Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών
Πανεπιστήμιο Θεσσαλίας

Comparison of databases with experimentally verified miRNA targets

Σύγκριση βάσεων δεδομένων με πειραματικά επιβεβαιωμένους στόχους miRNA

Διπλωματική εργασία

Μπεκιαρίδης Δημήτριος
Αλαμπάνου Αγγελική

Επιβλέποντες Καθηγητές:
Χατζηγεωργίου Άρτεμις
Καθηγήτρια Π.Θ.
Χούστη Αικατερίνη
Καθηγήτρια Π.Θ.

Βόλος, Σεπτέμβριος 2014

Acknowledgments

We would like to thank our supervisor Prof. Hatzigeorgiou Artemis for her valuable guidance and support until the completion of this thesis. We would also like to express our appreciation to Karagouni Dimitra for being extremely helpful and cooperative at all times during our research.

Finally, we would like to thank our families and friends for the immeasurable support, encouragement and motivation during our studies.

Abstract

The purpose of this thesis is the analysis and comparison of databases listing experimentally verified target microRNAs. MicroRNAs are small molecules of RNA (21 to 23 nucleotides in length) which control gene expression. They affect the expression of specific proteins, thereby impacting significantly many cell functions and even causing the development of various diseases. In recent years, the rapid growth of interest in miRNAs targets, as well as the experiments that verify them, raised the necessity for the creation of databases that gather and make publicly available this information. These databases, apart from the fact that they facilitate the research community's access to experimental data, are also a useful tool for the validation of programs that computationally predict the targets of miRNAs. The comparison of the databases was made with the implementation of code that parses the entire data of each base and with the processing of the outcome. The main criteria for the comparison was the number of miRNAs targets in each database per experiment category from which they were validated, as well as the number of publications in which they were referred. The results give an approximate picture of the correlation between the databases that we examine, while the comparison can be used as a guide to the features and advantages of each database.

Περίληψη

Σκοπός της παρούσης διπλωματικής εργασίας είναι η ανάλυση και σύγκριση βάσεων δεδομένων που περιέχουν πειραματικά επαληθευμένους στόχους microRNAs. Τα microRNAs είναι μικρά μόρια RNA (μήκους 21-23 νουκλεοτιδίων) τα οποία ελέγχουν την γονιδιακή έκφραση. Η δράση τους επηρεάζει την έκφραση συγκεκριμένων πρωτεϊνών, επιδρώντας έτσι σημαντικά σε πολλές κυτταρικές λειτουργίες και αποτελώντας ακόμα και βάση εκδήλωσης παθήσεων. Τα τελευταία χρόνια, η ραγδαία αύξηση του ενδιαφέροντος για τους στόχους των miRNAs, όπως επίσης και των πειραμάτων εξακρίβωσής τους, κατέστησε αναγκαία τη δημιουργία βάσεων δεδομένων που θα συγκεντρώνουν αυτή την πληροφορία. Οι βάσεις αυτές, πέρα από το γεγονός ότι διευκολύνουν την πρόσβαση της ερευνητικής κοινότητας στα πειραματικά δεδομένα, αποτελούν και χρήσιμα εργαλεία για την επαλήθευση των προγραμμάτων που προβλέπουν υπολογιστικά τους στόχους των miRNAs. Η σύγκριση των βάσεων έγινε με την δημιουργία κώδικα που διατρέχει το σύνολο των δεδομένων της κάθε βάσης και με την επεξεργασία των παραγόμενων αποτελεσμάτων. Τα αποτελέσματα δίνουν μια προσεγγιστική εικόνα του συσχετισμού μεταξύ των βάσεων που εξετάζουμε, ενώ η σύγκριση αυτή μπορεί να αποτελέσει έναν οδηγό για τα χαρακτηριστικά και τα πλεονεκτήματα της κάθε βάσης.

Table of Contents

Introduction

1.1 Biological background

This chapter is a general biological introduction to the topics regarding this thesis, focusing on basic mechanisms of molecular biology. MicroRNAs (miRNAs) and their contribution to gene expression will be discussed more thoroughly in a subsequent chapter.

1.1.1 The structure of DNA and RNA

In modern molecular biology, genome is described as a full set of hereditary material in a multi-cellular organism. Genes are all genetic data of a single cell. In the majority of living organisms, each gene consists of a specific sequence of nucleotides encoded in a DNA (deoxyribonucleic acid), or in some cases of viruses, a gene consists of RNA (Ribonucleic acid). This DNA is packed tightly into structures called chromosomes, which consist of long chains of DNA and associated proteins to adjust processes of genes.
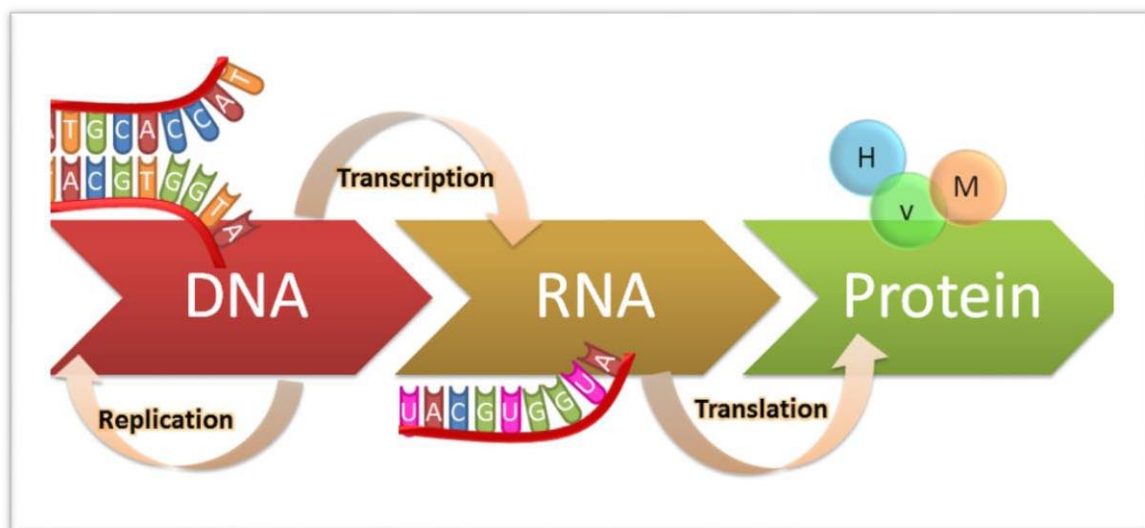

Figure 1. Central dogma of molecular biology: Information flow in biological systems. *(http://lit.genius.com/Biology-genius-the-central-dogma-annotated)*

3

In 1958, Francis Crick proposed the "central dogma" of Molecular Biology to describe the flow of genetic information in cells, also known as gene expression. The main idea is that protein synthesis proceeds from DNA to RNA and then to protein. This process includes three different stages, DNA replication, transcription and translation (Alberts 2008).

RNA is a more versatile version of DNA. Contrary to the DNA, it consists of only a single strand of nucleotides, thus being smaller and less stable, allowing it to bind to other molecules. In addition to single stranded RNA (ssRNA), double stranded RNA (dsRANA) also exists, which is a typical intermediate form during viral replication in a cell. The nucleotides in RNA are not all the same as those in DNA, as the complementary base to adenine (A) is uracil (U) instead of thymine (T).

There are many different types of RNA, all fulfilling different roles in the cell. The ones more relevant to the processes regarding miRNA are the messenger RNA (mRNA), which is used to transport the information encoded in the DNA out of the nucleus; the transfer RNA (tRNA), which is used to link nucleotide sequences with their corresponding amino acids during translation of mRNA into proteins; the ribosomal RNA (rRNA), RNA present in ribosomes, which mediate translation of mRNA into proteins; the micro RNA (miRNA), which is a small strand of non-coding RNA; and the small interfering RNA or silencing RNA (siRNA), which is a second type of small non-coding RNA, about the same size as miRNA, that target a very limited number of mRNAs, usually just one (Carthew et al., 2009).

1.1.2 Messenger RNA

Messenger RNA or mRNA is a large family of RNA molecules which play a major role in protein synthesis. The main function of mRNA is to transfer the genetic material from DNA to the ribosomes, where the genetic material templates carried by the messengers are converted into amino acids, the basis of the polypeptide chain.

1.1.3 Bioinformatics

Bioinformatics is a scientific field that uses computer science and biological data to understand molecular biology. By providing machine learning algorithms, analytical tools and computational methods, bioinformatics employ large data from sequence databases such as DNA sequences and amino acids sequences of proteins. The software tools for bioinformatics read the DNA

4

letters (A, T, G, C) and are able to compare DNA sequences, discover new genes and investigate the functionality of proteins and the genetic networks. The results of these processes help us understand the living organisms.

1.2 MicroRNA

1.2.1 Introduction to microRNAs

MicroRNAs (miRNAs) are non-protein coding single-stranded endogenous RNAs (~19-25 nt long) that can regulate gene expression in plants and animals. miRNAs can silence genes through transcript degradation and/or translation suppression in the case of protein coding genes. Almost two decades since the identification of lin-4 in *Caenorhabditis elegans*, miRNAs are now deemed central to the RNA revolution, exhibiting regulatory control to a multitude of crucial cell functions ranging from apoptosis, cell proliferation and differentiation to cell signalling. miRNAs have now been located in a multitude of organisms including plants, algae, viruses and metazoan (Vergoulis et al., 2011).

Although their biological importance has become clear, how they recognize and regulate target genes remains less well understood. Target sites can be grouped into two broad categories. 5′ dominant sites have sufficient complementarity to the miRNA 5′ end to function with little or no support from pairing to the miRNA 3′ end. An average miRNA has approximately 100 target sites, indicating that miRNAs regulate a large fraction of protein-coding genes and that miRNA 3′ ends are key determinants of target specificity within miRNA families.

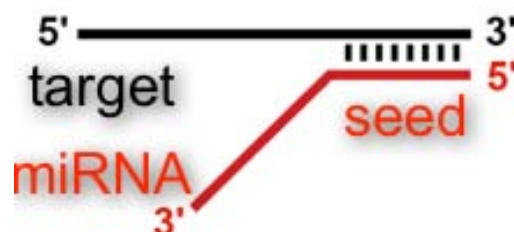To date, functions have been assigned to only a few of the hundreds of animal miRNA genes.

Figure 2. miRNA model. *(http://www.yale.edu/giraldezlab/miRNA.html)*

5

1.2.2 Discovery of miRNAs

In 1993, Victor Ambros, Rosalind Lee and Rhonda Feinbaum, observed MiRNAs in Caenorhabditis elegans (C. elegans) in genetic screens for mutants, due to the gene lin-4 which did not code for a protein (Lee et al., 1993). Lin-4 produced instead short RNA transcripts and further processing produced a smaller transcript that consisted of 22 nucleotides. The sequence complementarity to multiple sequences in the 3' UTR of the lin-4 mRNA was enough to inhibit the translation of lin-14 mRNA. Seven years later, in 2000, a second miRNA, let-7, was discovered in C. elegans (Reinhart et al., 2000). Similar to lin-4, it was found to regulate heterochronic genes such as lin-14, lin-41, lin-28 mRNA during L4 (the fourth larval stage)-adult transition in C. elegans. This short non-coding RNA has been found in ascidia, molluscs, vertebrates, arthrophodae, annelida and hemichordata. Since then, more than 4.000 miRNAs have been discovered in all eukaryotes, while over 700 have been identified in humans. It is predicted that more than 800 miRNA are yet to be discovered (Wu X et al., 2012).
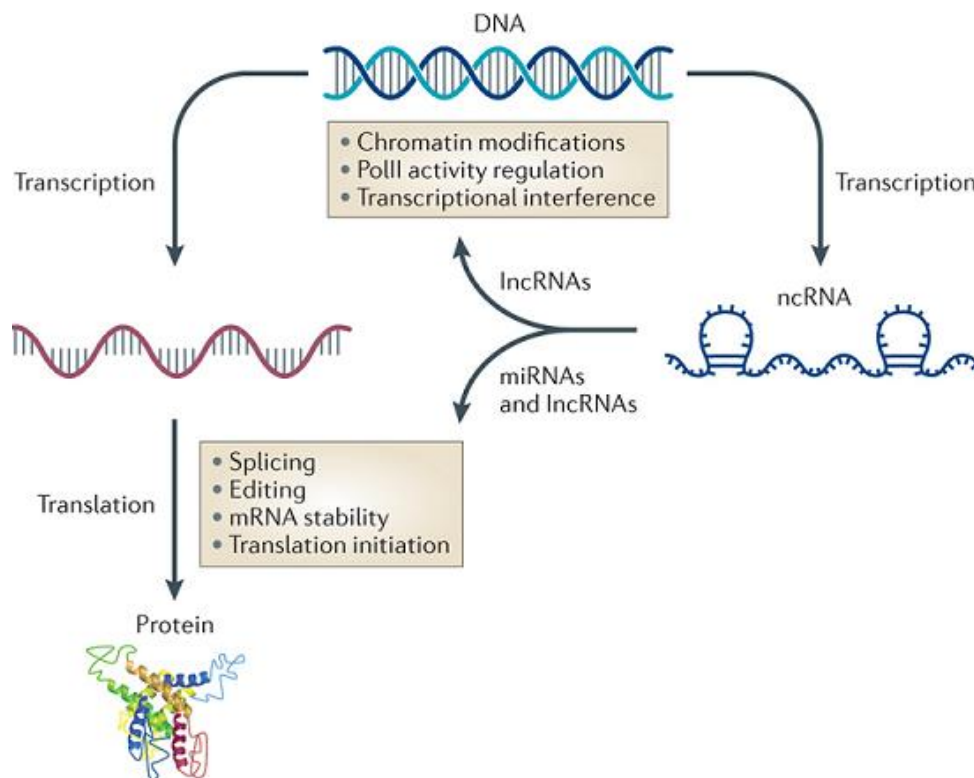

Figure 3. Overview of the changes in molecular biology due to the discovery of ncRNA transcripts. *(*Wahlestedt, 2013*)*

6

1.2.3 MicroRNA biogenesis

MicroRNAs are produced from either their own genes, as independent singular or clustered transcriptional units or from introns. As much as 40% of miRNA genes may lie in the introns of protein and non-protein coding genes, or even in exons. These are usually (but not exclusively) found in a sense orientation and are frequently regulated together with their host genes.

According to the canonical pathway for miRNAs biogenesis, a long poly-adenylated primary miRNA precursor (Cullen et al., 2004), the pri-miRNA, generates miRNA, which is transcribed by RNA polymerase type II (Pol-II). The pri-miRNA may be thousands of nucleotides in length and contains one or more hairpin structures, which are approximately 80 nucleotides long, the pre-miRNA. The polymerase often binds to a promoter found near the DNA sequence, encoding what will become the hairpin loop of the pre-miRNA. When a stem loop precursor is found in the 3' UTR, a transcript may serve as a pri-miRNA and mRNA. RNA polymerase III (Pol-III) transcribes some miRNAs, transfer RNAs (tRNAs), and mammalian wide interspersed repeat (MWIR) promoter units. There are several steps of transcript that are required to produce mature miRNAs from pri-miRNAs.

In the nucleus, the RNase-III enzyme Drosha recognizes the pri-miRNAs and then cleaves these into precursors which are 60 to 100 nt long. The cut performed by Drosha is done at the base of the pre-microRNA. After the initial cleavage by Drosha, pre-miRNAs are transported to the cytoplasm by Exportin 5. In the cytoplasm, the hairpin precursors are cleaved by Dicer, a second RNase-III enzyme, to generate a small double-stranded RNA molecule about 15 to 22 nucleotides in length that contains both the mature miRNA strand and its complementary strand which is called miRNA*.

In the cytoplasm the pre-miRNA is cleaved by another RNAse III type double stranded endonuclease called Dicer (Hutvagner et al., 2001; Ketting et al., 2001). Dicer cleavage of pre-miRNA results in an imperfect miRNA:miRNA* duplex around 20-25 nucleotides in size (Hutvagner et al., 2001; Ketting et al., 2001) containing the mature miRNA strand and its opposite complementary miRNA* strand.
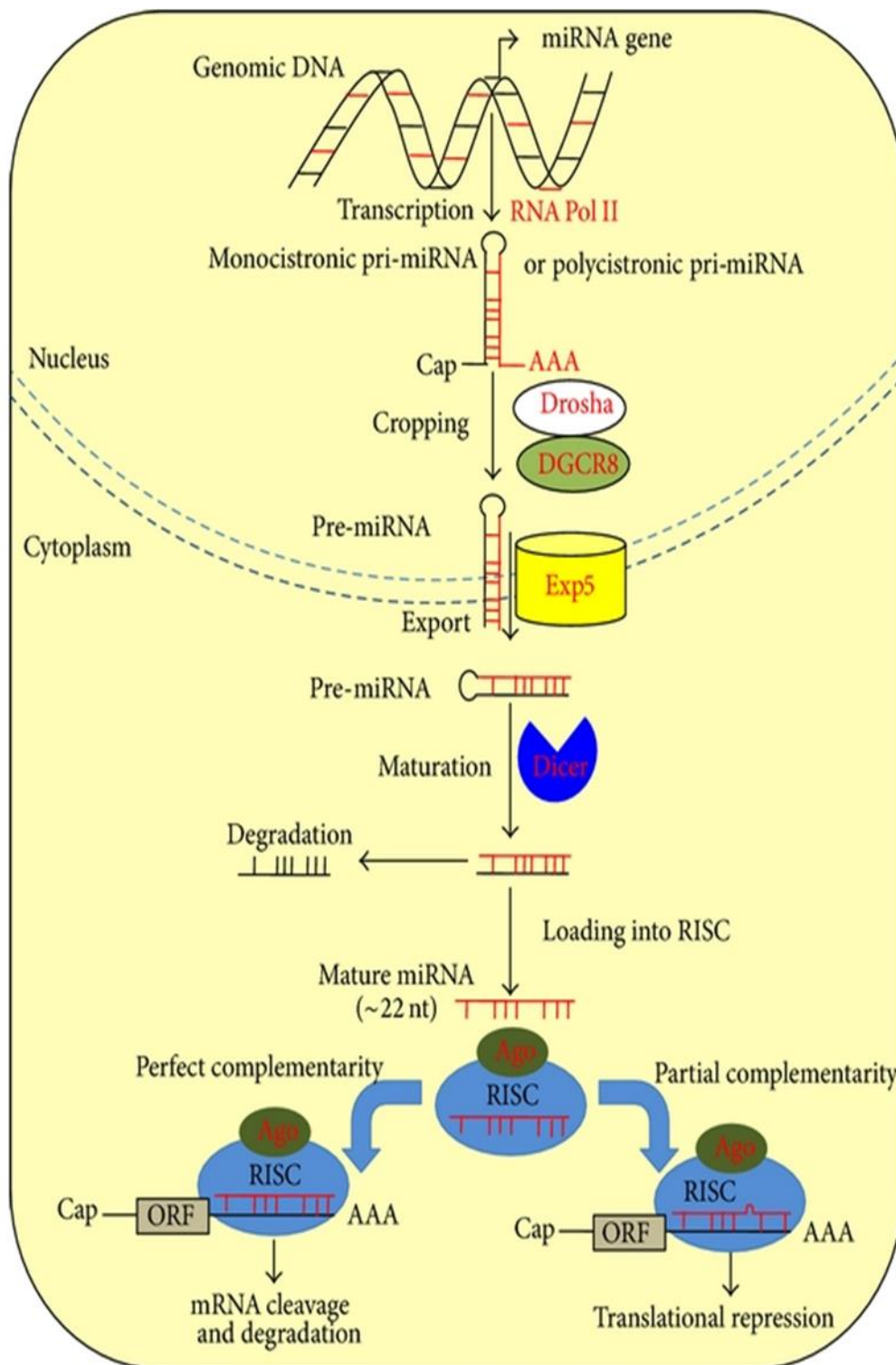
7

Figure 4. Schematic representation of a miRNA biogenesis pathway. *(*Gori et al., 2014*)*

1.3 miRNA target prediction

1.3.1 Prediction methods

An important aspect of miRNA biology is the identification of miRNA targets. Prediction of miRNA targets provides an alternative approach to assign biological functions. In recent years there has been made much progress in this area, as more than 10 miRNA target prediction programs have been established. However, so far there has not been found an efficient and definitive computational method to predict miRNA targets. The commonly accepted mechanism of miRNA targeting in animals involves an interaction between the 'seed region' of the miRNA (5′-end) and the 3′ untranslated region (3′-UTR) of the mRNA. Most of the target prediction algorithms are based on this model.

In order to develop computational algorithms identifying miRNA target genes, principles of miRNA target recognition are often established based on empirical evidences. For example, the importance of base pairing between miRNAs and their targets was suspected according to the observation that the 'target site' of the lin-14 UTR is complementary to the 5' region of the lin-4 miRNA (Lee et al., 1993). Below some of the features used by the mammalian target prediction programs are described:

Seed Sequence complementarity

The 5'end (nucleotide 2–7) of miRNA and 3'UTR region of mRNA have complementarity between them. There are three types of MiRNA binding sites of targets: (1) 5'-dominant canonical with 7 to 8 nucleotide match(base-pairing) in the seed region, (2) 5'-dominant seed with extensive match at the 5'end of the miRNA and limited (3) 3' compensatory target sites which can compensate for the seed region mismatch (Bartel DP, 2009).

Target site conservation

The conservation of miRNA-mRNA target site in the 3'-UTR sequence over many species is an important principle of miRNA target prediction. If the conservation is more extensive that result in a more reliable prediction. Furthermore, this method based on conserved site sequences decreases the false-positive rate of target prediction substantially (Bartel DP, 2004).

Thermodynamics

Several methods predict the target genes of a miRNA based on thermodynamic stability. Thermodynamic analysis of miRNA and mRNA interaction calculates the miRNA-mRNA binding free energy. However, datasets of miRNA-mRNA duplexes are very limited so it is not always easy to decide the right threshold of free energy. For example, when there is a stable miRNA-mRNA binding which means that free energy is lower, it is not certain that prediction of miRNA target genes is reliable (Watanabe et al., 2007).
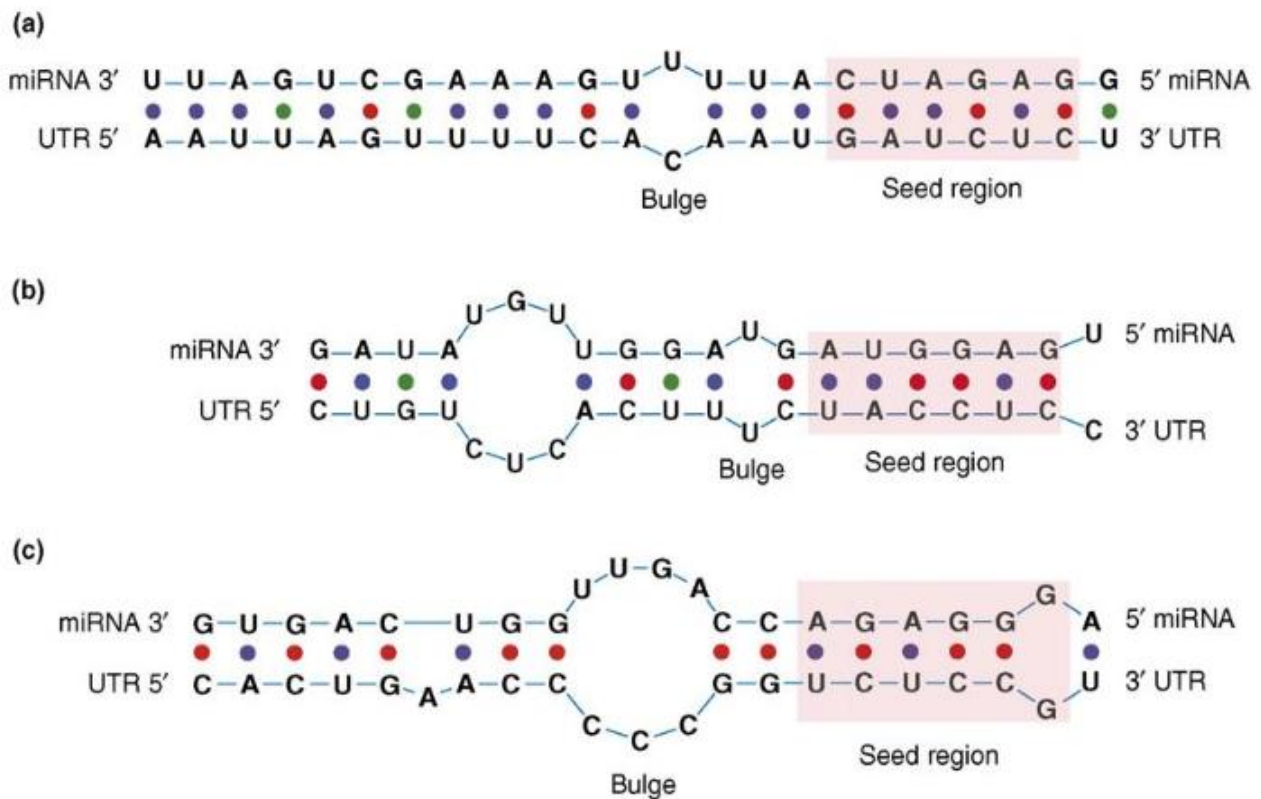

Figure 5. Secondary structures of the three types of miRNA - mRNA interaction
(a) Canonical sites (b) Dominant seed sites (c) Compensatory sites. *(Maziere et al., 2007)*

1.3.2 List of miRNA target prediction tools

Several computational programs have been developed to identify potential miRNA targets. These programs are based on the above different structures, observed between the mature miRNA and 3' UTRs. Current prediction tools are described briefly below.

- **MiRanda:** This tool identifies genomic targets for microRNAs by using weighted dynamic programming algorithm (John, B et al., 2004).
  Type of Method: Complementary

- **TargetScan/TargetScanS:** This algorithm requires the seed complementary at least for 6 nt and deals with the different seed types that have been determined, with a specific hierarchy (Grimson, A., et al., 2007).
  Type of Method: Seed Complementary

- **DIANA microT:** This algorithm searches in the UTRs for stringent seed paring to the miRNA. A signal to noise ratio (SNR) is considered for each interaction to measure the number of false positives (Kiriakidou et al., 2004).
  Type of Method: Thermodynamics

- **PicTar**: is website that provides details for microRNA target predictions (Chen et al., 2006).
  Type of Method: Thermodynamics

- **RNAhybrid**: is a tool for finding the minimum free energy hybridization of a long and a short RNA (Kruger et al., 2006).
  Type of Method: Thermodynamics

- **PITA**: is an algorithm that considers not only the duplex interaction information, but also considers the accessibility to the site in the mRNA (Kertesz et al., 2007).
  Type of Method: Target-site accessibility

- **ElMMo**: is an algorithm that obtains the sites based on the conservation score and uses a Bayesian method (Gaidatzis et al., 2007).
  Type of Method: Bayesian inference

- **RNA22**: A pattern-based method for the identification of microRNA-target sites independent of miRNA target conservation and a method that verifies their corresponding RNA/RNA complexes (Miranda et al., 2006).

11

1.4 miRNA experimental targeting

The use of different rules of targeting from the various miRNA target prediction programs produces rather different lists of predicted targets. Variations can arise because of the different sources of 3′-UTR sequences. Some programs use the Ensembl database to define 3′-UTRs, while some others use the University of California Santa Cruz database, a fact that differentiates significantly the prediction outcomes (Thomson et al., 2011). Moreover, with the identification of genuine miRNA targets lacking a complete 6-mer match and the further complications of RNA structure and RNA-binding proteins affecting site accessibility many predictions may not be *bona fide* targets and many genuine targets can be missed. Furthermore, the false positive rate of such prediction programs has been calculated to vary between 24% and 70%.

All these highlight the necessity of experimental data to demonstrate genuine miRNA function. The experimental validation is a low-throughput process, but is essential to identify authentic miRNA targets. However, there are many experimental methods with various competence and limitations, which have yet to be fully explored.

The mechanisms which are used for examining presumed target expression levels can be divided into five categories: Reporter gene assays, mRNA-level (or gene-level) measurements, protein-level measurements, high-throughput techniques and "others".

1.4.1 Reporter Gene Assay

Over the last years, genetic reporter systems have significantly affected the understanding of gene expression and regulation. A genetic reporter system consists of the genetic element under analysis and a reporter gene attached to it, in an expression vector. Reporter genes are used as an indication of whether the genetic element has been expressed in the cell or organism population. The most frequently used reporter genes involve fluorescent and luminescent proteins, due to their visually identifiable characteristics. For example, cells that express green fluorescent protein (GFP) glow green under blue light, while the E. coli lacZ gene causes bacteria expressing it to appear blue when grown on a medium that contains the substrate analog X-gal. One of the most common reporter genes is luciferase, which catalyzes a reaction with luciferin producing light.

12

1.4.1.1 Luciferase reporter assay

Luciferase is a generic term for the class of oxidative enzymes used in bioluminescence and is distinct from a photoprotein.

The luciferase reporter assay is a common tool to study gene expression at the transcriptional level. The fact that it is relatively inexpensive and convenient, while it gives quantitative measurements instantaneously makes it a widely used tool with broad applications across fields of cell and molecular biology.

Luciferases make up a class of oxidative enzymes found in several species that enable the organisms that express them to "bioluminesce," or emit light. The most famous one of these enzymes is the firefly luciferase. Fireflies are able to emit light via a chemical reaction in which luciferin is converted to oxyluciferin by the luciferase enzyme. Some of the energy released by this reaction is in the form of light.
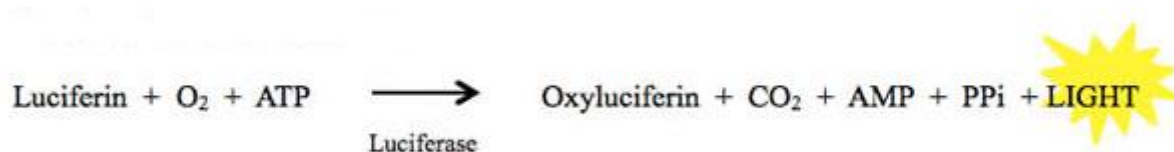
$$Luciferin + O_2 + ATP \xrightarrow{Luciferase} Oxyluciferin + CO_2 + AMP + PPi + LIGHT$$

Figure 6. Luciferase reaction
(http://bitesizebio.com/10774/the-luciferase-reporter-assay-how-it-works/)

This reaction is highly energetically efficient, meaning nearly all the energy put into the reaction is rapidly converted to light. This makes it extremely sensitive, which is great for a reporter assay.

This reporter assay can be used to study gene expression as well as other cellular components and events that are involved in gene regulation. Its extreme sensitivity allows quantification of even small changes in transcription, and the availability of results within minutes of completing your experiment makes it even more appealing.
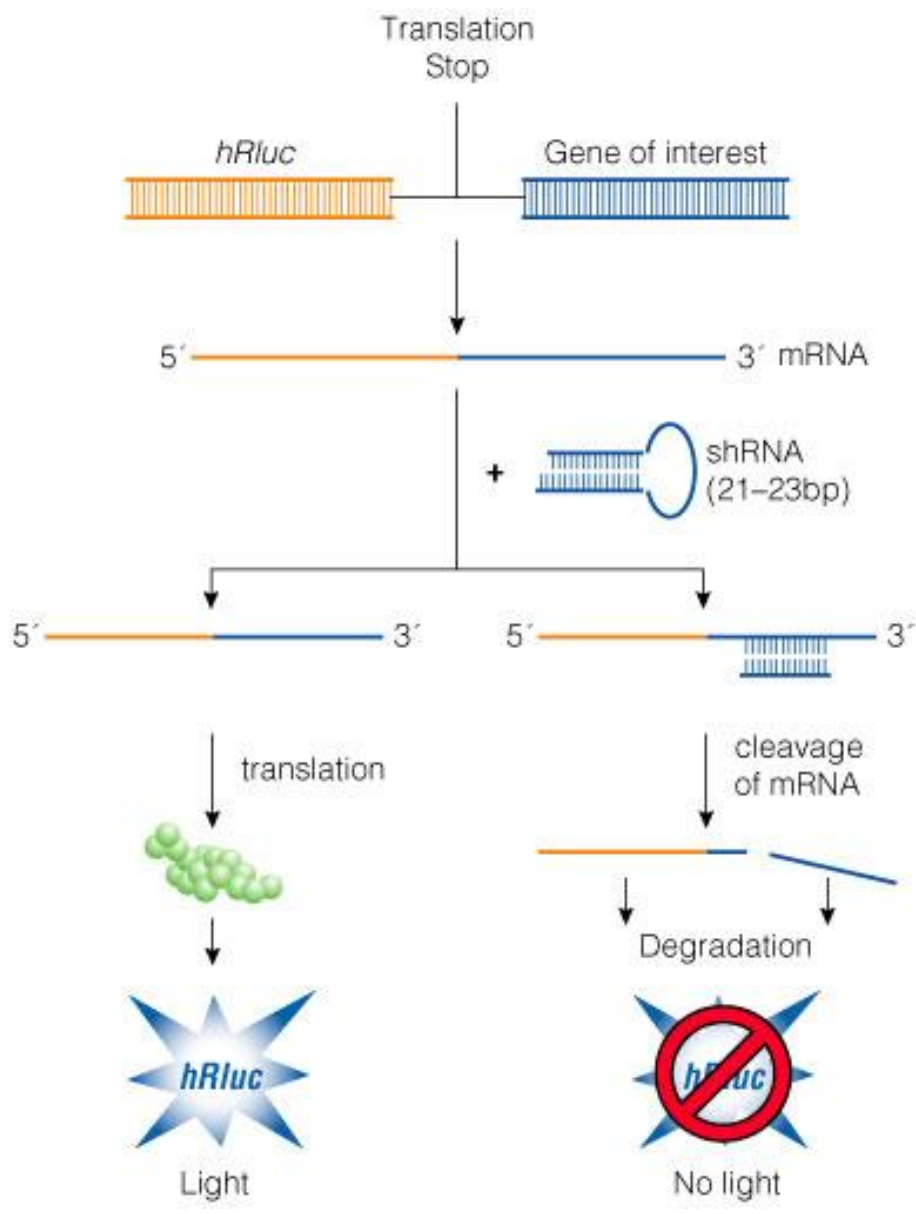
13

Figure 7. Luciferase reporter assay method
(http://mirtarbase.mbc.nctu.edu.tw/images/Luciferase_assay.jpg)

14

1.4.2 Techniques in Gene Level and Protein Level

1.4.2.1 Western Blot

The western blot (sometimes called the protein immunoblot or immunoblotting) is a common analytical technique used to detect specific proteins in a sample. It was introduced in the laboratory of Harry Towbin at the Friedrich Miescher Institute in 1979. The term "blotting" refers to the transfer of biological samples from a tissue homogenate or extract to a membrane and their detection on the surface of the membrane.

The procedure uses gel electrophoresis in order to separate the proteins of the sample by size. Then, the separated proteins are transferred to a solid support, generally a nitrocellulose or polyvinylidene difluoride (PVDF) membrane. The main method for the transfer is electroblotting, using an electric current to pull proteins from the gel into the membrane. The proteins move onto the membrane, while they maintain the arrangement they had within the gel.

After transfer, the target protein will be detected using appropriately matched and labeled antibodies. The typical detection procedure involves three distinct stages. The first stage describes how the blot prevents the antibodies to bind to the membrane. During the second stage, antibody incubation, the appropriately labeled antibody binds to the target protein following two different methods, direct detection and indirect detection. Finally, the label of antibody will be detected by a substrate in order to identify the location of the target protein on the blot.

Since western blotting is a quick procedure, which uses simple equipment and inexpensive reagents, while producing clear and unambiguous results, it is one of the most common laboratory techniques. An improved immunoblot method, Zestern analysis, is able to address this issue without the electrophoresis step, thus significantly improving the efficiency of protein analysis.

1.4.2.2 qRT-PCR experiment

Quantitative reverse transcription PCR (qRT-PCR) or quantitative real-time PCR is a high sensitive method for identification of microRNAs. qRT-PCR experiment involves RNA isolation, cDNA synthesis and verifies miRNA using

15

miRNA-specific primers. Moreover, there are two methods for miRNA expression analysis, the stem-loop method and the poly(A) method (Kang et al., 2012).
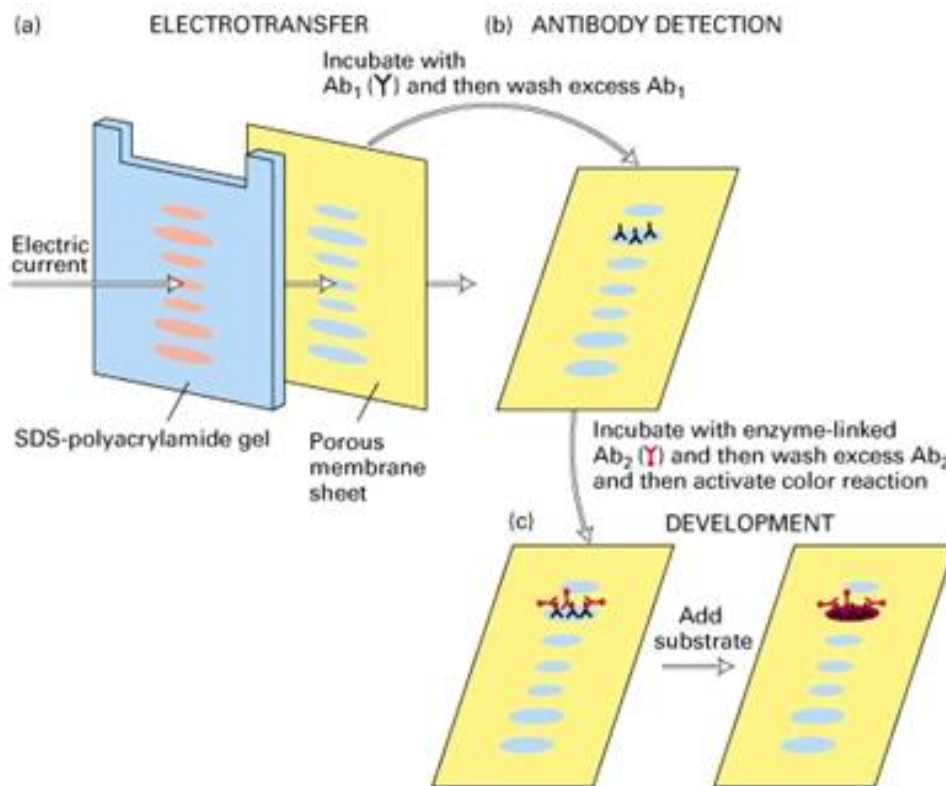

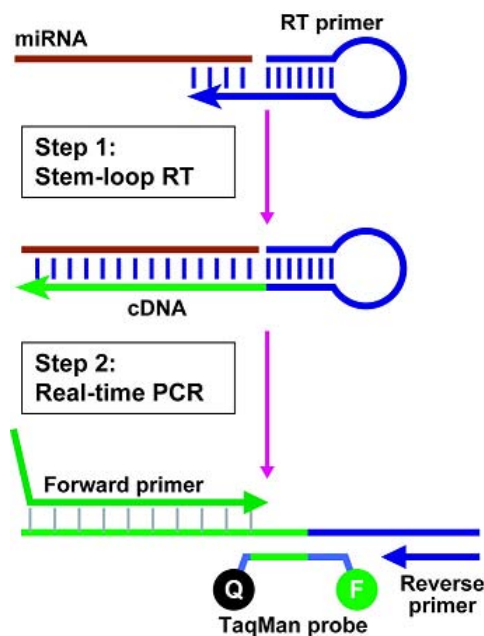**Figure 8.** Western blot procedure after gel electrophoresis. *(http://mirtarbase.mbc.nctu.edu.tw/images/Western.jpg)*

Figure 9. Schematic description of TaqMan miRNA assays. *(Caifu Chen, 2005)*

16

1.4.3 High Throughput Experiments

High throughput methods have been developed in order to identify miRNA targets by using several experimental techniques. These methods provide to scientists useful information about miRNA target sites and they use them as training sets to determine miRNA-mRNA interactions. High-throughput approach is based especially on producing many short target sequences simultaneously. High-throughput experiments enable to choose multiple miRNAs in an individual experiment and continue with parallel target verification. These experiments are more efficient and sensitive than others and are developed to reduce the cost of DNA and RNA sequencing. Some of the most frequently used high-throughput experiments are described below in detail.

1.4.3.1 Microarrays

Microarray experiments are used to indirectly detect miRNA-mRNA interactions by not observing the direct interaction between a single miRNA and mRNA sequence but the changes occurred in gene expression. A microarray is a high throughput tool that is able to measure the expression levels of multiple genes concurrently in a single experiment. Before the beginning of microarray analysis scientists assume that there is a large number of target genes and the expression level of every gene is stable. Compared to other high throughput methods microarray experiments are cheaper. The main principle of microarrays is the process of hybridization. During that process, a double stranded molecule is constructed, consisting of two complimentary sequences of DNA or one of DNA and RNA. The experiment begins with the isolation of mRNA in order to define the control sample and the experiment sample. The second step is to reverse transcribe and label the mRNA. The next step is to hybridize the labeled target mRNA to the microarray and finally combine equal amounts of target sequences and scan the microarray.
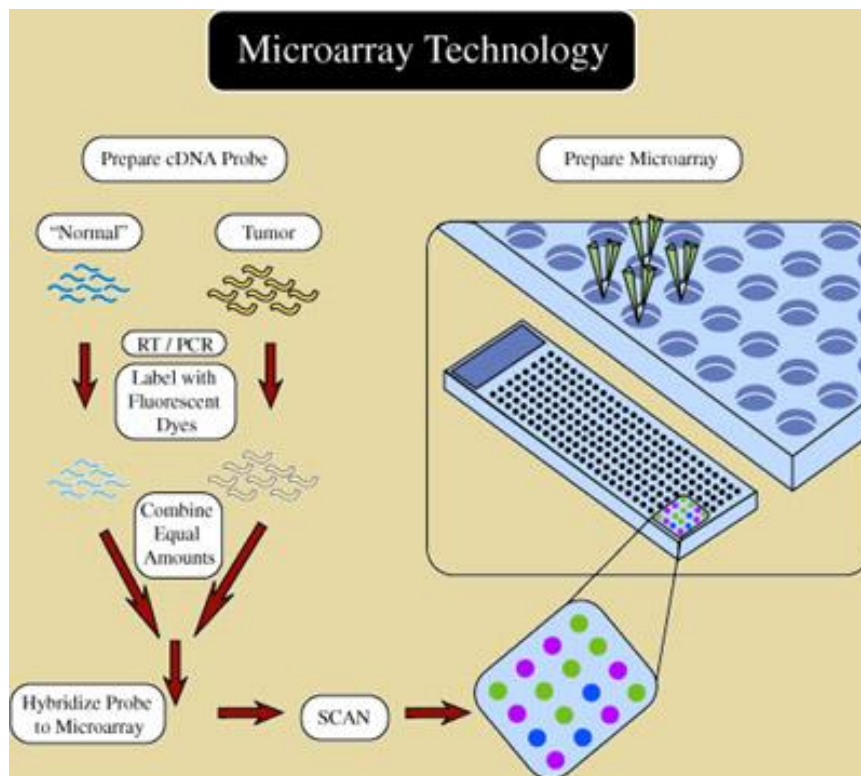
Figure 10: Microarray experiment. *(http://mirtarbase.mbc.nctu.edu.tw/php/help.php)*

1.4.3.2 Proteomics

The proteome is the set of proteins expressed by a genome at a specific time. A variety of proteomic methods have been developed for identification of miRNA targets. A proteomic experiment includes isolation of proteins from cells and the production of fractionated samples. The benefit of that process is that after changes in miRNA expression, it is easy to detect directly the changes in protein synthesis. A specific example of that experimental category is the method of pSILAC, stable isotope labeling by amino acids in cell culture, based on mass spectrometry. It is a popular method for quantitative proteomics.
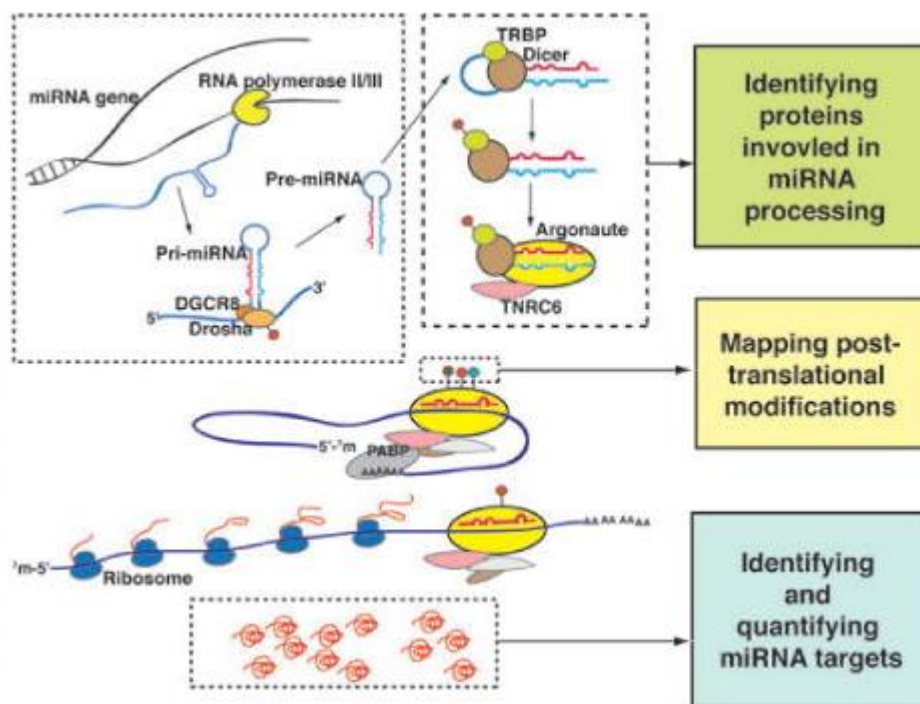
18

Figure 11: Proteomic experiment.
(http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3715049/?report=classic)

1.4.3.3 Sequencing

Among high-throughput techniques for identification miRNA-mRNA interaction is deep sequencing. Recent methods, such as photoactivable-ribonucleoside-enhanced crosslinking and immunoprecipitation, (PAR-CLIP) and high-throughput sequencing coupled with ultraviolet crosslinking and immunoprecipitation (HITS–CLIP), are based on detecting direct RNA-protein interaction sequences by immunoprecipitation of Ago protein but PAR-CLIP is more efficient in the of crosslinking between the RNA and the proteins and RNA recovery.[] A PAR-CLIP experiment involves a photoactivatable crosslinker, usually 4-thiouracil which is incorporated into transcripts and then are crosslinked with RNA binding proteins by UV radiation. The location of the crosslinks is now identified and the target sites are easily specified due to thymidine-cytidine transitions.
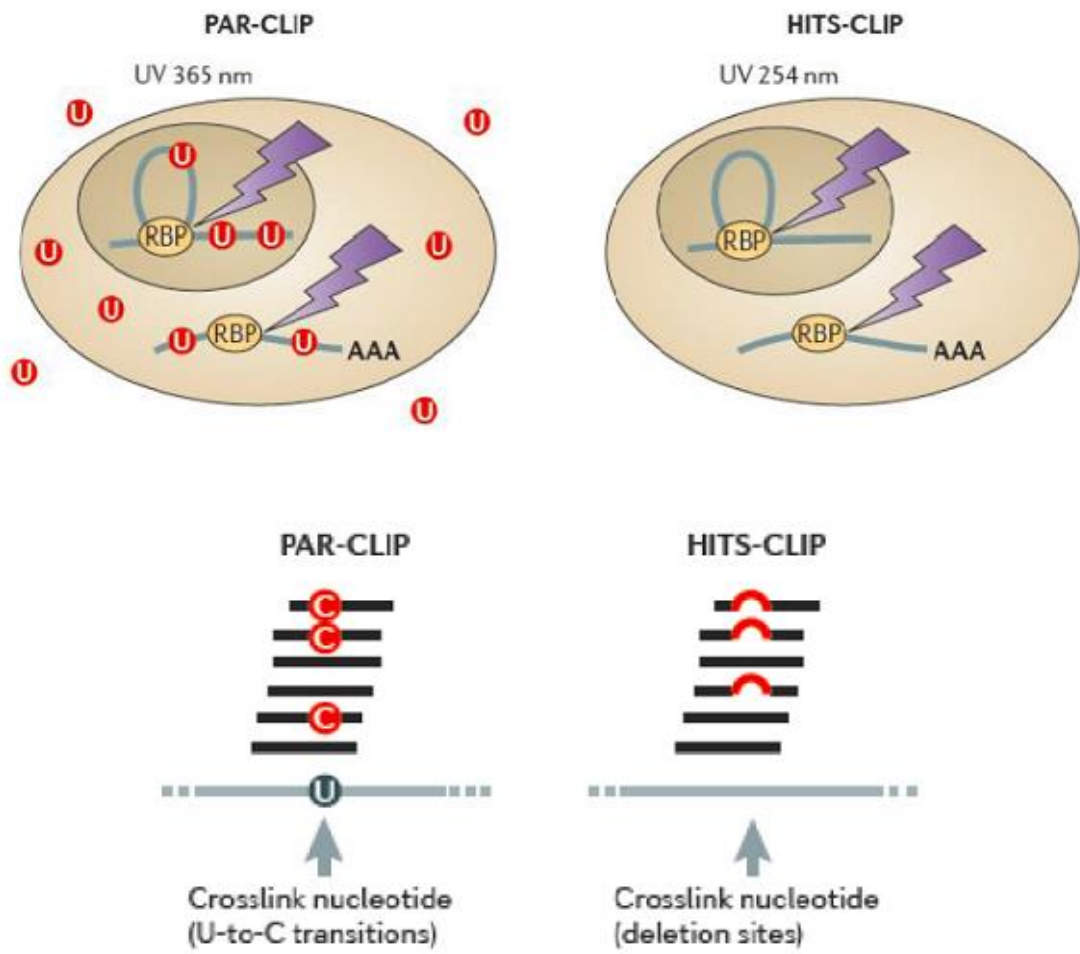
19

Figure 12: Outline of the PAR-CLIP experiment. *(König J et al., Nat Rev Genet., 2012 Jan)*

1.4.3.4 Clash

Another high-throughput based technique for miRNA identification is CLASH (cross linking, ligation and sequencing of hybrids), which was developed to directly map the miRNA-mRNA interactions in deep-sequencing data. A typical CLASH experiment involves identification of both target site and miRNA using UV crosslinking, purification of AGO and ligation of base paired RNA's. These target sites are used as a learning target set to compare other miRNA-mRNA interactions.
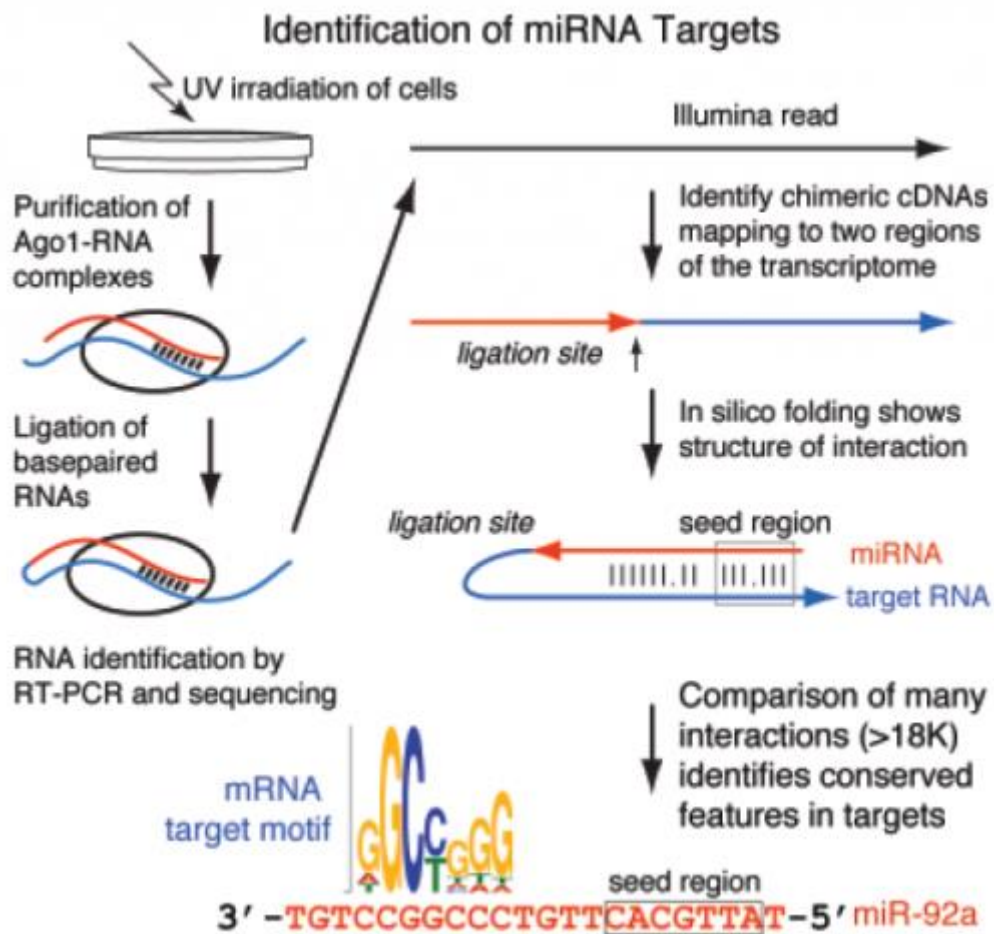

Figure 13: CLASH experiment. *(http://www.wcb.ed.ac.uk/research/tollervey)*

1.5 MicroRNA and microRNA target databases

Since 2001, when the term "microRNA" was formally introduced, various databases have been created to store miRNA information, such as their target mRNAs and their functionalities. In addition, the considerable growth of validation experiments raised the need for a database to store the results in a uniform way. These databases provide an efficient access to the great amount of experimental data and make them publicly available.

Databases such as those mentioned are miRBase (an archive of miRNA sequences and annotations) and also TarBase, miRTarBase, miRecords, miR2Disease, MirnaMAP, miRSel, TargetScan, miRwalk and StarBase (with the first four being the most comprehensive and most frequently updated). Below is a brief presentation of these five databases which we used in our research.

1.5.1 miRBase

miRBase (Kozomara and Griffiths-Jones, 2013) is the public repository for all published microRNA sequences and associated annotation. It was established in 2002 (then called the MicroRNA Registry), with the main purpose of assigning stable and consistent names to newly discovered microRNAs. MiRBase acts as a central repository for all published microRNA sequences, and also facilitates online searching and bulk download of all microRNA information. Moreover, the database aggregates microRNA validations and target predictions.

The newly discovered miRNAs are submitted to the database, when the article describing their identification is accepted for publication in a peer-reviewed journal. Then, miRBase assigns an official gene name to be used in the published version of the article. miRBase microRNA gene names are in the form of *dme-mir-100*. The prefix signifies the organism (in this case *Drosophila melanogaster*), while, the numbers are assigned sequentially.

The latest version of miRBase (v21, released in June 2014) contains 28.645 microRNA loci from 223 species, processed to produce 35.828 mature microRNA products.
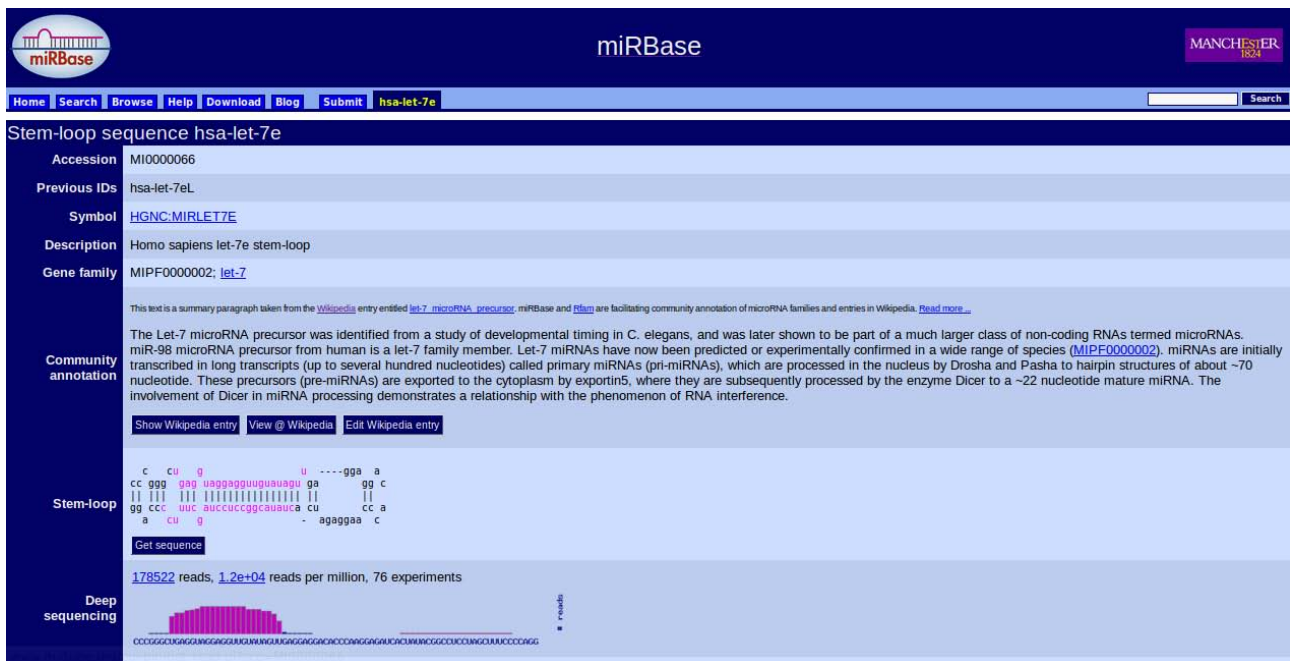
22

Figure 14. Screenshot of miRBase database, showing information about hsa-let-7e miRNA.
(http://www.mirbase.org/cgi-bin/mirna_entry.pl?acc=MI0000066)

1.5.2 TarBase

DIANA-TarBase v1.0 was the first available database of experimentally validated miRNA targets, created in 2006. Currently, the sixth version is the largest available manually curated target database, indexing more than 65,000 miRNA-gene interactions (Vergoulis et al., 2012). The database includes targets for 21 species, derived from specific, as well as high throughput experiments, such as microarrays, proteomics, HITS-CLIP and PAR-CLIP. Each target site is described by the miRNA that binds it, the gene in which it occurs, the nature of the experiments that were conducted to test it, the sufficiency of the site to induce translational repression and/or cleavage, and the paper from which all these data were extracted.

In the latest version of TarBase users can contribute to the database by submitting their own data, derived from their publications. TarBase also includes powerful searching and filtering capabilities, while is available for download for free.

Additionally, the database is functionally linked to several other relevant and useful databases such as Ensembl, Hugo, UCSC and SwissProt.

23

Figure 15. Screenshot of TarBase v6.0 database, depicting experimentally validated targets of hsa-let-7e-5p miRNA.

(http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=tarbase/index&mirnas=hsa-let-7e-5p)

1.5.3 miRTarBase

The miRTarBase database (Sheng-Da Hsu et al., 2014) was launched in 2010 with data sources for more than 100 published studies in the identification of miRNA targets, molecular networks of miRNA targets and systems biology. After the latest update (version 4, released in 2013), the database includes significant expansions and enhancements over the initial release.

As a database, miRTarBase has accumulated more than 50.000 miRNA-target interactions (MTIs), which are collected by manually surveying pertinent literature after data mining of the text systematically to filter research articles related to functional studies of miRNAs. Generally, the collected MTIs are validated experimentally by reporter assay, western blot, microarray and next-generation sequencing experiments.

24

Figure 16. Screenshot of miRTarBase database, depicting experimentally validated targets of hsa-let-7e-5p miRNA and the methods from which they resulted. *(http://mirtarbase.mbc.nctu.edu.tw/php/search.php?org=hsa&opt=mirna_id&kw=hsa-let-7e-5p&miFam=0)*

1.5.4 miRecords

miRecords (Xiao et al., 2009) is a database developed in 2008 by the group of Dr. Tongbin Li from the Department of Neuroscience, University of Minnesota. The validated targets component of this resource hosts a large, high-quality manually curated database of experimentally validated miRNA–target interactions with systematic documentation of experimental support for each interaction. The current release of this database includes 2.705 records of validated miRNA–target interactions between 644 miRNAs and 1.901 target genes in nine animal species. Among these records, 2.028 were curated from "low throughput" experiments.

miRecords consists of two components. The "Validated Targets" section is a large, high-quality database of experimentally validated miRNA targets resulting from meticulous literature curation. The "Predicted Targets" section is an integration of predicted miRNA targets produced by 11 established miRNA target prediction programs (DIANA-microT, MicroInspector, miRanda, MirTarget2, miTarget, NbmiRTar, PicTar, PITA, RNA22, RNAhybrid and TargetScan).

25

Figure 17. Screenshot of miRecords database, depicting experimentally validated targets of hsa-let-7e miRNA and the target prediction programs that predicted each interaction. (*http://c1.accurascience.com/miRecords/interactions.php?species=Homo+sapiens&mirna_acc=hsa-let-7e&targetgene_type=refseq_acc&targetgene_info=&v=yes&search_int=Search*)

1.5.5 miR2Disease

miR2Disease (Jiang et al., 2009) is a manually curated database, created in April 2008, which aims at providing a comprehensive resource of miRNA deregulation in various human diseases. Each entry in the miR2Disease contains detailed information on a miRNA-disease relationship, including miRNA ID, disease name, a brief description of the miRNA-disease relationship, miRNA expression pattern in the disease state, detection method for miRNA expression, experimentally verified miRNA target genes and literature reference. All entries can be retrieved by miRNA ID, disease name or target gene.

It is a collaboration project between the Harbin Institute of Technology (HIT) and the Center of Computational Biology and Bioinformatics at the Indiana University of Medical School. The current version of miR2Disease documents 3.273 curated relationships between 349 human microRNAs and 163 human diseases.

26

Figure 18. Screenshot of miR2Disease database, listing information about a specific experimentally validated target of hsa-let-7e miRNA. *(http://www.mir2disease.org/)*

1.6 Aim of the study

Since the discovery of miRNAs more than 10 years ago, there has been a huge progress in the identification of new miRNAs, as well as their respective targets. Their functions, involved in regulating the gene expression, caused the emerge of new bioinformatics resources to identify their targets. These resources include mainly computational methods of target prediction and experimental techniques validating the putative targets. Moreover, various databases have been created, storing the experimentally validated miRNA targets. Their development offered essential aid to the miRNA research community, as they constitute an effective tool for validation and assessment of the target prediction programs. However, as each database evolves independently and the number of new databases grows larger, an increasing level of specialization is

observed. On the other hand, as the databases' content is derived from the same literature and experimental data, the repetition and the overlap of the entries are inevitable, as well as redundancy. In this thesis we aim to compare some of the most significant databases (regarding their content and other features), giving a graphical association of the databases. We focused the comparison mainly in the number of interactions per experiment category, as well as the number of publications, from which these data were extracted. The present review can be an asset for any researcher interested in a more thorough study of these databases and their characteristics.

Methods

In this chapter we will elaborate on the way we compared the databases which contain the experimentally validated miRNAs. Initially we referred to the database official websites, downloading the corresponding files for each database. Namely, we downloaded the TarBase.xls file from http://diana.cslab.ece.ntua.gr, the miRTarBase.xls file from http://mirtarbase.mbc.nctu.edu.tw, the miRecords_version4.xls file from http://mirecords.biolead.org, and the miRtar.txt file from http://www.mir2disease.org. Accordingly, we converted any .txt files into .xls format in order to facilitate the reading and comparison of the data. Every file contains the headings to each column of information data, such as Gene Name, Spieces (miRNA), Target Gene and References. Each row below that represents a gene-miRNA interaction validated by an experiment. The experiments mentioned in each database constitute four groups: high- throughput techniques, reporter gene assay, techniques in protein/gene level, and other experiments. More specifically, the high-throughput techniques group consists of the experiments: Microarray, ClASH, Proteomics, Sequencing, Immunoprecipitation, Quantitative proteomic approach, ChIP-seq, Degradome, CLIP-seq, pSILAC, Mass spectrometry, Branched DNA probe assay, 2DGE. As for the techniques in protein/gene level the experiments are: Western blot, Northern blot, qRT-PCR, Flow, Immunocytochemistry, ELISA, intrarenal expression, Immunohistochemistry, Immunofluorescence, Immunoblot, In situ hybridization, FACS, RTPCR, RNase mapping analysis, EMSA, ASO assay, semi-qRT-PCR, quantitative PCR, Ribonuclease protection assay, mRAP, Real time PCR. Finally, for reporter gene assay category the experiments are: Luciferase reporter assay, Reporter assay, GFP reporter assay, phenotypic sensor assay, GUS reporter assay, LacZ reporter assay, B-globin reporter assay, EGFP reporter assay.

The fields used in our analysis are the miRNA, the Target Gene and the Experiment. Additionally, we looked into the Publications field to determine the distinct publications of each database and the experiment groups to which they correspond. In the case of miRecords database we had to modify the file, in order to make the Experiment field consistent with the respective fields of the other databases. Moreover, we deleted any double entries by removing repeated interactions from the databases, thus ensuring the accuracy of the comparison results.

We used Perl programming language (Perl 5, version 14) for parsing the .xls

files and obtaining our results, on a machine running Ubuntu Linux OS.

In order to run the Perl scripts we had to install the modules **Spreadsheet::ParseExcel** and **Spreadsheet::WriteExcel**. A more comprehensive analysis of the code we implemented follows.

In order to calculate the number of miRNA-target interactions in each database, we created a Perl script, which took as an input the .xls file with the entire content of the database. We did this for the three databases (TarBase, miRTarBase, miRecords) that include targets verified with more than one experiment category, as miR2Disease database's entries are all validated with Luciferase reporter assay.

The Spreadsheet::ParseExcel module can be used to read information from Excel 95-2003 binary files. The advantage is that Spreadsheet::ParseExcel is compatible with Linux and Windows OS. The **new()** method is used to create a new Spreadsheet::ParseExcel parser object. The spreadsheet is parsed into a top-level object called **$parser**. The workbook contains several worksheets. We iterate through them by using the workbook **worksheets()** property. Each worksheet has a **MinRow** and **MinCol** and corresponding **MaxRow** and **MaxCol** properties, which can be used to figure out the range the worksheet can access. Cells can be obtained from a worksheet through the **Cells** property. Each statement in the switch block is labeled with the experiments of the main categories high-throughput experiments, reporter gene assay, techniques in gene/protein level and other experiments. We declared four scalar variables (**my counter1** to **my counter4**), to store the number of times each case block is accessed. Function **get_cell($row, $col)** returns the "Cell" object at row **$row** and column **$col** if it is defined. Otherwise returns **undef**.

For the calculation of the number of publications in each database, we created four Perl scripts. There, we worked as before except that, apart from the counter variables, we declared five arrays in order to test if a Pubmed ID was already entered before. The four arrays **(@pubarray1** to **@pubarray4**) store the Pubmed IDs of each publication and **@pubarray5** stores the total sum of publications. We used Perl's **push()** function to push Pubmed ID values onto the end of a @pubarray, while we checked if a value was already in the array, using Perl's **grep()** function. At the end, we returned the size of the arrays by using them as scalars. An array used in a scalar context returns its size.

The first three Perl scripts can all be found in the Appendix of this thesis, while from the other four, we enclose those written for TarBase and miRTarBase, as the most distinctive.

30

Results

We present the results obtained from microRNA databases and running Perl code. The complete list of experimentally validated microRNA-target interactions are given in Table 1.

	TarBase	miRTarBase	miR2Disease	miRecords
Created in	Dec 2005	Oct 2010	Apr 2008	Nov 2008
Current version Last update	v6 Jan 2012	v4.5 1 Nov 2013	 2 Jul 2010	v4 27 Apr 2013
Different species	21	18	1	9
Number of miRNA-target interactions	32.318	52.462	805	2.189
Number of miRNA-target interactions validated by **reporter gene assay** experiments	819	4.585	805	974
Number of miRNA-target interactions validated by **techniques in gene level and protein level**	446	3.764	0	1.233
Number of miRNA-target interactions validated by **high-throughput techniques**	29.999	47.616	0	800
Number of miRNA-target interactions validated by **other** experiments	4.070	935	0	698

Table 1.

31

	TarBase	miRTarBase	miR2Disease	miRecords
Number of publications	1.104	2.210	433	867
Number of publications for **reporter gene assay** experiments	105	1.192	433	423
Number of publications for **techniques in gene level and protein level**	76	1.625	0	580
Number of publications for **high-throughput techniques**	976	363	0	18
Number of publications for **other experiments**	154	167	0	222

Table 2.
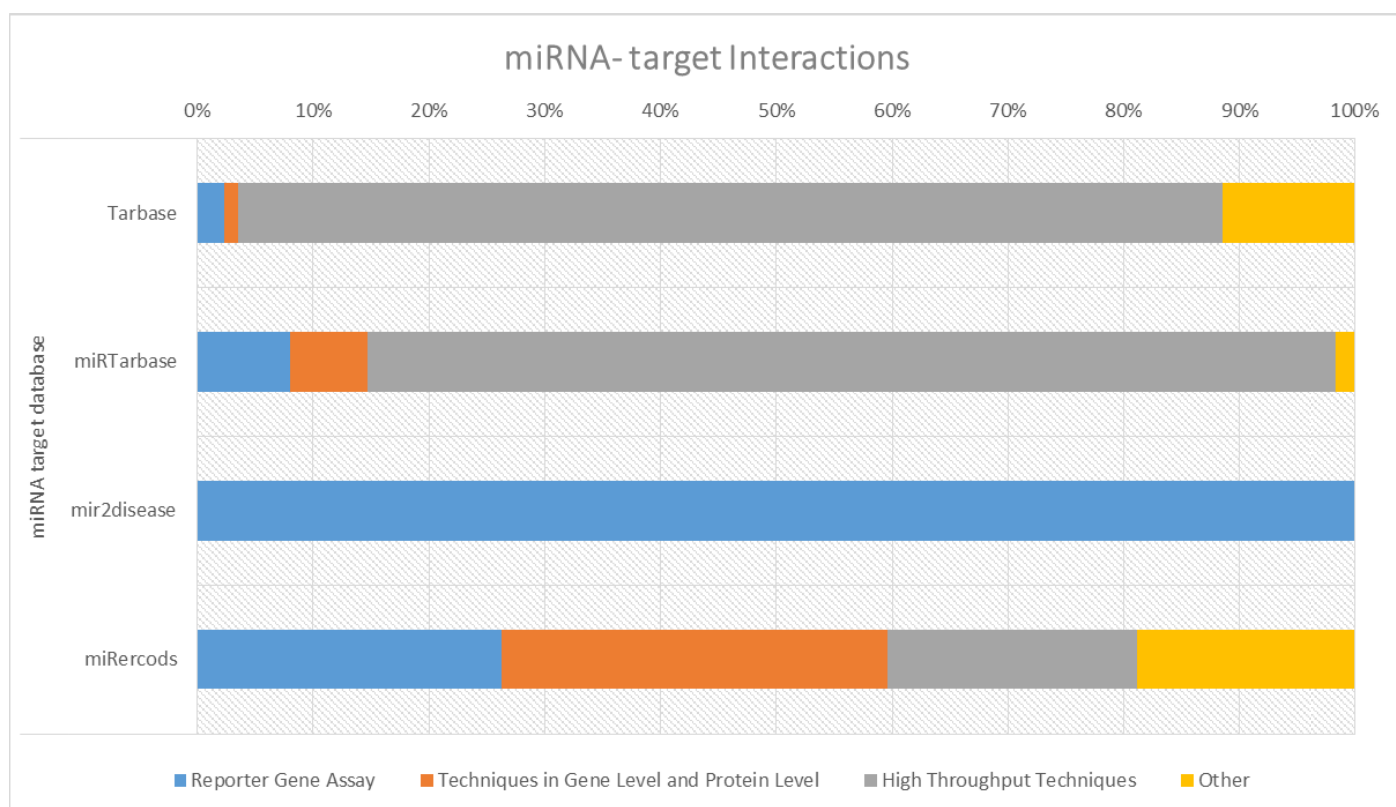
Chart 1. Total number of miRNA-target interactions

32

In chart 1, we observe that the majority of microRNA-target interactions in the first two databases, which also have the most interactions in total, resulted from high throughput experimental methods. This was expected as these methods produce more results than the others. On the contrary, in the other two databases the majority of interactions are the result of low throughput methods (miR2Disease database contains exclusively Reporter Gene Assay experimental results).
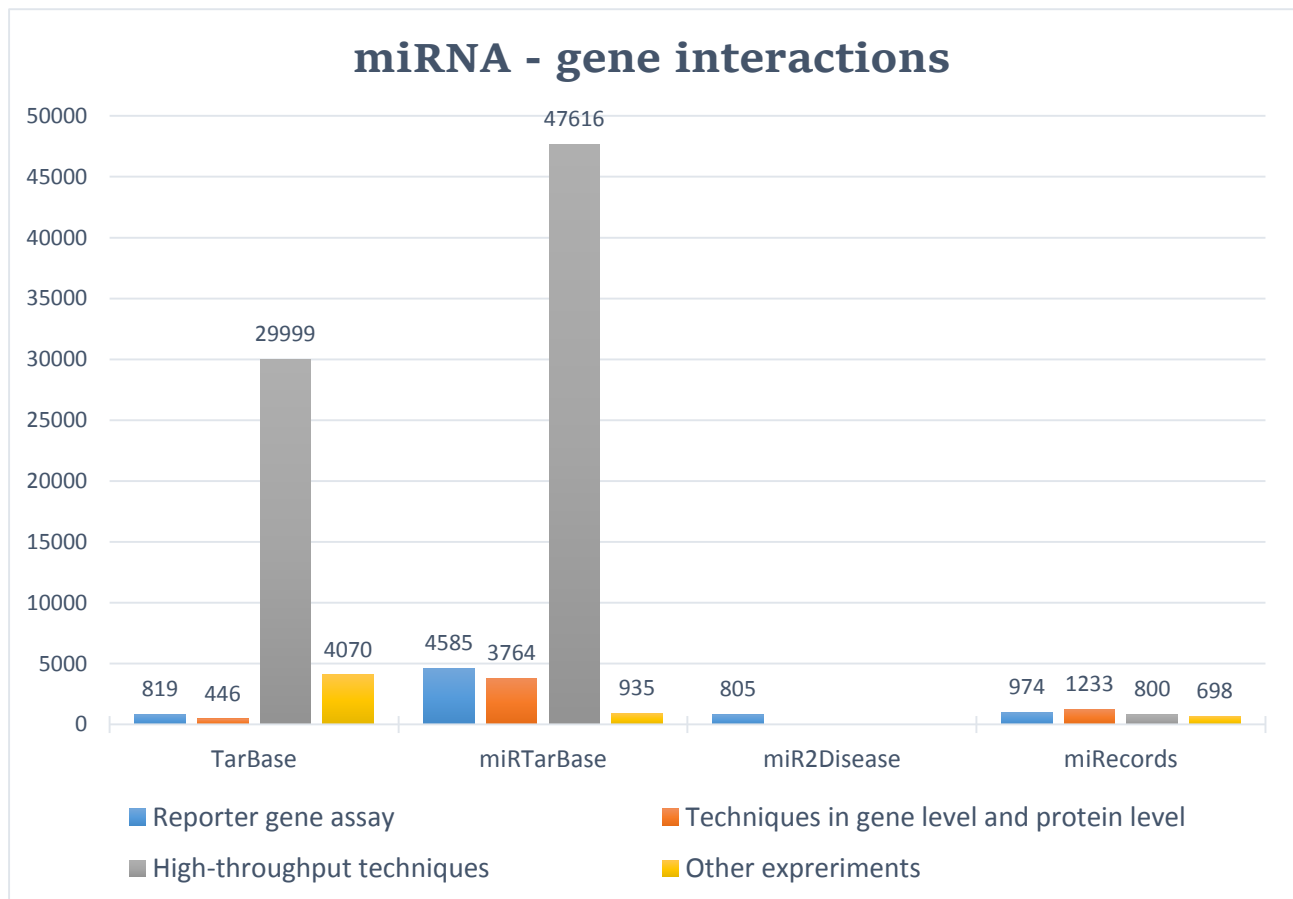
Chart 2.

Chart 2 depicts the total microRNA-target interactions of each database, per experimental method. The dominance of miRTarBase regarding the amount of interactions is evident. This database contains more interactions than the other three in the Reporter Gene Assay, Gene/Protein Level, and High Throughput Techniques. Nevertheless, TarBase has more microRNA-target interactions experimentally validated by a method other than these three.

Reporter gene assay

TarBase	819
miRTarBase	4585
miR2Disease	805
miRecords	974
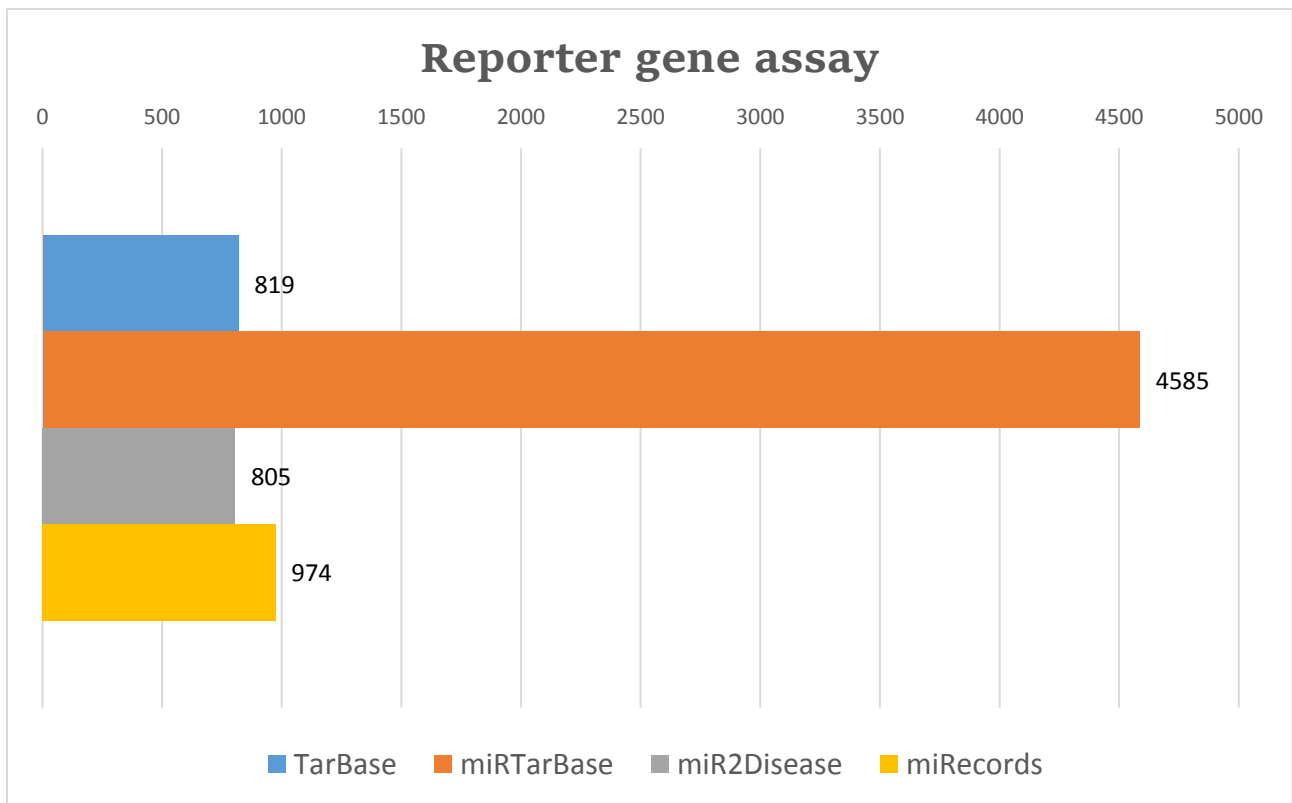
Chart 3.

Chart 3 depicts the miRNA-target interactions that resulted from the Reporter Gene Assay. We observe the wide margin between the miRTarBase interactions and the other databases. In all databases the luciferase reporter assays are the majority of reporter gene assays. MiR2Disease specifically consists only of luciferase experiments. Moreover, 84% of miRTarBase reporter assays (3832) are also luciferase assays, while GFP reporter assays come second in number (242; 5.3% of all the reporter assays).

Chart 4 is the respective diagram for Gene/Protein Level Techniques. Mir2Disease is evidently absent as it contains exclusively Reporter Gene Assays. MirTarBase again dominates the rest regarding the total interactions. In this experimental category, TarBase and miRTarBase contain western blot and PCR experiments, while miRecords primarily contains western blot.

Chart 5 shows the comparison of TarBase and miRTarBase regarding experiments of gene/protein level techniques.

34

Techniques in gene level and protein level

TarBase: 446
miRTarBase: 3764
miRecords: 1233

■ TarBase ■ miRTarBase ■ miRecords

Chart 4.

TarBase and miRTarBase comparison in gene/protein level techniques

Western blot — Tarbase: 227, miRTarbase: 2607
PCR — Tarbase: 219, miRTarbase: 2538

■ Tarbase ■ miRTarbase

Chart 5.

35

High-throughput techniques

	Value
TarBase	29999
miRTarBase	47616
miRecords	800

Chart 6.

TarBase and miRTarBase comparison for high-throughput experiments

	Tarbase	miRTarbase
Microarrays	11255	12736
Proteomics	3392	3178
Sequencing	15352	14415

Chart 7.

36

Chart 8.

In Chart 6, miRTarBase has evidently the most interactions in the high throughput category. A large part of the 800 High Throughput interactions of miRecords resulted from microarrays.

Chart 7 compares the TarBase and miRTarBase databases' high throughput interactions, as these two have the largest number of this kind of interactions. The compared interactions are grouped in the three experimental categories Microarrays, Proteomics, and Sequencing. TarBase has more Proteomics and Sequencing interactions, while miRTarBase prevails in the Microarrays category.

Chart 8 shows all the high throughput experiments belonging to the miRTarBase and each one's number of interactions. The chart is in logarithmic scale.

37
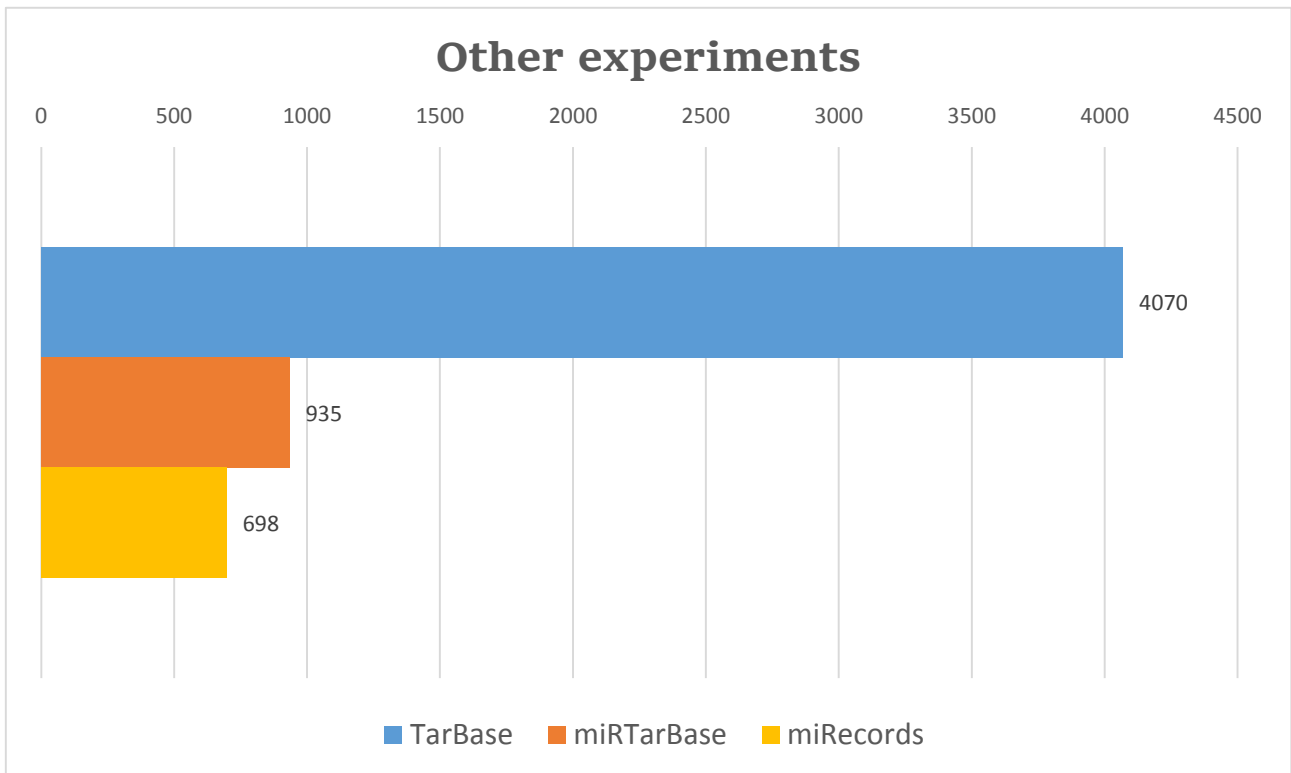
Chart 9.

In this diagram we can observe the number of interactions of each base verified by other experiments. TarBase database contains the most of this kind.

Chart 10.

38
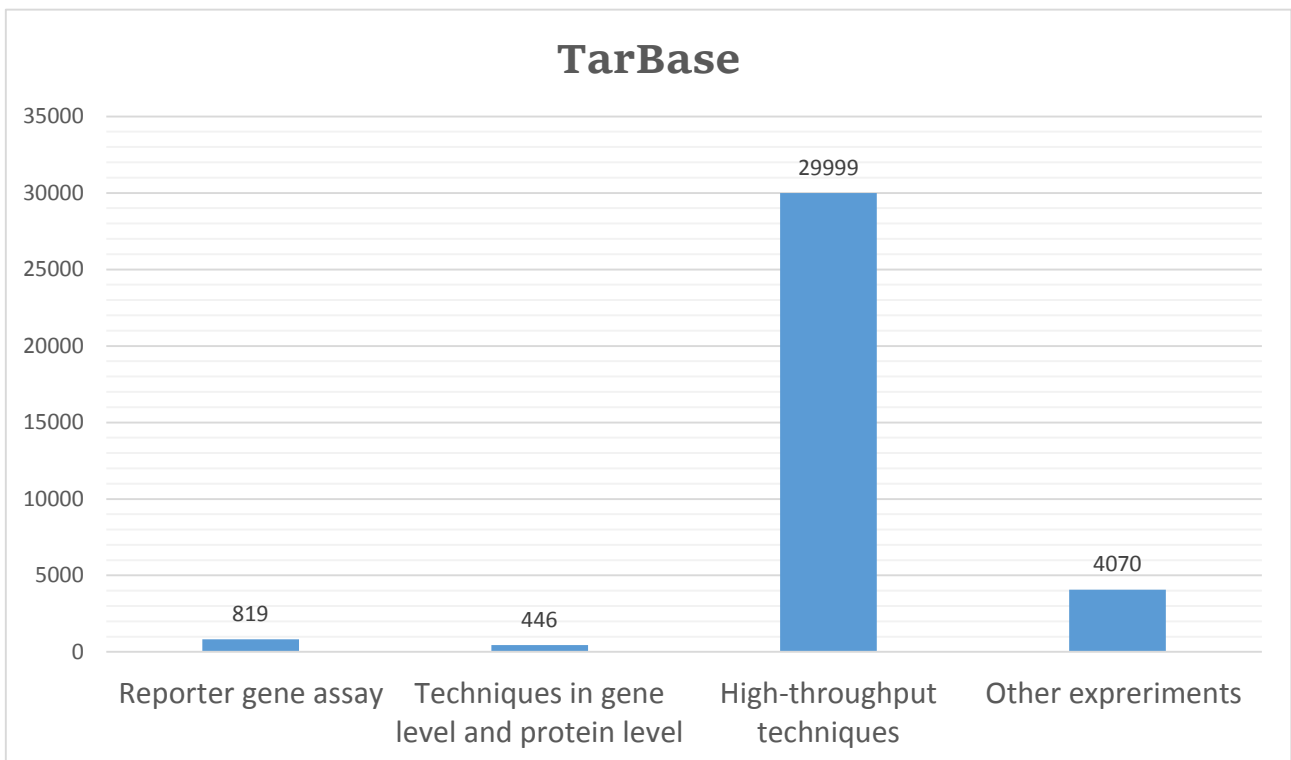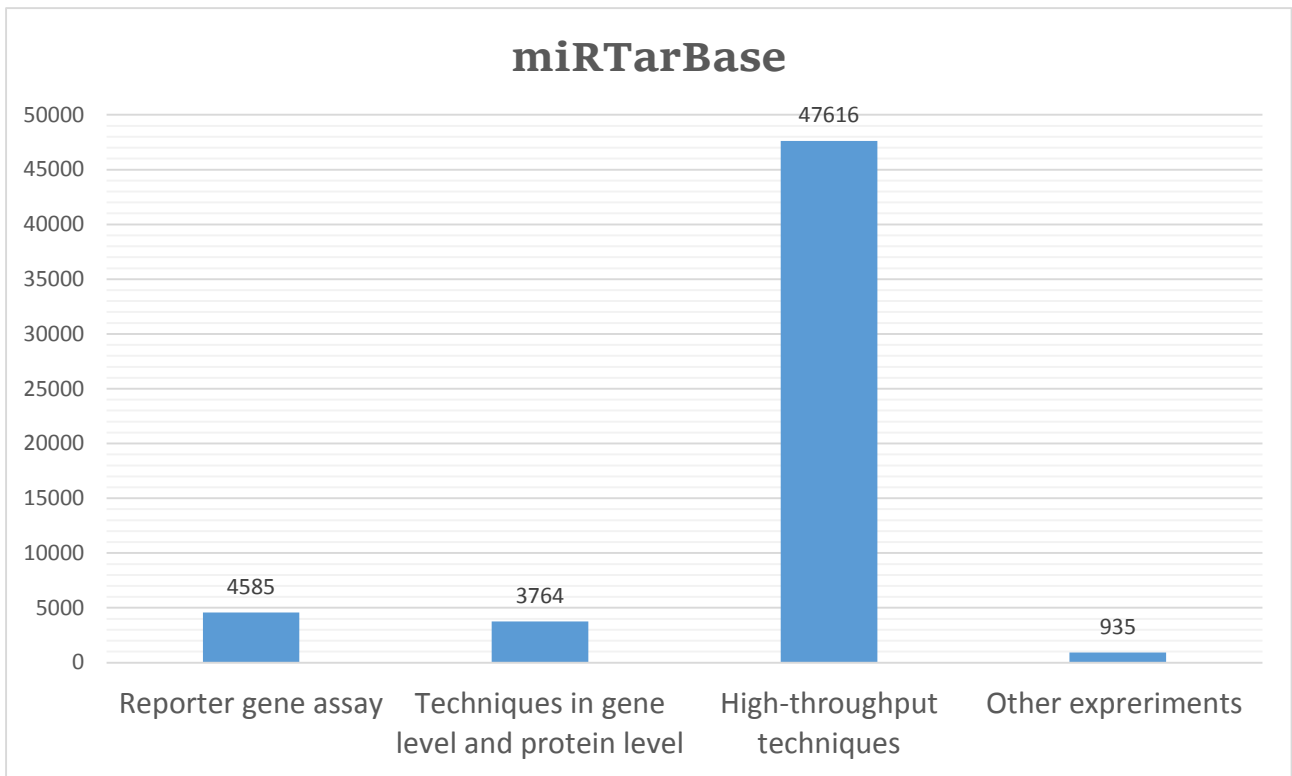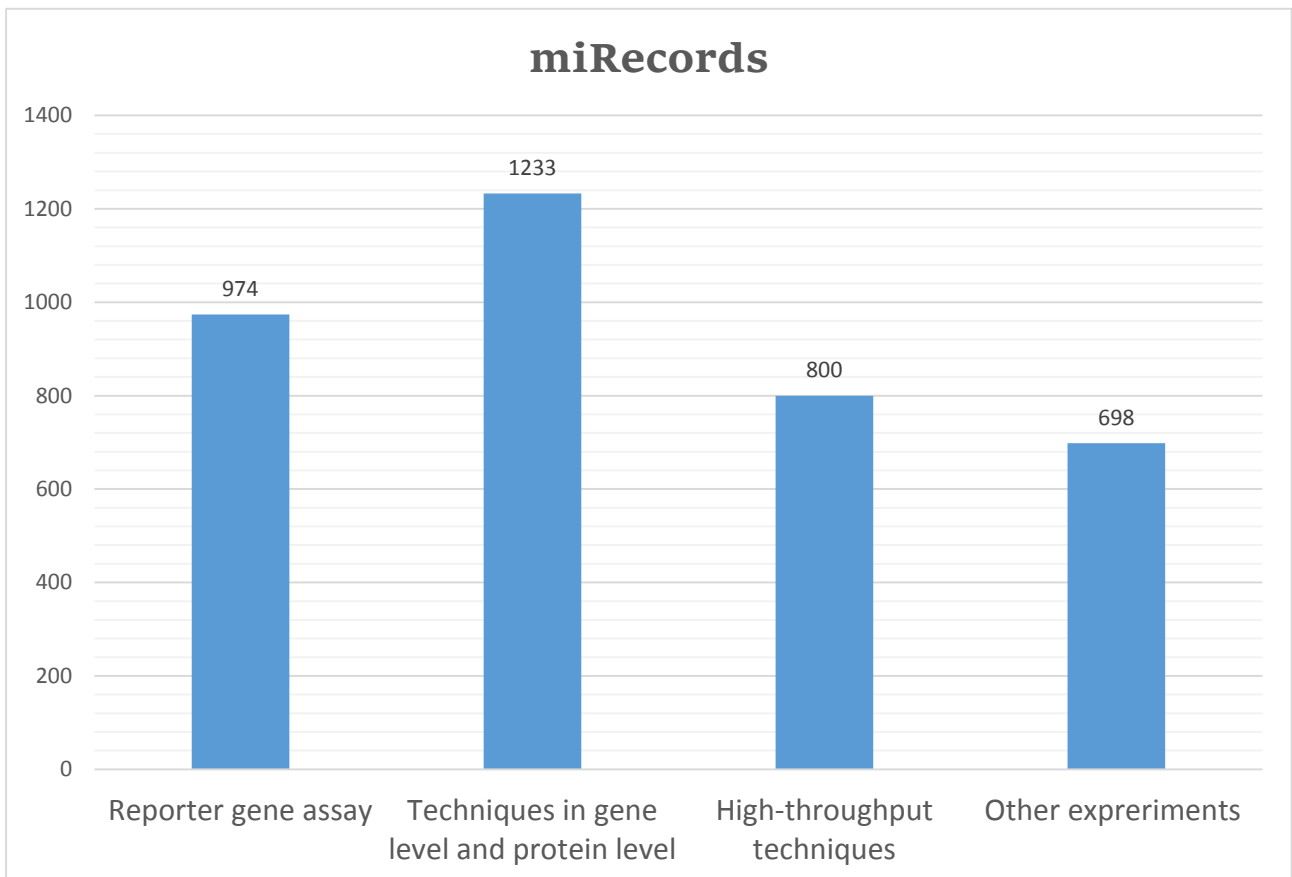
Chart 11.

Chart 12.

Charts 10, 11 and 12 present each database's total interactions by experimental method in a way that draws comparisons between the methods.
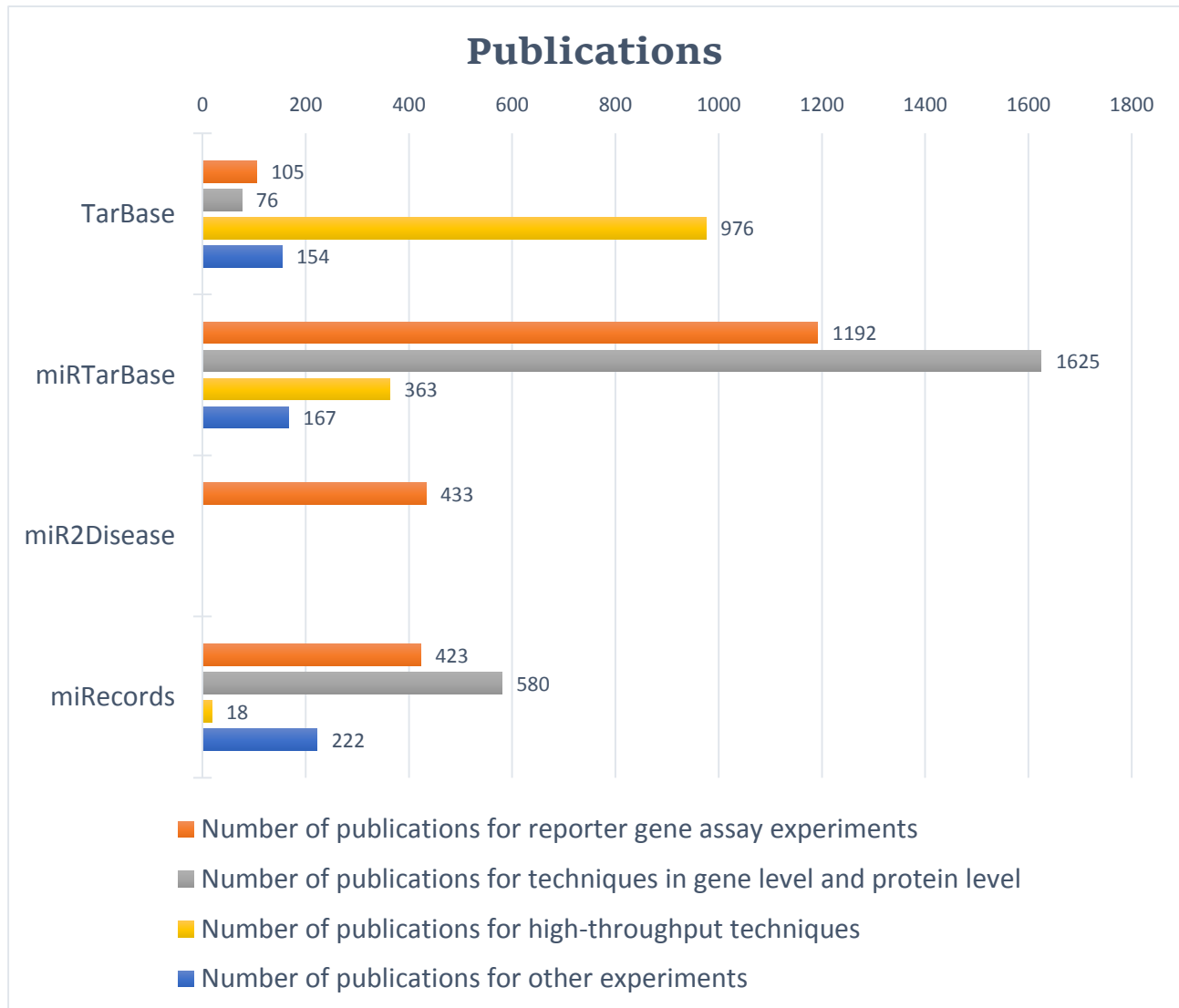
Chart 13.

This final chart shows the publications referring to the microRNA-targets contained in each database. The most publications relate to interactions verified by Gene/Protein Level techniques and contained in the miRTarBase database. We can note that there is a large ratio of high throughput interactions/publications as high throughput experiments result in more targets.

40

Conclusion and Discussion

The analysis of the results gives a graphical representation of the contents of each database and the characteristics that distinguish it from the others. Some databases are more focused on the quantity of the interactions that they contain, while others aim to specialization, including only targets validated with a particular category of experiments. The first approach can be useful for a prediction program that demands as input a large variety of experimental data, while the second looks more substantial if it concerns a category of experiments with more accurate identification of miRNA targets. Differences were also observed in the methods by which each database collects its interactions, as some are using text-mining techniques to retrieve the data from research literature, while others are manually curated, thus including more reliable interactions than the former. Moreover, frequent updates to the database, provide an additional advantage, as the content is enriched with the most current data.

Specifically, miRTarBase was found to be the database with the most interactions in total and in every experiment category. Taking advantage of the text-mining algorithms, miRTarBase has a great advantage in the total number of publications as well.

Concerning miR2Disease, it has limited interactions due to the fact that it specializes in only reporter gene assays but gives extra information on the relationship between microRNAs and diseases.

Tarbase as the first ever database of this kind, has the greatest variety of interactions among different species. It has the most total interactions from the databases that include strictly manually curated data, preserving a high reliable content. Additionally, it contains the greatest number of high-throughput techniques publications. Finally, being the oldest microRNA targets database has a major contribution in the field, influencing the development of other similar databases.

41

References

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P .2008. *Molecular Biology of the Cell*, Fifth Edition, Garland Science

Bartel DP. 2004. *MicroRNAs: genomics, biogenesis, mechanism, and function.* Cell. 116(2):281-297.

Bartel DP. 2009. *MicroRNAs: target recognition and regulatory functions*. Cell. 136: 215-233

Cai X, Hagedorn CH, Cullen BR; Hagedorn; Cullen. 2004. *Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs.* RNA. 10(12):1957-1966.

Gaidatzis D, van Nimwegen E, Hausser J, Zavolan M. 2007. *Inference of miRNA targets using evolutionary conservation and pathway analysis*. BMC Bioinformatics. 8(1):248.

Carthew RW, Sontheimer EJ. Feb 20, 2009. *Origins and Mechanisms of miRNAs and siRNAs*. Cell. 136(4): 642–655.

Chen K. and Rajewsky N. 2006. *Natural selection on human microRNA binding sites inferred from SNP data.* Nat Genet. 38(12):1452-1456

Gori M, Arciello M, Balsano C. 2014. *MicroRNAs in nonalcoholic fatty liver disease: novel biomarkers and prognostic tools during the transition from steatosis to hepatocarcinoma.* Biomed Res Int. 2014:741465.

Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP. 2007. *MicroRNA targeting specificity in mammals: determinants beyond seed pairing*. Mol Cell. 27(1):91-105.

Hsu SD, Tseng YT, Shrestha S, Lin YL, Khaleel A, Chou CH, Chu CF, Huang HY, Lin CM, Ho SY, Jian TY, Lin FM, Chang TH, Weng SL, Liao KW, Liao IE, Liu CC, Huang HD. 2014. *miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions.* Nucleic Acids Res. 42(Database issue):D78-85.

Hutvágner G, McLachlan J, Pasquinelli AE, Bálint E, Tuschl T, Zamore PD. 2001.

A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. Science. 293(5531):834-838.

Jiang Q, Wang Y, Hao Y, Juan L, Teng M, Zhang X, Li M, Wang G, Liu Y. 2009. *miR2Disease: a manually curated database for microRNA deregulation in human disease.* Nucleic Acids Res. 37(Database issue):D98-104.

John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS. 2004. *Human MicroRNA targets*. PLoS Biol. 2(11):e363

Kang K, Peng X, Luo J, Gou D. 2012. *Identification of circulating miRNA biomarkers based on global quantitative real-time PCR profiling.* J Anim Sci Biotechnol. 3(1):4.

Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E. 2007.*The role of site accessibility in microRNA target recognition*. Nat Genet. 39(10):1278-1284.

Ketting RF, Fischer SE, Bernstein E, Sijen T, Hannon GJ, Plasterk RH. 2001. *Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C. elegans.* Genes Dev. 15(20):2654-2659.

Kiriakidou M, Nelson PT, Kouranov A, Fitziev P, Bouyioukos C, Mourelatos Z, Hatzigeorgiou A. 2004. *A combined computational-experimental approach predicts human microRNA targets*. Genes Dev. 18(10):1165-1178

König J, Zarnack K, Luscombe NM, Ule J. 2012. *Protein-RNA interactions: new genomic technologies and perspectives*. Nat Rev Genet. 13(2):77-83.

Kozomara A, Griffiths-Jones S. 2013. *miRBase: annotating high confidence microRNAs using deep sequencing data.* Nucleic Acids Res. 42 (Database issue):D68-73

Kruger J. and Rehmsmeier M. 2006. *RNAhybrid: microRNA target prediction easy, fast and flexible.* Nucleic Acids Res. 34 (Web Server issue):W451-454.

Lee R.C., Feinbaum R.L., and Ambros V. 1993. *The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14*. Cell 75, 843–854.

Mazieres J, He B, You L, Xu Z, Lee AY, Mikami I, Reguart N, Rosell R, McCormick F, Jablons DM. 2004. *Wnt inhibitory factor-1 is silenced by promoter*

hypermethylation in human lung cancer. Cancer Res. 64(14):4717-4720.

Miranda KC, Huynh T, Tay Y, Ang YS, Tam WL, Thomson AM, Lim B, Rigoutsos I. 2006. *A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes.* Cell. 126(6):1203-1217.

Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, Horvitz HR, Ruvkun G. 2000. *The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans.* Nature 403:901
-906.

Thomson DW, Bracken CP, Goodall GJ. 2011. *Experimental strategies for microRNA target identification.* Nucleic Acids Res. 39(16):6845-6853.

Vergoulis T, Vlachos IS, Alexiou P, Georgakilas G, Maragkakis M, Reczko M, Gerangelos S, Koziris N, Dalamagas T, Hatzigeorgiou AG. 2012. *TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support.* Nucleic Acids Res. 40(Database issue):D222-229.

Wahlestedt C. 2013. *Targeting long non-coding RNA to therapeutically upregulate gene expression.* Nat Rev Drug Discov. 12(6):433-446.

Watanabe Y., Tomita M., Kanai A. 2007. *Computational methods for microRNA target prediction.* Meth Enzymol. 2765-2786.

Wu X, Somlo G, Yu Y, Palomares MR, Li AX, Zhou W, Chow A, Yen Y, Rossi JJ, Gao H, Wang J, Yuan YC, Frankel P, Li S, Ashing-Giwa KT, Sun G, Wang Y, Smith R, Robinson K, Ren X, Wang SE. 2012. *De novo sequencing of circulating miRNAs identifies novel markers predicting clinical outcome of locally advanced breast cancer.* J Transl Med.10:42.

Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T. 2009. *miRecords: an integrated resource for microRNA–target interactions.* Nucleic Acids Res.37 (Database issue):D105-110.

APPENDIX

Perl implementation code

```perl
#!/usr/bin/perl -w

# parse_Tarbase.pl
# Tarbase parsing to search for miRNA-target interactions (MTI)

    use strict;
    use Switch;

    use Spreadsheet::ParseExcel;

    my $parser   = Spreadsheet::ParseExcel->new();
    my $workbook = $parser->parse('Tarbase.xls');

    if ( !defined $workbook ) {
        die $parser->error(), ".\n";
    }

  #count MTI interactions for Reporter Gene Assay
    my $counter1=0;

  #count MTI interactions for Techniques in Gene Level και Protein
Level
    my $counter2=0;

  #count MTI interactions for High Throughput Techniques
    my $counter3=0;

  #count MTI interactions for Other experiments
    my $counter4=0;



    for my $worksheet ( $workbook->worksheets() ) {

        my ( $row_min, $row_max ) = $worksheet->row_range();
        my ( $col_min, $col_max ) = $worksheet->col_range();
```

45

```perl
my $cell = $worksheet->get_cell( $row, $col );
            next unless $cell;
        switch ($cell->value()){
          case "Reporter Gene Assay" {

                  $counter1=$counter1+1;
                  for my $col ( $col_min .. $col_max ) {
                          if ($col == 1 or $col == 3 or $col == 2 ){
                              my $cell = $worksheet->get_cell( $row,
$col );

              print "     ", $cell->value() ;
              print "   ";


                          }
                    }
                 print "\n";
                 next;
          }
          case ["Western Blot","PCR"] {

                  $counter2=$counter2+1;
                  for my $col ( $col_min .. $col_max ) {
                          if ($col == 1 or $col == 3 or $col == 2 ){
                              my $cell = $worksheet->get_cell( $row,
$col );

              print "     ", $cell->value() ;
              print "   ";


                          }
                    }
                 print "\n";
           next;
          }
          case ["Microarrays","Proteomics","Sequencing"]{

                  $counter3=$counter3+1;
                  for my $col ( $col_min .. $col_max ) {
                          if ($col == 1 or $col == 3 or $col == 2 ){
                          my $cell = $worksheet->get_cell( $row,
$col );
```

46

```perl
                print "     ", $cell->value() ;
                print "    ";
                          }
                   }
              print "\n";
           next;
         }
         case "Other" {

                  $counter4=$counter4+1;
                  for my $col ( $col_min .. $col_max ) {
                          if ($col == 1 or $col == 3 or $col == 2 ){
                              my $cell = $worksheet->get_cell( $row,
$col );

                print "      ", $cell->value() ;
                print "    ";
                          }
                   }
              print "\n";
           next;
         }
           }
         }
       }
    }

    print "counter1 is for Reporter Gene Assay: ",$counter1 ,"\n";
    print "counter2 is for Techniques in Gene Level and Protein Level:
",$counter2 ,"\n";
    print "counter3 is for High Throughput Techniques:
",$counter3 ,"\n";
    print "counter4 is for Other: ",$counter4 ,"\n";
```

47

```perl
#!/usr/bin/perl -w

# parse_mirtarbase.pl
# miRTarBase parsing to search for miRNA-target interactions (MTI)

    use strict;
    use Switch;

    use Spreadsheet::ParseExcel;

    my $parser   = Spreadsheet::ParseExcel->new();
    my $workbook = $parser->parse('miRTarBase.xls');

    if ( !defined $workbook ) {
        die $parser->error(), ".\n";
    }

  #count MTI interactions for Reporter Gene Assay
    my $counter1=0;

  #count MTI interactions for Techniques in Gene Level και Protein
Level
    my $counter2=0;

   #count MTI interactions for High Throughput Techniques
    my $counter3=0;

   #count MTI interactions for Other experiments
    my $counter4=0;


    for my $worksheet ( $workbook->worksheets() ) {

        my ( $row_min, $row_max ) = $worksheet->row_range();
        my ( $col_min, $col_max ) = $worksheet->col_range();

        for my $row ( $row_min .. $row_max ) {
            for my $col ( 1 .. 6 ) {

                my $cell = $worksheet->get_cell( $row, $col );
                next unless $cell;
```

```perl
                    if ($cell->value() =~ /Luciferase reporter assay/ or
$cell->value() =~ /Reporter assay/ or $cell->value() =~ /GFP reporter
assay/ or $cell->value() =~ /phenotypic sensor assay/ or
 $cell->value() =~ /GUS reporter assay/ or $cell->value() =~ /LacZ
reporter assay/ or $cell->value() =~ /B-globin reporter assay/ or
 $cell->value() =~ /EGFP reporter assay/){

                        $counter1=$counter1+1;
                        for my $col ( $col_min .. $col_max ) {
                                if ($col == 3 or $col == 1 or $col == 6 ){
                                    my $cell = $worksheet->get_cell( $row,
$col );

            print "    ", $cell->value() ;
            print "  ";

                                }
                            }
                        print "\n";
                    }

            if ($cell->value() =~ /Western blot/ or $cell->value() =~
/Northern blot/ or $cell->value() =~ /qRT-PCR/ or $cell->value() =~
/Flow/ or $cell->value() =~ /flow/ or
 $cell->value() =~ /Immunocytochemistry/ or $cell->value() =~ /ELISA/
or $cell->value() =~ /intrarenal expression/ or $cell->value() =~
/Immunohistochemistry/ or
 $cell->value() =~ /Immunofluorescence/ or $cell->value() =~
/Immunoblot/ or $cell->value() =~ /In situ hybridization/ or
$cell->value() =~ /FACS/ or $cell->value() =~ /RTPCR/ or
 $cell->value() =~ /RNase mapping analysis/ or $cell->value() =~
/EMSA/ or $cell->value() =~ /ASO assay/ or $cell->value() =~ /semi-
qRT-PCR/ or $cell->value() =~ /quantitative PCR/ or
 $cell->value() =~ /Ribonuclease protection assay/ or $cell->value()
=~ /mRAP/ or $cell->value() =~ /Real time PCR/){

                        $counter2=$counter2+1;

                        for my $col ( $col_min .. $col_max ) {
                                if ($col == 3 or $col == 1 or $col == 6 ){
```

```perl
            my $cell = $worksheet->get_cell( $row, $col );

        print "    ", $cell->value() ;
        print "   ";

                    }
                }
              print "\n";
        }

        if ($cell->value() =~ /Microarray/ or $cell->value() =~
/CLASH/ or $cell->value() =~ /Proteomics/ or $cell->value() =~
/Sequencing/ or
 $cell->value() =~ /Immunoprecipitaion/ or $cell->value() =~
/Quantitative proteomic approach/ or $cell->value() =~ /ChIP-seq/ or
$cell->value() =~ /ChIP/ or $cell->value() =~ /Degradome/ or
 $cell->value() =~ /CLIP-seq/ or $cell->value() =~ /pSILAC/ or
$cell->value() =~ /Mass spectrometry/ or $cell->value() =~ /Branched
DNA probe assay/ or $cell->value() =~ /2DGE/){
                    $counter3=$counter3+1;
                    for my $col ( $col_min .. $col_max ) {
                            if ($col == 3 or $col == 1 or $col == 6 ){
                                my $cell = $worksheet->get_cell( $row,
$col );

        print "    ", $cell->value() ;
        print "   ";

                    }
                }
              print "\n";
        }
        if ($cell->value() =~ /Other/ ){

                $counter4 = $counter4+1;
                for my $col ( $col_min .. $col_max ) {
                        if ($col == 3 or $col == 1 or $col == 6 ){
                            my $cell = $worksheet->get_cell( $row,
$col );
```

50

```perl
print "      ", $cell->value() ;
          print "   ";
                    }
               }
            print "\n";
        }
        }
      }
    }

    print "counter1 is for Reporter Gene Assay: ",$counter1 ,"\n";
    print "counter2 is for Techniques  Gene Level/Protein Level:
",$counter2 ,"\n";
    print "counter3 is for High Throughput Techniques:
",$counter3 ,"\n";
    print "counter4 is for Other experiments: ",$counter4 ,"\n";
```

```perl
 #!/usr/bin/perl -w


# parse_mirecords.pl
# miRecords parsing to search for miRNA-target interactions (MTI)
```

1

```perl
            my $cell = $worksheet->get_cell( $row, $col );

        print "     ", $cell->value() ;
        print "   ";
                    }
                }
            print "\n";
         }
         if ($cell->value() =~ /GL-PL/ ){
                $counter2=$counter2+1;
                for my $col ( $col_min .. $col_max ) {
                        if ($col == 0 or $col == 1 or $col == 2 ){
                            my $cell = $worksheet->get_cell( $row,
$col );


        print "      ", $cell->value() ;
        print "   ";
                    }

                }
            print "\n";
         }
         if ($cell->value() =~ /HTT/ ){
                $counter3=$counter3+1;
                for my $col ( $col_min .. $col_max ) {
                        if ($col == 0 or $col == 1 or $col == 2 ){
                            my $cell = $worksheet->get_cell( $row,
$col );

        print "      ", $cell->value() ;
        print "   ";


                    }

                }
            print "\n";

        }
         if ($cell->value() =~ /Other/ ){

                $counter4 = $counter4+1;
```

52

```perl
for my $col ( $col_min .. $col_max ) {

            if ($col == 0 or $col == 1 or $col == 2 ){
                my $cell = $worksheet->get_cell( $row, $col );

    print "      ", $cell->value() ;
    print "   ";
                }
            }
        print "\n";
        }
        }
    }
}

print "counter1 is for Reporter Gene Assay: ",$counter1 ,"\n";
print "counter2 is for Techniques  Gene Level και Protein Level:
",$counter2 ,"\n";
print "counter3 is for High Throughput Techniques:
",$counter3 ,"\n";
print "counter4 is for Other experiments: ",$counter4 ,"\n";
```

53

```perl
#!/usr/bin/perl -w

# tarbase_pub.pl
# Count number of publications of TarBase

    use strict;
    use Switch;

    use Spreadsheet::ParseExcel;

    my $parser   = Spreadsheet::ParseExcel->new();
    my $workbook = $parser->parse('Tarbase.xls');

    if ( !defined $workbook ) {
        die $parser->error(), ".\n";
    }
  #counter for Reporter Gene Assay
    my $counter1=0;

  #counter for Techniques Gene Level και Protein Level
    my $counter2=0;

  #counter for High Throuput Techniques
    my $counter3=0;

  #counter for Other
    my $counter4=0;


    my @pubarray1;
    my @pubarray2;
    my @pubarray3;
    my @pubarray4;
    my @pubarray5;



    for my $worksheet ( $workbook->worksheets() ) {

        my ( $row_min, $row_max ) = $worksheet->row_range();
        my ( $col_min, $col_max ) = $worksheet->col_range();
```

54

```perl
my $cell = $worksheet->get_cell( $row, $col );
                next unless $cell;

            my $pubvalue5 = $worksheet->get_cell( $row, 4);
            push(@pubarray5, $pubvalue5->value()) unless grep{$_ eq
$pubvalue5->value()} @pubarray5;


            switch ($cell->value()){
            case "Reporter Gene Assay" {
                    my $pubvalue1 = $worksheet->get_cell( $row, 4);
                    push(@pubarray1, $pubvalue1->value()) unless
grep{$_ eq $pubvalue1->value()} @pubarray1;
                    $counter1=$counter1+1;
                    for my $col ( $col_min .. $col_max ) {
                            if ($col == 1 or $col == 2 or $col == 3){
                                my $cell = $worksheet->get_cell( $row,
$col );

            print "      ", $cell->value() ;
            print "   ";


                                }

                        }
                    print "\n";
                 next;
                }

            case ["Western Blot","PCR"] {
                    my $pubvalue2 = $worksheet->get_cell( $row, 4);
                    push(@pubarray2, $pubvalue2->value()) unless
grep{$_ eq $pubvalue2->value()} @pubarray2;
                    $counter2=$counter2+1;
                    for my $col ( $col_min .. $col_max ) {
                            if ($col == 1 or $col == 2 or $col == 3 ){
                                my $cell = $worksheet->get_cell( $row,
$col );

            print "      ", $cell->value() ;
```

55

```perl
			print "	";

						}

					}
				print "\n";
		next;
		  }

			case ["Microarrays","Proteomics","Sequencing"]{
				my $pubvalue3 = $worksheet->get_cell( $row, 4);
				push(@pubarray3, $pubvalue3->value()) unless
grep{$_ eq $pubvalue3->value()} @pubarray3;

				$counter3=$counter3+1;
				for my $col ( $col_min .. $col_max ) {
					if ($col == 1 or $col == 2 or $col == 3 ){
						my $cell = $worksheet->get_cell( $row,
$col );

		print "	", $cell->value() ;
		print "	";

						}

					}
				print "\n";
		next;
		  }

			case "Other" {
				my $pubvalue4 = $worksheet->get_cell( $row, 4);
				push(@pubarray4, $pubvalue4->value()) unless
grep{$_ eq $pubvalue4->value()} @pubarray4;
				$counter4=$counter4+1;
				for my $col ( $col_min .. $col_max ) {
					if ($col == 1 or $col == 2 or $col == 3){
						my $cell = $worksheet->get_cell( $row,
$col );
```

56

```perl
    print "       ", $cell->value() ;
           print "    ";


                          }


                   }
                  print "\n";
          next;
            }
          }
          }
       }
    }

    print "counter1 is for Reporter Gene Assay: ",$counter1 ,"\n";
    print "total pubarray1: ",scalar @pubarray1, "\n";
    print "counter2 is for Techniques  Gene Level and Protein Level:
",$counter2 ,"\n";
    print "total pubarray2: ",scalar @pubarray2, "\n";
    print "counter3 is for High Throughput Techniques:
",$counter3 ,"\n";
    print "total pubarray3: ",scalar @pubarray3, "\n";
    print "counter4 is for Other: ",$counter4 ,"\n";
    print "total pubarray4: ",scalar @pubarray4, "\n";

    print "total pubmed_ids: ",scalar @pubarray5, "\n";
```

```perl
#!/usr/bin/perl -w

 # mirtarbase_pub.pl
 # Count number of publications of miRTarBase

    use strict;
    use Switch;

    use Spreadsheet::ParseExcel;

    my $parser   = Spreadsheet::ParseExcel->new();
    my $workbook = $parser->parse('miRTarBase.xls');

    if ( !defined $workbook ) {
        die $parser->error(), ".\n";
    }
  #count for Reporter Gene Assay
    my $counter1=0;

  #count for Techniques Gene Level και Protein Level
    my $counter2=0;

   #count for High Throughput Techniques
    my $counter3=0;

   #count for other
    my $counter4=0;


    my @pubarray1;
    my @pubarray2;
    my @pubarray3;
    my @pubarray4;
    my @pubarray5;



    for my $worksheet ( $workbook->worksheets() ) {

        my ( $row_min, $row_max ) = $worksheet->row_range();
        my ( $col_min, $col_max ) = $worksheet->col_range();
```

58

```perl
                    my $cell = $worksheet->get_cell( $row, $col );
                    next unless $cell;

                    my $pubvalue5 = $worksheet->get_cell( $row, 8);

                    push(@pubarray5, $pubvalue5->value()) unless grep{$_
eq $pubvalue5->value()} @pubarray5;



                if ($cell->value() =~ /Luciferase reporter assay/ or
$cell->value() =~ /Reporter assay/ or $cell->value() =~ /GFP reporter
assay/ or $cell->value() =~ /phenotypic sensor assay/ or
 $cell->value() =~ /GUS reporter assay/ or $cell->value() =~ /LacZ
reporter assay/ or $cell->value() =~ /B-globin reporter assay/ or
 $cell->value() =~ /EGFP reporter assay/){

                        my $pubvalue1 = $worksheet->get_cell( $row, 8);

                        push(@pubarray1, $pubvalue1->value()) unless
grep{$_ eq $pubvalue1->value()} @pubarray1;

                        $counter1=$counter1+1;

                        for my $col ( $col_min .. $col_max ) {

                                if ($col == 1 or $col == 3 or $col == 6 ){
                                    my $cell = $worksheet->get_cell( $row,
$col );

            print "     ", $cell->value() ;
            print "   ";

                                }

                    }
                  print "\n";

            }
```

```perl
            if ($cell->value() =~ /Western blot/ or $cell->value() =~
/Northern blot/ or $cell->value() =~ /qRT-PCR/ or $cell->value() =~
/Flow/ or $cell->value() =~ /flow/ or
 $cell->value() =~ /Immunocytochemistry/ or $cell->value() =~ /ELISA/
or $cell->value() =~ /intrarenal expression/ or $cell->value() =~
/Immunohistochemistry/ or
 $cell->value() =~ /Immunofluorescence/ or $cell->value() =~
/Immunoblot/ or $cell->value() =~ /In situ hybridization/ or
$cell->value() =~ /FACS/ or $cell->value() =~ /RTPCR/ or
 $cell->value() =~ /RNase mapping analysis/ or $cell->value() =~
/EMSA/ or $cell->value() =~ /ASO assay/ or $cell->value() =~ /semi-
qRT-PCR/ or $cell->value() =~ /quantitative PCR/ or
 $cell->value() =~ /Ribonuclease protection assay/ or $cell->value()
=~ /mRAP/ or $cell->value() =~ /Real time PCR/){


                    my $pubvalue2 = $worksheet->get_cell( $row, 8);

                    push(@pubarray2, $pubvalue2->value()) unless
grep{$_ eq $pubvalue2->value()} @pubarray2;

                    $counter2=$counter2+1;

                    for my $col ( $col_min .. $col_max ) {

                            if ($col == 1 or $col == 3 or $col == 6){
                                my $cell = $worksheet->get_cell( $row,
$col );


        print "     ", $cell->value() ;
        print "   ";

                            }
                        }
                    print "\n";

            }
```

```perl
            if ($cell->value() =~ /Microarray/ or $cell->value() =~
/CLASH/ or $cell->value() =~ /Proteomics/ or $cell->value() =~
/Sequencing/ or
 $cell->value() =~ /Immunoprecipitaion/ or $cell->value() =~
/Quantitative proteomic approach/ or $cell->value() =~ /ChIP-seq/ or
$cell->value() =~ /ChIP/ or $cell->value() =~ /Degradome/ or
 $cell->value() =~ /CLIP-seq/ or $cell->value() =~ /pSILAC/ or
$cell->value() =~ /Mass spectrometry/ or $cell->value() =~ /Branched
DNA probe assay/ or $cell->value() =~ /2DGE/){

                    my $pubvalue3 = $worksheet->get_cell( $row, 8);

                    push(@pubarray3, $pubvalue3->value()) unless
grep{$_ eq $pubvalue3->value()} @pubarray3;

                    $counter3=$counter3+1;

                    for my $col ( $col_min .. $col_max ) {

                            if ($col == 1 or $col == 3 or $col == 6){
                                my $cell = $worksheet->get_cell( $row,
$col );


        print "      ", $cell->value() ;
        print "   ";
                            }
                        }
                      print "\n";
                }

            if ($cell->value() =~ /Other/ ){

                    my $pubvalue4 = $worksheet->get_cell( $row, 8);

                    push(@pubarray4, $pubvalue4->value()) unless
grep{$_ eq $pubvalue4->value()} @pubarray4;

                    $counter4 = $counter4+1;

                    for my $col ( $col_min .. $col_max ) {
                            if ($col == 1 or $col == 3 or $col == 6 ){
```

61

```perl
  my $cell = $worksheet->get_cell( $row, $col );

          print "     ", $cell->value() ;
          print "   ";
                    }

              }
            print "\n";
         }
        }
      }
    }

    print "counter1 is for Reporter Gene Assay: ",$counter1 ,"\n";
    print "total pubarray1: ",scalar @pubarray1, "\n";
    print "counter2 is for Techniques  Gene Level and Protein Level:
",$counter2 ,"\n";
    print "total pubarray2: ",scalar @pubarray2, "\n";
    print "counter3 is for High Throughput Techniques:
",$counter3 ,"\n";
    print "total pubarray3: ",scalar @pubarray3, "\n";
    print "counter4 is for Other experiments: ",$counter4 ,"\n";
    print "total pubarray4: ",scalar @pubarray4, "\n";

    print "total pubmed_ids: ",scalar @pubarray5, "\n";
```

62