ABSTRACT

miRCancer: a microRNA-Cancer Association Database and Toolkit Based on Text Mining

by Boya Xie

November, 2010

Director: Dr. Qin Ding

DEPARTMENT OF COMPUTER SCIENCE

MicroRNAs (miRNAs) are a class of single-stranded small non-coding RNAs. Study of miRNAs has been a fast growing field since the first miRNA was discovered in 1993. Many findings indicate that miRNA plays an important role in regulating crucial processes in gene development. In this research, we designed and implemented a framework called miRCancer, which is the first comprehensive database for miRNA expression profiles in human cancers based on experimental results. miRCancer also includes an integrated microRNA sequence analysis toolkit to help researchers discover possible relationships and functionalities of miRNAs. Text mining was used to automatically extract miRNA expression information from online electronic literatures. A new coding system, EICD-O, was introduced to annotate miRNA expressions in cancers.

miRCancer: a microRNA-Cancer Association Database and Toolkit Based on Text Mining

A Thesis

Presented To

The Faculty of the Department of Computer Science

East Carolina University

In Partial Fulfillment

of the Requirements for the Degree

Masters of Science

by

Boya Xie

November, 2010

miRCancer: a microRNA-Cancer Association Database and Toolkit Based on Text Mining

miRCancer: a microRNA-Cancer Association Database and Toolkit Based on Text Mining

by
Boya Xie

APPROVED BY:

DIRECTOR OF THESIS: _____
Qin Ding, PhD

COMMITTEE MEMBER: _____
John Placer, PhD

COMMITTEE MEMBER: _____
Nasseh Tabrizi, PhD

COMMITTEE MEMBER: _____
Qin Ding, PhD

CHAIR OF THE DEPARTMENT OF COMPUTER SCIENCE: _____
Karl Abrahamson, PhD

DEAN OF THE GRADUATE SCHOOL: _____
Paul J. Gemperline, PhD

ACKNOWLEDGEMENTS

Firstly, I would like to express my deepest gratitude to my supervisor, Dr. Qin Ding, for her invaluable advice and constant guidance throughout my entire master degree study.

I also would like to extend my thanks to my friends Di Wu and Joyce for their delightful contribution in discussions and ideas exchanged. This research would not happen without their help.

Great thanks also go to my parents who are always being there and support me.

At last, I would like to thank all those who have, in one way or another, contributed to the success of this thesis.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1 INTRODUCTION

MicroRNAs (miRNAs) are a class of single-stranded small non-coding RNAs, many findings suggested that miRNA plays an important role in regulating crucial processes in gene development. In this research project, we designed and constructed the first comprehensive database for microRNA (miRNA) expression profiles in human cancers based on experimental results. We also developed a software system to make the database searchable. In addition, the system includes an integrated miRNA sequence analysis tool for miRNAs related studies as well. The entire framework is named as miRCancer, and the two components are named as miRCDB and miRSAT, standing for miRNA-Cancer DataBase, and miRNA Sequence Analysis Tool, respectively.

1.1. Motivation and Objectives

The main motivation of this research comes from the increasing interests in miRNAs and the demand for an instrument which can document and offer easy access to experimental results for miRNAs.

The study in miRNA is a new but rapidly increasing research area. Since the first miRNA was discovered in 1993, there were more than 6,000 publications about miRNAs before 2010 and about 3,000 published in 2010. The reason that miRNA has attracted so many interests is the important role that miRNAs are playing in the gene expression process. Many studies show that miRNA is involved in regulating crucial processes. One of the most exciting involvements of miRNA is its association with disease. Since the heart disease and cancer are the top two killers for human beings, many investigations are being carried out on miRNA-based diagnostics and

therapeutics for these two diseases. Although most of the experimental results from those studies were clearly documented and are available to public, they are scattered in literatures.

The large number of miRNAs and the many-to-many relationships between miRNA and cancers make it difficult to access miRNA-cancer research results. On one hand, the number of newly discovered miRNAs has been exponentially increased in recent years, from a few numbers of them before 2005 to over 17,000 nowadays. On the other hand, different miRNAs have different functionalities and may target different genes. Most miRNAs are shown to function in multiple cancers [1]. Some studies even pointed out that one miRNA may have different functions in the same type of cancer during its different development stages.

With the above motivations, this project has the following objectives:

- Develop an automatic text extraction module to get miRNA-cancer experimental results from literatures.
- Design and construct a database which stores miRNA expression profiles in human cancers and detailed related information.
- Develop an integrated miRNA sequence analysis tool including clustering and Chi-square analysis.
- Design and develop a graphic user interface to facilitate database search and provide sequence analysis tools to analyze selected sequences.

The online free database PubMed [2] was used as the literature source. Key words: “miRNA cancer” and “microRNA cancer” were used to search against the database. The open source database MySQL [3] was used for the server for database construction. The interface was

implemented in Java using NetBeans [4]. In order to communicate with the MySQL database, a JDBC driver and the MySQL Connector/J were used [5].

1.2. Structure of the Thesis

This thesis consists of six chapters. Chapter 1 briefly introduces the context of the research, as well as motivations and objectives of this study. Chapter 2 summarizes the background and studies of miRNA and the related research about miRNA-cancers relationships. Similar databases and miRNA function annotations are reviewed. Chapter 3 presents the system design which includes graphic user interface, logic layer, storage and the data retrieval module. Chapter 4 describes the database design in details. The automatic miRNA-cancer relationship data retrieval process is introduced. Chapter 5 presents two sequence analysis tools: clustering and chi-square analysis. Chapter 6 documents a step-by-step illustration to setup the system. It also provides instructions on how to use the system for both system administrator and normal users. Chapter 7 summarizes the achievements of this research, and suggests possible future work in this area.

Chapter 2 LITERATURE REVIEW

2.1. miRNA and Cancers

miRNAs are a class of non-coding, about 22-nucleotide, single-stranded RNAs that derive from a stem–loop precursor. They regulate gene expression in various ways, for example, by binding to a part of one or more messenger RNAs (mRNA), resulting in inhibiting mRNA translation or mRNA degradation [6-7]. Many findings suggest that miRNA plays an important role in regulating crucial processes, such as cell proliferation [8], apoptosis [9-10], development [11], differentiation [12-13], and metabolism [14-15]. Nevertheless, functions of most miRNAs remain undiscovered.

The first known miRNA, lin-4, was introduced in 1993. It was found in the nematode worm Caenorhabditis elegans and observed to down-regulate lin-14 protein translation [6]. Since then, several techniques have been developed to identify miRNAs. As of September 2010, over 17,000 mature miRNA sequences have been identified with a growth of more than 6,000 in the past year. Increasing interest in miRNA functionality motivated several studies; as a result, many strategies have been invented to deduce miRNA functions. The most common approach is via computational prediction algorithms, such as TargetScan, PicTar, miRanda, and miRecords [16-19]. Commonly-used experimental methods to reveal miRNA functions include over-expression and silencing [1].

Among the inferred functions, disease association has attracted not only scientific attention but also business interest; there is growing funding for commercializing miRNA-based diagnostics and therapeutics. Cancer causes about 13% of human deaths worldwide and 26% in developed countries [20]. Cancer is expected to pass heart disease as the number one killer.

Cancer increases the economic burden dramatically not only on patients but also on society. Carcinogenesis includes “unwanted” gene mutations which induce transformation of normal cells, e.g., by over-activation of pro-oncogenes such as ras, src, or abl, and in-activation of tumor suppressors such as p53 and PTEN [21], etc. Tumorigenic gene mutations can also be induced by harmful environmental factors such as tobacco over-usage and UV exposure. Although research demonstrates that carcinogens including tobacco, radiation, and chemicals may increase the chance of developing cancer, the mechanism of cancer formation is yet to be identified. The fact that miRNA expression levels vary significantly between normal and cancer cells suggests that miRNA might be associated with cancer development. Indeed, miRNA fingerprints are recognized in all types of analyzed cancers, such as breast cancer, cervical cancer, hepatocellular carcinoma, lung cancer, prostate cancer, and lymphoblastic leukemia [22-27].

Despite the fact that abundant research investigating miRNA expressions in cancer cells has been carried out and some experimental results are available, they remain scattered in the literature. Hence this project was carried out to develop the miRCDB database to include information about miRNA regulation in various human cancers. Additionally, the miRSAT tool was developed to provide various types of computational sequence analyses which may assist knowledge discovery in miRNA sequences.

2.2. miRNA Databases

Since the term “microRNA” was formally introduced in 2001, numerous databases and tools have been created to store miRNA information and predict their target mRNAs. Although computational target prediction methods are fast for identifying potential miRNA targets, experimental validation of miRNA functionalities is desired. The significant increase in validation experiments raises the need for a database to store these results in some uniform way.

However, compared to databases providing computationally-predicted miRNA functions, databases storing experimental miRNA functions are rare. To the best of our knowledge, only a few databases recording experimental miRNA functions exist.

2.2.1. TarBase

TarBase, developed by the DIANA lab, is the first database which stores experimentally-detected miRNA targets [28]. All the information in TarBase was manually curated and limited to only five kinds of species: human, mouse, fruit fly, worm, and zebrafish. Though TarBase provided quite detailed descriptions of each target site, such as the gene and the location within the 3' UTR where the miRNA binding occurs, it could not serve as a comprehensive data source for miRNA targeting studies due to the limited number of species and miRNAs. By October 2010, TarBase had collected about 170 miRNAs which is far less than the number of discovered miRNAs (over 17,000) or even the number of miRNAs discovered in human (1,038).

2.2.2. miRBase

The miRBase is a central online miRNA repository that is currently hosted and maintained at the University of Manchester [29]. miRBase provides three kinds of services: a searchable database of published miRNA sequences and annotation, naming service for unique novel miRNA genes, and miRNA target prediction. miRBase contains the most complete set of miRNAs than any other resources, therefore it is used as the miRNA sequence source in this research project.

Currently miRBase only provides computationally predicted miRNA targets, but it is expected to include experimental miRNA function annotation soon [29].

2.2.3. miR2Disease

miR2Disease records experimental miRNA expression profiles with human diseases [30]. Each entry in the database was manually curated which includes: miRNA id, disease name, a brief description of the relationship, detection method, miRNA target genes, and literature reference. As of November 2010, miR2Disease has recorded 349 miRNAs in 134 diseases. In total, it has 2,920 different entries. Although the number of miRNAs involved is far less than the number of human miRNAs documented, considering the limited number of experiments that has been done on miRNA expressions in diseases, it is likely that miR2Disease has provided the complete set of miRNA-disease relationships being studied so far.

Because all the information in miR2Disease was manually documented, errors were likely to be introduced into the database during the data collection process. In addition, manual collection is a time and effort consuming process.

2.3. miRNA Function Annotation

The miRNA functions are annotated differently in various databases. TarBase records miRNA functions as miRNA and target site pairs. From the target sites and target genes, miRNA functions are inferred. For miR2Disease, each miRNA is linked to one or more diseases, and the disease terminologies are organized according to disease ontology. Disease ontology identification numbers (DOID) are used in miR2Disease to identify human diseases.

There is no uniform way to indicate miRNA functions, but there are standard ways to annotate miRNAs and cancers [31-32]. The international classification of diseases for oncology (ICD-O) [33] published by the World Health Organization serves as the tumor or cancer registries. The ICD-O coding was extended in this research to annotate miRNA functions in cancers.

2.4. Text Mining in Biomedical Literatures

Text mining is the process to derive useful information from text using computational approaches or tools. The increasing number of literatures and the open access policies of many biomedical journals make text mining more useful for both hypothesis generation and biological discovery. Unlike text mining in other fields, the biomedical text mining requires not only computer specialists but also biologists. The biomedical text has very specialized and complex vocabularies. Sometimes, a good interpretation of the text literately is not enough to understand it and the biology background information may be needed in order to get the accurate meaning.

Text mining has been broadly used in biomedical literatures, especially in gene and protein related studies due to the need to process a large number of publications [34-36]. Nonetheless, text mining with miRNA literatures is rarely seen, but is expected to become popular and useful for the following reasons: 1) miRNAs have similar nature as genes; many studies on genes could be applied to miRNAs as well. 2) miRNA studies have dramatically increased in recent years leading to significant increase in the number of papers published on miRNAs. 3) Though many researches on miRNA functions have been carried out, most miRNAs functions remains unclear. All the above reasons lead to one conclusion that tools dealing with large number of literatures are in demand.

There are three text mining approaches that are commonly used in the biomedical realm: co-occurrence based, rule-based, and machine learning [37]. Co-occurrence is widely used as a simple preprocessor or filter. One example of using co-occurrence in mining miRNA functionalities is the miRSel [38]. Co-occurrence based systems are easy to build but have to deal with variants in how a concept is stated in human writings. For example, stomach cancer could be referred as gastric cancer, cancer in stomach, or stomach carcinoma. Rule-based system

keeps sets of rules, such as the structure of language, relevant biology facts, and collection of variant forms. The rule sets may be built either by hard-coding or linguistic/semantic analyzing. In either way, rule-based system usually takes significant amount of time to develop. Finally, the machine learning approach requires large number of training data which is normally expensive or even impossible to generate.

Chapter 3 SYSTEM DESIGN

The miRCancer system adopts a 3-tier architecture style which organizes subsystems into three layers: Graphical User Interface (GUI), Logic, and storage as shown in Figure 1. Details about each layer are given below.

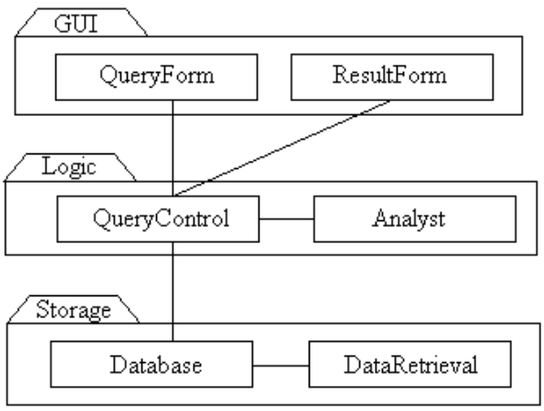


Figure 1 miRCancer System Architecture

3.1. User Interface Layer

The user interface layer contains two interfaces: the query interface and the result interface. These two interfaces allow the user to put criteria and search against the database, perform analysis on the sequences, and interactively view the search/analysis results.

The GUI layer captures user's selection and passes the information to the logic layer, which will further interact with the storage and compute suitable results. After the results are found, they are returned to GUI to display.

3.1.1. Query Interface

Query interface is divided into two sections: one for searching miRNA-cancer relationship, and the other for setting up analysis criteria and performing analyses. The first

section allows users to enter either a miRNA name or a cancer name, and search for miRNA profiles in cancers related to the input.

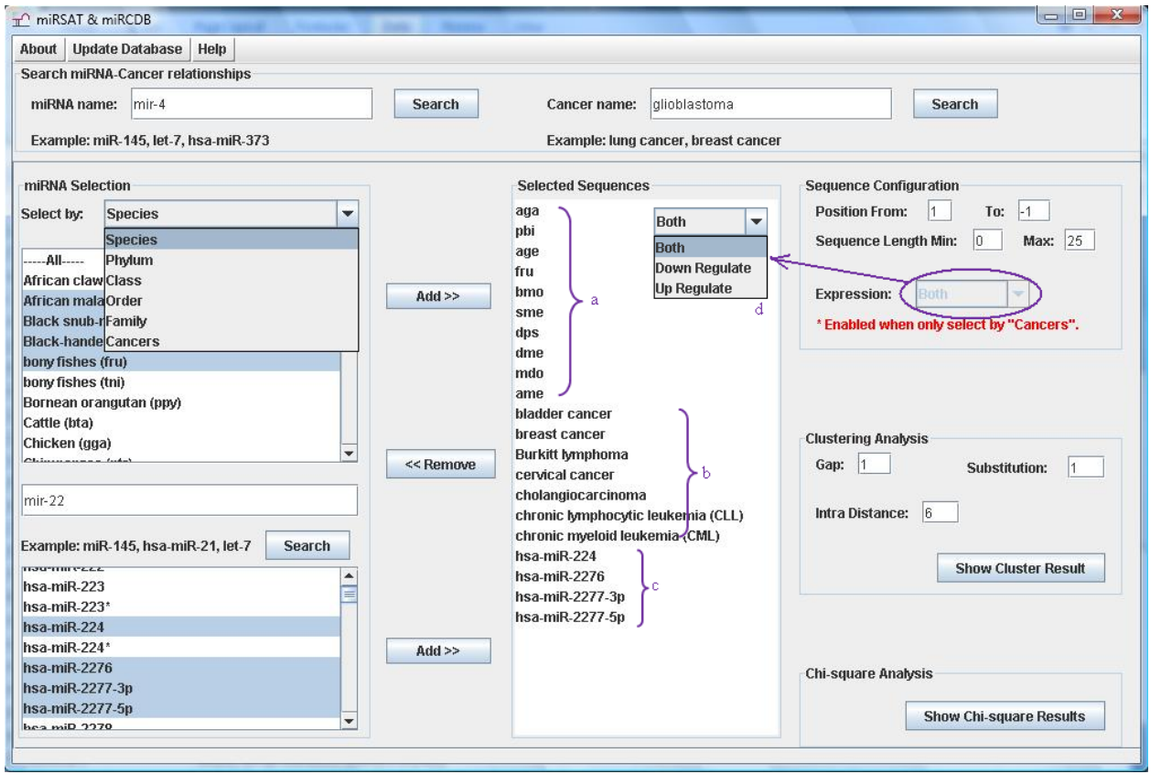


Figure 2 Query Interface.

(a). Selected by Species. (b). Selected by Associated Cancer Name.
(c). Selected by miRNA Name. (d). Expression Pattern Options

The second section, which is much more complicated, is formed with two parts: sequence selection and analysis tool selection. In sequence selection, miRNAs can be selected in multiple ways: by species, phylum, class, order, family, or related cancer, or by miRNA name. Selection could also be made by combining these approaches as shown in Figure 2. Besides selecting those entire sequences, a user is also allowed to specify sequence length and/or sections of sequences he or she is interested in. For example, one can choose sequences from the mouse that have 20 to 23 nucleotides each, and run analysis only on nucleotides located from position 3 to 18. This

query is submitted to the system by setting “position from=3, to=18, sequence length min=20, max=23”, and choosing mouse in the species list. In addition, when miRNA sequences are chosen by associated cancer name, the option “expression pattern” is enabled, so that the user could state what pattern to be examined, i.e., up regulated, down regulated, or both. In the analysis tool selection, a user can choose to search for miRNA-cancer relationship, perform clustering analysis, or perform chi-square analysis. By "clustering" we refer to grouping miRNAs by sequence similarity. Three parameters may be set for clustering analysis: gap score (GAP), substitution score (SUB), and intra cluster distance threshold (INTRA). The details of the parameters will be explained in Chapter 5.1.

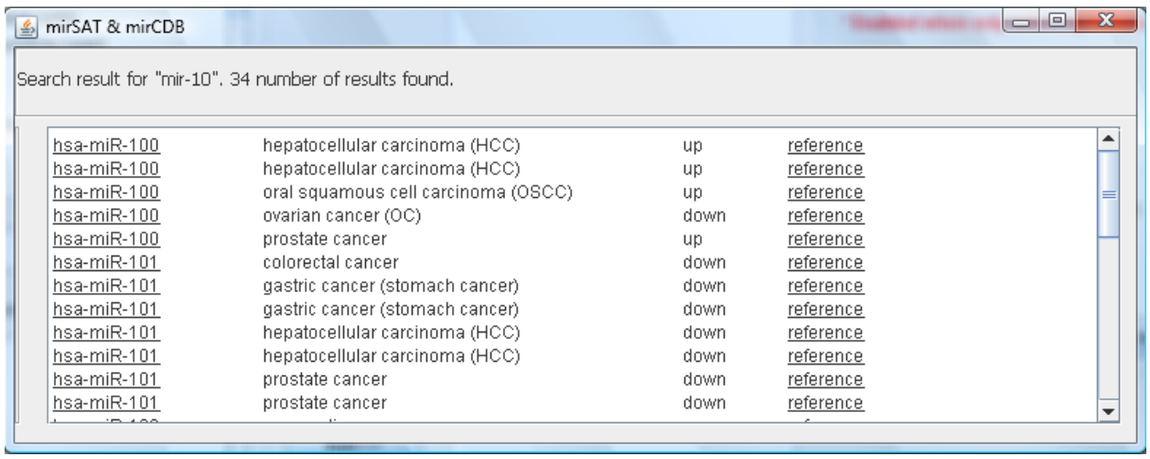


Figure 3 Result Interface for miRNA-Cancer Relationship Search.

3.1.2. Result Interface

Result Interface is extended into two pages: showing miRNA expressions with cancers, and displaying analysis output in an interactive way.

If the user enters keyword either in the miRNA name field or the cancer name field, the search result will be displayed in a similar way as shown in Figure 3. A summary showing the search keyword and number of results is displayed at the top of the window. Below that, each

relationship is presented with the miRNA name, cancer name, expression profile, and a link to the PubMed website for the relevant paper. Besides the linkage to PubMed, each miRNA name also can be clicked and will link to the miRBase web page where the details of the miRNA sequence are available.

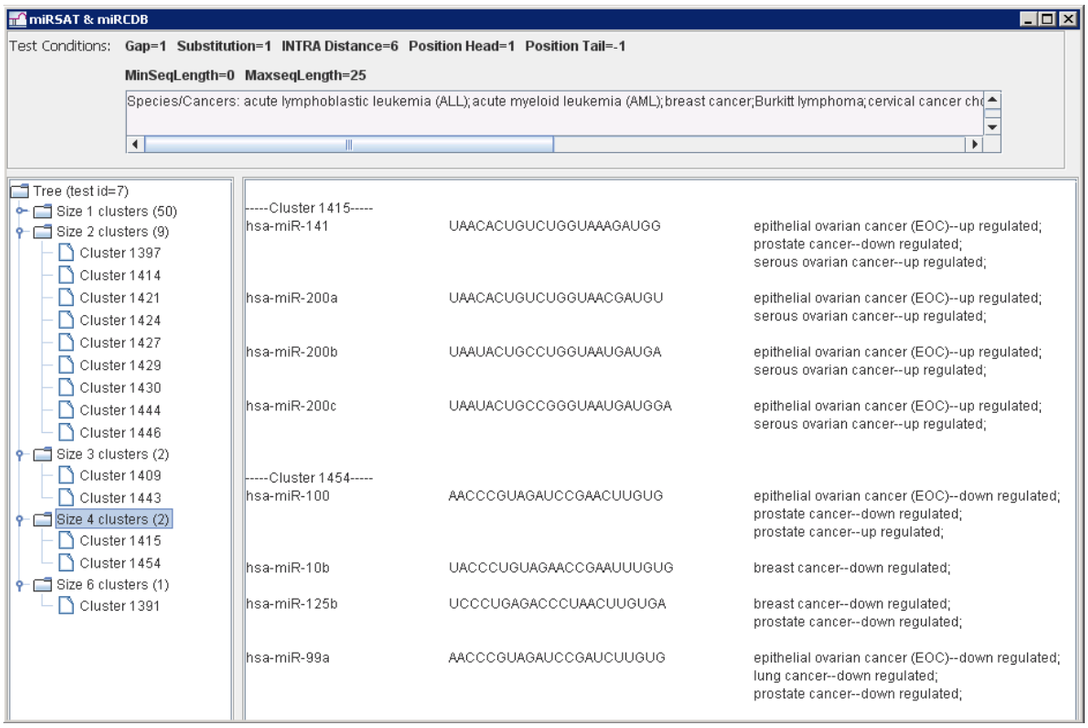


Figure 4 Cluster Analysis Result Interface.

There are two result interfaces corresponding to the two analysis tools we developed. Cluster results are presented in a summary tree and a detailed panel. By clicking on one or multiple tree nodes, corresponding detailed sequence information will be presented in the detail panel on the right, as shown in Figure 4. The detailed information includes: the miRNA sequence id, sequence, and related cancer names with the types of expressions. Chi-square analysis results are presented in a symmetric grid where each square represents a pair of nucleotide positions in the miRNAs. By moving the mouse over the grid, the three labels on the top of the grid will

change their values according to the square which is pointed to by the mouse. For instance, in Figure 5, the mouse is currently over the cell at the 14th column and the 12th row with chi-square score 8.64. Moreover, the grid is shaded with different intensities so that darker squares have smaller chi-square values.

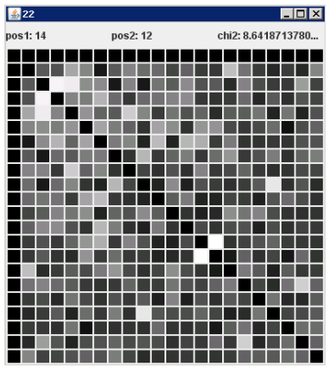


Figure 5 Chi-square Analysis Result Interface

3.2. Logic Layer

The logic layer includes all computational analysis utilities and controls information flow between the user interface and the storage layer. Generally it acquires queries from user interface, and then gathers information from the storage layer. After desired manipulation, it returns results back to the user through interface. The core for this layer is the implementation of analysis tools, which is discussed in Chapter 5. All the sequence analysis algorithms are structured following a strategy design pattern which decouples policy-deciding class from a set of algorithms, so that different algorithms can be changed transparently from the user.

3.3. Storage Layer

The database contains the most up-to-date reported miRNA sequences by periodic (once a month) update from miRBase website [29], which is the central online miRNA repository. The miRBase database could be downloaded in the Fasta format which is used by miRCDB to update miRNA sequences. The miRSAT & miRCDB currently stores 17,341 matured miRNA

sequences where 2,357 are minor miRNA sequences. Each sequence has information including id, accession number, sequence length, indicator for minor, and the belonging species. Having a database that is updated periodically will ensure that it remains comprehensive and as useful as possible for finding patterns within and across species.

In addition to the sequence data, all corresponding species are stored in the database as well. The 17,341 miRNAs are found in 142 species, and each of them is recorded into the database with its family, order, class, and phylum. Every miRNA sequence links with one species through a three to four letter species abbreviation code. For instance, any miRNA from a human being is linked through the code “hsa”, which is the abbreviation of homo sapiens. Besides miRNA sequences and species, as the name miRCDB (miRNA-Cancer DataBase) suggests, the database stores important information about miRNA expression profile in cancers. The miRNA and cancer relationships are automatically extracted from 484 published papers among a literature pool of size over 10,000. Together with miRSAT, the system explores information enclosed in the cancer-associated miRNA sequences. For example, characteristics of miRNAs associated with a certain cancer could be inferred by running a clustering test on the specific cancer.

Furthermore, the storage layer also includes a data retrieval module which automatically extracts miRNA expression profiles in cancers from literatures based on text mining technique. This module does not have any interface for users and neither is supposed to be used by users. The data retrieval is designed for the system administrator to update the database periodically with information from the PubMed. It is recommended to update the miRNA-cancer relationships as often as updating the miRNA sequences. Although there were more sequences identified in the past year than the number of researches on miRNA profiling in cancers, it is

important to keep the miRNA expression information comprehensive for miRCDB to serve its functionality as it is designed. More discussions about miRCDB design and the data retrieval module are presented in the next chapter.

Chapter 4 MIRCDB AND DATA RETRIEVAL

As mentioned previously, the miRCDB is the main component of this system which documents miRNA expression patterns in different cancers. It collects only human cancers in the current version with the possibility of extending to other species in the future. The data retrieval module is the major contribution component which made this study novel. This chapter will present the structure of the database and the data retrieval process.

4.1. miRNA-Cancer DataBase

The entity-relationship model of the database is presented in the figure below.

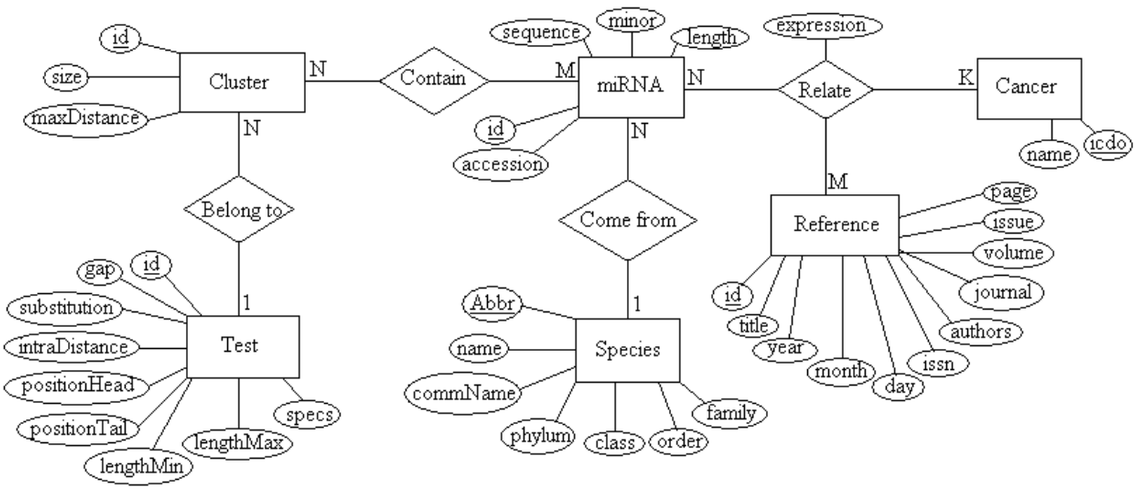


Figure 6 ER Diagram for miRCDB

The miRCDB contains eight tables, in which five are essential and the other three are auxiliaries for sequence analysis. Our previous paper has discussion about the three auxiliary tables [39]. The corresponding relational schema is:

- miRNA (<u>id</u>, accession, sequence, minor, length, speciesAbbr*)
- Cancer (<u>icdo</u>, name)

- Reference (id, title, year, month, day, issn, authors, journal, volume, issue, page)
- mirCancer (mirId*, refId*, icdo*, expression)
- Species (Abbr, name, commName, phylum, class, order, family)
- Cluster (id, size, maxDistance, testId*)
- Test (id, gap, substitution, intraDistance, positionHead, positionTail, lengthMin, lengthMax, specs)
- mirCluster (mirId*, clustered*)

Primary keys are marked with underline while foreign keys are marked with "*".

As described in section 3.3, all miRNA entries are provided by miRBase. Each entry has the identification number, accession number, sequence, an integer showing whether this entry is a minor sequence, the length of the sequence, and a three or four letter name indicating which species this sequence was identified.

All cancer data are gathered from international classification of diseases for oncology (ICD-O) [33]. The ICD-O has been used in cancer registries for nearly 25 years. Each instance includes the ICD-O code and a common name. ICD-O coding is extended in this research for its two substantial advantages: being widely used in medicine and biology leads to easy reference to the medical and biological literature; carefully structured 3rd edition ICD-O codes describe different cancer stages as well as morphology of the neoplasm. Some studies [40-42] found that a miRNA may have different expressions during different cancer development stages. More experiments on the miRNA expression patterns associated with stages may open a new research area. Descriptions of cancer behavior and morphology by ICD-O codes give the system extensibility for capturing miRNA expression profiles for different stages, types, and subtypes of cancers in the future.

Each reference refers to a published paper in the PubMed. The miRCDB stores the following information for each paper: the PubMed ID (PMID), article title, date the paper was published, ISSN number, authors' list, the journal name where the article was published with, volume, issue number, and page in the journal. With the PMID, one paper could be accessed online by appending the id to the following web address: http://www.ncbi.nlm.nih.gov/pubmed/. For example: a paper with the id of 21060679 can be found at http://www.ncbi.nlm.nih.gov/pubmed/21060679.

miRNA expressions in various cancers are extracted from PubMed with the search keywords: miRNA cancer and microRNA expression. The next section describes the extracting process in details. The EICD-O coding, or expression ICD-O coding, was introduced to annotate miRNA expressions in cancers. By adding one extra number (0 for uncertain, 1 for down-regulate, and 2 for up-regulate) in front of the original ICD-O code, EICD-O code is able to present the miRNA expression profile with not only a specific cancer but also a cancer stage. For example, a miRNA that up-regulates malignant serous ovarian carcinoma is coded as "2 C56.9 M-9090/3", in which the first digit "2" corresponds to up-regulate, while "C56.9" is the topography code for ovary, "M-9090" is the morphology code for struma ovarii, and the number "3" is the behavior meaning a malignant cancer stage. Each miRNA-to-cancer relationship is recorded with the EICD-O code and the reporting literature reference. Due to the fact that multiple experiments might be performed for the same miRNA with the same type of cancer, each relationship is identified by the combination of miRNA id, EICD-O code, and the PMID.

4.2. Data Retrieval

In order to perform the automatic data retrieval, the program should have the knowledge about miRNA name, cancer name, and expression vocabularies. With these three dictionaries,

the data retrieval module takes a collection of literatures, processes each literature depending on the keyword appearance, and then documents miRNA profiling in cancers with the associated reference. The Data Retrieval module is structured following the strategy design pattern that encapsulates algorithm implementations from client. This design allows the retrieval method to be modified or improved later without affecting users.

4.2.1. Dictionary Construction and Keyword Searching

Compared to other biology sequences, such as genes and proteins, miRNA names have been formalized early enough so that they are quite unified and relatively simple. Most miRNAs are named with the "miR" prefix, except for a few miRNAs: bantam, let-7 family, lin-4, and lsy-6. The genes that encode the miRNA are also named using the same prefix "mir". To better differentiate with genes, matured miRNA sequences are usually named with "miR", and minor sequences have a '*' at the end (e.g. miR-101, cel-miR-230*). The numerical numbers are assigned sequentially. Because miRNAs are highly conserved between species and organisms, identical sequences have the same id number regardless of species (e.g. mmu-miR-23a in mouse, has-miR-23a in human), and letter and/or numeral suffixes are used to differentiate identical sequences within a species (e.g. miR-6-1 and miR-6-2 are identical). Three or four letter prefix is used to indicate species (e.g. cfa-miR for dogs and cel-miR for nematode).

There are some variations of this name convention. The data retrieval process handles variations when searching miRNA names. The variants include: miRNA-, microRNA-, miR, microRNA for miR-; miRNA-let- for miR-let-; and cases without species prefix (e.g. miR-195 for has-miR-195).

For cancer name vocabulary, there are some deviations too. For example, gastric cancer is the same as stomach cancer; cancer and carcinoma are used interchangeably.

A list of 16 terms is compiled to describe relations between miRNA-cancer pairs. Seven of them refer to miRNA inhibit cancer: downregulation, down-regulation, downregulate, down-regulate, inhibit, repress, and down-regulating. Nine of them refer to oncogenic functions: overexpress, upregulation, upregulate, up-regulation, up-regulate, promote, high express, highly express, and up-regulating.

4.2.2. Text Mining

Because the article title and abstract usually captures the results of researches, each literature's title and abstract are used as the input for data extraction. The program will find a miRNA-cancer relationship only when the tri-occurrences of a miRNA, a cancer name, and an expression term are found in a title or in the abstract. However, the tri-occurences does not guarantee that a relationship will be found.

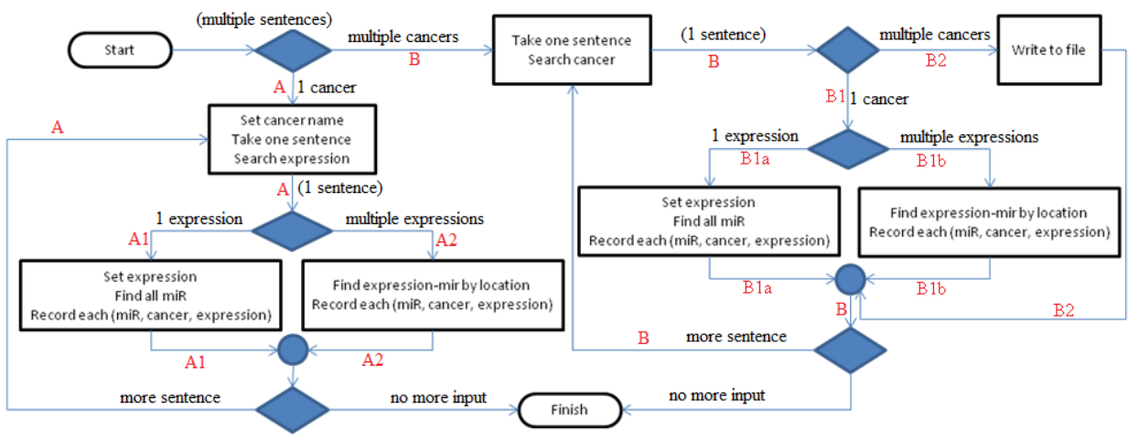


Figure 7 Data Retrieval Flow Chart

A simplified flow chart presented in Figure 7 illustrates the flow of text processing, with the complete flow chart attached in Appendix A. The retrieval process takes one abstract or one title at a time, and starts by looking for cancer names. In the case of a title, it is treated as a single

sentence. If there is only one kind of cancer found, this name is used as the master cancer name, and the abstract is further divided into single sentences to search for expressions and miRNAs (route A in Figure 7). Whenever there is a miRNA-expression pair found in a single sentence, it joins the master cancer name to form a tri-relation: miRNA, expression, and cancer, which are then recorded into miRCDB as one entry. Route A is then separated into two routes A1 and A2 to detect miRNA-expression pairs. A1 shows the case in which there is only one kind of expression, either up-regulate or down-regulate, within a single sentence. In this case, any miRNA present in the sentence is recorded with that expression and the master cancer. A2 shows the case in which both expressions, i.e., up-regulate and down-regulate, are found in that sentence. In this case, expressions and miRNAs are paired based on their locations following the rules below:

- Exp1, exp2, mir1, mir2 -> both mir are paired with exp2
- Exp1, mir1, exp2, mir2 -> mir1 paired with exp1, mir2 paired with exp2
- Mir1, exp1, mir2, exp2 -> mir1 paired with exp1, mir2 paired with exp2
- Mir1, mir2, exp1, exp2 -> both mir are paired with exp1
- Mir1, exp1, exp2, mir2 -> mir1 paired with exp1, mir2 paired with exp2

Exp1 and exp2 means different expressions. If two or more expression terms are present together and have the same meaning, either up-regulate or down-regulate, they are considered as one exp1. When multiple miRNA names are showing continuously and not interrupted by any expression terms, they are considered as one mir1/mir2 group.

Route B represents the case that multiple cancer names are mentioned in one abstract. The abstract is divided into single sentences, and cancer names are searched again within a sentence. Route B1 is similar to route A, both of which pair up expressions with miRNAs in each

individual sentence. The only difference is that route B1 takes the cancer name from a single sentence while route A has a master cancer name for the abstract. Route B2 is a special case that within one sentence, multiple cancer names exist. This case is not attempted by the retrieval module; instead it simply writes the input string to a file for human user to view. The results show that the case represented by route B2 is rare, where there are only 4 occurrences from more than 10,000 papers.

4.3. Results and Evaluation

The data retrieval module used PubMed query result in XML format as input. In order to get the most complete collection of literatures related to miRNA expressions in cancers, several keywords were used to search against the PubMed: miRNA cancer, microRNA expression, mir cancer and so on. In total, 10,219 papers were processed. The PubMed provides search result in XML format which include all literature details including the article abstract.

There were 827 valid miRNA expressions found from these papers. Among them, 484 papers contributed towards the valid expressions. 180 miRNAs and 33 cancers were involved in these profiling.

Compared with the miR2Disease which has 349 miRNAs in 134 diseases and totally 2,920 different entries, the result from this research looks reasonable. Due to limited time and resources, no formal result evaluation was carried out. However, 12 test cases were performed to ensure the correctness. Appendix B provides the complete list of test cases.

Chapter 5 MIRSAT

The miRSAT tookit implements various tools that allow the researchers to interact with the database and search for miRNA features that may be of interest. Once the user selects some subset of the sequences in the database, an analysis tool may be used to discover features that characterize the selected sequences as distinct from the entire database. If such a feature is discovered, then we might expect that feature to have predictive power when considering newly-discovered miRNA sequences.

5.1. Clustering Analysis

Clustering is the assignment of objects into groups so that objects from the same cluster are more similar to each other than objects from different clusters. In sequence analysis, clustering is used to group homologous sequences into gene families. This is a very important concept in bioinformatics and evolutionary biology in general. Once an analysis is done, the clusters are returned to the user who may then check for any biological relevance associated with the clusters. Recall that these clusters are within only those strings (i.e., sequences or subsequences) that the researcher has chosen. This enables the researchers to investigate questions such as “Among the vertebrates, how closely related are those miRNAs that up-regulate certain cancers?”

5.1.1. Distance Calculation

The Needleman-Wunsch algorithm[43] was implemented in this system for calculating distance between sequences. The algorithm performs a global alignment on two sequences, and is commonly used in bioinformatics to align protein or nucleotide sequences. It is a dynamic

programming based algorithm, and was the first application of dynamic programming to biological sequence comparison.

The scoring system could be as simple as some arbitrary numbers in this system, or be sophisticate using similarity matrix for aligned characters. In order to define a suitable similarity matrix, substitution penalty and gap penalty, expert knowledge is necessary. This system leaves the substitution (SUB) and gap (GAP) penalty as user input, and set the match score for all characters to be zero. However, it is possible to use the similarity matrix in the future, if a suitable similarity matrix is found.

The global alignment technique is said to be most useful when the sequences in the query set are similar and of roughly equal size. Since most miRNA sequences have 20 to 23 nucleotides, the choice of using Needleman-Wunsch is appropriate.

5.1.2. Clustering Algorithm

The simple single-link hierarchical clustering method is used in this project. Hierarchical clustering treats each data point as a singleton cluster, and then successively merges clusters until all points have been merged into a single remaining cluster. The distance of two closes sequences from each cluster is used as the distance of those two clusters. The threshold for joining two clusters is called INTRA in the project, and it is also left for the user to define.

5.2. Chi-Square Analysis

A glance at histograms of frequencies of the four RNA bases at each position in the miRNA sequences shows that “U” is significantly over-represented in positions 1 and 9, an observation that become more significant since they bound the “seed region” [44]. In this project, a tool was devised to promote discovery of other such phenomena beyond what a histogram could reveal. Here a chi-square test is used to infer correlation between the characters found in

two positions of the strings under consideration. For example, the user might have selected only those miRNAs related to mammals, and discover that among those strings, whenever there is an “A” in the 4th position, there is a statistically significantly high rate of occurrences of “T” in the 17th position. When this test is selected, the user is presented with a visual representation of the chi-square scores for each pair of positions. By hovering the mouse over any square of the grid the user can see the chi-square score for the corresponding pair of positions, a level of significance (p-value), and whether or not the expected frequencies in the table used to compute the chi-square statistic invalidate the test for that position pair. The statistics packages in the open-source Apache commons-math was used in this project. The package is available at http://commons.apache.org.

Chapter 6 SYSTEM SETUP AND USER MANUAL

The system includes three java projects developed in NetBeans: CreateTables, parseNCBIJ, and DiscoveryMicRNA. The CreateTables project creates necessary tables in the database and inserts cancer, literature information and miRNA expressions into the database. The parseNCBIJ project implements the data retrieval module. It takes XML file as input and generates three text files: one for miRNA expression with cancer and paper id, one for used paper information, and the third one records papers that need user's attention. The DiscoveryMicRNA project constructs the GUI layer as well as the logic layer. It is designed mainly for the users while the other two are for system administrator.

6.1. System Dependency and Setup

The miRCDB uses MySQL server, thus MySQL and the JDBC Driver for MySQL are needed. Since some of the inputs to the system are Excel files, the Jexcelapi library is also required. Below are the recommended steps to setup the system:

1. Download and install MySQL. MySQL Community Server version 5.1.52 is tested to work with the system.
2. Configure MySQL using its Instance Config Wizard, set password to 1, and local host to 3306. After successful configuration, create a database named miRCDB.
3. Download and install JDBC Driver follow the instruction from: http://www.developer.com/java/data/article.php/3417381 [Retrieved on Nov. 11, 2010]
4. Download and install JExcelApi following the instructions from: http://www.javanb.com/netbeans/1/19762.html [Retrieved on Nov. 10, 2010]
5. Compile and run project CreateTables, select option 1 to create tables.

At this point, the system has everything ready for data input. The next section illustrates the steps for administrator to prepare the data for miRCDB.

6.2. Setup miRCDB

There are four kinds of data that are essential for miRCDB: miRNA, cancer, literature, and relationships. System administrators are supposed to follow the steps below to prepare such data.

1. Go to miRBase website download page to download the file mature.fa.
2. Compile and run DiscoveryMicRNA, click “Update Database” -> “Load Species Table”, and choose the provided Species.xls file.
3. Continue from previous step, click “Update Database” -> “Load miRNA”, and choose mature.fa file downloaded in step 1.

   Now miRNA and species information are inserted to miRCDB.
4. Go to PubMed (http://www.ncbi.nlm.nih.gov/pubmed) to search for relevant keywords, such as miRNA expression. Send the result to a file and choose format as XML. There will be a file named pubmed_result.txt. Copy this file to the parseNCBIJ directory.
5. Compile and run the parseNCBIJ project. Three files will be generated as describe at the beginning of this chapter.
6. Copy the miRNA expressions and papers file generated from previous step to the given file microRNAcancer.xls, and put them into different sheets: mircan and reference, respectively.
7. Run CreateTables again, and select option 2 to insert literature references.
8. Run CreateTables again, and select option 4 to insert cancers.
9. Run CreateTables again, and select option 3 to insert miRNA-cancers relationships.

Chapter 7 FUTURE WORK AND CONCLUSION

The miRCancer toolkit utilizes the new EICD-O coding to indicate miRNA expression in human cancers, and form such database on experimental results. It provides computational tools for miRNA sequence analysis which will assist researchers in finding miRNA sequence information. By using the automatic data extraction module, the system was able to retrieve all miRNA expressions from published literatures. As a result, the miRCDB claims itself as the first comprehensive database of miRNA expression profiles in human cancers based on experiments results.

Current experimental researches focus mainly on human cancers; as a result, this research is limited to human cancers with the possibility of extending to other species in the future. Moreover, there are only a few numbers of studies that were carried out on miRNA functionalities in different cancer development stages. Therefore, this database stores miRNA profiles in cancer stages I to III without differentiation. However, the coding system introduced in this research has the possibility to capture miRNA expressions in different cancer stages. The miRNA expression in other animal cancers may also be included in the database in the future. The EICD-O coding will help to define which cancer stage the expression is at while miRNA sequence name itself tells the species information.

Besides frequently updating miRNA sequences and expression information from miRBase and PubMed respectively, the system also provides user friendly interface to search the database and apply computational tools. The system is properly designed so that correctness is guaranteed and extendibility is offered. The strategy design pattern used in data retrieval module and analysis tool subsystem allows the toolkit to be improved or even substituted independently to users. Although sequence analysis tools are not the focus of this project, the current

framework allows more analysis tools to be added. There are numerous sequence analysis tools and data mining tools available. All miRNA sequences stored in this database could be easily analyzed using these tools with or without a little preprocessing.

Last but not least, the system is currently implemented as a stand-alone Java application. It may be converted into a web-based version later to make it available to the public.

REFERENCES

1. Zhang, B., et al., *microRNAs as oncogenes and tumor suppressors.* Dev Biol, 2007. **302**(1): p. 1-12.

2. *PubMed*. [cited 2009 November 18]; Available from: http://www.ncbi.nlm.nih.gov/PubMed/.

3. *MySQL*. [cited 2010 Nov. 8]; Available from: http://dev.mysql.com/downloads/mysql/.

4. *NetBeans*. [cited 2010 November 8]; Available from: http://netbeans.org/index.html.

5. *Connector/J*. [cited 2010 November]; Available from: http://dev.mysql.com/downloads/connector/j/.

6. Lee, R.C., R.L. Feinbaum, and V. Ambros, *The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14.* Cell, 1993. **75**(5): p. 843-54.

7. Bartel, D.P., *MicroRNAs: genomics, biogenesis, mechanism, and function.* Cell, 2004. **116**(2): p. 281-97.

8. Hwang, H.W. and J.T. Mendell, *MicroRNAs in cell proliferation, cell death, and tumorigenesis.* Br J Cancer, 2006. **94**(6): p. 776-80.

9. Xu, P., M. Guo, and B.A. Hay, *MicroRNAs and the regulation of cell death.* Trends Genet, 2004. **20**(12): p. 617-24.

10. Jovanovic, M. and M.O. Hengartner, *miRNAs and apoptosis: RNAs to die for.* Oncogene, 2006. **25**(46): p. 6176-87.

11. Karp, X. and V. Ambros, *Developmental biology. Encountering microRNAs in cell fate signaling.* Science, 2005. **310**(5752): p. 1288-9.

12. Chen, C.Z., et al., *MicroRNAs modulate hematopoietic lineage differentiation.* Science, 2004. **303**(5654): p. 83-6.

13. Shivdasani, R.A., *MicroRNAs: regulators of gene expression and cell differentiation.* Blood, 2006. **108**(12): p. 3646-53.

14. Wienholds, E. and R.H. Plasterk, *MicroRNA function in animal development.* FEBS Lett, 2005. **579**(26): p. 5911-22.

15. Poy, M.N., et al., *A pancreatic islet-specific microRNA regulates insulin secretion.* Nature, 2004. **432**(7014): p. 226-30.

16. Lewis, B.P., et al., *Prediction of mammalian microRNA targets.* Cell, 2003. **115**(7): p. 787-98.

17. Krek, A., et al., *Combinatorial microRNA target predictions.* Nat Genet, 2005. **37**(5): p. 495-500.

18. John, B., et al., *Human MicroRNA targets.* PLoS Biol, 2004. **2**(11): p. e363.

19. Xiao, F., et al., *miRecords: an integrated resource for microRNA-target interactions.* Nucleic Acids Res, 2009. **37**(Database issue): p. D105-10.

20. Garcia, M., et al. (2007) *Global Cancer Facts & Figures 2007.*

21. Calin, G.A. and C.M. Croce, *MicroRNA signatures in human cancers.* Nat Rev Cancer, 2006. **6**(11): p. 857-66.

22. Iorio, M.V., et al., *MicroRNA gene expression deregulation in human breast cancer.* Cancer Res, 2005. **65**(16): p. 7065-70.

23. Lui, W.O., et al., *Patterns of known and novel small RNAs in human cervical cancer.* Cancer Res, 2007. **67**(13): p. 6031-43.

24. Gramantieri, L., et al., *Cyclin G1 is a target of miR-122a, a microRNA frequently down-regulated in human hepatocellular carcinoma.* Cancer Res, 2007. **67**(13): p. 6092-9.

25. Takamizawa, J., et al., *Reduced expression of the let-7 microRNAs in human lung cancers in association with shortened postoperative survival.* Cancer Res, 2004. **64**(11): p. 3753-6.

26. Porkka, K.P., et al., *MicroRNA expression profiling in prostate cancer.* Cancer Res, 2007. **67**(13): p. 6130-5.

27. Mi, S., et al., *MicroRNA expression signatures accurately discriminate acute lymphoblastic leukemia from acute myeloid leukemia.* Proc Natl Acad Sci U S A, 2007. **104**(50): p. 19971-6.

28. Sethupathy, P., B. Corda, and A.G. Hatzigeorgiou, *TarBase: A comprehensive database of experimentally supported animal microRNA targets.* RNA, 2006. **12**(2): p. 192-7.

29. Griffiths-Jones, S., et al., *miRBase: tools for microRNA genomics.* Nucleic Acids Res, 2008. **36**(Database issue): p. D154-8.

30. Jiang, Q., et al., *miR2Disease: a manually curated database for microRNA deregulation in human disease.* Nucleic Acids Res, 2009. **37**(Database issue): p. D98-104.

31. Ambros, V., et al., *A uniform system for microRNA annotation.* RNA, 2003. **9**(3): p. 277-9.

32. Griffiths-Jones, S., *The microRNA Registry.* Nucleic Acids Res, 2004. **32**(Database issue): p. D109-11.

33. Fritz, A., et al., *International Classification of Diseases for Oncology*. 3rd ed. 2000.

34. Hanisch, D., et al., *ProMiner: rule-based protein and gene entity recognition.* BMC Bioinformatics, 2005. **6 Suppl 1**: p. S14.

35. Fundel, K. and R. Zimmer, *Gene and protein nomenclature in public databases.* BMC Bioinformatics, 2006. **7**: p. 372.

36. Fundel, K., et al., *A simple approach for protein name identification: prospects and limits.* BMC Bioinformatics, 2005. **6 Suppl 1**: p. S15.

37. Cohen, K.B. and L. Hunter, *Getting started in text mining.* PLoS Comput Biol, 2008. **4**(1): p. e20.

38. Naeem, H., et al., *miRSel: automated extraction of associations between microRNAs and genes from the biomedical literature.* BMC Bioinformatics, 2010. **11**: p. 135.

39. B. Xie, R.H., Q. Ding, and D. Wu,, *miRSAT & miRCDB: An Integrated MicroRNA Sequesce Analysis Tool and a Cancer-Associated MicroRNA Database.*, in *BICoB*. 2010, ISCA: Honolulu, Hawaii.

40. Lu, J., et al., *MicroRNA expression profiles classify human cancers.* Nature, 2005. **435**(7043): p. 834-8.

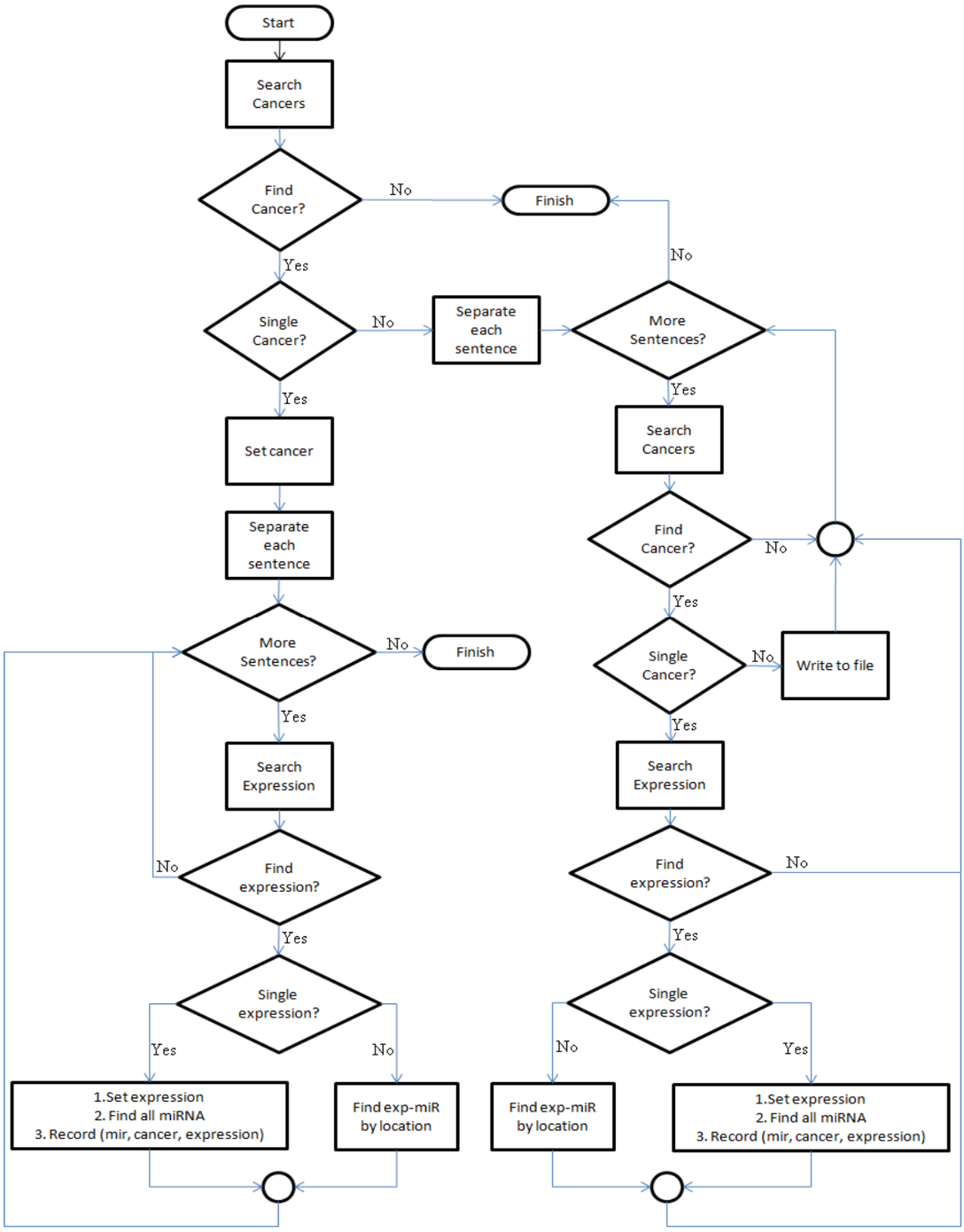
41. Yan, L.X., et al., *MicroRNA miR-21 overexpression in human breast cancer is associated with advanced clinical stage, lymph node metastasis and patient poor prognosis.* RNA, 2008. **14**(11): p. 2348-60.

42. Veerla, S., et al., *MiRNA expression in urothelial carcinomas: important roles of miR-10a, miR-222, miR-125b, miR-7 and miR-452 for tumor stage and metastasis, and frequent homozygous losses of miR-31.* Int J Cancer, 2009. **124**(9): p. 2236-42.

43. Needleman, S.B. and C.D. Wunsch, *A general method applicable to the search for similarities in the amino acid sequence of two proteins.* J Mol Biol, 1970. **48**(3): p. 443-53.

44. Lewis, B.P., C.B. Burge, and D.P. Bartel, *Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets.* Cell, 2005. **120**(1): p. 15-20.

APPENDIX A: Data Retrieval Flow Chart



APPENDIX B: Test Cases

Description	#1. Missing expression term
Input	MicroRNAs (miRNAs) in body fluids are candidate diagnostics for a variety of conditions and diseases, including breast cancer. One premise for using extracellular miRNAs to diagnose disease is the notion that the abundance of the miRNAs in body fluids reflects their abundance in the abnormal cells causing the disease. As a result, the search for such diagnostics in body fluids has focused on miRNAs that are abundant in the cells of origin. Here we report that released miRNAs do not necessarily reflect the abundance of miRNA in the cell of origin. We find that release of miRNAs from cells into blood, milk and ductal fluids is selective and that the selection of released miRNAs may correlate with malignancy. In particular, the bulk of miR-451 and miR-1246 produced by malignant mammary epithelial cells was released, but the majority of these miRNAs produced by non-malignant mammary epithelial cells was retained. Our findings suggest the existence of a cellular selection mechanism for miRNA release and indicate that the extracellular and cellular miRNA profiles differ. This selective release of miRNAs is an important consideration for the identification of circulating miRNAs as biomarkers of disease.
Output	None
Expected Output	None

Description	#2. Missing cancer name
Input	Aberrant overexpression of the miR-17-92 polycistron is strongly associated with B-cell lymphomagenesis. Recent studies have shown that miR-17-92 down-regulates the proapoptotic protein Bim, leading to overexpression of Bcl2, which likely plays a key role in lymphomagenesis. However, the fact that Jeko-1 cells derived from mantle cell lymphoma exhibit both homozygous deletion of BIM and overexpression of miR-17-92 suggests other targets are also involved in B-cell lymphomagenesis. To identify essential target(s) of miR-17-92 in lymphomagenesis, we first transfected miR-17-92 into 2 genetically distinct B-cell lymphoma cell lines: Raji, which overexpress c-Myc, and SUDHL4, which overexpress Bcl2. Raji transfected with miR-17-19b-1 exhibited down-regulated expression of Bim and a slight up-regulation in Bcl2 expression. On the other hand, SUDHL4 transfectants showed aggressive cell growth reflecting facilitated cell cycle progression at the G(1) to S transition and decreased expression of CDKN1A mRNA and p21 protein (CDKN1A/p21) that was independent of p53 expression. Conversely, transfection of antisense oligonucleotides against miR-17 and miR-20a into Jeko-1 led to up-regulation of CDKN1A/p21, resulting in decreased cell growth with G(1) to S arrest. Thus, CDKN1A/p21 appears to be an essential target of miR-17-92 during B-cell lymphomagenesis, which suggests the miR-17-92 polycistron has distinct targets in different B-cell

	lymphoma subtypes.
Output	None
Expected Output	None

Description	#3. Missing miRNA name
Input	Mammography is a powerful screening tool for early detection of breast cancer, but it has limitations in terms of both specificity and sensitivity. Imaging tools such as MRI that complement mammography are too costly to serve as first-line screens. Recently, progress has been made on blood markers, particularly microRNAs and proteins. There are new methods for protein marker discovery directly in blood, but they are limited in the number of patients that can be examined. An alternative is to discover markers as transcripts in tissues, followed by development of blood protein tests for those that perform best. To identify genes that are overexpressed in malignancy it is paramount to include normal control tissues from healthy individuals. Here we report the identification of potential breast cancer markers, including some that are overexpressed in aggressive disease.
Output	None
Expected Output	None

Description	#4. One miRNA, one cancer, one kind of expressions presents in an abstract.
Input	MicroRNAs (miRNAs), are ~22 nucleotides long, non-coding RNAs that control gene expression post-transcriptionally by binding to their target mRNA's 3'UTRs (untranslated regions). Due to their roles in various important regulatory processes and pathways, miRNAs have been implicated in disease mechanisms such as tumorigenesis when their expression is deregulated. To date, a significant number of miRNAs and their target messenger RNAs (mRNAs) have been identified and verified. It is generally accepted that miRNAs can potentially bind to many mRNAs, which brings the requirement of validation of these interactions. While understanding that such individual interactions is crucial to delineate the role of a specific miRNA, we took a holistic approach and analyzed global changes in the cell due to expression of a miRNA in a model cell line system. Our model consisted of MCF7 cells stably transfected with miR-125b (MCF7-125b) and empty vector (MCF7-EV). MiR-125b is one of the known down-regulated miRNAs in breast cancers. In this study we examined the global structural changes in MCF7 cells lacking and expressing miR-125b by Attenuated Total Reflectance Fourier Transform Infrared (ATR-FTIR) Spectroscopy and investigated the dynamic changes by more sensitive spin-labelling Electron Spin Resonance (ESR) spectroscopy. Our results revealed less RNA, protein, lipid, and glycogen content in MCF7-125b compared to MCF7-EV cells. Membrane fluidity and proliferation rate were shown to be lower in MCF7-125b cells. Based on these changes, MCF7-125b and MCF7-EV cells were discriminated successfully by cluster analysis. Here, we provide a novel

	means to understand the global effects of miRNAs in cells. Potential applications of this approach are not only limited to research purposes. Such a strategy is also promising to pioneer the development of future diagnostic tools for deregulated miRNA expression in patient samples.
Output	miR-125b, C50.9, down-regulate
Expected Output	miR-125b, C50.9, down-regulate

Description	#5. One miRNA, one cancer, 2 kinds of expressions in an abstract.
Input	Lung cancer is one of the leading causes of cancer-related death worldwide. Curcumin has been reported to have an antitumor effect by inducing apoptosis and suppressing growth of tumor cells. However, the mechanism by which curcumin exerts its anti-cancer effect needs further research. The purpose of the present study was to identify a miRNA-mediated mechanism which plays a role in the anti-cancer effects of curcumin. Alterations in miRNA expression were seen in curcumin-treated A549 cells, including significant downregulation of miRNA-186* expression by microarray analysis and real-time PCR. The miRNA-186* functions by overexpression or inhibition were investigated using biological assays in A549 cells. Additionally, caspase-10 was identified as a target of miRNA-186* using dual luciferase reporter assays and western blot analysis. These results demonstrate that curcumin induces A549 cell apoptosis through a miRNA pathway. Also, miRNA-186* could serve as a potential gene therapy target in curcumin treatment. furthermore, caspase-10 was shown to be a target of miR-186* regulation.<PMID>20878113
Output	miR-186*, C34.9, down-regulate
Expected Output	miR-186*, C34.9, down-regulate

Description	#6. One miRNA, multiple cancers, two kinds of expressions in an abstract.
Input	OBJECTIVE: MicroRNA (miRNA) plays an essential role in the progression of a variety of cancers, but its role in cervical cancer progression is not well defined. We aimed to test whether special miRNAs and their target mRNAs contribute to cervical cancer progression.METHODS: The expression profiles of 1145 microRNAs in cervical squamous cell carcinomas (CSCC) and adjacent non-tumor tissues were investigated using an Illumina microRNA microarray platform. Differentially expressed miRNAs were validated by RT-PCR. Downstream target validation was performed for miR-886-5p.RESULTS: We found that the expression levels of seven miRNAs differed significantly between CSCC tissues and adjacent non-tumor tissues. Forced expression of one miRNA, miR-886-5p, over-expressed in CSCC tissues lowered expression of the pro-apoptotic protein Bax, reduced apoptosis and promoted cell proliferation in H8, an HPV16-immortalized human cervical squamous epithelial cell line. Knockdown of miR-886-5p increased Bax protein and apoptotic cell death in cells of the cervical squamous carcinoma cell line, SiHa.CONCLUSION:

	MicroRNA miR-886-5p inhibits apoptosis of cervical cancer cells by down-regulating the production of Bax.
Output	miR-886-5p, C53.9, up-regulate
Expected Output	miR-886-5p, C53.9, up-regulate

Description	#7. Multiple miRNA, one cancer, two kinds of expressions in an abstract.
Input	This study aimed to investigate the microRNA (miRNA) profile in prostate carcinoma tissue by microarray analysis and RT-qPCR, to clarify associations of miRNA expression with clinicopathologic data and to evaluate the potential of miRNAs as diagnostic and prognostic markers. Matched tumor and adjacent normal tissues were obtained from 76 radical prostatectomy specimens. Twenty-four tissue pairs were analyzed using human miRNA microarrays for 470 human miRNAs. Differentially expressed miRNAs were validated by TaqMan RT-qPCR using all 76 tissue pairs. The diagnostic potential of miRNAs was calculated by receiver operating characteristics analyses. The prognostic value was assessed in terms of biochemical recurrence using Kaplan-Meier and Cox regression analyses. Fifteen differentially expressed miRNAs were identified with concordant fold-changes by microarray and RT-qPCR analyses. Ten microRNAs (hsa-miR-16, hsa-miR-31, hsa-miR-125b, hsa-miR-145, hsa-miR-149, hsa-miR-181b, hsa-miR-184, hsa-miR-205, hsa-miR-221, hsa-miR-222) were downregulated and 5 miRNAs (hsa-miR-96, hsa-miR-182, hsa-miR-182, hsa-miR-183, hsa-375) were upregulated. Expression of 5 miRNAs correlated with Gleason score or pathological tumor stage. Already 2 microRNAs classified up to 84% of malignant and nonmalignant samples correctly. Expression of hsa-miR-96 was associated with cancer recurrence after radical prostatectomy and that prognostic information was confirmed by an independent tumor sample set from 79 patients. That was shown with hsa-miR-96 and the Gleason score as final variables in the Cox models build in the 2 patient sets investigated. Thus, differential miRNAs in prostate cancer are useful diagnostic and prognostic indicators. This study provides a solid basis for further functional analyses of miRNAs in prostate cancer.
Output	hsa-miR-16, C61.9, down-regulate hsa-miR-31, C61.9, down-regulate hsa-miR-125b, C61.9, down-regulate hsa-miR-145, C61.9, down-regulate hsa-miR-149, C61.9, down-regulate hsa-miR-181b, C61.9, down-regulate hsa-miR-184, C61.9, down-regulate hsa-miR-205, C61.9, down-regulate hsa-miR-221, C61.9, down-regulate hsa-miR-222, C61.9, down-regulate hsa-miR-96, C61.9, up-regulate hsa-miR-182, C61.9, up-regulate

	hsa-miR-182, C61.9, up-regulate hsa-miR-183, C61.9, up-regulate hsa-375, C61.9, up-regulate
Expected Output	hsa-miR-16, C61.9, down-regulate hsa-miR-31, C61.9, down-regulate hsa-miR-125b, C61.9, down-regulate hsa-miR-145, C61.9, down-regulate hsa-miR-149, C61.9, down-regulate hsa-miR-181b, C61.9, down-regulate hsa-miR-184, C61.9, down-regulate hsa-miR-205, C61.9, down-regulate hsa-miR-221, C61.9, down-regulate hsa-miR-222, C61.9, down-regulate hsa-miR-96, C61.9, up-regulate hsa-miR-182, C61.9, up-regulate hsa-miR-182, C61.9, up-regulate hsa-miR-183, C61.9, up-regulate hsa-375, C61.9, up-regulate

Description	#8. Multiple miRNA, multiple cancers, one expression in an abstract.
Input	RESULTS: A set of 4 human micro-RNAs (miR-28, miR-185, miR-27, and let-7f-2) were found significantly up-regulated in renal cell carcinoma ($P < 0.05$) compared to normal kidney. Human micro-RNAs miR-223, miR-26b, miR-221, miR-103-1, miR-185, miR-23b, miR-203, miR-17-5p, miR-23a, and miR-205 were significantly up-regulated in bladder cancers ($P < 0.05$) compared to normal bladder mucosa. Of the kidney cancers studied, there was no differential micro-RNA expression across various stages, whereas with increasing tumor-nodes-metastasis staging in bladder cancer, miR-26b showed a moderate decreasing trend ($P = 0.082$).PMID: 17826655
Output	miR-103-1, C67.9, up-regulate miR-28, C64.9, M8312/3, up-regulate miR-27, C64.9, M8312/3, up-regulate miR-23a, C67.9, up-regulate miR-17-5p, C67.9, up-regulate miR-23b, C67.9, up-regulate miR-185, C64.9, M8312/3, up-regulate miR-26b, C67.9, up-regulate miR-205, C67.9, up-regulate miR-185, C67.9, up-regulate miR-221, C67.9, up-regulate miR-203, C67.9, up-regulate let-7f-2, C64.9, M8312/3, up-regulate miR-223, C67.9, up-regulate
Expected Output	miR-28, C64.9, M8312/3, up-regulate miR-185, C64.9, M8312/3, up-regulate

	miR-27, C64.9, M8312/3, up-regulate let-7f-2, C64.9, M8312/3, up-regulate miR-223, C67.9, up-regulate miR-26b, C67.9, up-regulate miR-221, C67.9, up-regulate miR-103-1, C67.9, up-regulate miR-185, C67.9, up-regulate miR-23b, C67.9, up-regulate miR-203, C67.9, up-regulate miR-17-5p, C67.9, up-regulate miR-23a, C67.9, up-regulate miR-205, C67.9, up-regulate

Description	#9. Tri-relation exists, but miRNA-expression pair not exists in any single sentences.
Input	In this study, we quantified 249 mature micro-RNA (miRNA) transcripts in estrogen receptor-positive (ER(+)) primary breast tumors of patients with lymph node-negative (LNN) disease to identify miRNAs associated with metastatic capability. In addition, the prognostic value of the candidate miRNAs was determined in ER(-)/LNN breast cancer. Unsupervised analysis in a prescreening set of 38 patients identified three subgroups predominantly driven by three miRNA signatures: an ER-driven luminal B-associated miRNA signature, a stromal miRNA signature, and an overexpressed miRNA cluster located on chromosome 19q23, but these intrinsic miRNA signatures were not associated with tumor aggressiveness. Supervised analysis in the initial subset and subsequent analysis in additional tumors significantly linked four miRNAs (miR-7, miR-128a, miR-210, and miR-516-3p) to ER(+)/LNN breast cancer aggressiveness (n = 147) and one miRNA (miR-210) to metastatic capability in ER(-)/LNN breast cancer (n = 114) and in the clinically important triple-negative subgroup (n = 69) (all $P < 0.05$). Bioinformatic analysis coupled miR-210 to hypoxia/VEGF signaling, miR-7 and miR-516-3p to cell cycle progression and chromosomal instability, and miR-128a to cytokine signaling. In conclusion, our work connects four miRNAs to breast cancer progression and to several distinct biological processes involved therein. PMID: 18755890
Output	None
Expected Output	None

Description	#10. Cancer name conventions. (non-small cell lung cancer vs. lung cancer)
Input	Following the identification of a set of hypoxia-regulated microRNAs (miRNAs), recent studies have highlighted the importance of miR-210 and of its transcriptional regulation by the transcription factor hypoxia-inducible factor-1 (HIF-1). We report here that miR-210 is overexpressed at late stages

	of non-small cell lung cancer. Expression of miR-210 in lung adenocarcinoma A549 cells caused an alteration of cell viability associated with induction of caspase-3/7 activity. miR-210 induced a loss of mitochondrial membrane potential and the apparition of an aberrant mitochondrial phenotype. The expression profiling of cells overexpressing miR-210 revealed a specific signature characterized by enrichment for transcripts related to 'cell death' and 'mitochondrial dysfunction', including several subunits of the electron transport chain (ETC) complexes I and II. The transcript coding for one of these ETC components, SDHD, subunit D of succinate dehydrogenase complex (SDH), was validated as a bona fide miR-210 target. Moreover, SDHD knockdown mimicked miR-210-mediated mitochondrial alterations. Finally, miR-210-dependent targeting of SDHD was able to activate HIF-1, in line with previous studies linking loss-of-function SDH mutations to HIF-1 activation. miR-210 can thus regulate mitochondrial function by targeting key ETC component genes with important consequences on cell metabolism, survival and modulation of HIF-1 activity. These observations help explain contradictory data regarding miR-210 expression and its putative function in solid tumors.Cell Death and Differentiation advance online publication, 1 October 2010; doi:10.1038/cdd.2010.119.pmid:20885442
Output	miR-210, C34.9, M8046/3, up-regulate
Expected Output	miR-210, C34.9, M8046/3, up-regulate

Description	#11. MiRNA name variations. MiRNA vs miR.
Input	PURPOSE: To investigate the different miRNA expression profiles of postoperative radiotherapy sensitive and resistant patients of non-small cell lung cancer, explore their potential role and find some radio-sensitivity markers. MATERIALS AND METHODS: Thirty non-small cell lung cancer patients who have been treated by postoperative radiotherapy were selected and were divided into radiotherapy sensitive group and resistant group according to overall survival and local or distant recurrence rate. Expression profile of miRNA in these two groups was detected by a microarray assay and the results were validated by quantitative RT-PCR and Northern blot. At the molecular level, the effect of one differently expressed miRNA (miR-126) on the growth and apoptosis of SK-MES-1 cells induced by irradiation was examined. RESULTS: Comparing with resistant patients, five miRNAs (miRNA-126, miRNA-let-7a, miRNA-495, miRNA-451 and miRNA-128b) were significantly upregulated and seven miRNAs (miRNA-130a, miRNA-106b, miRNA-19b, miRNA-22, miRNA-15b, miRNA-17-5p and miRNA-21) were greatly downregulated in radiotherapy sensitive group. Overexpression of miRNA-126 inhibited the growth of SK-MES-1 cells and promoted its apoptosis induced by irradiation. The expression level of p-Akt decreased in miRNA-126 overexpression group. After treating with phosphoinositidyl-3 kinase (PI3K) constitutively activator (IGF-1) and inhibitor (LY294002), miRNA-126 overexpression had no

	significant effects on the apoptosis of SK-MES-1 cells. CONCLUSION: We found 12 differently expressed miRNAs in the radiotherapy sensitive and resistant non-small cell lung cancer samples. Moreover, our results showed miRNA-126 promoted non-small cell lung cancer cells apoptosis induced by irradiation through the PI3K-Akt pathway.pmid:20728239
Output	miR-19b, C34.9, M8046/3, down-regulate miR-128b, C34.9, M8046/3, up-regulate miR-let-7a, C34.9, M8046/3, up-regulate miR-495, C34.9, M8046/3, up-regulate miR-126, C34.9, M8046/3, up-regulate miR-451, C34.9, M8046/3, up-regulate miR-106b, C34.9, M8046/3, down-regulate miR-17-5p, C34.9, M8046/3, down-regulate miR-130a, C34.9, M8046/3, down-regulate miR-21, C34.9, M8046/3, down-regulate miR-22, C34.9, M8046/3, down-regulate miR-15b, C34.9, M8046/3, down-regulate miR-126, C34.9, M8046/3, down-regulate
Expected Output	miR-126, C34.9, M8046/3, up-regulate miR-let-7a, C34.9, M8046/3, up-regulate miR-495, C34.9, M8046/3, up-regulate miR-451, C34.9, M8046/3, up-regulate miR-128b, C34.9, M8046/3, up-regulate miR-130a, C34.9, M8046/3, down-regulate miR-106b, C34.9, M8046/3, down-regulate miR-19b, C34.9, M8046/3, down-regulate miR-22, C34.9, M8046/3, down-regulate miR-15b, C34.9, M8046/3, down-regulate miR-17-5p, C34.9, M8046/3, down-regulate miR-21, C34.9, M8046/3, down-regulate

Description	#12. Multiple cancers in a single sentence.
Input	MicroRNAs (miRNAs) are short noncoding RNA molecules, which posttranscriptionally regulate genes expression and play crucial roles in diverse biological processes, such as development, differentiation, apoptosis and proliferation. Here, we investigated the possible role of miRNAs in the development of multidrug resistance (MDR) in human gastric and lung cancer cell lines. We found that miR-181b was downregulated in both multidrug-resistant human gastric cancer cell line SGC7901/vincristine (VCR) and multidrug-resistant human lung cancer cell line A549/cisplatin (CDDP), and the downregulation of miR-181b in SGC7901/VCR and A549/CDDP cells was concurrent with the upregulation of BCL2 protein, compared with the parental SGC7901 and A549 cell lines, respectively. In vitro drug sensitivity assay demonstrated that overexpression of miR-181b sensitized SGC7901/VCR and A549/CDDP cells to anticancer drugs,

	respectively. The luciferase activity of a BCL2 3'-untranslated region-based reporter construct in SGC7901/VCR and A549/CDDP cells suggests that a new target site in the 3'UTR of BCL2 of the mature miR-181s (miR-181a, miR-181b, miR-181c and miR-181d) was found. Enforced miR-181b expression reduced BCL2 protein level and sensitized SGC7901/VCR and A549/CDDP cells to VCR-induced and CDDP-induced apoptosis, respectively. Taken together, our findings suggest that miR-181b could play a role in the development of MDR in both gastric and lung cancer cell lines, at least in part, by modulation of apoptosis via targeting BCL2.
Output	Write to file
Expected Output	Write to file