Master Degree Project

Functional Analysis of Circulating microRNAs in Pancreatic Cancer

Master Degree Project in bioinofrmatics
One year Level 30 ECTS
Spring term 2018

Emmy Borgmästars

Supervisor: Zelmina Lubovac
External supervisor: Malin Sund, Umeå University
Examiner: Angelica Lindlöf

Abstract

MicroRNAs (miRNAs) are small RNA species that regulate gene expression at a post-transcriptional level and are emerging as potentially important biomarkers for various disease states. To understand the underlying role of miRNAs, functional analysis can be performed to enable discovery of novel molecular mechanisms that contribute to the pathogenesis of the disease and possibly can function as novel therapeutic targets. Functional analysis of miRNA consists of miRNA target prediction and functional enrichment analysis of target genes. MiRNA target predictions often generate many potential target genes and to validate all of these in an experimental setup is not possible. Hence, a validation approach is valuable to narrow down interesting candidate target genes. One method commonly used is to correlate miRNA expression to mRNA expression to assess the regulatory effect of a particular miRNA.

Finding novel non-invasive biomarkers for pancreatic cancer is highly desirable as the only clinically used biomarker today namely cancer antigen-19-9 (CA19-9) is not sensitive or specific enough. MiRNAs are emerging as potential biomarkers in pancreatic cancer and a dataset of fifteen circulating miRNAs identified as differentially expressed in pancreatic cancer was used in this study.

This study aimed to develop an automated bioinformatics pipeline in R software that performs miRNA target prediction, functional enrichment analysis and *in silico* evaluation of predicted miRNA-mRNA pairs by correlation of miRNA expression levels to its targets on mRNA and protein expression levels. The miRNA isoform expression data from the cancer genome atlas was utilized here to obtain the specific expression pattern for each mature miRNA. Since miRNAs mainly affect the protein levels, protein expression data from the cancer proteome atlas were also included in the correlation analyses. Fifteen significantly altered circulating miRNAs detected in pancreatic cancer patients were queried in the pipeline.

Contents

Abstract 2

1. Introduction 5

1.1. Biological role of miRNAs 5

1.2. Bioinformatics pipeline for miRNA functional analysis 6

1.2.1. miRNA target prediction 7

1.2.2. Functional enrichment analysis 7

1.2.3. Validation of miRNA functional analysis 8

1.2.4. R packages for miRNA functional analysis 8

1.3. Biomarkers in pancreatic cancer 9

1.4. Problem formulation 10

1.5. Aim 11

1.6. Limitations 11

2. Materials & Methods 11

2.1. Bioinformatics pipeline 11

2.2. Data 12

2.2.1. miRNAs 12

2.2.2. Expression data 13

2.3. miRNA target prediction 13

2.3.1. Alternative miRNA target prediction tools 14

2.4. Functional enrichment 15

2.4.1. Alternative functional enrichment tools 15

2.5. *In silico* evaluation 15

2.5.1. Pre-processing of expression data 15

2.5.2. Annotation of mature miRNAs 16

2.5.3. Correlation 17

2.5.4. Correlation settings 17

2.5.5. Post-processing 18

2.6. Identifying hub genes 18

3. Results 19

3.1. Number of target genes partially overlap between databases 19

3.2. Strong correlations were observed for miR-885-5p 20

3.2.1. miRNA-protein correlation 22

3.2.2. miRNA-mRNA-protein integration 23

3.3. Functional enrichment analysis 23

3.4. E2 and E3 ubiquitin ligases were identified as hub genes. 26

4. Discussion 27

4.1. miRNA target prediction 27

4.2. *In silico* evaluation of miRNA target prediction 28

4.2.1. Strong correlations were identified for miR-885-5p 29

4.2.2. Functional enrichment of correlated targets 30

4.3. Protein-protein interaction networks 30
4.3.1. E2 ubiquitin-conjugating enzymes 30
4.3.2. E3 ubiquitin ligases 30
4.4. The importance of annotating mature miRNAs 31
5. Conclusion 32
6. Future studies 32
6.1. Bioinformatics development 33
6.2. Experimental studies 33
7. Impact on society 33
8. Ethical aspects 34
9. Acknowledgements 34
10. References 34
Supplementary information 41
1. Figures & Tables 41
2. Programming code 44
2.1. SQL databases 44
2.2. Function for miRNA target prediction, functional enrichment & correlation analyses 50
2.3. Run function as described in '2.2. miRNA target prediction, functional enrichment & correlation analyses' for 15 miRNAs 57
2.4. Check how many samples express a specific miRNA 60
2.5. -3p and -5p annotation of mature miRNAs 61
2.6. Extract top enriched GO terms and KEGG pathways 65
2.7. Plot mature and miRNA gene expression 69

1. Introduction

1.1. Biological role of miRNAs

MicroRNAs (miRNAs) are small RNA oligonucleotides with a length of about 19-24 nucleotides (Bhaskaran and Mohan, 2014). The biogenesis of miRNAs starts with transcription of the miRNA gene by polymerase II (Fig. 1). After transcription, a stem loop called pri-miRNA is formed (Fig. 1A), which is cut in the nucleus by a complex called microprocessor containing Drosha enzyme. After transportation to the cytoplasm by Exportin-5 (Fig. 1C), two isoforms are formed, one -3p and one -5p mature miRNA, by the Dicer enzyme (Fig. 1D). Usually one of the mature miRNAs, called the passenger strand, is degraded and the other strand, often referred to as guide strand, is playing a role in miRNA-mediated regulation (Bhaskaran and Mohan, 2014). However, both strands could also act in miRNA-mediated regulation. The mature miRNAs are reverse complementary to each other and thus can have different messenger RNA (mRNA) targets.

The nomenclature of a miRNA consists of a three-letter code for species (http://www.mirbase.org/help/nomenclature.shtml), for instance in human it is 'hsa' (Griffiths-Jones, 2006; Griffiths-Jones et al., 2006; Griffiths-Jones et al., 2008; Kozomara and Griffiths-Jones, 2011, 2014). The 'miRNA gene' is referred to the gene or the stem-loop region that the miRNA originates from, designated as 'mir-885', whereas the mature miRNA is denoted as 'miR-885'. If there is a clear predominant miRNA species, it is simply represented as 'miR-885' for the guide strand and 'miR-885*' for the passenger strand, however the -3p and -5p extensions are used when both miRNA species can act as a regulatory miRNA.

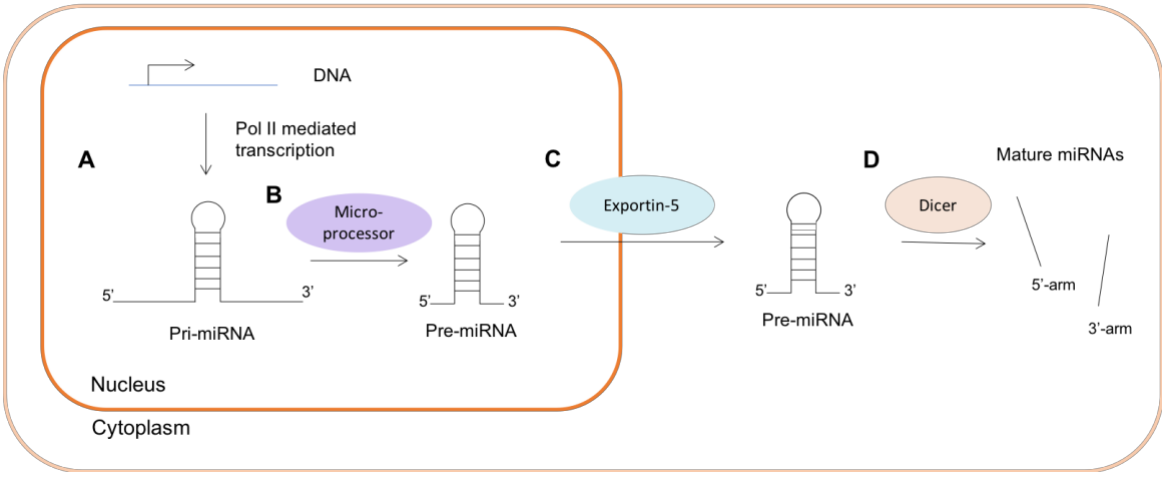


Figure 1. Biogenesis of mature miRNAs. Pri-miRNA is transcribed from the miRNA gene by polymerase II (pol II) (A) and a complex termed microprocessor processes pri-miRNA into pre-miRNA (B) (Bhaskaran and Mohan, 2014). Pre-miRNA is translocated from nucleus to cytoplasm by Exportin-5 (C) and Dicer enzyme cleaves pre-miRNA into two mature miRNA strands termed 3'- and 5'-arms (D).

MiRNAs are generally considered down-regulators of mRNAs at a post-transcriptional level, however, studies have shown that they can act as up-regulators as well (Rusk, 2008; Vasudevan et al., 2007). The translational regulation is not mediated by miRNAs alone but associates to effector complexes called micro-ribonucleoprotein (miRNP) (Orang et al.,

2014). In miRNA-mediated down-regulation, translational repression is usually the primary event followed by mRNA degradation (Wilczynska and Bushell, 2015). Repression also seems to be a reversible state, where repressed mRNA could become activated and translated instead of degraded. In other words, a decay on protein level but not on mRNA level could occur due to miRNA-mediated regulation. MiRNA-mediated up-regulation could occur indirectly by interfering with repressive miRNPs or directly by the activity of miRNPs (Orang et al., 2014). However, the positive regulation mediated by miRNAs seem to be restricted to certain cell conditions, for instance cells in G0 cell cycle state.

A 6-8 nucleotides long region located at the 5'-end of miRNAs is referred to as the 'seed region', which is used for binding to its mRNA targets (Ellwanger et al., 2011). MiRNAs regulate genes mainly by binding to the 3'-untranslated region (3'-UTR) region of mRNAs (Fig. 2), but interactions with other regions of mRNA have been identified as well, such as the 5'-UTR (Lytle et al., 2007) and the coding sequence (CDS) (Tay et al., 2008). In plants, miRNAs bind to their target in a perfect or near perfect fashion, whereas for animals the binding might not always be completely complementary between miRNA and mRNA (Millar and Waterhouse, 2005). This makes it challenging to design reliable target prediction tools for studies in animals. In addition, one miRNA can have multiple mRNA targets and one mRNA can be targeted by multiple miRNAs (Lee et al., 2015; Wu et al., 2010).

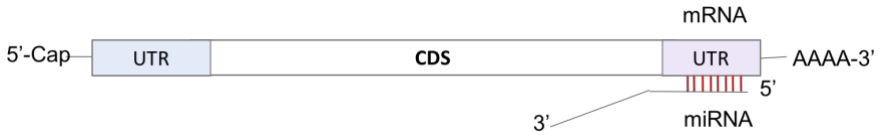


Figure 2. miRNA-mRNA interaction. MiRNA-mediated regulation involves binding between its seed region located at the 5'-end and miRNA target. Seed region is shown in red (Barrett et al., 2012; Peterson et al., 2014). mRNA = messenger RNA, miRNA = microRNA, UTR = untranslated region, CDS = coding sequence.

1.2. Bioinformatics pipeline for miRNA functional analysis

In order to annotate the functions of miRNAs, the bioinformatics pipeline usually consists of two steps; miRNA target prediction and functional enrichment of miRNA targets (Liu et al., 2014). There is a copious amount of miRNA target prediction tools and databases available and it can be difficult to choose one to work with, with the risk of missing out on important targets or getting a lot of false-positive target hits. One option is to use a combination of tools and this approach has been applied in many studies, including pancreatic cancer studies (Liang et al., 2018; Zhang et al., 2013). A previous approach used 12 different prediction programs and the overlapping targets were further studied for gene ontology (GO) terms enrichment in Omicshare tools (www.omicshare.org), Kyoto encyclopedia of genes and genomes (KEGG) and protein-protein interactions (PPI) for identifying prognostic miRNA markers in pancreatic ductal adenocarcinoma (Liang et al., 2018). Another study used TargetScan, PicTar and miRanda tools for predicting targets together with comparative proteomics analysis (Zhang et al., 2013).

For automated analyses including both miRNA target prediction and functional enrichment, DNA intelligent analysis (DIANA)-mirPath v3.0 can be used (Vlachos et al., 2015b). DIANA-mirPath is a web-based, automated bioinformatics pipeline where the user submits one or

more miRNAs as input and get miRNA target genes, KEGG pathways and GO terms as output (Vlachos et al., 2015b). The target prediction uses DIANA-microT-CDS v5.0 (Paraskevopoulou et al., 2013) and TargetScan v6.2 (Agarwal et al., 2015) as well as the experimentally validated target database DIANA-Tarbase v7.0 (Vlachos et al., 2015a). For functional enrichment statistical tests, the user can choose between Fisher's exact test and an unbiased empirical distribution (Bleazard et al., 2015). A Taverna plug-in was developed for the previous mirPath version 2.1, where the user can via the interface get direct access to the server and enable the integration of miRNA analysis into own-designed pipelines (Paraskevopoulou et al., 2013). DIANA-mirPath has previously been implemented for circulating miRNA functional analysis in various fields such as pediatric astrocytomas (López-Aguilar et al., 2017) and kidney disease (Xie et al., 2017).

1.2.1. miRNA target prediction

There has been a rapid development of new bioinformatics tools for target prediction with at least 75 databases of computational predictions available today, and a lot of them use a combination of existing tools (Tokar et al., 2018). The differences between the tools is that they use various algorithms and parameters for target prediction. MiRNA target prediction tools use different parameters, such as conservation, site accessibility, free energy and seed match (Peterson et al., 2014). The prediction tools often generate a large amount of false-positives and it is difficult to confirm the true targets without experimental validation (Singh, 2017).

Most of the algorithms that exist today search for targets in the 3'-UTR region of mRNAs, such as DIANA-microT (Paraskevopoulou et al., 2013; Reczko et al., 2012), microRNA.org (Betel et al., 2008) and miRDB (Wang, 2016; Wong and Wang, 2015). TargetScan also searches the 3'-UTR of targets for 6-8 mer matches and is among the mostly used tools for miRNA target prediction (Agarwal et al., 2015; Riffo-Campos et al., 2016). To further extend the search to other regions, the DIANA-microT-CDS algorithm was developed that also covers the coding sequence (Reczko et al., 2012).

Many tools combine already existing algorithms, such as miRWalk that combines 12 prediction tools and performs predictions in 3'-UTR, 5'-UTR, CDS and promoter regions, hence this tool covers all possible binding sites of mRNAs (Dweep et al., 2011). However, miRWalk has been regarded as unstable and it usually generates a lot of hits (Liu et al., 2014). In addition to prediction algorithm-based databases, sources that extract experimentally validated miRNA target genes from the literature also exist. A comparison between four such databases showed that Tarbase 7.0 outperformed miRWalk 2.0, miRTarBase and miRecords in terms of stability (Lee et al., 2015).

1.2.2. Functional enrichment analysis

In order to assign a function to a specific miRNA, the genes identified in the target prediction step can be used for functional enrichment analyses (Liu et al., 2014). The challenge of analyzing long lists of genes lead to the development of enrichment tools that can be used for identifying shared biological functions of the genes in a long list (Huang et al., 2009a). Differences between enrichment tools depends on the use of statistical analyses, for example Fisher's exact test, chi-square test or hypergeometric test. Examples of enrichment

tools that exist today are DAVID (Huang et al., 2009a, b), Funrich (Pathan et al., 2015) and GOMA (Huang et al., 2013).

When analyzing lists of miRNAs rather than a single miRNA, an unempirical distribution's algorithm developed by Bleazard et al. (Bleazard et al., 2015) is preferable to bring the enrichment analysis from the gene to the miRNA level. The bias identified by Bleazard et al. arise from hypergeometric tests or Fisher's exact test when the overlap of functional enrichment of each miRNA target gene set is not accounted for. The unempirical distribution's algorithm include adjustment by randomly generated functionally enriched terms.

1.2.3. Validation of miRNA functional analysis

The most preferable way of evaluating miRNA target predictions is experimental validation. However, this is not always possible for all predicted miRNA targets. Several databases containing experimentally validated targets exist and using these databases generates validated miRNA target genes (Lee et al., 2015). A previous thesis project in bioinformatics compared DIANA-TarBase, TargetScan and DIANA-microT-CDS target predictions as well as the enrichment tools DAVID and Toppfun for 18 miRNAs (Papagiannidis, 2015). The databases were evaluated by comparing the results to previous studies. Some of the miRNAs that were found by the DIANA-TarBase were not predicted by the prediction tools, indicating that the prediction algorithms also generate false-negatives.

Another validation approach is to correlate miRNA and mRNA expression levels in combination with miRNA target prediction. One source for this kind of data is the cancer genome atlas (TCGA, http://cancergenome.nih.gov/), which contains a lot of different types of data, such as miRNA, mRNA expression, mutation and phenotype data for various cancer types. Some studies have only focused on negative correlations (Bong et al., 2017; Zhang et al., 2014), whereas others have included both negative and positive correlations (Seo et al., 2017; Wang and Li, 2009).

Many studies have focused on mRNA level to see the effect of a specific miRNA, whereas the effect of miRNA-mediated regulation might in some cases only be visible on protein level (Wilczynska and Bushell, 2015). In addition, mRNA expression levels does not always correlate with protein expression levels as shown previously with an average Spearman correlation coefficient between mRNA and protein expression of 0.2 (Seo et al., 2017). Seo et al. (2017) integrated protein expression in addition to miRNA and mRNA levels to identify miRNAs related to glioblastoma. Protein expression levels performed on the same samples as found in TCGA can be accessed through the cancer proteome atlas (TCPA, http://bioinformatics.mdanderson.org/main/TCPA:Overview) containing expression analysis for around 200 proteins (Li et al., 2013). For miRNA target prediction, the study utilized only experimentally validated databases. The human protein atlas (HPA, www.proteinatlas.org) has also been used as support for interpretation of the miRNA functional role and its targets in different cancer types (Liang et al., 2018).

1.2.4. R packages for miRNA functional analysis

There are several R packages available for miRNA functional analysis. multiMiR (Ru et al., 2014) and RBiomirGS (Zhang and Storey, 2018) are packages that utilize an online repository

containing three experimentally validated databases and eight predicted databases. multiMiR is limited to perform miRNA target prediction, whereas RBiomirGS includes functional enrichment analysis as well. The R package MiRComb utilizes mRNA and miRNA expression assuming a negative correlation between miRNA and mRNA expression (Vila-Casadesús et al., 2016). MiRComb starts with miRNA-mRNA expression correlations and then performs miRNA target prediction on negatively correlated targets. For target prediction, it utilizes two databases; microcosm (Griffiths-Jones, 2006; Rehmsmeier et al., 2004) and TargetScan (Lewis et al., 2005), with restriction to conserved predictions. miRLAB is an R package that performs target prediction, enrichment analysis and includes mRNA and miRNA expression data from TCGA to infer regulatory relationships (Le et al., 2015). This package ranks the negative correlations highest but seems to include positive correlations as weak support for regulatory relationships. The retrieval of TCGA data is based on the TCGA-assembler package (Zhu et al., 2014), which downloads miRNA sequencing data from TCGA. This package is described as very flexible and the pipeline allows the user to include other datasets as well (Le et al., 2015).

1.3. Biomarkers in pancreatic cancer

Pancreatic cancer (PC) is often discovered in a late clinical stage, and the prognosis is very poor due to metastatic spread (Franklin et al., 2017). The most commonly used diagnostic biomarker today is carbohydrate antigen 19-9 (Ca19-9) but this biomarker has several disadvantages including suboptimal specificity, since it is also highly elevated in other diseases, and false negative detections (Poruk et al., 2013). Hence, a major goal is to find novel biomarkers. MiRNAs are highly stable in blood and have been studied as potential non-invasive biomarkers in numerous diseases, including pancreatic cancer (Franklin et al., 2017; Schultz et al., 2014; Xu et al., 2016). Recently, differentially expressed miRNAs in plasma were studied in PC patients before and at diagnosis (Franklin et al., 2017). Fifteen significantly altered miRNAs at diagnosis could be identified and a combination of these miRNA markers was shown to outperform Ca 19-9 at diagnosis.

Studies have used different approaches for miRNA functional analysis in the context of pancreatic cancer. A previous approach used miRWalk and overlapping targets were further studied for enriched GO terms in Omicshare tools (www.omicshare.org), enriched KEGG pathways and protein-protein interactions (PPI) for identifying prognostic miRNA markers in pancreatic ductal adenocarcinoma (Liang et al., 2018). The study utilized the TCGA database for identifying miRNAs with prognostic potential in pancreatic cancer. Ten miRNAs were identified, none of which overlaps with the results generated by Franklin et al. (2017) (Table 1). A combination of five of these miRNAs (miR-125a, miR-1301, miR-376b, miR-376c and miR-328) had the most predictive prognostic potential, with a higher expression level corresponding to a better prognosis. Hub genes were identified and validated by the human protein atlas repository. The validated hub genes examined in the human protein atlas were EGFR, HRAS, UBC, ESR1, AR, SMAD4 and MAPK1. Another study used TargetScan, PicTar and miRanda tools for predicting targets together with comparative proteomics analysis (Zhang et al., 2013).

Another study used miRcomb package for identifying miRNA-mRNA pairs in pancreatic cancer (Vila-Casadesus et al., 2018). The dataset studied consisted of 3 healthy pancreatic

tissue and 9 pancreatic cancer tissues. The top 50 correlations coefficients between miRNA and mRNA expression levels were between -0.85 and -0.97. For the correlation analyses, the healthy and cancerous tissues were combined and generated clustering in many cases contributing to the strong correlation coefficient. Among the strongest correlations were miR-106b, miR-144, miR-34a and miR-122, which were also identified in plasma by Franklin et al. (2017) (Table 1).

1.4. Problem formulation

As more high-throughput technologies become available and the sharing of high-throughput data increases, the pressure for developing bioinformatics tools also increases. There are several issues in the bioinformatics approach for miRNA functional analysis.

Firstly, the choice of miRNA target prediction tools and functional enrichment method has a great impact on the outcome and comparisons of tools are always of value for increasing the knowledge of important parameters in target predictions. As mentioned, only within the field of pancreatic cancer, a lot of different tools have been used and it does not seem to exist a consensus pipeline for this purpose.

Secondly, downloading data from DIANA-mirPath v3.0 (Vlachos et al., 2015b), as described in section “1.2. Bioinformatics pipeline for miRNA functional analysis”, is not optimal and could be developed further. Searching for one microRNA and downloading corresponding tables for target genes, KEGG pathways and GO terms and then structure them is time-demanding and could be optimized. Another problem is that the format for downloading KEGG pathways and GO terms is in comma-separated values (.csv), and since the terms themselves might contain commas, some rows get shifted when separated and manually changing these is not efficient.

Thirdly, validating the target prediction results. Validation of identified miRNA target genes is a challenge and an intermediate step from prediction to wet lab validation is of great benefit to narrow down interesting candidates. One possible *in silico* evaluation is to correlate miRNA expression levels to mRNA and/or protein expression levels. It is important to include protein levels since mRNA levels do not always reflect the protein levels (Seo et al., 2017) and due to the fact that miRNA could have a reversible repression mechanism (Wilczynska and Bushell, 2015), which in theory would not affect mRNA levels but only protein levels of the miRNA target gene. Many previous studies also focus only on negative correlations as miRNAs usually act as down-regulators of its target genes. However, positive regulation has occurred and thus it is important to not exclude positive correlation.

Fourthly, the TCGA miRNA-sequencing data does not contain defined 3p and 5p arms unless the miRNA isoform quantification data is used (Kuo et al., 2015). The TCGA repository has been utilized to obtain expression levels of miRNAs in several studies related to miRNA functional analysis. For those R packages that include TCGA miRNA-sequencing data, and in the miRNA, mRNA and protein integration study by (Seo et al., 2017), the restriction of the TCGA to miRNAs without -3p or -5p annotations of mature miRNAs is not addressed. Using miRNA gene expression levels could affect the results to a high degree depending on which mature miRNA is studied. Since mature miRNAs originating from the same stem loop have

different sequences (they are reverse complementary to each other) and thus different targets, a higher resolution of this data would provide a more specific evaluation.

1.5. Aim

The aim of this study is to develop an automated miRNA functional analysis and *in silico* evaluation pipeline in R that includes expression data of mature miRNA isoforms from the TCGA repository. This aim will be achieved by:

1) Defining -3p and -5p arms of miRNA isoform expression sequencing data from TCGA
2) Building a database for miRNA target prediction that is queried from R
3) Including functional enrichment functions available in R
4) Correlating expression levels of miRNA and its predicted target genes on mRNA and protein levels.

1.6. Limitations

This study will be limited to 15 miRNAs studied in the plasma of pancreatic cancer patients (Franklin et al., 2017). The 15 significantly altered miRNAs are measured in blood plasma whereas the miRNA, mRNA and protein expression data for *in silico* evaluation is derived from pancreatic cancer tissue. When detecting circulating biomarkers, it is possible that there is another origin than the disease organ. The protein expression levels from TCGA are limited to around 200 analyzed proteins (Li et al., 2013). The bioinformatics pipeline developed in R will be limited to the context of pancreatic cancer.

2. Materials & Methods

2.1. Bioinformatics pipeline

The bioinformatics pipeline (Fig. 3) was implemented in R version 3.4.3 (R core team, 2017). MiRNA target prediction of 15 miRNAs differentially altered in plasma of pancreatic cancer (PC) patients was performed as described in “2.3. miRNA target prediction” (Table 1) (Franklin et al., 2017). Enrichment analysis of miRNA target genes was performed twice; for all miRNA target genes and again for the correlated miRNA targets as described in section “2.4 Functional enrichment analysis”.

The -3p and -5p arms of mature miRNAs were annotated to miRNA-sequencing data as described in section “2.5.2. Annotation of mature miRNAs”. Correlations between expression levels of miRNA and mRNA, and miRNA and protein of identified miRNA targets were performed as an *in silico* evaluation, described in section “2.5. *In silico* evaluation”, to support miRNA target prediction in the context of pancreatic cancer.

Separate structured query language (SQL) databases were generated for miRNA target prediction (section “2.3. miRNA target prediction”), miRNA-mRNA correlation and miRNA-protein correlations (section “2.5.1. Pre-processing of expression data”). The databases were

developed outside the automated pipeline described (Fig. 3) that were then queried from the automated pipeline.

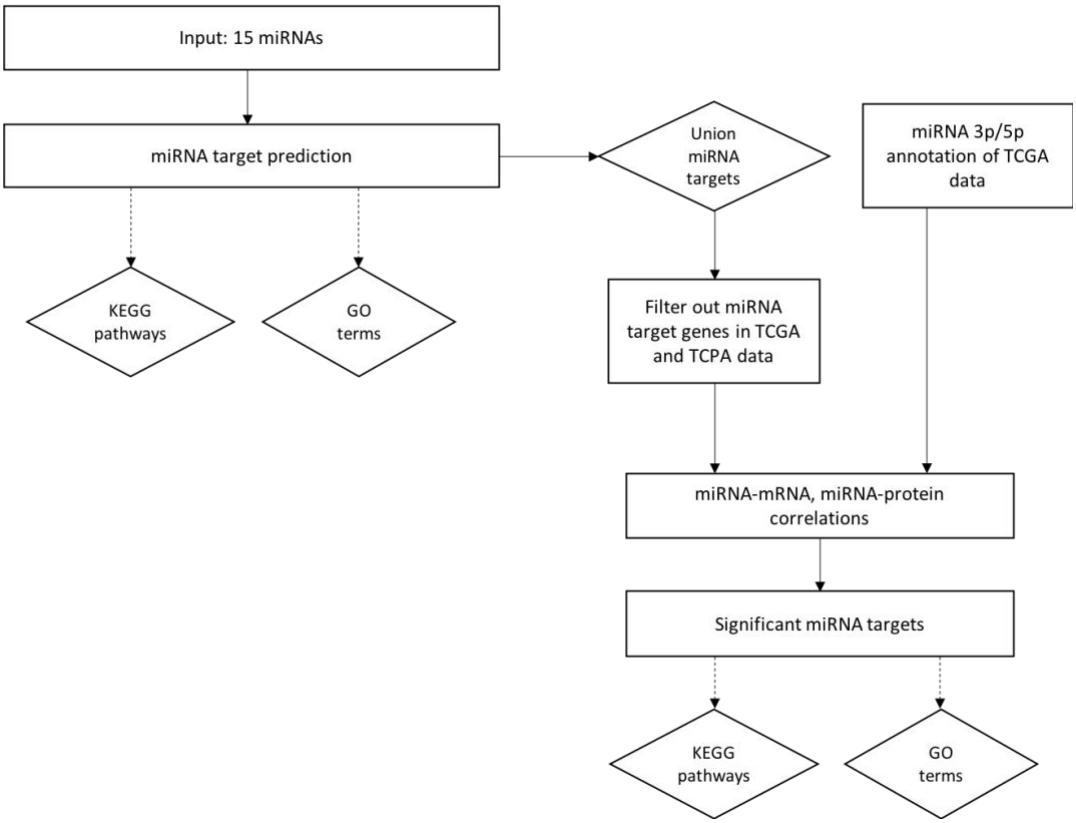


Figure 3. An overview of the bioinformatics pipeline. Fifteen microRNAs identified as differentially expressed in the plasma of pancreatic cancer (Franklin et al., 2017) were queried for miRNA target prediction and functional enrichment analysis for KEGG pathways and GO terms. The union miRNA targets identified in the three miRNA target prediction databases were used to filter out target genes from mRNA and protein expression lists. Before incorporating miRNA-sequencing data into the pipeline, -3p and -5p arms were annotated for miRNAs. To evaluate the target predictions, correlations were performed between miRNA-mRNA and miRNA-protein expression levels. The list of significant correlations was also subjected to functional enrichment analysis. miRNA = microRNA, KEGG = Kyoto encyclopedia of genes and genomes, GO = gene ontology, TCGA = the cancer genome atlas, TCPA = the cancer proteome atlas.

2.2. Data

2.2.1. miRNAs

Fifteen significantly altered miRNAs detected in plasma of pancreatic cancer (PC) patients at the time of diagnosis were used in this study (Franklin et al., 2017). These circulating miRNAs were identified in plasma samples from patients diagnosed with pancreatic cancer and admitted for surgery at the Department of Surgery, Umeå university hospital. The PC patients included in this study were diagnosed during 1990-2009. MiRNA isolates from 23 PC patients and 22 controls were analyzed by RT-qPCR for 372 validated miRNAs using Human Panel I (V.4, Exiqon, Vedbaek, Denmark). The combination of these 15 miRNAs generated an area under curve (AUC) of 0.96 compared to 0.92 for CA 19-9 at time of diagnosis.

Table 1. Significantly altered plasma miRNAs in pancreatic cancer patients (Franklin et al., 2017). Regulation describes whether the miRNA was found up- or down-regulated in pancreatic cancer patients.

miRNA	Regulation	FC
miR-144-3p	Down	0.4
miR-106b-5p	Down	0.8
miR-451a	Down	0.5
miR-101-3p	Down	0.7
miR-26a-5p	Down	0.6
miR-574-3p	Up	1.5
miR-885-5p	Up	3.9
miR-130b-3p	Up	1.5
miR-34a-5p	Up	2.2
miR-24-3p	Up	1.2
miR-22-5p	Up	1.4
let-7d-3p	Up	1.3
miR-197-3p	Up	1.4
miR-423-3p	Up	1.3
miR-122-5p	Up	2.5

FC= fold change

2.2.2. Expression data

MiRNA and mRNA expression have previously been generated by next-generation sequencing (seq) within the cancer genome atlas (TCGA) Research Network: http://cancergenome.nih.gov/. The miRNA-seq isoform expression quantification data files available at the GDC portal (https://portal.gdc.cancer.gov/) for pancreatic adenocarcinoma (PAAD) cases was used and processed according to section "2.5.1. Pre-processing of expression data". For mRNA expression data, log2-transformed count values for 177 cases were downloaded from xena browser (https://xenabrowser.net/datapages/). Protein expression levels, analyzed by reverse-phase protein arrays (RPPA), on tissue samples provided by TCGA were obtained (Li et al., 2013). Level 4 protein expression data was downloaded from the cancer proteome atlas (TCPA; http://bioinformatics.mdanderson.org/main/TCPA:Overview) by accessing the data portal (http://tcpaportal.org/tcpa/download.html). A total of 105 PAAD samples has previously been analyzed by RPPA for 218 proteins.

2.3. miRNA target prediction

For miRNA target prediction, three different databases were used; the experimentally validated database DIANA-TarBase v7 (Vlachos et al., 2015a) and two *in silico* target prediction databases, DIANA-microT-CDS (Reczko et al., 2012) and TargetScan v7.1 (Agarwal et al., 2015). Permission to download DIANA-TarBase v7 was given as well as a link for download (Vlachos et al., 2015a). DIANA-microT-CDS was downloaded from http://diana.imis.athena-innovation.gr/DianaTools/index.php. Three TargetScan v7.1 databases were downloaded (http://www.targetscan.org/vert_71/) containing predicted targets for conserved miRNA families, predicted conserved sites for miRNAs and predicted non-conserved sites miRNAs.

The downloaded databases were combined into a structured query language (SQL) database called 'mirna_database.sqlite' using sqlite3. The database mirna_database.sqlite contained the following five tables; 'Tarbase', 'microT_CDS','TagetScan_conserved', 'TargetScan_conserved_site' and 'TargetScan_nonconserved_site'. Since it was not of interest to look separately at the three different TargetScan databases, the results from the three TargetScan databases were immediately merged into one referred to as simply TargetScan. The SQL database was queried using dbConnect() and dbGetQuery() functions in RSQLite package (Müller et al., 2018). For each miRNA (Table 1), a search was performed against each table of the mirna_database.sqlite database. The overlap of identified miRNA target genes from Tarbase, microT-CDS and TargetScan was visualized with Venn diagrams, generated by using the R package VennDiagram (Chen, 2018).

DIANA-microT-CDS threshold can be set between 0.4-1 for *Homo sapiens*. No clear guidelines are provided for the choice of microT threshold, other than the fact that a lower threshold provides a higher number of targets and possibly higher number of false-positives. Vlachos et al. (Vlachos et al., 2012) suggested a higher threshold (0.9) if the user wants to find out pathways affected by a specific miRNA and a lower threshold if the user wishes to study if a group of miRNAs are involved in a specific pathway under definite conditions (0.7). Here, a threshold of 0.7 was chosen.

2.3.1. Alternative miRNA target prediction tools

According to Lee et al. (Lee et al., 2015), DIANA-Tarbase v7 outperformed three other experimentally validated tools; miRWalk 2.0 (Dweep et al., 2011), miRTarBase (Chou et al., 2018; Hsu et al., 2011) and miRecords (Xiao et al., 2009). The authors claimed miRWalk to be instable and to mirror miRTarBase. miRWalk 3.0 (http://129.206.7.150/) is now available but not complete yet and thus this database was not chosen for this project. Since the comparison was performed by (Lee et al., 2015), miRTarBase has been updated and therefore this database could have been chosen as well for identifying experimentally validated targets. However, according to database statistics, this database has 422,517 predicted targets in total, which is lower than that for Tarbase v7 that contains more than 500,000 targets.

Microrna.org is a resource that uses a miRanda target algorithm (Enright et al., 2003) and mir-SVR score (Betel et al., 2010). The miRanda algorithm is a target prediction based on sequence complementarity in 3'-UTR of mRNAs including thermodynamic analyses (Enright et al., 2003). The mir-SVR score is based on a machine learning method that ranks the targets according to a down-regulation score (Betel et al., 2010). The microrna.org database was not chosen since the -3p or -5p definition of the miRNA are not stated. This could lead to false positives if the user is specifically interested in only one of the -3p or -5p arms. This study uses miRNAs with defined -3p and -5p arms and thus it was important that the target prediction separates between these as well (Table 1). None of the above-mentioned miRNA target prediction tools include functional enrichment or correlation analyses based on mRNA or protein expression levels.

DIANA-miRpath v.3 is an automated pipeline that performs miRNA target prediction as well as functional enrichment of target genes (Vlachos et al., 2015b). A great advantage with this tool is that the user does not have to take the predicted target genes and use them as input

in another resource for functional enrichment. However, as described in section "1.4. Problem formulation", the tool is suboptimal when it comes to downloading the data.

2.4. Functional enrichment

Functional enrichment analysis was performed on miRNA target genes. Since a high number of genes in a functional enrichment input can lead to a lower biological significance, this pipeline was designed to also perform the functional enrichment after *in silico* evaluation, on the genes that were found to have a significant correlation (Vlachos et al., 2015b). Functional enrichment included enrichment analyses of GO terms and KEGG pathways. The enriched GO terms and KEGG pathways were achieved by functions goana() and kegga(), respectively, available in the edgeR package (McCarthy et al., 2012; Robinson et al., 2010). A cutoff at 5 genes and a false discovery rate of 0.05 was used.

2.4.1. Alternative functional enrichment tools

There are numerous enrichment tools to choose from, approximately 70 resources exist for this purpose (Huang et al., 2009a). The database for annotation, visualization and integrated discovery (DAVID, (Huang et al., 2009b) is a commonly used enrichment tool. Another resource is Omicshare tools previously used by Liang et al. (Liang et al., 2018). These were not chosen since it was more efficient to implement the edgeR package in R rather than analyzing the miRNA target genes outside of the pipeline.

2.5. *In silico* evaluation

In silico evaluation was performed to support the miRNA target genes identified by correlating miRNA expression levels to mRNA and protein expression levels of its target genes. Correlation analyses between miRNA and mRNA and protein expression levels were performed. Since miRNAs can function as up- or down-regulators of mRNAs, positive and negative correlations were included (Rusk, 2008; Vasudevan et al., 2007).

2.5.1. Pre-processing of expression data

A few pre-processing steps were necessary prior to correlation analysis (Fig. 4). Since a slightly different set of PAAD samples had been generated by miRNA-sequencing, mRNA-sequencing and RPPA, the pre-processing was performed prior to miRNA-mRNA correlations and prior to miRNA-protein correlations, separately. The setdiff() function was used on the column names, where the column names corresponded to sample names, to see which samples were present in only miRNA or mRNA data sets. Samples present in only one of miRNA or mRNA data sets were deleted to obtain the exact same sets of samples for miRNA and mRNA data (Fig. 4A). Only one sample was found present in only miRNA or mRNA expression data sets and this sample was deleted. For the protein expression data there was a difference of 85 samples between protein and miRNA expression data. A different approach was used here, which was to keep the columns present in both tables to automatically exclude the samples that only had expression data for either proteins or miRNAs (Fig. 4B). The remaining columns were ordered by column names to achieve the same order of samples to ensure correct correlation.

Two separate SQL databases were generated; miRNAmRNACor.sqlite and miRNAProteinCor.sqlite. miRNAmRNACor.sqlite contained miRNA and mRNA expression data tables for 177 samples and miRNAProteinCor.sqlite contained miRNA and protein expression data tables for 98 samples (Fig. 4). The mRNA expression levels consisted of Ensembl-IDs and the output from miRNA target prediction was HUGO Gene Nomenclature Committee (HGNC) symbols, therefore the transcriptome names were translated to HGNC symbols using org.Hs.eg.db package with the mapIds() function (Carlson, 2017) kindly provided by Hendrik Arnold de Weerd (University of Skövde, Skövde, Sweden). For protein expression data, the gene nomenclature used HGNC symbols so this procedure was not necessary for protein-miRNA correlation database.

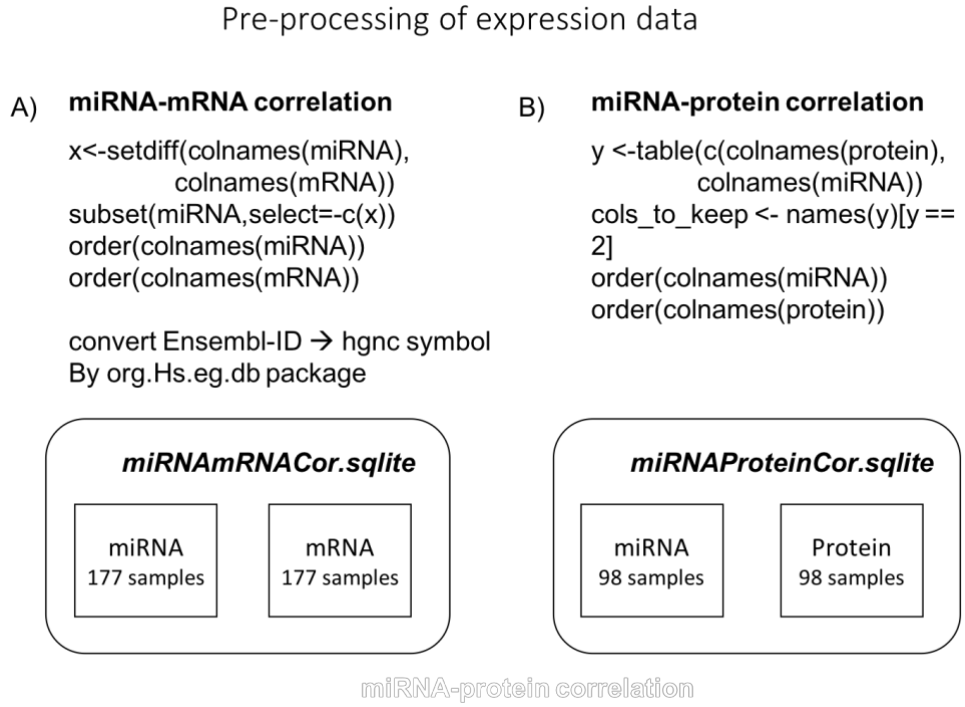


Figure 4. Overview of the pre-processing and structure of the databases for correlation analyses. Pre-processing steps were necessary to obtain the same samples and order to enable proper correlation analyses. Two separate databases were generated; one for miRNA-mRNA correlations (A) and one for miRNA-protein correlations (B).

2.5.2. Annotation of mature miRNAs

Mature miRNAs derived from the same stemloop may have different targets, and thus it is important to annotate the -3p and -5p to its respective expression pattern when possible. However, the miRNA sequencing expression data found in TCGA does not contain information about -3p or -5p arm, a problem addressed by Kuo et al. (2015) that also developed a Python script to interrogate this information. Their idea was applied here using R. The miRNA isoform expression quantification data was utilized from TCGA. A reads per million (rpm) threshold of 1 for calling a gene expressed was applied (Hebenstreit et al., 2011; Wagner et al., 2013). Kuo et al (Kuo et al., 2015) only included the top three expressions of each isoforms for each mature miRNA, whereas here, all values with rpm ≥ 1 was summarized for each miRNA. A 'gdc sample sheet', containing information such as file ID and sample ID for each PAAD case, was also downloaded (Fig. 5A). The gdc sample sheet was used as input for the R function. For each quantification file (Fig. 5B), the function consisted

of summarizing the reads per million (rpm) counts ≥ 1 for each MIMA-ID using the function ddply() from the plyr package (Wickham, 2011). The rpm values were log2-transformed. All samples were merged into one table using merge() function, with 'all'=T, resulting in samples as colnames and MIMAT-IDs in the first column. The MIMAT-ID is an identity for each unique, mature miRNA. A file containing information on MIMAT-ID and mature miRNA names (hsa.gff3) was downloaded from miRBase version 22 (www.mirbase.org). The MIMAT-IDs were translated using Perl version 5.18.2 and the miRNA nomenclature file (hsa.gff3). The final output was a table of the expression levels for each mature miRNA and PAAD sample (Fig. 5C).

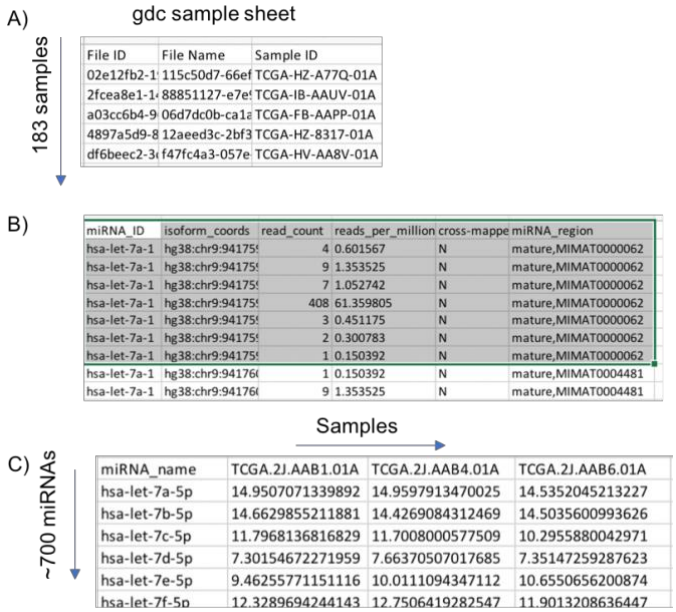


Figure 5. Overview of -3p and -5p annotation of TCGA miRNA-seq data and structure of output table. A) The function was applied to pancreatic adenocarcinoma samples using each row of the gdc sample sheet as input. B) For each sample, the reads per million counts ≥ 1 for each unique MIMAT-ID was summarized and log2-tranformed. C) The expression data for all samples were merged into one table and the MIMAT-IDs converted to mature miRNA names resulting in one output table containing expression of ~700 miRNAs for 183 samples.

2.5.3. Correlation

Correlation is a method for investigating the association between two numerical variables (Hazra and Gogtay, 2016). The correlation coefficient can be either positive or negative and between -1 to +1, where -1 is perfect, negative correlation and 1 is perfect, positive correlation, whereas 0 indicates no correlation. Another method for dependency studies between two variables is mutual information which gives a score between 0-1, where 0 is no dependency and 1 is perfect dependency (Lindlof and Lubovac, 2005). Mutual information was not suitable for this study as it cannot calculate for negative dependency.

2.5.4. Correlation settings

The unique() function was applied to the target gene lists to remove duplicate target genes information. The intersection of genes would imply strong evidence for miRNA-mRNA relationships but at the same time exclude potentially important targets. The union() function was chosen here since the pipeline contains an *in silico* evaluation of the target genes and, as mentioned, targets only predicted by one of the databases could still indicate a true target.

Correlations between expression levels for miRNA and mRNA, and miRNA and protein expression levels were performed. The cor() and cor.test() functions were applied using the

Pearson correlation method, a commonly used method for normally distributed data, with option to exclude data points with NA values. A significance threshold was set at $\alpha = 0.05$ and multiple testing adjustment using the Benjamini-Hochberg method was performed. A consensus correlation coefficient threshold is difficult to find, however a value below 0.3 indicates a "poor" correlation and this cutoff was used here for filtering out correlated pairs (Hazra and Gogtay, 2016).

One challenge for correlation methods is choosing a correlation coefficient threshold. An appropriate threshold is important for avoiding false positives but at the same time being able to identify related variables (Perkins and Langston, 2009). Various approaches have been proposed, for example a method using spectral graph theory where the underlying data structure is taken into account (Perkins and Langston, 2009). In this study, a value of above 0.3 or below (-0.3) will be used as threshold for positive or negative correlation. Correlation coefficients was divided into different strengths of correlation as suggested by Hazra and Gogtay (Hazra and Gogtay, 2016):

- Strong correlation: 0.7 - 1.0 or (-0.7) - (-1.0)
- Good correlation: 0.5 - 0.7 or (-0.5) - (-0.7)
- Moderate correlation: 0.3 - 0.5 or (-0.3) - (-0.5)

Hazra and Gogtay (Hazra and Gogtay, 2016) highlight the importance of creating scatterplots prior to correlation calculations in order to see if any correlation is indicated and to detect outliers that can affect the correlation, rather than solely relying on the correlation coefficient. Since this study included many correlations, only the strong correlations (PCC ≥ 0.7; PCC ≥ -0.7) were visualized by the *plot()* function in R.

2.5.5. miRNA-mRNA-protein integration

To find miRNA-target gene pairs that correlate on both mRNA and protein levels, the union() function was first used to combine positive and negative correlation, separately on mRNA and protein levels. The intersect() function was used on the mRNAs and proteins to generate lists of target genes correlated on both levels for each miRNA.

2.6. Identifying hub genes

There are many possible analyses that can be performed on the output genes from this pipeline. Here, protein-protein networks (PPI) with identified highly interacting proteins, so called hub genes, were studied for the supported target genes. Since there were many correlated miRNA-target pairs on mRNA level, PPI networks were generated separately for negatively and positively correlated target genes, on target genes of all 15 miRNAs together.

Protein-protein interaction network of positive correlations was only performed on the top 2000 genes since this was the limit of input genes in the STRING database. Gene lists were submitted to the STRING database version 10.5 (https://string-db.org/cgi/input.pl) with *Homo sapiens* as organism and an interaction score of 0.7. The network was downloaded as tab-separated values (tsv) file, pre-processed to contain four columns; Node1, interaction, Node2 and combined_score. The interaction column contained only the value 'combined_score'. The cytohubba plugin (Chin et al., 2014) was used in Cytoscape version

3.6.0 to show the ranked, top 10 hub genes (Fig. 3). Due to a low number of significant correlations on protein level, this step was only performed on correlated miRNA-mRNA expression levels.

3. Results

3.1. Number of target genes partially overlap between databases

MiRNA target prediction was performed in three databases; TargetScan v7.1, DIANA-TarBase v7 and DIANA-microT-CDS. Most number of predicted targets was generally identified from TargetScan, exceeding 3000 predicted target genes for many of the miRNAs (Fig. 6). In contrast, no target gene was found in TargetScan for hsa-miR-101-3p. For five of the miRNAs (hsa-miR-106b-5p, hsa-miR-101-3p, hsa-miR-26a-5p, hsa-miR-130b-3p, hsa-miR-22-5p), the number of genes was comparable between DIANA-TarBase and DIANA-microT-CDS.

The generated Venn diagrams showed that many of the target genes in DIANA-TarBase were not identified by the *in silico* prediction tools (Fig. S1). A threshold for DIANA-microT-CDS was set at 0.7 which affects how many targets are predicted. Lowering this threshold could perhaps generate more targets that are also in DIANA-TarBase, but on the other hand it also generates a higher number of false positives. The opposite scenario also occurs, that targets predicted by TargetScan or DIANA-microT-CDS have not been experimentally validated.

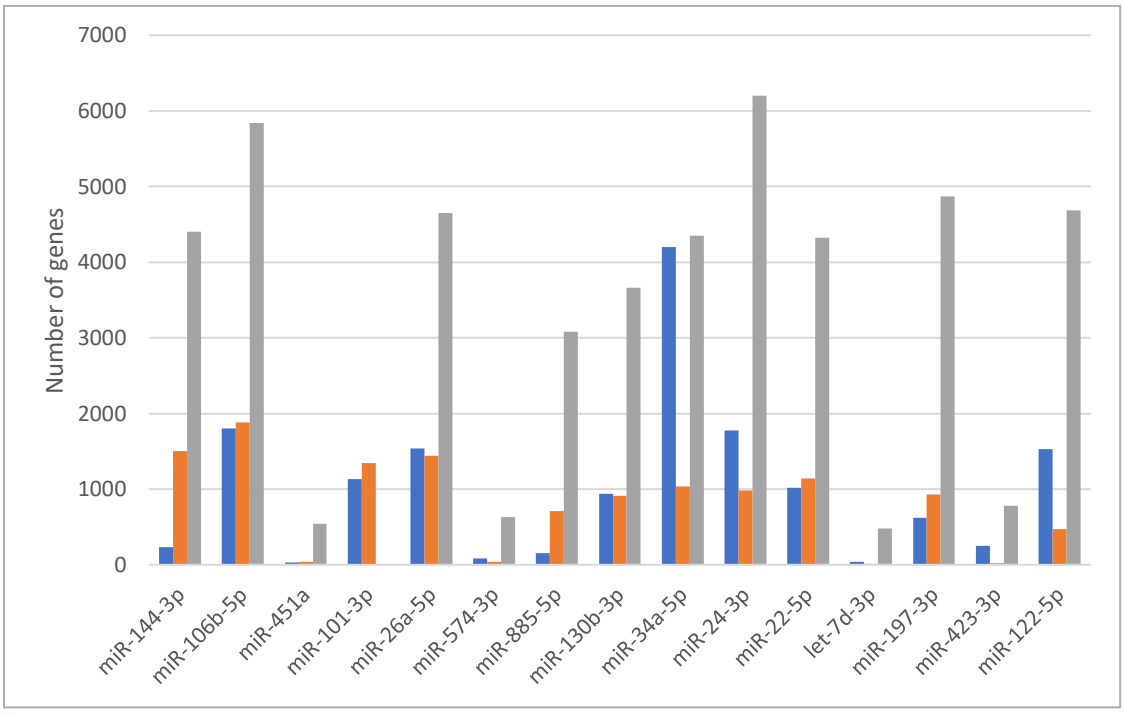


Figure 6. Number of target genes predicted in DIANA-TarBase v7, DIANA-microT-CDS and TargetScan v7.1 for fifteen miRNAs. The mirna_database.sqlite was queried for fifteen miRNAs and the number of genes were recorded for every miRNA and miRNA target prediction database. The x axis shows every miRNA queried and the y axis the number of identified genes. DIANA-TarBase v7 is shown in blue, DIANA-microT-CDS is shown in orange and TargetScan v7.1 is shown in grey.

3.2. Strong correlations were observed for miR-885-5p

As miRNA target prediction tools can result in a lot of false positives or targets that do not have an essential role for the disease or condition of interest, *in silico* evaluation data is useful to narrow down interesting gene candidates. To identify target genes that might have an importance in pancreatic cancer progression, expression data of miRNAs, mRNAs and proteins from PC tissue was used to find correlations between the query miRNA and its corresponding target genes on mRNA and protein levels.

For all 15 miRNAs combined, a total of 3526 significant correlations (*p*-value < 0.05) were discovered, of which 1331 were positively correlated (PCC ≥ 0.3), and 2195 negatively correlated (PCC ≤ -0.3). The number of unique targets was compared to the number of positive and negative correlations on mRNA level. In general, the percentage correlations of the total unique miRNA targets was low. miR-885-5p showed the highest number of significant correlations (PCC ≥ 0.3; PCC ≤ -0.3) followed by miR-106b-5p and miR-24-3p, whereas miR-144-3p and miR-122-5p showed no significant correlations (Fig. 7).

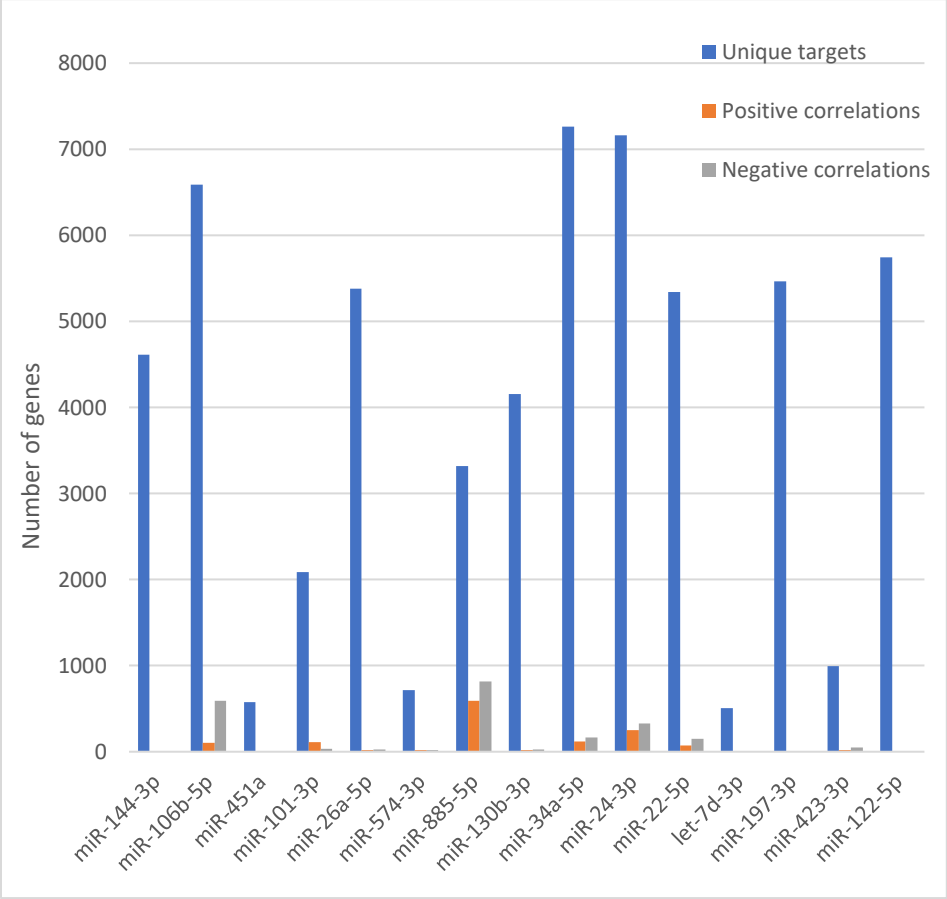


Figure 7. Number of unique miRNA targets, positively correlated and negatively correlated miRNA target genes for 15 miRNAs. Unique targets (shown in blue) for each miRNA was plotted together with the specific number of positive correlations (shown in orange, PCC ≥ 0.3) and negative correlations (shown in grey, PCC ≤ -0.3). The y axis shows number of target genes and the x axis shows each miRNA.

An additional filtering step was performed to extract strong correlations (PCC ≥ 0.7; PCC ≤ -0.7), which resulted in 81 strong positive correlations and 172 strong negative correlations. This referred only to hsa-miR-885-5p and some of its target genes (Tables S1 and S2). This miRNA had the highest average fold change (3.9) in the plasma of pancreatic cancer patients (Table 1) (Franklin et al., 2017). The strong correlations were plotted to be able to detect outliers that could affect the correlation coefficient (Fig. 8). As seen in the number of target predictions (Fig. 6), the number of targets predicted for hsa-miR-885-5p is not particularly higher than the other miRNAs. By plotting the strong correlations, it was observed that clustering occurred as well as outliers. There is also a possibility that these four outliers are of biological relevance and that hsa-miR-885-5p is a potential prognostic biomarker.

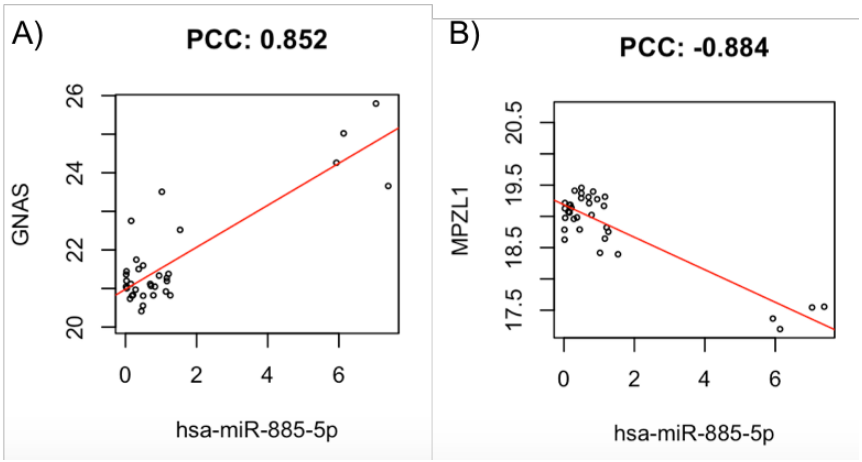


Figure 8. Scatterplots of the strongest correlations for miR-885-5p. Correlation was performed using Pearson method and the strong correlations (PCC ≥ 0.7; PCC ≤ -0.7) were plotted. Correlations between the strongest positive correlation between GNAS and miR-885-5p (A) and the strongest negative correlation between MPZL1 and miR-885-5p (B) are shown demonstrating clustering and four outliers. GNAS = GNAS complex locus, MPZL1 = myelin protein zero like 1, hsa = Homo sapiens, PCC = Pearson correlation coefficient.

The low number of expression values could be one possible reason for the strong correlations observed for hsa-miR-885-5p. To test this hypothesis, the miRNA expression levels from each of the miRNA were separately extracted from the database used for miRNA-mRNA correlations and the number of patients showing expression of each miRNA was counted. This was performed by creating a vector and extracting the length with na.omit as option to only count the values and not the NA elements. All miRNAs except for hsa-miR-144-3p (89 samples), hsa-miR-885-5p (31 samples) and hsa-miR-122-5p (9 samples) show expression for all PAAD samples (Table 2).

Table 2. Number of miRNA-sequencing expression values for each miRNA.

miRNA	Number of samples with miRNA expression
hsa-miR-144-3p	89
hsa-miR-106b-5p	177
hsa-miR-451a	177
hsa-miR-101-3p	177
hsa-miR-26a-5p	177
hsa-miR-574-3p	177
hsa-miR-885-5p	31
hsa-miR-130b-3p	177
hsa-miR-34a-5p	177
hsa-miR-24-3p	177
hsa-miR-22-5p	177
hsa-let-7d-3p	177
hsa-miR-197-3p	177
hsa-miR-423-3p	177
hsa-miR-122-5p	9

3.2.1. miRNA-protein correlation

Correlations between miRNA and protein expression levels were performed on 98 samples. In total, nine significant correlations (p-value < 0.05) were identified on protein level, of which 5 were positively correlated (PCC ≥ 0.3) and 4 negatively correlated (PCC ≤ -0.3). Only four miRNAs (miR-24-3p, miR-885-5p, miR-101-3p and miR-22-5p) showed significant correlations on protein expression level (Table 3). miR-24-3p had a total of 4 negative and 2 positive correlations, and thus had the highest number of significant correlations on protein level.

Table 3. Significant correlations (PCC ≥ 0.3; PCC ≤ -0.3) between miRNA and its target gene on protein level.

miRNA	Protein	PCC
hsa-miR-24-3p	ASNS	-0.31
hsa-miR-24-3p	JAK2	-0.34
hsa-miR-24-3p	ATM	-0.30
hsa-miR-24-3p	CDK1	-0.36
hsa-miR-101-3p	PDCD4	0.31
hsa-miR-885-5p	MSH6	0.71
hsa-miR-24-3p	INPP4B	0.35
hsa-miR-24-3p	EGFR	0.34
hsa-miR-22-5p	PEA15	0.38

One of the miRNA-protein expression pairs showed strong correlation (PCC ≥ 0.7); hsa-miR-885-5p and MutS homolog 6 (MSH6, PCC = 0.712, Fig. 9). Plotting this correlation showed one outlier causing a strong correlation.

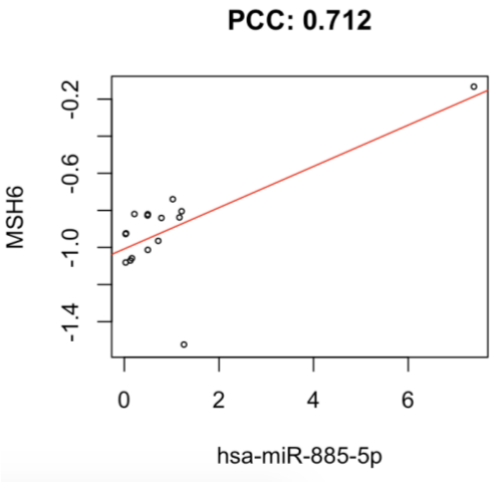


Figure 9. Scatterplot of hsa-miR-885-5p and MSH6. Correlation was performed using Pearson method and the strong positive correlation (PCC ≥ 0.7) between MSH6 and miR-885-5p was plotted. hsa = Homo sapiens, PCC = Pearson correlation coefficient, MSH6 = MutS homolog 6.

3.2.2. miRNA-mRNA-protein integration

Three miRNA-target gene pairs were found on both mRNA and protein levels (PCC ≥ 0.3; PCC ≤ -0.3, Table 4). Inositol polyphosphate-4-phosphatase type II B (INPP4B) and hsa-miR-24-3p was found correlated on both mRNA and protein level. The correlation coefficient was nearly on the same level as mRNA (PCC = 0.38) and protein (PCC = 0.35) levels. For miR-24-3p-INPP4B and miR-101-3p-PDCD4, the correlation coefficient was positive on both mRNA and protein level. On the contrary, miR-24-3p-CDK1 was found positively correlated on mRNA level and negatively correlated on protein level.

Table 4. Number of miRNA-sequencing expression values for each miRNA.

miRNA	Gene	PCC (mRNA)	PCC (protein)
miR-24-3p	CDK1	0.38	-0.36
miR-24-3p	INPP4B	0.38	0.35
miR-101-3p	PDCD4	0.42	0.31

PCC = Pearson correlation coefficient

3.3. Functional enrichment analysis

Functional enrichment analysis was performed on all miRNA target genes before and after *in silico* evaluation, *i.e.* on correlated (PCC ≥ 0.3; PCC ≤ -0.3) miRNA target genes. The highest number of GO terms and KEGG pathways was seen for hsa-miR-34a-5p and hsa-miR-24-3p before *in silico* evaluation, respectively (Tables 5-8). After *in silico* evaluation, the miRNAs (miR-885-5p, miR-106b-5p and miR-24-3p) with a high number of significant correlations (PCC ≥ 0.3; PCC ≤ -0.3, Fig. 7) also turned out as the top 3 in number of GO terms (Table 6) and KEGG pathways (Table 8). The total number of GO terms and KEGG pathways decreased after *in silico* evaluation compared to before. As not all of the identified target genes generated correlated pairs, it was expected to see a decrease in the functional enrichment analysis.

The GO term and KEGG pathway with lowest P-value before and after *in silico* evaluation were extracted along with the number of miRNA target genes for each term and total number of GO terms or KEGG pathways observed (Tables 5-8). GO terms (Table 6) and KEGG pathways (Table 8) after *in silico* evaluation were only generated for 10 and 7 out of 15 miRNAs, respectively. The pipeline did not generate GO terms or KEGG pathway results for miR-144-3p or miR-122-5p since no significant correlations were identified for these two miRNAs. This explains the "NA" values in number of KEGG pathways, whereas a value of 0 indicates that significant correlations were found but no enriched KEGG pathways for these miRNA target genes could be identified.

Table 5. Top GO term for each miRNA before *in silico* evaluation. The number of genes enriched for each term is shown along with the P-value for the top GO term and the total number of GO terms obtained for each miRNA.

miRNA	Top GO term	Number of genes	P-value	Number of GO terms
hsa-miR-144-3p	regulation of cellular macromolecule biosynthetic process	1172	8.56E-30	1725
hsa-miR-106b-5p	cellular macromolecule metabolic process	3078	1.59E-39	1554
hsa-miR-451a	positive regulation of cellular process	196	5.93E-10	643
hsa-miR-101-3p	regulation of macromolecule metabolic process	898	5.45E-48	1938
hsa-miR-26a-5p	cellular macromolecule metabolic process	2586	2.20E-40	1470
hsa-miR-574-3p	regulation of macromolecule metabolic process	283	5.85E-10	601
hsa-miR-885-5p	gene expression	1039	3.14E-16	913
hsa-miR-130b-3p	regulation of metabolic process	1579	9.42E-30	1744
hsa-miR-34a-5p	cellular macromolecule metabolic process	3176	8.28E-39	2118
hsa-miR-24-3p	cellular metabolic process	4020	5.51E-42	1514
hsa-miR-22-5p	cellular component organization	1853	8.07E-26	1656
hsa-let-7d-3p	histone mRNA metabolic process	5	0.0002543	198
hsa-miR-197-3p	cellular metabolic process	3098	2.37E-35	1220
hsa-miR-423-3p	regulation of nitrogen compound metabolic process	309	5.43E-14	980
hsa-miR-122-5p	cellular metabolic process	3180	1.35E-24	959

Table 6. Top GO term for each miRNA after *in silico* evaluation. The number of genes enriched for each term is shown along with the P-value for the top GO term and the total number of GO terms obtained for each miRNA.

miRNA	Top GO term	Number of genes	P-value	Number of GO terms
hsa-miR-144-3p	NA	NA	NA	NA
hsa-miR-106b-5p	single-multicellular organism process	297	1.99E-10	722
hsa-miR-451a	NA	NA	NA	0
hsa-miR-101-3p	intracellular signal transduction	45	5.63E-09	551
hsa-miR-26a-5p	cell projection assembly	5	0.00381475	11
hsa-miR-574-3p	cellular response to nitrogen compound	5	0.0028732	11
hsa-miR-885-5p	developmental process	508	5.92E-07	703
hsa-miR-130b-3p	negative regulation of gene expression	10	0.00326584	26
hsa-miR-34a-5p	macrophage activation	7	2.04E-05	277
hsa-miR-24-3p	cell division	44	6.21E-09	736
hsa-miR-22-5p	chromosome organization	35	3.45E-07	260
hsa-let-7d-3p	NA	NA	NA	0
hsa-miR-197-3p	NA	NA	NA	0
hsa-miR-423-3p	positive regulation of intracellular signal transduction	14	3.77E-06	152
hsa-miR-122-5p	NA	NA	NA	NA

Table 7. Top KEGG pathway for each miRNA before *in silico* evaluation. The number of genes enriched for each term is shown along with the P-value for the top KEGG pathway and the total number of KEGG pathways obtained for each miRNA.

miRNA	Top KEGG pathway	Number of genes	P-value	Number of KEGG pathways
hsa-miR-144-3p	Proteoglycans in cancer	82	1.60E-09	95
hsa-miR-106b-5p	FoxO signaling pathway	67	7.15E-07	95
hsa-miR-451a	FoxO signaling pathway	11	0.00121457	28
hsa-miR-101-3p	Proteoglycans in cancer	46	2.11E-07	76
hsa-miR-26a-5p	Cellular senescence	75	5.62E-09	92
hsa-miR-574-3p	Hippo signaling pathway	16	0.00022937	55
hsa-miR-885-5p	Ubiquitin mediated proteolysis	38	0.00017599	61
hsa-miR-130b-3p	TGF-beta signaling pathway	40	1.35E-08	101
hsa-miR-34a-5p	Cellular senescence	94	7.90E-11	113
hsa-miR-24-3p	Proteoglycans in cancer	104	3.37E-07	119
hsa-miR-22-5p	ErbB signaling pathway	50	3.45E-10	105
hsa-let-7d-3p	EGFR tyrosine kinase inhibitor resistance	9	9.73E-05	19
hsa-miR-197-3p	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	33	0.00037731	73
hsa-miR-423-3p	Focal adhesion	29	1.48E-06	84
hsa-miR-122-5p	Regulation of actin cytoskeleton	78	0.0003571	68

Table 8. Top KEGG pathway for each miRNA after *in silico* evaluation. The number of genes enriched for each term is shown along with the P-value for the top KEGG pathways and the total number of KEGG pathways obtained for each miRNA.

miRNA	Top KEGG pathway	Number of genes	P-value	Number of KEGG pathways
hsa-miR-144-3p	NA	NA	NA	NA
hsa-miR-106b-5p	Rap1 signaling pathway	22	3.64E-06	38
hsa-miR-451a	NA	NA	NA	0
hsa-miR-101-3p	Focal adhesion	9	7.00E-06	12
hsa-miR-26a-5p	NA	NA	NA	0
hsa-miR-574-3p	NA	NA	NA	0
hsa-miR-885-5p	Cell adhesion molecules (CAMs)	25	1.74E-05	44
hsa-miR-130b-3p	NA	NA	NA	0
hsa-miR-34a-5p	Adherens junction	5	0.00278269	4
hsa-miR-24-3p	Cell cycle	15	1.06E-05	26
hsa-miR-22-5p	Cushing's syndrome	7	0.0024336	5
hsa-let-7d-3p	NA	NA	NA	0
hsa-miR-197-3p	NA	NA	NA	0
hsa-miR-423-3p	Human papillomavirus infection	5	0.00971859	2
hsa-miR-122-5p	NA	NA	NA	NA

3.4. E2 and E3 ubiquitin ligases were identified as hub genes.

The cytohubba plugin generates a ranking of the hub genes where rank 1 is the hub genes with most connections. For the negatively correlated genes, all hub genes were of rank 1 (Fig. 10A). For positively correlated genes, UBE2C was of rank 1 and the rest rank 2 (Fig. 10B). Members of ubiquitin-conjugating enzymes (UBE), involved in protein degradation (Xie et al., 2014), was ranked as 1 for positive (UBE2C) and was also found among top 10 negatively correlated genes (UBE2D4). UBE2D4 correlated negatively with miR-106b-5p and miR-24-3p. Interestingly, miR-24-3p also correlated positively with UBE2C, connecting these two hub genes. Some similarities were identified between the two PPIs. All hub genes identified in both PPI are in different ways involved in protein degradation. Protein degradation can be divided into three enzymatic steps containing the following; 1) ubiquitin-activating enzyme E1, ubiquitin-conjugating enzyme E2 and ubiquitin ligase E3 (Sun, 2006). For both PPIs, 1 ubiquitin-conjugating enzyme and 9 ubiquitin E3 ligases were identified.

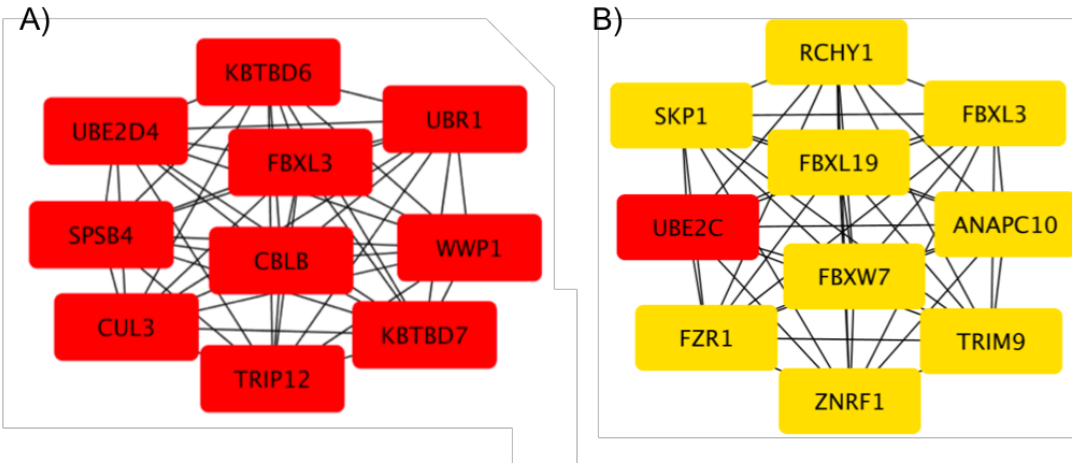


Figure 10. Top 10 hub genes for significant A) negative and B) positive correlations for all 15 miRNAs. Hub genes were identified by the cytohubba plugin (Chin et al., 2014) in cytoscape version 3.6.0. Red = rank 1, yellow = rank 2.

4. Discussion

4.1. miRNA target prediction

Many miRNA target prediction algorithms and combinations of these are available today; however, this pipeline was restricted to the use of three databases; DIANA-TarBase v7 (Vlachos et al., 2015a), DIANA-microT-CDS (Reczko et al., 2012) and TargetScan v7.1 (Agarwal et al., 2015). DIANA-TarBase, an experimentally validated database, was previously shown to outperform other validated target databases (Lee et al., 2015). We still depend on target prediction algorithms since experimental validation takes time, and thus TargetScan v7.1 (Agarwal et al., 2015) and DIANA-microT-CDS (Reczko et al., 2012) were implemented as well. Among the most common algorithms used today, TargetScan has been regarded as the most robust one, covering 3'-UTR of mRNAs (Riffo-Campos et al., 2016). In addition, DIANA-microT-CDS was chosen to further cover the coding region as well (Reczko et al., 2012). The promoter region or 5'-UTR of target mRNAs were not covered by any prediction tools implemented here.

Similar to previous studies (Papagiannidis, 2015), the miRNA target prediction in DIANA-TarBase v7, DIANA-microT-CDS and TargetScan v7.1 generated different number of targets and an overlap was seen between these databases. However, many of the targets were only predicted by one or two of the databases. This is not surprising since databases use different algorithms and the mechanisms behind miRNA-mediated regulation are not yet fully understood (Agarwal et al., 2015). In addition, experimental validation takes more time than *in silico* prediction and some of the predicted targets might not have been experimentally validated yet. The generated Venn diagrams (Fig. S1) also indicate that the mechanisms behind miRNA-mRNA binding is not yet fully understood and that the algorithms need further development to include the experimentally validated targets (Agarwal et al., 2015).

As no consensus pipeline for miRNA functional analysis exists today and choosing a tool can be difficult, there is still a need for unbiased, comprehensive benchmarking studies of

miRNA target prediction algorithms. Another challenge is how to validate the large amount of predicted miRNA targets in these benchmarking studies.

4.2. *In silico* evaluation of miRNA target prediction

To validate every predicted miRNA target experimentally is impossible, it can also be difficult to choose a target to start with. For this purpose, an *in silico* evaluation consisting of correlation analyses between miRNA and its predicted targets was implemented to extract the miRNAs with highest impact in a disease-specific manner. As miRNAs are considered down-regulators, some studies have only included negative correlations (Bong et al., 2017; Zhang et al., 2014). However, miRNAs have also been shown to act as up-regulators (Rusk, 2008; Vasudevan et al., 2007), and thus some studies have included positive correlations as well (Seo et al., 2017; Wang and Li, 2009). (Seo et al., 2017) was one of the first studies to include protein expression data in the correlations as miRNA-mediated regulation acts post-transcriptionally and thus mainly affects the protein expression levels. miRNAs regulate their targets by degradation or repression and an effect on the protein level might not always be noticed on mRNA level (Wilczynska and Bushell, 2015). Hence, when possible, the protein expression levels should be included in correlation-based *in silico* evaluation.

The mechanisms behind miRNA-mediated regulation are highly complex and act in a disease- or tissue-specific manner (Wu et al., 2017). MiRNA target prediction provides miRNA target genes but no information on other factors that could affect the extent to which miRNA regulate its targets. Hence, validation is needed to provide additional support to the targets identified and extract genes of interest in a disease-specific context, in this case pancreatic cancer. Correlation analysis is one way of determining the dependency between two variables (Hazra and Gogtay, 2016) and was performed on mRNA and protein expression levels of the target genes to its corresponding miRNA, analyzed in pancreatic cancer tissue. In this study, it was applied on expression levels of miRNA and its target genes on both mRNA and protein levels. Correlation analyses do not automatically indicate that the dependency is a direct mechanism, however since the miRNA-gene pairs were predicted to interact, it gives a strong support that the effect might be direct. The expression analyses were performed on the same tissue samples and should provide a good support for the translational regulation by miRNAs. Many significant correlations were identified, and it is difficult to assess their values based only on bioinformatics methods. Experimental validation would be useful to further support this method as an *in silico* evaluation.

One limitation for using correlation analyses based on mRNA and protein expression data is that there is a possibility that the predicted target is not among the expression data. This means that an otherwise strong target could be excluded only due to the fact that it is not available for performing correlation. For the protein assay analyzing around 200 proteins this problem becomes larger since the chance is very small of finding the predicted target among the protein expression levels. Hence, there is a need for more high-throughput proteomics. No feature was included in the pipeline to show which targets were not available among mRNA or protein expression data.

The pipeline takes more time to run when correlation analyses are included, however the hands-on time is the same since the pipeline is automated. In fact, including the correlation

analyses saves time in post-processing steps of extracting interesting miRNA target candidates since the output list of interesting candidates becomes shorter after *in silico* evaluation. Performing functional enrichment of significantly correlated miRNA targets did not result in more significance when comparing the top GO terms or KEGG pathways for each miRNA (Tables 5-8). It also made the list of GO terms and KEGG pathways shorter and for some of the miRNAs, no enrichment terms were generated for correlated targets. No systematic review was performed, but it was observed that for some of the miRNAs the KEGG pathways that remained as enriched after *in silico* evaluation were more significant compared to before *in silico* evaluation.

A correlation coefficient threshold was set at ≥ 0.3 or ≤ -0.3, whereas a consensus threshold for this purpose does not exist. Experimental studies could be an aid in developing a threshold for correlation analyses between miRNA and its target genes, how strong correlation could one expect from a regulatory miRNA? Is it more interesting to look at a target with a correlation coefficient of 0.2 compared to 0.5 or 0.7 for instance? Another question is whether the same level of correlation coefficient threshold should be used for positive and negative correlations. For developing therapeutics applications, it would probably be of more importance to use a miRNA that gives a strong effect on its targets rather than targeting a fine-tuning miRNA. However, this also depends on the cellular context and condition.

4.2.1. Strong correlations were identified for miR-885-5p

Plotting the strong correlations (PCC ≥ 0.7;PCC ≤ -0.7), which only occurred for miR-885-5p, it was clear that clustering was present and four outliers of high expression (Fig. 8, log2(reads per million counts) > 5). In total, only 31 patients expressed this miRNA in TCGA pancreatic adenocarcinoma and thus the correlation only included these patients (Table 2). The four outliers could be experimental artifacts or indicate some biological importance, as supported by previous studies. In the original data of significantly altered circulating miRNAs in pancreatic cancer patients (Franklin et al., 2017), miR-885-5p had the highest fold change in pancreatic cancer patients and has been identified in several other studies as a circulating marker for pancreatic cancer (Ganepola et al., 2014; Hussein et al., 2017; Schultz et al., 2014). This study utilized input miRNAs detected in plasma samples of PC patients, whereas the expression data used for evaluation originated from tumor tissue. The circulating miRNAs could be a leakage from the tumor or a systematic response to the cancer state (Franklin et al., 2017). The information on how many of the PAAD patients in TCGA express each miRNA in the tumor tissue could help in understanding the origin of the circulating miRNA markers (Table 2). This result could support a hypothesis that some of the miRNA markers, such as hsa-miR-885-5p have another origin other than pancreas for causing the upregulation in PC plasma.

Some functional studies of miR-885-5p have been performed in the context of liver diseases and colorectal cancer. miR-885-5p has been found to act as a tumor suppressor in the context of hepatocellular carcinoma (Zhang et al., 2016) and hepatoblastoma (Indersie et al., 2017), by targeting β-catenin, which is involved in wnt/β-catenin signaling pathway. Further, a decay in expression correlated to a more progressed hepatocellular carcinoma by correlation to tumor-node-metastasis (TNM) stages (Zhang et al., 2016). Significant negative correlation was identified in this study between miR-885-5p and β-catenin (CTNNB1) on

mRNA level (PCC = -0.68). Correlation between miR-885-5p and CTNNB1 on protein level could not be performed since this protein was not included in the TCPA data. CTNNB1 was identified as a miR-885-5p target by TargetScan v7.1 and DIANA-microT-CDS only so it was not included in DIANA-TarBase v7, which highlights the need for including miRNA target prediction tools and not only validated databases. miR-885-5p was found up-regulated in liver metastases compared to the primary tumor in colorectal cancer (Vychytilova-Faltejskova et al., 2016). Another study confirmed that miR-885-5p was up-regulated in distant metastases in colorectal cancer, and suggested a regulation involving its predicted target gene CPEB2 (Lam et al., 2017). CPEB2 was identified as a miR-885-5p target gene by TargetScan and DIANA-microT-CDS in this study, however it did not result in any significant correlations. No functional studies have been performed in the context of pancreatic cancer, however miR-885-5p has been identified as a potential circulating miRNA biomarker by several studies (Franklin et al., 2017; Ganepola et al., 2014; Hussein et al., 2017; Schultz et al., 2014).

4.2.2. Functional enrichment of correlated targets

Functional enrichment was performed on miRNA target genes before and after *in silico* evaluation. Union of significant negative and positive correlations on mRNA and protein level was studied for enriched GO terms and KEGG pathways. GO terms were generated for 10 miRNAs and KEGG pathways for 7 miRNAs after *in silico* evaluation (Table S3). Terms associated to pancreas and pancreas cancer were searched for to extract the most interesting miRNAs. Twelve miRNAs (miR-144-3p, miR-106b-5p, miR-101-3p, miR-26a-5p, miR-885-5p, miR-130b-3p, miR-34a-5p, miR-24-3p, miR-22-5p, miR-197-3p, miR-423-3p and miR-122-5p) associated to KEGG pathway pancreatic cancer (path:hsa05212) before evaluation and only 2 miRNAs (miR-885-5p and miR-24-3p) after evaluation. The GO term pancreas development (GO:0031016), which could be interesting in the context of pancreatic cancer, was found associated to 7 miRNAs (miR-144-3p, miR-451a, miR-101-3p, miR-26a-5p, miR-130b-3p, miR-22-5p and miR-423-3p) before evaluation and zero after evaluation.

4.3. Protein-protein interaction networks

The specific roles of the identified hub genes (Fig. 10) need to be further evaluated, but interestingly, both positive and negative correlations generate similar results with respect to the roles of the hub genes, with involvement in protein degradation.

4.3.1. E2 ubiquitin-conjugating enzymes

Ubiquitin-conjugating enzyme E2 D4 (UBE2D4, Fig. 13A) and UBE2C (Fig. 13B) were identified as the only two hub genes belonging to the category of E2 ubiquitin-conjugating enzymes. UBE2D4 is being investigated as a therapeutic target in cancer and candidate ligands have been identified with affinity for this protein (Ramatenki et al., 2017). UBE2C was recently identified as a hub gene in pancreatic ductal adenocarcinoma and high expression correlated to a worse prognosis (Zhou et al., 2018).

4.3.2. E3 ubiquitin ligases

Many ubiquitin ligases E3 have been studied as cancer biomarkers and targets (Sun, 2006). The largest family of E3 ubiquitin ligases is the SCF and SCF-like proteins. Among the hub

genes identified, S-phase kinase associated protein 1 (Skp1), F-box proteins (FBXL3, FBXL19 and FBXW7) and cullin 3 (CUL3) would classify as these enzymes (Sun, 2006). Kelch repeat and BTB domain containing 6 (KBTBD6), KBTBD7 and cullin 3 (CUL3) cooperate in T cell lymphoma invasion and metastasis 1 (TIAM1) expression and an up-regulation of TIAM1 leads to Ras-related C3 botulinum toxin substrate 1 (Rac1) up-regulation (Genau et al., 2015). Elevated Rac1 expression resulted in impaired cell proliferation and increased cell invasion. Rac1 is often hyperactivated in pancreatic cancer (Yan et al., 2014). FBXL19 is involved in regulating cell migration through degradation of Rac1 (Zhao et al., 2013).

FBXW7 has been shown to have a metastasis-suppressive role in the host microenvironment, where stromal FBXW7-deficient mice developed more metastasis in a melanoma xenograft (Yumimoto et al., 2015). This deficiency caused increased chemokine CCL2 serum levels, following accumulation of macrophages and myeloid-derived suppressor cells, which in turn promoted metastasis. WW domain containing E3 ubiquitin protein ligase 1 (WWP1) is overexpressed in many cancer types and knockdown of this mRNA causes decreased cell invasion, cell growth, apoptosis and G1-phase arrest in osteosarcoma (Wu et al., 2015). Depletion of CBLB (casitas B-lineage lymphoma-b) leads to metastasis rejection by natural killer cells (Paolino et al., 2014). SPSB4 (SplA/Ryanodine Receptor Domain and SOCS Box Containing 4) has a role in cell repulsive responses of Eph receptor tyrosine kinases which are often overexpressed in various cancer types (Okumura et al., 2017).

Mutations in ubiquitin protein ligase E3 component N-Recognin 1 (UBR1) can cause Johanson-Blizzard syndrome where one among many symptoms is exocrine pancreatic insufficiency (Sukalo et al., 2014). A decreased expression of thyroid hormone receptor-interacting protein 12 (TRIP12) caused elevated levels of pancreas transcription factor 1a (PTF1a), a protein that is involved in early development of pancreas and in the conservation of acinar cell types (Hanoun et al., 2014). Ring finger and CHY zinc finger domain containing 1 (RCHY1) has several substrates and some of these such as p53 and c-myc are involved in cell death and proliferation (Bridoux et al., 2015). ANAPC10 is an anaphase promoting complex (APC) subunit and these are involved in cytokinesis, chromatid segregation and establishing G1 phase in the cell cycle (Chang et al., 2014). In late M phase and early G1 phase of the cell cycle, fizzy and cell division cycle 20 related 1 (FZR1) protein associates with APC complex, and when de-associated, FZR1 can suppress B-Raf proto-oncogene, serine/threonine kinase (BRAF) (Wan et al., 2017). Tripartite Motif Containing 9 (TRIM9) is involved in hippocampal development (Winkle et al., 2016). Zinc and ring finger 1 (ZNRF1) is involved in neuronal/axonal degeneration (Wakatsuki et al., 2015).

4.4. The importance of annotating mature miRNAs

Instead of using miRNA-sequencing data from TCGA, this study utilized miRNA isoform quantification data from TCGA to separate between expression levels of -3p and -5p arms of mature miRNAs. The importance of gaining a higher resolution of miRNA-sequencing data when available is illustrated by an example comparing log2-expression (reads per million counts) values for 182 TCGA patients for the miRNA gene mir-144 and the mature miRNAs; miR-144-3p and miR-144-5p (Fig. 11). This example illustrates the difference between miRNA gene and mature miRNA isoform expression and highlights the importance of deriving the expression specifically for the mature -3p miRNA in this case. In addition, a lot of the TCGA

samples showed no expression of miR-144-3p but did indeed show expression for miR-144-5p, something that would cause false-positive expression values if we would have relied on mir-144 expression as this miRNA gene expression pattern is more similar to the -5p arm expression in this case. Thus, the results will differ depending on whether the miRNA gene or mature miRNA isoform expression from TCGA is analyzed.

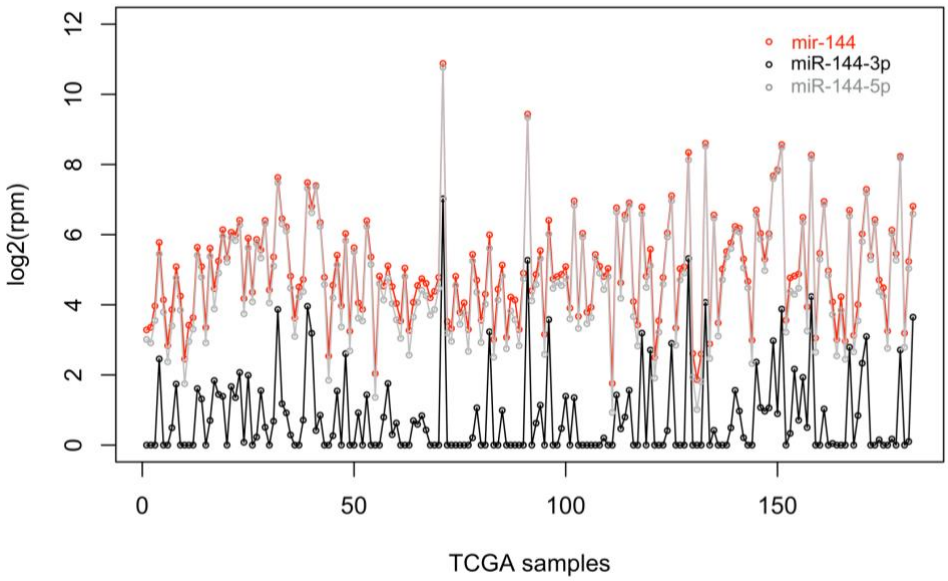


Figure 11. The difference between mir-144 (red), miR-144-3p (black) and miR-144-5p (grey). Log2-expression values were plotted for 182 TCGA samples. Rpm = reads per million counts, TCGA = the cancer genome atlas, mir = miRNA gene, miR = mature miRNA.

5. Conclusion

An automated bioinformatics pipeline for miRNA functional analysis and *in silico* evaluation of predicted targets by correlation was developed. Positive and negative correlations between miRNAs and targets, on mRNA and protein levels, were included as well as functional enrichment before and after *in silico* evaluation. Predicted miRNA-mRNA interactions in conjunction with correlation analyses of expression levels provides high support for the regulatory mechanisms of a specific miRNA. The pipeline was studied for 15 miRNAs but is applicable to any miRNA in the context of pancreatic cancer. In addition to the miRNA functional analysis pipeline, hub genes were identified of supported targets narrowing down the list of interesting candidate biomarkers or crucial players in the development of pancreatic cancer to analyze further. The top 10 identified hub genes, regardless of positive or negative correlations, are involved in protein degradation processes.

6. Future studies

6.1. Bioinformatics development

This pipeline was implemented in the context of pancreatic adenocarcinoma but could be extended to other disease modules. It is restricted to pancreatic cancer since the TCGA dataset for PAAD only was translated into expression levels of annotated mature miRNAs with defined -3p and -5p arms. It would also be possible to further extend the pipeline to include other target prediction databases. This pipeline included miRNA target prediction on three different databases, an interesting statistic would be to include a score from 1-3 where score 3 means the target is predicted by three databases, score 2 by 2 databases and score 1 by 1 database. The low number of analyzed proteins was a limitation in this study and more high-throughput proteomics data would be desirable to include for miRNA-protein correlations. As mentioned, the pipeline does not provide information about predicted targets that could not be found in mRNA or protein expression data. Such a feature should be implemented in the future to avoid excluding potentially interesting target candidates.

The pipeline needs to be developed to make fast and easy extraction of interesting enrichment terms since this step also generates high amounts of data (Table 1). A tool that detects and removes outliers could also be implemented as the correlation for miR-885-5p contained 4 outliers and might mask other important correlations. Lastly, the pipeline should be compared to other R packages or pipelines in terms of miRNA target genes, enrichment terms and hub genes of interesting candidate targets.

6.2. Experimental studies

Expression analyses of four biomarkers (cancer antigen 19-9, endostatin, collagen IV and carcinoembryonic antigen) is available for the PC patients that were used to identify the 15 altered plasma miRNAs (Franklin et al., 2017). Further analyses could also include correlation analysis between these four biomarkers, if they are identified as targets, and their corresponding miRNA. Due to its role in other cancer types, functional analysis of miR-885-5p in the context of pancreatic cancer could be of interest.

7. Impact on society

Cancer is a complex genetic disease that affects a lot of people (Urbach et al., 2012). Early detection is crucial to obtain better prognosis for pancreatic cancer patients (Kenner et al., 2017). However, pancreatic cancer is usually discovered late with a spread disease where surgery is not an option anymore and thus early biomarkers are needed (Franklin et al., 2017). MiRNAs are stable in blood and the interest of using these RNA species as biomarkers in various diseases have expanded lately. MiRNA functional analysis help in understanding the biological role, the origin of the miRNA and the impact of different miRNAs. MiRNA target prediction is also important for testing toxicity of novel miRNA therapeutic to predict other possible unwanted miRNA-target interactions. One miRNA therapeutic named miravirsen, targeting miR-122, is currently involved in several clinical trials for the treatment of hepatitis C virus infection (Chakraborty et al., 2017).

8. Ethical aspects

In research, there are always some ethical aspects when using data or specimens derived from humans. Consent must always be given in order to participate in research. The data used in this study were publicly available, hence no ethical approval was needed. Another aspect to remember is that using different bioinformatics tools other than the ones planned from the beginning or adding new analyses to the data might generate novel results that might be useful for the participants to know or not know about and then there might be a discussion on how to handle the new information. At the same time as we improve the analyses and mapping of the genome or epigenome and finding the functions for specific gene variants, we might ask ourselves whether this type of information in the future could be misused, for example by insurance companies. The research community strives to have transparency and it is not impossible that misuse of personal data might occur in some way in the future.

9. Acknowledgements

I would like to thank my supervisors Zelmina Lubovac (University of Skövde) and Malin Sund (Umeå University) for great support and their enthusiasm throughout this project. I also thank Hendrick Arnold de Weerd for great help with programming and my examinator Angelica Lindlöf at University of Skövde for valuable feedback. Thanks are also due to Tejaswi Venkata Satya Badam at the University of Skövde for assisting me with the implementation of 3p/5p annotation and Thomas Hillerton for help with Perl programming. I also thank my colleagues Hanna Nyström, Ola Billing and Pär Jonsson at Umeå University for great input.

10. References

Agarwal, V., Bell, G.W., Nam, J.-W., and Bartel, D.P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. eLife *4*, 101.

Barrett, L.W., Fletcher, S., and Wilton, S.D. (2012). Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. Cellular and molecular life sciences : CMLS *69*, 3613-3634.

Betel, D., Koppal, A., Agius, P., Sander, C., and Leslie, C. (2010). Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. Genome Biol *11*, R90.

Betel, D., Wilson, M., Gabow, A., Marks, D.S., and Sander, C. (2008). The microRNA.org resource: targets and expression. Nucleic acids research *36*, D149-153.

Bhaskaran, M., and Mohan, M. (2014). MicroRNAs: history, biogenesis, and their evolving role in animal development and disease. Veterinary pathology *51*, 759-774.

Bleazard, T., Lamb, J.A., and Griffiths-Jones, S. (2015). Bias in microRNA functional enrichment analysis. Bioinformatics *31*, 1592-1598.

Bong, I.P.N., Ng, C.C., Baharuddin, P., and Zakaria, Z. (2017). MicroRNA expression patterns and target prediction in multiple myeloma development and malignancy. Genes Genomics *39*, 533-540.
Bridoux, L., Deneyer, N., Bergiers, I., and Rezsohazy, R. (2015). Molecular Analysis of the HOXA2-Dependent Degradation of RCHY1. PLoS One *10*, e0141347.
Carlson, M. (2017). org.Hs.eg.db: Genome wide annotation for Human.
Chakraborty, C., Sharma, A.R., Sharma, G., Doss, C.G.P., and Lee, S.S. (2017). Therapeutic miRNA and siRNA: Moving from Bench to Clinic as Next Generation Medicine. Mol Ther Nucleic Acids *8*, 132-143.
Chang, L.F., Zhang, Z., Yang, J., McLaughlin, S.H., and Barford, D. (2014). Molecular architecture and mechanism of the anaphase-promoting complex. Nature *513*, 388-393.
Chen, H. (2018). VennDiagram: Generate High-Resolution Venn and Euler Plots.
Chin, C.H., Chen, S.H., Wu, H.H., Ho, C.W., Ko, M.T., and Lin, C.Y. (2014). cytoHubba: identifying hub objects and sub-networks from complex interactome. BMC Syst Biol *8 Suppl 4*, S11.
Chou, C.H., Shrestha, S., Yang, C.D., Chang, N.W., Lin, Y.L., Liao, K.W., Huang, W.C., Sun, T.H., Tu, S.J., Lee, W.H*., et al.* (2018). miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. Nucleic Acids Res *46*, D296-D302.
Dweep, H., Sticht, C., Pandey, P., and Gretz, N. (2011). miRWalk--database: prediction of possible miRNA binding sites by "walking" the genes of three genomes. Journal of biomedical informatics *44*, 839-847.
Ellwanger, D.C., Büttner, F.A., Mewes, H.-W., and Stümpflen, V. (2011). The sufficient minimal set of miRNA seed types. Bioinformatics (Oxford, England) *27*, 1346-1350.
Enright, A.J., John, B., Gaul, U., Tuschl, T., Sander, C., and Marks, D.S. (2003). MicroRNA targets in Drosophila. Genome biology *5*, R1.
Franklin, O., Jonsson, P., Billing, O., Lundberg, E., Öhlund, D., Nyström, H., Lundin, C., Antti, H., and Sund, M. (2017). Plasma Micro-RNA Alterations Appear Late in Pancreatic Cancer. Annals of surgery, 1.
Ganepola, G.A., Rutledge, J.R., Suman, P., Yiengpruksawan, A., and Chang, D.H. (2014). Novel blood-based microRNA biomarker panel for early diagnosis of pancreatic cancer. World J Gastrointest Oncol *6*, 22-33.
Genau, H.M., Huber, J., Baschieri, F., Akutsu, M., Dotsch, V., Farhan, H., Rogov, V., and Behrends, C. (2015). CUL3-KBTBD6/KBTBD7 ubiquitin ligase cooperates with GABARAP proteins to spatially restrict TIAM1-RAC1 signaling. Mol Cell *57*, 995-1010.
Griffiths-Jones, S. (2006). miRBase: the microRNA sequence database. Methods Mol Biol *342*, 129-138.
Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A., and Enright, A.J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res *34*, D140-144.
Griffiths-Jones, S., Saini, H.K., van Dongen, S., and Enright, A.J. (2008). miRBase: tools for microRNA genomics. Nucleic Acids Res *36*, D154-158.
Hanoun, N., Fritsch, S., Gayet, O., Gigoux, V., Cordelier, P., Dusetti, N., Torrisani, J., and Dufresne, M. (2014). The E3 ubiquitin ligase thyroid hormone receptor-interacting protein 12 targets pancreas transcription factor 1a for proteasomal degradation. J Biol Chem *289*, 35593-35604.
Hazra, A., and Gogtay, N. (2016). Biostatistics Series Module 6: Correlation and Linear Regression. Indian J Dermatol *61*, 593-601.

Hebenstreit, D., Fang, M., Gu, M., Charoensawan, V., van Oudenaarden, A., and Teichmann, S.A. (2011). RNA sequencing reveals two major classes of gene expression levels in metazoan cells. Mol Syst Biol *7*, 497.
Hsu, S.D., Lin, F.M., Wu, W.Y., Liang, C., Huang, W.C., Chan, W.L., Tsai, W.T., Chen, G.Z., Lee, C.J., Chiu, C.M.*, et al.* (2011). miRTarBase: a database curates experimentally validated microRNA-target interactions. Nucleic Acids Res *39*, D163-169.
Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic acids research *37*, 1-13.
Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols *4*, 44-57.
Huang, Q., Wu, L.Y., Wang, Y., and Zhang, X.S. (2013). GOMA: functional enrichment analysis tool based on GO modules. Chin J Cancer *32*, 195-204.
Hussein, N.A., Kholy, Z.A., Anwar, M.M., Ahmad, M.A., and Ahmad, S.M. (2017). Plasma miR-22-3p, miR-642b-3p and miR-885-5p as diagnostic biomarkers for pancreatic cancer. J Cancer Res Clin Oncol *143*, 83-93.
Indersie, E., Lesjean, S., Hooks, K.B., Sagliocco, F., Ernault, T., Cairo, S., Merched-Sauvage, M., Rullier, A., Le Bail, B., Taque, S.*, et al.* (2017). MicroRNA therapy inhibits hepatoblastoma growth in vivo by targeting beta-catenin and Wnt signaling. Hepatol Commun *1*, 168-183.
Kenner, B.J., Go, V.L.W., Chari, S.T., Goldberg, A.E., and Rothschild, L.J. (2017). Early Detection of Pancreatic Cancer: The Role of Industry in the Development of Biomarkers. Pancreas *46*, 1238-1241.
Kozomara, A., and Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res *39*, D152-157.
Kozomara, A., and Griffiths-Jones, S. (2014). miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res *42*, D68-73.
Kuo, W.T., Su, M.W., Lee, Y.L., Chen, C.H., Wu, C.W., Fang, W.L., Huang, K.H., and Lin, W.C. (2015). Bioinformatic Interrogation of 5p-arm and 3p-arm Specific miRNA Expression Using TCGA Datasets. J Clin Med *4*, 1798-1814.
Lam, C.S., Ng, L., Chow, A.K., Wan, T.M., Yau, S., Cheng, N.S., Wong, S.K., Man, J.H., Lo, O.S., Foo, D.C.*, et al.* (2017). Identification of microRNA 885-5p as a novel regulator of tumor metastasis by targeting CPEB2 in colorectal cancer. Oncotarget *8*, 26858-26870.
Le, T.D., Zhang, J., Liu, L., Liu, H., and Li, J. (2015). miRLAB: An R Based Dry Lab for Exploring miRNA-mRNA Regulatory Relationships. PLoS One *10*, e0145386.
Lee, Y.J.D., Kim, V., Muth, D.C., and Witwer, K.W. (2015). Validated MicroRNA Target Databases: An Evaluation. Drug development research *76*, 389-396.
Lewis, B.P., Burge, C.B., and Bartel, D.P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell *120*, 15-20.
Li, J., Lu, Y., Akbani, R., Ju, Z., Roebuck, P.L., Liu, W., Yang, J.-Y., Broom, B.M., Verhaak, R.G.W., Kane, D.W.*, et al.* (2013). TCPA: a resource for cancer functional proteomics data. Nature methods *10*, 1046-1047.
Liang, L., Wei, D.-M., Li, J.-J., Luo, D.-Z., Chen, G., Dang, Y.-W., and Cai, X.-Y. (2018). Prognostic microRNAs and their potential molecular mechanism in pancreatic cancer: A study based on The Cancer Genome Atlas and bioinformatics investigation. Molecular medicine reports *17*, 939-951.
Lindlof, A., and Lubovac, Z. (2005). Simulations of simple artificial genetic networks reveal features in the use of Relevance Networks. In Silico Biol *5*, 239-249.

Liu, B., Li, J., and Cairns, M.J. (2014). Identifying miRNAs, targets and functions. Briefings in bioinformatics *15*, 1-19.
López-Aguilar, J.E., Velázquez-Flores, M.A., Simón-Martínez, L.A., Ávila-Miranda, R., Rodríguez-Florido, M.A., and Ruiz-Esparza Garrido, R. (2017). Circulating microRNAs as Biomarkers for Pediatric Astrocytomas. Archives of medical research *48*, 323-332.
Lytle, J.R., Yario, T.A., and Steitz, J.A. (2007). Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5' UTR as in the 3' UTR. Proceedings of the National Academy of Sciences of the United States of America *104*, 9667-9672.
McCarthy, D.J., Chen, Y., and Smyth, G.K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res *40*, 4288-4297.
Millar, A.A., and Waterhouse, P.M. (2005). Plant and animal microRNAs: similarities and differences. Functional & integrative genomics *5*, 129-135.
Müller, K., Wickham, H., James, D.A., Falcon, S., [ctb], S.A., and Healy, L.c. (2018). RSQLite: 'SQLite' Interface for R.
Okumura, F., Joo-Okumura, A., Obara, K., Petersen, A., Nishikimi, A., Fukui, Y., Nakatsukasa, K., and Kamura, T. (2017). Ubiquitin ligase SPSB4 diminishes cell repulsive responses mediated by EphB2. Mol Biol Cell *28*, 3532-3541.
Orang, A.V., Safaralizadeh, R., and Kazemzadeh-Bavili, M. (2014). Mechanisms of miRNA-Mediated Gene Regulation from Common Downregulation to mRNA-Specific Upregulation. Int J Genomics.
Paolino, M., Choidas, A., Wallner, S., Pranjic, B., Uribesalgo, I., Loeser, S., Jamieson, A.M., Langdon, W.Y., Ikeda, F., Fededa, J.P.*, et al.* (2014). The E3 ligase Cbl-b and TAM receptors regulate cancer metastasis via natural killer cells. Nature *507*, 508-512.
Papagiannidis, F. (2015). Exploring the functions of miRNA targets in association with cadmium exposure on breast cancer. Unpublished thesis report, University of Skövde.
Paraskevopoulou, M.D., Georgakilas, G., Kostoulas, N., Vlachos, I.S., Vergoulis, T., Reczko, M., Filippidis, C., Dalamagas, T., and Hatzigeorgiou, A.G. (2013). DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. Nucleic acids research *41*, W169-173.
Pathan, M., Keerthikumar, S., Ang, C.S., Gangoda, L., Quek, C.Y., Williamson, N.A., Mouradov, D., Sieber, O.M., Simpson, R.J., Salim, A.*, et al.* (2015). FunRich: An open access standalone functional enrichment and interaction network analysis tool. Proteomics *15*, 2597-2601.
Perkins, A.D., and Langston, M.A. (2009). Threshold selection in gene co-expression networks using spectral graph theory techniques. BMC Bioinformatics *10 Suppl 11*, S4.
Peterson, S.M., Thompson, J.A., Ufkin, M.L., Sathyanarayana, P., Liaw, L., and Congdon, C.B. (2014). Common features of microRNA target prediction tools. Frontiers in genetics *5*, 23.
Poruk, K.E., Gay, D.Z., Brown, K., Mulvihill, J.D., Boucher, K.M., Scaife, C.L., Firpo, M.A., and Mulvihill, S.J. (2013). The clinical utility of CA 19-9 in pancreatic adenocarcinoma: diagnostic and prognostic updates. Current molecular medicine *13*, 340-351.
R core team (2017). R: A Language and Environment for Statistical Computing (Vienna, Austria).
Ramatenki, V., Dumpati, R., Vadija, R., Vellanki, S., Potlapally, S.R., Rondla, R., and Vuruputuri, U. (2017). Targeting the ubiquitin-conjugating enzyme E2D4 for cancer drug discovery-a structure-based approach. J Chem Biol *10*, 51-67.

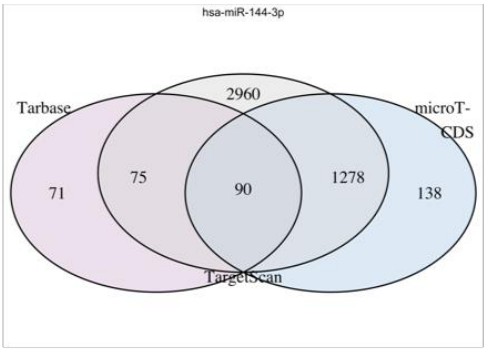
Reczko, M., Maragkakis, M., Alexiou, P., Grosse, I., and Hatzigeorgiou, A.G. (2012). Functional microRNA targets in protein coding sequences. Bioinformatics (Oxford, England) *28*, 771-776.
Rehmsmeier, M., Steffen, P., Hochsmann, M., and Giegerich, R. (2004). Fast and effective prediction of microRNA/target duplexes. RNA *10*, 1507-1517.
Riffo-Campos, A.L., Riquelme, I., and Brebi-Mieville, P. (2016). Tools for Sequence-Based miRNA Target Prediction: What to Choose? Int J Mol Sci *17*.
Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139-140.
Ru, Y., Kechris, K.J., Tabakoff, B., Hoffman, P., Radcliffe, R.A., Bowler, R., Mahaffey, S., Rossi, S., Calin, G.A., Bemis, L.*, et al.* (2014). The multiMiR R package and database: integration of microRNA-target interactions along with their disease and drug associations. Nucleic Acids Res *42*, e133.
Rusk, N. (2008). When microRNAs activate translation. Nature methods *5*, 122-123.
Schultz, N.A., Dehlendorff, C., Jensen, B.V., Bjerregaard, J.K., Nielsen, K.R., Bojesen, S.E., Calatayud, D., Nielsen, S.E., Yilmaz, M., Holländer, N.H.*, et al.* (2014). MicroRNA biomarkers in whole blood for detection of pancreatic cancer. JAMA *311*, 392-404.
Seo, J., Jin, D., Choi, C.-H., and Lee, H. (2017). Integration of MicroRNA, mRNA, and Protein Expression Data for the Identification of Cancer-Related MicroRNAs. PloS one *12*, e0168412.
Singh, N.K. (2017). miRNAs target databases: developmental methods and target identification techniques with functional annotations. Cellular and molecular life sciences : CMLS *74*, 2239-2261.
Sukalo, M., Fiedler, A., Guzman, C., Spranger, S., Addor, M.C., McHeik, J.N., Oltra Benavent, M., Cobben, J.M., Gillis, L.A., Shealy, A.G.*, et al.* (2014). Mutations in the human UBR1 gene and the associated phenotypic spectrum. Hum Mutat *35*, 521-531.
Sun, Y. (2006). E3 ubiquitin ligases as cancer targets and biomarkers. Neoplasia *8*, 645-654.
Tay, Y., Zhang, J., Thomson, A.M., Lim, B., and Rigoutsos, I. (2008). MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation. Nature *455*, 1124-1128.
Tokar, T., Pastrello, C., Rossos, A.E.M., Abovsky, M., Hauschild, A.-C., Tsay, M., Lu, R., and Jurisica, I. (2018). mirDIP 4.1-integrative database of human microRNA target predictions. Nucleic acids research *46*, D360-D370.
Urbach, D., Lupien, M., Karagas, M.R., and Moore, J.H. (2012). Cancer heterogeneity: origins and implications for genetic association studies. Trends Genet *28*, 538-543.
Vasudevan, S., Tong, Y., and Steitz, J.A. (2007). Switching from repression to activation: microRNAs can up-regulate translation. Science (New York, NY) *318*, 1931-1934.
Vila-Casadesús, M., Gironella, M., and Lozano, J.J. (2016). MiRComb: An R Package to Analyse miRNA-mRNA Interactions. Examples across Five Digestive Cancers. PloS one *11*, e0151127.
Vila-Casadesus, M., Vila-Navarro, E., Raimondi, G., Fillat, C., Castells, A., Lozano, J.J., and Gironella, M. (2018). Deciphering microRNA targets in pancreatic cancer using miRComb R package. Oncotarget *9*, 6499-6517.
Vlachos, I.S., Kostoulas, N., Vergoulis, T., Georgakilas, G., Reczko, M., Maragkakis, M., Paraskevopoulou, M.D., Prionidis, K., Dalamagas, T., and Hatzigeorgiou, A.G. (2012). DIANA miRPath v.2.0: investigating the combinatorial effect of microRNAs in pathways. Nucleic Acids Res *40*, W498-504.

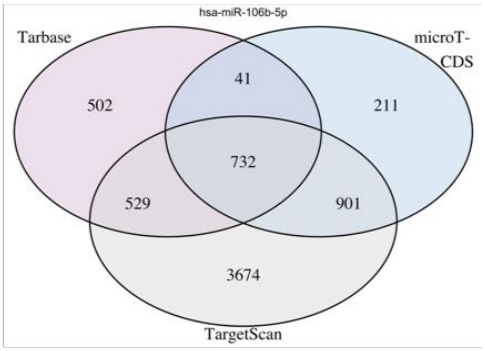
Vlachos, I.S., Paraskevopoulou, M.D., Karagkouni, D., Georgakilas, G., Vergoulis, T., Kanellos, I., Anastasopoulos, I.L., Maniou, S., Karathanou, K., Kalfakakou, D.*, et al.* (2015a). DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. Nucleic Acids Res *43*, D153-159.
Vlachos, I.S., Zagganas, K., Paraskevopoulou, M.D., Georgakilas, G., Karagkouni, D., Vergoulis, T., Dalamagas, T., and Hatzigeorgiou, A.G. (2015b). DIANA-miRPath v3.0: deciphering microRNA function with experimental support. Nucleic acids research *43*, W460-466.
Vychytilova-Faltejskova, P., Pesta, M., Radova, L., Liska, V., Daum, O., Kala, Z., Svoboda, M., Kiss, I., and Slaby, O. (2016). Genome-wide microRNA Expression Profiling in Primary Tumors and Matched Liver Metastasis of Patients with Colorectal Cancer. Cancer Genomics Proteomics *13*, 311-316.
Wagner, G.P., Kin, K., and Lynch, V.J. (2013). A model based criterion for gene expression calls using RNA-seq data. Theory Biosci *132*, 159-164.
Wakatsuki, S., Furuno, A., Ohshima, M., and Araki, T. (2015). Oxidative stress-dependent phosphorylation activates ZNRF1 to induce neuronal/axonal degeneration. J Cell Biol *211*, 881-896.
Wan, L., Chen, M., Cao, J., Dai, X., Yin, Q., Zhang, J., Song, S.J., Lu, Y., Liu, J., Inuzuka, H.*, et al.* (2017). The APC/C E3 Ligase Complex Activator FZR1 Restricts BRAF Oncogenic Function. Cancer Discov *7*, 424-441.
Wang, X. (2016). Improving microRNA target prediction by modeling with unambiguously identified microRNA-target pairs from CLIP-ligation studies. Bioinformatics (Oxford, England) *32*, 1316-1322.
Wang, Y.P., and Li, K.B. (2009). Correlation of expression profiles between microRNAs and mRNA targets using NCI-60 data. BMC Genomics *10*, 218.
Wickham, H. (2011). The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software *40*, 1-29.
Wilczynska, A., and Bushell, M. (2015). The complexity of miRNA-mediated repression. Cell Death Differ *22*, 22-33.
Winkle, C.C., Olsen, R.H., Kim, H., Moy, S.S., Song, J., and Gupton, S.L. (2016). Trim9 Deletion Alters the Morphogenesis of Developing and Adult-Born Hippocampal Neurons and Impairs Spatial Learning and Memory. J Neurosci *36*, 4940-4958.
Wong, N., and Wang, X. (2015). miRDB: an online resource for microRNA target prediction and functional annotations. Nucleic acids research *43*, D146-152.
Wu, S., Huang, S., Ding, J., Zhao, Y., Liang, L., Liu, T., Zhan, R., and He, X. (2010). Multiple microRNAs modulate p21Cip1/Waf1 expression by directly targeting its 3' untranslated region. Oncogene *29*, 2302-2308.
Wu, W.-S., Tu, B.-W., Chen, T.-T., Hou, S.-W., and Tseng, J.T. (2017). CSmiRTar: Condition-Specific microRNA targets database. PloS one *12*, e0181231.
Wu, Z., Zan, P., Li, S., Liu, J., Wang, J., Chen, D., Wang, H., Qian, Y., Luo, L., and Huang, X. (2015). Knockdown of WWP1 inhibits growth and invasion, but induces apoptosis of osteosarcoma cells. Int J Clin Exp Pathol *8*, 7869-7877.
Xiao, F., Zuo, Z., Cai, G., Kang, S., Gao, X., and Li, T. (2009). miRecords: an integrated resource for microRNA-target interactions. Nucleic Acids Res *37*, D105-110.
Xie, C., Powell, C., Yao, M., Wu, J., and Dong, Q. (2014). Ubiquitin-conjugating enzyme E2C: a potential cancer biomarker. Int J Biochem Cell Biol *47*, 113-117.

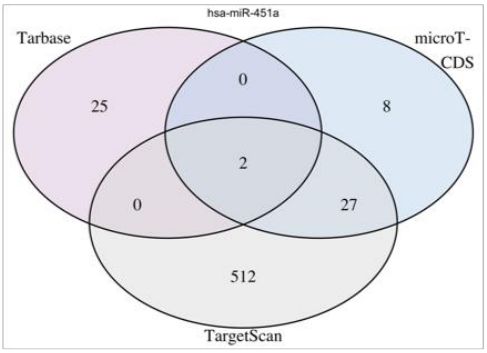
Xie, J.X., Fan, X., Drummond, C.A., Majumder, R., Xie, Y., Chen, T., Liu, L., Haller, S.T., Brewster, P.S., Dworkin, L.D.*, et al.* (2017). MicroRNA profiling in kidney disease: Plasma versus plasma-derived exosomes. Gene *627*, 1-8.
Xu, J., Cao, Z., Liu, W., You, L., Zhou, L., Wang, C., Lou, W., Sun, B., Miao, Y., Liu, X.*, et al.* (2016). Plasma miRNAs Effectively Distinguish Patients With Pancreatic Cancer From Controls: A Multicenter Study. Annals of surgery *263*, 1173-1179.
Yan, Y., Hein, A.L., Etekpo, A., Burchett, K.M., Lin, C., Enke, C.A., Batra, S.K., Cowan, K.H., and Ouellette, M.M. (2014). Inhibition of RAC1 GTPase sensitizes pancreatic cancer cells to gamma-irradiation. Oncotarget *5*, 10251-10270.
Yumimoto, K., Akiyoshi, S., Ueo, H., Sagara, Y., Onoyama, I., Ueo, H., Ohno, S., Mori, M., Mimori, K., and Nakayama, K.I. (2015). F-box protein FBXW7 inhibits cancer metastasis in a non-cell-autonomous manner. J Clin Invest *125*, 621-635.
Zhang, J., and Storey, K.B. (2018). RBiomirGS: an all-in-one miRNA gene set analysis solution featuring target mRNA mapping and expression profile integration. PeerJ *6*, e4262.
Zhang, T.T., Sun, Y., Jia, C.W., Yu, S.N., Lu, Z.H., and Chen, J. (2013). [Target prediction and verification of miR-27a in pancreatic cancer]. Zhonghua Bing Li Xue Za Zhi *42*, 392-396.
Zhang, W., Zang, J., Jing, X., Sun, Z., Yan, W., Yang, D., Shen, B., and Guo, F. (2014). Identification of candidate miRNA biomarkers from miRNA regulatory network with application to prostate cancer. J Transl Med *12*, 66.
Zhang, Z., Yin, J., Yang, J., Shen, W., Zhang, C., Mou, W., Luo, J., Yan, H., Sun, P., Luo, Y.*, et al.* (2016). miR-885-5p suppresses hepatocellular carcinoma metastasis and inhibits Wnt/beta-catenin signaling pathway. Oncotarget *7*, 75038-75051.
Zhao, J., Mialki, R.K., Wei, J., Coon, T.A., Zou, C., Chen, B.B., Mallampalli, R.K., and Zhao, Y. (2013). SCF E3 ligase F-box protein complex SCF(FBXL19) regulates cell migration by mediating Rac1 ubiquitination and degradation. FASEB J *27*, 2611-2619.
Zhou, Z., Cheng, Y., Jiang, Y., Liu, S., Zhang, M., Liu, J., and Zhao, Q. (2018). Ten hub genes associated with progression and prognosis of pancreatic carcinoma identified by co-expression analysis. Int J Biol Sci *14*, 124-136.
Zhu, Y., Qiu, P., and Ji, Y. (2014). TCGA-assembler: open-source software for retrieving and processing TCGA data. Nat Methods *11*, 599-600.

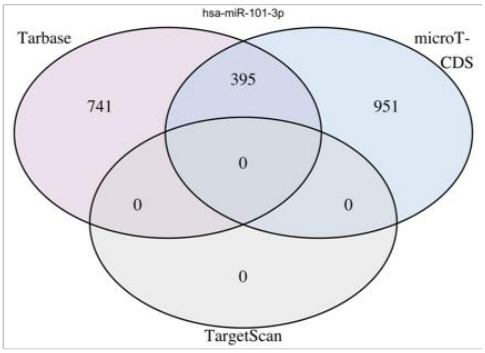
Supplementary information

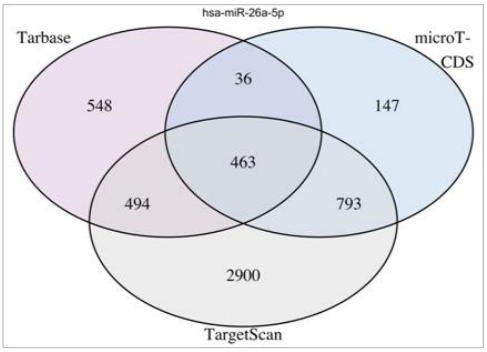
1. Figures & Tables

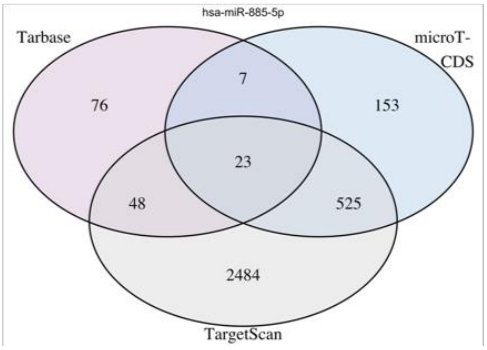


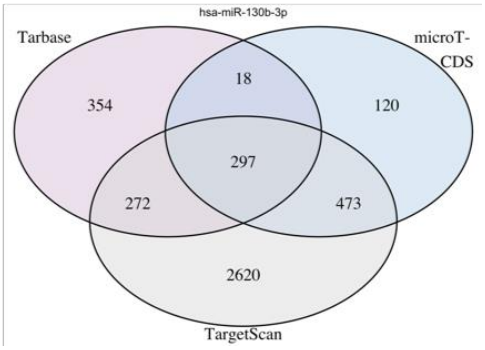












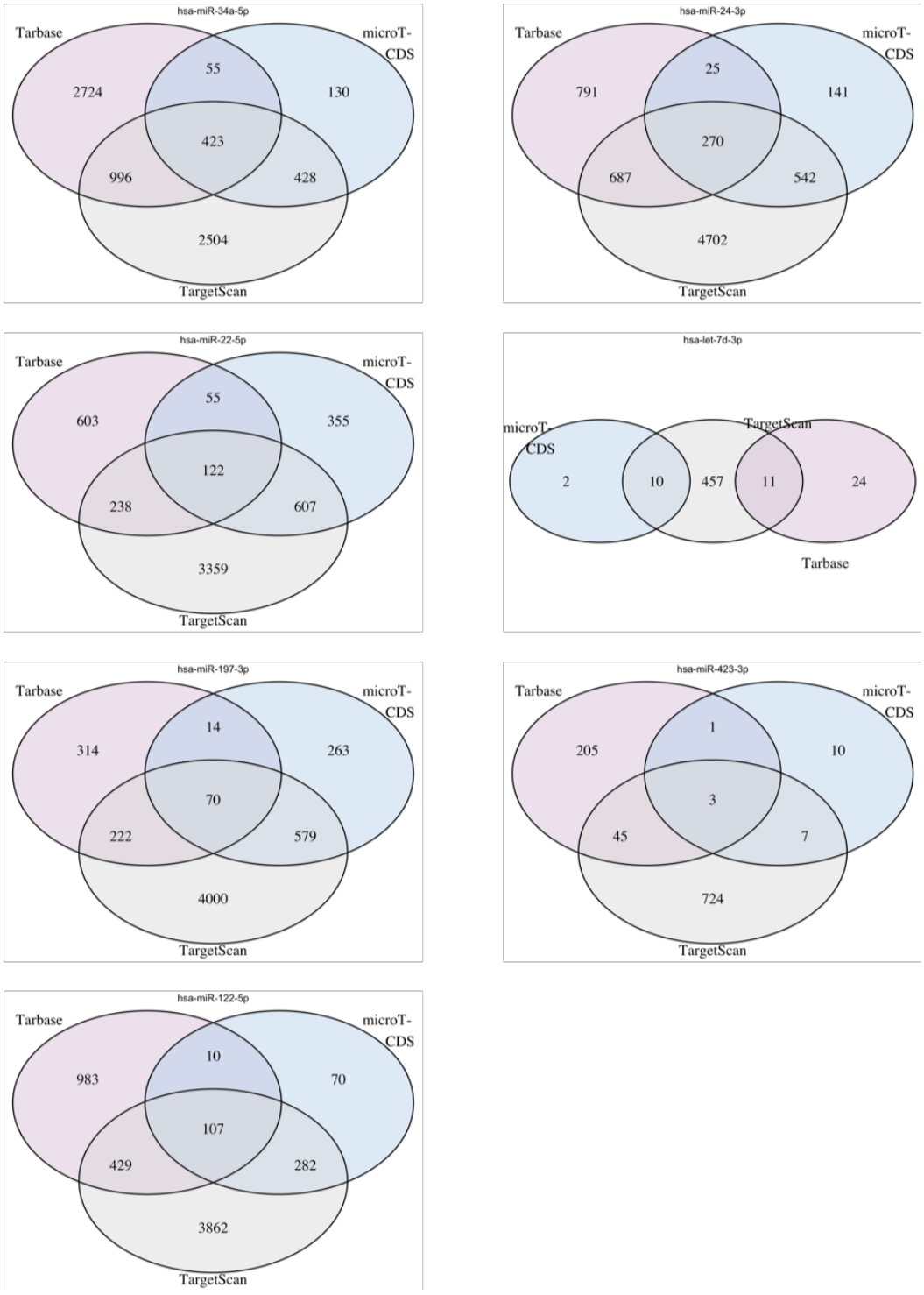


Figure S1. Venn diagram of DIANA-TarBase v7, DIANA-microT-CDS and TargetScan v7.1 for each miRNA. The R package VennDiagram was used to generate the overlap of identified miRNA target genes. DIANA-TarBase v7 = Tarbase (pink), DIANA-microT-CDS = microT-CDS (blue), TargetScan v.7.1 = TargetScan (grey).

Table S1. Top 10 strongest negative correlations for hsa-miR-885-5p.

miRNA	Gene	PCC	P_value	adj_P_value_PCC
hsa-miR-885-5p	MPZL1	-0.88	9E-12	3E-08
hsa-miR-885-5p	VANGL1	-0.88	1E-11	3E-08
hsa-miR-885-5p	KIAA1217	-0.88	1E-11	3E-08
hsa-miR-885-5p	AJUBA	-0.87	8E-11	1E-07
hsa-miR-885-5p	ALPK1	-0.86	1E-10	1E-07
hsa-miR-885-5p	LAMC2	-0.86	1E-10	2E-07
hsa-miR-885-5p	GPRC5A	-0.85	3E-10	3E-07
hsa-miR-885-5p	EREG	-0.85	4E-10	3E-07
hsa-miR-885-5p	PROSER2	-0.85	6E-10	4E-07
hsa-miR-885-5p	SYTL4	-0.84	9E-10	6E-07

Table S2. Top 10 strongest positive correlations for hsa-miR-885-5p.

miRNA	Gene	PCC	P_value	adj_P_value_PCC
hsa-miR-885-5p	GNAS	0.85	3E-10	3E-07
hsa-miR-885-5p	GIN1	0.84	6E-10	4E-07
hsa-miR-885-5p	NECAP1	0.82	6E-09	2E-06
hsa-miR-885-5p	CRADD	0.82	6E-09	2E-06
hsa-miR-885-5p	MCMDC2	0.81	9E-09	3E-06
hsa-miR-885-5p	NAPB	0.80	2E-08	5E-06
hsa-miR-885-5p	BTBD3	0.80	2E-08	5E-06
hsa-miR-885-5p	GDAP1	0.80	2E-08	6E-06
hsa-miR-885-5p	NDRG3	0.79	5E-08	1E-05
hsa-miR-885-5p	ARG2	0.79	5E-08	1E-05

2. Programming code

2.1. SQL databases

Three structured query language (SQL) databases were generated. One for miRNA target prediction (mirna_database.sqlite), one for miRNA and mRNA expression levels (miRNAmRNACor.sqlite) and one for miRNA and protein expression levels (miRNAProteinCor.sqlite). Before creating databases, pre-processing steps were performed as described.

Pre-processing of miRNA databases

###Explanation to 'regex.pl':
#The original 2 first lines of the DIANA-microT-CDS looked like the following:

TranscriptId,GeneID(gene_name) ,mirna(mirbase_v),miTG_score
F52F10.2,F52F10.2(F52F10.2),cel-miR-62(18),0.488

#After extraction of only human miRNAs it looked like the following:

ENST00000535623,ENSG00000128595(CALU),hsa-let-7c(18),0.637
ENST00000378078,ENSG00000107185(RGP1),hsa-let-7c(18),0.530

#In order to get rid of the (18) adjacent to the miRNA name, regex.pl was used for
#translating '(' and ')' to ',' to get them separated and the same was applied to the new
#column-name line to separate them accordingly. The ouput after this step looked like this:

TranscriptId,GeneID,gene_name, ,mirna,mirbase_v,,miTG_score
ENST00000535623,ENSG00000128595,CALU,,hsa-let-7c,18,,0.637

Perl script: regex.pl

```
#!/usr/bin/perl

use warnings;

#Open .csv file & if not able to open an error message
#is shown and the program is terminated
if(!open(HANDLE,"human_microT.csv")){
        die "Could not open file\n";
}

#Open new file to write to
if(!open(WRITE, ">DIANA_microT.csv")){
        die "Could not create file\n";
}

print (WRITE "TranscriptId,GeneID,gene_name, ,mirna,mirbase_v,,miTG_score\n");
```

```
#All file lines are read and concatenated into a single
#string called $content
$content=join("",<HANDLE>);

#Regex to
$content =~ tr/\(\)/,,/;
print (WRITE $content);

#Dissociate filehandles
close(HANDLE);
close(WRITE);
```

Bash terminal: Pre-processing steps for miRNA target prediction database

```
#Change colnames of Targetscan conserved & nonconserved site tables
sed -i -e '1s/Gene Symbol/Gene.Symbol/g' Conserved_Site_Context_Scores.txt
sed -i -e '1s/Gene Symbol/Gene.Symbol/g' Nonconserved_Site_Context_Scores.txt


#For DIANA-microT-CDS, only human miRNAs were extracted
grep "hsa-" microT_CDS_data.csv > DIANA_microT.csv

#Since ‘hsa' is not in the colnames, a new colname-line was printed into the file using Perl
script: regex.pl as described above
```

Bash terminal: Create miRNA target prediction SQLite database

```
#Start sqlite3
sqlite3

.mode tabs

.import
/Users/emmyborgmastars/Documents/Rstats/databases/Conserved_Site_Context_Scores.tx
t TargetScan_conserved_site
.import
/Users/emmyborgmastars/Documents/Rstats/databases/Nonconserved_Site_Context_Scor
es.txt TargetScan_nonconserved_site

.mode csv

.import /Users/emmyborgmastars/Documents/Rstats/databases/DIANA_microT.csv
microT_CDS
```

```
.import
/Users/emmyborgmastars/Documents/Rstats/databases/targetscan_default_pred.csv
TargetScan_conserved

.separator “\t”

.import /Users/emmyborgmastars/Documents/Rstats/databases/tarbase_data.csv Tarbase

###Create new database
.save /Users/emmyborgmastars/Documents/Rstats/databases/mirna_database.sqlite

.quit
```

Pre-processing for miRNA, mRNA and protein correlations

Script: 180320_mRNA_miRNA_harm.R

```
##Also contains miRNA-protein ‘harmonization’

setwd("/Users/emmyborgmastars/Documents/Rstats/databases")

#See difference between colnames miRNA and mRNA

mRNA<-read.table("TCGA-PAAD.htseq_fpkm-uq.tsv",sep="\t",header=T)
a<-colnames(mRNA)

miRNA<-read.table("all_mirnas.csv",sep=";",header=T)
miRNA<-miRNA[,-1]
miRNA<-miRNA[,-1]
b<-colnames(miRNA)
head(b)

setdiff(a,b)
setdiff(b,a)

#Remove "TCGA.IB.7647.01A" from miRNA and then order both tables by colnames
miRNA<-subset(miRNA, select=-c(TCGA.IB.7647.01A))

ny_miRNA<-miRNA[,order(colnames(miRNA))]
ny_mRNA<-mRNA[,order(colnames(mRNA))]

write.csv(ny_miRNA$miRNA_region,"translatemiRNAs.csv")

ensemblids<-mRNA$Ensembl_ID

#Add miRNA_names
setwd("/Users/emmyborgmastars/Documents/Perlstats")
```

```
name<-read.table("180323_MIMA_miRNAnames.csv",sep=";",header=T)

new_miRNA<-cbind(ny_miRNA,data.frame(miRNA_name=name$miRNA_name))
w_miRNA<-new_miRNA[,-1]
w_miRNA<-w_miRNA[,c(ncol(w_miRNA),1:(ncol(w_miRNA)-1))]

#Move the last column to first in miRNA
ny_miRNA<-ny_miRNA[,c(ncol(ny_miRNA),1:(ncol(ny_miRNA)-1))]

setwd("/Users/emmyborgmastars/Documents/Rstats/databases/for_correlation")
#Write tables
write.table(w_miRNA,"180323_miRNA_formRNA.txt",sep="\t",row.names = F)
write.table(ny_mRNA,"180322_mRNA.txt",sep="\t")

#Convert ensembl_ID to hgnc_symbol
library(org.Hs.eg.db)

prot_to_symbol <- function(ids){
  prot_ids_list <-  mapIds(org.Hs.eg.db,
                    keys=ids,
                    column="SYMBOL",
                    keytype="ENSEMBL",
                    multiVals="list")
  prot_ids_unlisted <- lapply(X = 1:length(prot_ids_list), FUN = function(x){
    cbind(names(prot_ids_list[x]), unlist(prot_ids_list[x], use.names = F))
  })

  prot_ids_final <- base::Reduce(f = rbind, x = prot_ids_unlisted)

  colnames(prot_ids_final) <- c("ensembl_gene_id", "hgnc_symbol")

  write.table(prot_ids_final,"prot_ids_final.txt",sep="\t")
}

prot_to_symbol(ids=ensembl)

#MERGE
setwd("/Users/emmyborgmastars/Documents/Perlstats")
new_ensembl<-read.csv("180322_ensembl.csv",header=T)
new_table<-read.csv("180322_mRNA.txt",header=T,sep="\t")
colnames(new_ensembl)<-"Ensembl_ID"
vector<-as.vector(new_ensembl$Ensembl_ID)
new_table<<-cbind(new_table,data.frame(Ensembl_ID=vector))
new<-new_table[,-1]
new<-new[,c(ncol(new),1:(ncol(new)-1))]

hgnc<-read.table("prot_ids_final.txt",sep="\t",header=T)
```

```
colnames(hgnc)<-c("Ensembl_ID","hgnc_symbol")
final_table<-merge(hgnc,new,all=TRUE,by="Ensembl_ID")
final<-final_table[,-1]
setwd("/Users/emmyborgmastars/Documents/Rstats/databases/for_correlation")
write.table(final,"180323_mRNA.txt",sep="\t",row.names = F)

#See difference between colnames miRNA and protein
setwd("/Users/emmyborgmastars/Documents/Rstats/databases/for_correlation")
protein<-read.table("protein_to_setdiff.csv",sep=";",header=T)
protein<-protein[,-c(2,3,4)]
t<-t(protein)
write.table(t,"180322_t_protein.csv",sep=";")
protein<-read.table("180322_t_protein.csv",sep=";",header=T)
a<-colnames(protein)

#miRNA for protein correlation
setwd("/Users/emmyborgmastars/Documents/Rstats/databases")
miRNA<-read.table("all_mirnas.csv",sep=";",header=T)
miRNA<-miRNA[,-c(1,2)]
b<-colnames(miRNA)

setdiff(a,b)
setdiff(b,a)

diff_v<-setdiff(b,a)
head(diff_v)
diff_tcga<-diff_v[-1]
dif_prot<-setdiff(a,b)
dif_prot<-dif_prot[-1]

tabled_cols <-table(c(colnames(protein), colnames(miRNA)))
cols_to_keep <- names(tabled_cols)[tabled_cols == 2]

mirna_new<-subset(miRNA, select=cols_to_keep)
protein_new<-subset(protein,select=cols_to_keep)

mirna_new<-mirna_new[,order(colnames(mirna_new))]
protein_new<-protein_new[,order(colnames(protein_new))]

protein_final<-cbind(protein_name,protein_new)

#Remove from miRNA and then order both tables by colnames
miRNA<-subset(miRNA, select=-(diff_tcga))

ny_miRNA<-miRNA[,order(colnames(miRNA))]
ny_mRNA<-mRNA[,order(colnames(mRNA))]
```

```
#Add miRNA_names
setwd("/Users/emmyborgmastars/Documents/Rstats/databases")
name<-read.table("180323_MIMA_miRNAnames.csv",sep=";",header=T)
mirna_name<-as.data.frame(name$miRNA_name)
colnames(mirna_name)<-"mirna_name"
mirna_final<-cbind(mirna_name,mirna_new)


setwd("/Users/emmyborgmastars/Documents/Rstats/databases/for_correlation")
#Write tables
write.table(mirna_final,"180328_miRNA_forprotein.txt",sep="\t",row.names = F)
write.table(protein_final,"180328_protein.txt",sep="\t",row.names = F)
```

Bash terminal: Create miRNA-mRNA correlation SQLite database

```
sqlite3
.separator ”\t”
.import
/Users/emmyborgmastars/Documents/Rstats/databases/for_correlation/180323_mRNA.txt
mRNA
.import
/Users/emmyborgmastars/Documents/Rstats/databases/for_correlation/180323_miRNA_fo
rmRNA.txt miRNA
.save /Users/emmyborgmastars/Documents/Rstats/databases/miRNAmRNACor.sqlite
.quit
```

Bash terminal: Create miRNA-protein correlation SQLite database

```
sqlite3
.separator ”\t”
.import
/Users/emmyborgmastars/Documents/Rstats/databases/for_correlation/180328_protein.tx
t Protein
.import
/Users/emmyborgmastars/Documents/Rstats/databases/for_correlation/180328_miRNA_fo
rprotein.txt miRNA
.save /Users/emmyborgmastars/Documents/Rstats/databases/miRNAProteinCor.sqlite
.quit
```

2.2. Function for miRNA target prediction, functional enrichment & correlation analyses

Script: mirna_search.R

```
#Load libraries
#Required packages: RSQLite,VennDiagram,gridExtra
library(DBI)
library(RSQLite)
library(VennDiagram)
library(gridExtra)
library(edgeR)
library(GO.db)
library(org.Hs.eg.db)
setwd("/Users/emmyborgmastars/Documents/Rstats/databases")

#Create a data frame number of genes
number_of_genes<-data.frame(matrix(ncol=4,nrow=0))
colnames(number_of_genes)<-c("miRNA_ID","Tarbase","microT_CDS","TargetScan")

#Create data frame for PCC
PCC_table<-data.frame(matrix(ncol=4,nrow=0))
colnames(PCC_table)<-c("miRNA","Gene","PCC","P_value")

correlate_mRNA<-function(gene, miRNA){
setwd("/Users/emmyborgmastars/Documents/Rstats/databases")
mRNA_con <- dbConnect(SQLite(),'miRNAmRNACor.sqlite')
ex_mRNA<-dbGetQuery(mRNA_con,
            paste0('SELECT * FROM mRNA WHERE hgnc_symbol IS"',   gene, '"'))
v_mRNA<-suppressWarnings(as.numeric(ex_mRNA[1,]))

PCC_value<-
suppressWarnings(cor(v_miRNA,v_mRNA,method="pearson",use="na.or.complete"))

if (!is.na(PCC_value)){
pval<-suppressWarnings(cor.test(v_miRNA,v_mRNA,method="pearson")$p.value)

#Add number of genes information to data frame
PCC_table<-rbind(PCC_table,
            data.frame(miRNA=miRNA,Gene=gene,PCC=PCC_value,P_value=pval))
}
return(PCC_table)
dbDisconnect(mRNA_con)
}

#Create data frame for protein
PCC_p_table<-data.frame(matrix(ncol=4,nrow=0))
colnames(PCC_p_table)<-c("miRNA","Protein","PCC","P_value")
```

```
correlate_prot<-function(prot,miRNA){
 setwd("/Users/emmyborgmastars/Documents/Rstats/databases")
 prot_con <- dbConnect(SQLite(),'miRNAProteinCor.sqlite')
 ex_prot<-dbGetQuery(prot_con,
              paste0('SELECT * FROM Protein WHERE protein_name IS"',   prot, '"'))

 v_prot<-suppressWarnings(as.numeric(ex_prot[1,]))

 PCC_value<-
suppressWarnings(cor(p_miRNA,v_prot,method="pearson",use="na.or.complete"))

 if (!is.na(PCC_value)){
 p_val_s<-suppressWarnings(cor.test(p_miRNA,v_prot,method="pearson")$p.value)

 #Add number of genes information to data frame
 PCC_p_table<-rbind(PCC_p_table,
             data.frame(miRNA=miRNA,Protein=prot,PCC=PCC_value,P_value=p_val_s))
 }
 return(PCC_p_table)
 dbDisconnect(prot_con)
}

mirna_search<-function(miRNA){

 #Connect to mirna_database
 con <- dbConnect(SQLite(),'mirna_database.sqlite')

 ##Search query mirna

 #Tarbase
 tarbase<-dbGetQuery(con, paste0('SELECT * FROM Tarbase WHERE mirna IS"',   miRNA, '"'))
 tarbase_g<-unique(tarbase$geneName)
 n_tarbase<-length(tarbase_g)

 #microT_CDS
 microT_CDS<-dbGetQuery(con, paste0('SELECT * FROM microT_CDS WHERE mirna IS"',
miRNA, '"','AND miTG_score>=0.7'))
 microT_g<-unique(microT_CDS$gene_name)
 n_microT<-length(microT_g)

 ###TargetScan
 #Conserved miRNA family
 targetscan1<-dbGetQuery(con, paste0('SELECT * FROM TargetScan_conserved WHERE
miRNA IS"',   miRNA, '"'))
 targetscan1_g<-unique(targetscan1$Gene.Symbol)
```

```
 #Conserved site
 targetscan2<-dbGetQuery(con, paste0('SELECT * FROM TargetScan_conserved_site WHERE
miRNA IS"',   miRNA, '"'))
 targetscan2_g<-unique(targetscan2$Gene.Symbol)

 #Nonconserved site
 targetscan3<-dbGetQuery(con, paste0('SELECT * FROM TargetScan_nonconserved_site
WHERE miRNA IS"',   miRNA, '"'))
 targetscan3_g<-unique(targetscan3$Gene.Symbol)

 dbDisconnect(con)

 targetscan_g<-union(targetscan1_g,union(targetscan2_g,targetscan3_g))
 n_targetscan<-length(targetscan_g)
 setwd("/Users/emmyborgmastars/Documents/Rstats/databases/DATA.DIR")
 write.table(targetscan_g,paste0('Target_genes/TargetScan_',miRNA,'.txt'))
 write.table(tarbase_g,paste0('Target_genes/Tarbase_',miRNA,'.txt'))
 write.table(microT_g,paste0('Target_genes/microT_CDS_',miRNA,'.txt'))

 ####Create Venn diagram

 #Calculate overlap
 overlap12 <- calculate.overlap(
  x = list(
   "Tarbase" = tarbase_g,
   "microT.CDS" = microT_g
  )
 );

 overlap23 <- calculate.overlap(
  x = list(
   "microT.CDS" = microT_g,
   "TargetScan" = targetscan_g
  )
 );

 overlap13 <- calculate.overlap(
  x = list(
   "Tarbase" = tarbase_g,
   "TargetScan" = targetscan_g
  )
 );

 overlap123 <- calculate.overlap(
  x = list(
   "Tarbase" = tarbase_g,
   "microT.CDS" = microT_g,
```

```
    "TargetScan" = targetscan_g
  )
);

area1<-length(tarbase_g)
area2<-length(microT_g)
area3<-length(targetscan_g)
n12<-length(overlap12$a3)
n23<-length(overlap23$a3)
n13<-length(overlap13$a3)
n123<-length(overlap123$a5)

#create venndiagram with three sets

color=c("thistle","slategray2","gainsboro")
size=1.5
grid.newpage()
g=draw.triple.venn(area1, area2, area3, n12, n23, n13, n123,
                category = c("Tarbase","microT-\n\t\t\t\t\t\tCDS","TargetScan"),
                fill=color,cat.cex=rep(size,3),cex=rep(size,7))

tiff(width = 7, height = 5, file=paste0('Venndiagrams/',miRNA,'.tiff'), units="in", res=600)
grid.arrange(gTree(children=g), top=paste0(miRNA))
dev.off()


#Add number of genes information to data frame
number_of_genes<-rbind(number_of_genes,
               data.frame(miRNA_ID=miRNA,Tarbase=n_tarbase,
               microT_CDS=n_microT,TargetScan=n_targetscan))

#Union of target genes
uni<-union(tarbase_g,union(microT_g,targetscan_g))

###Functional enrichment before evaluation

##Convert hgnc symbols to entrez gene id
deID<-as.numeric(unlist(mapIds(org.Hs.eg.db,keys=uni,
                      column="ENTREZID",keytype="SYMBOL",multiVals="list")))

#GO terms
go<-goana(deID, FDR = 0.05, trend = FALSE, species="Hs")
sign_GO<-subset(go,P.DE<=0.05,select=c(Term,Ont,N,DE,P.DE))
GO_terms<-subset(sign_GO,DE>=5,select=c(Term,Ont,N,DE,P.DE))
attach(GO_terms)
write.table(GO_terms[order(-
DE),],paste0('Functional_enrichment/Before_evaluation/GO_',miRNA,'.txt'),
```

```
          sep=";")
 detach(GO_terms)

 #KEGG pathways
 kegg<-kegga(deID, FDR = 0.05, trend = FALSE, species="Hs")
 sign_kegg<-subset(kegg,P.DE<=0.05,select=c(Pathway,N,DE,P.DE))
 kegg_paths<-subset(sign_kegg,DE>=5,select=c(Pathway,N,DE,P.DE))
 attach(kegg_paths)
 write.table(kegg_paths[order(-
DE),],paste0('Functional_enrichment/Before_evaluation/KEGG_',miRNA,'.txt'),
          sep=";")
 detach(kegg_paths)

 ###Correlation analyses
 setwd("/Users/emmyborgmastars/Documents/Rstats/databases")

 #proteins
 #Create table
 prot_con <- dbConnect(SQLite(),'miRNAProteinCor.sqlite')
 e_miRNA<-dbGetQuery(prot_con, paste0('SELECT * FROM miRNA WHERE mirna_name IS"',
miRNA, '\n"'))
 p_miRNA<<-suppressWarnings(as.numeric(e_miRNA[1,]))
 cor_prot_list<-lapply(X=uni,FUN=correlate_prot,miRNA=miRNA)
 dbDisconnect(prot_con)

 #Multiple testing
 prot_df<-do.call(rbind.data.frame, cor_prot_list)
 adj_P_s<-p.adjust(prot_df$P_value, method="BH")
 prot_df<-cbind(prot_df,adj_P_s)
 temp_sign_prot<-subset(prot_df, adj_P_s < 0.05,
select=c(miRNA,Protein,PCC,P_value,adj_P_s))
 pos_temp_sign_prot<-subset(temp_sign_prot, PCC >= 0.3,
select=c(miRNA,Protein,PCC,P_value,adj_P_s))
 neg_temp_sign_prot<-subset(temp_sign_prot, PCC <= -0.3,
select=c(miRNA,Protein,PCC,P_value,adj_P_s))
 proteins<-
union(as.vector(pos_temp_sign_prot$Protein),as.vector(neg_temp_sign_prot$Protein))

 #mRNAs
 #Create table
 mRNA_con <- dbConnect(SQLite(),'miRNAmRNACor.sqlite')
 ex_miRNA<-dbGetQuery(mRNA_con, paste0('SELECT * FROM miRNA WHERE miRNA_name
IS"',   miRNA, '\n"'))
 v_miRNA<<-suppressWarnings(as.numeric(ex_miRNA[1,]))
 cor_mRNA_list<-lapply(X=uni,FUN=correlate_mRNA,miRNA=miRNA)
 dbDisconnect(mRNA_con)
```

```
 #Multiple testing
 mRNA_df<-do.call(rbind.data.frame, cor_mRNA_list)
 adj_P_p<-p.adjust(mRNA_df$P_value, method="BH")
 mRNA_df<-cbind(mRNA_df,adj_P_p)
 temp_sign_mRNA<-subset(mRNA_df, adj_P_p < 0.05,
select=c(miRNA,Gene,PCC,P_value,adj_P_p))
 pos_temp_sign_mRNA<-subset(temp_sign_mRNA, PCC >= 0.3,
select=c(miRNA,Gene,PCC,P_value,adj_P_p))
 neg_temp_sign_mRNA<-subset(temp_sign_mRNA, PCC <= -0.3,
select=c(miRNA,Gene,PCC,P_value,adj_P_p))
 mRNAs<-
union(as.vector(pos_temp_sign_mRNA$Gene),as.vector(neg_temp_sign_mRNA$Gene))

 uni2<-union(mRNAs,proteins)

 if(length(uni2)>0) {

 ###Functional enrichment after evaluation - union of mRNA and protein
 deID<-as.numeric(unlist(mapIds(org.Hs.eg.db,keys=uni2,
                     column="ENTREZID",keytype="SYMBOL",multiVals="list")))

 setwd("/Users/emmyborgmastars/Documents/Rstats/databases/DATA.DIR")

 #GO
 go<-goana(deID, FDR = 0.05, trend = FALSE, species="Hs")
 sign_GO<-subset(go,P.DE<=0.05,select=c(Term,Ont,N,DE,P.DE))
 GO_terms<-subset(sign_GO,DE>=5,select=c(Term,Ont,N,DE,P.DE))
 attach(GO_terms)
 write.table(GO_terms[order(-
DE),],paste0('Functional_enrichment/After_evaluation/GO_',miRNA,'.txt'),
         sep=";")
 detach(GO_terms)

 #KEGG
 kegg<-kegga(deID, FDR = 0.05, trend = FALSE, species="Hs")
 sign_kegg<-subset(kegg,P.DE<=0.05,select=c(Pathway,N,DE,P.DE))
 kegg_paths<-subset(sign_kegg,DE>=5,select=c(Pathway,N,DE,P.DE))
 attach(kegg_paths)
 write.table(kegg_paths[order(-
DE),],paste0('Functional_enrichment/After_evaluation/KEGG_',miRNA,'.txt'),
         sep=";")
 detach(kegg_paths)
 }
 setwd("/Users/emmyborgmastars/Documents/Rstats/databases")

return(list("number_of_genes"=number_of_genes,"pos_temp_sign_mRNA"=pos_temp_sign
_mRNA,
```

```
"neg_temp_sign_mRNA"=neg_temp_sign_mRNA,"pos_temp_sign_prot"=pos_temp_sign_pr
ot,
          "neg_temp_sign_prot"=neg_temp_sign_prot))
}
```

2.3. Run function as described in '2.2. miRNA target prediction, functional enrichment & correlation analyses' for 15 miRNAs

Script: masterpipeline.R

```
setwd("/Users/emmyborgmastars/Documents/Rstats/databases")

#Run the mirna search function
source("mirna_search.R")

#Import list of mirnas
mirnas<-read.csv("mirna_list.txt",header=F)

#Start time
start_time<-Sys.time()

#loop searches with each mirna using test function, margin = 1 (row), 2 (col)
whole_pipeline<-apply(X = mirnas, MARGIN = 1, FUN = mirna_search)

#Stop time
end_time<-Sys.time()
end_time-start_time

#Import gdc samplesheet - only for miRNA annotation 3p and 5p
gdc_samples<-read.csv("gdc_sample_sheet_2018-02-13-03-23-11.tsv.2018-02-
13T14_23_27.088749.tsv",sep="\t",header=T)

setwd("/Users/emmyborgmastars/Documents/Rstats/databases/DATA.DIR/Significant_corr
elations")

#Add number of miRNAs to 'vector'
vector<-c(1:15)

number_genes<-c()

#Create summary table with number of target genes
for (i in vector) {
  number_genes<-rbind(number_genes,whole_pipeline[[i]]$number_of_genes)
}

#All miRNA combined significant correlations

pos_mRNA<-c()
neg_mRNA<-c()
neg_prot<-c()
pos_prot<-c()
```

```
for (i in vector) {
  pos_mRNA<-rbind(pos_mRNA,whole_pipeline[[i]]$pos_temp_sign_mRNA)
}

for (i in vector) {
  neg_mRNA<-rbind(neg_mRNA,whole_pipeline[[i]]$neg_temp_sign_mRNA)
}

neg_prot<-c()
for (i in vector) {
  pos_prot<-rbind(pos_prot,whole_pipeline[[i]]$pos_temp_sign_prot)
}

for (i in vector) {
  neg_prot<-rbind(neg_prot,whole_pipeline[[i]]$neg_temp_sign_prot)
}

#Write to tables
#Save number of genes
setwd("/Users/emmyborgmastars/Documents/Rstats/databases/DATA.DIR")
write.table(number_genes,"number_of_genes.txt",sep="\t")
write.table(pos_mRNA,"pos_mRNA.txt",sep="\t")
write.table(neg_mRNA,"neg_mRNA.txt",sep="\t")
write.table(pos_prot,"pos_prot.txt",sep="\t")
write.table(neg_prot,"neg_prot.txt",sep="\t")

################################################################################
#180324
#Check the correlations
setwd("/Users/emmyborgmastars/Documents/Rstats/databases")
par(mfrow=c(2,2))
check_cor<-function(X){
  gene<-X[2]
  miRNA<-X[1]

  mRNA_con <- dbConnect(SQLite(),'miRNAmRNACor.sqlite')
  ex_miRNA<-dbGetQuery(mRNA_con, paste0('SELECT * FROM miRNA WHERE miRNA_name
IS"',   miRNA, '\n"'))
  v_miRNA<<-as.numeric(ex_miRNA[1,])

  ex_mRNA<-dbGetQuery(mRNA_con,
                      paste0('SELECT * FROM mRNA WHERE hgnc_symbol IS"',   gene, '"'))
  v_mRNA<-as.numeric(ex_mRNA[1,])

  PCC_value<-cor(v_miRNA,v_mRNA,method="pearson",use="na.or.complete")
  plot(v_miRNA,v_mRNA,xlab=paste0(miRNA),ylab=paste0(gene),main=paste0('PCC:
',round(PCC_value, digits = 3)),
```

```
    cex=0.5,pch=1)
 abline(lm(v_mRNA~v_miRNA), col="red")

 dbDisconnect(mRNA_con)
}

#Check correlations
strong_pos_mRNA<-
subset(pos_mRNA,PCC>=0.7,select=c(miRNA,Gene,PCC,P_value,adj_P_p))
strong_neg_mRNA<-subset(neg_mRNA,PCC<=-
0.7,select=c(miRNA,Gene,PCC,P_value,adj_P_p))
strong_pos_prot<-subset(pos_prot,PCC>=0.7,select=c(miRNA,Protein,PCC,P_value,adj_P_p))
strong_neg_prot<-subset(neg_prot,PCC<=-
0.7,select=c(miRNA,Protein,PCC,P_value,adj_P_p))

#Extract strong correlations
apply(X=strong_pos_mRNA,MARGIN = 1,FUN=check_cor)
apply(X=strong_neg_mRNA,MARGIN = 1,FUN=check_cor)
apply(X=strong_pos_prot,MARGIN = 1,FUN=check_cor)
apply(X=strong_neg_prot,MARGIN = 1,FUN=check_cor)

#Intersection of correlations on both mRNA and protein levels

for (i in vector) {
 setwd("/Users/emmyborgmastars/Documents/Rstats/databases/DATA.DIR")
 neg_prot<-rbind(neg_prot,whole_pipeline[[i]]$neg_temp_sign_prot)
 neg_prots<-whole_pipeline[[i]]$neg_temp_sign_prot
 pos_prots<-whole_pipeline[[i]]$pos_temp_sign_prot
 neg_mRNAs<-whole_pipeline[[i]]$neg_temp_sign_mRNA
 pos_mRNAs<-whole_pipeline[[i]]$pos_temp_sign_mRNA

 prots<-union(neg_prots$Protein,pos_prots$Protein)
 mRNAs<-union(neg_mRNAs$Gene,pos_mRNAs$Gene)
 both_levels<-intersect(prots,mRNAs)
 write.table(both_levels,paste0(i,'_both_levels.txt'),sep="\t")
}
```

2.4. Check how many samples express a specific miRNA

```
setwd("/Users/emmyborgmastars/Documents/Rstats/databases")

#Create table
table<-data.frame(matrix(ncol=2,nrow=0))
colnames(table)<-c("miRNA_ID","expr_numb")


count_expr<-function(miRNA){
  mRNA_con <- dbConnect(SQLite(),'miRNAmRNACor.sqlite')
  count_miRNA<<-dbGetQuery(mRNA_con, paste0('SELECT * FROM miRNA WHERE
miRNA_name IS"',   miRNA, '\n"'))
  v_miRNA<<-as.numeric(count_miRNA[1,])
  table<<-
rbind(table,data.frame("miRNA_ID"=miRNA,"expr_numb"=length(na.omit(v_miRNA))))
  dbDisconnect(mRNA_con)
  }

#Import list of mirnas
mirnas<-read.csv("mirna_list.txt",header=F)

#loop searches with each mirna using test function, margin = 1 (row), 2 (col)
apply(X = mirnas, MARGIN = 1, FUN = count_expr)

setwd("/Users/emmyborgmastars/Documents/Rstats/databases/DATA.DIR")
write.table(table,"miRNA_expression_count.txt",sep="\t")
```

2.5. -3p and -5p annotation of mature miRNAs

Script: merge_miRNA.R

```
#Function for merging

merge_mirna<-function(X){
 sample_ID<-X[7]

 #Import data
 x<-read.csv(paste0(sample_ID,'.csv'),header=T)
 x<-data.frame(x[1],x[3])
 m<<-merge(x,m,all=TRUE,by="mature.miRNA_region")
}
```

Script: 3p_5p_annotation.R

```
library(plyr)

#function
convert_to_3p5p<-function(X){
   file_ID<-X[1]
   file_name<-X[2]
   sample_ID<-X[7]
 address<-(paste0('/Users/emmyborgmastars/Documents/Pythonstats/miRNA_isoform/',
file_ID))
setwd(address)

data<-read.csv(file_name, header=TRUE,sep="\t")


#Extract only rpm and miRNA
df<-data.frame(data$reads_per_million_miRNA_mapped,data$miRNA_region)
dfr<-subset(df,data.reads_per_million_miRNA_mapped >= 1)

#Summarize all rpm values of same miRNA identity
ny<-
 ddply(dfr,.(data.miRNA_region),summarize,sum_rpm=sum(data.reads_per_million_miRNA_
 mapped))

#Create log2 values, with sample_ID as colname
ny[sample_ID]<-log2(ny$sum_rpm)
colnames(ny)<-c("mature,miRNA_region","sum_rpm",sample_ID)

setwd("/Users/emmyborgmastars/Documents/Rstats/miRNA_data")
#Write to csv file
write.table(ny,paste0(sample_ID,'.csv'),sep=",")
```

```
setwd("/Users/emmyborgmastars/Documents/Rstats")
return(nrow(ny))

}
```

Script: 3p5ppipeline.R

```
setwd("/Users/emmyborgmastars/Documents/Rstats")

library(plyr)

##This package does not seem to be necessary
library(stringr)

#Run function script
source("3p_5p_annotation.R")

#Import data
df<-read.csv("gdc_sample_sheet_2018-02-13-03-23-11.tsv.2018-02-
 13T14_23_27.088749.tsv",sep="\t",header=T)

#create data frame
mirna_expr<-data.frame(matrix(ncol=1,nrow=1))
colnames(mirna_expr)<-"MIMA_ID"

#Run loop over each data row
apply(X = df, MARGIN = 1, FUN = convert_to_3p5p)

####MERGE
setwd("/Users/emmyborgmastars/Documents/Rstats")
source("merge_mirna.R")
setwd("/Users/emmyborgmastars/Documents/Rstats/miRNA_data")

m<-data.frame(matrix(ncol=1,nrow=1))
colnames(m)<-"mature.miRNA_region"


apply(X = df, MARGIN = 1, FUN = merge_mirna)

write.csv(m,"all_mirnas.csv")

#Transpose data frame
df<-read.csv("all_mirnas.csv",header=TRUE,sep=";")
df<-df[,-1]
df<-df[,-1]

tr<-t(df)
```

```
write.csv(tr,"t_mirnas.csv")
```

Perl script: MIMA_tr.pl

```
#For translating MIMAT-ID:s to miRNA names

#!/usr/bin/perl

use warnings;

#Open files
open(my $mRNA_file, '<', "mirb_IDANDname.csv");

#Open .csv file, only first line

open my $file, '<', "MIMA_t_mirnas.csv";
my $firstLine = <$file>;
close $file;


my %hash;

#Loop over each MIMA-hsa pair and save in hash
while (<$mRNA_file>){
  my $line = $_;
  my @split = split(";", $line);
                    $hash{$split[0]} = $split[1];

}



#Create file
if(!open(WRITE,">pat1.csv")){
die "Could not open file for writing\n";
}


#foreach key change to its corresponding value
foreach (keys %hash){
$firstLine =~ s/$_/$hash{$_}/;

}

print (WRITE $firstLine);
```

```
#Dissociate filehandles
close(WRITE);
close $mRNA_file;
```

2.6. Extract top enriched GO terms and KEGG pathways

Bash script: 180608_sortGOKEGG.sh

```
#! /bin/bash

#Start with the files before evaluation
cd
/Users/emmyborgmastars/Documents/Rstats/databases/DATA.DIR/Functional_enrichment/
Before_evaluation


#GO terms For loop, sort all files in the directory 'Before_evaluation'
for file in GO_*.txt
do

 sort -k6,6g -n -t ";" "$file" > sorted_"$file"

done

#KEGG pathways
for file in KEGG_*.txt
do

 sort -k5,5g -n -t ";" "$file" > sorted_"$file"

done

#Same procedure for after evaluation
cd
/Users/emmyborgmastars/Documents/Rstats/databases/DATA.DIR/Functional_enrichment/
After_evaluation


#GO terms For loop, sort all files in the directory 'Before_evaluation'
for file in GO_*.txt
do

 sort -k6,6g -n -t ";" "$file" > sorted_"$file"

done

#KEGG pathways
for file in KEGG_*.txt
do

 sort -k5,5g -n -t ";" "$file" > sorted_"$file"
```

done

Script: 180608_topGOKEGG.R

```
#Create empty data frames
tab_GO<-as.data.frame(matrix(0, ncol = 1, nrow = 1))
tab_KEGG<-as.data.frame(matrix(0, ncol = 1, nrow = 1))

#Function that extracts the top GO term and KEGG pathway
extract_top_BF<-function(miRNA){
 f<-paste0('sorted_GO_',miRNA,'.txt')
 mirna_name<-as.data.frame(miRNA)

 #GO terms
 if (file.exists(f)){
  GO_terms<-read.table(f,header=T, fill=T, sep=";")
  GO_BP<-subset(GO_terms,Ont=='BP',select=c(Term,Ont,N,DE,P.DE))
  number_of_BP<-as.data.frame(nrow(GO_BP))
  colnames(number_of_BP)<-"Number_of_BP"
  topGO<-GO_BP[1,]
  tab_GO<-cbind(tab_GO,mirna_name,topGO,number_of_BP)
  tab_GO<-tab_GO[,-4]
  tab_GO<-tab_GO[,-4]
 }
 else {
  Term<-"NA"
  DE<-as.integer("NA")
  P.DE<-as.integer("NA")
  Number_of_BP<-as.integer("NA")
  tab_GO<-cbind(tab_GO,mirna_name,Term,DE,P.DE,Number_of_BP)
 }

 #KEGG pathways

 if (file.exists(f)){
  KEGG_terms<-read.table(paste0('sorted_KEGG_',miRNA,'.txt'),header=T,fill=T, sep=";")
  topKEGG<-KEGG_terms[1,]
  number_of_KEGG<-as.data.frame(nrow(KEGG_terms))
  colnames(number_of_KEGG)<-"Number_of_KEGG"
  tab_KEGG<-cbind(tab_KEGG,mirna_name,topKEGG,number_of_KEGG)
  tab_KEGG<-tab_KEGG[,-4]
 }
 else {
  Pathway<-"NA"
  DE<-as.integer("NA")
```

```
   P.DE<-as.integer("NA")
   Number_of_KEGG<-as.integer("NA")
   tab_KEGG<-cbind(tab_KEGG,mirna_name,Pathway,DE,P.DE,Number_of_KEGG)
  }

  return (list("tabGO"=tab_GO,"tabKEGG"=tab_KEGG))
}

setwd("/Users/emmyborgmastars/Documents/Rstats/databases")

#Import list of mirnas
mirnas<-read.csv("mirna_list.txt",header=F)

setwd("/Users/emmyborgmastars/Documents/Rstats/databases/DATA.DIR/Functional_enric
hment/Before_evaluation")

#Run function
topterms_before_eval<-apply(X = mirnas, MARGIN = 1, FUN = extract_top_BF)

setwd("/Users/emmyborgmastars/Documents/Rstats/databases/DATA.DIR/Functional_enric
hment/After_evaluation")
topterms_after_eval<-apply(X = mirnas, MARGIN = 1, FUN = extract_top_BF)

vector<-c(1:15)

#GO terms extraction
summary_GO_before<-c()

for (i in vector) {
  summary_GO_before<-rbind(summary_GO_before,topterms_before_eval[[i]]$tabGO)
}

summary_GO_before<-summary_GO_before[,-1]
colnames(summary_GO_before)<-c("miRNA","Top GO term", "Number of genes", "P-
value","Number of GO terms")
setwd("/Users/emmyborgmastars/Documents/Rstats/databases/DATA.DIR/Functional_enric
hment/Before_evaluation")
write.table(summary_GO_before,"topGOterms.txt",sep=";",row.names = F)

summary_GO_after<-c()

for (i in vector) {
  summary_GO_after<-rbind(summary_GO_after,topterms_after_eval[[i]]$tabGO)
}

summary_GO_after<-summary_GO_after[,-1]
```

```
colnames(summary_GO_after)<-c("miRNA","Top GO term", "Number of genes", "P-value","Number of GO terms")
setwd("/Users/emmyborgmastars/Documents/Rstats/databases/DATA.DIR/Functional_enrichment/After_evaluation")
write.table(summary_GO_after,"topGOterms_after.txt",sep=";",row.names = F)

#KEGG pathways extraction

#Before evaluation
summary_KEGG_before<-c()
for (i in vector) {
  summary_KEGG_before<-
rbind(summary_KEGG_before,topterms_before_eval[[i]]$tabKEGG)
}

summary_KEGG_before<-summary_KEGG_before[,-1]
colnames(summary_KEGG_before)<-c("miRNA","Top KEGG pathway","Number of genes","P-value","Number of KEGG pathways")
setwd("/Users/emmyborgmastars/Documents/Rstats/databases/DATA.DIR/Functional_enrichment/Before_evaluation")
write.table(summary_KEGG_before,"topKEGGterms.txt",sep=";",row.names = F)

#After evaluation
summary_KEGG_after<-c()
for (i in vector) {
  summary_KEGG_after<-rbind(summary_KEGG_after,topterms_after_eval[[i]]$tabKEGG)
}

summary_KEGG_after<-summary_KEGG_after[,-1]
colnames(summary_KEGG_after)<-c("miRNA","Top KEGG pathway","Number of genes","P-value","Number of KEGG pathways")
setwd("/Users/emmyborgmastars/Documents/Rstats/databases/DATA.DIR/Functional_enrichment/After_evaluation")
write.table(summary_KEGG_after,"topKEGGterms_after.txt",sep=";",row.names = F)
```

2.7. Plot mature and miRNA gene expression

```
#Make plots of hsa-mir-144, hsa-miR-144-3p and hsa-miR-144-5p

#Import tables
setwd("/Users/emmyborgmastars/Documents/Rstats/databases")
mir_144<-read.table("TCGA-PAAD.mirna.tsv",sep="\t",header=T)

setwd("/Users/emmyborgmastars/Documents/Rstats/databases/for_correlation/backup
 gammalt")

miR_144<-read.table("180323_miRNA_formRNA.txt",sep="\t",header=T)

diff<-setdiff(colnames(mir_144),colnames(miR_144))
mir_144<-subset(mir_144, select=-c(TCGA.IB.7647.01A))

ny_mir_144<-mir_144[,order(colnames(mir_144))]
ny_miR_144<-miR_144[,order(colnames(miR_144))]

im_mir_144<-subset(ny_mir_144,miRNA_ID=="hsa-mir-144",select=colnames(ny_mir_144))
miR_144_3p<-subset(ny_miR_144,miRNA_name=="hsa-miR-144-
 3p\n",select=colnames(ny_miR_144))
miR_144_5p<-subset(ny_miR_144,miRNA_name=="hsa-miR-144-
 5p\n",select=colnames(ny_miR_144))

miR_144_3p[is.na(miR_144_3p)] <- 0
miR_144_5p[is.na(miR_144_5p)] <- 0

im_mir_144<-subset(im_mir_144, select=-c(miRNA_ID))
miR_144_3p<-subset(miR_144_3p, select=-c(miRNA_name))
miR_144_5p<-subset(miR_144_5p, select=-c(miRNA_name))

v_im_mir_144<-as.numeric(im_mir_144)
v_miR_144_3p<-as.numeric(miR_144_3p)
v_miR_144_5p<-as.numeric(miR_144_5p)

c<-0.5
p<-1

setwd("/Users/emmyborgmastars/Documents/Rstats/databases/DATA.DIR")
tiff(width = 7, height = 5, file=paste0("180421_mirnas.tiff"), units="in", res=600)

plot(v_im_mir_144,ylab="log2(rpm)",xlab="TCGA
 samples",col="red",ylim=range(0:12),pch=p,cex=c)
lines(v_im_mir_144,col="red")
text(161,11.5,"mir-144",col="red",cex=0.8)
```

```
points(v_miR_144_3p,ylab="log2(rpm)",xlab="TCGA samples",col="black",pch=p,cex=c)
lines(v_miR_144_3p,col="black")
text(165,10.85,"miR-144-3p",col="black",cex=0.8)

points(v_miR_144_5p,ylab="log2(rpm)",xlab="TCGA samples",col="dark grey",pch=p,cex=c)
lines(v_miR_144_5p,col="grey")
text(165,10.2,"miR-144-5p",col="azure4",cex=0.8)

points(148,11.5,pch=p,cex=c,col="red")
points(148,10.85,pch=p,cex=c,col="black")
points(148,10.2,pch=p,cex=c,col="azure4")
dev.off()
```