TURUN YLIOPISTO

miRNA data analysis workflow

Minna Kyläniemi
Master's thesis
Department of Future Technologies
Bioinformatics
University of Turku
June 2019

UNIVERSITY OF TURKU
Department of Future Technologies

MINNA KYLÄNIEMI: miRNA data analysis workflow

Master's thesis, 68 p., 4 p. appendices
Master's Degree Programme in Bioinformatics

Micro RNAs (miRNA) have been shown to regulate many biological processes by silencing the expression of their target genes. They are small non-coding RNAs that have been found in all types of organisms from eukaryotes to viruses. It has been shown that one miRNA can have several target genes and on the other hand, one gene can be targeted by several different miRNAs. Thus, the analysis of miRNA data is complicated. The aim of this project was to develop a workflow for miRNA functional analysis and test its functionality with some published datasets.

The workflow for the functional analysis of miRNAs includes miRNA differential expression analysis, target gene identification and functional enrichment analysis. The first aim of this project was to find a suitable database or a set of databases to retrieve miRNA target predictions. By literature search, information was gathered about different miRNA target prediction databases that are currently available. mirDIP4.1, which collects predictions from 30 different resources and is updated frequently, was selected as the source of miRNA target predictions.

For the functional analysis, two different tools were tested. First one of these, R/Bioconductor package mdgsa is based on gene set enrichment analysis. The other one, BUFET is a python script that performs overrepresentation analysis with empirical correction for bias that is often observed in miRNA functional analysis. For the testing of these algorithms, datasets from three different publications were used with miRNA target predictions from various sources. As expected, the results from different approaches differed both from the original publications and from each other. One reason for the differences observed in results compared to those of the original method publications was the different target prediction database that was used here.

Keywords: micro RNA, miRNA, target prediction, functional enrichment analysis, gene set enrichment analysis, overrepresentation analysis, pathway analysis

Table of contents

ABBREVIATIONS

Bp	basepair
DNA	deoxyribonucleic acid
GO	Gene Ontology
KEGG	Kyoto Encyclopedia of Genes and Genomes
miRNA	micro RNA
mRNA	messenger RNA
MsigDB	Molecular signatures database
NGS	next generation sequencing
Nt	nucleotide
ORA	overrepresentation analysis
RNA	ribonucleic acid

1 Introduction

The development of technologies such as microarrays and next generation sequencing has enabled analysis of DNA and RNA molecules from cells and tissues on the whole genome or transcriptome level. These analyses have revealed differences in amounts of RNA molecules between samples related to proliferation, differentiation, cell type, disease or treatments of samples (Hausser and Zavolan 2014; Lin and Gregory 2015; Mehta and Baltimore 2016; Riffo-Campos, Riquelme, and Brebi-Mieville 2016). In addition to messenger RNAs (mRNA) that are translated into proteins, other types of RNAs called non-coding RNAs, have been described. Non-coding RNAs can be divided into long and small non-coding RNAs. One class of small non-coding RNAs are micro RNAs (miRNAs) that were described first time in early 1990s in *Caenorhabditis elegans* (R. C. Lee, Feinbaum, and Ambros 1993). However, it took about a decade before their role in biology was revealed and there is still plenty to learn.

miRNAs have been discovered from most eukaryotes and even from viruses. This indicates that their role in regulating biological processes is important for many types of organisms. There are currently more than 2600 different mature human miRNA sequences listed in the miRBase v22 database (Kozomara, Birgaoanu, and Griffiths-Jones 2019), which is a public repository of known miRNA sequences and annotation. In addition to human miRNAs, there are more than 48 000 mature miRNAs from 271 organisms, for example mouse, *Caenorhabditis elegans* and plants such as *Arabidopsis thaliana* (Kozomara, Birgaoanu, and Griffiths-Jones 2019). miRNAs regulate cellular processes mainly by silencing their target genes (Hausser and Zavolan 2014) and the majority of mammalian mRNAs have been predicted to be targeted by at least one miRNA (R. C. Friedman et al. 2009).

Changes in the miRNA expression levels is one regulating factor for many biological processes, such as differentiation, proliferation, cell death, cancer and the regulation of immune system (Hausser and Zavolan 2014; Lin and Gregory 2015; Mehta and

Baltimore 2016; Riffo-Campos, Riquelme, and Brebi-Mieville 2016). For example, aberrant miRNA expression occurs in many cancer cells (Lin and Gregory 2015). Some miRNAs like let-7 miRNA family, are known to function as tumour suppressors by regulating the expression of oncogenes such as *MYC* (Lin and Gregory 2015). If the expression of such miRNA is downregulated, it can have severe consequences. It has been proposed that miRNA expression profiles could be used as diagnostic markers in some types of cancers (Lin and Gregory 2015) and miRNA expression profiles can be altered also in other diseases. Therefore, understanding the impact of miRNAs on biological processes on both healthy tissues and in disease is important.

The biological actions of miRNAs are complex. Usually one miRNA has up to hundreds of different target genes and on the other hand, one gene can have several different miRNAs that can regulate its expression. Determining the target genes of miRNAs is crucial for understanding their function in biological processes. The target genes of miRNAs can be identified by experimental validation or predicted computationally. Experimental validation of miRNA target interactions is demanding and expensive and the data is still limited. Therefore, miRNA target prediction tools are important resources in interpreting the impact that miRNAs have on regulating biological processes. Different miRNA target prediction tools are based on different combinations of biological properties of miRNAs and even tools that are based on same biological properties may have varying results. Therefore, the selection of target prediction tool is important step for miRNA data analysis.

In the cell, single molecules and genes act together as pathways. To study these pathway level changes, functional enrichment analysis is performed. For the analysis, defined gene sets for pathways or other biological terms are used. Functional enrichment analysis can be performed by over-representation analysis or gene set enrichment analysis. Because the gene sets are defined on gene level, the miRNA data must be translated to gene level or then the gene set information to miRNA level to be able to do the analysis.

miRNA functional analysis is challenging, since the target predictions by different algorithms are not uniform. In addition, some recent studies (Bleazard, Lamb, and

Griffiths-Jones 2015; Godard and Van Eyll 2015) have shown that commonly used approach to perform miRNA pathway analysis based on the list of their target genes by overrepresentation analysis has a bias towards certain pathways and might not give biologically relevant results. There is no standardised way for miRNA data analysis, and therefore it is important to gain more insight how choice of target prediction tool and functional enrichment analysis tool affect the results.

2 Review of the literature

2.1 Biogenesis of miRNAs

miRNAs are small non-coding RNAs, that have been shown to regulate many processes in the cells. miRNA genes in human genome are usually located on introns and intergenic regions, but some miRNA sequences are located on the coding regions of protein encoding genes (Treiber, Treiber, and Meister 2019). miRNAs are transcribed as precursor molecules that are cleaved to form the mature miRNAs that are about 22 nucleotides (nt) long (Lin and Gregory 2015). miRNAs are encoded in the genome either as clusters of many miRNA genes or as individual genes (Treiber, Treiber, and Meister 2019). The first, unprocessed miRNA molecule is called the primary miRNA (Pri-miRNA), which forms a loop structure in the nucleus (Y. Lee et al. 2004) (Figure 1). Pri-miRNA is cleaved in the nucleus to form a precursor miRNA (pre-miRNA) and then further processed by protein called DICER in the cytoplasm to release the mature miRNA molecule (Treiber, Treiber, and Meister 2019). Mature miRNA binds to a protein complex to form a miRNA induced silencing complex (miRISC), which binds to complementary mRNA sequences preventing the translation of mRNA into protein (Lin and Gregory 2015). The binding of miRISC into mRNA can also induce the cleavage of target mRNA, its accelerated degradation or to mRNA deadenylation, which destabilizes the mRNA (Jonas and Izaurralde 2015) and leads to the loss of target mRNA. However, miRNA mediated gene regulation is described more as gene silencing than inhibition of target gene expression since miRNA mediated gene regulation leads usually only to decreased protein expression and not to the total abolishment of the target protein (Jonas and Izaurralde 2015).

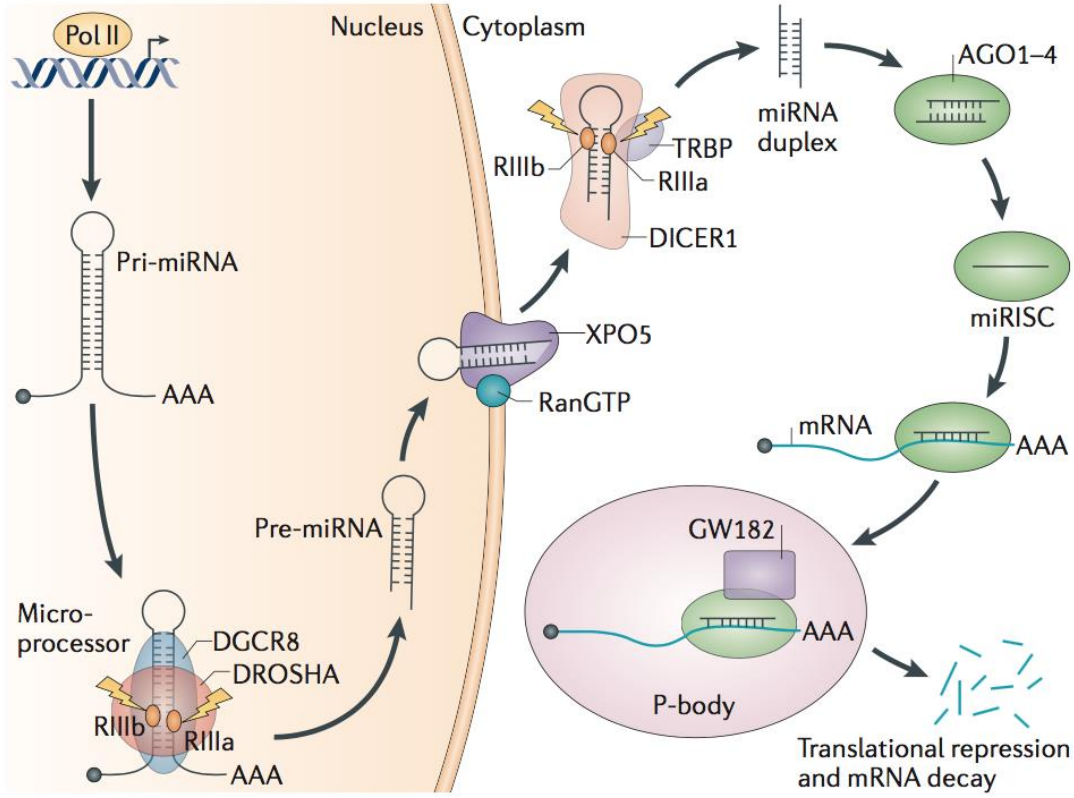

Figure 1. miRNA biogenesis. miRNAs are expressed as primary miRNAs from the genes encoded in the genome. Pri-miRNA is further processed by a protein complex formed by DROSHA and DGCR8, that cleave the pri-miRNA to produce precurosor miRNA (pre-miRNA). Pre-miRNA is then exported to cytoplasm and further processed by DICER1 to release the mature miRNA, which is about 22 nucleotides long. This short RNA forms a miRISC protein complex, which finds its target mRNAs by complementary binding to the miRNA. miRNA binding to its target mRNA leads to the inhibition of protein translation and degradation of target mRNA. Picture adopted from (Lin and Gregory 2015)

The regulation of gene expression by miRNAs is a complex process. The binding sites of miRNAs are usually located on untranslated 3' ends of mRNA and one miRNA can have several binding sites in the same target mRNA, but miRNAs can also have several different target genes (Jonas and Izaurralde 2015; Riffo-Campos, Riquelme, and Brebi-Mieville 2016). In addition, many genes are targeted by several different miRNAs (Riffo-Campos, Riquelme, and Brebi-Mieville 2016). It has been also predicted that most mammalian mRNAs are targets of some miRNAs (R. C. Friedman et al. 2009). Therefore, miRNAs can either antagonize or strengthen the effect of each other on the gene expression pattern of a cell. Although the target site of miRNA is

usually located in the 3'-end of the mRNA, some studies have shown that miRNAs can also bind to 5'-ends of their target mRNAs (Ørom, Nielsen, and Lund 2008) or to the coding region of gene (Guo et al. 2015; Schnall-Levin et al. 2011). Many possible target sites on mRNAs lead to more difficult prediction of miRNA target genes.

The miRNA sequence can be divided into seed sequence and 3' end sequence, which have different importance in determining the target site of miRNA (Figure 2). The seed sequence of miRNA is the eight nucleotides at the 5'- end of the miRNA. This seed sequence is the most important in determining the target sites of miRNAs, but even the seed complementarity to target mRNA does not need to be perfect. Some mismatches are also allowed in the 3'-end of the miRNA. This imperfect complementarity needed for binding increases the number of potential target genes for any miRNA. (Peterson et al. 2014; Riffo-Campos, Riquelme, and Brebi-Mieville 2016)

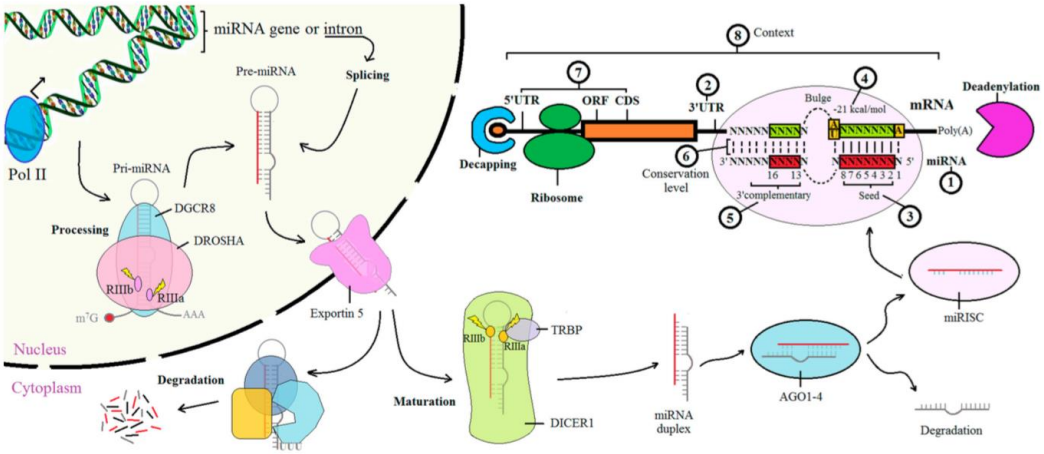

Figure 2. The biological basis of miRNA target prediction algorithms. miRNA binding sites are usually located on 3'UTR regions (2) of mRNAs. ~22 nucleotides long miRNA can be divided into seed sequence (3) and to 3'complementary sequence (5). Many of the miRNA-mRNA sequences are conserved across species (6) and there is more conservation on the seed sequence than on 3'complementary sequence. The stability of miRNA-mRNA binding can be estimated by calculating the free energy of miRNA-mRNA duplex (4). miRNA-miRISC binding to its target site can lead to deadenylation

of mRNA, which destabilizes mRNA. In addition, it can lead to the degradation of mRNA. Image from (Riffo-Campos, Riquelme, and Brebi-Mieville 2016).

The seed region can be used to classify miRNAs into families. Information about miRNA families is collected into Rfam database (rfam.xfam.org) (Kalvari, Argasinska, et al. 2018; Kalvari, Nawrocki, et al. 2018). The latest version of Rfam, version 13.0, lists 298 miRNA families in human genome. Some miRNA families have been shown to have an important role in specific developmental stages or in other biological processes. For example, let-7 miRNA family have been shown to have a role in development and in tumour suppression in vertebrates, but also in *C. elegans* and *Drosophila* (Lin and Gregory 2015; Rougvie 2001). In addition, this miRNA family is highly conserved among species (Rougvie 2001). The evolutionary conservations of miRNAs highlight the importance they have in the regulation of different biological processes.

2.2 Analysis of miRNA data

From a cell or tissue sample all RNA molecules can be extracted. However, if using a kit optimised for mRNA isolation, the small non-coding RNAs may be excluded from the sample (Brown et al. 2018). Therefore, the sample preparation and RNA extraction needs to be designed properly, if small non-coding RNAs are the focus of interest. Similarly, the library preparation of next generation sequencing needs to be designed specifically for these small RNAs as regular protocols are not suitable for these 22 nt sequences (Giraldez et al. 2018). For microarray analysis, there is specific arrays available for miRNA analysis, but Next Generation Sequencing (NGS) is becoming more widely used as well.

After the RNA isolation of the sample, the expression of different miRNAs is analysed by sequencing or microarrays (Figure 3). NGS allows the analysis of the whole miRNA population in a sample whereas microarrays are limited to those miRNAs that have predesigned probes on the array for their detection. The raw data from sequencing or array is then processed and the quality of data assured after which differential expression of miRNAs between samples can be calculated. This list of differentially

expressed miRNAs can then be used for further analyses. Similarly, as with gene expression data, the miRNA list itself is not usually very informative, but further analysis is necessary for the biological interpretation of data. In comparison to mRNA data, which is on the gene level, the problem with miRNA data is that gene sets of biological pathways are on the gene level whereas the miRNA data is not on gene level. Therefore, the lists of differentially expressed miRNAs need to be translated to the lists of genes that are affected by the miRNAs or then the pathway data would need to be translated from gene level to miRNA level. The target genes of miRNAs are either predicted by some of the available target prediction tools or then experimentally validated targets for the list of miRNAs are searched from databases. However, the data for experimentally validated miRNA target gene interactions is still limited, and therefore target prediction is often used. Functional enrichment analysis of miRNA data can be performed either by overrepresentation analysis or gene set enrichment analysis. miRNA target prediction and functional enrichment analysis are described in more detail in the following sections.

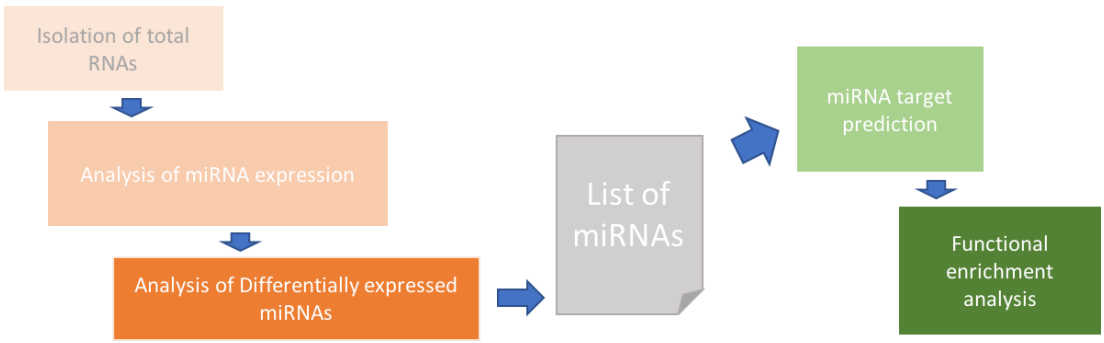

Figure 3. Typical miRNA analysis workflow. The analysis of miRNAs starts with the isolation of total RNAs from samples. Analysis of miRNA expression from these samples can be done by microarray analysis or by high throughput sequencing. Then the raw data is processed, quality is analysed and the differential expression of miRNAs calculated between different samples. The result is a list of miRNAs, which can be filtered by certain criteria or then the full list of miRNAs with test statistics is used for further analysis. miRNA target genes can be predicted by several different tools or then experimentally validated targets from databases can be used. Finally, functional enrichment analysis can be performed and for that gene sets can be selected from for example Gene Ontology (GO) or Kyoto Encyclopedia of Genes and

Genomes (KEGG). Functional enrichment analysis can be done by overrepresentation analysis or gene set enrichment analysis.

2.3 Identifying miRNA target genes

Determining the target genes of miRNAs is important for understanding the function of miRNAs in biological processes. Target genes can be either predicted based on different biological properties of miRNAs or interactions of miRNAs with their target mRNAs can be experimentally validated in laboratory.

2.3.1 Target prediction algorithms

Development of accurate target prediction algorithms has been active for more than 15 years and currently there is more than 180 different resources listed in the OMIC-tools database for miRNA target prediction (www.omictools.com; 04/2019). Another resource for miRNA analysis tools, Tools4miR (www.tools4mirs.org), lists 59 different tools for miRNA target prediction. Tools4miR is manually curated and frequently updated (last update in March 2019) platform listing tools for miRNA analysis (Lukasik, Wójcikowski, and Zielenkiewicz 2016). Thus, there are plenty of tools for miRNA target prediction to choose from, but different algorithms are based on different biological properties of miRNAs (Figure 2) and the results of individual tools differ.

One of the common features that are used by several target prediction algorithms is a seed match, which refers to Watson-Crick base pairing between the seed region of miRNA and the mRNA. Seed match can be considered as 6mer, 7mer or 8mer match, depending on the length of seed that has full complementarity to target mRNA. Second, the conservation of the miRNA sequence, 3'UTR binding sites on mRNA and 5' UTR binding sites on mRNA are used either separately or in combination. Conservation of these regions is considered across species. Third biological property is the free energy of miRNA-mRNA binding, which relates to the stability of the binding. In addition, mRNA secondary structure, which affects the accessibility of the miRNA target site is commonly considered by target prediction algorithms. These

features are used by many target prediction tools in different combinations and together with other features such as the number of miRNA binding sites in mRNA, complementarity of 3' end of miRNA to mRNA, the position of the miRNA target site on mRNA and machine learning approaches. (Peterson et al. 2014; Riffo-Campos, Riquelme, and Brebi-Mieville 2016)

Some of the earliest or most used tools for miRNA target prediction are listed in Table 1. These algorithms have all been published 2003-2007 and while targetScan (updated 03/2018) (Agarwal et al. 2015; Grimson et al. 2007; Lewis, Burge, and Bartel 2005), RNA22 (updated 04/2015) (Miranda et al. 2006) and DIANA-microT (updated 2013) (M. Maragkakis et al. 2009) are still updated, miRanda (Enright et al. 2003; John et al. 2004) is deprecated and PITA (Kertesz et al. 2007) and PicTar (Krek et al. 2005) have not been updated for 10 years. These algorithms rely on some combination of the commonly used biological features of miRNAs and are mainly predicting miRNA interaction to 3'UTR sequences of mRNAs. However, some of these algorithms, such as miRanda and RNA22, use miRNA sequence and target gene 3'UTR or whole sequences or intronic regions as input and thus these tools can be used to find miRNA target sites anywhere from the genome.

Table 1. Widely used tools for miRNA target prediction

Target prediction tool	Species	Features used in prediction	Target sites on mRNA	References
TargetScan	mammals (8), drosophila, roundworm (*C. elegans)*	seed, conservation	3'UTR	Lewis *et al.* (2005), Grimson *et al.* (2007), Agarwal *et al.* (2015)
RNA22	human, mouse, roundworm (*C. elegans), drosophila*	seed, energy, accessibility	3'UTR	Miranda *et al.* (2006)
PITA	human, mouse, roundworm (C. elegans), drosophila	seed, energy, accessibility	3'UTR	Kertesz *et al.* (2007)
miRanda*	any (search done by miRNA sequence + genomic sequence)	seed, conservation, energy		Enright *et al.* (2003), John *et al.* (2004)
PicTar	human, mouse, roundworm (C. elegans), drosophila	seed, conservation, energy, accessibility	3'UTR	Krek *et al.* (2005)
DIANA-microT	human, mouse, roundworm (C. elegans), drosophila	seed, conservation, energy, accessibility, ML	3'UTR	Maragkakis *et al.* (2009)

* miRanda tool not available from May 2018
ML = machine learning

One problem with target prediction algorithms is that many of the tools that are published are not updated after their publication. Therefore, new information is not added when new miRNAs are found or more information is gained on factors affecting miRNA target interaction. In addition, every time that mirBase, the database for micro RNAs (Griffiths-Jones 2004; Kozomara, Birgaoanu, and Griffiths-Jones 2019), is updated to a new version, some miRNAs that have been earlier identified as miRNAs are removed and some new miRNAs are usually added. Some of the prediction tools are also based on a specific version of mirBase and miRNA names need to be translated to the correct version with some suitable tool such as miRNAmeConverter (R/Bioconductor package).

Another problem in using a single algorithm for miRNA target prediction is that the results of prediction tools differ and even algorithms that are based on the same biological properties of miRNAs have varying results (Figure 4) (Tokar et al. 2017). For example, mirBase and RepTar both are based on the sequence analysis and the binding energy of miRNA-mRNA duplex, but their results have very low Jaccard index indicating highly different results (Figure 4). The biggest overlap that was observed by Tokar *et al.* was between BCmicro and TargetRank and resulted in Jaccard index of 0.3 showing that these target predictions have some overlap, but the similarity is not very strong (Tokar et al. 2017). In the Figure 4, sequence analysis refers to the seed match and binding energy is the same as the free energy of miRNA-mRNA duplex. Another problem with miRNA target prediction algorithms is that it is still not known which of the biological features of miRNAs are the most important for their interaction with the target genes and their regulatory function. However, it has been shown that target prediction algorithms that are based only on the seed match and the free energy of binding, give less confident predictions than tools that are more advanced (Tokar et al. 2017). In addition, it has been recently demonstrated that prediction algorithms predict many false positive targets (Pinzón et al. 2017), which makes the selection of the target prediction algorithm more difficult. However, if a target gene is predicted by many different algorithms, it is more probable that miRNA target interaction occurs *in vivo.*

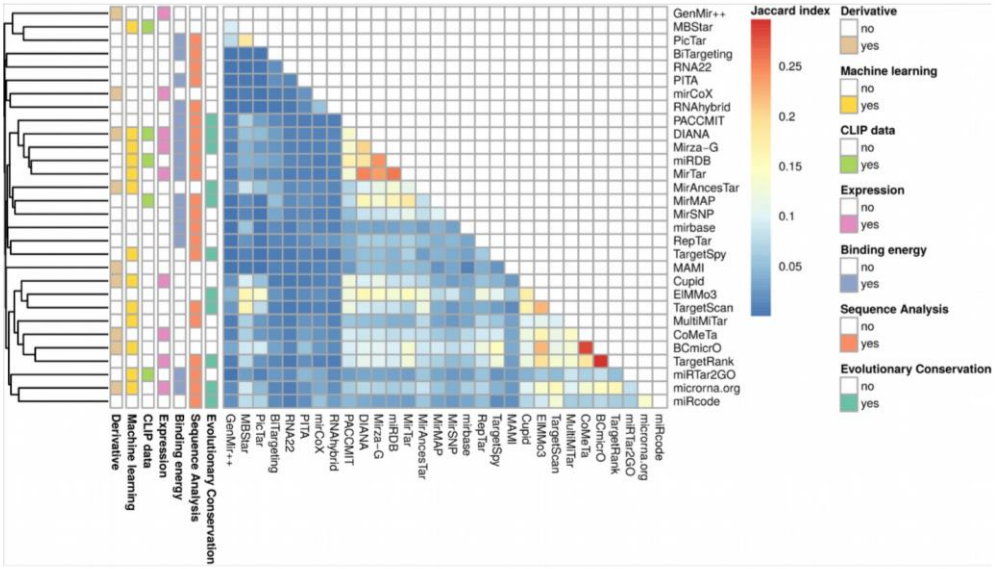

Figure 4. The results of different miRNA target prediction algorithms are not uniform. The colours indicate Jaccard indexes between the pairs of algorithms and information on the left of the panel shows some of the features that are common between algorithms. Figure adopted from (Tokar et al., 2017).

The number of target genes predicted for each miRNA can vary greatly by different resources. Currently, it is believed that one miRNA can target hundreds of genes, but it has been also shown that not all predicted miRNA targets experimentally validated and the number of false positive predictions can be high (Pinzón et al. 2017). If the target prediction algorithm results with too many target genes, the functional analysis will be skewed. As the target prediction is the first and crucial step in understanding the function of miRNAs, the accuracy of predictions is very important. It has been also stated that reliable target prediction prevents the bias in functional enrichment analysis (Tokar et al. 2017).

2.3.2 Experimentally validated targets

Although the number of experimentally validated targets for miRNAs is constantly increasing, the data is still limited. Methods that are used for the validation of miRNA-target gene interactions can be divided into low- and high-throughput methods. By

using low yield methods, interactions and/or regulation by only a few miRNAs can be studied at a time. With the high-throughput methods, miRNA binding sites throughout the genome or changes in the gene expression by miRNA on the whole transcriptome level can be explored. The data concerning the miRNA-mRNA interactions acquired by different techniques does not have equal confidentiality as some of the methods cannot distinguish direct from indirect regulation. (Karagkouni et al. 2017; Vlachos et al. 2015)

For example, quantitative RT-PCR (qRT-PCR), western blotting and ELISA can be used to verify miRNA targets on mRNA (qRT-PCR) or protein level (western blotting, ELISA). Another low throughput technique is using reporter gene assays, which can be used to directly and reliably verify miRNA target site interactions. Most of the high throughput methods are based on novel NGS techniques. Cross-linking immunoprecipitation (CLIP) sequencing is one widely used high-throughput method for miRNA target detection. The problem with traditional CLIP-seq is that it can identify the miRNA binding sites on genome wide level, but the interacting miRNA needs to be identified bioinformatically. In addition to CLIP-seq, other NGS based methods used are RNA immunoprecipitation sequencing (RIP-seq) and ribosome profiling sequencing (RPF-seq), which are used together with miRNA over-expression or silencing to elucidate the target sequences. Most recent techniques used for miRNA target identification include CLEAR-CLIP and CLASH, which include a ligation step where miRNAs are ligated with their target binding sites allowing detection of miRNA-mRNA duplexes. (Chou et al. 2017; Hausser and Zavolan 2014; Karagkouni et al. 2017; Vlachos et al. 2015)

The two biggest databases collecting information on experimentally validated miRNA target interactions are actively updated. The latest version of TarBase database is 8.0 and it has miRNA-mRNA interactions from 18 different species including for example human, mouse, rat and chicken (Karagkouni et al. 2017). TarBase includes information on 670 000 unique miRNA target interactions and collects data acquired from several different techniques such as CLEAR-CLIP, CLASH, RPF-seq, CLIP-seq and RIP-seq that all utilize NGS techniques, and from low-throughput techniques such as reporter assays and western blot (Karagkouni et al. 2017). Another database for

experimentally validated miRNA targets is mirTarBase and its latest version, V7.0, was published in 2017 (Chou et al. 2017). mirTarBase is slightly smaller than Tarbase with its 420 000 miRNA target interactions and it has interactions from 23 different species (Chou et al. 2017). Both databases use the mirBase v21 miRNA information while the latest mirBase version is v22 (released in October 2018).

Experimentally validated targets can be used alone or together with predicted targets for functional analysis, but they can be also used to train the target prediction algorithms to become more reliable. In addition, knocking out or silencing selected miRNAs is a way to study their functional role *in vivo* to validate the results obtained from bioinformatics analyses. Confirmation of target genes and validation of the role of miRNAs may have in regulating cellular processes are important steps in learning the mechanisms of action of miRNAs.

2.4 Functional enrichment analysis

Genes and other molecules in the cells act together as networks and pathways. Change of expression level of a single molecule should be considered in the context of the whole network to predict the effect on the cellular level. High-throughput methodologies such as NGS produce large quantities of data from which the differences of expression of the whole miRNA population can be studied. The analysis of these datasets has been moved from individual genes to the level of gene sets or networks that act together. This type of analysis can be called functional enrichment analysis or pathway analysis.

Similarly to mRNAs, some of the miRNA transcripts may have big differences on their expression levels, but some and often many are changed only mildly (Garcia-Garcia et al. 2016). These small changes can, however, be significant for the regulation of cellular processes, if many of the regulated genes belong to the same pathway and alter its function. Identically, number of miRNAs that are changed only slightly, but all regulate the same gene or genes belonging to the same pathway, might affect the function of some biological cascade. Therefore, a list of differentially expressed

miRNAs is not necessarily very informative and more advanced data mining tools such as functional enrichment analysis are needed for the interpretation of data. Functional enrichment analysis can be used to analyse the effect of transcriptional changes on a gene set level. For this, typically a list of differentially expressed miRNAs is used as input. The list can be filtered by for example fold change and p-value or then the full ranked list of transcripts can be used (Garcia-Garcia et al. 2016).

For analysis on pathway level defined gene sets are necessary and these sets can be formed by utilising existing biological knowledge. These gene sets can be derived for example from two of the most used databases; Gene Ontology (GO) (Ashburner et al. 2011; The Gene Ontology Consortium 2019) or Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al. 2019; Kanehisa and Goto 2000). In addition to these two very popular databases, there are many others that can be used such as the Molecular Signatures database (MSigDB) (Subramanian et al. 2005), which is a collection of gene sets maintained by GSEA team in Broad Institute. Gene sets can be formed for example based on their location on genome, biological function or cellular location.

In GO, genes are divided into three categories based on their biological properties. These categories are *Biological processes, Molecular Functions* and *Cellular Components.* GO terms are organized in a hierarchical way, where the top terms are more general and can have hundreds of genes. They are organized in a tree like structure and more general terms are followed by more specialised terms, which usually also have less genes in them. GO terms that have hundreds of genes or only a few are not very helpful in predicting the biological function of differentially expressed miRNAs. Enrichment of more specific terms is usually more informative for the understanding of biological functions. (Ashburner et al. 2011; The Gene Ontology Consortium 2019)

KEGG is a manually curated database for molecular networks. Unlike GO, it has pathway maps that show the interactions of different molecules in the pathway. KEGG pathways can be divided into networks related to metabolic functions,

regulatory functions and human diseases. (Kanehisa et al. 2019; Kanehisa and Goto 2000)

At present, there is no standardized way available for miRNA functional analysis. The functional analysis of miRNA data is more difficult than functional analysis of mRNA data since gene sets are defined on the gene level and thus either the data of differentially expressed miRNAs need to be translated to the gene level or then the gene sets to the miRNA level to be able to perform the analysis. The selection of miRNA target prediction tool or database for validated targets affects the results of the functional analysis (Tokar et al. 2017). However, to be able to interpret the biological functions of miRNAs, target gene prediction or validation of miRNA target genes and functional enrichment analysis is often necessary.

If an algorithm for functional enrichment analysis is dependent on miRNA target predictions from one specific source, the deprecation of prediction tool leads to unusable functional analysis tool as well. This happened when one of the most cited and used miRNA target prediction algorithms, miRanda, stopped working in May 2018. Functional analysis tools that allow the user to select the source of target predictions are more flexible and more robust to changes in the prediction algorithms. In addition, as several miRNAs can have binding sites in the same gene and on the other hand same miRNA can bind to several genes, these combined effects of miRNAs are not involved in all functional analysis tools. For the biological function of miRNAs, this synergistic effect is important and might affect the results of functional analysis (Garcia-Garcia et al. 2016).

Similar methods that are used for gene expression data are mainly used for miRNA functional enrichment analysis. In the next sections, the most used functional enrichment analysis tools, gene set analysis and overrepresentation analysis are introduced. Functional enrichment analyses of miRNA data have been thus far mainly performed with overrepresentation analysis, but a few algorithms for gene set analysis have been developed recently as well (Garcia-Garcia et al. 2016).

2.4.1 Overrepresentation analysis

Overrepresentation analysis is performed from a list of differentially expressed genes or miRNAs where the order of the list is not important. Hypergeometric test or Fisher's exact test are commonly used statistical methods to test whether genes belonging to certain pathways are overrepresented among the list of genes in question. The test is performed to evaluate whether the genes of a certain gene set are present in the list of differentially expressed genes more than by random chance would be expected Figure 5. (Gusev et al. 2007)

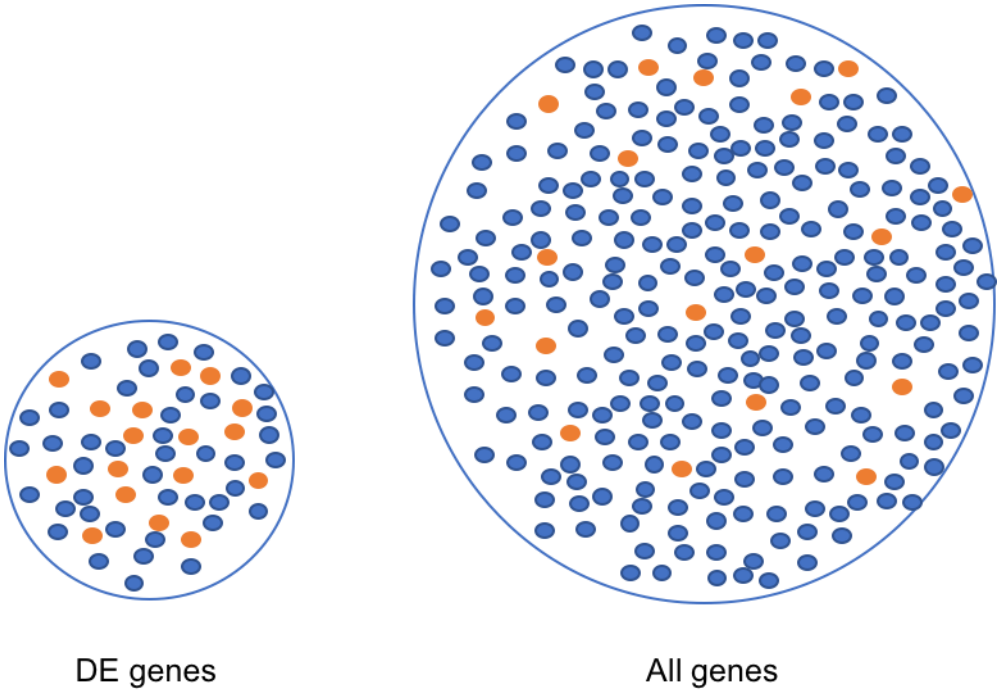

Figure 5. A schematic presentation of basis for the overrepresentation analysis. Orange dots present genes that belong to a certain gene set and blue dots represent all other genes that can be either the genes that have been analysed on a microarray or the whole genome if Next Generation Sequencing has been used. Among differentially expressed (DE) genes there is a certain number of genes belonging to this gene set. Among all genes, there is several genes belonging to this orange gene set, but not all of them are present in DE genes. Statistical test is performed to analyse whether there are more genes belonging to orange gene set in DE genes than would randomly be expected.

However, recently it has been shown that the traditional method of overrepresentation analysis on miRNA data (Gusev et al. 2007) leads to bias towards some pathways and enrichment can be seen even with random lists of miRNAs (Bleazard, Lamb, and Griffiths-Jones 2015). Therefore, the traditional overrepresentation analysis methods need to be modified to give unbiased results for miRNA data (Bleazard, Lamb, and Griffiths-Jones 2015; Zagganas et al. 2017). This bias in functional analysis of miRNA data is discussed more in Section 2.4.3.

2.4.2 Gene set analysis

In gene set analysis, the enrichment of genes belonging to a certain set or group in top or bottom of the ranked list of genes is studied. The ranking of genes can be done for example based on their expression level or differential expression. Statistical analysis is then performed to elucidate whether genes from one set are enriched on one end of the ranked list (Figure 6). The gene set enrichment method was described originally by Mootha *et al.* (Mootha et al. 2003) and further developed and named gene set enrichment analysis (GSEA) by Subramanian *et al.* (Subramanian et al. 2005). The GSEA method is based on estimating the statistical significance of the enrichment by empirical permutation test.

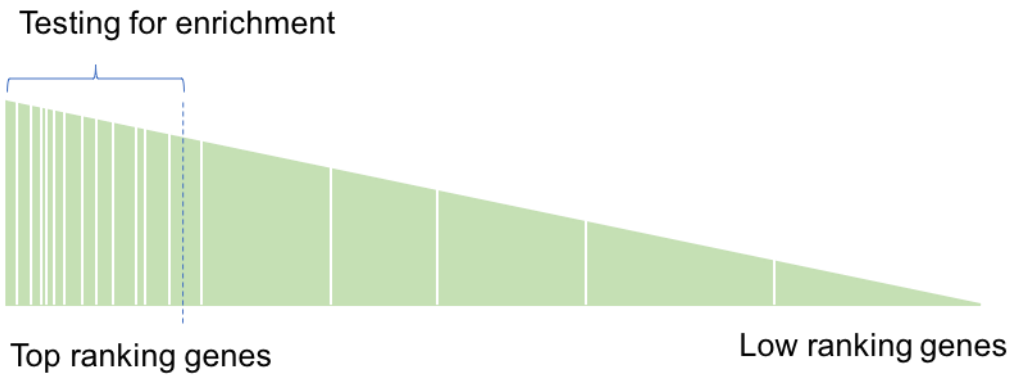

Figure 6. A schematic presentation of gene set enrichment analysis. All the genes that have been analysed are ranked according to for example their differential expression. Genes belonging to a certain gene set are presented as white lines and the rest of the genes are green. Statistical testing is performed to analyse whether genes belonging to a set (white lines) are more enriched among the top-ranking genes than expected by chance. The test can be also performed to see whether the genes of some gene set

are more enriched among low ranking genes, if genes are ranked according to their expression level from high expression to low expression.

There are several methods available for performing gene set analysis and it is a widely-used approach in analysing differentially expressed genes from NGS or microarray data. Gene set analysis is also useful for analysing miRNA data, but the data needs to be converted to gene level first. Gene set analysis has been used less than overrepresentation analysis in analysis of miRNA data (Garcia-Garcia et al. 2016). In a recently published method, the enrichment of a certain set of genes is calculated from the ranking statistics of genes by logistic regression models (Montaner and Dopazo 2010) and the synergistic effect of miRNAs regulating the same genes is included in the model (Garcia-Garcia et al. 2016).

2.4.3 Bias in miRNA functional analysis

Algorithms developed for the functional enrichment analysis of gene expression data are commonly used for miRNA data as well but recent studies have shown that in this case there is a bias towards certain pathways (Bleazard, Lamb, and Griffiths-Jones 2015; Godard and Van Eyll 2015). One reason for this observed bias is the step where miRNA data is translated to gene level for analysis (Bleazard, Lamb, and Griffiths-Jones 2015). It has been shown that many false positive targets are predicted by miRNA target prediction tools (Pinzón et al. 2017) that lead to unreliable enrichment results. In addition, the enrichment pattern can be seen by even random lists of miRNAs indicating that traditional way of functional analysis, which is often based on hypergeometric test, does not produce reliable results for miRNA data (Bleazard, Lamb, and Griffiths-Jones 2015). Because the identification of miRNA target genes is usually the first step in the functional analysis, confident predictions or experimental validation of target genes is important. It has also been stated that if the target prediction method is good, even hypergeometric test produces reliable results (Tokar et al. 2017). However, as different target prediction algorithms produce variable results, defining a good target prediction algorithm is difficult.

In the studies, which reported the bias in the functional analysis of miRNAs, new approaches for miRNA analysis were suggested (Bleazard, Lamb, and Griffiths-Jones

2015; Garcia-Garcia et al. 2016; Godard and Van Eyll 2015; Zagganas et al. 2017). In the study by Bleazard *et al.* (Bleazard, Lamb, and Griffiths-Jones 2015), an algorithm called empiricalGO was introduced. This algorithm is based on overrepresentation analysis of gene sets, but the analysis is improved by calculating an empirical p-value for the enrichment using iterations of random miRNA lists of the same size as the original list. They showed that after this correction, most of the published enriched pathways were not significant any longer. They also present a modified algorithm that consideres also the number of predicted target sites that input miRNAs have in each predicted target gene. In the datasets used in the analysis, this modified multi-hit empiricalGO algorithm resulted in different results than the original empiricalGO (Bleazard, Lamb, and Griffiths-Jones 2015). The empiricalGO algorithm was further developed in another study in which the algorithm was named BUFET. In BUFET algorithm the computational efficiency of the analysis was improved by replacing the use of hash tables with bitsets (Zagganas et al. 2017). However, the BUFET algorithm does not include the information of multiple miRNA target sites in the same gene (Zagganas et al. 2017).

In the study by Godard and van Eyll (Godard and Van Eyll 2015), overrepresentation analysis is performed in such a way that gene sets (pathways) are translated to the miRNAs that regulate these genes and then hypergeometric test is performed in miRNA level. They showed that with this approach the enrichment bias can be avoided. Overrepresentation analysis in general has been criticized for the loss of information since only a fraction of the whole set of analysed genes is used (Garcia-Garcia et al. 2016). Gene set enrichment analysis, which has not been widely used for miRNA data, was proposed as another solution to overcome the problem of biased results and the possible drawback in overrepresentation analysis (Garcia-Garcia et al. 2016). This algorithm is called mdgsa and in this algorithm, the miRNA differential expression is transferred to gene level by utilizing an inhibition score, which also includes the information of multiple miRNA target sites in the same gene. The whole list of differentially expressed miRNAs is used and both the direction of the change and the strength (p-value) are used to give a ranking index to the miRNAs. These indexes are then translated on gene level regulation incorporating the synergistic

effects that miRNAs can have on the genes that they are regulating. The mdgsa algorithm developed by Garcia-Garcia *et al.* (Garcia-Garcia et al. 2016) can also be modified to include mRNA expression information of the same samples, if available, to restrict the analysis on only those genes that are expressed on the sample. The results by mdgsa positively correlated with results obtained by Godard and van Eyll (Godard and Van Eyll 2015) and randomised miRNA lists did result in only a very small number of significant GO terms (Garcia-Garcia et al. 2016), indicating that mdgsa algorithm does not show significant bias in the functional analysis of miRNA data.

3 Aims of the study

The aim of this project was to develop a workflow for miRNA functional analysis. To do this, first step was to find a good miRNA target prediction tool for identifying miRNA target genes. Second aim was to test this target prediction tool with some published datasets and then use the obtained predictions in functional enrichment analysis. The third aim was to select two algorithms for functional enrichment analysis and run the analyses with the selected datasets and predictions obtained from the target prediction tool and to compare the obtained results with original results from method publication and between the algorithms. Algorithms for functional analysis were selected so that the recently shown bias in miRNA functional analysis (Bleazard, Lamb, and Griffiths-Jones 2015; Godard and Van Eyll 2015) would be avoided.

4 Materials and methods

R version 3.4.3 with R studio version 1.1.423 on Mac OS X 10.12.6 was used for the analyses done with R. Python version 2.7.11 on Mac OS X 10.12.6 was used for the analyses done with Python.

4.1 Datasets

Several datasets were used to compare the results from both the target prediction tools and from the functional analyses (Table 2). Datasets were selected from the publications in which the functional enrichment analysis tools selected for this study were presented (Bleazard, Lamb, and Griffiths-Jones 2015; Garcia-Garcia et al. 2016) and data from one independent study (Kassambara et al. 2017).

Table 2. Datasets used in this study.

Publication	Name in publication	Name in thesis	Number of miRNAs in filtered list
Kassambara *et al.* 2016	Cluster 1	Kassambara 1	23
	Cluster 2	Kassambara 2	14
	Cluster3	Kassambara 3	12
Bleazard *et al.* 2015	Tanic *et al.*	Tanic	46
	Raponi *et al.*	Raponi	15
Garcia-Garcia *et al. 2016*	KICH cancer dataset*	Kich01**	335
		Kich05**	386
		Kich05FC2**	274
	KIRP cancer dataset*	Kirp01**	400
		Kirp05**	459
		Kirp05FC2**	283

*dataset in original publication not filtered; ** dataset filtered with 01 = pval <0.01, 05 = pval <0.05, 05FC2 = pval < 0.05 and |Fold change| >2

Two different cancer dataset KICH (Kidney Chromophobe) and KIRP (Kidney renal papillary cell carcinoma), that were also analysed by Garcia-Garcia (Garcia-Garcia et al. 2016) were selected from the cancer genome atlas project (https://cancergenome.nih.gov). Both datasets were paired data, that had data from controls and patients. For this study, the differential expression paired analysis done by Garcia-Garcia *et al.* was downloaded from github (https://github.com/dmontaner-papers/gsa4mirna): tables res_edger_paired_kich.csv and res_edger_paired_kirp.csv for KICH and KIRP datasets, respectively. This data consists of p-values and test

statistics on miRNA level. In the original publication the target prediction data was collected from TargetScan (validated targets) (R. C. Friedman et al. 2009) and the gene sets used were defined from Gene Ontology (GO) terms (biological processes, cellular components and molecular functions) and downloaded from http://ensembl.org.

Two datasets that were analysed also by Bleazard *et al.* (Bleazard, Lamb, and Griffiths-Jones 2015) were used. These datasets are from human samples and they have been published earlier (Raponi et al., 2009; Tanic et al., 2013) and are available in GEO (Gene Expression Omnibus, www.ncbi.nlm.nih.gov/geo/); GSE16025 and GSE44899, respectively. In the publication by Bleazard *et al.* they had selected one of the miRNA lists presented in the original papers and to be able to compare results in this study with the results in Bleazard *et al.* I used the same comparisons. From Raponi *et al.* (Raponi et al. 2009), miRNAs differentially expressed in comparison of lung squamous cell carcinoma (SCC) samples to normal lung were used. This miRNA list is shown in Table 2 of original publication. For BUFET analysis this list was used directly, but for the mdgsa analysis the data was analysed using geo2r from ncbi.nlm.nih.gov and Benjamini-Hochberg correction for p-value. Similarly for the Tanic *et al.* (Tanic et al. 2013), miRNA list from cluster 2 (Figure 3 and Supplementary table S3 of original publication) was used in Bleazard *et al*. and directly for BUFET analysis in this study, but for the mdgsa analysis the data was analysed with geo2r. In geo2r, BRCAX samples were compared with normal samples for finding the differentially expressed miRNAs and Benjamini-Hochberg correction was used for p-value. To convert the miRNA names to current mirBase version (v21) mirBaseConverter (v1.0.4, R/Bioconductor) and miRNAmeConverter (v1.6.0, R/Bioconductor) were used. Two of the miRNAs, hsa-miR-220 and hsa-miR-1286, in Raponi *et al.* dataset (Raponi et al. 2009) were not available in the current mirBase v21 version and were discarded from further analysis. In Bleazard *et al.* (Bleazard, Lamb, and Griffiths-Jones 2015) the target predictions were derived from the miRanda tool with the following settings: free energy < 20 kcal/mol and score >155. In the original study, the gene ontology data of biological processes for all human genes was downloaded from http://ensembl.org.

One dataset that was not used by Bleazard *et al.* (Bleazard, Lamb, and Griffiths-Jones 2015) or by Garcia-Garcia *et al.* (Garcia-Garcia et al. 2016) was also selected. This dataset has miRNAs that are differentially expressed during human plasma cell differentiation (Kassambara et al. 2017). The dataset has three miRNA clusters that are differentially expressed during plasma cell differentiation. These clusters are presented in Figure 2C of the original publication. All these three clusters were used for BUFET analysis. The full data including p-values and test statistics for each miRNA were available as supplementary data and this data for Cluster 1 was used for mdgsa analysis. In the original paper the target predictions were from miRTarget tool that integrates predictions from miRTarbase and miRecords (Kassambara et al. 2017). In the original publication, functional enrichment analysis was done by ClusterProfiler (R/Bioconductor) and Molecular Signatures Database v5 (MSigDB) canonical pathways were used as gene sets (Subramanian et al. 2005). miRNA names in the dataset were converted to mirBase v21 format by miRNameConverter (v1.6.0, R/Bioconductor). Altogether 5 of the miRNAs from the original dataset were not available in the newest version of mirBase and were discarded from the analysis. These miRNAs were hsa-miR-768-5p, hsa-miR768-3p, hsa-miR-886-5p, hsa-miR-1308 and hsa-miR-886-3p.

4.2 miRNA target prediction tools

4.2.1 mirDIP4.1

mirDIP unidirectional search v4.1 dataset was downloaded from the mirDIP4.1 webpage (March 27th 2018, http://ophid.utoronto.ca/mirDIP/). mirDIP4.1 database combines human miRNA-target prediction information from 30 different resources (Tokar et al., 2017), which are listed in

Table 4. The original version of mirDIP database was published in 2011 (Shirdel et al. 2011) and the database is actively updated (the latest update on September 2018). Databases used for mirDIP4.1 were selected so that they have been updated or published between 2006 and 2017. Target predictions from different resources have

been standardized and normalized and they use integrative score derived from the individual resources to classify the target predictions. In mirDIP4.1 dataset, predictions are ranked into four different classes: very high (top 1%), high (top 5%, excluding top 1%), medium (top 33%, excluding top 1%) and low (the rest of predictions). However, for analyses performed in this study, the mirDIP4.1 "High" dataset was filtered so, that it contains both "Very High" and "High" confidentially class and similarly, "Medium" dataset contains all interactions except those classified as "Low". Subsetting of the full mirDIP unidirectional search v4.1 dataset was done in R by column "SCORE_CLASS".

The mirDIP4.1 database has unique target interactions from 27667 unique genes and 2586 unique miRNAs (http://ophid.utoronto.ca/mirDIP/). The gene names in the mirDIP4.1 database are standardized according to the Hugo Gene Nomenclature Committee (HGNC, April 2017) and miRNA names are in the mirBase V21 format (www.mirbase.org). R/Bioconductor package miRNAmeConverter (version 1.6.0) was used to convert miRNA names from the used datasets into the correct format.

4.2.2 TargetScan

TargetScan algorithm was published originally in 2005 (Lewis, Burge, and Bartel 2005) and it has been updated actively since then (R. C. Friedman et al. 2009; Garcia et al. 2012; Grimson et al. 2007). Version 7.0 of TargetScan introduced a new improved algorithm that calculates context ++ scores for human and mouse target predictions (Agarwal et al. 2015). The context ++ score is calculated based on 14 different target site features and it is concentrating on target site matches on 3'UTR regions in human genes. TargetScan algorithm uses the conserved 8mer, 7mer and 6mer seed region matches to predict biological targets of miRNAs. TargetScan version 7.1 dataset is included in the mirDIP 4.1 dataset.

For this study, the dataset "Conserved site context++ scores" was downloaded that has all the conserved miRNA sites from www.targetscan.org (version 7.2, downloaded in October 18th 2018). This dataset has all the conserved miRNA sites and file has 1468778 rows. To choose only the rows that have human miRNA target

gene data, I selected the rows that have Homo Sapiens species id "9606". There were 265217 human miRNA-target gene interactions in this dataset.

4.2.3 miRanda

The miRanda algorithm was originally published in 2003 and it has been one of the most used algorithms for miRNA target prediction (Enright et al. 2003; John et al. 2004). However, the webpage for miRanda (www.microrna.org) has stopped working during May 2018 so the algorithm is not available for use any longer. miRanda target prediction is based on a seed match, conservation and the free energy of alignment. It is using the 3'UTR sequences of genes to find the miRNA target sites. Target predictions from miRanda algorithm were used in Bleazard *et al.* (Bleazard, Lamb, and Griffiths-Jones 2015) and they have also published the miRanda dataset as supplementary data. This miRanda dataset is version 3.3a filtered by free energy < -20 kcal/mol and score > 155. For analyses done in this study, this miRanda dataset was used.

4.2.4 mirTarget

miRTarget target prediction tool was used and developed by the authors of Kassambara *et al.* (Kassambara et al. 2017) from which one the used datasets was selected. For this study, miRTarget version 1.0.0 was downloaded from https://github.com/kassambara/miRTarget and used to get miRNA target predictions for Kassambara data to replicate the analysis done in the original publication (Kassambara et al., 2017). The full miRTarget dataset was also used to analyse the other datasets used in this study. miRTarget tool uses experimentally validated targets from the miRTarbase (http://mirtarbase.mbc.nctu.edu.tw/, release 6.0) and data from miRecords database (http://c1.accurascience.com/miRecords/, version: 27.4.2013), which contains also experimentally validated targets together with predicted targets. Target genes are selected as union of the two above mentioned databases and user can select how many of the 11 available target prediction tools in the miRecords database are used. In the original publication, only the experimentally validated targets were selected together with those targets that were predicted by at least 5 out of 11 target prediction tools available in miRecords. In this study, the

miRTarget full dataset was constructed with the same settings as in the original publication. miRecords webpage has been deprecated (tested 04/2019), but download of the dataset using mirTarget code was working when the analysis was done 04/2018. However, miRTarget is not a good choice for target prediction anymore as deprecation of miRecords database leads to outdated data eventually.

4.2.5 DIANA-microT database

DIANA-microT target predictions use 3'UTR regions of genes and mirBase annotated miRNAs (Manolis Maragkakis et al. 2011). Target predictions from DIANA-microT v4 algorithm were used in Bleazard et al. (Bleazard, Lamb, and Griffiths-Jones 2015) and they have also published microT dataset as supplementary data and in this file miRNA names have been converted to miRBase v21. For analyses done in this study, this dataset was used for finding target gene predictions for the lists of miRNAs.

4.3 Functional enrichment analysis tools

4.3.1 BUFET

BUFET (boosting the unbiased miRNA functional enrichment using bitsets) is a tool for miRNA functional enrichment analysis (Zagganas et al. 2017). It is an improved version of the method called EmpiricalGO that was developed by Bleazard *et al.* (Bleazard, Lamb, and Griffiths-Jones 2015). Both BUFET and EmpiricalGO algorithms are based on analysing several random miRNA lists that are the same size as the input miRNA list. The overlap of the target genes of these miRNA lists and genes of pathways is calculated and p-value for the input miRNA list is calculated from the proportion of random miRNA lists that produced an equal or greater pathway overlap. BUFET uses Benjamini-Hochberg correction for p-value. Iterations of 10 000 random sets of miRNAs are done by default, but this can also be defined by the user. EmpiricalGO is based on the use of hash tables for the overrepresentation analysis. This makes it more computationally expensive than the BUFET method, that uses bitsets instead of hash tables. BUFET is an open source python code, that can be run

on linux or unix operating system. BUFET was downloaded from https://github.com/diwis/BUFET/ and compiled as instructed. For running BUFET python version 2.7.11 was used.

BUFET can be used with pathway information from GO, KEGG or other user defined sources, so it is a flexible method. BUFET analysis has an option to be done using the miRNA target prediction from miRanda by choosing the –miRanda option, but this option can only be utilised if the user has installed miRanda locally before May 2018. However, any desired source of miRNA target prediction information can be used as a csv file in format miRNA_name|target_gene one pair per row. The miRNA names in the interactions file and miRNA input file need to be according to the same mirBase version. BUFET analysis needs the full miRNA-target gene interaction file to be able to do the permutations with the random lists of miRNAs.

Input files necessary for BUFET analysis are a list of miRNAs, a gene synonym file from for example NCBI, a gene annotation file (gene_name|pathway_ID|pathway _name) and an interaction file that lists the interactions between genes and miRNAs. In addition to these obligatory files, there are some options that can be changed by the user. For example, the number of iterations and other options such as the number of processors, the name of output file, species name if mouse (human as default) and option --ensGO if GO ontology data is downloaded from http://ensembl.org/, can be selected by the user. For BUFET analysis, I used the gene synonym data file from NCBI (http://ftp.ncbi.nih.gov/gene/DATA/GENE_INFO/Mammalia/All_Mammalia.gene_info.gz), pathway data (Gene ontology (GO)) from http://ensembl.org with the --ensGO option and other pathway data modified to the correct format. To download the GO data table from http://ensembl.org, necessary columns were selected based on instructions given on BUFET manual (https://github.com/diwis/BUFET/).

4.3.2 Mdgsa

Mdgsa algorithm is for gene set enrichment analysis. Mdgsa is a R/ Bioconductor package and it was developed by Garcia-Garcia *et al.* (Garcia-Garcia et al. 2016). For

this study, version 1.10.0 of Mdgsa was used. It is based on transfer indexes that are calculated from the test statistics and p-values. The sign of test statistics (fold change) gives the direction of the change and p-value the strength of differential expression of miRNA. This index (r) is calculated for each miRNA as:

r = -sign(statistic)*log(p-value)

These indexes for each miRNA are used to calculate transfer index for each gene based on the information from differentially expressed miRNAs and miRNA target information. In this calculation, the indexes of all miRNAs targeting the same gene are summarized together. This means that the information of differentially expressed miRNAs is transferred to the gene level and the genes can be ranked according to their transfer indexes. This allows analysis to account for the additive effects of different miRNAs, since several miRNAs can target the same gene. Genes that have indexes close to zero are not showing regulation by miRNAs. If the mRNA expression data of the same samples is also available, the transfer index calculation can be modified so that only those genes that are expressed in the sample are considered in the calculations. The gene set enrichment analysis is then done for the gene list ranked by transfer index using logistic regression.

Mdgsa analysis needs the miRNA differential expression data with test statistics and p-values, miRNA-gene interaction information and gene annotations from GO or other sources. For mdgsa analysis, I used raw p-values to calculate indexes similarly as in Garcia-Garcia *et al.* (Garcia-Garcia et al. 2016).

Mdgsa analysis was done with default settings in which the pathway size is limited between 10 and 500 genes and with modified settings with pathway size from 5 to 500 genes.

4.3.3 ClusterProfiler

R/Bioconductor package ClusterProfiler has been used in several publications to perform the functional enrichment analysis of miRNA data (Yu et al. 2012). It was used in the analysis of one of the datasets selected for this study (Kassambara et al.

2017) and in the mirDIP4.1 publication (Tokar et al. 2017) and thus selected as an example of traditional functional analysis algorithm (v 3.6.0 used for this study). ClusterProfiler can perform overrepresentation analysis by hypergeometric distribution or gene set enrichment analysis with a permutation test. For this study, overrepresentation analysis by ClusterProfiler was used. The algorithm uses Bioconductor annotation data GO.db and KEGG.db as a source of gene sets, but it can be used with other ontologies as well. For other pathway data sources, data frames need to be modified so that one data frame has term ID and gene information and the other data frame term ID and term name information. These data frames are then used as a source of pathway information for analysis. The enrichGO function calculates the enrichment for GO terms for a given list of genes by overrepresentation analysis. This package has also functions for visualizing the enrichment results. For this study, I used ClusterProfiler algorithm with Bioconductor annotation data GO.db and with Molecular Signatures Database data.

4.4 Gene set data

Two different source of canonical pathways data for functional analyses were used. The first set was gene ontology (GO) data downloaded from Ensembl using their BioMart tool (www.ensembl.org/biomart). Version 94, which has the data for human genome assembly 38 (GRCh38.p12), was used in this study. To be able to use this dataset directly for BUFET analysis the following attributes were downloaded in this order: Ensembl Gene ID, Ensembl Transcript ID, Associated Gene Name, GO Term Accession, GO Term Name, GO Term Definition and GO Domain. The dataset was further divided by the GO domain to use only the Biological processes (BP) GO terms in the analysis by BUFET and BP, Molecular functions (MF) and Cellular components (CC) separately for the mdgsa analysis. The other canonical pathways dataset was Molecular Signatures Database v6.1 (MSigDB) canonical pathways data (Subramanian et al. 2005), which was downloaded from http://software.broadinstitute.org/gsea/msigdb. For ClusterProfiler, GO.db from Bioconductor/R and MSigDB were used.

5 Results

The aim of this study was to develop an analysis work flow for miRNA functional enrichment analysis and test its functionality. The functional analysis of miRNAs includes the prediction of miRNA target genes followed by gene set enrichment analysis or overrepresentation analysis to find the affected pathways. Different target prediction tools were compared together with three different functional analysis algorithms. For this study, five different miRNA datasets were selected to be used for the comparison of target prediction tools and functional analysis tools.

5.1 Target prediction

Target gene prediction for miRNAs is an important step in the analysis of miRNA data and the confidentiality of miRNA target predictions also affects the reliability of the results of functional analysis. There are numerous tools and databases for miRNA target predictions, which are based on different algorithms and different biological properties of miRNAs. Different algorithms and their biological background were described in detail in Section 2.3.

To select the target prediction tools for this analysis, different tools and databases for miRNA target predictions were searched from the literature. Because miRNA target prediction tools are based on different biological properties of miRNAs, the results of different algorithms are not uniform. For this study, I used target predictions from some individual resources to compare the results obtained in this study to the original studies. In addition to the individual target prediction tools, one aim for this study was to find a database that collects and combines data from many different target prediction tools. A target prediction tool that combines data from many different sources may be more biologically relevant since miRNA target gene interactions that are found by many different target prediction algorithms are more reliable than interaction predictions that are based on a single algorithm. A table of integrative tools for miRNA target prediction is presented in Table 3. From these tools, mirDIP4.1 and RAIN are the only ones that are using integrative scores to combine the data from their individual sources and other databases are only

collecting individual data from different tools and the user can filter the data based on how many and/or which individual tools have predicted or have experimentally validated data for the same interaction. RAIN has 8 target databases as its source information, whereas mirDIP4.1 is collecting data from 30 different databases. Both databases can be used as web-based tools or then the full database can be downloaded for local searches or to be integrated into a work flow. RAIN has target predictions for human, mouse, rat and yeast, whereas mirDIP4.1 has data only for human miRNA gene interactions.

Table 3. Integrative tools for miRNA target prediction.

Integrative tool	species	Target prediction databases	in silico	experimentally validated	type	Special	Reference
mirDIP4.1	human	**30**	30	0	web, downloadable dataset, API*	integrative score for target prediction , benchmarked with 2 experimentally validated databases (Tarbase v.7.0 and Npinter v.3.0)	Tokar *et al.* (2017)
miRGate	human, mouse, rat	**9**	5	4	web, API	only 3'UTR targets, all isoforms, pseudogenes, non coding genes	Andres-Leon *et al.(*2015)
miRecords	human, mouse, rat, drosophila, roundworm (*C. elegans),* zebrafish, chicken, sheep, dog	**11**	11	1	web, experimental database downloadable	last update 2013, two parts: target predictions from 11 tools and other database with expreimentally validated targets	Xiao *et al.* (2009)
miRWalk2.0	human, mouse, rat	**16**	12	4	web	miRNA targets of CDs, 3'UTR, 3'UTR, promoter	Dweep & Gretz (2015) , Dweep *et al.* (2011)
multiMiR	human, mouse	**11+3***	8	3	R package	many target filtering options, drug/disease-related miRNA databases	Ru *et al.* (2014)
miRGator v3.0	human	**9**	6	3	web	miRNA expression profiles from diseases, organs and tissues; miR-seq browser for alignment and visualization of reads from NGS data	Cho *et al.* (2013)
miRRor-suite	human, mouse, rat, drosophila, roundworm (*C. elegans*), zebrafish	**12**	12	0	web	user can select which databases to use	Friedman *et al.* (2014)
RAIN	human, mouse, rat, yeast	**8**	5	3	web, downloadable dataset	integrative score	Junge *et al.* (2017)
RegNetWork	human, mouse	**8**	5	3	web	Database has TF-TF, TF-gene, miRNA-gene, TF-miRNA and miRNA-TF interactions. Data collected from 25 databases.	Liu *et al.* (2015)

***** API available from September 2018, API = application programming interface, web = web tool, NGS= next generation sequencing, UTR= untranslated region, CD= coding region, TF= transcription factor

(Andrés-León, Núñez-Torres, and Rojas 2016; Cho et al. 2013; Dweep et al. 2011; Dweep and Gretz 2015; Y. Friedman, Karsenty, and Linial 2014; Junge et al. 2017; Liu et al. 2015; Ru et al. 2014; Tokar et al. 2017; Xiao et al. 2009)

Although mirDIP4.1 has only data from human miRNA target predictions, it is still very useful since many researchers are interested in data from human samples. miRecords database has the widest selection of different species having target prediction data for 9 different species from 11 target prediction databases and from 1 experimentally validated dataset. Experimentally validated miRNA target gene interaction data is still

restricted and not available for all miRNAs or cell types. However, combining information on experimentally validated and predicted target interactions is a good approach and can lead to more reliable miRNA target gene data. Although mirDIP4.1 database does not contain data from any database that has only data from experimentally validated targets, it contains data from resources that include both experimentally validated targets and target predictions. In addition, it collects data from the widest selection of individual resources and was therefore selected to be used for this study. The mirDIP unidirectional search database v4.1 was downloaded and used locally to search for the target predictions. In addition to mirDIP4.1 database, other target prediction tools targetScan, miRanda, microT and miRTarget were used to compare the results obtained by different target prediction algorithms and tools.

5.1.1 mirDIP4.1

The mirDIP4.1 integrative tool (Tokar et al. 2017) is an updated version of the mirDIP database, that was originally published in 2011 (Shirdel et al. 2011). mirDIP4.1 integrates data from 30 different prediction tools and uses integrative score to classify the predictions.

Table 4 lists the individual resources and their versions that have been used to build this database. Depending on the source, there is a variable number of genes, miRNAs and predictions.

Table 4. Sources of miRNA target prediction data used in mirDIP4.1 database. Number of predictions, genes and miRNAs. (Table adopted from Tokar et al. 2017)

Resource	Version/Date	Predictions	Genes	miRNAs
BCmicrO	March, 2017	10 682 301	18 418	580
BiTargeting	April, 2017	5 314 760	18 517	2582
CoMeTa	March, 2017	640 586	10 969	643
Cupid	March, 2017	298 163	8411	1181
DIANA	v5.0	7 112 061	18 529	1909
ElMMo3	March, 2017	2 837 861	18 179	997
GenMir++	March, 2017	5579	872	99
MAMI	March, 2017	95 408	14 285	309
MBStar	April, 2017	11 925 118	18 041	2031
microrna.org	January, 2008	684 192	18 424	241
MirAncesTar	March, 2017	36 116 591	18 532	2568
mirbase	March, 2017	498 128	17 913	684
miRcode	March, 2017	997 836	25 656	124
mirCoX	March, 2017	1 716 865	21 749	79
miRDB	v5.0	4 739 198	16 588	2571
MirMAP	v.1.1	11 392 502	18 574	2031
MirSNP	March, 2017	849 897	17 180	1909
MirTar	March, 2017	686 222	16 556	1897
miRTar2GO	March, 2017	1 164 371	10 890	366
Mirza-G	April, 2016	4 348 927	16 790	2564
MultiMiTar	March, 2017	429 258	10 986	473
PACCMIT	February, 2012	363 717	11 735	1905
PicTar	March, 2017	14 160	2430	114
PITA	v6.0	685 848	18 141	295
RepTar	March, 2017	2 996 265	17 280	1066
RNA22	v.2.0	3 127 672	1927	2584
RNAhybrid	v2.1.2	41 306 832	17 448	2584
TargetRank	March, 2017	342 703	14 241	525
Targetscan	v7.1	210 146	11 952	369
TargetSpy	April, 2016	286 654	15 485	356

Seven of the 30 individual resources, that have been used to construct the mirDIP4.1 database, combine data from experimentally validated targets and target predictions. These resources are BCmicrO, Cupid, DIANA, MBStar, miRDB, MirMAP and miRTar2GO. The experimentally validated interaction data of all these seven resources come from cross-linking immunoprecipitation (CLIP), which is a method used for miRNA binding site detection.

To unify the data from individual resources, the gene symbols have been standardized according to Hugo Gene Nomenclature Committee (HGNC) and miRNA names according to miRBase v21. Because of this standardization, I used R/Bioconductor package miRNAmeConverter to convert miRNA names from the datasets used in this study to miRBAse v21. This converter can change miRNA names to their current mirBase v21 names or to some other mirBase version based on selection by user.

For the construction of the mirDIP4.1 database, only resources that had the evaluation of target prediction by some type of quantitative measure have been selected. These quantitative measures of confidence of interaction included.

statistical significance, binding energy or a score. For each individual target prediction set, predictions were first normalized by ranking them based on their prediction confidentially from 0 to 1, where 0 was given to the most confident prediction of the set. If prediction dataset contained multiple predictions for the same miRNA-target gene pair, these were replaced by one prediction. For this single prediction, its rank was calculated as a product of the three most confident ranks (the lowest ranks) of individual predictions. Thus, genes that had multiple binding sites on the same miRNA were favoured and finally only one prediction for each miRNA-target gene pair in each individual dataset was used for the final mirDIP4.1 database.

For integrating the results from individual resources, mirDIP4.1 uses a benchmarking approach to allow a direct comparison of the confidence of individual miRNA-target prediction from different resources. As a benchmarking dataset, experimentally validated miRNA-target interactions from TarBase v.7.0 (Vlachos et al. 2015) and NPinter v.3.0 (Hao et al. 2016) were used. For each individual prediction dataset, the precision of target prediction was calculated using benchmark dataset and precision together with its associated ranks were used to form a function that was then applied to interpolate the precision of each prediction in that dataset. Using this approach, confidence score (S_{ij}) was assigned for each prediction of individual resources. This allows direct comparisons of prediction confidence of different resources. These confidence scores were further divided into four confidence classes; very high, high, medium and low. Obtained confidence scores where then used to calculate integrative score for each miRNA-target interaction:

$$S_j = 1 - \prod_i (1 - s_{ij}),$$

In this equation, s_{ij} is the confidence score of j-th miRNA-target interaction from i-th resource.

mirDIP4.1 can be used from the web interface to perform searches on either miRNA name list or gene list to find the target genes or targeting miRNAs, respectively. This type of search is called unidirectional search and can be filtered to desired confidence class. The results can be exported as csv or txt file. mirDIP4.1 search can also be

performed in bidirectional mode, in which both miRNA and gene list are provided and the search is restricted to the target genes that are present in the gene list. This can be useful, if both miRNA and mRNA expression data are available from the experiment. In such case, the number of target genes can be restricted to those that are expressed in the sample. Different versions of the mirDIP4.1 database can also be downloaded to be used locally. For this study, I downloaded the mirDIP unidirectional search v4.1 database to be used as my target prediction data. The number of interactions on different confidence classes from this dataset are shown on Table 5. Confidence class very high consists of top 1% of interactions, class high top 5% (excluding top 1%), class medium top 1/3 (excluding top 5%) and class low has all the remaining interactions.

Table 5. Classification of miRNA target gene interactions in mirDIP4.1 database.

Confidence class	Rank	No of interactions
Very high	top 1%	486572
High	top 5%	1946285
Medium	top 33%	30005226
Low	rest	16219050
	total	**48657133**

For analyses done in this study, mirDIP unidirectional search v4.1 database was divided into three parts that were used as individual prediction resources for analyses. For functional analysis, datasets contained interactions of confidence class “Very High” (mirDIP very high) or interactions of classes “Very High” and “High” (mirDIP high). For the comparison of target prediction tools, mirDIP4.1 dataset was divided according to confidence classes to mirDIP “Very High”, mirDIP “High” and mirDIP “Medium” and these datasets have a number of interactions as shown in Table 5.

When analyses for this study were done, mirDIP4.1 was only available as a web interface or downloadable dataset, but from September 2018, it has also been provided as API to be integrated into analysis work flow.

5.1.2 Comparison of target prediction tools

To analyse how the results of mirDIP4.1 and other tools for target prediction used in this study differ, the same miRNA lists were used to search for target predictions from all of these tools. The miRNA datasets that were used in this study are listed in Table 2. Full target prediction datasets were downloaded from mirDIP4.1 and targetScan web pages. R/Bioconductor package miRTarget uses the target predictions from the miRTarbase and miRecords. Both datasets were downloaded to be used locally. miRTarbase has the data for experimentally validated miRNA targets and miRecords is a dataset of predicted miRNA target genes. These datasets were combined to get the full dataset used by miRTarget. miRanda dataset was downloaded from the supplementary data of Bleazard *et al.* (Bleazard, Lamb, and Griffiths-Jones 2015). First, I compared how the number of target genes by each prediction tool differ from each other (Figure 7). Only mirDIP “Very high” and mirDIP “High” target predictions from the mirDIP4.1 dataset were used, since the number of mirDIP4.1 target genes for mirDIP “Medium” dataset for a list of 12 miRNAs was already 224614. In addition, since mirDIP “Medium” dataset contains 1/3 of all interactions in this database, it is not biologically the most meaningful dataset to be used for functional analysis. The full table that was used to create Figure 7 with addition of results from mirDIP4.1 “Medium” interactions is shown in Appendix 1.

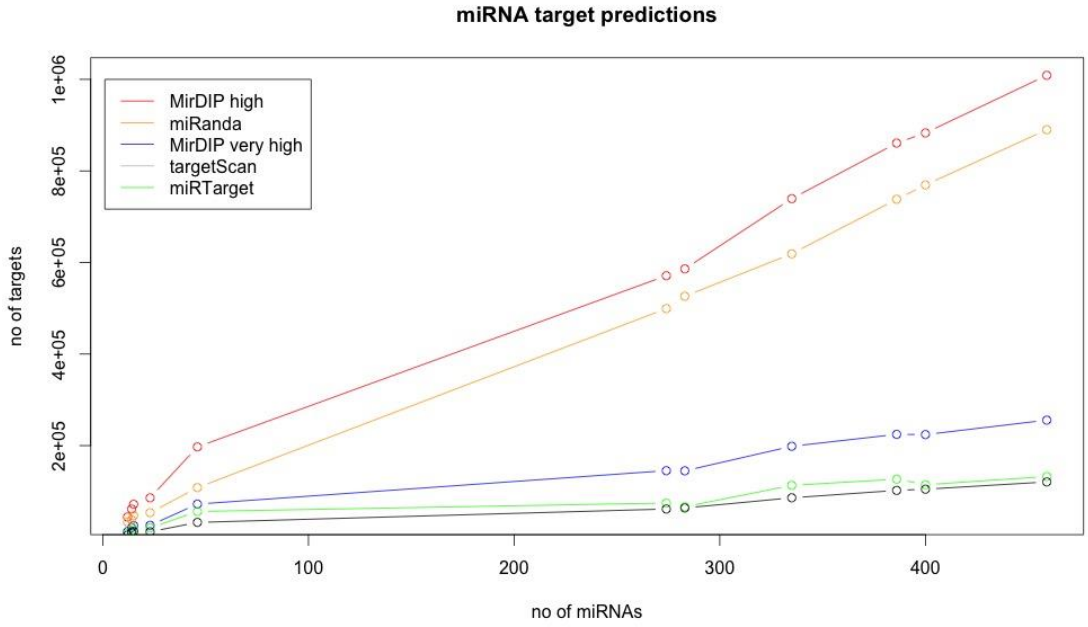

Figure 7. Number of miRNA-target gene interactions predicted by target prediction tools differ greatly. Lists of target genes for miRNA lists from the different datasets used in this study were extracted using R. Target gene lists were used without modifications, so same gene can be listed more than once, if it is affected by many different miRNAs. miRNA lists (number of miRNAs in list) in ascending order: Kassambara cluster3 (12), Kassambara cluster 2 (14), Raponi et al. (15), Kassambara cluster1 (23), Tanic et al. (46), KICH p<0.05+ |FC| >2 (274), KIRP p<0.05+|FC| >2 (283), KICH p<0.01 (335), KICH p<0.05 (386), KIRP p <0.01 (400) and KIRP p<0.05 (459).

As can be seen from the Figure 7, the size of miRNA list is proportional to the number of targets, which was expected. Two of the target prediction datasets, mirDIP "High" and miRanda had remarkably more targets predicted for the same lists of miRNAs than the other tools. For KIRP p<0.01 miRNA list mirDIP high predicted 8.4-fold more target genes than miRtarget and miRanda 7.3-fold more. On the other hand, for the shortest miRNA lists the number of predictions varied a lot and mirDIP high had 10.1-fold more targets than miRtarget for Kassambara3, but only about 5-fold more than miRTarget for Kassambara2 or Raponi miRNA lists. Similarly, miRanda predicted 7.6-fold more targets than miRTarget for kassambara3, but 3-4-fold more for Kassambara2 or Raponi miRNA lists. Therefore, it seems that the number of predicted target genes is highly dependent on which specific miRNAs are on the list. targetScan and miRTarget gave the most similar numbers of targets for all lists of miRNAs.

mirDIP4.1 integrates data from 30 different databases (

Table 4), but of these tools only targetScan was used individually in this study. However, since mirDIP4.1 is using integrative score to combine data from different sources, target gene lists by targetScan and mirDIP4.1 are not uniform. To compare the target gene predictions of these 4 different tools, I selected two different miRNA lists, Kassambara1 and Raponi to compare the predicted targets in more detail. The Kassambara1 list has 23 miRNAs and Raponi list has 15 differentially expressed miRNAs. To compare the differences of mirDIP4.1 data of confidence class "Very High" and "High" with the other datasets, I filtered the data to separate "Very High" and "High" interactions. Number of unique target genes from these mirDIP datasets together with data from miRanda, miRTarget and targetScan target predictions are shown in Table 6.

Table 6. Number of unique target genes.

Target prediction tool	Raponi	Kassambara1
mirDIP_VH	7294	7917
mirDIPHigh	12326	14263
targetScan	4931	5577
miRanda	7828	10068
miRTarget	3123	4926

mirDIP4.1 dataset filtered for "High" confidentially predicts more than 12 000 interactions for the lists of 15 or 23 miRNAs. This number of targets is very high, if the estimated total number of genes in human genome is around 20 000 and number of miRNAs about 2600 (mirBase). To illustrate the differences and similarities of these miRNA target predictions, I used R/CRAN package VennDiagram (version 1.6.20) to create venn diagrams. Data from Raponi miRNA list is shown in Figure 8 and data from Kassambara1 is shown in Figure 9. Target gene list obtained from miRTarget is the shortest one with both Kassambara1 and Raponi datasets (Figure 7, appendix 1)

and thus it is expected to show the lowest number of unique target genes compared with other target prediction datasets. For both miRNA lists, mirDIP high data, which has the highest number of target genes, has more than 3300 unique targets that are not listed in other datasets. In addition, miRanda has almost thousand target genes that are not present in any of the other datasets.

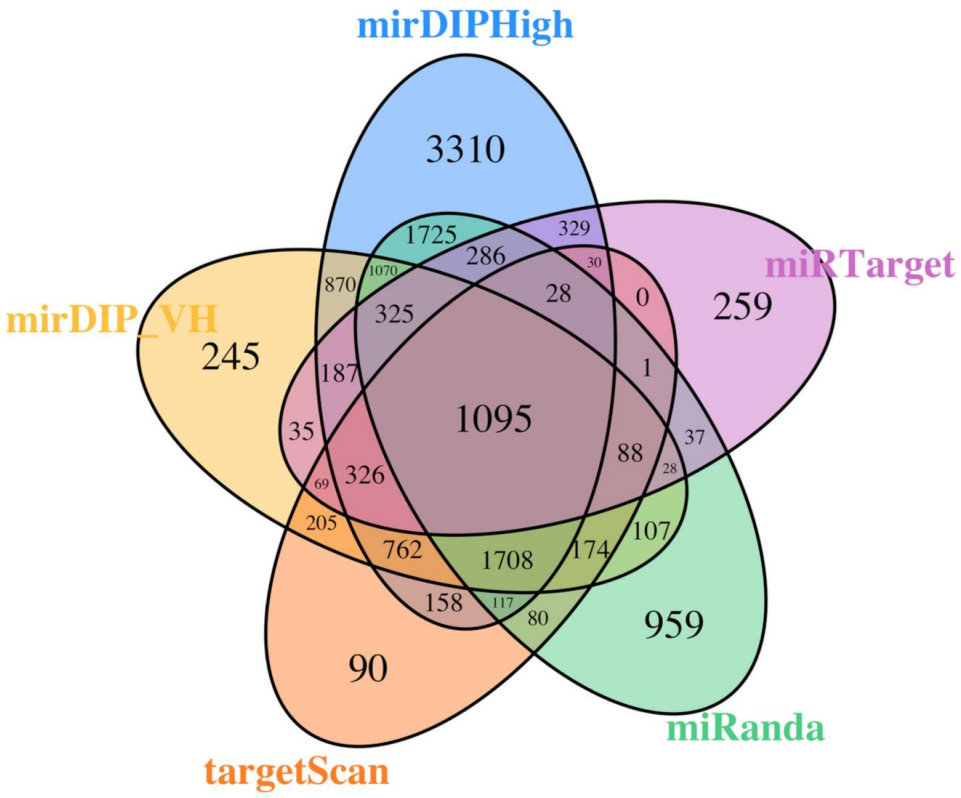

Figure 8. Venn Diagram of target predictions from Raponi et al. data. Raponi et al. dataset has 15 miRNAs. Their target genes were searched from targetScan, miRTarget, microT, miRanda and mirDIP datasets. From the mirDIP dataset mirDIP4.1 dataset filtered for confidence class "Very High" (mirDIP_VH) and "High" (mirDIPHigh) were used.

In the Raponi dataset, there were 1095 common target genes. It is surprising, that target predictions from mirDIP with confidence classes "Very High" and "High" have 6343 common genes, which is half of all the unique genes in the "High" dataset and 87% of genes in "Very High" dataset. As this comparison is only for the list of target genes that are predicted for a list of miRNAs, there is possibility that "Very High" and "High" classified interactions are from different miRNAs.

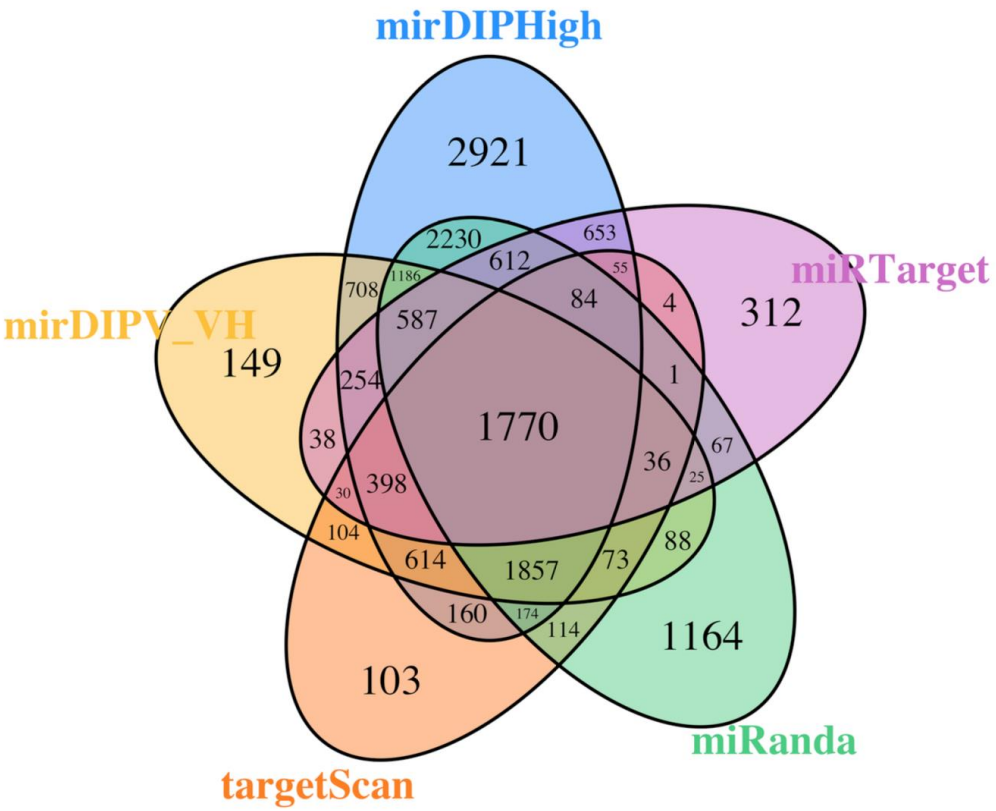

Figure 9. Venn diagram of target predictions from Kassambara cluster 1. Kassambara1 dataset has 23 miRNAs. Their target genes were searched from targetScan, miRTarget, microT, miRanda and mirDIP datasets. From the mirDIP dataset mirDIP4.1 dataset filtered for confidence class "Very High" (mirDIP_VH) and "High" (mirDIPHigh) were used.

The venn diagram of Kassambara1 dataset is similar to that from Raponi dataset. Kassambara1 miRNA list has 23 miRNAs and there were 1770 unique target genes, that were found by all target prediction tools. Similarly as for Raponi dataset, mirDIP "Very High" and "High" targets had a great overlap. There were 7394 common target genes between these two datasets, which is half of the unique genes in mirDIP "High" dataset and 93% of the genes in mirDIP "Very High" dataset. mirDIP "High" and miRanda have the highest number of target genes that are not found by any of the other prediction tools, but this is also expected as these two prediction sets have the highest total number of interactions.

Target prediction for miRNA data is important, but also challenging step of analysis. A list of only 12 miRNAs gives from 4358 (miRTarget) to 44058 (mirDIP high+very high) or even to 224614 (mirDIP medium) target gene interactions. Without complete experimentally validated target gene data for each miRNA, it is difficult to estimate what the correct number of target genes for each miRNA list is. However, mirDIP with medium confidentially predicted so high number of target gene interactions, that it is most probably not biologically relevant. Thus, target predictions from mirDIP "Medium" confidentially class were discarded from the further analyses.

5.2 Functional enrichment analysis

Overrepresentation analysis for miRNA data has been mainly performed by statistical analysis of pathway associated gene sets that are overrepresented in the genes targeted by miRNA list. Statistical analysis has been performed usually by Fisher's exact test. The traditional way of functional analysis of miRNA data has been criticized of being biased to certain biological pathways (Bleazard, Lamb, and Griffiths-Jones 2015).This bias is explained in more detail in section 2.4.3.

To avoid this bias, two different algorithms for functional analysis were selected based on recent literature. The first one, BUFET (Zagganas et al. 2017), is an over-representation analysis tool that improves the accuracy of p-values for functional enrichment by calculating empirical p-value from the pathway overlap of random miRNA lists that is higher than for the input miRNA list. The other algorithm selected for this study, mdgsa, is based on gene set enrichment analysis (Garcia-Garcia et al. 2016). Differential expression statistics of miRNAs are transferred to gene level so that both the direction of change (the sign of fold change) and strength of change (p-value) are included. Genes are then ranked according to their regulation by miRNAs and gene set enrichment is analysed by logistic regression.

Several different datasets were used to evaluate the functionality of BUFET and mdgsa algorithms for the functional analysis of miRNA data. Tanic and Raponi datasets were analysed in the original study of empirical algorithm from which BUFET

is a modified version (Bleazard, Lamb, and Griffiths-Jones 2015). In the publication describing BUFET only random miRNA lists were used, and thus data from this article was not useful for this analysis (Zagganas et al. 2017). In the original publication of the mdgsa algorithm, different cancer datasets were used to test the functionality of the algorithm (Garcia-Garcia et al. 2016). For this study, two of these datasets were selected, KICH and KIRP. One dataset unrelated to neither of the algorithms was also selected. This dataset consists of miRNAs differentially expressed during human plasma cell differentiation (Kassambara et al. 2017) and in the original paper ClusterProfiler package for overrepresentation analysis was used. Overrepresentation analysis by ClusterProfiler is based on hypergeometric distribution and more traditional way of functional analysis of miRNAs. All these datasets were used for BUFET and mdgsa analyses and miRNA lists from the Kassambara *et al.* paper were also analysed with ClusterProfiler. To compare mirDIP "Very high" and "High" target prediction with the target predictions of the original publications of these datasets, analyses were performed with different target prediction datasets.

5.2.1 Functional enrichment analysis with BUFET algorithm

To be able to compare the results obtained from functional analyses done for this study with the original data, I did analyses with combinations of different target prediction data and pathway annotations. Summary of analyses done with BUFET is shown on Table 7.

Table 7. Summary of analyses performed with BUFET.

Dataset	Predictions	Pathway data
Kassambara 1,2,3	miRTarget*	MSigDB*, GO
	mirDIP high, very high	MSigDB*, GO
	targetScan	MSigDB*, GO
	miRanda	MSigDB*, GO
Tanic	miRanda*	GO* (BP)
	targetScan	GO* (BP), MSigDB
	mirDIP high, very high	GO* (BP)
	microT	GO* (BP)
Raponi	miRanda*	GO*
	mirDIP high, very high	GO*
	targetScan	GO* (BP), MsigDB
KICH	targetScan*	GO*, MSigDB
	mirDIP high, very high	GO*
	miRanda	GO*
	microT	GO*
KIRP	targetScan*	GO*, MSigDB
	mirDIP high, very high	GO*
	miRanda	GO*
	microT	GO*

*original source of predictions and pathway data. GO = full set of gene ontology data including biological processes, cellular compartments and molecular functions. GO (BP) = gene ontology data of biological processes. MSigDB= Molecular Signatures Database pathways.

For the Kassambara data (Kassambara et al. 2017), each of three miRNA lists (1,2 and 3) was analysed separately. Data for KICH and KIRP cancers were in original publication (Garcia-Garcia et al. 2016) analysed by gene set enrichment analysis and thus not filtered. For this analysis, these datasets were filtered by fold change and p-value. Filtering was done with three different thresholds: 1) p-value < 0.01, 2) p-value < 0.05 and 3) |fold change| >2 + p-value < 0.05. As a summary, BUFET analyses were performed with 11 different miRNA lists using predictions from three to four different tools. Pathway data (Gene sets) were downloaded from http//ensembl.org/biomart (GO = gene ontology data) or from Molecular Signatures Database (MSigDB). Gene ontology data was further filtered to biological processes data, which was used in part of the analyses.

The default setting for the number of iterations performed by BUFET is 10 000. For some of the analyses, I used 100 000 or 1 000 000 iterations as well, but the number of iterations did not change the results. Thus 10 000 iterations seem to be sufficient. Most of the analyses done with BUFET did not result in any statistically significant pathways. There were only 14 analyses that resulted in any pathways with significant p-value. These analyses and the number of GO terms with p-value < 0.05 are listed in Table 8. For these analyses, I used the gene sets from gene ontology, either the full list of GO terms or biological processes GO terms.

Table 8. BUFET analyses with significant GO terms.

Dataset	no of miRNA	Target predictions	no of targets	GO terms	Significant GO terms	GO terms >5 genes
Tanic	46	miRanda	108234	full	4045	529
				BP	4048	532
Tanic	46	microT	33658	full	2	2
				BP	3	3
Raponi	15	miRanda	47150	full	1	0
KICH FC2, p-value <0.05	274	miRanda	499308	full	8339	1665
		microT	114622	full	6646	935
KICH p-value <0.01	335	miRanda	618998	full	8624	935
		microT	147482	full	7009	799
KICH p-value <0.05	386	miRanda	738057	full	8888	2000
		microT	169855	full	7533	1286
KIRP FC2, p-value <0.05	283	mirDIP high	586054	full	1	1
KIRP p-value <0.01	400	mirDIP high	883026	full	2	2
KIRP p-value <0.05	459	mirDIP high	1008752	full	5	5

Analyses with significantly enriched GO terms had target predictions from miRanda and microT for Raponi, Tanic and KICH datasets. For the KIRP dataset, the target predictions were from mirDIP4.1 high, which contains all the interactions that are classified as “High” or “Very High”. Another interesting point in this data is that the number of significant GO terms is either very few or then very high. Partly the high number of GO terms with significant p-value in some of the analysis is because there is a huge number of pathways that have 5 or less genes in them. These terms are the

majority of significant GO terms. In addition, a list of 46 miRNAs (Tanic) affects 4045 pathways and a list of 300-400 miRNAs affects up to almost 9000 pathways. It is difficult to estimate which of these pathways are biologically significant.

5.2.2 Functional enrichment analysis with mdgsa algorithm

Similarly, as for BUFET algorithm, I used different combinations of miRNA datasets and target predictions as well as both GO and MSigDB pathway annotations to perform analyses with the mdgsa algorithm. Summary of the performed analyses is shown in Table 9.

Table 9. Summary of analyses performed with mdgsa algorithm.

Dataset	Predictions	Pathway data
Kassambara 1	miRTarget*	MSigDB*, GO
	mirDIP high, very high	MSigDB*, GO
	targetScan	MSigDB*, GO
Tanic	miRanda*	GO* (BP, MF, CC)
	targetScan	GO* (BP, MF, CC), MSigDB
	mirDIP high, very high	GO* (BP, MF, CC), MSigDB
Raponi	miRanda*	GO* (BP, MF, CC)
	targetScan	GO* (BP, MF, CC), MsigDB
	mirDIP high, very high	GO* (BP, MF, CC), MSigDB
KICH	targetScan*	GO*(BP, MF, CC)
	mirDIP high, very high	GO* (BP, MF, CC)
	miRTarget	GO* (BP, MF, CC)
KIRP	targetScan*	GO* (BP, MF, CC)
	mirDIP high, very high	GO* (BP, MF, CC)
	miRTarget	GO* (BP, MF, CC)

*original source of predictions and pathway data. GO = full set of gene ontology data including biological processes, cellular compartments and molecular functions. GO (BP) = gene ontology data of biological processes, GO (MF) = gene ontology data of molecular functions, GO (CC) = gene ontology data of cellular components. MSigDB= Molecular Signatures

For mdgsa analysis, only one of the miRNA lists in the Kassambara data (Kassambara et al. 2017) Kassambara cluster 1 was used. The other datasets were the same as for BUFET analyses. miRNA lists selected from the original mdgsa publication, KICH and KIRP, where downloaded from the supplementary files of mdgsa article (Garcia-Garcia et al. 2016) and used for analyses in this study. For Tanic and Raponi data the full datasets with p-values and fold changes were not available from the article and therefore, I downloaded the original data for these studies from GEO (Gene Expression Omnibus, www.ncbi.nlm.nih.gov/geo/) and analysed it with NCBI's geo2r tool to get the full list of differentially expressed miRNAs with test statistics. In the Tanic data, BRCAX samples were compared with normal samples and in Raponi data, SCC (squamous cell carcinoma) samples were compared with normal lung samples to calculate fold changes and p-values for each miRNA in the data. These comparisons were the same as which were analysed by Bleazard *et al.* (Bleazard, Lamb, and Griffiths-Jones 2015). Both Tanic and Raponi data are from microarray analyses and in both analyses an array with 325 miRNAs have been used (Raponi et al. 2009; Tanic et al. 2013). Table 10 shows the number of miRNAs in each dataset that was used and how many of the miRNAs in each list had target genes in different miRNA-target interaction datasets.

Table 10. Number of miRNAs with target genes in different target prediction datasets.

		miRNAs with targets				
Dataset	**no of miRNAs**	**mirDIP very high**	**mirDIP high**	**TargetScan**	**miRanda**	**miRTarget**
Kassambara 1	815	799	815	302	na	815
Tanic	325	263	266	132	300	na
Raponi	325	300	300	216	300	na
KICH	609	603	603	270	na	na
KIRP	698	692	692	287	na	na

na= not analysed

TargetScan dataset used for this study consists of conserved miRNA target sites and it is smaller than the other target prediction datasets used in this study. miRanda and miRTarget target predictions were included for those datasets that were analysed by using these target predictions in the original publications. Both mirDIP datasets, very high and high, have targets for most of the miRNAs in the analysed lists. Kassambara

1 dataset was the only one to be analysed with miRTarget target predictions and the miRTarget algorithm found targets for all the miRNAs in the analysed list. miRTarget algorithm was developed and used in the article by Kassambara *et al.* (Kassambara et al. 2017) from which the dataset is from.

Although the number of miRNAs with target genes in the target prediction data varies, this does not seem to correlate directly with the number of enriched pathways found by analysis. The results of analyses performed with mdgsa from GO terms of biological processes (GO BP) is shown in Table 11. The results from GO terms of molecular functions and cellular components and analyses done with pathway data from MSigDB were similar to these results. They are shown in Appendix 2. The default setting for the mdgsa analysis is to limit the pathways to those that have 10-500 genes. As there is a huge number of pathways that have less than 10 genes, I ran the analysis also by limiting the pathways to those that have 5-500 genes. In contrast to BUFET analyses, most of the mdgsa analyses resulted in the enrichment of one or more significant pathways with all the studied datasets. Enrichment of significant pathways was also seen with all target prediction datasets used.

Table 11. Results of mdgsa analyses with GO Biological processes pathways.

Dataset	Target prediction tool	no of genes in pathway	pathways with p-val <0.05[#]
Kassambara1	mirDIP H	5-500	4
	mirDIP H	default**	1
	mirDIP VH	5-500	11
	mirDIP VH	default	1
	targetScan	5-500	3
	targetScan	default	0
	mirTarget*	5-500	3
	mirTarget*	default	1
KICH	mirDIP H	5-500	19
	mirDIP H	default	15
	mirDIP VH	5-500	5
	mirDIP VH	default	1
	targetScan*	5-500	14
	targetScan*	default	8
KIRP	mirDIP H	5-500	42
	mirDIP H	default	14
	mirDIP VH	5-500	22
	mirDIP VH	default	17
	targetScan*	5-500	16
	targetScan*	default	8
Raponi	mirDIP H	5-500	87
	mirDIP H	default	94
	mirDIP VH	5-500	16
	mirDIP VH	default	16
	targetScan	5-500	15
	targetScan	default	13
	miRanda*	5-500	1
	miRanda*	default	0
Tanic	mirDIP H	5-500	209
	mirDIP H	default	207
	mirDIP VH	5-500	65
	mirDIP VH	default	67
	targetScan	5-500	57
	targetScan	default	49
	miRanda*	5-500	1
	miRanda*	default	0

[#] Benjamini-Hochberg adjusted p-value, *target prediction data used in original analysis, ** default = pathways of 10-500 genes.

Comparison of results of mdgsa analyses from different miRNA datasets is difficult as it is expected that enriched pathways differ, if miRNA datasets are from distinct samples. To compare how the source of miRNA target prediction affects the results of enrichment analysis by mdgsa, I collected more detailed data from the analysis of Kassambara1 miRNA cluster. These results are shown in Table 12. If the same GO term or MSigDB pathway is present more than once, it is coloured on the table. Analyses with MSigDB pathways gave very similar results with both mirDIP very high and targetScan miRNA target predictions. MiRTarget interactions resulted in one significantly enriched pathway that was not seen with other target interactions. However, pathways that were significantly enriched were different that shown in the original publication by Kassambara *et al.* (Kassambara et al. 2017).

Those GO terms that had p-value <0.05 and were seen with different target prediction data were; GO:0007156 Homophilic cell adhesion via plasma membrane adhesion molecules, GO:0060527 Prostate epithelial cord arborization involved in prostate glandular acinus morphogenesis, GO:0005581 Collagen trimer and GO:0018242 protein O-linked glycosylation via serine. All the other terms resulted from only one analysis. In both GO and MSigDB pathways there were pathways related to collagen.

Table 12. Summary of results of Kassambara1 mdgsa analysis.

Target prediction tool	pathways	GO domain	No of genes in pathway	GO terms with p-val <0.05	GO terms	no of genes in GO term
mirDIP H	GO	BP	default*	1	GO:0007156	152
mirDIP H	GO	CC	default	0		
mirDIP H	GO	MF	default	0		
mirDIP H	GO	BP	5-500	4	GO:0007156	152
					GO:0051573	5
					GO:0036035	5
					GO:0060527	6
mirDIP H	GO	CC	5-500	0		
mirDIP H	GO	MF	5-500	2	GO:0008532	5
					GO:0033592	5
mirDIP VH	GO	BP	default	1	GO:0007156	145
mirDIP VH	GO	CC	default	2	GO:0005581	80
					GO:0016442	10
mirDIP VH	GO	MF	default	0		
mirDIP VH	GO	BP	5-500	11	GO:0035020	6
					GO:0071286	5
					GO:0009642	5
					GO:0072540	5
					GO:0086045	5
					GO:0050847	8
					GO:0050658	5
					GO:0007156	145
					GO:0018242	6
					GO:0060527	6
					GO:0051224	5
mirDIP VH	GO	CC	5-500	1	GO:0005581	80
mirDIP VH	GO	MF	5-500	0		
mirTarget	GO	BP	default	1	GO:0000338	10
mirTarget	GO	CC	default	1	GO:0005938	126
mirTarget	GO	MF	default	0		
mirTarget	GO	BP	5-500	3	GO:0006089	6
					GO:1902961	5
					GO:0060509	5
mirTarget	GO	CC	5-500	0		
mirTarget	GO	MF	5-500	5	GO:0097100	5
					GO:0034046	7
					GO:0045296	274
					GO:0008440	5
					GO:0004532	5
targetScan	GO	BP	default	0		
targetScan	GO	CC	default	2	GO:0005581	68
					GO:0031901	105
targetScan	GO	MF	default	0		
targetScan	GO	BP	5-500	3	GO:0021517	7
					GO:0018242	6
					GO:0048841	6
targetScan	GO	CC	5-500	1	GO:0005581	68
targetScan	GO	MF	5-500	3	GO:0043559	5
					GO:0032182	5
					GO:0008237	98
mirDIP H	MSigDB		default	0		
mirDIP H	MSigDB		5-500	0		
mirDIP VH	MSigDB		default	3	NABA_COLLAGENS|http://www.broadinstitute.org/gsea/msigdb/cards/NABA_	42
					REACTOME_COLLAGEN_FORMATION|http://www.broadinstitute.org/gsea/ms	53
					REACTOME_EXTRACELLULAR_MATRIX_ORGANIZATION|http://www.broadinst	72
mirDIP VH	MSigDB		5-500	4	NABA_COLLAGENS|http://www.broadinstitute.org/gsea/msigdb/cards/NABA_	42
					PID_ECADHERIN_NASCENT_AJ_PATHWAY|http://www.broadinstitute.org/gse	37
					REACTOME_COLLAGEN_FORMATION|http://www.broadinstitute.org/gsea/ms	53
					REACTOME_EXTRACELLULAR_MATRIX_ORGANIZATION|http://www.broadinst	72
mirTarget	MSigDB		default	1	REACTOME_UNWINDING_OF_DNA|http://www.broadinstitute.org/gsea/msig	11
mirTarget	MSigDB		5-500	1	REACTOME_UNWINDING_OF_DNA|http://www.broadinstitute.org/gsea/msig	11
targetScan	MsigDB		default	3	REACTOME_COLLAGEN_FORMATION|http://www.broadinstitute.org/gsea/ms	54
					NABA_COLLAGENS|http://www.broadinstitute.org/gsea/msigdb/cards/NABA_	42
					REACTOME_EXTRACELLULAR_MATRIX_ORGANIZATION|http://www.broadinst	68
targetScan	MsigDB		5-500	3	REACTOME_COLLAGEN_FORMATION|http://www.broadinstitute.org/gsea/ms	54
					NABA_COLLAGENS|http://www.broadinstitute.org/gsea/msigdb/cards/NABA_	42
					REACTOME_EXTRACELLULAR_MATRIX_ORGANIZATION|http://www.broadinst	68

* default = 10-500 genes. BP = biological processes, MF = molecular functions, CC = cellular compartments. mirDIP VH = mirDIP very high, mirDIP H = mirDIP High

5.2.3 Comparison of functional enrichment analysis results from BUFET, MDGSA and ClusterProfiler

R/Bioconductor package ClusterProfiler was used in Kassambara *et al.* (Kassambara et al. 2017) for overrepresentation analysis. ClusterProfiler can be used also to perform gene set enrichment analysis, but for this study it was used only for overrepresentation analysis. Only Kassambara cluster1 dataset was analysed with ClusterProfiler to elucidate how the source of target predictions affects the results of analysis. Summary of these analyses is shown in Table 13.

Table 13. Effect of target predictions to Cluster profiler analysis of Kassambara datasets.

Cluster	No of pathways	Predictions	Pathway data	Overlap with original data	Overlap between analyses using different predictions*
1	14	miRTarget	MsigDB	8/9 pathways	11
1	480	mirDIP VH	MsigDB	7/9 pathways	
1	1605	mirDIP VH	GO BP		15
1	23	miRTarget	GO BP		
2	18	miRTarget	MsigDB	18/ 18 pathways	16
2	434	mirDIP VH	MsigDB	16/ 18 pathways	
2	1486	mirDIP VH	GO BP		65
2	89	miRTarget	GO BP		
3	1	miRTarget	MsigDB	1/1 pathways	-
3	446	mirDIP VH	MsigDB	no overlap	
3	233	mirDIP VH	GO BP		
3	-	miRTarget	GO BP	no significant pathways	

**number of pathway terms that were same with using miRTarget or mirDIP VH target gene predictions*

ClusterProfiler analysis with target predictions from miRTarget and gene sets from MSigDB as in original publication resulted in great overlap with the results presented in the original publication. Analysis with target predictions from mirDIP VH and gene sets from MsigDB had also most of the pathways seen in the original publication, but the overall number of significant pathways were many folds higher. Similarly, analyses with mirDIP VH target predictions and GO biological processes resulted in

high number of significant pathways and much higher than the same analyses with target predictions from miRTarget

There were only a few significant pathways that were identified with mdgsa analysis of Kassambara dataset and these had no overlap with the pathways reported in the original publication. BUFET analysis of Kassambara datasets did not result in any significant pathways.

Overall, BUFET analyses resulted in very few significant pathways in most of the datasets, but there were also few analyses that resulted in thousands of significant pathways especially for the KICH dataset (Table 8). In contrast, the analysis of KICH dataset with mdgsa resulted in more reasonable numbers of enriched pathways (Table 11). The overlap of analyses of KICH dataset with targetScan prediction similarly as in original publication resulted in only very little overlap as there were 2 pathways in GO BP and only one in GO MF the same as in original publication. On the other hand, there was no overlap with the significant pathways of GO CC with any of the target predictions and 2 pathways that were same for GO MF with target predictions from mirDIP high or very high. Analyses of KIRP data with mdgsa did not result in any overlap with the data presented in the original publication of mdgsa algorithm (Garcia-Garcia et al. 2016). Those pathways that were seen in BUFET analysis with KIRP FC2, $p<0.05$ and KIRP $p<0.01$ datasets were also among the significant pathways in mdgsa analysis with mirDIP high predictions. Two out of five significant pathways that resulted from the analysis of KIRP $p<0.05$ dataset with mirDIP high prediction were among the significant pathways of mdgsa analysis. These results are summarized in Appendix 3.

BUFET was tested with 10 000, 100 000 and 1 000 000 permutations, but the number of permutations did not change the results, so most of the analyses were performed with the default of 10 000 permutations. Datasets that were selected as original datasets for BUFET analysis gave similar results than in the original publication (Bleazard, Lamb, and Griffiths-Jones 2015). Raponi dataset did not result in any significant pathways in the original publication and in the analyses performed in this study, only target predictions from miRanda resulted in one significant pathway. On

the other hand, Tanic dataset showed more than 3300 significant pathways in the original publication and similarly in analysis in this study, there was more than 4000 significant pathways. However, enrichment was only seen with the target predictions from miRanda. As a summary, there was not much overlap between the results from the different algorithms and only slight overlap with the original studies either. The source of target predictions clearly affected the results of functional enrichment analysis as expected.

6 Discussion and conclusions

The aim of this study was to develop and test a workflow for miRNA functional enrichment analysis. As a first step, target prediction databases were searched from literature and then tested and compared. mirDIP4.1 database, which collects target predictions from 30 independent resources was selected for the work. The original idea was that a database integrating data from as many as possible resources would lead to more reliable target predictions. Comparison of target predictions from mirDIP4.1 to other target prediction tools used in this study showed that the results of different target prediction tools differ greatly, which has been shown already before (Tokar et al. 2017). In the present study, it was also seen that the number of predicted targets was proportional to the number of miRNAs, but the number of targets predicted by different tools was also dependent on the specific miRNAs on the list. This indicates that different target prediction tools may have a different coverage of target genes for some miRNAs.

Because it is still not known which of the biological properties of miRNA are the most important for their regulatory function, more knowledge is needed to define confidentially the target genes of miRNAs. It has been also shown, that miRNAs do not regulate all of their predicted target genes *in vivo* (Pinzón et al. 2017), which complicates the predictions even more. Good target predictions have been shown to prevent bias in functional analysis (Tokar et al. 2017) and thus it would have been good to include data from experimentally validated targets into this study to compare how the use of experimentally validated targets would affect the results. Although the data of experimentally validated miRNA targets may still be limited, it could be an useful approach to compare these results with such obtained with for example mirTarbase targets.

One technical problem which I observed while preparing this study was that some of the tools and databases may not be updated after publication and some of the tools are deprecated at some point. This has happened for example to miRanda tool, which has not been available since May 2018 and for the miRecords database that is used by miRTarget target prediction tool. miRecords database has been deprecated quite

recently because at the time of the analyses done in this study it was still available for use. Another problem that may arise from these deprecated algorithms is that if a functional analysis algorithm is using target prediction from only one source as is the case for empiricalGO (Bleazard, Lamb, and Griffiths-Jones 2015), that tool may become unusable at least without modification of the code. EmpiricalGO requires a local installation of miRanda algorithm. However, empiricalGO is an open source python script, so it could be modified to overcome this problem.

Recent data shows that the functional analysis of miRNA data tends to have a bias towards certain pathways and this bias is seen even with random lists of miRNAs (Bleazard, Lamb, and Griffiths-Jones 2015; Godard and Van Eyll 2015). These results were the motivation to test such algorithms for functional analysis that would avoid this bias. The first algorithm that was selected for this study is based on the work by Bleazard *et al.* and this BUFET algorithm corrects analysis by empirical testing to avoid bias(Zagganas et al. 2017). The results of BUFET analyses in this study showed that the majority of miRNA lists with different sources of target predictions did not lead to any significantly enriched pathways. However, the analyses of some of the miRNA datasets, KICH and Tanic, resulted in thousands of significant GO terms. The Tanic dataset has only 46 miRNAs, so it does not seem biologically very likely that so few miRNAs would significantly regulate over 4000 GO gene sets. Most of these GO terms had less than five genes and only somewhat more than 500 had more than five genes in them. Interestingly, the analysis of Tanic dataset with target predictions from miRanda showed enrichment of over 4000 GO terms, but analysis with target predictions from microT only two to three GO terms. In addition, for BUFET analysis the full set of GO terms was used for most of the analyses. For some analyses, only GO terms of biological processes were used to study whether the number of GO terms in the dataset would affect the results. However, this did not affect the number of significant pathways. To increase the number of meaningful results it would have been good to filter the GO terms so that only those terms that have 5-500 genes would have been selected for the analysis.

It is usual that in biological processes changes in single genes or miRNAs might be small, but affect the same pathway having a clear impact on regulating some cellular

process. Another approach for functional enrichment analysis, gene set enrichment analysis, can show the impact of these small changes more clearly than the overrepresentation analysis. The other algorithm chosen for this study is gene set analysis algorithm mdgsa (Garcia-Garcia et al. 2016). Unlike BUFET, mdgsa incorporates the additive effect that several miRNAs targeting the same gene can have. Most of the analyses with mdgsa with different target predictions showed an enrichment of up to 200 pathways. Default settings for mdgsa analysis restrict the number of genes in a gene set to 10-500 genes. Analyses were done also for the gene sets of 5-500 genes. In contrast to BUFET analysis, KICH dataset had 1-19 enriched BP GO terms depending on the target prediction data and Tanic dataset 0-209 BP GO terms. For most of the analyses, gene sets of 10-500 (default) and 5-500 gave a similar number of enriched pathways indicating that the majority of the enriched pathways had 10-500 genes.

The source of target prediction data affected the results of functional analysis greatly. This was expected as the predicted targets from different sources were not the same. Since the number of validated miRNA target genes is still limited it is difficult to know expected number of target genes for each miRNA. As the target predictions from different algorithms differ (Tokar et al. 2017), integrative approach, such as mirDIP4.1, is justified to combine the data from many different sources. However, the mirDIP4.1 database predicts very huge number of target genes if the cut-off is not set to “Very high”. In addition, as the size of the human genome is about 20 000 genes and number of mature miRNAs about 2600, it seems questionable that a list of 23 miRNAs (Kassambara1) would regulate 1/3 of these genes as predicted by mirDIP4.1 with the cut-off “Very high”. However, even the smallest set of these target prediction tools, miRTarget, predicted almost 5000 target genes for this same list of 23 miRNAs.

One other aim of this study was to compare the results of these different functional analyses with each other and with the original data. Because the BUFET analyses resulted in only a few significant results, comparison with mdgsa could only be done for some of the analyses. In addition, BUFET analyses of Tanic and Raponi datasets with target predictions from miRanda lead to similar numbers of enriched pathways

as seen in the original publication (Bleazard, Lamb, and Griffiths-Jones 2015). However, it is questionable whether 46 differentially expressed miRNAs can regulate over 4000 pathways as predicted. The analysis of KICH and KIRP datasets that were analysed in the original study of mdgsa (Garcia-Garcia et al. 2016), showed only a little overlap with the original results with the target prediction from TargetScan as in the original study. However, the TargetScan predictions that were used in the original study have been different from those used in this study where the newest version of TargetScan was used (release 7.2, March 2018).

Reproduction of the data from Kassambara *et al.* (Kassambara et al. 2017) by ClusterProfiler and target predictions from miRTarget gave very similar results to original publication. The ClusterProfiler analyses resulted in a much higher number of enriched pathways with target prediction data from mirDIP4.1 compared with the original data. The analysis of Kassambara1 dataset with BUFET did not result in any significant enrichment, but few enriched pathways were seen in mdgsa analyses with all of the different target prediction datasets. However, there was no overlap to original data. This could be also due to the correction of bias in analysis that might have been in the original analysis, but the enriched pathways were not the same with different target predictions although there was some overlap between the results.

In conclusion, the functional analysis of miRNA data is complicated and the prediction of miRNA target genes is a crucial step for the analysis. In addition, as miRNAs can have synergistic and antagonistic effects on their target genes, the functional analysis algorithm should incorporate this feature of miRNAs into the analysis. The BUFET algorithm does not include the synergistic effect of miRNAs, but the multi-hit version of empiricalGO from which the BUFET is developed would incorporate the synergistic effect of miRNAs and thus be a better option. However, that would need to be modified, since it requires a local installation of deprecated miRanda target prediction algorithm. Because the gene set analysis such as mdgsa is not excluding any of the miRNAs, it can reveal the effects of more subtle changes than overrepresentation analysis. In addition, mdgsa incorporates the additive effects of miRNAs into its transferred index, which moves the miRNA level information into gene level. Therefore, mdgsa seems to be more relevant in biological sense in the

functional analysis of miRNA data. For the target prediction, confidential target prediction is crucial for the following analysis steps and interpretation of the data, but confidentiality of target predictions is difficult to estimate. Number of target interactions predicted by different target prediction tools were quite high even for a relatively short list of miRNAs. The integrative database of miRNA target interactions, mirDIP4.1, that was selected for this study predicted high numbers of target interactions for all the miRNA lists used. For this study, it would have been good to include also data from experimentally validated targets.

As a summary, the use of mirDIP4.1 database as a source for miRNA target predictions is justified as it is integrating the data from several different sources. However, the number of targets predicted by mirDIP4.1 is quite high, but as the true number of target genes for each miRNA is not verified yet, it is difficult to estimate whether it is correct or not. For the functional analysis, gene set analysis by mdgsa can integrate the synergistic effect of miRNAs that BUFET is not taking into account. This is important and makes mdgsa analysis biologically more relevant. In addition, the BUFET analysis resulted in either very few or no enrichment or then hundreds or thousands of enriched pathways. This does not seem to be biologically reliable. The multi-hit version of empiricalGO algorithm might have been a better choice, as it can integrate the synergistic effects of miRNAs. However, and as expected, both of the tested algorithms, mdgsa and BUFET, gave clearly different results than ClusterProfiler, which is based on overrepresentation analysis by a hypergeometric test without correction for bias. Of these algorithms, mdgsa would be a good choice for the functional analysis but the tool for miRNA target prediction affects the results. Based on work done in this study a workflow for miRNA data analysis is shown in Figure 10.

Figure 10. A miRNA data analysis workflow based on work done in this study. Different target prediction tools and functional enrichment analysis tools were tested in this study. Of these, mirDIP4.1 database for target prediction with Very High target classification and mdgsa algorithm for gene set enrichment are the tools suggested to be used for miRNA data analysis.

7 References

Agarwal, Vikram, George W. Bell, Jin Wu Nam, and David P. Bartel. 2015. "Predicting Effective microRNA Target Sites in Mammalian mRNAs." *eLife* 4(AUGUST2015): 1–38.

Andrés-León, Eduardo, Rocío Núñez-Torres, and Ana M. Rojas. 2016. "miARma-Seq: A Comprehensive Tool for miRNA, mRNA and circRNA Analysis." *Scientific Reports* 6: 1–7. http://dx.doi.org/10.1038/srep25749.

Ashburner, Michael et al. 2011. "Gene Ontology : Tool for the Unification of Biology." *Nature genetics* 25(1): 25–29. http://www.nature.com/ng/journal/v25/n1/abs/ng0500_25.html.

Bleazard, Thomas, Janine A. Lamb, and Sam Griffiths-Jones. 2015. "Bias in microRNA Functional Enrichment Analysis." *Bioinformatics* 31(10): 1592–98.

Brown, Rikki A.M. et al. 2018. "Total RNA Extraction from Tissues for microRNA and Target Gene Expression Analysis: Not All Kits Are Created Equal." *BMC Biotechnology* 18(1): 16.

Cho, Sooyoung et al. 2013. "MiRGator v3.0: A microRNA Portal for Deep Sequencing, Expression Profiling and mRNA Targeting." *Nucleic Acids Research* 41(D1): 252–57.

Chou, Chih-Hung et al. 2017. "miRTarBase Update 2018: A Resource for Experimentally Validated microRNA-Target Interactions." *Nucleic Acids Research* 46(D1): D296–302.

Dweep, Harsh, and Norbert Gretz. 2015. "miRWalk2.0: A Comprehensive Atlas of microRNA-Target interactions_Supplement." *Nature methods* 12(8): 697.

Dweep, Harsh, Carsten Sticht, Priyanka Pandey, and Norbert Gretz. 2011. "MiRWalk - Database: Prediction of Possible miRNA Binding Sites by ' Walking' the Genes of Three Genomes." *Journal of Biomedical Informatics* 44(5): 839–47. http://dx.doi.org/10.1016/j.jbi.2011.05.002.

Enright, Anton J et al. 2003. "MicroRNA Targets in Drosophila." *Genome biology*

5(1): 1–14.

Friedman, Robin C, Kyle Kai-how Farh, Christopher B Burge, and David P Bartel. 2009. “Most Mammalian mRNAs Are Conserved Targets of miRNAs.” *Genome Research* 19: 92–105.

Friedman, Yitzhak, Solange Karsenty, and Michal Linial. 2014. “MiRror-Suite: Decoding Coordinated Regulation by microRNAs.” *Database* 2014(December 2017): 1–12.

Garcia-Garcia, Francisco, Joaquin Panadero, Joaquin Dopazo, and David Montaner. 2016. “Integrated Gene Set Analysis for microRNA Studies.” *Bioinformatics* 32(18): 2809–16.

Garcia, David et al. 2012. “Weak Seed-Pairing Stability and High Target-Site Abundance Decrease the Proficiency of Lsy-6 and Other miRNAs.” *Nature Structural and Molecular Biology* 18(10): 1139–46.

Giraldez, Maria D. et al. 2018. “Comprehensive Multi-Center Assessment of Small RNA-Seq Methods for Quantitative miRNA Profiling.” *Nature Biotechnology* 36(8): 746–57.

Godard, Patrice, and Jonathan Van Eyll. 2015. “Pathway Analysis from Lists of microRNAs: Common Pitfalls and Alternative Strategy.” *Nucleic Acids Research* 43(7): 3490–97.

Griffiths-Jones, S. 2004. “The microRNA Registry.” *Nucleic Acids Research* 32(90001): 109D–111. https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkh023.

Grimson, Andrew et al. 2007. “MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing.” *Molecular Cell* 27(1): 91–105.

Guo, Zhi-Wei et al. 2015. “MtiBase: A Database for Decoding microRNA Target Sites Located within CDS and 5′UTR Regions from CLIP-Seq and Expression Profile Datasets.” *Database* 2015: bav102.

Gusev, Yuriy et al. 2007. “Computational Analysis of Biological Functions and

Pathways Collectively Targeted by Co-Expressed microRNAs in Cancer." *BMC Bioinformatics* 8(SUPPL. 7): 1–17.

Hao, Yajing et al. 2016. "NPInter v3.0: An Upgraded Database of Noncoding RNA-Associated Interactions." *Database* 2016: 1–9.

Hausser, Jean, and Mihaela Zavolan. 2014. "Identification and Consequences of miRNA-Target Interactions-beyond Repression of Gene Expression." *Nature Reviews Genetics* 15(9): 599–612.

John, Bino et al. 2004. "Human MicroRNA Targets." *PLoS Biology* 2(11): e636.

Jonas, Stefanie, and Elisa Izaurralde. 2015. "Towards a Molecular Understanding of microRNA-Mediated Gene Silencing." *Nature Reviews Genetics* 16(7): 421–33. http://dx.doi.org/10.1038/nrg3965.

Junge, Alexander et al. 2017. "RAIN: RNA-Protein Association and Interaction Networks." *Database* 2017(1): 1–9.

Kalvari, Ioanna, Eric P. Nawrocki, et al. 2018. "Non-Coding RNA Analysis Using the Rfam Database." *Current Protocols in Bioinformatics* 62(1): 1–27.

Kalvari, Ioanna, Joanna Argasinska, et al. 2018. "Rfam 13.0: Shifting to a Genome-Centric Resource for Non-Coding RNA Families." *Nucleic Acids Research* 46(D1): D335–42.

Kanehisa, Minoru et al. 2019. "New Approach for Understanding Genome Variations in KEGG." *Nucleic acids research* 47(D1): D590–95.

Kanehisa, Minoru, and Susumu Goto. 2000. "KEGG: Kyoto Encyclopedia of Genes and Genomes." *Nucleic Acids Research* 28(1): 27–30.

Karagkouni, Dimitra et al. 2017. "DIANA-TarBase v8: A Decade-Long Collection of Experimentally Supported miRNA–gene Interactions." *Nucleic Acids Research* 46(D1): D239–45.

Kassambara, Alboukadel et al. 2017. "Global miRNA Expression Analysis Identifies Novel Key Regulators of Plasma Cell Differentiation and Malignant Plasma

Cell.” *Nucleic Acids Research* 45(10): 5639–52.

Kertesz, Michael et al. 2007. “The Role of Site Accessibility in microRNA Target Recognition.” *Nature Genetics* 39(10): 1278–84.

Kozomara, Ana, Maria Birgaoanu, and Sam Griffiths-Jones. 2019. “miRBase: From microRNA Sequences to Function.” *Nucleic acids research* 47(D1): D155–62.

Krek, Azra et al. 2005. “Combinatorial microRNA Target Predictions.” *Nature Genetics* 37(5): 495–500.

Lee, Rosalind C., Rhonda L. Feinbaum, and Victor Ambros. 1993. “The C. Elegans Heterochronic Gene Lin-4 Encodes Small RNAs with Antisense Complementarity to Lin-14.” *Cell* 75(5): 843–54. https://www.sciencedirect.com/science/article/pii/009286749390529Y?via%3Dihub (April 5, 2019).

Lee, Yoontae et al. 2004. “MicroRNA Genes Are Transcribed by RNA Polymerase II.” *EMBO Journal* 23(20): 4051–60.

Lewis, Benjamin P., Christopher B Burge, and David P. Bartel. 2005. “Conserved Seed Pairing , Often Flanked by Adenosines , Indicates That Thousands of Human Genes Are MicroRNA Targets We Predict Regulatory Targets of Vertebrate microRNAs.” *Cell* 120(1): 15–20.

Lin, Shuibin, and Richard I. Gregory. 2015. “MicroRNA Biogenesis Pathways in Cancer.” *Nature Reviews Cancer* 15(6): 321–33. http://dx.doi.org/10.1038/nrc3932.

Liu, Zhi Ping, Canglin Wu, Hongyu Miao, and Hulin Wu. 2015. “RegNetwork: An Integrated Database of Transcriptional and Post-Transcriptional Regulatory Networks in Human and Mouse.” *Database* 2015: 1–12.

Lukasik, Anna, Maciej Wójcikowski, and Piotr Zielenkiewicz. 2016. “Tools4miRs - One Place to Gather All the Tools for miRNA Analysis.” *Bioinformatics* 32(17): 2722–24.

Maragkakis, M. et al. 2009. “DIANA-microT Web Server: Elucidating microRNA

Functions through Target Prediction." *Nucleic Acids Research* 37: 273–76.

Maragkakis, Manolis et al. 2011. "DIANA-microT Web Server Upgrade Supports Fly and Worm miRNA Target Prediction and Bibliographic miRNA to Disease Association." *Nucleic Acids Research* 39(SUPPL. 2): 145–48.

Mehta, Arnav, and David Baltimore. 2016. "MicroRNAs as Regulatory Elements in Immune System Logic." *Nature Reviews Immunology* 16(5): 279–94.

Miranda, Kevin C. et al. 2006. "A Pattern-Based Method for the Identification of MicroRNA Binding Sites and Their Corresponding Heteroduplexes." *Cell* 126(6): 1203–17.

Montaner, David, and Joaquín Dopazo. 2010. "Multidimensional Gene Set Analysis of Genomic Data." *PLoS ONE* 5(4): e10348.

Mootha, Vamsi K et al. 2003. "PGC-1α-Responsive Genes Involved in Oxidative Phosphorylation Are Coordinately Downregulated in Human Diabetes." *Nature Genetics* 34: 267. https://doi.org/10.1038/ng1180.

Ørom, Ulf Andersson, Finn Cilius Nielsen, and Anders H. Lund. 2008. "MicroRNA-10a Binds the 5′UTR of Ribosomal Protein mRNAs and Enhances Their Translation." *Molecular Cell* 30(4): 460–71.

Peterson, Sarah M. et al. 2014. "Common Features of microRNA Target Prediction Tools." *Frontiers in Genetics* 5(FEB): 1–10.

Pinzón, Natalia et al. 2017. "MicroRNA Target Prediction Programs Predict Many False Positives." *Genome Research* 27(2): 234–45.

Raponi, Mitch et al. 2009. "MicroRNA Classifiers for Predicting Prognosis of Squamous Cell Lung Cancer." *Cancer Research* 69(14): 5776–83.

Riffo-Campos, Ángela L., Ismael Riquelme, and Priscilla Brebi-Mieville. 2016. "Tools for Sequence-Based miRNA Target Prediction: What to Choose?" *International Journal of Molecular Sciences* 17(12).

Rougvie, Anne E. 2001. "Control of Developmental Timing in Animals." *Nature*

Reviews Genetics 2(September): 690–701. http://www.wormbase.org/db/misc/paper?name=WBPaper00004850.

Ru, Yuanbin et al. 2014. "The multiMiR R Package and Database: Integration of microRNA-Target Interactions along with Their Disease and Drug Associations." *Nucleic Acids Research* 42(17): 1–10.

Sanchez-Diaz, Patricia C. et al. 2013. "De-Regulated MicroRNAs in Pediatric Cancer Stem Cells Target Pathways Involved in Cell Proliferation, Cell Cycle and Development." *PLoS ONE* 8(4).

Schnall-Levin, Michael et al. 2011. "Unusually Effective microRNA Targeting within Repeat-Rich Coding Regions of Mammalian mRNAs." *Genome Research* 21(9): 1395–1403.

Shirdel, Elize A., Wing Xie, Tak W. Mak, and Igor Jurisica. 2011. "NAViGaTing the Micronome - Using Multiple microRNA Prediction Databases to Identify Signalling Pathway-Associated microRNAs." *PLoS ONE* 6(2).

Subramanian, Aravind et al. 2005. "Gene Set Enrichment Analysis : A Knowledge-Based Approach for Interpreting Genome-Wide." *PNAS* 102(43): 15545–50.

Tanic, M. et al. 2013. "MicroRNA-Based Molecular Classification of Non-BRCA1/2 Hereditary Breast Tumours." *British Journal of Cancer* 109(10): 2724–34.

The Gene Ontology Consortium. 2019. "The Gene Ontology Resource: 20 Years and Still GOing Strong." *Nucleic acids research* 47(D1): D330–38.

Tokar, Tomas et al. 2017. "mirDIP 4.1—integrative Database of Human microRNA Target Predictions." *Nucleic Acids Research* 46(D1): D360–70. http://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkx1144/4670951.

Treiber, Thomas, Nora Treiber, and Gunter Meister. 2019. "Regulation of microRNA Biogenesis and Its Crosstalk with Other Cellular Pathways." *Nature Reviews Molecular Cell Biology* 20(1): 5–20. http://dx.doi.org/10.1038/s41580-018-0059-1.

Vlachos, Ioannis S. et al. 2015. “DIANA-TarBase v7.0: Indexing More than Half a Million Experimentally Supported miRNA:mRNA Interactions.” *Nucleic Acids Research* 43(D1): D153–59.

Xiao, Feifei et al. 2009. “miRecords: An Integrated Resource for microRNA-Target Interactions.” *Nucleic Acids Research* 37(SUPPL. 1): 105–10.

Yu, Guangchuang, Li-Gen Wang, Yanyan Han, and Qing-Yu He. 2012. “clusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters.” *OMICS: A Journal of Integrative Biology* 16(5): 284–87. http://online.liebertpub.com/doi/abs/10.1089/omi.2011.0118.

Zagganas, Konstantinos et al. 2017. “BUFET: Boosting the Unbiased miRNA Functional Enrichment Analysis Using Bitsets.” *BMC Bioinformatics* 18(1): 399. http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1812-8.

Appendix 1. Number of miRNA-target gene interactions in different miRNA target prediction tools.

		No of target gene interactions					
Dataset	**no of miRNA**	**mirDIP very High**	**mirDIP high**	**mirDIP medium**	**targetScan**	**miRanda**	**miRTarget**
Kassambara3	12	12298	44058	224614	8147	33090	4358
Kassambara2	14	20932	61459	256965	13510	38017	11505
Raponi	15	25646	71914	265245	16754	47150	12314
Kassambara1	23	25919	85944	422337	20320	53535	11789
Tanic	46	72305	197154	853950	56132	108234	32262
KICH p-value <0.05, FC >2	274	145228	571044	4327227	74542	499308	61518
KIRP p-value <0.05, FC >2	283	144911	586054	4453289	66459	526280	63809
KICH p-value <0.01	335	198584	739438	5282557	113189	618998	86000
KICH p-value <0.05	386	224595	860842	6101847	126663	738057	102149
KIRP p-value <0.01	400	224052	883026	6372821	114600	769340	104817
KIRP p-value <0.05	459	255665	1008752	7308775	132493	889952	120689

Appendix 2. Results of mdgsa analyses

A. Results of mdgsa analyses with GO Cellular compartments terms.

Dataset	Target prediction tool	no of genes in pathway	pathways with p-val <0.05[#]
Kassambara1	mirDIP H	5-500	0
	mirDIP H	default**	0
	mirDIP VH	5-500	1
	mirDIP VH	default	2
	mirTarget*	5-500	0
	mirTarget*	default	1
	targetScan	5-500	1
	targetScan	default	2
KICH	mirDIP H	5-500	7
	mirDIP H	default	7
	mirDIP VH	5-500	1
	mirDIP VH	default	1
	targetScan*	5-500	1
	targetScan*	default	2
KIRP	mirDIP H	5-500	22
	mirDIP H	default	23
	mirDIP VH	5-500	5
	mirDIP VH	default	5
	targetScan*	5-500	7
	targetScan*	default	7
Raponi	miRanda*	5-500	0
	miRanda*	default	0
	mirDIP H	5-500	51
	mirDIP H	default	59
	mirDIP VH	5-500	5
	mirDIP VH	default	5
	targetScan	5-500	2
	targetScan	default	5
Tanic	miRanda*	5-500	0
	miRanda*	default	0
	mirDIP H	5-500	91
	mirDIP H	default	90
	mirDIP VH	5-500	22
	mirDIP VH	default	21
	targetScan	5-500	15
	targetScan	default	12

[#] Benjamini-Hochberg adjusted p-value, *target prediction data used in original analysis, ** default = pathways of 10-500 genes.

B. Results of mdgsa analyses with GO Molecular functions terms.

Dataset	Target prediction tool	no of genes in pathway	pathways with p-val <0.05[#]
Kassambara1	mirDIP H	5-500	2
	mirDIP H	default**	0
	mirDIP VH	5-500	0
	mirDIP VH	default	0
	mirTarget*	5-500	5
	mirTarget*	default	0
	targetScan	5-500	3
	targetScan	default	0
KICH	mirDIP H	5-500	6
	mirDIP H	default	12
	mirDIP VH	5-500	6
	mirDIP VH	default	7
	targetScan*	5-500	6
	targetScan*	default	8
KIRP	mirDIP H	5-500	16
	mirDIP H	default	15
	mirDIP VH	5-500	18
	mirDIP VH	default	15
	targetScan	5-500	12
	targetScan	default	15
Raponi	miRanda*	5-500	2
	miRanda*	default	0
	mirDIP H	5-500	56
	mirDIP H	default	59
	mirDIP VH	5-500	10
	mirDIP VH	default	13
	targetScan	5-500	11
	targetScan	default	13
Tanic	miRanda*	5-500	2
	miRanda*	default	0
	mirDIP H	5-500	89
	mirDIP H	default	93
	mirDIP VH	5-500	37
	mirDIP VH	default	125
	targetScan	5-500	26
	targetScan	default	29

[#] Benjamini-Hochberg adjusted p-value, *target prediction data used in original analysis, ** default = pathways of 10-500 genes.

Appendix 3. Enriched Biological processes GO terms in KIRP dataset

	KIRP orig pub. (Garcia-Garcia et al.)	mdgsa, default settings			BUFET		
target Predictions	TargetScan	targetScan	midDIP high	mirDIP very high	mirDIP high		
	KIRP full data set	KIRP full data set	KIRP full data set	KIRP full data set	Kirp01	Kirp05	Kirp05FC2
	GO:0002064	GO:0006954	GO:0006412	GO:0000786	**GO:0006614**	GO:0019083	**GO:0006614**
	GO:0008286	GO:0009952	**GO:0006413**	GO:0043005	**GO:0000184**	**GO:0006413**	
	GO:0010837	GO:0016569	GO:0030177	GO:0014069		GO:0050911	
	GO:0022898	GO:0048598	GO:0043010	GO:0045211		**GO:0000184**	
	GO:0032409	GO:0006955	GO:0010613	GO:0036464		GO:0007608	
	GO:0032412	GO:0032956	**GO:0000184**				
	GO:0032869	GO:0045892	**GO:0006614**				
	GO:0042659	GO:0007411	GO:0016477				
	GO:0071709		GO:0042742				
	GO:2000027		GO:0060041				
	GO:0001655		GO:0006816				
	GO:0032486		GO:0007507				
	GO:0034762		GO:0016569				
	GO:0034765		GO:0031424				
	GO:0005126						