of the miRNA or mRNA dataset.

Examples

```
data(data.obj)
plotDensity(data.obj, "miRNA")
plotDensity(data.obj, "mRNA")
```

plotGO *Plot GO enriched terms*

Description

From RamiGO package, plots the hierarchy of the significant GO terms

Usage

```
plotGO(obj, type, ontology, fdr = 0.05, filename = "GO_tree_default")
```

Arguments

obj a corObject.
 type "GO" or "KEGG".
 ontology For GO terms, one of "BP", "CC" or "MF". For KEGG terms, use "KEGG".
 fdr FDR cutoff.
 filename name of the TIFF figure.

Value

A file with the GO terms plotted, organised hierarchically and highlighting the significant ones.

Examples

```
### do not run
# data(data.obj)
# plotGO(data.obj, "GO", "BP", fdr = 0.05, filename = "GO_example")
### this will create the file "GO_example.png"
```

plotHeatmap *Plot hierarchical clustering of miRNA or mRNA samples*

Description

Plot the hierarchical clustering tree of the miRNA or mRNA samples according to the euclidean distance between them.

Usage

```
plotHeatmap(obj, subset)
```

Arguments

obj a corObject.
 subset "miRNA" or "mRNA"

Value

A plot of the hierarchical clustering tree of the miRNA or mRNA samples according to the euclidean distance between them.

See Also

[plot3d](#), [plotPca](#), [dist](#), [plotCordist](#)

Examples

```
data(data.obj)
plotHeatmap(data.obj, "miRNA")
```

plotHeatmap *Plot heatmaps*

Description

Plot heatmaps of the top "n" miRNAs or mRNAs from selected miRNA-mRNA pairs

Usage

```
plotHeatmap(obj, class, n = 50, col.color = NULL, min.exp = NULL, main = NULL, pval.cutoff = NULL, groupi
```


Arguments

obj a corObject
 class "miRNA", "miRNA" or "both". When "both", the heatmap is a mixture of miRNAs and mRNAs.
 n number of items to plot.
 col.color character vector, column name or number which contains that will be used to label the samples.
 min.exp (optional) minimum mean expression of the items to be plotted
 main (optional) title of the plot
 pval.cutoff (optional) maximum adjusted pvalue
 grouping.col (optional) factor or character vector indicating predefined groups
 grouping.row (optional) factor or character vector indicating predefined groups
 order criteria used to select the "top" miRNAs or mRNAs. "pval" or "adj.pval" (default), or "logratio" or "FC". Only for class="miRNA" or "mRNA"
 cex.lab cex of the label axis. Default is 0.75

Value

A plot with the heatmap of the selected miRNAs or mRNAs.

See Also

[plotCorrelation](#)

Examples

```
data(data.obj)
plotHeatmap(data.obj, "miRNA", n=100)
```

plotMA

MA plot

Description

Plots a MA plot of the 2 selected samples.

Usage

```
plotMA(obj, subset, sample1 = NULL, sample2 = NULL, cex.lab = 1, cex.axis = 1)
```

Arguments

obj a "corObject"
 subset "miRNA" or "mRNA"
 sample1 number of the column or column name of sample 1 to use
 sample2 number of the column or column name of sample 2 to use
 cex.lab cex of the label axis. Default is 1
 cex.axis cex of the axis. Default is 1

Value

A MA plot of the selected two samples

See Also

[plot3d](#), [plotPca](#), [boxplotSamples](#)

Examples

```
data(data.obj)
plotMA(data.obj, "miRNA", sample1=1, sample2=2)
plotMA(data.obj, "mRNA", sample1="H_1", sample2="D_7")
```

plotNetwork

Plot a network

Description

Plot the network of the selected miRNA-mRNA interactions, with selected features

Usage

```
plotNetwork(obj, pval.cutoff = 0.05, score.cutoff = NULL, sub.miRNA = NULL,
sub.mRNA = NULL, names = TRUE, dat.sum = obj@irfo[["dat.sum"]],
add.other = NULL, vertex.cex = NULL, n = NULL, node.size = 1.5)
```

Arguments

obj a corObject.
 pval.cutoff *p* value cutoff of the selected miRNA-mRNA interactions.
 score.cutoff maximum score of the selected miRNA-mRNA interactions.
 sub.miRNA (optional) character vector with the names of the miRNAs to be included on the plot.
 sub.mRNA (optional) character vector with the names of the mRNAs to be included on the plot.

names Plot the names of the miRNAs and mRNAs. TRUE or FALSE.
 dat.sum minimum occurrences between databases of the selected miRNA-mRNA interactions.
 add.other Optional, character vector: name of the dataframe containing additional interactions (usually miRNA-mRNA interactions) that will also be displayed.
 vertex.cex Optional, character vector: name of the dataframe containing the relative size for each node in the network. For example "interact.table".
 n maximum number of interactions.
 node.size Size of the node

Details

The colors are representative of the interactions. "interact.table" can be loaded from miRData package using `data(interact.table)`.

Value

A network of the selected miRNA-mRNA interactions. Circles represent miRNAs and squares miRNAs; red fill means upregulated miRNA/mRNA, while green fill means downregulated miRNA/mRNA; lines indicate the miRNA-mRNA pairs; red line means positive score and green line means negative score; arrow width is proportional to the number of appearances on the databases.

See Also

[boxplotCorrelation](#), [plotCorrelation](#)

Examples

```
data(data.obj)
plotNetwork(data.obj, pval.cutoff=0.01, dat.sum=1,
  vertex.cex="Interact.table", names=FALSE)
```

plotPca

PCA with miRNA or mRNA data

Description

Plot Principal Components Analysis of miRNA or mRNA data

Usage

```
plotPca(obj, subset, col.color = 1, colors = c("turquoise", "violet"), pos.legend="topleft",
  names = FALSE, ...)
```

Arguments

obj a corObject.
 subset "miRNA" or "mRNA".
 col.color name of the column in the pheno slot used for coloring the samples.
 colors In case of a two-level factor grouping, colors to be used.
 pos.legend legend position, "bottomright", "topright", "bottomleft" or "topleft"
 names FALSE or TRUE. Plot names of the samples.
 ... further arguments.

Value

A pca plot showing the principal components analysis (PCA) of miRNA or mRNA data.

See Also

[plot3d](#)

Examples

```
data(data.obj)
plotPca(data.obj, "miRNA")
plotPca(data.obj, "mRNA")
```

plotSurv

Plot Kaplan-Meier curve

Description

Plot Kaplan-Meier curve of miRNA or mRNA

Usage

```
plotSurv(obj, subset, item, time, event)
```

Arguments

obj a corObject.
 subset "miRNA" or "mRNA"
 item name of the miRNA or mRNA
 time name or number of the column in the pheno slot that contains the time.
 event name or number of the column in the pheno slot that contains the censoring (0=censored, 1=event).

`plotVolcano`

43

Value

A plot with the Kaplan-Meier.

References

survival package: <https://cran.r-project.org/web/packages/survival/index.html>

See Also

[addSurv](#)

Examples

```
#data(data.obj)
#plotSurv(data.obj, "miRNA", item = ...)
```

44

`removeSamp`

See Also

[addSurv](#)

Examples

```
data(data.obj)
plotVolcano(data.obj, "miRNA")
plotVolcano(data.obj, "miRNA")
```

`removeSamp`

Remove samples or miRNA/miRNA

Description

Remove samples or miRNA/miRNA from a corObject.

Usage

```
removeSamp(obj, dataset, samples = Na, genes = Na, keep=FALSE)
```

Arguments

<code>obj</code>	a corObject
<code>dataset</code>	"miRNA" or "miRNA"
<code>samples</code>	colnames of the samples to be removed.
<code>genes</code>	rownames of the genes (miRNA or miRNA) to be removed.
<code>keep</code>	TRUE (keep given colnames/rownames) or FALSE. By default, FALSE.

Details

Genes are removed from miRNAdata/miRNAdata slots. Samples are removed from both pheno.miRNA/pheno.miRNA and miRNAdata/miRNAdata slots.

Take into account that minimum number of required samples is 2, otherwise the function will give an error.

Value

a corObject with the selected samples already removed.

`plotVolcano`

43

Value

A plot with the Kaplan-Meier.

References

survival package: <https://cran.r-project.org/web/packages/survival/index.html>

See Also

[addSurv](#)

Examples

```
#data(data.obj)
#plotSurv(data.obj, "miRNA", item = ...)
```

`plotVolcano`

Volcano plot

Description

Volcano plot of miRNA or mRNA differential expression analysis

Usage

```
plotVolcano(obj, subset, FC1 = 1.5, FC2 = 2, FDR = 0.05, cex = 1, cex.lab = 1, cex.axis = 1 )
```

Arguments

<code>obj</code>	a "corObject" with a "diffexp.miRNA" or "diffexp.mRNA" slot already defined
<code>subset</code>	"miRNA" or "miRNA"
<code>FC1</code>	first FoldChange cutoff
<code>FC2</code>	second FoldChange cutoff
<code>FDR</code>	significance cutoff (FDR)
<code>cex</code>	cex value for the dots. Default is 1
<code>cex.lab</code>	cex of the label axis. Default is 1
<code>cex.axis</code>	cex of the axis. Default is 1

Value

A volcano plot of the miRNA or mRNA data from differential expression analysis. Items with FDR lower than the limit are highlighted in yellow. Items with FDR lower than the limit and higher than FC1 are highlighted in orange. Items with FDR lower than the limit and higher than FC2 are highlighted in red.

Examples

```
data(data.obj)

dim(data.obj$dat.miRNA)
data.obj<-removeSamp(data.obj,"miRNA",samples="D_3",genes="hsa-miR-200c")
dim(data.obj$dat.miRNA)

colnames(data.obj$dat.miRNA)
data.obj<-removeSamp(data.obj,"miRNA",samples=c("D_1","D_2"),keep=TRUE)
colnames(data.obj$dat.miRNA)
```

selOutliers *select outliers based on PCA analysis*

Description

select outliers from miRNA or mRNA samples based on PCA analysis

Usage

```
selOutliers(obj, subset, method = "aq.plot", delete = FALSE, add.pheno = TRUE, n.dim = 2)
```

Arguments

obj a corObject
 subset "miRNA" or "mRNA"
 method method used to select the outliers "aq.plot"
 delete TRUE or FALSE. If TRUE, outlier samples are removed.
 add.pheno TRUE or FALSE. If TRUE, "is.outlier" column is added to pheno slot and then a PCA plot highlighting outlier samples is produced.
 n.dim number of components of the PCA to use.

Details

This is an implementation of mvoutlier package. Check the original source for more information.

Value

- If delete=FALSE and add.pheno=FALSE, character vector with the names of the outlier samples.
- If delete=FALSE and add.pheno=TRUE, a corObject with a column called is.outlier added to the pheno slot indicating if a sample is outlier or not.
- If delete=TRUE, a corObject without the outlier samples in the data and pheno slot.

See Also

[mvoutlier.plotPca](#)

Examples

```
data(data.obj)
selOutliers(data.obj,"miRNA",add.pheno=FALSE)
#data.obj.out <- selOutliers(data.obj,"miRNA",add.pheno=TRUE)
#head(data.obj$pheno.miRNA)
#head(data.obj$out@pheno.miRNA)
```

selSubsetCor *Select relevant miRNA-mRNA interactions*

Description

Select differentially expressed miRNAs or mRNAs, according to your criteria

Usage

```
selSubsetCor(obj, pval.cutoff = 1, dat.sum = 0, sub.miRNA = NULL,
             sub.mRNA = NULL)
```

Arguments

obj a corObject
 pval.cutoff maximum adj.pval of the selected miRNA-mRNA pairs
 dat.sum minimum number of concurrences across databases of the selected miRNA-mRNA pairs
 sub.miRNA (optional) character vector, limit to these miRNAs
 sub.mRNA (optional) character vector, limit to these mRNAs

Value

A data.frame with the selected miRNA-mRNA pairs.

See Also

[writeCsv.writeExcel.writeSif](#)

Examples

```
data(data.obj)
selSubsetCor(data.obj, pval.cutoff=0.05, dat.sum=2)
```

selSubsetExprs *Select differentially expressed miRNAs or mRNAs*

Description

Select differentially expressed miRNAs or mRNAs, according to your criteria

Usage

```
selSubsetExprs(obj, dataset, FC = NA, logratio = foldchange2logratio(FC), slope = NA, pval = NA, adj = 1, up = FALSE, dw = FALSE)
```

Arguments

obj a corObject
 dataset "miRNA" or "miRNA".
 FC minimum absolute FoldChange of selected miRNAs or mRNAs
 logratio minimum absolute logratio cutoff of selected miRNAs or mRNAs
 slope for longitudinal-regression analysis, minimum absolute slope of selected miRNAs or mRNAs
 pval maximum p value of selected miRNAs or mRNAs
 adj, pval maximum adjusted p value of selected miRNAs or mRNAs
 min.meanExp minimum mean expression of selected miRNAs or mRNAs
 up TRUE or FALSE. Select only upregulated miRNAs or mRNAs.
 dw TRUE or FALSE. Select only upregulated miRNAs or mRNAs.

Value

A data.frame with the selected differentially expressed miRNAs or mRNAs and their characteristics.

See Also

[writeCsv](#), [writeExcel](#), [writeSif](#)

Examples

```
data(data.obj)
selSubsetExprs(data.obj, "miRNA", adj.pval=0.05, FC=1.5)
selSubsetExprs(data.obj, "miRNA", adj.pval=0.05, up=TRUE)
```

summary.corObject *Brief report of a corObject*

Description

Tests if a corObject is valid and prints the information about the tests performed.

Usage

```
## S3 method for class 'corObject'
summary(Object, ...)
```

Arguments

object a corObject.
 ... other.

Value

Simple version of printing of the report of a corObject.

See Also

[mkReport](#)

Examples

```
data(data.obj)
summary(data.obj)
```

topTable

Print or plot the top connected miRNA/mRNA

Description

Print or plot the top connected miRNA/mRNA

Usage

```
topTable(obj, class, pval.cutoff = 0.05, dat.sum =
  obj@rfo[["dat.sum"]], score.cutoff = NULL, plot = FALSE, names = FALSE, n = NULL, remove.names = F
```

Arguments

obj corObject
 class "miRNA" or "miRNA".
 pval.cutoff *p* value to cutoff.
 dat.sum number of minimum occurrences across databases.
 score.cutoff maximum score allowed
 plot FALSE or TRUE.
 names FALSE or TRUE. Apart from the frequency, add the names of the hits. (In this case the object returned is a data.frame).
 n maximum number of pairs to consider.
 remove.names in case of a plot, omit x-axis labels.

Details

If `pLot=FALSE` then the table is displayed, if `pLot=TRUE` the table is not displayed and a barplot is plotted. If `names=FALSE` a vector is returned, if `names=TRUE`, a data.frame is returned.

Value

A data.frame

See Also

[plotheatmap](#)

Examples

```
data(data.obj)

# get the names
topTable(data.obj, "miRNA", names=TRUE, plot=FALSE)
topTable(data.obj, "miRNA", names=TRUE, plot=FALSE)

# plot
topTable(data.obj, "miRNA", names=TRUE, plot=TRUE)
```

translatemiRNAs *Convert miRNA names between versions*

Description

Convert miRNA names between versions

Usage

```
translatemiRNAs(x, from = NULL, to = "21")
```

Arguments

x character vector with the names of the miRNAs to translate.
 from character vector, version of the miRBase of the miRNAs to translate. Must be one of: "6.0", "7.0", "7.1", "8.0", "8.1", "8.2", "8.3", "9.0", "9.1", "9.2", "10.0", "10.1", "11.0", "12.0", "13.0", "14", "15", "16", "17", "18", "19", "20", "21" or "unknown".
 to character vector, version of the miRBase of the miRNAs to translate. Must be one of: "6.0", "7.0", "7.1", "8.0", "8.1", "8.2", "9.0", "9.1", "9.2", "10.0", "10.1", "11.0", "12.0", "13.0", "14", "15", "16", "17", "18", "19", "20", "21"

Value

A data frame containing the original name of the translated name for each miRNA.

See Also

[checkmiRNAs](#)

Examples

```
translatemiRNAs(c("hsa-let-7b", "not-a-miRNA", "hsa-miR-200c"), from="6.0", to="21")
```

writeCsv *Write a csv file*

Description

Export the cytofile slot to a csv file.

Usage

```
writeCsv(obj, name, pval.cutoff = 1, dat.sum =
  obj@infol["dat.sum"]), slot = "net", pval = "adj.pval")
```

Arguments

obj a corObject.
 name the name of the file to write.
 pval.cutoff maximum corrected *p* value to take.
 dat.sum number of minimum occurrences across databases.

slot name of the slot to write. "net" (default), "diffexp.mirna", "diffexp.mrna", "dat.mirna", "dat.mrna", "pheno.mirna", or "pheno.mrna". P-value and dat.sum filtering is applied to "net" slot. P-value filtering is applied to "diffexp.mirna" and "diffexp.mrna" slots. No filtering is applied to "dat.mirna", "dat.mrna", "pheno.mirna" or "pheno.mrna" slots.

pval name of the p.value column to select.

Value

A csv file.

See Also

[writeExcel](#), [writeSif.seSubsetCor](#), [seSubsetExprs](#)

Examples

```
## do not run
#data(data.obj)
#writeCsv(data.obj, "results.csv.csv")
```

writeExcel

Write an Excel file

Description

Export the cytofile slot to an Excel 2003 file.

Usage

```
writeExcel(obj, name, pval.cutoff = 0.05, dat.sum = obj@info[["dat.sum"]],
slot = "net", pval = "adj.pval")
```

Arguments

obj a corObject.

name the name of the file to write.

pval.cutoff maximum corrected *p* value to take.

dat.sum number of minimum occurrences across databases.

slot name of the slot to write. "net" (default), "diffexp.mirna", "diffexp.mrna", "dat.mirna", "dat.mrna", "pheno.mirna", or "pheno.mrna". P-value and "diffexp.mirna" and "diffexp.mrna" slots. No filtering is applied to "diffexp.mirna" and "diffexp.mrna" slots. No filtering is applied to "dat.mirna", "dat.mrna", "pheno.mirna" or "pheno.mrna" slots.

pval name of the column where to take the *p* values.

Details

It writes an excel document with the selected pairs.

Value

An excel file.

See Also

[writeCsv](#), [writeSif.seSubsetCor](#), [seSubsetExprs](#)

Examples

```
## do not run
#data(data.obj)
#writeCsv(data.obj, "results.csv.csv")
```

writeSif

Write a SIF file

Description

Export the network (from cytofile slot) to a SIF file.

Usage

```
writeSif(obj, file, pval.cutoff = 0.05, dat.sum =
obj@info[["dat.sum"]], add.other = NULL, sub.mirna =
NULL, sub.mrna = NULL, expand = FALSE, vertex.cex =
"interact.table")
```

Arguments

obj a corObject.

file file to write.

pval.cutoff maximum corrected *p* value to take.

dat.sum number of minimum occurrences across databases.

add.other a character vector. Name of the data frame that contains extra interactions (usually mRNA-mRNA interactions) that will be added to the network.

sub.mirna character vector. Restrict to these miRNAs.

sub.mrna character vector. Restrict to these mRNAs.

expand expand with another table. TRUE or FALSE.

vertex.cex table to use to expand

`writeSif`

Value

A sif file, plus a node attributes file ("the node attributes still in preparation")

References

www.cytoscape.org

See Also

`writeCsv`, `writeSif`, `selSubsetCor`, `selSubsetExprs`, `openCytoscape`

Examples

```
## do not run
#data(data.obj)
#writeSif(data.obj, "results_sif")
```

Index

- *Topic **Cook distance**
- addCorrelation, R, 5
- *Topic **GO**
- GOanalysis, 26
- plotGO, 37
- *Topic **KEGG**
- GOanalysis, 26
- *Topic **LaTeX**
- mkrReport, 28
- *Topic **MA plot**
- plotMA, 39
- *Topic **Reactome**
- GOanalysis, 26
- *Topic **Type II error**
- combinePval, 21
- correctPval, 24
- *Topic **add**
- addDatabase, 7
- addFolDchanges, 10
- addNet, 14
- *Topic **barplot**
- topTable, 48
- *Topic **boxplot**
- boxplotCorrelation, 18
- boxplotSamples, 19
- *Topic **circos**
- plotCircos, 33
- *Topic **classes**
- corObject-class, 22
- *Topic **cluster**
- plotClust, 38
- *Topic **correlation**
- addCorrelation, 4
- addCorrelation, R, 5
- addGlimnet, 11
- boxplotCorrelation, 18
- pearson, 31
- plotCorrelation, 35
- *Topic **csv**
- writeCsv, 50
- *Topic **cytoscape**
- openCytoscape, 30
- *Topic **database**
- addDatabase, 7
- *Topic **datasets**
- data.obj, 24
- miRNA, 28
- miRNA, 29
- pheno, miRNA, 31
- pheno, miRNA, 32
- *Topic **density**
- plotDensity, 36
- *Topic **differential expression**
- addFolDchanges, 10
- addScore, 15
- *Topic **differential**
- addDiffExp, 8
- addLong, 12
- *Topic **distance**
- plotCorDist, 34
- plotClust, 38
- *Topic **elastinet**
- addGlimnet, 11
- *Topic **enrichment**
- GOanalysis, 26
- *Topic **excel**
- writeExcel, 51
- *Topic **expression**
- selSubsetCor, 46
- selSubsetExprs, 47
- *Topic **external file**
- plotGO, 37
- writeCsv, 50
- writeExcel, 51
- writeSif, 52
- *Topic **glimnet**
- addGlimnet, 11
- *Topic **heatmap**

plotHeatmap, 38
 *Topic **meta-analysis**
 combInefVal, 21
 *Topic **miRBase**
 checkmiRNAs, 20
 transLatemiRNAs, 49
 *Topic **miRNA**
 addDatabase, 7
 checkmiRNAs, 20
 transLatemiRNAs, 49
 *Topic **network**
 plotNetwork, 40
 *Topic **net**
 addDatabase, 7
 addFolDchanges, 10
 addNet, 14
 *Topic **outlier**
 selOutliers, 45
 *Topic **output**
 writeCsv, 50
 writeExcel, 51
 *Topic **p.value**
 combinePval, 21
 *Topic **package**
 Information-package, 3
 *Topic **pca**
 plot3d, 32
 plotPca, 41
 selOutliers, 45
 *Topic **pdf**
 mkReport, 28
 *Topic **pearson**
 addCorrelation, R, 5
 *Topic **plot**
 boxplotCorrelation, 18
 boxplotSamples, 19
 checkmiRNAs, 20
 plot3d, 32
 plotCircos, 33
 plotCordist, 34
 plotCorrelation, 35
 plotDensity, 36
 plotFolD, 37
 plotHeatmap, 38
 plotMA, 39
 plotNetwork, 40
 plotPca, 41

addSurv, 17, 43, 44
 aov, 9
 boxplotCorrelation, 18, 19, 41
 boxplotSamples, 19
 checkmiRNAs, 20, 50
 circIize, 4
 combinePval, 21, 22, 24
 cooks.distance, 6
 cor, 3, 6
 corObject-class, 22
 correctPval, 14, 24
 coxph, 17
 data(genes_human_h37), 34
 data.obj, 24
 DESeq, 4
 dist, 35, 38
 evaluate, 25
 fgsea, 4
 glmnet, 4
 GO.db, 4
 GOanalysis, 26
 GOSTats, 4
 gtools, 4
 Hmisc, 4
 information (information-package), 3
 information-package, 3
 KEGG.db, 4
 Limma, 4
 mclust, 4
 methods, 4
 miRData, 4
 miRNA, 28
 mirnas_human_17_h37, 34
 mkReport, 28, 48
 miRNA, 29
 mvoutlier, 4, 46
 network, 4
 openCytoscape, 30, 53
 p.adjust, 9, 12, 22, 24
 package:glmnet, 11
 package:GOSTats, 27
 package:Limma, 9
 package:RankProd, 9
 package:ReactomePA, 27
 package:survival, 17
 pearson, 31
 pheatmap, 4
 pheno_miRNA, 31
 corObject-class, 22
 correctPval, 14, 24
 coxph, 17
 plotCircos, 33
 plotCordist, 19, 34, 38
 plotCorrelation, 18, 19, 35, 39, 41
 plotDensity, 36
 plotFolD, 37
 plotHeatmap, 38, 49
 plotMA, 39
 plotNetwork, 15, 40
 plotPca, 33, 35, 38, 40, 41, 46
 plotSurv, 17, 42
 plotVolcano, 43
 pROC, 4
 RamiGO, 4
 RankProd, 4
 ReactomePA, 4
 removeSamp, 44
 scatterplot3d, 4
 selOutliers, 45
 selSubsetCor, 46, 57–53
 selSubsetExprs, 47, 51–53
 summary (summary_corObject), 48
 summary_corObject, 48
 survival, 4
 t.test, 9
 text, 36
 topTable, 48
 transLatemiRNAs, 8, 20, 49
 VennDiagram, 4
 verification, 4
 vroom, 9

INDEX

57

wilcoxon, 9
writeCsv, 46, 47, 50, 52, 53
writeExcel, 46, 47, 51, 51
writeSIF, 30, 46, 47, 51, 52, 52, 53
WriteXLS, 4
xtable, 4

