1-1-2014

Human miRNA Sequence Based Variations Database

Nadim Georges Bou Zeidan

University of South Florida, nadimbz@icloud.com

Follow this and additional works at: http://scholarcommons.usf.edu/etd

Part of the Computer Engineering Commons

Scholar Commons Citation

Bou Zeidan, Nadim Georges, "Human miRNA Sequence Based Variations Database" (2014). Graduate Theses and Dissertations.
http://scholarcommons.usf.edu/etd/5350

This Thesis is brought to you for free and open access by the Graduate School at Scholar Commons. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Scholar Commons. For more information, please contact scholarcommons@usf.edu.
Human miRNA Sequence Based Variations Database

by

Nadim Georges Bou Zeidan

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Computer Engineering
Department of Computer Science and Engineering
College of Engineering
University of South Florida

Co-Major Professor: Yicheng Tu, Ph.D.
Co-Major Professor: Hong Huang, Ph.D.
Feng Cheng, Ph.D.

Date of Approval:
October 22, 2014

Keywords: CNV, Human Genome, Nucleotide Polymorphisms, Frequency

Copyright © 2014, Nadim Georges Bou Zeidan
DEDICATION

I dedicate my thesis work to my family with love.
ACKNOWLEDGMENTS

I wish to acknowledge my committee members for their guidance and support. Thank you Dr. Yicheng Tu and Dr. Hong Huang for working with me diligently on my thesis. Also, thank you Dr. Feng Cheng for agreeing to serve on the committee. I would also like to acknowledge my school department and everyone who has helped me in the process.
# TABLE OF CONTENTS

**LIST OF FIGURES** ........................................................................................................................... ii

**ABSTRACT** ....................................................................................................................................... iii

**CHAPTER 1: INTRODUCTION** ........................................................................................................... 1

**CHAPTER 2: HMIRNASQV DATABASE** ............................................................................................ 9
  2.1 Database Structure .......................................................................................................................... 9
  2.2 Tables Structure ............................................................................................................................ 11
    2.2.1 Human MicroRNA Excel ........................................................................................................ 12
    2.2.2 Human MicroRNA Chromosome .......................................................................................... 14
    2.2.3 Human MicroRNA .................................................................................................................. 15
    2.2.4 Human MicroRNA Variants Assignment .............................................................................. 17
    2.2.5 Human MicroRNA Family Information .............................................................................. 20
    2.2.6 Predicted Target ....................................................................................................................... 21

**CHAPTER 3: HMIRNASQV WEB APPLICATION** ............................................................................. 25
  3.1 Web Application Requirements .................................................................................................... 25
  3.2 HmiRNA Infrastructure ............................................................................................................... 26
  3.3 Web Application Infrastructure .................................................................................................... 27
  3.4 Web Application Functionalities .................................................................................................. 29
    3.4.1 Search ................................................................................................................................. 29
    3.4.2 Detail View .......................................................................................................................... 32

**REFERENCES** .................................................................................................................................. 36
LIST OF FIGURES

Figure 1:  HmiRNAsqv Web Application ................................................................. 7
Figure 2:  HmiRNAsqv Database Architecture ....................................................... 9
Figure 3:  hsa-mir-17 miRNA Sequence Structure ............................................... 13
Figure 4:  HmiRNAsqv miRNA Table ................................................................. 14
Figure 5:  HmiRNAsqv Human miRNA Table ......................................................... 16
Figure 6:  HmiRNAsqv Variants Table .................................................................. 17
Figure 7:  Copy Number Variants Example .......................................................... 18
Figure 8:  CNVs Associated with hsa-let-7a-1 ....................................................... 19
Figure 9:  MicroRNA Family Information Table .................................................... 20
Figure 10: MicroRNA Predicted Targets Table ..................................................... 22
Figure 11: MicroRNA Predicted Targets .............................................................. 23
Figure 12: HmiRNA Infrastructure ...................................................................... 26
Figure 13: HmiRNAsqv n-Tier Structure ............................................................. 28
Figure 14: HmiRNAsqv Search Functionality ....................................................... 30
Figure 15: miRNA-Var List View ...................................................................... 30
Figure 16: Human MicroRNA Export .................................................................. 31
Figure 17: hsa-mir-107 Detail View ..................................................................... 32
Figure 18: hsa-mir-101-2 Copy Number Variants ............................................... 33
Figure 19: hsa-mir-107 Predicted Targets ............................................................. 34
ABSTRACT

MicroRNAs (miRNAs) are studied as key genetic elements that regulate the gene expression involved in different human diseases. Clinical sequence based variations like copy number variations (CNVs) affect miRNA biogenesis, dosage and target recognition that may represent potentially functional variants and relevant target bindings.

To systematically analyze miRNA-related CNVs and their effects on related genes, a user-friendly free online database was developed to provide further analysis of co-localization of miRNA loci with human genome CNV regions. Further analysis pipelines such as miRNA-target to estimate the levels or locations of variations for genetic duplications, insertions or deletions were also offered. Such information could support the simulation of miRNA-target interactions.
CHAPTER 1: INTRODUCTION

MicroRNAs also known as miRNAs are considered to be very significant and highly important genetic elements which play the role in regulating thousands of human genes expressions[1]. MicroRNAs are a result of their precursors maturation which are considered a class of short non coding RNAs that are between 21 and 23 nucleotides long [2][4]. Ryazansky, Mikhaleva, and Olenkina characterize MicroRNAs as having a hairpin like secondary structure [2]. The hairpins’ structures are a result of the microRNAs formation which are processed from longer transcripts that can fold back on themselves to give it that shape [3].

Gene regulation is very important because it makes cells different from each other. The human body has many different cell types and all these cells have the same genes which all originated from the zygote. Human Micro RNAs regulate most of the human genes. MicroRNAs are not only important to what make the normal cell, but also they play a crucial role in cells that contain diseases such as the cancer cell [3].

Researchers claim that the very first MicroRNA was not discovered in humans but instead it was discovered in a model organism. This discovery was made in 1993 by Victor Ambros, Rosalind Lee and Rhonda Feinbaum during a study of the organism nematode caenorhabditis elegans (C. elegans) regarding the gene lin-14 [3]. In their discovery, Victor Ambros, Rosalind Lee and Rhonda Feinbaum concluded that a short RNA product from lin-4 had the ability to regulate the lin-14 mRNA which causes less protein to be produced from the
lin-14 mRNA. This mainly occurs through the interaction between the lin-4 RNA and lin-14 mRNA [3].

Several years later, in 2000, following the Ambros’ lab discovery, Ann Rougvie’s lab discovered another small RNA which was involved in the control of development timing in *caenorhabditis elegans* [5]. Rougvie’s discovery made it apparent that Ambros’ earliest discovery was part of a wider phenomenon.

Bentwich I, Avniel A, and Bentwich stated that there are over 1000 miRNAs that may be encoded by the human genome [6]. Those 1000 miRNAs are said to target roughly about 60 percent of the genes found in humans and some other mammals. Also, it was claimed that these miRNAs are abundant in many cell types found in mammalian [6].

Many studies and research have investigated the role and the importance that miRNAs have on a lot of biological functions as well as the broad spectrum of human diseases. Immune response [13], tumor development [14], cell proliferation, differentiation, apoptosis [11][12], cardiac diseases [15] are just a few of the many functions/diseases that the miRNAs have influence over.

The regulation of target genes and miRNAs interactions is said to be sophisticated. Various miRNAs are able to regulate one target gene however it is stated through research that only one miRNA may affect multiple target genes [18]. As a matter of a fact, a growing number of studies continue to demonstrate that on average, one miRNA can regulate several hundred genes [16].
Targeted genes expression levels are affected negatively by MiRNAs through two distinct mechanisms. The affect level is dependent on the MiRNAs degree of complementarity to the target sequences of the genes [8].

The first mechanism involves the induction of the RNA-mediated interference pathway which occurs at the point where within the 3’ untranslated regions a perfect or near-perfect match transpires between miRNAs and their target mRNAs. Upon the miRNA-mRNA interaction recognition by the RNA-induced silencing complex, mRNA sundering occurs through an endonuclease activity. The second mechanism involves imperfect target matching which occurs when gene expression is controlled by miRNAs at the translational level [8][3].

Some researchers are hypothesizing that bladder cancer susceptibility is affected through miRNA biogenesis pathway and the common sequence variants in genes of miRNA [9].

According to Edgar G. and Vanvouri T. genes are divided into two separate coding groups, there are the non-coding genes which do not code functional proteins and the coding genes group which involves protein coding [10].

MicroRNAs (miRNAs) are studied as key genetic elements that regulate the gene expression involved in different human diseases. According to the miRBase.org public repository which is hosted at the University of Manchester, there has been over 1881 miRNAs sequences published for the homo sapiens organism [18]. The miRBase is the home to all published microRNA sequences and associated annotations [18]. miRBase does not accept the submission of newly discovered miRNAs until an article that describes their identification is accepted for publication in peer-reviewed journals [18]. The primary aim for the miRNA public repository is to implement a uniform naming scheme to newly discovered miRNAs by assigning stable and
consistent names [18]. Under the standard nomenclature here is how the following miRNA hsa-mir-124 precursor family example is broken out. The prefix “hsa” in this example signifies the organism which in this case is homo sapiens. The prefix hsa is then followed by a dash and a “mir”, and in this case the uncapitalized “mir” refers to the pre-miRNA whereas the capitalized “miR” refers to the mature miRNA form. mir or miR is then followed by a dash and a number which indicates the order of the name [19].

There exists miRNA annotations with additional lower case letters at the end of their name such as hsa-mir-106a and hsa-mir-106b, this nomenclature indicates that the latter have almost identical sequences with exception for either one or two nucleotides [19]. miRNA annotations that end with additional dash-number suffix usually imply that genes, pri-miRNAs and miRNAs are residing at different locations in the genome, and usually they are the ones that lead to one hundred percent identical mature miRNAs [19].

MicroRNAs that contain the following notation -3p and -5p as suffixes are usually found in comparable amounts and generally they tend to originate from opposite arms of the same pre-miRNA. Furthermore, miRNAs that contain an asterisk following the name are usually the indication of the discovery of mature species at lower levels from the opposite arm of what is called the hairpin. Also an important note to remember is that the abundance of the mature microRNA from one arm of the hairpin is higher than that found from the other arm [3].

Copy number variants also know CNVs are major contributors to the diversity to human genetics. CNVs alter the genomic sequence, and not only limited to the homo sapiens organism, by insertions, deletions, and duplications. These alterations on the genomic sequence range from a kilobase to multiple megabasepair in length [20].
Through studies and research, it was apparent that copy number variants are acknowledged to impact both disease as well as normal variation [21]. There exist multiple recognizable copy number variant-phenotype associations models, however they are nonexclusive [20]. The first model, comprised of the deletion and duplication of larger genomic segment as well as the existence in fewer allelic states. Those type of copy number variants are considered to be rare. Amidst low copy repeats or duplication of segments, a large portion of these copy number variants emerge by nonallelic homologous recombination [20]. Research indicates that these copy number variants tend to endure for a few generation within a pedigree or newly occurred ones. They are considered to have very short life span and are highly penetrant in the population [20].

The second model, involves copy number polymorphisms which are also known as CNPs. These copy number variant phenotypes are defined by the structure of genome and/or by the variation in copy number which define the multiple allelic state [20]. Copy number polymorphisms genes are considered to be very vital and embellishing for biology. They have major affect on immunity and drug responses to name just a few [22][23]. The following examples, Crohn's disease and HIV-1/AIDS susceptibility are considered a few of many phenotypes that are influenced by functional elements or genes dosage change [24][25].

MicroRNAs have the capability to bind to target messenger RNA transcripts of protein-coding genes. Their binding can negatively control the protein coding genes translation or they might cause degradation to messenger RNA. miRNAs and their mRNA target sequences are considered to be partially complementary [26]. Due to the nature of such imperfection in base
matching, it is very important as well as very difficult to be able to identify the targets prediction of miRNA accurately [27].

In recent years, the bioinformatics field has been developed into an important field which contributes heavily in many areas of biology. The term bioinformatics is associated with the reference to the creation of public and private databases which hold raw data that is related to the genetics, experimental molecular biology field and many more. For many biologists, researchers, and any person with interest in the biological field the very important function faced with is the analysis and interpretation of various data types. Analysis and interpretation can vary and includes but not limited to protein structures, protein domains, genomes, genomes mutations and sequencing nucleotide and amino acid [27]. It is safe to state that the fundamental focus of bioinformatics is to increase the biological processes understanding through the use of computers and computer programs.

Human miRNA has been a very hot topic since early 2000. There has been over 1800 miRNA sequences associated with the homo sapiens organism published to date [18]. This information is currently stored in many sophisticated and advanced public databases such as miRBase [18], microRNA [29], mirGen 2.0 [30] and many more. Furthermore, the National Center for Biological Information also known as NCBI, is the home to many biological databases such as and not limited to CloneDB, Conserved Domain Database. However, our interest will mainly be focused on the Database of Genomic Structural Variation also known as dbVar. The dbVar database is designed in order to store phenotype information and defined variants association in addition to archive information such as large insertion, translocations, inversions and deletion information associated with large scale genomic variations [31].
Analyzing the databases mentioned in the previous paragraph, one can easily determine the gap that they all share. Although these databases provide great deal of very important data, it is very apparent that as a researcher one must have to visit multiple database locations in order to gather the appropriate data for analysis and interpretation due to the fragmentation of the data. As an example, a biologist who is studying the structure of hsa-mir-17, this information can be available to him from the various databases mentioned previously. However, assuming that the biologists require more information such as retrieving the co-localization of miRNA loci with human genome CNV regions, he then would have to visit the location of another database and search for the information required for his study. This process can be very cumbersome at times. To contribute and help bridge the gap of data fragmentation for the researcher who is analyzing miRNA, we developed the HmiRNAsqv user interface.

![HmiRNAsqv Web Application](image)

**Figure 1.** HmiRNAsqv Web Application.
HmiRNAsqv, is a web application with a user friendly interface powered by a database backend. The aim of this web application is to provide biologists, researchers and every person with interest in the human miRNA studies, a user-friendly tool to help further the analysis of co-localization of miRNA loci with human genome copy number variant regions. Further, analysis pipelines such as miRNA-target to estimate the levels or locations of variations for genetic duplications, insertions or deletions were also offered. Such information could support the simulation of miRNA-target interactions.

HiMRNAsqv currently maintains 1881 miRNAs records along with their target predictions and copy number variants information records. The web interface provides the end user the ability to search our system and retrieve human miRNA records using the search functionality provided. miRNAs entries can be searched using specific keywords: miRNA name, Accession ID, Variant ID and Chromosome number. Upon the search execution, the user interface will return the appropriate result. Each record of the result contains detailed information of a specific miRNA, including miRNA identifier name, accession id, chromosome number, the start and end location of the chromosome, description, sequence structure, comment, mature sequences (-5p and -3p), copy number variants associations along with a hyperlink reference to the ncbi database, and finally targets prediction association along with hyperlinks for each association to the ncbi database. The data available in HmiRNAsqv database is extracted from the miRBase [18] and NCBI [31]. The HmiRNAsqv user interface is freely accessible from the URL of http://www.hmiRNA.com.
CHAPTER 2: HMIRNASQV DATABASE

2.1 Database Structure

Figure 2. HmiRNAsqv Database Architecture
The HmiRNAsqv database is designed using cutting edge technology. The database resides on a Microsoft SQL Server 2014. The Microsoft SQL server has been embraced to host the database due to the fact that it allows us to construct Big Data solution. Due to the nature of the size of the data that the HmiRNAsqv application will process, it is necessary to construct a responsive system which will react to the user’s request in a rapid manner.

The HmiRNAsqv database is comprised of six tables, mirna_chromosome_build, mirna_excel, mir_family_info, predicted_target, human_micro_rna and hmrna_var_ass. Each table is designed in order to hold specific data. In order to establish and enforce links between the appropriate tables, foreign keys have been applied. Foreign keys are fields or collection of fields which uniquely identify a row of another table from the table where it is being applied.

Along with the construction of the tables, stored procedures were also created. Stored procedures are considered as subroutines which are stored in the relational database data dictionary. Stored procedures facilitate data transaction between the relational database and web applications by allowing programmers to encapsulate the business logic of his/her application directly in the database. This methodology cuts down overhead in the application by eliminating the need to encode this logic in the program. By doing so, data integrity and consistency are preserved. Delegating the task to the database for data manipulation provides advantage to the application by eliminating the cost of computation from the web server and shifts it down to the database server level. One of the benefits for this delegation is the avoidance of network communication cost especially for structured query language statement that are complex in their nature. The second benefit is the fact that DBMS provide optimization through their optimizer when executing the structured query languages statements.
2.2 Tables Structure

In sections 2.21 through 2.28, we will be discussing the table structures in greater specificity as well as the nature of the data that are residing in each and every table. The need of data separation plays a very important role in formatting structure.

The HmiRNAsqv database model consists of look-up, operational, and association tables. An associative table or entity is considered to be a relationship between one or many tables/entities due to the fact that it can serve the purpose of joining them together. The table encapsulates properties from both. A successful association table must always contain the primary keys of all the adjoining entities. Also, an associative table can contain some other information that is directly relevant to the relationship.

A look-up table holds definitions mapping. For instance, a look-up table can be used to store colors. Colors can have many attributes such as color name, color description, RGB value, hexadecimal value and so on. It is required to have a primary key on the table to ensure uniqueness. The unique identifier can then be used in other tables such as an association or operation in order to reference a color. This methodology will eliminate the need to type in the color information in every row for every table it is being used in as well as ensure integrity and implement a standardization.

An operational table is designed mainly to process data such as deletion, insertion, update in real time. Usually the operational table is accessed by means of windows forms or web sites/applications mainly for the purposes stated before. The transactions that occur between the client face application and the database is usually handled through some sort of a stored procedure which are stored under the programability in the database.
2.2.1 Human MicroRNA Excel

The first table in our database model is named mirna_excel. This table is designed in order to hold information directly related to the miRNA records. This table is considered a look up table as it contains definitions mapping for miRNAs. The table is comprised of 10 fields. The first field in the table is mirna_id and has a type of varchar of length 300 character, this field is the representation of the miRNA identifier. Also, mirna_id is set to be a primary key in that table as duplicate miRNAs are not allowed. The implementation of a primary key ensures the uniqueness is not violated. An example of a mirna_id would be “hsa-mir-17”. The second field is named mina_acc and has a type of varchar of length 20 characters. This field is the accession id for specific miRNA records in the miRBase database.

According to the miRBase website, the accession number is considered to be the most stable identifier for an miRNA entry. It is reported that as relationships between sequences become clear, miRNA names might change from the ones that were published previously. This identifier will allow a traceable route to evolving names in the database with a consistent manner. Also, accession IDs convey little biological meaning [18].

The next field is the status, status has a type of varchar of length 100 characters, this field indicates whether a miRNA record has been altered/changed. When changes occur, the appropriate miRNAs records which are affected by the change will be flagged in the database and their status will change to “CHANGED” otherwise it will remain as “UNCHANGED”. The next field represents the sequence of miRNA named as sequence with the type varchar of length maximum characters. Figure 3 serves as an example to show the sequence of the hsa-mir-17 miRNA.
Figure 3. hsa-mir-17 miRNA Sequence Structure

The next field in the table is named mature_1_acc which is of type varchar of length 100 characters. This field contains the accession number of the mature miRNA. The accession identifier of the latter has the same concept as the mina_acc field. Next is the mature_1_id which is of type varchar as well with length comprised of 20 characters. This field represents the identifier number of the mature miRNA of a specific miRNA. miRNAs can have up to two mature miRNAs sequences, one on each arm of the hairpin. Mature miRNAs tend to have dash 3p or dash 5p as a suffix at the end of their miRNA names. miRNAs that contain the following notation -3p and -5p as suffixes are usually found in comparable amounts and generally they tend to originate from opposite arms of the same pre-miRNA [3].

This is followed by the mature_1_sequence field of type varchar and a length of maximum characters. This field represents the mature miRNA sequence of a specific miRNA. Mature microRNAs considered to be roughly about 21 to 23 nucleotides in length with small non-coding RNA molecules that naturally occur. The next field in that table is named mature_2_acc which is a varchar typed with length of 100 characters. It is then followed by the mature_2_id that has identical type as the latter however the length of characters differ as 20. Lastly, mature_2_sequence is the last field in that table architecture which contains the sequence structure of the mature miRNA. The type of that field is varchar with maximum characters in length.
2.2.2 Human MicroRNA Chromosome

The second table in our model is named mirna_chromosome_build. This table is designed in order to hold information directly related to the miRNA and chromosome. miRNAs play a very important role in genes regulation in specific chromosomes.

The authors of “Tuning into the signals: noncoding sequence conservation in vertebrate genomes”, Edgar and Vavouri have indicated the fact that the a large percentage of the human genome which is roughly 98% is composed of non-coding DNA. Also, the authors reveal that genes are composed of two groups. The first group is the non-coding genes which do not code functional proteins. The second group is the coding genes which are mailing genes that code proteins [10].

Depending on the organism type, it is revealed that the number of miRNA genes in some is higher compared to others. Observing multiple species, Atanu, Ghorai, and Ghosh Utpal pinpointed the fact that some species had the least absence of miRNA in the Y chromosome whereas in other species it was reported that the highest number of miRNAs was present on the X chromosome [32].
Evidently, it is clear that miRNA have great affect in many species including homo sapiens. Research in that area is growing at a rapid rate and the commitment to the data collected from scientific methods is needed at a much rapid pace. HmiRNAsqv database requires the raw data for miRNA and chromosome in order to be able to provide the user a complete picture of the miRNAs controlled genes in specific chromosomes.

The mirna_chromosome_build table is comprised of five fields. The first field, auto_mirna, is considered the primary key. The latter has an integer type, this field contains the auto mirna identifier. This identifier is used in order to combine rows from multiple tables, based on a common field between them which in this case is the auto_mirna. The next field in the table is named xsome, it is of type varchar and has a length of 20 characters. This field is designed in order to contain the chromosome number. The next two fields, are contig_start and contig_end and are of type bigint. A contig, short for contiguous, represents a consensus region of DNA which is the product of a set of overlapping DNA segments [33]. The last field in this table is called strand and it has a type of char with length of only two characters.

2.2.3 Human MicroRNA

The third table in our database model is called human_micro_rna. This table contains larger amount of meta data associated with each miRNA. The table structure is comprised of 22 data points. The first field in the table is named human_micro_rna_id which has an integer data type, it is considered the primary key in the table to ensure uniqueness. The next field is named auto_mina and it holds the data type of an integer. The field auto_mina is a foreign key to the table mina_chromosome_build described in section 2.2.2. This table also contains pre_mature_sequence_text, this field has a varchar type with length set to maximum characters.
The data in the latter is related directly to the pre-mature miRNA cursor and contains the sequence of the pre-mature cursor. The next data point in that table is named `pre_mature_comment` which has the type of `varchar` with maximum characters in length. The pre-mature comment field contains comments associated with the appropriate miRNA records recorded by the researchers and made available to the public in the miRBase database [18]. This table also contains the meta data of the mature sequences which falls on both sides of the hairpins in a miRNA with mature cursors. The metadata comprised of the identifier of
the mature sequence, the sequence structure as well as the start and end locations. The metadata is available for both -5p and -3p annotations. The start and end locations demonstrate the real estate certain mature sequences occupy on the appropriate arms of the miRNA hairpin.

2.2.4 Human MicroRNA Variants Assignment

The fourth table in the HmiRNAsqv database is named hmrna_var_assn. The table is designed in order to store the copy number variants associated with specific miRNAs. The db_var table is considered an association table considering the fact that it is housing the metadata for copy number variants along with a foreign key called human_micro_rna_id which is used to join on the human_micro_rna table to be able to associate the copy number variants with their appropriate miRNAs.

![Table Image]

**Figure 6.** HmiRNAsqv Variants Table

The table is composed of six data points as shown in figure 6. The first field in the table is named human_micro_rna_id and it possesses the type int. This field is set as a primary key. The next field is named var_id which contains the copy number variants identifier. Copy number variants also know CNVs are major contributor to the diversity to human genetics. CNVs alter
the genomic sequence by insertions, deletions, and duplications. These alterations on the genomic sequence range from a kilobase to multiple megabasepair in length [20]. Given the fact that their effect on the genome is important, it is apparent that the data about their location occupancy on chromosomes is very crucial to obtain. Considering the relevance of importance of the location ranges that the copy number variants inhabit, the data has been collected and stored in the NCBI databases. The National Institute of Health has made the latter raw data available to the biologists, researchers and the public. The raw data has been downloaded through the file transfer protocol and upon download, the data files have been parsed and properly inserted in the hmrna_var_assn table.

![Figure 7. Copy Number Variants Example](image)

The next data point in the table is named chromosome_num, the field is given a varchar data type with maximum characters in length. Followed by the chromosome_num, the next two data points, chromosome_range_from_num and chromosome_range_to_num, are given the type bigint.
The latter three data points reveal the information about the chromosome number and the location for which copy number variants inhabits. Figure 7 represents a screen shot of the top 10 records of the raw data set in the human_var_assn table. Copy number variants can easily be grouped by the human miRNA identifier. In the figure 7 example, a simple select statement on the human_var_assn where the human_micro_rna_id is equal to one will yield a set of 8 records containing only the data related to the human miRNA “hsa-let-7a-1”. MicroRNA (miRNA) hsa-let-7a-1 has a total of 8 copy number variants associated with it. The copy number variants reside on the 9th chromosome and inhabit different range number as stated in the data set in figure 7.


![Figure 8. CNVs Associated with hsa-let-7a-1]
The miRNA hsa-let-7a-1 association is only complete when the two tables, human_micro_rna and human_var_assn, are joined on each other using the human_micro_rna_id as a key. Figure 8 represents the data points associated with the human miRNA example used in figure 7.

### 2.2.5 Human MircoRNA Family Information

The fifth table in the HmiRNAsqv database is named mir_family_info. The latter is designed in order to hold data points related to the family ancestors of miRNAs. A MicroRNA family is comprised of a common ancestor which is derived from a group of miRNAs. MicroRNA family members are not always maintained in secondary structure or even primary sequence, however it is claimed through studies that those members have physiological functions that might be very similar [34].

![MicroRNA Family Information Table](image)

**Figure 9.** MicroRNA Family Information Table

The table mir_family_info is comprised of a collection of seven data points. The first two data points are used as primary keys. The mir_family_id contains the family identifier of specific miRNA. The latter has a type of varchar with a length that amounts to 200 characters. The next
field in that table is named mirna_id which has a type of varchar as well restricted to 200 characters in length. This field is designed in order to hold the miRNA identifier and can be used to join on the mina_excel or the human_micro_rna to query information that is related to a specific miRNA family. The following data point is named seed_plus_m8 which has a type of varchar and maximum length of 200 characters. MicroRNAs are classified into families based on the seed regions sequence similarities along with manual adjustment [34]. The next data point is named species_id and has a type of integer. This field is designed as an identifier which can be mapped to an enumeration list of species. In our relational database, HmiRNAsqv, the only specie we are concerned with is homo sapiens which has a specie identify of 9606. Following the species_id, the next three fields identify the mature sequence along with the family conservation and the mature accession identifier. The mature_sequence_text and mature_accession_id both have a varchar data type with maximum number of characters consisting of 200 in length. The family_conservation_id field has an integer data type.

2.2.6 Predicted Target

The sixth and last table in our database model is named predicted_target. The predicted target table contains eleven fields. The first field in that table is called mir_family_id and has data type of varchar consisting of maximum length of 300 characters. The latter is designed in order to hold the miRNA family identification described in section 2.2.5. The following field is named gene_id with the same data type and character length as the latter.

Gene ID represents the identification number of a specific gene. A gene can be defined as a genomic sequence that inhabits a certain locatable region. The genomic sequence corresponds to a unit of inheritance. Transcribed regions, regulatory regions and/or other functional sequence
regions are associated with the unit of inheritance. Genes are very important to human beings due to their importance on function ribonucleic acids (RNA) chains as well as on all protein determination [35][36].

![MicroRNA Predicted Targets Table](image)

**Figure 10. MicroRNA Predicted Targets Table**

The next field in the table is named gene_symbol_id followed by transcript_id and species_id. The three fields share the same data varchar data type as well as the 300 characters length. The gene_gene_symbol_id field holds gene symbol identifiers. Similar to the miRNA nomenclature, gene nomenclature also exists. In 1957 genetics symbols and nomenclature recommendations were published by an international committee [37]. In the year of 1960 formal guidelines for human gene names and symbols were widely accepted. Almost 19 years later, at the Edinburgh Human Genome meeting, full guidelines on the gene names and symbols were presented [38].
Figure 11. MicroRNA Predicted Targets

The next four fields have the same data type in common. The first field is named `utr_start_num` followed by `utr_end_num` then `msa_start_num` and `msa_end_num`. The UTR start and end numbers fields are related directly to the untranslated regions (UTR) of a coding sequence on miRNA strand. The fields indicate the start and end location of each gene. As discussed in section 2.2.3, miRNAs have two strands: the 5’ or 3’ side. Coding sequences that inhabit the 3’ strand side of miRNAs are called 3’UTR whereas the ones that are found on the 5’ side are called 5’ UTR [39].

The fields MSA start and end represent the multiple sequence alignments. The Multiple Sequence Alignment is very important in molecular biology and it plays a very important role in structurally important regions in proteins sequences as well as determining distances between sequences [40].

Finally, the last two data points in the predicted_target table are named `seed_match_text` and `prct_text` holding varchar data type with maximum character length of 300.
Lewis et al, discovered the existence of four different types of selectively conserved seed-matched types [41]. In the predicted target table, the four types of the conserved seed-matched types are: 6mer, 7mer-m8, 7mer-A1 and 8mer. Lewis et al. indicate that the 6mer site matches the 6 nucleotides miRNA seed compared to the 7mer-m8 site that matches the 8 nucleotide of miRNA. Moreover, the 7mer-A1 site is a representation of an A across from miRNA nucleotide 1 and finally the 8mer is a representation of both the A1 and m8 [41].

The tables in the HmiRNAsqv database described in section 2.21 through 2.2.6 play a vital role in data storage. The relational database provides us with great features and advantages which will facilitate data retrieval of miRNA information at granular levels. MicroRNA data can be retrieved from our database through the means of structural query language (SQL). However to ensure a pleasant and seamless experience to the end user of HmiRNAsqv, a web application has been developed in order to complement the database. The HmiRNAsqv web application is a great user interface developed with user experience in mind. Chapter three is dedicated in furtherance of explaining the intricacies of the development process, data gathering and parsing, data insertion in the database and search functionalities.
CHAPTER 3: HMIRNASQV WEB APPLICATION

3.1 Web Application Requirements

The HimRNAsqv web application is designed using cutting edge technologies powered by Microsoft. Microsoft Windows Server 2012 and Internet Information Server 7 has been chosen to represent our web platform to host our application. HmiRNAsqv is developed using the C# programming language and the .NET framework. The .NET framework is developed by Microsoft and primarily operates on Microsoft platform. The .NET framework includes the Framework Class Library which offers the programmer a variety of precompiled libraries readily available for usage such as cryptography, network communication, database connectivity to name a few. Presented with the CFL, the programmer can incorporate certain references to a specific dynamic-link library (DLL) in his/her code which will make its methods available for usage.

In addition to the C# programming language, the most popular client side programming language, JavaScript, has also been used in the development of the application. Client side scripting is vital to application responsiveness and provide the programmer with an array of advanced features that can be used to maximize the user experience when interacting with a website or a web application. Although, the use of JavaScript in our HmiRNAsqv application is not extensive, it has enhanced a few functionalities such as the auto complete and few functionalities in the back end (admin) section of the application.
3.2 HmiRNA Infrastructure

The infrastructure of the HmiRNAsqv has been thought of very carefully. The user experience is very important and requests responsiveness is vital. To ensure that the application is able to deliver the latter, the following methodology has been adopted. The web application has been published to the world wide web access using a dedicated web server running IIS 7. The HmiRNAsqv database has been allocated its own Microsoft SQL Server (MSSQL) in order to handle any transactions occurrences between the MSSQL and the web application.

![Figure 12. HmiRNA Infrastructure](image)

Figure 12 illustrate the infrastructure of the HmiRNAsqv web application. The separation of functionalities is a good practice which will ensure the desired output consisting of data integrity, optimized code, responsive request and a seamless user experience.
3.3 Web Application Infrastructure

In the previous section, 3.2, we discussed the importance of a well designed overall system infrastructure through the departmentalization of different responsibilities. Furthermore, to complement the latter, code infrastructure is also vital and failure to abide by best practices provided by Microsoft and other reputable sources will jeopardize the quality of the final product.

The HmiRNAseq web application is comprised of two sections: data processing and database search. The latter section is as important as the data processing however it requires less computation as most of the work is consumed on the database server. Nevertheless, the data processing section of the application is computationally expensive and it requires more CPU and memory resources. To ensure optimal implementation, the application is designed with one goal in mind, accuracy and optimal performance.

Microsoft n-tier structure concept is a great fit for our application. The n-tier structure ensure many great features and allow us to built a robust system. We decided to segment the application in three different tiers: presentation layer (UI), business layer, and the database layer.

The presentation layer primary responsibility is to administer the user interaction with the system and ensure a secure bridge between the presentation layer and the business layer through data encapsulation.

The business layer tier is considered a fundamental section of the application as it implements core functionalities and recognized for the business logic. This layer is designed in order to facilitate the interaction between the presentation and database layer. The business layer is considered the middle tier in our design.
The third and last tier in our application is the database layer. This layer has a distinct functionality which is to provide access to a data source through many different forms. The data source can either be exposed over a network which requires special connection or perhaps sources such as reading from excel spreadsheet, flat file or even a local access database. The database layer should never be exposed to the presentation layer and vice-versa, communication between the latter should only occur through a third component such as the business layer or some other means. The database layer in only designed to handle transactions between the application and the data source. In this layer, data manipulation is prohibited, the database layer should be a mapping process of the data source architecture.

Figure 13. HmiRNAsqv n-Tier Structure
The database layer in the HmiRNAsqv architecture relies heavily on the usage of stored procedures. Transactions such as select, insert, update and delete are all handled through stored procedures which are considered precompiled structural query languages. This concept eliminates the need to have the structural query languages embedded in the application as well as eliminating any convoluted code.

The n-tier structure is a very powerful concept which has been adopted widely. However, to take a full advantage of the concept one must implement his/her application using an object oriented design also known as (OOD) which we fully adopt in our design.

3.4 Web Application Functionalities

HmiRNAsqv has a user friendly interface. The aim of this web application is to provide biologists, researchers and every person with interest in the human miRNA studies, a user-friendly tool to help further the analysis of co-localization of miRNA loci with human genome copy number variant regions. To provide the service, we provided various functionalities in the application. A robust search has been implemented. In addition to the search, a detail view is also available. We also provided an export functionality which will allow users to export the the data set requested by them. The system permits users to also link out to various external data sources for in depth data on specific data points that are available in the HmiRNAsqv.

3.4.1 Search

The search functionality plays a vital role in the HmiRANsqv web application. It provides the user of the application the ability to search our database using four parameters. The search is designed using the AND operator. The user is presented with the option to search by the miRNA identifier, accession identifier, variant identifier and the chromosome number.
Figure 14. HmiRNAsqv Search Functionality

The original view on the search page is defaulted to display all of the 1881 records of the miRNA. A paging option is available to user in order to page the result set. Also, items per page are also present which will allow the user to display a specific number of items per page or all of the items at once.

Figure 15. miRNA-Var List View
Along with the robust search filter, the user is able to click on the export button to be able to export his search result into a comma separated value file. This functionality comes in very handy to the user of the application as it gives them the opportunity to save their searches. Figure 16 provides a glimpse at how the comma separated value file looks upon export.

![Image of Human MicroRNA Export](image)

**Figure 16. Human MicroRNA Export**

As noted in figure 16, a search for all the human microRNA that are associated with chromosome number 9 yielded 72 records. The exported file reflects that search. This is a very simple operation yet very powerful to the individual who is using the HmiRNAsqv application.
3.4.2 Detail View

The detail view section of the application grants the user the functionality of analyzing a specific human microRNA from the detailed view page in HmiRNAsqv. From the search page, the user can simply choose the opportunity to click on the view link associated with the record that he/she desires to inspect. Upon clicking on the link, the user is directed to the detailed view where more analysis on the human microRNA can be performed.

Figure 17. hsa-mir-107 Detail View

Figure 17 is a great demonstration of the detailed view of the human microRNA hsa-mir-107. The application provides important information such as the accession identifier, description, the sequence structure and comment. Furthermore, the application also displays the mature information related to the hsa-mir-107 human microRNA. The information can either represent the 3’ and 5’ or simply none depending on the availability. The example in figure 17
appears to have a mature sequence associated with the microRNA. The mature sequence falls on the 3’ strand of the hsa-mir-107 hairpin.

Figure 18. hsa-mir-101-2 Copy Number Variants

Copy Number Variants (CNV) are also available under the detailed view. Figure 18 represents the CNVs associated with the hsa-mir-101-2 human microRNA. As shown in figure 18, there exists six CNVs associated with that specific miRNA. Each CNV has the chromosome number along with the range from and to indicate the position of the CNV. Also, the variant identifier is a hyperlink which upon clicking it, the system will redirect you to the http://www.ncbi.nlm.nih.gov/dbvar/variants where there exist more detailed information regarding that
specific copy number variant. This functionality makes it very convenient to the researcher to be able to quickly retrieve his answer and spend more time on his analysis. The algorithm implemented ensures the CNVs match with the appropriate miRNA in a quick manner. The database is queried using the miRNA identifier to scan various tables and collect the data points that match the miRNA. The result of the query is returned to the application and displayed to the user in a friendly manner.

### Figure 19. hsa-mir-107 Predicted Targets

The final part of the detailed view in the application is the target prediction functionality. The application is able to return to the users the targets predicted and display them in a list view format that is both pageable and exportable. The user has the ability from within the detailed view to analyze the miRNA targets predicted and if desired, the user can simply click on the

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Gene Symbol ID</th>
<th>Transcript ID</th>
<th>Species ID</th>
<th>UTR Start</th>
<th>UTR End</th>
<th>MSA Start</th>
<th>MSA End</th>
<th>Seed Match</th>
<th>PCT</th>
</tr>
</thead>
<tbody>
<tr>
<td>87769</td>
<td>A2LD1</td>
<td>NM_033110</td>
<td>13616</td>
<td>1712</td>
<td>1718</td>
<td>3126</td>
<td>3132</td>
<td>7mer-1a</td>
<td>0.63</td>
</tr>
<tr>
<td>87769</td>
<td>A2LD1</td>
<td>NM_001195087</td>
<td>13616</td>
<td>1712</td>
<td>1718</td>
<td>3126</td>
<td>3132</td>
<td>7mer-1a</td>
<td>0.63</td>
</tr>
<tr>
<td>87766</td>
<td>A2LD1</td>
<td>NM_001195087</td>
<td>9665</td>
<td>1786</td>
<td>1792</td>
<td>3126</td>
<td>3132</td>
<td>7mer-1a</td>
<td>0.63</td>
</tr>
<tr>
<td>87769</td>
<td>A2LD1</td>
<td>NM_033110</td>
<td>9259</td>
<td>1011</td>
<td>1017</td>
<td>3126</td>
<td>3132</td>
<td>7mer-1a</td>
<td>0.63</td>
</tr>
<tr>
<td>87769</td>
<td>A2LD1</td>
<td>NM_033110</td>
<td>42254</td>
<td>179</td>
<td>185</td>
<td>3126</td>
<td>3132</td>
<td>7mer-1a</td>
<td>0.63</td>
</tr>
<tr>
<td>87769</td>
<td>A2LD1</td>
<td>NM_001195087</td>
<td>9785</td>
<td>2007</td>
<td>2013</td>
<td>3126</td>
<td>3132</td>
<td>7mer-1a</td>
<td>0.63</td>
</tr>
<tr>
<td>87769</td>
<td>A2LD1</td>
<td>NM_033110</td>
<td>37347</td>
<td>1476</td>
<td>1482</td>
<td>3126</td>
<td>3132</td>
<td>7mer-1a</td>
<td>0.63</td>
</tr>
<tr>
<td>87769</td>
<td>A2LD1</td>
<td>NM_033110</td>
<td>1616</td>
<td>1616</td>
<td>1625</td>
<td>3126</td>
<td>3132</td>
<td>7mer-1a</td>
<td>0.63</td>
</tr>
<tr>
<td>87769</td>
<td>A2LD1</td>
<td>NM_033110</td>
<td>42254</td>
<td>179</td>
<td>185</td>
<td>3126</td>
<td>3132</td>
<td>7mer-1a</td>
<td>0.63</td>
</tr>
<tr>
<td>87769</td>
<td>A2LD1</td>
<td>NM_033110</td>
<td>9615</td>
<td>1813</td>
<td>1819</td>
<td>3126</td>
<td>3132</td>
<td>7mer-1a</td>
<td>0.63</td>
</tr>
<tr>
<td>87769</td>
<td>A2LD1</td>
<td>NM_033110</td>
<td>9861</td>
<td>203</td>
<td>209</td>
<td>3126</td>
<td>3132</td>
<td>7mer-1a</td>
<td>0.63</td>
</tr>
<tr>
<td>87769</td>
<td>A2LD1</td>
<td>NM_033110</td>
<td>9861</td>
<td>1813</td>
<td>1819</td>
<td>3126</td>
<td>3132</td>
<td>7mer-1a</td>
<td>0.63</td>
</tr>
<tr>
<td>87769</td>
<td>A2LD1</td>
<td>NM_001195087</td>
<td>10116</td>
<td>1816</td>
<td>1825</td>
<td>3126</td>
<td>3132</td>
<td>7mer-1a</td>
<td>0.63</td>
</tr>
<tr>
<td>87769</td>
<td>A2LD1</td>
<td>NM_001195087</td>
<td>9785</td>
<td>2007</td>
<td>2013</td>
<td>3126</td>
<td>3132</td>
<td>7mer-1a</td>
<td>0.63</td>
</tr>
<tr>
<td>87769</td>
<td>A2LD1</td>
<td>NM_001195087</td>
<td>9785</td>
<td>2007</td>
<td>2013</td>
<td>3126</td>
<td>3132</td>
<td>7mer-1a</td>
<td>0.63</td>
</tr>
<tr>
<td>87769</td>
<td>A2LD1</td>
<td>NM_001195087</td>
<td>9785</td>
<td>2007</td>
<td>2013</td>
<td>3126</td>
<td>3132</td>
<td>7mer-1a</td>
<td>0.63</td>
</tr>
<tr>
<td>87769</td>
<td>A2LD1</td>
<td>NM_001195087</td>
<td>9785</td>
<td>2007</td>
<td>2013</td>
<td>3126</td>
<td>3132</td>
<td>7mer-1a</td>
<td>0.63</td>
</tr>
<tr>
<td>87769</td>
<td>A2LD1</td>
<td>NM_001195087</td>
<td>9785</td>
<td>2007</td>
<td>2013</td>
<td>3126</td>
<td>3132</td>
<td>7mer-1a</td>
<td>0.63</td>
</tr>
<tr>
<td>87769</td>
<td>A2LD1</td>
<td>NM_001195087</td>
<td>9785</td>
<td>2007</td>
<td>2013</td>
<td>3126</td>
<td>3132</td>
<td>7mer-1a</td>
<td>0.63</td>
</tr>
<tr>
<td>87769</td>
<td>A2LD1</td>
<td>NM_001195087</td>
<td>9785</td>
<td>2007</td>
<td>2013</td>
<td>3126</td>
<td>3132</td>
<td>7mer-1a</td>
<td>0.63</td>
</tr>
<tr>
<td>87769</td>
<td>A2LD1</td>
<td>NM_001195087</td>
<td>9785</td>
<td>2007</td>
<td>2013</td>
<td>3126</td>
<td>3132</td>
<td>7mer-1a</td>
<td>0.63</td>
</tr>
<tr>
<td>87769</td>
<td>A2LD1</td>
<td>NM_001195087</td>
<td>9785</td>
<td>2007</td>
<td>2013</td>
<td>3126</td>
<td>3132</td>
<td>7mer-1a</td>
<td>0.63</td>
</tr>
<tr>
<td>87769</td>
<td>A2LD1</td>
<td>NM_001195087</td>
<td>9785</td>
<td>2007</td>
<td>2013</td>
<td>3126</td>
<td>3132</td>
<td>7mer-1a</td>
<td>0.63</td>
</tr>
<tr>
<td>87769</td>
<td>A2LD1</td>
<td>NM_001195087</td>
<td>9785</td>
<td>2007</td>
<td>2013</td>
<td>3126</td>
<td>3132</td>
<td>7mer-1a</td>
<td>0.63</td>
</tr>
</tbody>
</table>

**Rows per page:** 10

[Link to download the file](\(\text{https://example.com}\.\text{Prediction of MicroRNA Targets.csv}\))
Gene ID column and at a click of a button he/she will be redirected to the National Institute Health database where more information can be acquired. Also the export to comma separated value file is readily available and at a click of a button the data can be downloaded.

The HmiRNAsqv database and web application combined create a great user friendly system that can be accessed freely at a click of a button. In the web application, the functionalities provide biologists, researchers and every person with interest in the human miRNA studies, a user-friendly tool to help further the analysis of co-localization of miRNA loci with human genome copy number variant regions. Further, analysis pipelines such as miRNA-target to estimate the levels or locations of variations for genetic duplications, insertions or deletions were also offered. Such information could support the simulation of miRNA-target interactions.

The HmiRNAsqv is an advanced and robust system with many great features. Furthermore, the system can be expanded to include more functionalities such as identifying the copy number variants and target prediction for organisms that are beyond homo sapiens. This enhancement can be incorporated into the system as a future work. The HmiRNAsqv is a great platform built using the object oriented design keeping in mind future enhancements and integration of new code to be able to accommodate the integration of many different organisms.
REFERENCES


36


[38] “HGNC.” GENE NAMES.ORG HUGO Gene Nomenclature Committee, National Human Genome Research Institute (NHGRI) grant U41HG003345 and Wellcome Trust grant 081979/Z/07/Z Web. 11 September 2014.


