DESIGN OF A SEQUENCE BASED MIRNA CLUSTERING METHOD; 
ANALYSIS OF FUNGAL MILRNAS AND HOST ORGANISM TARGET GENES

A THESIS SUBMITTED TO 
THE GRADUATE SCHOOL OF INFORMATICS 
OF 
MIDDLE EAST TECHNICAL UNIVERSITY

BY

KÜBRA NARCI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE 
OF MASTER OF SCIENCE 
IN 
BIOINFORMATICS

AUGUST 2014
DESIGN OF SEQUENCE BASED MIRNA CLUSTERING METHOD; PLANT FUNGAL MILRNAS AND HOST ORGANISM TARGET GENE ANALYSIS

Submitted by Kübra Narçi in partial fulfillment of the requirements for the degree of Master of Science in Bioinformatics, Middle East Technical University by,

Prof. Dr. Nazife Baykal
Director, Informatics Institute

Assist. Prof. Dr. Yeşim Aydın Son
Head of Department, Health Informatics

Prof. Dr. Mahinur S. Akkaya
Supervisor, Chemistry, METU

Assoc. Prof. Dr. Hasan Öğul
Co-Supervisor, Computer Eng., Başkent University

Examining Committee Members:

Assoc. Prof. Dr. Tolga Can
Computer Eng., METU

Prof. Dr. Mahinur S. Akkaya
Chemistry, METU

Assist. Prof. Dr. Bala Gür Dedeoğlu
Biotechnology, Ankara University

Assoc. Prof. Dr. Özlem Darcansoy İşeri
Institute of Transplantation and Gene Sciences, Başkent University

Assist. Prof. Dr. Yeşim Aydın Son
Information Institute, METU

Date: 27 August 2014
I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : Kübra Narcı

Signature :
ABSTRACT

DESIGN OF A SEQUENCE BASED MIRNA CLUSTERING METHOD; ANALYSIS OF FUNGAL MILRNAS AND HOST ORGANISM TARGET GENES

Narci, Kübra
MSc., Bioinformatics Program
Supervisor: Prof. Dr. Mahinur S. Akkaya
Co-Supervisor: Assoc. Prof. Dr. Hasan Öğul

August 2014, 94 Pages

Micro RNAs (miRNA) are small non-coding RNA molecules regulating transcription machinery of a cell. They are 20-22 nucleotide RNAs and they involve in gene silencing events by targeting specific regions of mRNA complementing to the miRNA. miRNAs show similarity over species in function and sequence as well, and they polycistronically expressed to enrol in the same processes. In current applications, cluster analysis of micro RNA sequences are only investigated by projection on their expression patterns. Such researches are still in focus of many scientists to discover functional annotations of miRNAs. However, as an important sequence similarity is detected in some of the functional groups of miRNAs and the sequence of miRNA is highly important in recognition of the target mRNA sequences in the RISC (RNA-induced silencing complex) binding regions, the new miRNA clusters targeting the same kind of mRNA sequences can be found by sequence similarity information. By representing the miRNA sequence as a mathematical number, it is possible to find significant similarities between miRNA sequences. In this perspective, a variety of clustering methods can be applied onto the informative matrix constructed by metrics of mature miRNA sequences. Here, we present a study that considers only mature miRNAs to obtain functionally relevant miRNA clusters. To this end, various machine learning methods are employed with different sequence representation schemes. Moreover, the data obtained by sequencing small RNAs fungal pathogens of plants were analyzed by the tool generated in this thesis to functionally annotate novel miRNA sequences for further
studies. Furthermore, small RNAs predicted by sequencing analysis, and some predicted fungal candidate targeting host plant messages studied as well.

**Keywords:** Micro RNA (miRNA), sequence clustering, small RNA sequencing, obligate plant fungus, mutualistic plant fungus.
ÖZ

SEKANSA DAYALI BİR MİRNA KÜMELEME YÖNTEMİ TASARIMI; BİTKİ MANTAR MİRNAALARI ANALİZLERİ VE KONAKTAKİ HEDEF GENLERİ

Narci, Kübra
Yüksek Lisans, Biyoenformatik Programı
Tez Yöneticisi: Prof. Dr. Mahinur S. Akkaya
Ortak Tez Yöneticisi: Doç. Dr. Hasan Oğul

Ağustos 2014, 94 Sayfa

saglamistir. Ayrıca, aday mikro RNA’ların bitki genomundaki olası hedef genleri saptanmıştır.

**Anahtar Kelimeler:** Mikro RNA, Sekansla kümeleme, küçük RNA sekanslanması, Biyotrofik bitki pathogeni, mutualistik bitki mantarı.
To my family,

To myself...
ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my advisor Dear Prof. Dr. Mahinur S. Akkaya. I am grateful to her for an original thesis topic and her continuous support, motivation and immense knowledge. Her continuous guidance helped me in my research and writing of this thesis.

I am very glad to have Assoc. Prof. Dr. Hasan Oğul as my co-supervisor. He shared brilliant ideas with me and helped me through generating algorithms as part of my thesis.

I am grateful to Assist. Prof. Dr. Yeşim Aydın Son and Assist. Prof. Dr. Aybar Can Acar for their knowledgeable courses helped me to learn what I know to complete this thesis.

Examining committee members Assoc. Prof. Dr. Tolga Can, Assist. Prof. Dr. Bala Gür Dedeoğlu and Assoc. Prof. Dr. Özlem Darcansoy İşeri are greatly acknowledged for their participation and valuable comments.

I am thankful to all my lab friends; Bayants Dadvadorj, Adnan Yaramış, Bahtiyar Yılmaz, and Sait Erdoğan for their friendship. I am also very thankful to my dearest friends and labmates Çağlar Özketen and Ayşe Andaç for preparing small RNA samples to sequence small RNAs and, to Burak Demiralay for his help in writing the codes, and to Zemran Mustafa not only for his friendship but also for all of the coffees he prepared for me during my struggle in data analyses in the lab.

I am also thankful to my housemate Fatma Akıncı and our honorary guests Mehtap Yonca and Serap Adakulu, for their never-ending patience to my all shortcomings.

I also acknowledge Scientific and Technological Research Council of Turkey (TÜBİTAK) for its master thesis grant under 2210 BIDEB, and the research grand of Akkaya (113Z038 and 110T445).

This study is dedicated to my family, my father Hayati Narcı, my mother Fatma Narcı, my sister Ebru Narcı Bulut and my brother D. Eren Narcı. I would like to thank to them for supporting me spiritually throughout my life.
# TABLE OF CONTENTS

ABSTRACT ........................................................................................................................................ iv  
ÖZ ................................................................................................................................................. vi

ACKNOWLEDGEMENTS ............................................................................................................... ix

TABLE OF CONTENTS .................................................................................................................. x

LIST OF TABLES ........................................................................................................................... xiv

LIST OF FIGURES ........................................................................................................................ xv

LIST OF ABBREVIATIONS ............................................................................................................. xvi

CHAPTER

INTRODUCTION ............................................................................................................................. 1

1.1 Motivation ................................................................................................................................ 1

1.2 Goal .......................................................................................................................................... 2

1.3 Sequence similarity .................................................................................................................... 2

1.3.1 Pairwise sequence alignment ............................................................................................... 3

1.3.2 K-mer substrings .................................................................................................................. 4

1.4 Clustering .................................................................................................................................. 5

1.4.1 K-means algorithm ............................................................................................................. 5

1.4.2 CLuster Aggregation algorithm ........................................................................................... 6

1.4.3 Self- Organizing Three Algorithm ....................................................................................... 7

1.4.4 Markov Clustering algorithm .............................................................................................. 8

1.5 Micro RNAs ............................................................................................................................. 9
1.5.1 Micro RNA biogenesis

1.5.2 Micro RNA target interaction

1.5.3 Micro RNA nomenclature

1.5.4 Micro RNAs in fungi

1.5.5 Micro RNA hijacking

1.6 Small non-coding RNA sequencing

1.7 Plant infecting fungi

2.1 Materials

2.1.1 Data: Human miRNA sequences

2.1.2 Data: Fungal small RNA sequences

2.2 Methods

2.2.1 Small RNA sequencing and sequence analysis

2.2.2 Sequence representation

2.2.2.1 Pairwise sequence alignments

2.2.2.2 K-mer counting

2.2.3 Clustering algorithms

2.2.3.1 Vecor based clustering algoritms

2.2.3.2 Graph based clustering algorithm

2.2.4 Statistical analysis of the clusters

2.2.5 Qualification of the efficiency of the clusters
2.2.6 milRNA target prediction ......................................................... 26
  2.2.6.1 psRNATarget tool ............................................................. 27
  2.2.6.2 miRBase database .......................................................... 28

2.2.7 Experimental analysis of the fungal milRNAs by the pipeline .... 28

RESULTS .......................................................................................... 29

3.1 Outline ....................................................................................... 29
3.2 Identification of small RNAs ..................................................... 29
3.3 Novel miRNA predictions ......................................................... 32
3.4 Nucleotide bias analysis of novel miRNAs ............................... 32
3.5 Matrices ..................................................................................... 34
3.6 Cluster algorithms ..................................................................... 36
3.7 Dunn Indeces ............................................................................. 37
3.8 TAM enrichment analysis .......................................................... 38
3.9 Sequence similarity based miRNA clustering pipeline .......... 40
3.10 Analysis of the milRNAs by the pipeline ................................. 41
3.11 milRNA selection for target gene analysis ......................... 41
3.12 psRNATarget tool predictions ............................................... 43

DISCUSSION ..................................................................................... 49

4.1 Why miRNAs ............................................................................ 49
4.2 Significance of miRNA representation ..................................... 49
4.3 Decision on which clustering algorithm ................................. 50
4.4 Importance of validation of the clusters ................................. 51
4.5 milRNA application to the pipeline ......................................... 52
4.6 Significance of miRNA prediction ........................................... 53
4.7 Importance of target analysis .................................................. 54
CONCLUSION ........................................................................................................57

5.1 Overview ........................................................................................................57

5.2 Conclusion ........................................................................................................57

5.3 Future studies ..................................................................................................59

REFERENCES ....................................................................................................62

APPENDICES ....................................................................................................70

APPENDIX A: SMALL RNA SEQUENCING REPORTS ....................................70

APPENDIX B: CLUSTER ANALYSIS OF MIRNAS ........................................72

APPENDIX C: SMALL RNA SEQUENCES SELECTED FOR TARGET GENE
ANALYSIS AND TARGET GENE ANALYSIS RESULTS ............................73

APPENDIX D: PLANT DEFENCE PROTEIN KEYWORDS .................................75

APPENDIX E: PERL SCRIPTS ............................................................................77

APPENDIX F: R CODES .....................................................................................88

APPENDIX G: CURRENT LIST OF PSRNATARGET LIBRARY .....................91
### LIST OF TABLES

**Table 1.** Fungi reference genome descriptions used in sequencing ......................... 20
**Table 2.** psRNATarget tool functions ........................................................................ 27
**Table 3.** Libraries for the organisms selected from psRNATarget catalogue ............ 27
**Table 4.** Total read, small reads, adaptors and clean reads for fungi ....................... 29
**Table 5.** Small RNA annotations for each organism ................................................. 31
**Table 6.** Clean reads, predicted miRNA and hairpin structure numbers per fungi .... 32
**Table 7.** MCL matrix by Smith-Waterman algorithm ............................................. 35
**Table 8.** MCL matrix by Smith-Waterman algorithm ............................................. 35
**Table 9.** Distance matrix by Smith-Waterman algorithm (SW-Distance) ................. 35
**Table 10.** Similarity matrix by Needleman-Wunsch algorithm (NW-Similarity) ....... 35
**Table 11.** Distance matrix by Needleman-Wunsch algorithm (NW-Distance) ........... 35
**Table 12.** Randomly Filled matrix ............................................................................ 35
**Table 13.** 3-mer distribution matrix (K-mer) .......................................................... 36
**Table 14.** Cluster numbers and data coverages of groupings by different methods ... 37
**Table 15.** Dunn Indeces of the groups for cluster algorithms applied ..................... 37
**Table 16.** Enrichment results of the clusters calculated by TAM tool ...................... 39
**Table 17.** miRNA cluster results .............................................................................. 41
**Table 18.** Number of predicted miRNA by more than 500 count number .............. 41
**Table 19.** Number of hits predicted by psRNATarget tool ........................................ 43
**Table 20.** psRNATarget predictions sorted by plant defense related proteins ........... 43
**Table 21.** Selected novel miRNA sequences ............................................................ 44
**Table 22.** Selected known miRNA sequences and their best homologs ................... 45
**Table 23.** Novel miRNAs’ predicted target regions in plant genomes ..................... 46
**Table 24.** Known miRNAs’ predicted target regions in plant genomes .................... 47
LIST OF FIGURES

Figure 1. Dynamic programming representation ........................................... 4
Figure 2. CLAG flowchart .............................................................................. 7
Figure 3. Stem loop of *C. elegans* lin-4 miRNA ........................................... 10
Figure 4. Sequence alignment of *let-7* family members ................................ 10
Figure 5. Plant and animal miRNA biogenesis pathways .................................. 11
Figure 6. A general approach for using high-throughput sequencing data in search of small RNAs ...................................................................... 16
Figure 7. Wheat yellow rust ............................................................................ 18
Figure 8. Powdery Mildew .............................................................................. 18
Figure 9. Vector representations of a miRNA sequence .................................... 23
Figure 10. Transition of a graph to a matrix by MCL algorithm ................. 24
Figure 11. Example R code to calculate Dunn Index ...................................... 25
Figure 12. TAM output with an example dataset ............................................ 26
Figure 13. psRNATarget tool result page illustration ...................................... 28
Figure 14. Graph demonstrating length distribution of read small RNAs of A) Bgh, B) Pst, and C) Pi ................................................................. 30
Figure 15. Pie charts representing small RNA annotations for Bgh (*B. graminis*), Pst (*P. striiformis*), and Pi (*P. indica*) ........................................ 31
Figure 16. Graph representing first nucleotide bias for novel miRNA predictions of A) Bgh, B) Pst, and C) Pi ................................................................. 33
Figure 17. Graph for position base bias of novel miRNA predictions of A) Bgh, B) Pst, and C) Pi ................................................................. 34
Figure 18. Workflow of miRNA sequence based clustering .............................. 40
Figure 19. Novel miRNA and known miRNA distributions in fungi ............... 42
Figure 20. Novel miRNA precursors shown in Table 21 ............................... 44
LIST OF ABBREVIATIONS

miRNA: micro RNA
UTR: Untranslated Region
RISC: RNA-induced silencing complex
MiRNA: miRNA-like small RNA
RNAi: RNA interference
NcRNA: Non-Coding RNA
NcRNA-Seq: Next Generation Sequencing
NextGen-Seq: Next Generation Sequencing
SNP: Single Nucleotide Polymorphism
Pst: *Puccinia striiformis* f. sp. *tritici*
Bgh: *Blumeria graminis* f.sp. *hordei*
Pi: *Piriformospora Indica*
CLAG: CLuster Aggregation
SOTA: Self-Organizing Three Algorithms
MCL: Markov Clustering
DI: Dunn Index
TAM: Tool for Annotations of miRNA
HMDD: Human MicroRNA Disease Database
UPE: Unpaired Energy
SOAP: Short Alignment Program
MFE: Minimal Free Energy
MSA: Multiple Sequence Alignment
CHAPTER I

INTRODUCTION

1.1 Motivation

Non-coding RNA (ncRNA) term is used for functional RNA products which are not translated into protein. NcRNAs separated into two groups: long and small non-coding RNAs. Small ncRNAs include highly copious and functionally vital RNAs like micro RNAs, small nucleolar RNAs, small interfering RNAs, and piwi-interacting RNAs (Zymański, Arciszewska, & Ywicki, 2002). Micro RNAs (miRNAs) as an abundant group of small non-coding RNA molecules involve in post-transcriptional gene silencing events. This mechanism is highly conserved in most of the organisms (D. P. Bartel, 2004). Some of the MIR genes, miRNA transcribing genes, are found to be polycistronically transcribed into miRNAs and located into the same chromosomal positions; they are called as miRNA clusters. In some of the miRNA clusters, a recognizable sequence similarity is known (Altuvia et al., 2005). Therefore, instead of the traditional clustering approaches that uses expression levels of miRNAs, sequence based clustering methods can be established.

Whenever the importance of small RNA machinery is realized, scientists are directed to discover many small RNA sequences. The process for the exploration of a new small RNA is speeded by New Generation Sequencing (NextGen-Seq) methods. Small non-coding RNA sequencing (ncRNA-Seq) is a NextGen-Seq technique. It is specialized to measure and degree of small RNA amount within the transcriptome (Collins, 2011). With this technological breakdown, high amount of novel small RNA sequences need to be proven is generated. Therefore, requirement of new bioinformatic tools are initiated.

Recently, in an interesting study miRNAs from different species travelling in the human body fluid is found. This brings the idea that miRNAs are drifting in a cross kingdom manner (M. Jiang et al., 2012). Recently, for the first time a fungal pathogen controls its host gene expression to defeat host immunity by locating itself in to the host cell (Weiberg et al., 2013). Therefore, it would be the case for other fungal pathogens as well as their target genes of pathogen small RNA molecules can be search in plant genomes progress as being follower of that topic.
1.2 Goal

My thesis focused on two very essential objectives; the first aim of the study is to develop a method for grouping mature miRNAs by their only sequence information. Clustering algorithms are applied into a matrix filled with sequence metrics. Sequence metric is either calculated through dynamic programming pairwise sequence alignment algorithms; Smith-Waterman (Smith & Waterman, 1981) and Needleman-Wunsch (Needleman & Wunsch, 1970) or by calculating their k-mer presences. The performance of the groups created by these methods is statistically analyzed by using Dunn Index (Dunn, 1973) calculation. The functionality of the pipeline is tested with a well-known human miRNA dataset. Tool for Annotations of miRNA (TAM) (Lu et al., 2010) is used to test the groups, annotate them into functional categories and thus calculate the enrichment of the miRNA groups with any purposeful similarities. As a result, the established workflow of the study is presented as part of this thesis.

The second objective of the thesis was to hunt down the candidate miRNA like small RNA (miRNA) sequences of the plant pathogenic fungi; Puccinia striformis f. sp. tritici and Blumeria graminis f. sp. hordei, and the symbiotic fungus Piriformospora indica. The pipeline established was applied into the candidate miRNA sets. Additionally, plant mRNAs of fungal miRNAs were detected using known bioinformatics tools such as psRNATarget tool (Dai & Zhao, 2011) and miRBase database (Kozomara & Griffiths-Jones, 2011). In conclusion, the target gene information and miRNA clusters are presented to find most potential miRNA candidates to work for further studies.

1.3 Sequence similarity

Similarity is the measure of how two objects are alike, with respect to the opposite of it; dissimilarity is the measure of how different two objects are. In molecular biology, two biomolecules are said to be homologous if they share a common ancestor and habitually the question whether two nucleic acid sequences are homologous or not are search with molecular biological tools (Jeremy M Berg, 2002). In which, similarity is used as a mark of the homology of the sequences, since over time evolutionary changes like mutations, insertions and deletions on the sequences make two bimolecular objects dissimilar (Pascarella & Argos, 1992).

The similarity of four letters of sequences (DNA or RNA) like any words cannot be predicted directly from the letter itself. Computational biologists have attempted to understand this problem after 1879 works of Lewis Carrol who only interested in similarity of the ordinary words. To better understand how two biomolecular sequences similar, whether they are homolog or not, the necessary of transformation of the sequence similarity content into a mathematical measure is understood. Thus, sequence alignment algorithms were born from this requirement.
1.3.1 Pairwise sequence alignment

To degree two sequence distances between each other pairwise sequence alignment approach is used. In general there are two considerations; global vs local alignment. Global alignment attempts to align every single word in two sequences. Needleman-Wunsch algorithm is developed first for global alignment search (Needleman & Wunsch, 1970) (Equation 1).

**Equation 1.** Needleman-Wunsch algorithm (Durbin et al.1998).

\[
F(i, j) = \max \left\{ \begin{array}{l}
F(i - 1, j - 1) + s(x_i, y_j), \\
F(i - 1, j) - d, \\
F(i, j - 1) - d.
\end{array} \right.
\]

Later, an implementation to global alignment, local alignment method is established, Smith-Waterman Algorithm (Smith & Waterman, 1981). Local alignment, with respect to global alignment, does not effort to align every single world but only support significantly similar words to find local similarities (Equation 2).

**Equation 2.** Smith-Waterman algorithm (Durbin et al., 1998).

\[
F(i, j) = \max \left\{ \begin{array}{l}
0, \\
F(i - 1, j - 1) + s(x_i, y_j), \\
F(i - 1, j) - d, \\
F(i, j - 1) - d.
\end{array} \right.
\]
The both of the algorithms use a dynamic programming method to find an optimal solution. The idea of dynamic programming is that dividing the problem into sub problems, solving the sub problems individually to reach the eventual solution (Figure 1). In general dynamic programs find the best optimal solution. Commonly, alignment algorithms find the best alignment between two sequences by simple scoring schemes; giving a positive penalty for match conditions and negative penalties for mismatch and gap conditions. The result of this computation shows the weight of the similarity between the sequences (Durbin et al., 1998).

1.3.2 K-mer substrings

Representation of the distribution of k-mer substrings of a sequence is an alignment free approach with respect to pairwise sequence alignment tools independent from
sequence order. It aims to produce a sequence model definite on the distribution of k-mers, namely all probable k length substrings (Oğul & Mumcuoğlu, 2007).

1.4 Clustering

Clustering is a machine learning method widely used for data mining studies. It is an unsupervised learning job, aims to classify the objects by using their characteristic features. Clustering is defined as grouping with a given set of objects based on their similarity and dissimilarity measures. In a well developed clustering, intra-clusters needed to show great correspondence, with respect to there are small coherence in inter-clusters (Sisodia, 2012)

The final purpose of clustering the data into partitions is to understand the structure of the data set, like outliers, patterns, and distribution of the data, and to see the natural groupings. There are several algorithms performing clustering according to size and shape of the data. In bioinformatics, since there is variety of data structures it is a multi-objective optimization problem. The problem why there is so many cluster algorithms explained by Vlademir Estivil-Castra by the fact that the notion of “cluster” cannot be defined altogether (Sisodia, 2012).

In general clustering algorithms can be classified into 2 groups; vector and graph based clustering algorithms.

1.4.1 K-means algorithm

K-means (Macqueen, 1967) is the classical yet one of the most used methods of partitional clustering. It clusters the dataset into k number of groups. Grouping is done by minimizing the sum of squares of distances between data points and the corresponding cluster centroids. The logic of the method depends on the iterations of these steps; first determination of the centroid coordinate, evaluating the distance of each object to the centroids and last grouping into the objects based on minimum distance (Macqueen, 1967).

“Stated informally, the k-means procedure consists of simply starting with k groups each of which consists of a single random point, and thereafter adding each new point to the group whose mean the new point is nearest. After a point is added to a group, the mean of that group is adjusted in order to take account of the new point. Thus at each stage the k-means are, in fact, the means of the groups they represent (hence the term k-means) (Macqueen, 1967)”.

Prior to these steps however, k (number of clusters) must be specified. Actually, if the dataset is unknown and analyzer doesn’t know how many grouping will be done, optimization of k becomes one of the weaknesses of this method. Regardless of the fact that, k-means is widely used method, if the numbers of data are not high enough,
initial groupings will determine the cluster contents significantly. Therefore, with different centroids, different classifications are possible and the evaluation of validity of these clusters becomes substantial (Rawlins, Lewis, Hettenhausen, & Mirjalili, 2012).

Furthermore, k-means has problems when the data clusters are different in sizes, density, with non-globular shapes and with outlier points. Nonetheless, as the size of the data increases, some of the weaknesses of the method can be overcome. Also, running the set with different initial centers can be a solution to shape problems (Jain, 2010).

1.4.2 CLuster Aggregation algorithm

CLuster Aggregation (CLAG), (Dib & Carbone, 2012) is a clustering method specially established for large non-uniform biological datasets. It is an unsupervised non-hierarchical method aiming to zoom in only compressed regions in the uneven datasets by given parameters. The algorithm iterates for suitable aggregations on the dense regions. Therefore, the algorithm does not group whole data; instead, only finds best similarities in the particular correlation metrics.

CLAG handles with two types of metrics; feature metrics like classical cluster algorithms uses and symmetric metric output of all-to-all kind of relation. One of the benefits of the algorithm is that the cluster number is not specified by the user. Depending on two input parameters, ∆ and the threshold for environmental and symmetric scores, clustered amount of data points is arranged. Parameter ∆ decides the proximity of the clusters, namely distance between intra and inter classes restrained by this parameter. By this way, for each present group a signal value is calculated (after each aggregation of clusters), and the strength of the group is decided by the threshold for environmental and symmetric score (only for symmetric matrices). Subsequently, only most momentous clusters remain. Furthermore, since the algorithm does not samples the data with initial centroids, it does not suffer from the problems of k-means, like yielding different clusters for repeat runs and dealing with dense-shaped data points (Dib & Carbone, 2012). Figure 2 represents two steps and iterations of CLAG algorithm.
Figure 2. CLAG flowchart. Representation of the different stages of the algorithm is shown. CLAG has three inputs from the user: Matrix, $\Delta$ and the scores threshold N. Environmental score differs between the feature matrix (solid line) and symmetric matrix. Directly taken from (Dib & Carbone, 2012).

1.4.3 Self-Organizing Three Algorithm

Hierarchical Clustering algorithms are one of the classical ways to find homogeneous regions in the datasets. The Self-Organizing Three Algorithm (SOTA) (Dopazo et al., 1997; Herrero et al., 2001) is a hierarchical clustering method unusually using neural network (Self-Organizing Map- SOM) centered on a distance function well fit to the nature of the data. Neural network propagates to fit the topology of the set into a binary tree.
The algorithm aims to integrate advantages of both methods hierarchical clustering and SOM without suffering from their problems. SOTA is a divisive method, clusters form from a growing neural network, with respect to agglomerative approach of hierarchical algorithms. This feature of SOTA has led to stop at any desired level of hierarchy until cluster numbers reach to equality with data points, and so, arrangement of the homogeneity of the clusters is arrived. Prior to the analysis, the algorithm evaluates the distances between the elements and choses two main nodes. The following divisions calculated up to homogeneity of these nodes are absolute not change. This makes the centroids of the data fixed; re-runs of the data do not change the position of the centroids and thus, with respect to k-means algorithm, cluster members remains fixed (Dopazo et al., 1997; Javier Herrero et al., 2001).

In the case of SOM clustering, the topology is hexagonal or rectangular; there is no toleration to obtain clusters at preferred stages. Instead, SOTA uses hierarchical approach to turn topology into binary to gather clusters at preferred stages. SOTA method is proven to cluster large gene expression patterns like microarray analysis results. The method is efficient to be able to isolate the real clusters from the noise of the data (Herrero et al., 2003).

Different from K-means and CLAG algorithms which are using partition method, SOTA uses hierarchical evaluating system on SOMs, so topology of the data becomes binary tree. However, when the matrix is given as input to the SOTA, miRNA representation becomes a vector like in K-means and CLAG. Each vector represented as a cell prior in the algorithm. Distances between the cells are evaluated through a distance function as default Euclidean distance. Two main nodes are chosen by the algorithm as the division creates two sets showing the best homogeneities inside. After that, on each set again two nodes are chosen and homogeneity calculated. Growing network works in the same way by means of re-evaluating the binary tree at each step.

One advantage of the method is that cluster centroids shows homogeneity of the dataset, and centroids do not change by re-runs. Furthermore, only end-user given parameter is cluster number, therefore there is no need to make any optimizations.

1.4.4 Markov Clustering algorithm

Markov Clustering Algorithm (MCL algorithm) (Dongen, 2000) is a graph clustering method developed by Stijn Von Dongen at 2000 as his PhD thesis. This algorithm has been widely used in bioinformatics to find functional relations in protein datasets. Such as OrthoMCL (Li et al., 2003) and TribeMCL (Enright, et al., 2002) use MCL algorithm applied into all-to-all BLAST results of protein sequences. Proteins become the nodes of the graph and the distance objects between them are considered as pairwise comparisons of the proteins.
MCL algorithm uses a weighted symmetry matrix which shows the pairwise distances between the items in the dataset. The pairwise weights are turned into transition probabilities with normalization. The algorithm makes random walks using probability matrix to find inter connected elements namely the clusters. In general, the algorithm has two steps; normalization and inflammation. Normalization step is responsible for calculating probabilities of each connection for each node in the graph. After each normalization step, inflammation is taking square of the matrix simply to overweight current strength connections and on the contrary underweight the weak ones. Inflammation value can be arranged by the behavior and the structure of the dataset. Inflammation value can be increased to find more strength connected clusters and to observe biconnected groupings, it can also be decreased to find naturally big connections or to present well separated groups. These two steps iterate on the graph until the convergence is fixed (Enright et al., 2002; L. Li et al., 2003).

The algorithm is very gainful on classical vector based cluster algorithms when the distance metric is considered as important between objects. The method instantly found the cluster number unlike the classical methods. Unlike k-means and SOTA cluster number is not provided by the end user. This behavior leads to find nature of the dataset and remove the noise from the actual clusters.

1.5 Micro RNAs

miRNAs are genes of which small noncoding RNA products regulate encoding machinery. They are quite small sequences, 20-22 nucleotides in length and they involve in cleavage or translational repression events by precise sequence complementarity to their target RNA sequence (D. P. Bartel, 2004; Lagos-Quintana, Rauhut, Lendeckel, & Tuschl, 2001)

The first miRNA, lin 4, is discovered in 1993 and is found as getting complementary into the gene lin-14 3’UTR region in the organism C. elegans (Lee, Feinbaum, & Ambros, 1993). Hairpin structure of miRNA lin-4 is represented in Figure 3.

From that time the first miRNA is found as repressing transcription of a gene, miRNA studies become one of the hottest field of investigation. miRNAs were also found in other organisms, such as flies, fishes, mammals, plants, viruses and pathogens with various functions like developmental processes, cell death, proliferation, and fat storage (D. P. Bartel, 2004; Lagos-Quintana et al., 2001; Lee & Ambros, 2001).
Figure 3. Stem loop of *C. elegans* lin-4 miRNA. Hairpin structure of miRNA lin-4 is represented with 5’ and 3’ overhangs. The red sequence represents mature miRNA. Adapted from (D. P. Bartel, 2004).

The other functionally important miRNA is *let*-7, which was found controlling differentiation in *C. elegans*. The *let*-7 was a well conserved sequence (Figure 4) in animals as well its role regulating cellular processing. miRNA family members generally involved in the same function in the cell. For example, *let*-7 family members function together to control developmental timing. The high sequence similarity in this family provides a baseline for further experiments to explore similar miRNA sequences in various organisms (Abbott et al., 2005).

![miRNA hairpin structure](image)

**Figure 4.** Sequence alignment of *let*-7 family members.

1.5.1 Micro RNA biogenesis

RNA polymerase II and III (pol II and III) transcribe micro RNA genes (MIR genes) into pri-miRNA hairpin structures (Borchert et al., 2006; Y. Lee et al., 2004). Pri-miRNA is nearly 70 nucleotides in length and consists of a hairpin loop with 5’ cap and 3’ poly-A tail. The hairpin structure with the cap and the tail process into pre-miRNA (precursor miRNA) structures through clearing off the cap and the tail with the enzyme, Drosha, which cuts RNA, in the nucleus of animals cells (Cai et al., 2004; Y. Lee et al., 2004).
Figure 5. Plant and animal miRNA biogenesis pathways. a) miRNA biogenesis pathway in animals b) in plants. MIR gene is transcribed by PolII and pri-miR which is attached by cap proteins and polA tail is formed in the nucleus. Pri-miR is transformed into pre-miRNA after 1\textsuperscript{st} processing and miRNA-miRNA* duplex is formed after 2\textsuperscript{nd} processing. In animal, precursor miRNA is exported into cytoplasm by Exp5/ Ran-GTP complex and Dicer cleares off the stem in nucleus to form miRNA duplex. In plant, methylated miRNA duplex is exported into the cytoplasm and miRNA* is released when RISC complex met with the duplex. In the RISC complex, base pairing between miRNA and its target RNA occurs and eventual mRNA cleavage occurs. Directly taken from Zeng, 2006.

Pre-miRNA is exported into cytoplasm by Exportin receptors that recognize 3’ end of the stem loop, and hairpin loop is sliced off by RNase III enzyme Dicer. In the end, mature miRNA duplex (miRNA:miRNA*) is formed (Figure 5). In the cytoplasm, the yield miRNA:miRNA* duplex is imperfectly interacted, this feature
has affects in miRNA processing by Dicer and leads miRNA sequence to match target RNA sequence (He & Hannon, 2004; Lund & Dahlberg, 2006).

miRNA biogenesis shows some differences in plants (Figure 5). Dicer homolog, Dicer-like, both micro processes the pri-miRNA and cleaves the hairpin loop off to process pre-miRNA sequences in the nucleus. 3’ overhangs of the miRNA: miRNA* duplex is methylated by RNA methlytransferase Hua-Enhancer1 to be get exported into the cytoplasm. Quickly after the miRNA duplex is formed and un-wound by the enzyme Dicer; one of the strands the miRNA, Dicer and many other proteins form the RNA-induced silencing complex (RISC) for posttranscriptional repression (Lelandais-Brière et al., 2010).

1.5.2 Micro RNA target interaction

By RISC region, miRNA sequence is used as template to complement the target mRNA sequence, and as a result, the fate of miRNA targeting is often gene silencing. There are two modes of the silencing, one is RNA degradation (miRNA cleavage) and the other is blocking the mRNA being translated (translational repression). However, recent reports are suggesting that that would be a positive effect of miRNA targeting like sponsoring transcription or translation and stabilization of transcription (Asgari, 2011).

miRNA binding sites in target mRNA region is generally in 3 ‘UTR region, occasionally in 5 ‘UTR region of the gene in animals and typically in a coding region in plants. The percentage of the complementarities changes by, and depends on type of the organisms (Pratt & MacRae, 2009).

In animals, miRNA to target complementarity is nearly partial. This situation makes miRNA target prediction a hard occasion. However, the seed region is always important for determination of the sequence. Second to eight nucleotides of the pre-miRNA sequence are accepted as key nucleotides (D. P. Bartel, 2013). miRNAs in plants are different than animal miRNAs regardless of the fact that the same RNA polymerases are used in synthesis of them from MIR genes. Pri-miRNAs in plants are more variable in size (64 to 303 nt) (B. Bartel & Bartel, 2003) than in animals (nearly 70 nt) (Rhoades et al., 2002) and MIR gene location differs. These changes have effects on the formation of the hairpin structure. Arms of the hairpins are respectively well matched and contain smaller amount mismatches in plants than animals (Rhoades et al., 2002). There is also nearly perfect match between the miRNA sequence and its corresponding target RNA (B. Bartel & Bartel, 2003; D. P. Bartel, 2004; Pratt & MacRae, 2009).

Furthermore, it is found that there is evolutionary importance of these mismatches and they are well conserved in plants. The mismatches are tolerating the easier release from the RISC complexes when perfectly located into its target (B. Bartel & Bartel, 2003).
1.5.3 Micro RNA nomenclature

Since many of miRNAs across various species are being discovered, a system of nomenclature was required in a systematic manner thus improved nomenclaturing is progressively established to cover all the miRNAs discovered. Experimentally confirmed miRNAs are named with a mir following a dash and a number. If the –r letter in mir is uncapitalised, it refers to the pre-miRNA, and otherwise it is the mature form of that pre-miRNA (Abbott et al., 2005; Lee & Ambros, 2001; Lee et al., 2002; Pratt & MacRae, 2009). For example, mir-1 is the pre-miRNA of miR-1 which is a mature miRNA.

The lower case letter, followed by the number, shows similar miRNA structures. Such as, miR-1a and miR-1b can be in the same family. The exactly same miRNAs from different chromosomal loci are indicated by an additional number like; miR-1-1 and miR-1-2.

A prefix is added for annotations of organism specification. For example, ‘hsa’ prefix is used for Homo sapiens. miRNA of the human is presented as hsa-miR-xx. Often, from the same pre-miRNA stem-loop more than one mature miRNA is formed. miRNAs originated from the 3’ end or 5’ end of the hairpin is indicated with a -3p or -5p suffix (miR-1-3p, miR-1-5p). This suffices only added if the stem loop is producing mature miRNAs from both overhangs. Yet, often, only one arm of the hairpin is abundant in all the species (Bartel, 2004). This distinction is introduced by an asterics (*). The less abundant mature miRNA is indicated by an asterics. For example, miR-1 can be more abundant than miR-1*, most often the one with the asterics is also the anti-miRNA.

1.5.4 Micro RNAs in fungi

Some RNA interference (RNAi) mechanism is conserved in nearly all eukaryotes regardless if the discovered dicer components are determined or not. The miRNA pathway is one of the RNAi mechanism is proven in fungi, yet. This commonly raises a belief that miRNAs are not present in them (Lee et al., 2011). However, diverse mechanisms generating miRNA-like small RNAs (milRNA) and their stem-loop hairpins were observed in filamentous fungus like Neurospora crassa, Sclerotinia sclerotiorum and Metarhizium anisopliae independent from the dicer reliant common pathway (Kang et al., 2013). In fact, existence of miRNA like small RNAs (milRNA) in fungi lead scientists to consider that RNAi mechanism is as ancient as unicellular organism evolution (Kang et al., 2013; Lee et al., 2011).

Nevertheless, neither miRNA nor milRNA presence is verified in obligate fungal pathogens like Puccinia striiformis f. sp. tritici and Blumeria graminis f.sp. hordei or fungal root endophyte Piriformospora indica, which makes this thesis uniquely original.
1.5.5 Micro RNA hijacking

Communication of organisms with hormones or growth factors is valid for the whole ecosystem. However, discovery that miRNAs are used for cross species signalling is very innovative. MiRNAs travelling in the human body fluid and serum in a cross kingdom manner is found. In fact, these miRNAs are found to be very stable at broad range of pH, RNases, and changing temperatures (Jiang et al., 2012). Therefore, in addition to the existing miRNA regulation pathways in all the eukaryotes, the discovery of miRNA sequences found as regulating foreign cells from diverse kingdoms is a very exciting one. Thus suggesting that cross kingdom regulation is indeed possible through an unknow miRNA delivery pathway. The first fungal small RNA supressing a plant protein synthesis is reported by Weiberg and collegues at 2003. They suggest that small RNAs of the fungal pathogen Botrytis cinerea regulates host RNA interference (RNAi) machinery by interfering of gene transcripts of the protein Argonate (AGO) by silencing it (Weiberg et al., 2013). Thus, this pathogen is able to relocate its infectious sRNAs into the host cell to defeat host immunit to achieve successful infection by interfering the plant’s own miRNA generation pathway.

1.6 Small non-coding RNA sequencing

There are numerous technologies to infer and measure transcriptome. One of them is hybridization based quantification approach, microarrays. High throughput microarray analysis used to quantify transcriptome and map them to corresponding genes in the genome in relatively inexpensive way. However, microarray analysis often includes several normalization and standardization steps to remove the bias with in the data (Wang et al., 2010).

In contrast to microarray technology, RNA sequencing (RNA-Seq) methods made it possible to directly deduce from cDNA fragments. There are old fashion approaches which usually based on Sanger sequencing in RNA sequencing technology, but it is proven that these methods are both expensive and imprecise. New established High Throughput Deep Sequencing methods (Next generation sequencing: NextGen Seq) for RNA sequencing like Illimuna Genome Analyzer overcome all these problems and result accurate and fast read sequences. Furthermore, there are some key advantages of using Next generation sequencing methods. Unlike microarray based method NextGen Seq makes possible to study Metagenomics. It is possible to sequence novel transcriptomics without mapping them into any reference genomes namely by De-nova sequencing. Also, genome variation studies (SNP) are also promising (Haas & Zody, 2010; Wang et al., 2010).

Characterization of miRNAs, siRNAs, small nucleolar RNAs, and long ncRNAs on a genomic degree is possible with high throughput sequencing methods. Small non-coding RNA sequencing (ncRNA-Seq) is a method of RNA-Seq specialized to only quantify and degree of small RNA content of transcriptome. Figure 6 displays the
pathway to explore small RNAs. Recently, ncRNA-Seq is very popular to establish small RNAs of respectively unknown organisms to publish and discover miRNA sequences.

Nevertheless, sequencing ncRNA content of an organism is only creating a lot of small RNA sequences need to be analyzed through bioinformatic tools. Often, the raw reads are a prior characterized by means of their length, and small length reads are removed in filtering. The output is further mapped into the reference genome, if available and only consensus sequences are handled. The second approach depends on the expression profiles. Time series are examined individually with their technical replicates to diminish possible bias (Collins, 2011). Furthermore, nowadays several methods are established to predict precursor and mature miRNAs from deep sequencing data. For example, miRDeep* (An et al., 2013) is an integrated miRNA identification tool, currently used for both known miRNA quantity determination from miRBase and novel miRNA prediction. miRNAfold (Tempel & Tahi, 2012), unlike miRDeep, is an ab-initio method specialized with a fast algorithm to predict potential pre-miRNA hairpin structures in genomic sequences. Moreover, miPred (Jiang et al., 2007) is a toolkit used also to separation of the real pre-miRNAs from pseudo ones based on RF (random forest) function and using their differences in hybrid feature in stem loop structure.

However, big amount of data is established and several biases in the sequencing of ncRNA-Seq are observed like some RNA levels are clearly enhanced and some is even not detectable. Bias describes systematic errors in the experiment and directs the truth. Therefore, whether while the experimental stages or during the bioinfortaical data generation bias should be avoided, or some analytical techniques need to be developed to uncover the bias (Raabe et al., 2014).
Figure 6. A general approach for using high-throughput sequencing data in search of small RNAs. Adaptor sequences are added into the pool of total RNA containing ncRNA sequences. Sequencing data will produce ncRNA sequences with adaptor sequence, and long RNA sequences without adaptor. When adaptor sequences are clipped off, ncRNA sequences are ready for further bioinformatic analysis. Adapted from Collins, 2011.
1.7 Plant infecting fungi

Kingdom Fungi including yeasts, molds, and mushrooms is member of eukaryotes. They differ from other organisms like plants and animals with their chitin composed cell wall. There are various types of fungi with different associations and occupations. Plant is a common wide habitat for fungus. A fungus can be saprophytic, pathogenic or mutualistic to its host plant. Saprophytic fungi do not feed on living cells instead they decompose dead materials. They are harmless and even valuable to the environment. Pathogenic fungi use the sources of its host plant and cause serious diseases. If fungi strictly depend on host sources, can be a specific nutrient only, to live this interaction is called obligate paracitsim. Mutualistic fungi includes likens allied with cyanobacterium and mycorrhiza associated with plant root (Michael T. Madigan, 2009). In this study, we will only focus on obligate parasitic fungi *Puccinia striiformis* f. sp. *tritici* and *Blumeria graminis* f. sp. *hordei*, and mutualistic living mycorrhiza *Piriformospora indica* which are infecting crops.

1.7.1 *Piriformospora indica*

*Piriformospora indica* (Pi) is a fungal root endophyte interacting with a broad range of plants in mutualistic symbiosis. Endophyte interactions are advantageous to organism, fungus and host. The beneficial effect to the host plant can be improvement of the nutrient supply or even progress of the resistance to pathogens (Varma et al., 1999). As a result of the investigations, it is found that Pi can also locate into dead roots of barley, and removes the dead cells helping the growth of its host (Deshmukh et al., 2006). Therefore, it is suggested that Pi can be used to cultivate plants to raise crop yield (Varma et al., 1999).

1.7.2 *Puccinia striiformis* f. sp. *tritici*

*Puccinia striiformis* f. sp. *tritici* (Pst) is a filamentous fungus lives as obligate biotrophic parasite. It causes stripe or yellow rust disease (Figure 8), which is a severe infection of wheat worldwide. Disease cultivation favored by cold regions while growing of the wheat occurs. Because of its early occurrence and easy spread, it can destroy 10% to 70% of the yield on risky circumstances. Primary host of Pst is wheat (*Triticum aestivum* L.), but it can affect some cultivated barleys (*Hordeum vulgare* L.) without any serious epidemic (Chen et al., 2000).
1.7.3  *Blumeria graminis f. sp. hordei*

Powdery mildew *Blumeria graminis* attracts the cereals wheat and barley. *Blumeria graminis f.sp. hordei* grows selectively on barley causing barley powdery mildew disease. Powdery mildew appears as white particles on leaves and stems (Figure 8). Powdery mildews are obligate biotrophs; their necessity to a host organism to grow is inevitable. *Blumeria graminis f. sp. hordei* is a well-studied organism since its destruction is worldwide and economic importance is significant (Bindschedler et al., 2009; Glawe, 2008).

![Figure 8: Powdery Mildew. Photos are taken by A. Çağlar Özketen, with Leica microscope systems.](image)
CHAPTER II

MATERIALS AND METHODS

2.1 Materials

Micro RNA sequences of human dataset downloaded from miRBase Database (Kozomara & Griffiths-Jones, 2011) and small RNA sequencing results of the organisms *Puccinia striiformis* f. sp. *tritici* (Pst), *Blumeria graminis* f. sp. *hordei* and *Piriformospora indica* (Pi) are used in thesis study. Moreover, online server psRNATarget’s online genome database (Dai & Zhao, 2011) was used for the search of plant targets of the miRNAs.

Clustering algorithms were applied by their corresponding R packages or R codes. The bioinformatics tools were used from their online servers. For batch jobs, self-written Perl scripts were also established. Perl scripts for various purposes are inserted into Appendix E and Appendix F includes R codes written for clustering.

2.1.1 Data: Human miRNA sequences

Current Tool for Annotations of miRNA (TAM) miRNA catalogue, 238 miRNA sets with 413 distinct miRNAs, is downloaded. miRNA names were not specialized with their 3’ or 5’ overhangs. Therefore, miRNA names are matched with their corresponding sequences in miRBase tool. When both overhangs were stored, in total 666 miRNA sequences were utilized. These Human miRNA sequences are presented in Supplementary 2 data folder as in fasta format.

2.1.2 Data: Fungal small RNA sequences

Total small RNA sequencing of the fungi, *Puccinia striiformis* f. sp. *tritici* (Pst), *Blumeria graminis* f. sp. *hordei* (Bgh) and *Piriformospora indica* (Pi), were performed by an Illumina Genome Analyzer (BGI, Shezhan, China). Clean small RNA
sequences with advanced bioinformatic analyses like statistical analyses are provided by the company. Methodology of their sequencing technology and their strategy, and small RNA annotations will be discussed in the method part. These miRNA sequences are presented in Supplementary 1 data folder.

2.2 Methods

2.2.1 Small RNA sequencing and sequence analysis

RNA isolation of the fungi (Pst, Bgh, and Pi) were done following the manufacturer's protocol with addition of PVP-T in the grounding step (Trizol-invitrogen). Total RNA isolates of fungi were sent for sequencing. Before, the analysis BGI ensures the quality of the RNAs. Deep sequencing analysis was done by Illumina genome Analyzer (Illumina, SanDiego, CA, USA) at the Beijing Genomics Institute (BGI, Shenzhen, China). The total RNA isolates was size-fractionated and small RNA sequences (10-30 nt) was isolated. 5' and 3' adaptors were ligated to sRNAs and RT-PCR were applied to acquire double stranded DNA ready for sequencing. PCR products are sequenced by high technology Solexa sequencing strategy, and then, low quality tags, primers and adaptor sequences were removed. After reading the sequence, adaptor sequences; 5' and 3' primers, were removed. The final clean reads were mapped into corresponding genomes described in Table 1 by Short Oligonucleotide Alignment Program (SOAP) (Li et al., 2008). To annotate small RNAs into their types like rRNA, tRNA, snRNA, snoRNA, their deposite at NCBI genbank database (http://www.ncbi.nlm.nih.gov/genbank/) and Rfam database(http://rfam.xfam.org/) were used. Small RNAs matching repeat deposited at hg18 database (http://hgdownload.cse.ucsc.edu/downloads.html) were identified as repeat-associated small RNAs. Small RNAs sequences were aligned into the genomic regions. Small RNAs were also aligned into miRBase database (http://www.mirbase.org/) for identification of homologous small RNAs into the known species.

<table>
<thead>
<tr>
<th>Fungi</th>
<th>Reference genome description</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Blumeria graminis f. sp. hordei</em></td>
<td>Blumeria graminis f. sp. hordei DH14, whole genome sequence, sequenced by Blumeria Genome Consortium (Bioproject : PRJNA28821)</td>
</tr>
<tr>
<td><em>Piriformospora indica</em></td>
<td>Piriformospora indica DSM 11827, whole genome sequence, (Bioproject: PRJEA76339) .Sequencing was performed by MWG (Germany)</td>
</tr>
<tr>
<td><em>Puccinia striiformis f. sp. tritici</em></td>
<td>2K41-Yr9 isolate, race PST-78, a Great Plains isolate from 2000 (Bioproject : PRJNA123765) sequenced by Broad Institute</td>
</tr>
</tbody>
</table>
Then, a miRNA prediction algorithm, MIREAP (BGI, n.d.), was applied into the sRNA sequences targeted to the genome. The algorithm is both used to identify known and novel miRNAs from a small RNA library constructed by deeply sequenced RNA fragments. miRNAs are predicted by calculation of minimal free energy (MFE) of their corresponding hairpin structure. These miRNAs will be assigned as novel miRNA. Remaining small RNAs are aligned to miRBase (Kozomara & Griffiths-Jones, 2011) and most similar ones were selected as best homologs (known miRNAs). These small RNAs are considered as candidates of milRNAs.

BGI small RNA reports for the three organisms are included in the Supplementary material 1 data file. For explanations and data organizations of the report see Appendix A.

2.2.2 Sequence representation

miRNA sequences are represented by using two different approaches. One is by using pairwise sequence alignment tools and the other method is presenting k-mer substring distributions of the miRNA sequences.

2.2.2.1 Pairwise sequence alignments

In our study we used pairwise sequence alignments to degree the distance between two miRNA sequences. The distance values as metric were stored into matrices. A sequence is composed of a set of vector elements, each of which denotes the similarity of current miRNA sequence with any other miRNA sequence in the repository. To test different methodologies, Smith-Waterman algorithm as local and Needleman-Wunsch algorithm as global alignment are both applied. All sequences in a list are pair wisely aligned to each other, and their alignment scores are stored into a symmetric all-to-all matrix (Section 3.1).

Alignment scores are the measure of how two sequences similar to each other. In the matrix, the nodes are demonstrating the similarity vectors with respect to edges are miRNA sequences (Similarity Matrix). However, to generate a matrix showing distance measures (Distance Matrix), the scores for a miRNA sequence aligned to other miRNAs is subtracted from the score produced from the self-alignment of that miRNA, basically, it is the maximum score a miRNA sequence can produce.

Negative scores are also possible in Needleman–Wunsch with respect to Smith-Waterman that creates only positive scores. Therefore, the similarity and dissimilarity (distance) matrices should not be thought as real representative graphical distances, instead, they are the metric values showing how two pairings are alike or distant.
When both algorithms were applied, scoring schemes were the same, scores were calculated according to $\text{Gap}=-1$, $\text{Mismatch}=-1$, and $\text{Match}=+1$ values. Perl scripts were used for these jobs. Input files for both scripts are FASTA files containing miRNA sequences. Perl scripts used to construct matrix with Needleman-Wunsch and Smith-Waterman are inserted into Appendix E.

Furthermore, to test the efficiency of the matrices on clustering algorithms randomly filled matrices were generated. For that purpose Perl rand function is used to assign each of the matrices. Depending on the length of miRNA sequence which is in between 20-22 nucleotides (Lagos-Quintana et al., 2001), each pairwise similarity is assigned. The random matrix cells are thus designed as to simulate Smith-Waterman algorithm score results. Ultimate aim of this analysis is to estimate the change factor in representing similarity information of miRNAs.

2.2.2.2  K-mer counting

We chose k as 3 for a 3-mer representation in the analysis. On a defined RNA alphabet (A, G, U, C), when k equals to 3, there is $4^3$, 64 distinct count values. The presence of 3 length substrings (like AGU, CAA, GAU...) can be controlled and their presence indications can be stated as 1 or unlikely situation can be 0. Consequently, number of miRNAs versus $4^k$ dimension matrix is filled by 1 and 0. By this method, sequence information becomes independent from nucleotide triplet order, and the sequences are not affected from each other. Perl code is written to construct the matrix filled in by 3-mer features.

Perl Script written to calculate k-mers and construct a matrix from their precedence is inserted into Appendix E.

2.2.3  Clustering algorithms

Three vector based (K-means, CLAG, and SOTA) and one graph (MCL) based clustering algorithms are applied in this study.

2.2.3.1  Vecor based clustering algorithms

In vector based clustering algorithms each object should be represented with a vector whose components are a set of features. Therefore, demonstration of a miRNA becomes the vector showing the set (n number) of scores (S) from the results of its pairwise sequence similarity alignments or the set of numbers (1 or 0) showing $4^k$ number (N) of k-mer distributions for different test of trials. Figure 9 shows the vector representation of a miRNA sequence which is the input of the algorithms.
K-means algorithm is applied into our study by using the R functions defined in R documentations. The code established is inserted into Appendix F. In K-means analysis, k cluster number was always choosen as 30 and covering the whole data. However, as we know that at each run K-means choose different centroids leading to different groupings. To overcome different partition problem, re-run results were generated for the same input of cluster number 30. After memberships of the clusters defined, each trial was compared to each other in order to detect most stable groups. Therefore, some extension was made through the clusters, and less stable groups divided if their membership is not convincing enough for other trials or if the member is unstable for an affiliation it is eluted from the set.

For CLAG analysis, we directly used the CLAG package and manuels announced by the developers. R functions and the code used is inserted to the Appendix F. CLAG parameters used in the study are arranged by the behavior of the dataset and optimized for close number of clusters to 30.

Sota R function in clValid library is used as described in R documentation in our analysis, and R codes are included in to Appendix F. Cluster number is settled as 30 for all of the analysis.

### 2.2.3.2 Graph based clustering algorithm

As a graph based clustering algorithm, Markov clustering (MCL) is applied in our study. MCL analyses are done through R codes, adapted from original MCL codes written by Sylvain Brohee and inserted into Appendix F.

Figure 10 represents how a graph object can be translated into a matrix. Conversely, a symmetric matric representing pairwise relations between objects (the objects are miRNA sequences in our study) can also be figured into a graph. If all of the objects (N number) are connected to each other, graph will produce N number of nodes, and N*(N-1)/2 number of edges. It complements with a complete graph. However, since in an all-versus-all type of a matric all pairs are shown, as the number of objects increases, the compactness of the graph also increases. This situation will lead to
increase in noise, and unwanted connections through the clusters as the algorithm implements. In that direction, I proposed to use second or fourth derivatives of the scores produced from Smith-Waterman algorithm to mark the distances of miRNA sequences. Smith-Waterman algorithm is chosen since it only produces positive results, and derivative of the matrices are taken to a prior weaken the most unconnected edges while strengthen the powerful connections. Furthermore, in the analysis, different inflammation values are applied as the optimization of the dataset handled.

Figure 10. Transition of a graph to a matrix by MCL algorithm. Example of a distance graph for four objects (A-D), circles represents nodes (objects) and lines represent distances. Weighted transition matrix is shown generated from the graph.

2.2.4 Statistical analysis of the clusters

After cluster algorithms are applied into the matrixes, the resulting groups are tested according to their Dunn Index (DI). DI is the ratio of the smallest distance between the observations in the different clusters to the largest distance of the observations in the same cluster. DI metric aims to signify how compact and well separated the clusters is. The value of DI is 0 when all of the objects are in the same cluster and infinite when all the objects present for a cluster. To get a better result, DI needs to be maximized. The distance metric in DI can be classical Euclidian or Manhattan distance. There are ready applications of DI in MATLAB or R languages (Dunn, 1973; Handl, Knowles, & Kell, 2005). R code is shown in Figure 11.

Dunn Index calculation is carried out through R codes established in clValid library of R packages.
data(mouse)
express <- mouse[1:25,c("M1","M2","M3","NC1","NC2","NC3")]
rownames(express) <- mouse$ID[1:25]
## hierarchical clustering
Dist <- dist(express,method="euclidean")
clusterObj <- hclust(Dist, method="average")
nc <- 2  ## number of clusters
cluster <- cutree(clusterObj,nc)
dunn(Dist, cluster)

Figure 11. Example R code to calculate Dunn Index. Mouse data is used for hierarchical clustering; average distance is calculated by Euclidean and final Dunn value calculated with cluster assignments and Dist (distance) object. Figure is directly taken from R documentation of the function Dunn package clValid http://artax.karlin.mff.cuni.cz/r-help/library/clValid/html/dunn.html.

2.2.5 Qualification of the efficiency of the clusters

In order to test the efficiency of the clusters, Tool for annotations of miRNA (TAM) is used to assign human miRNA clusters into categories. TAM (Lu et al., 2010) is developed in order to accurately allocate eloquent human miRNAs into categories. There are five kinds of assignments in the tool; miRNA family, miRNA function, miRNA cluster, miRNA associated disease and miRNA tissue specificity. Family and Cluster clarifications are based to miRBase (Kozomara & Griffiths-Jones, 2011) classes, Human MicroRNA Disease Database (HMDD) (Lu et al., 2008) is used for disease specific associations, and function and specific tissue relations is collected from literature.

One of the advantages of the tool is that it can allocate a novel miRNA sequence into a category with the enrichment analysis. For an agreed set of miRNA, the tool estimates the significance (p value) of enrichment of these miRNAs in the given categories. P value is calculated in a correspondence with the size of the given set of miRNA and size of the dataset. Therefore, percentage of how many given miRNAs are in the consistent cluster and the significance of it are outputs of the tool (Lu et al., 2010). Web accession to TAM tool is http://202.38.126.151/hmdd/tools/tam.html/.

As mentioned before, TAM calculates the significance, p value, of enrichment for a given list of miRNA for each category of TAM. Each grouping for each clustering methods are given to TAM tool and results are saved as text files. Figure 12 represents an example output of the tool. In our analysis p-value and percentage coverage is used. For each clustering method, number of clusters validated to be enriched in each category of TAM is calculated. Then, percentage value of enrichment is established. In our study, to show functional relations in our clusters
we used only the hits of more than two miRNA sequences with p value becomes bigger than 0.005, and percentage coverage more than twenty percent.

<table>
<thead>
<tr>
<th>Term</th>
<th>Count</th>
<th>Percent</th>
<th>Fold</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsa-mir-106b cluster</td>
<td>2</td>
<td>0.67</td>
<td>33.296</td>
<td>1.06e-3</td>
</tr>
<tr>
<td>hsa-mir-17 cluster</td>
<td>2</td>
<td>0.33</td>
<td>18.6298</td>
<td>5.15e-3</td>
</tr>
<tr>
<td>mir-17 family</td>
<td>2</td>
<td>0.25</td>
<td>12.4722</td>
<td>9.41e-3</td>
</tr>
<tr>
<td>mir-15 family</td>
<td>1</td>
<td>0.5</td>
<td>24.5444</td>
<td>0.0397</td>
</tr>
<tr>
<td>mir-26 family</td>
<td>1</td>
<td>0.33</td>
<td>16.6306</td>
<td>0.0591</td>
</tr>
<tr>
<td>Akt pathway</td>
<td>2</td>
<td>0.12</td>
<td>5.8833</td>
<td>0.0416</td>
</tr>
<tr>
<td>Apoptosis</td>
<td>2</td>
<td>0.02</td>
<td>1.9574</td>
<td>0.0676</td>
</tr>
<tr>
<td>Bone regeneration</td>
<td>2</td>
<td>0.05</td>
<td>2.2077</td>
<td>0.2174</td>
</tr>
<tr>
<td>Cell cycle related</td>
<td>3</td>
<td>0.05</td>
<td>2.2077</td>
<td>0.1325</td>
</tr>
<tr>
<td>Cell proliferation</td>
<td>1</td>
<td>0.04</td>
<td>1.7017</td>
<td>0.4428</td>
</tr>
<tr>
<td>Chemoresponsivity of tumor cells</td>
<td>1</td>
<td>0.25</td>
<td>12.4722</td>
<td>0.0761</td>
</tr>
<tr>
<td>HIV latency</td>
<td>4</td>
<td>0.19</td>
<td>9.0226</td>
<td>3.68e-4</td>
</tr>
<tr>
<td>Hormones regulation</td>
<td>5</td>
<td>0.08</td>
<td>4.0235</td>
<td>3.49e-3</td>
</tr>
</tbody>
</table>

**Figure 12.** TAM output with an example dataset. Count number shows how many given miRNAs are enriched in that cluster, p-value is the significance of that enrichment, and Percent is the percentage of enriched miRNA number in that cluster. Figure is adapted from TAM web site.

TAM analyses are done through its online server. Results for each grouping are saved as text documents, and the enrichment analyses are done through a self-written Perl script which is shown in Appendix E.

### 2.2.6 miRNA target prediction

Plant miRNAs matches perfectly to their complementarity sites that make the difference between animal based target prediction sites and plant miRNA target prediction tools. psRNATarget tool (Dai & Zhao, 2011) is developed to find targets locations of plant small RNAs. Since, our one hypothesis is to test whether fungal miRNAs are targeting plant genomes; we have decided to use this server to find targets of miRNAs of the fungi.
2.2.6.1  psRNATarget tool

psRNATarget is a web application to find possible target locations in plant genomes of a given small RNA sequence. The algorithm of the tool uses a dynamic programming approach, a modified version of Smith-Waterman algorithm. Basically, it scans the targets base pairing to miRNA sequences. The tool has a very wide database containing plant genomes. The current records of genome releases is listed in Appendix G. PsRNATarget is a valuable tool since it is only designed for plant-miRNA relations, taking into account that mammalian miRNA base pairing is different than plants (B. Bartel & Bartel, 2003; Pratt & MacRae, 2009). The tool is highly user friendly with high throughput small sequencing data with user friendly result pages. Figure 13 shows an example to its result page. Web accession to psRNATarget is http://plantgrn.noble.org/psRNATarget/.

The tool has two important functions. First one is reverse complementary matching between small RNA and the target site is calculated, and the second is unpaired energy (UPE) that is needed to open hairpin structure. Thresholds of the tool can be viewed from Table 2. In our analysis the analysis carried by the parameters specified in the table. The libraries selected per organism through our study represented in Table 3.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Thresholds</th>
<th>Explanations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum expectation</td>
<td>3.0</td>
<td>Threshold of the score for complementing miRNA and its target. Score is calculated by algorithm of miRU.</td>
</tr>
<tr>
<td>Length for complementarity scoring (hspsize)</td>
<td>20.0</td>
<td>Limit for the complementary region, it should be taken by maximum length of miRNA sequence.</td>
</tr>
<tr>
<td>Target accessibility (UPE)</td>
<td>25.0</td>
<td>It is the maximum energy (delta G) to open hairpin structure of the miRNA. Less energy means more possibility to target the region. RNAup algorithm is used for calculation.</td>
</tr>
<tr>
<td>Flanking Length around target site for target accessibility analysis</td>
<td>17-13 bp</td>
<td>Flanking regions are need for RISC complex, miRNA to target pairing.</td>
</tr>
<tr>
<td>Range of central mismatch leading to translational inhibition</td>
<td>9-11 nt</td>
<td>It is important for decision making when miRNA cleaves target site or inhibits the translation of target gene.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fungi</th>
<th>Target Organism; Libraries selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bgh</td>
<td><em>Triticum aestivum</em> (wheat) by DFCI Gene Index (TAGI)</td>
</tr>
<tr>
<td></td>
<td><em>Hordeum vulgare</em> (barley) by DFCI Gene Index (HVGI)</td>
</tr>
<tr>
<td>Pst</td>
<td><em>Triticum aestivum</em> (wheat) by DFCI Gene Index (TAGI)</td>
</tr>
<tr>
<td></td>
<td><em>Hordeum vulgare</em> (barley) by DFCI Gene Index (HVGI)</td>
</tr>
<tr>
<td>Pi</td>
<td>All of the library except Human dataset*</td>
</tr>
</tbody>
</table>

*psRNATarget library list is presented in Appendix G
2.2.6.2 miRBase database

MiRBase is the largest database for the glossary of published miRNA sequences. Currently, there are thousands of miRNA entries from various organisms. The database also has portable search options and links to other databases like the Pubmed (Kozomara & Griffiths-Jones, 2011).

2.2.7 Experimental analysis of the fungal miRNAs by the pipeline

Fungal (*Puccinia striiformis* f. sp. *tritici* (Pst), *Blumeria graminis* f. sp. *hordei* (Bgh) and *Piriformospora indica* (Pi)) miRNA predictions and small RNAs homologous to miRBAse are combined individually, and analyzed through the workflow created by this study.
CHAPTER III

RESULTS

3.1 Outline

In materials part, the data used in the study was briefly summarized. In sections 2.2.2 to 2.2.5 representation miRNA sequences, cluster application methods and their statistical and functional analysis methods was shown. In section 2.2.1 small RNA sequencing of the fungi and the sequence analysis methods were described. Section 2.2.6 display miRNA candidate prediction and target analysis methods and section 2.2.7 shows how the application of the method was applied into miRNA sequences of the pathogens analysed in our laboratory. In results part, between the sections 3.1 and 3.3 small RNA identification reports, novel miRNA predictions, and nucleotide bias analysis will be shown and miRNA prediction and target analysis results will be shown between sections 3.9 and 3.11. Section 3.2 example matrices created by different sequence representation method will be presented.

3.2 Identification of small RNAs

Deep high throughput sequencing resulted with in total between 11 million to 33 million reads for Pst, Bgh and Pi shown in Table 4. From total reads small reads lower than 18 nt, 3’ and 5’ adaptors, poly A sequences, and null low quality reads are filtered out. The remains were clean reads ready for annotation. From those sequences between 2 million to 5 million were the unique clean sequences.

<table>
<thead>
<tr>
<th>Fungi</th>
<th>Total Reads</th>
<th>&lt; 18 nt*</th>
<th>3’ adaptor</th>
<th>5’ adaptor</th>
<th>Clean Reads</th>
<th>Unique clean reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bgh</td>
<td>26317125</td>
<td>349950</td>
<td>15002</td>
<td>16424</td>
<td>25767253</td>
<td>2539594</td>
</tr>
<tr>
<td>Pst</td>
<td>32490108</td>
<td>205066</td>
<td>81350</td>
<td>242498</td>
<td>31802410</td>
<td>4863345</td>
</tr>
<tr>
<td>Pi</td>
<td>11744418</td>
<td>289475</td>
<td>19992</td>
<td>19548</td>
<td>11299452</td>
<td>2095999</td>
</tr>
</tbody>
</table>

*Sequence reads smaller than 18 nucleotides
The length distributions of the unique clean reads for the fungi are shown in Figure 14. For each organism's length distribution is calculated by the number of reads with respect to length of small RNAs. Distributions are often in between 13 to 32 nucleotides. miRNA distribution can be inferred from the nucleotides between 20 to 24. Length of the small RNA is crucial to decide its annotation type.

**Figure 14.** Graph demonstrating length distribution of read small RNAs of A) Bgh, B) Pst, and C) Pi.
The clean unique reads, sRNAs, were mapped to the known miRNAs, rRNAs, tRNAs, snRNAs, snoRNAs, repeat associated RNAs, introns and exons. Number of unique small RNAs that were annotated to the known sRNAs is represented in Table 5. Figure 15 is illustrating pie charts for annotations of small RNAs for the organisms. Small RNA sequences are included in Supplumentary file 1.

<table>
<thead>
<tr>
<th>Annotation</th>
<th>Bgh</th>
<th>Pi</th>
<th>Pst</th>
</tr>
</thead>
<tbody>
<tr>
<td>exon_antisense</td>
<td>686284</td>
<td>23829</td>
<td>494490</td>
</tr>
<tr>
<td>exon_sense</td>
<td>259388</td>
<td>837215</td>
<td>456115</td>
</tr>
<tr>
<td>intron_antisense</td>
<td>59169</td>
<td>1722</td>
<td>142719</td>
</tr>
<tr>
<td>intron_sense</td>
<td>35723</td>
<td>10869</td>
<td>138276</td>
</tr>
<tr>
<td>miRNA</td>
<td>108534</td>
<td>55336</td>
<td>222494</td>
</tr>
<tr>
<td>rRNA</td>
<td>75152</td>
<td>152198</td>
<td>26180</td>
</tr>
<tr>
<td>repeat</td>
<td>15952</td>
<td>4476</td>
<td>2269</td>
</tr>
<tr>
<td>scRNA</td>
<td>0</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>snRNA</td>
<td>2041</td>
<td>4650</td>
<td>849</td>
</tr>
<tr>
<td>snoRNA</td>
<td>491</td>
<td>482</td>
<td>2855</td>
</tr>
<tr>
<td>tRNA</td>
<td>8255</td>
<td>2971</td>
<td>324164</td>
</tr>
<tr>
<td>unannotated</td>
<td>1288605</td>
<td>1002238</td>
<td>3809675</td>
</tr>
</tbody>
</table>

Figure 15. Pie charts representing small RNA annotations for Bgh (B. graminis), Pst (P. striiformis), and Pi (P. indica).
3.3 Novel miRNA predictions

Whole genome sequences, defined in corresponding method part, of the fungi were used for novel miRNA predictions. MIREAP algorithm is used for hairpin predictions. 2,588 Bgh, 491 Pst, and 61 Pi miRNA precursors were identified. Among the predicted hairpins, some of them were only producing from only 3’ overhangs or 5’ overhangs while some producing from both overhangs (Table 6). It is known that the precursors expressing from both overhangs are more likely real miRNA hairpins and their one overhang always express more than the other.

<table>
<thead>
<tr>
<th>Fungi</th>
<th>miRNA precursors</th>
<th>5’ *</th>
<th>3’ **</th>
<th>5’ and 3’</th>
<th>MFE range (kcal/mol)</th>
<th>Known miRNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bgh</td>
<td>2588</td>
<td>1565</td>
<td>1609</td>
<td>586</td>
<td>-18 to -42.2</td>
<td>3219</td>
</tr>
<tr>
<td>Pst</td>
<td>491</td>
<td>311</td>
<td>256</td>
<td>76</td>
<td>-18.1 to -215.5</td>
<td>7737</td>
</tr>
<tr>
<td>Pi</td>
<td>61</td>
<td>38</td>
<td>36</td>
<td>13</td>
<td>-21.8 to -149.12</td>
<td>2169</td>
</tr>
</tbody>
</table>

*3’ overhang miRNA, **5’ overhang miRNA

3.4 Nucleotide bias analysis of novel miRNAs

The nucleotide bias at the first position of predicted novel miRNAs with a definite length is analyzed (Figure 16). In all of the organisms, the miRNAs shows a dominant bias to uracil (U) at the first nucleotide. The typical characteristic that miRNAs usually begin with a U at the 5’ terminus approves with that observation.

The percentage of the four nucleotides appealing at each position is also analyzed (Figure 17). In general, five positions, 1st, 6th, 8th, 16th, and 18th nucleotides, display the most leading bias to U. The region collapsing the 2nd to the 8th nucleotides of miRNAs is called as the seed region, and pair perfectly with their target sites. Thus, the 1st and 8th nucleotides are at the boundaries of the seed region, 6th nucleotide is within the seed region, and the 16th and 18th nucleotides are near the 5’ terminus of the miRNAs. The bias to U at these positions may contribute to the miRNA regulation (Zhang, Stellwag, & Pan, 2009).
Figure 16. Graph representing first nucleotide bias for novel miRNA predictions of A) Bgh, B) Pst, and C) Pi.
Figure 17. Graph for position base bias of novel miRNA predictions of A) Bgh, B) Pst, and C) Pi.

3.5 Matrices

In materials part, construction of the matrices is briefly summarized. Tables (7, 8, 9, 10, 11, 12, 13) illustrate the matrices created by different algorithms or different approaches on human miRNA set. Only six of the miRNA is figured as an example.
### Table 7. MCL matrix by Smith-Waterman algorithm.

<table>
<thead>
<tr>
<th>matrix</th>
<th>miR-153-3p</th>
<th>miR-502-3p</th>
<th>miR-519c-3p</th>
<th>miR-21-3p</th>
<th>miR-216b-3p</th>
<th>miR-520g-5p</th>
</tr>
</thead>
<tbody>
<tr>
<td>miR-153-3p</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>miR-502-3p</td>
<td>4</td>
<td>0</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>miR-519c-3p</td>
<td>3</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>miR-21-3p</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>miR-216b-3p</td>
<td>2</td>
<td>6</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>miR-520g-5p</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

### Table 8. MCL matrix by Smith-Waterman algorithm.

<table>
<thead>
<tr>
<th>matrix</th>
<th>miR-153-3p</th>
<th>miR-502-3p</th>
<th>miR-519c-3p</th>
<th>miR-21-3p</th>
<th>miR-216b-3p</th>
<th>miR-520g-5p</th>
</tr>
</thead>
<tbody>
<tr>
<td>miR-153-3p</td>
<td>22</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>miR-502-3p</td>
<td>4</td>
<td>22</td>
<td>4</td>
<td>1</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>miR-519c-3p</td>
<td>3</td>
<td>4</td>
<td>22</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>miR-21-3p</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>21</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>miR-216b-3p</td>
<td>2</td>
<td>6</td>
<td>2</td>
<td>3</td>
<td>24</td>
<td>1</td>
</tr>
<tr>
<td>miR-520g-5p</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>23</td>
</tr>
</tbody>
</table>

### Table 9. Distance matrix by Smith-Waterman algorithm (SW-Distance).

<table>
<thead>
<tr>
<th>matrix</th>
<th>miR-153-3p</th>
<th>miR-502-3p</th>
<th>miR-519c-3p</th>
<th>miR-21-3p</th>
<th>miR-216b-3p</th>
<th>miR-520g-5p</th>
</tr>
</thead>
<tbody>
<tr>
<td>miR-153-3p</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>miR-502-3p</td>
<td>18</td>
<td>0</td>
<td>18</td>
<td>21</td>
<td>16</td>
<td>21</td>
</tr>
<tr>
<td>miR-519c-3p</td>
<td>19</td>
<td>18</td>
<td>0</td>
<td>19</td>
<td>20</td>
<td>19</td>
</tr>
<tr>
<td>miR-21-3p</td>
<td>20</td>
<td>20</td>
<td>0</td>
<td>18</td>
<td>19</td>
<td>23</td>
</tr>
<tr>
<td>miR-216b-3p</td>
<td>22</td>
<td>22</td>
<td>20</td>
<td>21</td>
<td>22</td>
<td>0</td>
</tr>
<tr>
<td>miR-520g-5p</td>
<td>22</td>
<td>22</td>
<td>20</td>
<td>21</td>
<td>22</td>
<td>0</td>
</tr>
</tbody>
</table>

### Table 10. Similarity matrix by Needleman-Wunsch algorithm (NW-Similarity).

<table>
<thead>
<tr>
<th>matrix</th>
<th>miR-153-3p</th>
<th>miR-502-3p</th>
<th>miR-519c-3p</th>
<th>miR-21-3p</th>
<th>miR-216b-3p</th>
<th>miR-520g-5p</th>
</tr>
</thead>
<tbody>
<tr>
<td>miR-153-3p</td>
<td>22</td>
<td>1</td>
<td>-1</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
</tr>
<tr>
<td>miR-502-3p</td>
<td>1</td>
<td>22</td>
<td>3</td>
<td>0</td>
<td>4</td>
<td>-4</td>
</tr>
<tr>
<td>miR-519c-3p</td>
<td>-1</td>
<td>3</td>
<td>22</td>
<td>-3</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>miR-21-3p</td>
<td>-2</td>
<td>0</td>
<td>-3</td>
<td>21</td>
<td>2</td>
<td>-2</td>
</tr>
<tr>
<td>miR-216b-3p</td>
<td>-2</td>
<td>4</td>
<td>-1</td>
<td>2</td>
<td>24</td>
<td>-4</td>
</tr>
<tr>
<td>miR-520g-5p</td>
<td>-2</td>
<td>-4</td>
<td>-1</td>
<td>-2</td>
<td>-4</td>
<td>23</td>
</tr>
</tbody>
</table>

### Table 11. Distance matrix by Needleman-Wunsch algorithm (NW-Distance).

<table>
<thead>
<tr>
<th>matrix</th>
<th>miR-153-3p</th>
<th>miR-502-3p</th>
<th>miR-519c-3p</th>
<th>miR-21-3p</th>
<th>miR-216b-3p</th>
<th>miR-520g-5p</th>
</tr>
</thead>
<tbody>
<tr>
<td>miR-153-3p</td>
<td>0</td>
<td>23</td>
<td>22</td>
<td>24</td>
<td>24</td>
<td>24</td>
</tr>
<tr>
<td>miR-502-3p</td>
<td>21</td>
<td>0</td>
<td>19</td>
<td>22</td>
<td>18</td>
<td>26</td>
</tr>
<tr>
<td>miR-519c-3p</td>
<td>23</td>
<td>19</td>
<td>0</td>
<td>25</td>
<td>23</td>
<td>23</td>
</tr>
<tr>
<td>miR-21-3p</td>
<td>23</td>
<td>21</td>
<td>24</td>
<td>0</td>
<td>19</td>
<td>23</td>
</tr>
<tr>
<td>miR-216b-3p</td>
<td>26</td>
<td>20</td>
<td>25</td>
<td>22</td>
<td>0</td>
<td>28</td>
</tr>
<tr>
<td>miR-520g-5p</td>
<td>25</td>
<td>27</td>
<td>24</td>
<td>25</td>
<td>27</td>
<td>0</td>
</tr>
</tbody>
</table>

### Table 12. Randomly Filled matrix.

<table>
<thead>
<tr>
<th>matrix</th>
<th>miR-153-3p</th>
<th>miR-502-3p</th>
<th>miR-519c-3p</th>
<th>miR-21-3p</th>
<th>miR-216b-3p</th>
<th>miR-520g-5p</th>
</tr>
</thead>
<tbody>
<tr>
<td>miR-153-3p</td>
<td>0</td>
<td>21</td>
<td>13</td>
<td>24</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>miR-502-3p</td>
<td>6</td>
<td>0</td>
<td>18</td>
<td>1</td>
<td>17</td>
<td>4</td>
</tr>
<tr>
<td>miR-519c-3p</td>
<td>19</td>
<td>17</td>
<td>0</td>
<td>12</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>miR-21-3p</td>
<td>18</td>
<td>22</td>
<td>8</td>
<td>0</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>miR-216b-3p</td>
<td>15</td>
<td>14</td>
<td>4</td>
<td>10</td>
<td>0</td>
<td>15</td>
</tr>
<tr>
<td>miR-520g-5p</td>
<td>25</td>
<td>24</td>
<td>5</td>
<td>19</td>
<td>13</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 13. 3-mer distribution matrix (K-mer)

<table>
<thead>
<tr>
<th>matrix</th>
<th>AAA</th>
<th>AAU</th>
<th>AAC</th>
<th>AAG</th>
<th>AUA</th>
<th>AUG</th>
<th>ACA</th>
</tr>
</thead>
<tbody>
<tr>
<td>miR-153-3p</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>miR-502-3p</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>miR-519c-3p</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>miR-21-3p</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>miR-216b-3p</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>miR-520g-5p</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

3.6 Cluster algorithms

Table 14 represents cluster numbers and data coverages for different matrices clustered with altered algorithms. Cluster number is the number of grouping made for human miRNA dataset, and data coverage is the percentage of the miRNA sequences included in the clusters. In a structural manner classical K-means algorithm clusters the whole dataset. However, district objects are required to be removed. Therefore, re-runs for K-means algorithm is arranged, so high coverage of the data is seen for K-means algorithm. After several arrangements by DI calculation, cluster number is optimally found as 43. Random matrix results with 47 clusters and 100% coverage.

CLAG algorithm has the smallest data coverage among the other stated methods. CLAG algorithm tends to find condense regions, most strength clusters. Thus, cluster numbers created and data coverage are very small, between 9% to 18% coverage. Cluster numbers and data coverage for Random matrix is different than the real matrices. Since randomly assignment of numbers generates highly condensed regions which CLAG cannot directly cluster the data. SOTA algorithm clusters the whole dataset. Initial cluster number is given by the user, and settled as 30.

MCL algorithm has a different methodology than other algorithms since it is a graphical clustering method. As described in methods part, as the objects gets distant each other clustered number of data changes. Inflammation value also has effects on data coverage. At least 15 number of clusters with 86% is found for the matrix powered by 2 and inflamed by 4, and the most 56 number of clusters with 73% is found for the matrix powered by four and inflamed by 2. There is no big change between Similarity, Distance and K-mer matrices by cluster numbers and data coverages for all clustering methods.
Table 14. Cluster numbers and data coverages of groupings by different methods.

<table>
<thead>
<tr>
<th>Matrix*</th>
<th>Cluster Number</th>
<th>Data Coverage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-means</td>
<td>K-mer</td>
<td>47</td>
</tr>
<tr>
<td></td>
<td>NW-Similarity</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>NW-Distance</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>SW-Similarity</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>SW-Distance</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>Random Matrix</td>
<td>47</td>
</tr>
<tr>
<td>CLAG</td>
<td>K-mer</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>NW-Similarity</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>NW-Distance</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>SW-Similarity</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>SW-Distance</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>Random Matrix</td>
<td>104</td>
</tr>
<tr>
<td>SOTA</td>
<td>K-mer</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>NW-Similarity</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>NW-Distance</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>SW-Similarity</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>SW-Distance</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Random Matrix</td>
<td>30</td>
</tr>
<tr>
<td>MCL</td>
<td>A</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>E</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>46</td>
</tr>
</tbody>
</table>

*A, B and C are the 2nd; D, E, and F are the 4th power of the original Smith Waterman applied MCL matrix. 4, 5.6,2,3, and 4 inflation values are applied into the matrices respectively A, B, C, D, E and F.

3.7 Dunn Indeces

Statistical validation of the clusters is done with Dunn Index (DI) calculations. In short, DI indicates how well the clusters are separated depending on the cluster and object numbers.

Table 15. Dunn Indeces of the groups for cluster algorithms applied.

<table>
<thead>
<tr>
<th>Matrix</th>
<th>Dunn index</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-means</td>
<td>0.3511</td>
</tr>
<tr>
<td>NW-Similarity</td>
<td>0.2641</td>
</tr>
<tr>
<td>NW-Distance</td>
<td>0.2297</td>
</tr>
<tr>
<td>SW-Similarity</td>
<td>0.4539</td>
</tr>
<tr>
<td>SW-Distance</td>
<td>0.3507</td>
</tr>
<tr>
<td>Random Matrix</td>
<td>0.7920</td>
</tr>
<tr>
<td>CLAG</td>
<td>K-mer</td>
</tr>
<tr>
<td>NW-Similarity</td>
<td>0.6257</td>
</tr>
<tr>
<td>NW-Distance</td>
<td>0.4498</td>
</tr>
<tr>
<td>SW-Similarity</td>
<td>0.4867</td>
</tr>
<tr>
<td>SW-Distance</td>
<td>0.4789</td>
</tr>
<tr>
<td>Random Matrix</td>
<td>0.8311</td>
</tr>
<tr>
<td>SOTA</td>
<td>K-mer</td>
</tr>
<tr>
<td>NW-Similarity</td>
<td>0.2430</td>
</tr>
<tr>
<td>NW-Distance</td>
<td>0.2043</td>
</tr>
<tr>
<td>SW-Similarity</td>
<td>0.2766</td>
</tr>
<tr>
<td>SW-Distance</td>
<td>0.2369</td>
</tr>
<tr>
<td>Random Matrix</td>
<td>0.8396</td>
</tr>
</tbody>
</table>
Table 15 shows DI values for each clustering approach. DI is not calculated with MCL algorithm since MCL is graph clustering method, similarity or dissimilarity metric between objects can not indicate the real distance values.

When DI results are analysed, it is observed that Random matrices are well grouped than real matrices. The reason underlying the situation is that Random matrix is filled unsystematically; however this makes this matrix so homogenise that clusters are found to be even, containing less noise.

In general, DI for CLAG is better than other methods. Unlike other algorithms, CLAG only objects into the condense regions, finds small number of clusters, removing nearly 90% of the data. Therefore, clusters are very strength. K-means algorithm can produce better clusters than SOTA for some methods. SOTA DIs are low since SOTA tends to cluster whole data unlike altered K-means algorithm. Therefore, SOTA parcels the data without rendering out the separated objects.

### 3.8 TAM enrichment analysis

Table 16 represents cluster percentages of TAM tool enrichment results. All five categories of TAM tool is shown; Clusters, Function, Family, HMDD and Tissue. The ALL category was added to show the percentages for the categories altogether. Calculation of the percentages and cut-off values are well explained in method part.

To show TAM analyses of the clusters and the cut-off values are significant, random samples were taken from the given set of miRNA and analyzed through TAM tool (10, 30, and 150 grouping) to set up a cut-off value. Depending on sampling values, miRNA enrichment changes expressively. 10 clusters, for example, tend to show increase in enrichment, while 150 clusters nearly do not give any enrichment results. Actually, it is about the size of the cluster, as size of a cluster increase, it is more likely it gets hit from TAM tool. However, our cluster analyses in this study mostly run by thirty numbers of clusters. So, 40% percentage for ALL result can be cut-off value to underline significance.

Two different similarity detection approaches, k-mer counting and pairwise similarity detection, were analyzed with TAM tool. Results show that clustering algorithms SOTA and k-means produce less percent of enrichment with k-mer method than pairwise similarity techniques. However, a general overview to the outputs of TAM tool signifies that there is no significant change by modification in similarity methods, at least 70% of clusters display enrichment. Between the categories of TAM, it is found out that the most enrich one is family categorie. This result is expected and it proves the initial hypothesis that the sequence similarity of miRNA sequences represents functional similarity. Cluster enrichments are respectively smaller since clusters of miRNAs are found by expression analysis and proximity in location. Furthermore, tissue analysis of TAM was not complete and it
includes slightly less information, which affects our analysis with decreasing enrichment results.

MCL method is applied only by Smith-Waterman distance matrix. Various optimizations are needed to make TAM enrichment analysis. Outputs show that when a prior data inflation is increased, more clusters are found but less functional annotation is possible through TAM. Best functional annotation for Clusters category (60%) is found by MCL algorithm with the matrix powered by 2 and inflated with 4. This matrix was the also less covered dataset, probably only found the best relations in the dataset.

Data coverage found to be also related with enrichment analysis. CLAG analysis represents that as data coverage decrease, more similarity can be found in miRNA sequences. Since only pairwise similarities are detected at least 75% of the clusters enriched in all categories and in families. K-means algorithm with respect to CLAG tends to cover whole data and show more enrichment. At least 80% of the clusters enriched in all categories. SOTA like k-means also show at least 80% enrichment, but in cluster category, K-means better than SOTA and any other cluster algorithms.

### Table 16. Enrichment results of the clusters calculated by TAM tool

<table>
<thead>
<tr>
<th>Matrix</th>
<th>Clusters</th>
<th>Function</th>
<th>Family</th>
<th>HMDD</th>
<th>Tissue</th>
<th>All*</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-means</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>K-mer</td>
<td>44.68</td>
<td>12.76</td>
<td>72.34</td>
<td>44.68</td>
<td>10.63</td>
</tr>
<tr>
<td></td>
<td>NW-Similarity</td>
<td>57.78</td>
<td>22.22</td>
<td>82.22</td>
<td>40.00</td>
<td>11.11</td>
</tr>
<tr>
<td></td>
<td>NW-Distance</td>
<td>52.17</td>
<td>19.56</td>
<td>76.09</td>
<td>45.65</td>
<td>6.52</td>
</tr>
<tr>
<td></td>
<td>SW-Similarity</td>
<td>63.16</td>
<td>23.68</td>
<td>78.94</td>
<td>55.26</td>
<td>15.79</td>
</tr>
<tr>
<td></td>
<td>SW-Distance</td>
<td>56.76</td>
<td>32.43</td>
<td>70.27</td>
<td>40.54</td>
<td>10.81</td>
</tr>
<tr>
<td></td>
<td>Random Matrix</td>
<td>8.51</td>
<td>6.38</td>
<td>8.51</td>
<td>14.89</td>
<td>4.25</td>
</tr>
<tr>
<td>CLAG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>K-mer</td>
<td>24.13</td>
<td>13.79</td>
<td>75.86</td>
<td>27.59</td>
<td>13.79</td>
</tr>
<tr>
<td></td>
<td>NW-Similarity</td>
<td>30.00</td>
<td>16.67</td>
<td>80.00</td>
<td>36.67</td>
<td>13.33</td>
</tr>
<tr>
<td></td>
<td>NW-Distance</td>
<td>22.58</td>
<td>16.13</td>
<td>77.42</td>
<td>41.94</td>
<td>9.68</td>
</tr>
<tr>
<td></td>
<td>SW-Similarity</td>
<td>20.00</td>
<td>14.00</td>
<td>70.00</td>
<td>24.00</td>
<td>10.00</td>
</tr>
<tr>
<td></td>
<td>SW-Distance</td>
<td>16.67</td>
<td>20.83</td>
<td>79.17</td>
<td>45.83</td>
<td>4.17</td>
</tr>
<tr>
<td></td>
<td>Random matrix</td>
<td>4.81</td>
<td>2.88</td>
<td>4.81</td>
<td>8.63</td>
<td>0</td>
</tr>
<tr>
<td>MCL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>60.00</td>
<td>26.67</td>
<td>86.66</td>
<td>33.33</td>
<td>13.33</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>38.89</td>
<td>22.22</td>
<td>66.67</td>
<td>33.33</td>
<td>16.67</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>41.17</td>
<td>23.53</td>
<td>58.82</td>
<td>29.41</td>
<td>17.65</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>32.14</td>
<td>14.29</td>
<td>51.79</td>
<td>25.00</td>
<td>7.14</td>
</tr>
<tr>
<td></td>
<td>E</td>
<td>39.13</td>
<td>13.04</td>
<td>47.03</td>
<td>21.74</td>
<td>8.70</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>39.13</td>
<td>15.22</td>
<td>47.83</td>
<td>21.74</td>
<td>10.87</td>
</tr>
<tr>
<td>SOTA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>K-mer</td>
<td>50.00</td>
<td>33.33</td>
<td>63.33</td>
<td>36.67</td>
<td>10.00</td>
</tr>
<tr>
<td></td>
<td>NW-Similarity</td>
<td>40.00</td>
<td>23.33</td>
<td>73.33</td>
<td>26.67</td>
<td>6.67</td>
</tr>
<tr>
<td></td>
<td>NW-Distance</td>
<td>36.67</td>
<td>13.33</td>
<td>70.00</td>
<td>50.00</td>
<td>6.67</td>
</tr>
<tr>
<td></td>
<td>SW-Similarity</td>
<td>50.00</td>
<td>30.00</td>
<td>70.00</td>
<td>43.33</td>
<td>10.00</td>
</tr>
<tr>
<td></td>
<td>SW-Distance</td>
<td>43.33</td>
<td>23.33</td>
<td>60.00</td>
<td>50.00</td>
<td>10.00</td>
</tr>
<tr>
<td></td>
<td>Random Matrix</td>
<td>10.00</td>
<td>10.00</td>
<td>10.00</td>
<td>20.00</td>
<td>3.33</td>
</tr>
<tr>
<td>Random Clusters</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10 groups</td>
<td>10.00</td>
<td>13.33</td>
<td>6.67</td>
<td>33.33</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>30 groups</td>
<td>14.44</td>
<td>10.00</td>
<td>11.11</td>
<td>22.22</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>150 groups</td>
<td>3.16</td>
<td>4.28</td>
<td>2.48</td>
<td>9.93</td>
<td>0.68</td>
</tr>
</tbody>
</table>

* Five categories of TAM tool is shown; clusters, function, family, HMDD, and Tissue. All represents the percentage result annotated by any of the categories at least one time. The results are given as percentage.

K-means algorithm is able to cluster whole data, with significant DI values. By using the classical K-means algorithm, in fact, it is possible to generate clusters 92% of them enriched in at least one of TAM categories, and also 82% of them significantly
enriched in family category. However, CLAG algorithm only projects into condense regions of the data, and found small major shrink clusters visualized by low data coverage with high DI value. Yet, it is proven that, these small clusters are well enriched in function (80% in ALL). Therefore, there is no harm to use CLAG a priori to cluster analysis to shape the centroids of the data. SOTA like K-means also clusters whole dataset with a given cluster number. The algorithm shows significant enrichment with TAM tool (83%). MCL algorithm with respect to other algorithms uses graph theory for grouping indeed able to generate well group of miRNAs separated from noise with 86% of enrichment in function.

3.9 Sequence similarity based miRNA clustering pipeline

Depending on the results of TAM analysis, the hypothesis stating that miRNA sequences with similar nucleotide content enriched in the same function is proven. To generalize the study, a pipeline illustrating the workflow is created. Figure 18 represents the flowchart of the pipeline.

![Figure 18. Workflow of miRNA sequence based clustering.](image)
3.10 Analysis of the miRNAs by the pipeline

Predicted miRNA sequences (Novel miRNAs) of Fungi, Bgh, Pst and Pi, are combined with known miRNAs which are homologs of small RNA sequences (Known miRNAs). The sets are questioned through the pipeline individually. Especially, different similarity detection ways and clustering algorithms are used. For example, Bgh study is carried by MCL algorithm by using distance measuring with Smith Waterman algorithm, Pst is examined by CLAG algorithm by Smith-Waterman similarity matrix construction, and Pi is analysed by SOTA algorithm by using k-mer counting approach. Final clusters are inserted into Supplementary Material 3 Data folder and Appendix B includes explanations of data organisation. Table 17 signifies the analysis results.

Table 17. miRNA cluster results.

<table>
<thead>
<tr>
<th>Fungus</th>
<th>Novel miRNA Count</th>
<th>Known miRNA Count</th>
<th>Total Input Number</th>
<th>Cluster Number</th>
<th>Data Coverage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bgh</td>
<td>2593</td>
<td>180</td>
<td>2773</td>
<td>25</td>
<td>76.5</td>
</tr>
<tr>
<td>Pst</td>
<td>492</td>
<td>301</td>
<td>793</td>
<td>2</td>
<td>0.9</td>
</tr>
<tr>
<td>PI</td>
<td>61</td>
<td>54</td>
<td>115</td>
<td>11</td>
<td>75.7</td>
</tr>
</tbody>
</table>

The final clusters presented are the only groupings enriched with known miRNA sequences. These groups are especially chosen to functionally annotate miRNA predictions of fungus. miRNA sequences in these clusters can be annotated by the known miRNAs’ functions.

3.11 miRNA selection for target gene analysis

In sequencing analysis bias is seen commonly. To remove the possible noise, from the catalogue of the predicted miRNA sequences, novel miRNA sets and small RNAs own homologue to miRBase, BGI provided, only the sequences read more than 500 times are selected a prior. Table 18 presents number of miRNAs selected for Target analysis by psRNATarget tool.

Table 18. Number of predicted miRNA by more than 500 count number.

<table>
<thead>
<tr>
<th>Detected miRNAs</th>
<th>Bgh</th>
<th>Pst</th>
<th>Pi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Novel miRNA</td>
<td>111</td>
<td>34</td>
<td>12</td>
</tr>
<tr>
<td>Known miRNA</td>
<td>41</td>
<td>54</td>
<td>18</td>
</tr>
<tr>
<td>Total</td>
<td>152</td>
<td>88</td>
<td>30</td>
</tr>
</tbody>
</table>
Count number distributions of the selected miRNA are shown in Figure 19. Supplementary Material 4 Data folder includes only small RNA sequences with more than 500 count respectively; novel miRNAs and small RNAs with their miRBase homologs. Data organization is explained through Appendix C.

**Figure 19.** Novel miRNA and known miRNA distributions in fungi.
3.12 psRNATarget tool predictions

Selected milRNA candidates are analyzed through psRNATarget tool with its default values as seen in corresponding material part. Raw outputs of the tool can be found from Supplementary Material 4 Data folder, and data organization is detailed in Appendix C. Table 19 represents number of hits per organism. As seen from Table 9, milRNA candidates are targeting to several regions in plant genomes. As the results are carefully analyzed, for one milRNA targeting to more than one target region, and vice versa is identified. Moreover, some of the targets were not annotated to be coding or the coding gene was not functionally known. Consequently, a sorting algorithm to mine only the best milRNA candidates is written as filtering out the plant defence related proteins. That was essential to remove the non-necessary and non-annotated hits from the data. Furthermore by this way we narrow the search area to a reasonable numbers. Thus a perl script is written. Key words used to sort targets are inserted into Appendix D. In total 3029 hits (1519, 831, 6079) was analysed for sorting and at the end 510 number of hits found to be correlated to plant defence mechanism.

Table 20 represents the numbers for analyzed milRNAs, their target prediction by psRNATarget tool that sorted by plant defense related proteins and unique milRNA and target sites. Supplementary Material 4 detailed in Appendix C includes a table showing candidate milRNAs, their target regions, and scoring scheme of psRNATarget tool.

Table 19. Number of hits predicted by psRNATarget tool.

<table>
<thead>
<tr>
<th>Hit Number</th>
<th>Bgh</th>
<th>Pst</th>
<th>Pi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Novel miRNA</td>
<td>991</td>
<td>295</td>
<td>2130</td>
</tr>
<tr>
<td>Known miRNA</td>
<td>528</td>
<td>536</td>
<td>3949</td>
</tr>
<tr>
<td>Total</td>
<td>1519</td>
<td>831</td>
<td>6079</td>
</tr>
</tbody>
</table>

Table 20. psRNATarget predictions sorted by plant defense related proteins.

<table>
<thead>
<tr>
<th>Fungi</th>
<th>Analyzed miRNA</th>
<th>Number of Hit</th>
<th>Unique miRNA</th>
<th>Unique Target Entry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Novel miRNA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bgh</td>
<td>111</td>
<td>114</td>
<td>59</td>
<td>47</td>
</tr>
<tr>
<td>Pst</td>
<td>34</td>
<td>24</td>
<td>10</td>
<td>16</td>
</tr>
<tr>
<td>Pi</td>
<td>12</td>
<td>111</td>
<td>12</td>
<td>111</td>
</tr>
<tr>
<td>Known miRNA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bgh</td>
<td>41</td>
<td>35</td>
<td>20</td>
<td>26</td>
</tr>
<tr>
<td>Pst</td>
<td>54</td>
<td>43</td>
<td>18</td>
<td>38</td>
</tr>
<tr>
<td>Pi</td>
<td>18</td>
<td>203</td>
<td>16</td>
<td>156</td>
</tr>
</tbody>
</table>

For further experimental analysis most likely candidate milRNAs should be selected. Therefore, unique milRNAs and unique target regions are also sorted by maximum
expectation value and target accessibility (UPE). At the end, 14 candidate miRNA sequences are selected and shown. Novel miRNA sequences are shown in Table 21 by predicted hairpin structures and selected known small RNAs with their best homologs in miRBase database is shown in Table 22.

Table 21. Selected novel miRNA sequences.

<table>
<thead>
<tr>
<th>miRNA name</th>
<th>Sequences</th>
<th>MFE kcal/mol</th>
<th>L*</th>
<th>Count**</th>
</tr>
</thead>
<tbody>
<tr>
<td>P.indica-m0005_3p</td>
<td>TTTGAATTTCTTGACTGATGCA</td>
<td>-52.30</td>
<td>22</td>
<td>1146</td>
</tr>
<tr>
<td>P.indica-m0038_3p</td>
<td>AGAATTATGGGGGTTGTGTCTTCT</td>
<td>-94.40</td>
<td>22</td>
<td>27648</td>
</tr>
<tr>
<td>P.indica-m0004_5p</td>
<td>TCTCTCGCTGCATGCTTTCTCTG</td>
<td>-26.80</td>
<td>22</td>
<td>143930</td>
</tr>
<tr>
<td>B1202A-m0614_5p</td>
<td>TGTGTGTGGACGTACGGGCA</td>
<td>-20.10</td>
<td>20</td>
<td>570</td>
</tr>
<tr>
<td>B1202A-m0367_5p</td>
<td>TGGCCGGTAATGAATGGGCA</td>
<td>-24.90</td>
<td>21</td>
<td>561</td>
</tr>
<tr>
<td>PSTmix-m0195_3p</td>
<td>TGCAGGGAGGGGTAGGGA</td>
<td>-164.80</td>
<td>20</td>
<td>736</td>
</tr>
<tr>
<td>PSTmix-m0459_3p</td>
<td>TGGCAAGGACTCCGGAGGCC</td>
<td>-120.30</td>
<td>21</td>
<td>50514</td>
</tr>
</tbody>
</table>

*Length of the miRNA.** Number of reads of the miRNA

Figure 20. Novel miRNA precursors shown in Table 21.
Table 22. Selected known miRNA sequences and their best homologs.

<table>
<thead>
<tr>
<th>Small RNA ID</th>
<th>L*</th>
<th>Count**</th>
<th>Sequence</th>
<th>Best homolog (mirBASE)***</th>
<th>L*</th>
<th>Family</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bgh-t0005282</td>
<td>21</td>
<td>566</td>
<td>Small RNA TGAAGGATGAGTTGATGGAA Homolog TGGAAATGATTGAGCTTGTGAGGA</td>
<td>cel-miR-1819-3p</td>
<td>22</td>
<td>undef</td>
</tr>
<tr>
<td>Bgh-t0006075</td>
<td>22</td>
<td>503</td>
<td>Small RNA TTGGAAGACCTTGTATATTATT Homolog TGGGAAGACCTTGTATATTTT</td>
<td>dre-miR-7b</td>
<td>21</td>
<td>mir-7</td>
</tr>
<tr>
<td>Pi-t0000089</td>
<td>22</td>
<td>7825</td>
<td>Small RNA AGGAATGAAAGAAGTATGTAT Homolog TGGGAATGAAAGAAGTATGTAT</td>
<td>mmu-miR-1a-3p</td>
<td>22</td>
<td>mir-1</td>
</tr>
<tr>
<td>Pi-t0001651</td>
<td>22</td>
<td>611</td>
<td>Small RNA TGAGGTAGAGTTGTATAGTT Homolog TGAGGTAGAGTTGTATAGTT</td>
<td>cel-let-7-5p</td>
<td>22</td>
<td>let-7</td>
</tr>
<tr>
<td>Pi-t0000286</td>
<td>25</td>
<td>3028</td>
<td>Small RNA GCGAGATCTTGGGTGTAGCAAAAT Homolog AGATCTTGGGTGTAGCAAAAT</td>
<td>sha-miR-716b</td>
<td>25</td>
<td>mir-716</td>
</tr>
<tr>
<td>Pst-t0000930</td>
<td>21</td>
<td>2740</td>
<td>Small RNA TTGGAAGAAGAAGTATGAAGTATGTAT</td>
<td>cel-miR-4814-3p</td>
<td>22</td>
<td>Undef</td>
</tr>
<tr>
<td>Pst-t0002068</td>
<td>21</td>
<td>1444</td>
<td>Small RNA TGGGAAGAAGAAGTATGAAGTATGTAT</td>
<td>mtr-miR5239</td>
<td>21</td>
<td>Undef</td>
</tr>
</tbody>
</table>

* Length of the miRNA, ** Read number of the miRNA, *** Best homolog of the small RNA in mirBase database found by alignment algorithms

Table 23 and Table 24 represent psRNATarget results of selected candidate miRNAs. For more detailed information raw results Appendix should be seen. The selected miRNA to target regions are sorted by Target accessibility (UPE) and Expectation value. For Pi analysis different organisms targeting are in purpose selected to show wide range of Pi for targeting.
Table 23. Novel miRNAs’ predicted target regions in plant genomes.

<table>
<thead>
<tr>
<th>miRNA name</th>
<th>Accession**</th>
<th>Exp*</th>
<th>UPE</th>
<th>miRNA/Target Alignment</th>
<th>Target Description</th>
<th>Target Organism</th>
</tr>
</thead>
<tbody>
<tr>
<td>P.indica-m0005_3p</td>
<td>GE484550</td>
<td>2.0</td>
<td>14.835</td>
<td>miRNA 20 GUAGUCAGUUCUUAAAGUUU Target 230 CAUGACUCAAGAAAUUCAAA 249</td>
<td>Proline-rich protein</td>
<td>Helianthus annuus</td>
</tr>
<tr>
<td>P.indica-m0038_3p</td>
<td>LOC_Os03g42020.1</td>
<td>1.5</td>
<td>20.066</td>
<td>miRNA 22 UCUGUGUUGGUGGUAUAAAGA Target 567 AAGACACAACCUUAAUUCU 588</td>
<td>Calcium-transporting ATPase 2</td>
<td>Oryza sativa</td>
</tr>
<tr>
<td>P.indica-m0004_5p</td>
<td>TC143937</td>
<td>2.0</td>
<td>23.293</td>
<td>miRNA 21 UCUUUUCGUACGUUCUCUCU Target 505 GGAGCAUGCAGGUGUUGUGU 525</td>
<td>Pyruvate kinase</td>
<td>Populus trichocarpa</td>
</tr>
<tr>
<td>B1202A-m0614_5p</td>
<td>CD058210</td>
<td>1.0</td>
<td>17.178</td>
<td>miRNA 20 CGGGCAUGCAGGUGUUGUGU Target 123 GCCCGUACGUCCACACACU 142</td>
<td>RND efflux system outer membrane lipoprotein NodT</td>
<td>Hordeum vulgare</td>
</tr>
<tr>
<td>B1202A-m0367_5p</td>
<td>CA682639</td>
<td>2.0</td>
<td>12.132</td>
<td>miRNA 21 ACGGAGUAUGUAUGCCCGGU Target 82 UGCCUCUAAUUAUACCCGACCA 102</td>
<td>A-kinase anchor protein 11</td>
<td>Tritium aestivum</td>
</tr>
<tr>
<td>PSTmix-m0195_3p</td>
<td>TC241290</td>
<td>2.5</td>
<td>16.720</td>
<td>miRNA 20 AGCUGAUGGAGGGAGACGU Target 454 CGAGACACCCUCUUGCCCCCCU 473</td>
<td>Serine carboxypeptidase family protein</td>
<td>Hordeum vulgare</td>
</tr>
<tr>
<td>PSTmix-m0459_3p</td>
<td>TC403010</td>
<td>3.0</td>
<td>18.670</td>
<td>miRNA 20 GGAGAGGCUCAGGAAAGG Target 299 CUUCUCUGGAAUUCUCUCUCAC 318</td>
<td>Protein kinase</td>
<td>Tritium aestivum</td>
</tr>
</tbody>
</table>

*Expectation **Target gene accession ID, ***psRNATarget details can be viewed through Table 2


<table>
<thead>
<tr>
<th>miRNA name</th>
<th>Accession**</th>
<th>Exp*</th>
<th>UPE</th>
<th>miRNA/Target Alignment</th>
<th>Target Description</th>
<th>Target Organism</th>
</tr>
</thead>
<tbody>
<tr>
<td>PI-t0000089</td>
<td>TC12570</td>
<td>2.0</td>
<td>16.779</td>
<td>miRNA 21 AUGUAUGAAGAAUGUAAGGA 1 ::::: ::::::::::::: Target 802 UAAUAAUUUCUUUACAUUCUU 822</td>
<td>Clp protease proteolytic subunit</td>
<td>Phaseolus coccineus</td>
</tr>
<tr>
<td>PI-t0001651</td>
<td>BQ848540</td>
<td>2.0</td>
<td>4.681</td>
<td>miRNA 21 UGAUAUGUUGGAUGGAGU 1 :::::::::: ::::::::: Target 51 AUUAUAUACCUACUACUCA 71</td>
<td>ATPase subunit 7</td>
<td>Lactuca sativa</td>
</tr>
<tr>
<td>PI-t0000286</td>
<td>TC400366</td>
<td>0</td>
<td>14.232</td>
<td>miRNA 25 UAAACGAUGAUGGGUUCAGACG 1 :::::::::: ::::::::: Target 560 AUUAUGCUACUACCAAGAUCUGC 584</td>
<td>Pspzf zinc finger protein-like</td>
<td>Arabidopsis thaliana</td>
</tr>
<tr>
<td>Bgh-t0005282</td>
<td>TC251661</td>
<td>3.0</td>
<td>13.112</td>
<td>miRNA 21 AAGGUAGUUGAGAAAGGAU 1 :::::::: :: ::::::::: Target 3197 UUGUAUCAGCUUUGAUCUUCA 3217</td>
<td>NB-ARC domain containing protein</td>
<td>Hordeum vulgare</td>
</tr>
<tr>
<td>Bgh-t0006075</td>
<td>CA698277</td>
<td>2.5</td>
<td>8.53</td>
<td>miRNA 20 UAAUUAUGUUGUACAGAAGGU 1 :::::::: :: ::::::::: Target 473 ACAUAACACAAGCUCUCA 492</td>
<td>Pyruvate kinase</td>
<td>Tritium aestivum</td>
</tr>
<tr>
<td>Pst-t0000930</td>
<td>AJ476982</td>
<td>1.5</td>
<td>10.593</td>
<td>miRNA 21 UGAUAGUGUUGGAAAGCGU 1 :::::::: :: ::::::::: Target 137 AUUAUUGCUACUACGCUCCAA 157</td>
<td>Protein kinase like protein</td>
<td>Hordeum vulgare</td>
</tr>
<tr>
<td>Pst-t0002068</td>
<td>CJ641074</td>
<td>3.0</td>
<td>19.941</td>
<td>miRNA 21 UGGAAGAUGAAGAAAGGGGG 1 :::::::: :: ::::::::: Target 274 CUCUUCUACUUCUCUUCGC 254</td>
<td>Phosphatidylinositol 4-kinase</td>
<td>Tritium aestivum</td>
</tr>
</tbody>
</table>

*Expectation **Target gene accession ID, ***psRNATarget details can be viewed through Table 2
Table 23 and 24 indicates that important genes are found as targets of the predicted milRNAs. *Piriformospora indica* (Pi) is a symbiotic fungus very beneficial to plant roots. As a result of the analysis, we found that candidate milRNAs of Pi targeting to very important regions in diverse organisms. Clp protease proteolytic subunit in *Phaseolus coccineus*, ATPase subunit 7 in *Lactuca sativa*, Pspzf zinc finger protein-like *Arabidopsis thaliana*, proline rich protein like in *Helianthus annuus*, calcium transporting ATPase 2 in *Oryza sativa* and Pyruvate kinase in *Populus trichocarpa* are predicted as most probable regions targeting by candidate milRNAs of Pi.

*Puccinia striiformis* f. sp. tritici (Pst) is an obligate parasitic fungi infecting *Triticum aestivum* and *Hordeum vulgare*. We identified candidate milRNAs targeting to serine carboxypeptidase family protein, protein kinase, protein kinase like protein and Phosphatidylinositol 4-kinase. The candidate targets are all related to kinase proteins. It is such an important investigated since kinase proteins are fuels of the pathways, and they have significant roles in defence mechanism.

*Blumeria graminis* f.sp. hordei is also an obligate parasitic fungi of *Hordeum vulgare* and *Triticum aestivum* is nonselective host of it. NB-ARC domain containing protein, pyruvate kinase, RND efflux system outer membrane, and A-kinase anchor protein 11 are identified as target regions. Again kinase type proteins are found as most likely target regions.
CHAPTER VI

DISCUSSION

4.1 Why miRNAs

miRNAs are part of a large group, small non-coding RNAs. Through a various RNAi (RNA interference) pathways they regulate important processes in the cell such as, developmental changes, cellular timing, fat storage, etc. (D. P. Bartel, 2004). The investigations also show that miRNA mechanism itself is very significant in cancer development. For example, when miRNA biogenesis is interrupted by down regulation of Dicer complexes cancer development is observed. Also, over expression of e.g. miR-155 leading to various cancer types like breast, lung and colon is known (Jansson & Lund, 2012). miRNAs are significantly linked to the key cellular processes and very fundamental to cell survival in Eukaryotes. Thus, their identification and their functional annotations are very imperative in all the aspects of biology.

4.2 Significance of miRNA representation

miRNAs are small 20-24 nucleotide length RNA sequences. Similarity and dissimilarity measures are usually used to compare miRNA sequences by each other, but operating directly on the sequences would be discarding. Thus, representation of a miRNA letter as a mathematical metric is required, and algorithms operating on the sequence need to be developed. Most of the algorithms in literature are using similarity as measure to represent a letter, like amino acid sequence of a protein. In that perspective, to compare miRNA similarities all together pairwise sequence algorithms, Smith-Waterman as local and Needleman-Wunsch as global, were used in this study. Furthermore, as a novel perspective we also counted k-mers in that sequences and represented miRNA with a vector present for each distinctive “k”s. The method is significant as it is independent from sequence order and the similarity itself. In our analysis we used directly the scores of pairwise alignments. In literature other methods using the same strategy on amino acid sequences exist. The tools OrthoMCL (L. Li et al., 2003) and TribeMCL (Enright et al., 2002) are two
important examples to them. They evaluate the score of protein similarities directly from sequence and takes negative logarithms of p-values of these scores. However, p-value calculation and the negative logarithms of it are not applicable into our study. The reason behind is mostly because of the small frame miRNAs. p-value becomes very high decimal small numbers and close to each other. Hence, even the smaller numbers makes the comparison of cluster algorithms harder. Nevertheless, a new attitude can be developed to score the alignments, and to evaluate the significance of these scores.

As mentioned before, there is a nearly perfect base pairing between miRNA and their target mRNAs in plants (D. P. Bartel, 2004; Pratt & MacRae, 2009; Rhoades et al., 2002). This situation conserves the sequence of miRNA even with mutations or deletions over its evolution. At the same time, miRNAs within the same family involve in same functional processes and show very significant similar expression patterns and these miRNAs in the same family show high level of sequence similarity (B. Bartel & Bartel, 2003). Therefore, in our analysis, when their sequence information documented as metric values of which as independent from the module miRNAs in the same family are clustered in the same group. As the similarity between sequences increased the probability involving into the same cluster also rises.

4.3 Decision on which clustering algorithm

The use of application of machine learning algorithms into biological datasets is very common. For example, there are various algorithms developed to cluster transcriptome series. Yet, these algorithms are not applied into a matrix representing directly the sequence itself. Markov Clustering Algorithm (MCL) is one of the machine learning methods used in some applications to cluster proteins with directly amino acid sequence metric matrices. In this study, MCL algorithm was also applied and found to be enlightening to group miRNA sequences.

Other than MCL algorithm, we also applied K-means, Cluster Aggregation Algorithm (CLAG) and Self-Organizing Three Algorithm (SOTA). These algorithms all have diverse procedures, working with different methodologies. Indeed, by using different algorithms we tested our hypothesis with different measures. Depending on the results by these algorithms, unless different cluster algorithms produce different partitions of the data, there are minor changes in their function enrichments by TAM analysis and the results are correlated. The classical clustering algorithm K-means and SOTA produces predetermined number of groups and so data coverage of these algorithms are really high (Table 14). Thus, prior to their run cluster numbers should be known. CLAG algorithm works with predetermination of inflammation and normalization values, but it is independent from setting of k number. It is found that CLAG is only useful to find the most condensed groups. Data coverage of MCL
algorithm changes by inflammation value also. In our analysis 73% to 86% (Table 14) coverage is identified. MCL algorithm is the most beneficial tool when pairwise similarities of miRNAs are shown as real distance values. Certainly using different methods to cluster miRNA sequences is convenient as often clustering depends on type, shape and size of the dataset. In fact, deciding which clustering algorithm is most valuable for the dataset is subtle problem of bioinformatics.

Moreover, hierarchical clustering on nucleic sequences itself is possible through multiple sequence alignment (MSA) (Larkin et al., 2007). However, MSA method directly operates on sequences, but not necessarily on a metric derived from similarity information. That cause extensive steps of iterations to detect global patterns of similarities. Moreover, MSA algorithms are greedy to collect errors by more than a few iterations with huge number of sequences. As the size of the dataset increase, algorithm complexity also increases. Therefore, MSA is not relevant for large datasets. In that direction our method is very favorable, since it is able to shorten the time required for clustering and uses also non-hierarchical algorithms like MCL unless classical expression based techniques are used. Our method with respect to MSA methods, can handle very large datasets without any collected errors, and both global and local similarities can be used to group the sequences.

4.4 Importance of validation of the clusters

A prior to functional analysis of the miRNA clusters, the strength of the groupings are determined with Dunn Index (DI) calculation. As mentioned before, when the dataset character is unknown it is hard to decide which algorithm at which conditions required applying. Therefore, DI calculation is a way to understand whether the clusters separated well. In literature, other statistical analysis exists to control and estimate cluster importance. DI calculation is one of the most classical and commonly used algorithms. In our study, DI calculation could not applied into for only MCL algorithm, as MCL is not a vector clustering method, and the distances of MCL algorithm are simulated distance values not real Euclidean metrics. The DI results for other algorithms can be viewed through Table 15. DI results for random matrices were different and higher than the real matrices. That was because that randomly created matrices are more homogenenic than the real matrices. CLAG also resulted with high DI values since the algorithm only zoom into the compressed regions keeping off the unstable clusters.

TAM tool consists of a large collection of human miRNAs. miRNA categories are constructed with diverse relations of these miRNAs, so currently it is the most useful tool to understand relations within the human miRNA clusters. We used TAM site for functional categorization of output clusters of our methods. However, some of the TAM categories were poor in data, and there were not quite enough information. Actually, it is mostly because those miRNA experimental studies require time. For
example, tissue and HMMD categories two of them, and in our analysis their enrichment results were very low. However, this does not reflect the quality of the clusters.

TAM tool validation analyses have shown that clusters generated by these metric systems results functionally meaningful relations and can be viewed through Table 14, 15, and 16. The random samplings from the data were applied to measure the significance of the cut-off values used in TAM calculations. The results show that at most 46% enrichment can be handled by random groups. No other clusters analysed in the study resulted as low value as random matrices. This indicates that the TAM validation was meaningful for analysis of enrichment in miRNA clusters.

Classical K-means with Smith-Waterman algorithm applied matrix found to be the most useful method by 92% enrichment in ALL categories. Indeed, percentage enrichments of all methods were correlated. MCL algorithm result with 60% best enrichment by 2-4 inflammation values. For CLAG algorithm, k-mer application was the most useful method giving 80% enrichment by Needleman-Wunsch algorithm. In general, Smith-Waterman algorithm is found as best to identify the similar regions in clustering (Table 16). The most enriched category was family as suspected before as in family members of a miRNA family there can be a good sequence similarity. The result shows that the sequence representation methods are eloquent also. For cluster and function categories the enrichment results are not high as family category. That is because sequence similarity in that manner is not correlated as family category does.

In fact, different clustering algorithms result with different clusters with differently splitting the data as their algorithms are not the same. Optimization of the algorithms covering all results may be a solution, or a new clustering algorithm can be developed. Therefore, if a clustering algorithm grouping metrics of mature miRNA sequences is developed, it would be very innovative.

4.5 miRNA application to the pipeline

The workflow generated through our clustering analysis was shown in Section 3.10, Figure 18. Novel and known candidate miRNA sequences of fungal species *Puccinia striiformis* f. sp. *tritici*, *Blumeria graminis* f.sp. *hordei* and *Piriformospora indica* are combined and analyzed through the workflow presented in this study. Based on the size of the data set, an appropriate method were chosen and applied for each organism since the data size for each organism are different and the effect of the of any given method is unknown. Thus, diverse number of clusters with changed coverage of data was produced for each organism. The cluster numbers for each organism were represented in Table 17. The purpose of this application is to estimate functions of novel miRNA sequences from enrichments of them with known miRNAs. Therefore, some clusters with no enrichment were eliminated and only the
clusters represent enrichment were presented. The pipeline is found to be informative in assignment of these novel miRNAs based on sequence similarity matrices. Thereby, novel miRNAs were able to be assigned into miRNA families. However, for further examination by the pipeline, an unknown dataset can be analysed through different methods and the best clusters with high dunn indexes can be chosen. The next step for the study of the unknown miRNAs is to conduct molecular experiments. Furthermore, computational assignment of the known miRNA sequences into target genes and target gene enrichment studies would be innovative for investigation of the role of novel miRNAs.

4.6 Significance of miRNA prediction

High throughput sequencing tools like illumine solexa technology were the most recent breakthrough of in the state of the art sequencing technologies. Development of reliable and fast sequencing tools allowed many projects to be initiated to uncover the mystery of both well known species and least studied species like fungi. Small RNA sequencing projects ignite many brilliant research ideas, such as search of miRNA sequences in fungi. In this study, we also interested in prediction of miRNA elements in fungi. Small RNAs were sequenced by the Illumina technology overall analyses were provided by the BGI. Two obligate plant parasites (Bgh and Pst) and one mutualistic living fungus Pi were sequenced as a part of the study. Small RNA annotations were different (Table 5 and Figure 15) as small RNA length distributions (Figure 14). The length distribution of the Pi appreared to very different than that of Bgh and Pst. This result might suggest a different and unknown miRNA or miRNA generation mechanism in Pi. A different mechanism can be considered likey since Bgh and Pst are obligate parasites as part of bariosides, Pi on the other hand is a mutualistic living symbiont fungal root endophyte. The symbiotic life of Pi and the small RNA mechanism is currently unknown. In length distribution of Pi, there are two peaks different than Bgh and Pst has one in 22 nucleotides. The peak in 22 nucleotides illustrates the miRNA existence. However, two peaks in 22 and 25 nucleotides in Pi distribution graph may signify a different small RNA mechanism in Pi. This also could explain why there are less predicted novel miRNA precursors in Pi than Bgh and Pst.

Selection of candidate miRNA sequences for further experimental analysis is an important step, as experiemntal analysis in laboratories can be costy through several steps of the experiemt. Computational programs predicting miRNA stem loops and mature miRNA sequences are developed and they are very useful to assign small RNAs as candidate miRNAs. Nevertheless, prediction of miRNA sequence is only the first phase; also some tools to predict the functions of these miRNA sequences need to be developed. In our analysis, besides these novel miRNA predictions by MIREAP algorithm (BGI, n.d.), we also analysed small RNA sequences having homology to miRBase, and we used read numbers of these sequences as a
preliminary limit to filter out the noise of incorrect readings. Thus, in total 270 (152 Bgh, 88 Pst, and 30 Pi) miRNAs were used (Table 18) for target analysis. We used 500 reads (indicating the level of expression) as cut off value. Therefore in our analysis some miRNA sequences with low expression levels are lost. We found that some of the small RNAs showed homology to some important miRNA families like miR-1 and miR-7 (Table 22). miR-7 family has roles in regulating expression of messenger RNA and their miRNA structure is conserved over the species (Reddy, Ohshiro, Rayala, & Kumar, 2008). Bgh-t0006075 has high homology to miR-7 family, thus a similar function in Blumeria graminis f. sp. hordei is expected. miR-1 family, especially the members, miR-1-1 and miR-1-2, are vital in physiological and development of heart tissues and they are related to important heart diseases in human (Koutsoulidou, Mastroyiannopoulos, Furling, Uney, & Phylactou, 2011). The functions of miRNAs must be confirmed by target analysis and experimental analyses.

4.7 Importance of target analysis

The candidate fungal miRNA sequences were analyzed by online tool psRNATarget to find their target RNA locations in their host organisms. psRNATarget tool is specific for miRNA alignment of plant genomes only. This makes the tool most suitable for our analysis. Through the analysis, default optimization values were applied as fungi miRNA character on target region is unknown. The tool has some important functions while identifying the target locations. Target accessibility and maximum expectation values are two of them. From several candidate target locations (Table 19), we only selected immunity related targets by a careful sorting with keywords described in Appendix D. In total, 81 unique novel miRNAs targeting to 174 unique regions, and 54 known miRNAs targeting to 220 unique regions were detected (Table 20). For further experimental studies, 7 novel and 7 known miRNA sequences were proposed. These sequences are the ones having the highest potential to be miRNA candidates with their low Maximum Expectation and high UPE (target accessibility) values.

The selected miRNA target regions found to be very valuable in determination of function of miRNAs found in the study (Table 23 and 24). The target of P.indica-m0005_3p, proline rich protein, resides in cell wall protecting the cell from fungal attacks (Williamson, 1994). Pyruvate kinase target of P.indica-m0004_5p and Bgh-t0006075 is one of the key regulatory enzymes. Pyruvate kinase triggers important signaling pathways (Duggleby & Dennis, 1973). Furthermore, another candidate miRNA Pst-t0002068 predicted as targeting to phosphatidylinositol 4-kinase which is associated with plant cell wall too. Pathogen miRNA can attack the genes encoding proteins in the cell wall thereby regulating first defence layer of host cell; the cell wall (Xu et al., 1992).
It is well known that whenever the pathogen interacts with the host plant, in the cell wall plant defence is activated producing reactive oxygen species. The start of the cascade of signal transduction are regulated with ATPase activities (Elmore & Coaker, 2011). In our study, we found that a candidate miRNA, PI-t0001651, is targeting to ATPase subunit 7 (Table 24).

Protein kinases are known as vital having central role in plant defence to recognize the pathogen enters and to initiate signaling. In our analysis, we identified many Protein kinases as miRNA target genes; such as Pyruvate kinase in *Triticum aestium* and protein kinase like protein in *Hordeum vulgare* by miRNA Pst-t0000930 *Puccinia striiformis* f. sp. *tritici*, miRNA Bgh-t0006075 of *Blumeria graminis* f.sp. *hordei*, respectively (Table 23 and 24).

Plant miRNA target prediction tools vary, and the performance of each tool is often very diverse. In the paper presented by Srivastava et al. there are nearly 11 dissimilar tools with changed activities (Srivastava et al., 2014). They propose to use these methods in combination instead of using a single tool. For example, with psRNATarget tool that we used can be optimized by using “Tapirhybrid” (Billiau et al., 2010) to increase accuracy and precision of the predictions.
CHAPTER V

CONCLUSION

5.1 Overview

In search of finding miRNA families with predicted functions regulating their target genes in turn pathways such as development, immunity, environmental responses and many more. The expression levels of miRNAs with time series are commonly preferred approach, since experimental analyses are considered most reliable and promising. However, they are indeed costly and time consuming. Therefore, there is an urgent need in generating computational tools for cluster analyses to determine miRNAs families with individuals having similar functions. Toward this end, this thesis is focused on developing a novel approach using the data available in databanks of human genome with experimentally determined mature miRNA sequences. Given a list of mature miRNA sequences, sequence content translated into a metric system and clustered by available clustering algorithms. Moreover, as a part of this study, miRNAs were hunt down by obtaining small non-coding RNA sequencing dataset in our laboratory, and their target regions were identified, therefore candidate miRNAs with their target regions were computationally achieved.

5.2 Conclusion

In this thesis study, we addressed two important questions. The first object of the study was to provide a workflow for clustering miRNA sequences independent from their expression profiles. The pipeline presented here accurately clustered miRNA groups using sequence clustering approach by means of existing machine learning algorithms, K-means, CLAG, SOTA and MCL. Given a list of mature miRNA sequences, similarity relations were detected by methods like k-length substring counting and pairwise sequence alignment algorithms. To detect pairwise
similarities between two sequences Smith-Waterman and Needleman-Wunsch algorithms were used. As a result, three different sequence representation methodologies were utilized to detect sequence similarities. Pairwise sequence algorithms were used to construct a matrix filled by scores of descriptive scores. An all-to-all approach is used and all sequences in the list compared to each other. Thus, the filled matrix becomes the representations of distances between all miRNAs, and it is used as input of cluster algorithms. The other approach was k-mer counting, independent from the order which is a priority in pairwise alignment algorithms. It is also a novel approach for representation of a sequence as input of clustering algorithms.

Preexisting clustering methods used in this study have been not previously applied into a miRNA sequence metric matrices. In that perspective too, this study has also an innovative outcome. Only, MCL algorithm which is a graphical clustering method indeed was originated to cluster protein sequence score metrics, which is very useful for sequences represented as distance values. Hierarchical clustering on nucleic sequences is possible through multiple sequence alignment (MSA). However, MSA methods directly operate on sequences, but not on a metric in matrix. Thus, this thesis study developed a new clustering approach specifying the detection of miRNA sequence groups since currently there is no miRNA sequence based clustering algorithm. By using various existing clustering algorithms, we were able to instruct appropriate optimizations to chose best possible one most fitting for miRNA functional clustering analysis.

Statistical evaluation of clusters was completed through DI calculations. Only the clusters significantly showed strength of clusters used in the study. The functional enrichments in that clusters were calculated by very effective bioinformatics tool, Tool for annotations of miRNA; TAM uses a given set of miRNAs by calculating p-values of enrichment in the set and it shows the number of sequences in the cluster found in the same category. With a self-written Perl codes, percentage of significant enrichments were calculated and presented in this study. Table 14, 15 and 16 present superiority of the clusters and represent TAM results. Our analyses have shown the clustering approaches used in the study represent important functional enrichments. Although, there are some minor changes compared to TAM results when similarity detection method changed. Most significantly, in family category we saw the highest enrichments indicating that sequence similarity in miRNA families is predictable. Since, our method yielded significant similarities it is applicable to sequence clustering for miRNAs regardless of the small differences that were observed in comparison to TAM output. Thus, our results indicate that a higher enrichment was obtained compared to any random matrix that is used.

The final results of our analyses show that biologically important patterns do exist in miRNA sequences and they can be found by similarity detecting tools. Moreover, there is important sequence similarities in miRNAs and this likeness are directly related to function. Consequently, a novel dataset, deep sequencing (small RNASEq)
of fungi were analyzed by our original pipeline created and novel groups were found for further studies. The functional assignment of the novel miRNA sequences by the known small RNAs is possible by this way. We suggest that functional relations through experimental analysis should be sought in the constructed clusters. We hope that this study will comprise a baseline for future studies.

RNA interference mechanism is found to be conserved in all eukaryotes. Nonetheless, miRNA pathway elements are not shown to exist in fungi. This raised a suggestion that fungi do not produce miRNA (Lee et al., 2011). However, recently scientists discovered that some miRNA like small elements (milRNA) are produced in filamentous fungus suggesting a presence of an independent from commonly known Dicer dependent pathways is existing in fungi (Kang et al., 2013). Hence, we hypothesized out that; plant pathogens such as *Puccinia striiformis* f. sp. *tritici*, *Blumeria graminis* f.sp. *hordei* and a mutualistic symbiotic fungi, *Piriformospora indica*, are likely to produce milRNA elements. Since, cross kingdom regulation through miRNA signalling is highly reformist issue, we have initiated this study. Since all plant fungi are interacting with its hosts by sending effector proteins, we raised the question if mature milRNAs are sent by the invading fungi into the plant cell. In our analysis, we identified milRNA in Pst, Bgh and Pi. As the reports sent from BGI suggested, the clean read numbers and small RNA annotations of the organisms were different. The reason behind that the milRNA generation mechanism can be diverse in these organisms as like in Pi there could also be novel small RNA generation machinery as Pi’s different living conditions. The small RNA distribution and nucleotide bias analysis had been shown that the general prospect of the graphs like to the results representing miRNA generation.

Targets of these small RNA sequences predicted as candidate milRNA sequences were searched on the psRNATarget tool database. As a result many candidate target regions were branded and reported with this study. Yet, for further studies 7 known and 7 novel milRNA sequences and their target regions were presented after application of several filtrations to find only the target regions related to plant defence mechanism. These sequences are predicted by their psRNATarget tool options explained in corresponding methods part.

In conclusion, we found candidate milRNA sequences predicted in the fungi that are commonly used in our labrotary with the candidate target regions in plant genomes.

### 5.3 Future studies

As future perspectives, with parallel to miRNA clusters generated by various methods and with our original pipeline and candidate target genes can be utilized to generate target gene clusters. Since, similar miRNA sequences targeting to similar targets, the targets will most likely be part of the same pathways. GO enrichment
tool (Gene Consortium, 2000) can be used to enrich set of genes, and like TAM p-values can be taken advantage of. Therefore, the study can be further developed by specific target analysis. Moreover, as the sequence similarity in miRNA family members were searched, a correlation study between the miRNA clusters and expression levels (as the number of reads, data available) can be conducted which will innovative and informative if a reasonable correlation is found.

Candidate miRNAs proposed in this study also required to be validated with experimental analysis. To demonstrate miRNAs in fungi, miRNA quantification analysis need to be carried out. For this purpose, qRT-PCR can be considered as the most valuable tool. Indeed, the Akkaya lab has initiated such analyses, already some preliminary results were obtained using the technique developed by (C. Chen et al., 2005) for real time quantification of miRNAs by stem-loop producing primers.

The second step in miRNA analysis is to define miRNA targets. For that purpose co-expression of the miRNA by its target gene need to be investigated with invivo. Northern blot method using total RNA isolates can be applied. Furthermore, qPCR experiments using TaqMan assays (Applied Biosystems, Foster City, CA) are very sensitive for specific miRNA and its target. Also, in situ experiments are possible to show the specific miRNA is biologically effecting cells or tissue (Kuhn et al., 2008).

The next step for validation of miRNA and its target would be designing knockdown assay. Silencing the miRNA in the cell through mutation on miRNA gene itself can be one approach.
This master thesis supported by TÜBİTAK 2210 BİDEB thesis grant and research grand of Akkaya project 113Z038 and 110T445.
REFERENCES


66


67


APPENDICES

Some supplementary material of this study is provided in electronic format in one CD. Organizations of the contents in the CD are shown in Appendix A, Appendix B and Appendix C.

APPENDIX A: SMALL RNA SEQUENCING REPORTS

Small RNA sequencing experiment is carried on by the firm BGI, Chine. The bioinformatics analysis reports are inserted into a CD as .rar files. Supplementary Material 1 includes the files for Bgh, Pt and PI. Arrangements of the files are the same. Contents of the files are explained item by item:

1. Graph annotation: Small RNA annotation statistics and graphs, distributions, miRNA Variant analysis, miRNA prediction results.
   1.1. Annotation: Total and unique small RNA statistics and pie charts.
   1.2. Match exon: Statistics and pie chart for small RNA reads matching to exon or intron regions.
   1.3. Match_genome: Chromosome distributions of small RNA reads
   1.5. Match_repeat: Repeat small RNA analysis
   1.6. Match_Rfam: Statistical analysis for small RNAs matching to Rfam which is very valuable database including information about non-coding RNAs, and their structures (Burge et al., 2013).
   1.7. miVariant: Position and first nucleotide bias analysis for miRNA annotations.
   1.9. Primary: Length distribution analysis.

2. Target prediction: Novel and known miRNA target analysis. Prediction strategy can be viewed through README file in the same director.
3. Result_advance: Advance bioinformatics analysis reports, small RNA annotations, novel miRNA predictions. Bioinformatic analysis details can be viewed through README file in the same directory.

- **clean.fa**: clean small RNA reads, the raw result in fasta format.
- **match Genome.txt**: Small RNA sequences perfectly matching reference genome sequence.
- **match Genome.stat**: The summary of small RNAs matching reference genome.
- **match_ncgb.txt**: Small RNAs matching rRNA, tRNA, snRNA, snoRNA deposited at NCBI genbank database.
- **match_ncgb.stat**: The summary of match_ncgb.txt
- **match_Rfam.txt**: Small RNAs matching rRNA, tRNA, snRNA, snoRNA deposited at Rfam database.
- **match_Rfam.stat**: Summary file for match_Rfam.txt
- **match_repeat.txt**: Small RNAs matching repeat deposited at hg18 database.
- **match_repeat.stat**: The summary of match_repeat.txt
- **overlap exon sense.txt**: Small RNAs matching exon sense strand.
- **overlap exon antisense.txt**: Small RNAs matching exon antisense strand.
- **overlap intron sense.txt**: Small RNAs matching intron sense strand.
- **overlap intron antisense.txt**: Small RNAs matching intron antisense strand.

3.1. Annotation: Annotations assigned to each small RNA sequence.

3.2. Prediction: Novel miRNAs identified by Solexa sequencing technology and Computational analysis and hairpin structures of predicted novel miRNAs.

- **match_mirbase.txt**: Small RNAs whose sequences are similar with miRNAs deposited at miRBase.
- **miRNA.fa**: This file is derived from "match_mirbase.txt".

3.3. Variant Prediction: Variant analysis of the novel miRNAs, hairpin structures of predicted variants.

3.4. BGI_exp_dif: Count Level of novel miRNA reads. This file does not exist for Bgh analysis.

3.5. BGI_family_analysis: miRNA family analysis. This file does not exist for Bgh analysis.

3.6. BGI_function: GO and KO enrichment analysis of known and novel miRNA predictions. This file does not exist for Bgh analysis.

4. Results_primary: Clean small RNA reads and length to number of read stati
APPENDIX B: CLUSTER ANALYSIS OF MIRNAS

Supplementary Material 3 folder includes three excel files containing cluster analysis results of Bgh, Pst and Pi miRNAs. In the first column of the excel file is novel and homolog miRNA names. The second column is the cluster associations of these miRNAs. The third column shows whether the miRNA is novel or homolog.
APPENDIX C: SMALL RNA SEQUENCES SELECTED FOR TARGET GENE ANALYSIS AND TARGET GENE ANALYSIS RESULTS

Supplementary Material 4 folder includes files concerning miRNA target gene analysis by psRNATarget tool. The organization of the files detailed item by item:

1. psRNATarget_raw_results: The folder contains raw psRNATarget predictions for Known and Novel miRNA selections for Pst, Bgh and PI. All of the results are organized in excel files.

   miRNA_Acc: miRNA name
   Target_Acc: Accession ID of target gene
   Expectation: Score for complementing miRNA and its target. Score is calculated by algorithm of miRU.
   UPE: It is the maximum energy (delta G) to open hairpin structure of the miRNA. Less energy means more possibility to target the region.
   miRNA aligned fragment: miRNA sequence fragment aligning to the target region
   miRNA start: Start position of aligned miRNA sequence
   miRNA end: End position of aligned miRNA sequence
   Target start: Start position of aligned target gene fragment
   Target end: End position of aligned target gene fragment
   miRNA_aligned_fragment: miRNA fragment aligning to target gene
   Target_aligned_fragment: Target fragment aligning to miRNA
   Inhibition: Repression way of the miRNA
   Target_Description: Target gene details and annotations
   Multiplicity: Number for miRNA targeting to the same target region
   Organism: Target Organism

2. Novel_miRNA_sequences: The file includes novel miRNAs by more than 500 number of count for Bgh, Pst and PI.

   miRNA name: miRNA name given by BGI
   Sequence: miRNA sequence
   Count: Count level of the miRNA

3. Known_miRNA_sequences: The file contains small RNAs of Bgh, Pst and PI owe homology to miRBase with more than 500 number of count.

   Small RNA ID: miRNA name
   Sequence: miRNA sequence
   Count: Count level of the miRNA
   Homolog (miRBase): miRBase homolog of the corresponding miRNA
Homolog Sequence: Sequence of the homolog miRNA

4. psRNATarget_Predictions: Targets genes of miRNAs searched by psRNATarget tool for Bhg, Pst and Pi are included in this file. Target genes are sorted with their plant defence system associations.

miRNA name: miRNA names given by BGI or modified from small RNA IDs.
Accesion: Target gene accession ID.
Expectation: Score for complementing miRNA and its target. Score is calculated by algorithm of miRU.
UPE: It is the maximum energy (delta G) to open hairpin structure of the miRNA. Less energy means more possibility to target the region.
miRNA aligned fragment: miRNA sequence fragment aligning to the target region
Target aligned fragment: Target region aligning to the miRNA sequence
Target Description: Target gene annotations
Target Organism: Target gene Organism
**APPENDIX D: PLANT DEFENCE PROTEIN KEYWORDS**

Plant defence proteins as keywords used in target gene analysis.

| Beta-fructofuranosidase  |
| Auxin                   |
| Trev                    |
| Transcription           |
| PEX14                   |
| Acid phosphatase        |
| Dnase                   |
| ABC transporter permease |
| Potassium transporter   |
| Thiamine biosynthetic enzyme |
| Apolipoprotein B        |
| MIPC synthase           |
| EIF                     |
| Monosaccharide transporter |
| Integrase               |
| PEARLI                  |
| Pol polyprotein         |
| Peptidase M1 family protein |
| Chalcone synthase       |
| Aggregation promoting factor related surface protein |
| Secretion protein HlyD precursor |
| RND efflux system outer membrane lipoprotein |
| NodTCaleosin            |
| ACC synthase            |
| Crossover junction endodeoxyribonuclease RuvC |
| TE4                     |
| PHD domain containing   |
| peptidase U7 family precursor |
| Phosphoribosylformylglycinamidine |
APPENDIX E: PERL SCRIPTS


```perl
use strict;
use warnings;
my $output_file="output.txt";
open(OUTPUT,"">".$output_file);
my @seq_array;
my @sequences;
my @headers;
#read the text file, separate sequences and headers into 2 arrays
open (my $inFile, '<', "hsa-fasta.txt") or die $!;
while (<$inFile>)
{  
   push(@seq_array,split /\s+/);
}
close ($inFile);
for (my$i=0; $i <= $#seq_array-1; $i+=2) { push (@headers, $seq_array[$i]) }
for (my$i=1; $i <= $#seq_array+1; $i+=2) { push (@sequences, $seq_array[$i]) }

#blast matrix initiation
@blastmatrix=();
for (my $i=0; $i< $#sequences+1;$i++)
{  
   for (my $j=0; $j< $#sequences+1;$j++)
   {  
      $blastmatrix[$i][$j]='';
   }
}
#call blast and fill the matrix blast with the scores
for (my $e=0;$e< $#sequences+1;$e++)
{  
   for (my $k=0;$k< $#sequences+1;$k++)
   {  
      $blastmatrix[$e][$k]=makeblast($sequences[$e],$sequences[$k]);
   }
}
##subroutine blast
my $seq1;
my $seq2;
sub makeblast
{  
   $seq1=shift;
   $seq2=shift;
}
```
#print "1:$seq1"."2:$seq2"."\n";
my $MATCH = 1; # +1 for match
my $MISMATCH = -1; # -1 for mismatch
my $GAP = -1; # -1 for gap
# initialization
my @matrix;
$matrix[0][0]{score} = 0;
$matrix[0][0]{pointer} = "none";
for(my $j = 1; $j <= length($seq1); $j++)
{
    $matrix[0][$j]{score} = $GAP * $j;
    $matrix[0][$j]{pointer} = "left";
}
for (my $i = 1; $i <= length($seq2); $i++)
{
    $matrix[$i][0]{score} = $GAP * $i;
    $matrix[$i][0]{pointer} = "up";
}
# fill
for(my $i = 1; $i <= length($seq2); $i++)
{
    for(my $j = 1; $j <= length($seq1); $j++)
    {
        my ($diagonal_score, $left_score, $up_score);
        # calculate match score
        my $letter1 = substr($seq1, $j-1, 1);
        my $letter2 = substr($seq2, $i-1, 1);
        if ($letter1 eq $letter2)
        {
            $diagonal_score = $matrix[$i-1][$j-1]{score} + $MATCH;
        }
        else
        {
            $diagonal_score = $matrix[$i-1][$j-1]{score} + $MISMATCH;
        }
        # calculate gap scores
        $up_score = $matrix[$i-1][$j]{score} + $GAP;
        $left_score = $matrix[$i][$j-1]{score} + $GAP;
        # choose best score
        if ($diagonal_score >= $up_score)
        {
            if ($diagonal_score >= $left_score)
            {
                $matrix[$i][$j]{score} = $diagonal_score;
                $matrix[$i][$j]{pointer} = "diagonal";
            } else
            {
                $matrix[$i][$j]{score} = $left_score;
                $matrix[$i][$j]{pointer} = "left";
            }
        }
    }
}
else
{
    if ($up_score >= $left_score)
    {
        $matrix[$i][$j]{score} = $up_score;
        $matrix[$i][$j]{pointer} = "up";
    }
    else
    {
        $matrix[$i][$j]{score} = $left_score;
        $matrix[$i][$j]{pointer} = "left";
    }
}

my $ak = length($seq1);
my $kara = length($seq2);
return $matrix[$kara][$ak]{score};

#write the distance/blast matrix on outfile

print OUTPUT "matrix
@headers
n";
for(my $i = 0; $i < $#sequences+1; $i++)
{
    # $#array_2d gives the highest index from the array
    print OUTPUT $headers[$i]."\t";
    for(my $j = 0; $j < $#sequences+1; $j++)
    {
        print OUTPUT "$blastmatrix[$i][$j]\t" ;
    }
    print OUTPUT "\n";
}
print "\n#

2. Perl Script to construct All-vs-All Matrix by Smith-Waterman Algorithm from FASTA file of miRNA sequences. The code is adapted from etutorials.org site http://etutorials.org/Misc/blast/Part+II+Theory/Chapter+3.+Sequence+Alignment/3.2 +Local+Alignment+Smith-Waterman/.

use strict;
use warnings;
my $output_file="output.txt";
open(OUTPUT,">".$output_file);
my @seq_array;
my @sequences;
my @headers;

#read the text file, separate sequences and headers into 2 arrays
open (my $inFile, '<', 'hsa-fasta.txt') or die $!
while (<$inFile>)
{
push(@seq_array, split /\s+/);
}
close ($inFile);
for (my $i=0; $i <= $#seq_array-1; $i+=2) { push (@headers, $seq_array[$i]) ;
for (my $i=1; $i <= $#seq_array+1; $i+=2) { push (@sequences, $seq_array[$i]) ;
print OUTPUT2, @headers;
my @blastmatrix = ();
for (my $i=0; $i< $#sequences+1;$i++)
{
    for (my $j=0; $j< $#sequences+1;$j++)
    {
        $blastmatrix[$i][$j] = '';
    }
}
# call blast and fill the matrix blast with the scores
for (my $e=0;$e<=$#sequences+1;$e++)
{
    for (my $k=0;$k<=$#sequences+1;$k++)
    {
        $blastmatrix[$e][$k] = makeblast($sequences[$e], $sequences[$k]);
    }
}
### subroutine blast #################
my $seq1;
my $seq2;
sub makeblast
{
    $seq1=shift;
    $seq2=shift;
    # scoring scheme
    my $MATCH = 1; # +1 for letters that match
    my $MISMATCH = -1; # -1 for letters that mismatch
    my $GAP = -1; # -1 for any gap

    # initialization
    my @matrix;
    $matrix[0][0]{score} = 0;
    $matrix[0][0]{pointer} = "none";
    for (my $j = 1; $j <= length($seq1); $j++)
    {
        $matrix[0][$j]{score} = 0 * $j;
        $matrix[0][$j]{pointer} = "none";
    }
    for (my $i = 1; $i <= length($seq2); $i++)
    {
        $matrix[$i][0]{score} = 0 * $i;
        $matrix[$i][0]{pointer} = "none";
    }
    # fill
    my $max_i = 0;
    my $max_j = 0;
    my $max_score = 0;
    for (my $i = 1; $i <= length($seq2); $i++)
    {
for my $j (1..length($seq1)) {
    my ($diagonal_score, $left_score, $up_score);
    # calculate match score
    my $letter1 = substr($seq1, $j-1, 1);
    my $letter2 = substr($seq2, $i-1, 1);
    if ($letter1 eq $letter2) {
        $diagonal_score = $matrix[$i-1][$j-1]{score} + $MATCH;
    } else {
        $diagonal_score = $matrix[$i-1][$j-1]{score} + $MISMATCH;
    }
    # calculate gap scores
    $up_score = $matrix[$i-1][$j]{score} + $GAP;
    $left_score = $matrix[$i][$j-1]{score} + $GAP;
    if ($diagonal_score <= 0 and $up_score <= 0 and $left_score <= 0) {
        $matrix[$i][$j]{score} = 0;
        $matrix[$i][$j]{pointer} = "none";
        next; # terminate this iteration of the loop
    }
    # choose best score
    if ($diagonal_score >= $up_score) {
        if ($diagonal_score >= $left_score) {
            $matrix[$i][$j]{score} = $diagonal_score;
            $matrix[$i][$j]{pointer} = "diagonal";
        } else {
            $matrix[$i][$j]{score} = $left_score;
            $matrix[$i][$j]{pointer} = "left";
        }
    } else {
        if ($up_score >= $left_score) {
            $matrix[$i][$j]{score} = $up_score;
            $matrix[$i][$j]{pointer} = "up";
        } else {
            $matrix[$i][$j]{score} = $left_score;
        }
    }
}`
$matrix[$i][$j]{pointer} = "left";
}
# set maximum score
if ($matrix[$i][$j]{score} > $max_score)
{
    $max_i = $i;
    $max_j = $j;
    $max_score = $matrix[$i][$j]{score};
}
}
# trace
my $align1 = "";
my $align2 = "";
# start at last cell of matrix
my $j = length($seq1);
my $i = length($seq2);
while (1)
{
    last if $matrix[$i][$j]{pointer} eq "none"; # ends at
    first cell of matrix
    if ($matrix[$i][$j]{pointer} eq "diagonal")
    {
        $align1 .= substr($seq1, $j-1, 1);
        $align2 .= substr($seq2, $i-1, 1);
        $i--;
        $j--;
    }
    elsif ($matrix[$i][$j]{pointer} eq "left")
    {
        $align1 .= substr($seq1, $j-1, 1);
        $align2 .= "-";
        $j--;
    }
    elsif ($matrix[$i][$j]{pointer} eq "up")
    {
        $align1 .= "-";
        $align2 .= substr($seq2, $i-1, 1);
        $i--;
    }
    }
my $ak = length($seq1);
my $kara = length($seq2);
return $matrix[$kara][$ak]{score};
}
# write the distance/blast matrix on outfile
$s = "\t";
print OUTPUT "matrix\t@headers\n";
for (my $i = 0; $i <$#sequences + 1; $i++)
{
    # $#array_2d gives the highest index from the array
    print OUTPUT $headers[$i]."\t";
    for (my $j = 0; $j <$#sequences + 1; $j++)
    {
        print OUTPUT "$blastmatrix[$i][$j]\t";
    }
}
3. Perl Script to count 3-mers and build a feature matrix from FASTA files of miRNA sequences. The code is self-written.

```perl
use strict;
use warnings;
my @seq_array;
my @sequences;
my @headers;
my $output_file="output.txt";
open(OUTPUT,">".$output_file);
open (my $inFile, '<', 'hsa-fasta.txt') or die $!
while (<$inFile>)
{
    push(@seq_array,split /

s+/);
}
close ($inFile);
for (my$i=1; $i < $#seq_array+1; $i+=2) { push (@sequences, $seq_array[$i]) }; 
for (my$i=0; $i < $#seq_array; $i+=2) { push (@headers, $seq_array[$i]) };
my @kmer_array= qw/AAA AAT AAC AAG ATA ATT ATC ATG
ACA ACT ACC ACG AGA AGT AGC
AGG TAA TAT TAC TAG TTA TTT TTC TTG TCA
TCT TCC TCG TGA TGT TGC TGG TAA CAT
CAC CAG CTA CTT CTC CTG CCA CCT CCC CCG
CGA CGT CGC CGG GAA GAT GAC GAG GTA
GTT GTC GTG GCA GCT GCC GCG GGA GGT GGC
GGG/;
my @kmer_matrix;
$kmer_matrix[0][0]{score}= 0;
for my $i (0.. $#sequences){
    for my $j (0.. $#kmer_array){
        if ( substr($sequences[$i],0) =~ /$kmer_array[$j]/) {
            $kmer_matrix[$i][$j]{score}=1;
        }
    }
}
no warnings 'uninitialized';
#write the distance/blast matrix on outfile
"="t";
print OUTPUT "matrix\t@kmer_array\n";
for($i = 0; $i < $#headers+1; $i++)
{
}
```
4. Perl Script to calculate TAM percentage enrichments of the groupings. The code is self-written.
my $hitClu=0;
my $hitFunc=0;
my $hitFam=0;
my $hitHM =0;
my $hitTissue =0;
for (my $i=0; $i < $#mydata; $i+=8) { push (@category, $mydata[$i]) }
for (my $i=1; $i < $#mydata; $i+=8) { push (@term, $mydata[$i]) }
for (my $i=2; $i < $#mydata; $i+=8) { push (@count, $mydata[$i]) }
for (my $i=3; $i < $#mydata; $i+=8) { push (@percent, $mydata[$i]) }
for (my $i=4; $i < $#mydata; $i+=8) { push (@fold, $mydata[$i]) }
for (my $i=5; $i < $#mydata; $i+=8) { push (@pvalue, $mydata[$i]) }
for (my $i=6; $i < $#mydata; $i+=8) { push (@bonferroni, $mydata[$i]) }
for (my $i=7; $i < $#mydata; $i+=8) { push (@FDR, $mydata[$i]) }
for my $e(0..$#category)
{
    if ($category[$e] eq "Cluster")
    {
        if ( $count[$e]!=1 && $pvalue[$e] <$pvalue_Control && ($percent[$e]*$count[$e])>$percentage_Control)
        {
            $hitClu=1;
        }
    }
    if ($category[$e] eq "Function")
    {
        if ( $count[$e]!=1 && $pvalue[$e] <$pvalue_Control && ($percent[$e]*$count[$e])>$percentage_Control)
        {
            $hitFunc=1;
        }
    }
    if ($category[$e] eq "Family")
    {
        if ( $count[$e]!=1 && $pvalue[$e] <$pvalue_Control && ($percent[$e]*$count[$e])>$percentage_Control)
        {
            $hitFam=1;
        }
    }
    if ($category[$e] eq "HMDD")
    {
        if ( $count[$e]!=1 && $pvalue[$e] <$pvalue_Control && ($percent[$e]*$count[$e])>$percentage_Control)
        {
            $hitHM=1;
        }
    }
    if ($category[$e] eq "TissueSpecific")
    {
    }
if ( $count[$e]!=1 && $pvalue[$e] < $pvalue_Control && ($percent[$e]*$count[$e]) > $percentage_Control)
{
    $hitTissue=1;
}

if ($hitClu == 1){ $SumClust++;}
if ($hitFunc == 1){ $SumFunc++;}
if ($hitFam == 1){ $SumFam++;}
if ($hitHM == 1){ $SumHM++;}
if ($hitTissue == 1){ $SumTissue++;}
if ($hitHM == 1 || $hitClu == 1 || $hitFunc == 1 || $hitFam == 1 || $hitTissue == 1 ){ $SumAll++;}

my $CluP = ($SumClust*100/$groupnumber);
my $FuncP = ($SumFunc*100/$groupnumber);
my $FamP = ($SumFam*100/$groupnumber);
my $HMP = ($SumHM*100/$groupnumber);
my $TissueP= ($SumTissue*100/$groupnumber);
my $AllP = ($SumAll*100/$groupnumber);

print OUTPUT "Clust Per = ". $CluP. "\n";
print OUTPUT "Function Per =" .$FuncP."\n";
print OUTPUT "Family Per = ".$FamP."\n";
print OUTPUT "HMDD Per = ".$HMP."\n";
print OUTPUT "Tissue Per = ".$TissueP."\n";
print OUTPUT "All Per = ".$AllP."\n";

5. Perl script to mine targets in psRNATarget outputs. The code is self-written.

use strict;
use warnings;
open (IN, "infile.txt");
my $text;
{ local $/ =undef; $text =<IN>; }

my @a = split /
/,$text ;
close (IN);
open(OUT, '>', "outfile.txt") or die "Coul dn't open: \$!"; foreach (@a)
{
    if ($_ =~/Ribosomal protein/ |/Malate dehydrogenase/|/Polyalpylglutamate synthetase/|/Serine carboxypeptidase/ |
/Chalcone synthase/ |/Lipoxygenase/ |/ATPase/ |/Geranylgeranyl hydrogenase/ |/Prokaryal protein/ |
/Immunoglobulin lambda-like polypeptide/ |/acetyltransferase/ |
/zinc finger protein/ |/Proline-rich/ |
/Heat shock protein/ |/Plasma membrane/ |/Glutathione/ |/Ribulose bisphosphate carboxylase/|
/myosin heavy chain/ |/kinase/ |/Sugar transferase/ |
/Fusicoccadiene synthase/ |/Preprotein translocase secA subunit/ |
6. The perl code to read MCL clustering results. The code is self-written.
APPENDIX F: R CODES


```r
mydata=read.csv("matrix.csv", row.names=1)
mydata = na.omit(mydata)
mydata = scale(mydata)

wss <- (nrow(mydata)-1)*sum(apply(mydata,2,var))
for (i in 2:300) wss[i] <- sum(kmeans(mydata, centers=i)$withinss)
plot(1:300, wss, type="b", xlab="Number of Clusters",ylab="Within groups sum of squares")

fit <- kmeans(mydata, 30)
aggregate(mydata,by=list(fit$cluster),FUN=mean)

mydata2 <- data.frame(mydata, fit$cluster)
sink("C:\\ws\\kmeans-clusters.txt")
lapply(fit$cluster,print)
sink()
```

2. R code for CLAG analysis. The code is adapted from CLAG manuel and CLAG R package.

```r
mydata=read.csv("matrix.csv", header=TRUE)
library(CLAG)
M=mydata
RES=CLAG.clust(M, delta=0.05, threshold=0, analysisType=1, normalization="affine-global",rowId=row.names(M))
PCA <- prcomp(M)
clusterColors <- c("black", rainbow(RES$ncluster))
plot(PCA$x[,1], PCA$x[,2], col=clusterColors[RES$cluster+1], main=paste(RES$nclusters, "clusters"))
sink("clag-RES.txt")
lapply(RES,print)
sink()
sink("clag-clusters.txt")
lapply(RES$cluster,print)
sink()
```

3. R code for SOTA clustering. SOTA clustering is described under clValid R package (Brock, Pihur, Susmita, & Somnath, 2008). R functions are adapted from helpful R documentation.

```r
library(clValid)
```
mymatrix = read.csv("matrix.csv", header=TRUE)
sotaCl <- sota(as.matrix(mymatrix), 29)
plot(sotaCl, cl=6)
distan2 = dist(NW.dis)
dunn2 = dunn(distan2, sotaCl$clust, method = "euclidean")
sink("C:\\ws\\sota-clusters.txt")
lapply(sotaC2$clust, print)
sink()


```
add.one <- function (M) {
  for (i in 1:dim(M)[1]) {
    if (M[i,i] == 0) {
      M[i,i] <- M[i,i] + 1;
    }
  }
  return (M);
}
# Inflation step of MCL
inflate <- function (M, inf) {
  M <- M^(inf);
  return (M);
}
# Normalize the matrix by column
norm <- function (M) {
  colum.sum <- apply(M,2,sum)
  M <- t(M) / colum.sum
  return (t(M))
}
# MCL procedure
mcl <- function (M, # Matrix
  inf, # Inflation value
  iter, # Number of iterations
  verbose = F)
{
  for (i in 1:iter) {
    old.M <- M;
    M.norm <- norm(M);
    M <- M.norm%*%M.norm;
    M <- inflate(M, inf);
    M <- norm(M);
    if (sum(old.M == M) == dim(M)[1]*dim(M)[2]) {
      break;
    }
    if (verbose) {
      print (paste ("iteration", i));
    }
  }
}
```
return (M);

} # Matrix (mcl result)

collect.mcl.clusters <- function (M) {
  M.names <- row.names(M);
  clustered.nodes <- vector(mode = "logical", length = dim(M)[1])
  for (i in 1:dim(M)[1]) {
    nodes <- M.names[which(M[i,] != 0)];
    if (length(nodes) > 0 && !clustered.nodes[which(M[i,] != 0)]) {
      print (nodes);
      clustered.nodes[which(M[i,] != 0)] = T;
    }
  }
  return (clustered.nodes);
}

mcl.data=read.csv("matrix.csv", header=TRUE)
mcl.data <- as.matrix(mcl.data)
mcl.data=mcl.data*mcl.data
inf <- 4.0
mcl.clusters <- mcl(mcl.data,inf,200, verbose = T);
x=collect.mcl.clusters(mcl.clusters)
sink("MCL.txt")
lapply(mcl.clusters,print)
sink()
write.matrix(mcl.clusters, file = "MCL-clusters.txt", sep = " ")
APPENDIX G: CURRENT LIST OF PSRNATARGET LIBRARY

Organized as the organism, genome type, project name, version, release date.

<table>
<thead>
<tr>
<th>Organism</th>
<th>Genome Type</th>
<th>Project Name</th>
<th>Version</th>
<th>Release Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allium cepa (Onion)</td>
<td>unigene</td>
<td>DFCI Gene Index (ONGI)</td>
<td>2</td>
<td>2008_07_17</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>transcript</td>
<td>TAIR</td>
<td>10</td>
<td>2010_12_14</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>unigene</td>
<td>DFCI Gene Index (AGI)</td>
<td>15</td>
<td>2010_04_08</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>genomic DNA</td>
<td>3.4K segments from strand with 0.4K overlapped region</td>
<td>TAIR</td>
<td>released on 2004_01_22</td>
</tr>
<tr>
<td>Aquilegia (columbine)</td>
<td>unigene</td>
<td>DFCI Gene Index (AQGI)</td>
<td>2</td>
<td>2008_06_06</td>
</tr>
<tr>
<td>Beta vulgaris (beet)</td>
<td>unigene</td>
<td>DFCI Gene Index (BVGI)</td>
<td>4</td>
<td>2011_03_17</td>
</tr>
<tr>
<td>Brachypodium distachyon</td>
<td>transcript</td>
<td>JGI genomic project</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brachypodium distachyon</td>
<td>unigene</td>
<td>DFCI Gene Index (BDGI)</td>
<td>1</td>
<td>2010_05_26</td>
</tr>
<tr>
<td>Brassica napus (rape)</td>
<td>unigene</td>
<td>DFCI Gene Index (BNGI)</td>
<td>5</td>
<td>2011_03_18</td>
</tr>
<tr>
<td>Brassica rapa (turnip)</td>
<td>cds</td>
<td>de novo scaffolds assembly</td>
<td>1.1</td>
<td>released on 2011_08_30</td>
</tr>
<tr>
<td>Capsicum annuum</td>
<td>454</td>
<td>unigene</td>
<td>454</td>
<td></td>
</tr>
<tr>
<td>Capsicum annuum (Pepper)</td>
<td>unigene</td>
<td>DFCI Gene Index (CAGI)</td>
<td>4</td>
<td>2009_05_21</td>
</tr>
<tr>
<td>Carica papaya (Papaya)</td>
<td>unigene</td>
<td>DFCI Gene Index (CAPAGI)</td>
<td>1</td>
<td>2010_05_27</td>
</tr>
<tr>
<td>Chlamydomonas reinhardtii</td>
<td>unigene</td>
<td>DFCI Gene Index (CHRGI)</td>
<td>8</td>
<td>2011_03_21</td>
</tr>
<tr>
<td>Chlamydomonas reinhardtii</td>
<td>transcript</td>
<td>JGI genomic project</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chlamydomonas reinhardtii</td>
<td>mRNA</td>
<td>augustus gene model prediction, JGI-Assembly v4, augustus u9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cicer arietinum (chickpea)</td>
<td>transcriptome sequence, nipgr</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cicer arietinum (chickpea)</td>
<td>cds</td>
<td>ICGGC draft genome sequencing project</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Citrus clementina (Clementine)</td>
<td>unigene</td>
<td>DFCI Gene Index (CICLGI)</td>
<td>2</td>
<td>2009_05_22</td>
</tr>
<tr>
<td>Citrus sinensis (Orange)</td>
<td>unigene</td>
<td>DFCI Gene Index (CSGI)</td>
<td>1</td>
<td>2008_06_25</td>
</tr>
<tr>
<td>Coffea canephora (coffee)</td>
<td>unigene</td>
<td>SGN unigene, SGN</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Coffea canephora (coffee)</td>
<td>unigene</td>
<td>DFCI Gene Index (COCAGI)</td>
<td>3</td>
<td>2011_03_18</td>
</tr>
<tr>
<td>Cucumis sativus (cucumber)</td>
<td>cds</td>
<td>Cucumber genome sequencing project</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Cuscuta pentagona</td>
<td>unigene</td>
<td>Cuscuta pentagona, de-novo assembly of NGS RNA-seq reads, Sinha Lab, Department of Plant Biology, Life Sciences Addition, UC Davis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ectocarpus siliculosus (Brown algae)</td>
<td>unigene</td>
<td>DFCI Gene Index (BAGI)</td>
<td>1</td>
<td>2010_05_27</td>
</tr>
<tr>
<td>Eucalyptus grandis</td>
<td>(Flooded gum or Rose gum)</td>
<td>transcript, JGI genomic project, Phytozome, phytozome v8.0, internal number 201</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eucalyptus grandis</td>
<td>EST contig</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Euphorbia esula (Leafy spurge)</td>
<td>unigene</td>
<td>DFCI Gene Index (EUESGI)</td>
<td>1</td>
<td>2008_06_30</td>
</tr>
<tr>
<td>Festuca arundinacea (tall fescue)</td>
<td>unigene</td>
<td>DFCI Gene Index (FAGI)</td>
<td>3</td>
<td>2010_04_07</td>
</tr>
<tr>
<td>Festuca pratensis (Meadow ryegrass)</td>
<td>unigene</td>
<td>DFCI Gene Index(MRGI)</td>
<td>1</td>
<td>2008_06_06</td>
</tr>
<tr>
<td>Date</td>
<td>Organism/Microorganism</td>
<td>Type</td>
<td>Source/Database</td>
<td>Version</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------------------------</td>
<td>---------------</td>
<td>----------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>2010_05_28</td>
<td>Fusarium oxysporum (Fusarium oxysporum)</td>
<td>Transcript</td>
<td>Broad Institute</td>
<td>4287</td>
</tr>
<tr>
<td>2011_03_31</td>
<td>Glycine max (soybean)</td>
<td>Unigene</td>
<td>DFCI Gene Index (GMGI)</td>
<td>16</td>
</tr>
<tr>
<td>2011_03_31</td>
<td>Glycine max (soybean)</td>
<td>Transcript</td>
<td>JGI genomic project</td>
<td>v8.0</td>
</tr>
<tr>
<td>2011_03_22</td>
<td>Gossypium (cotton)</td>
<td>Unigene</td>
<td>DFCI Gene Index (CGI)</td>
<td>11</td>
</tr>
<tr>
<td>2011_03_22</td>
<td>Gossypium raimondii (cotton_raimondii)</td>
<td>Transcript</td>
<td>JGI genomic project</td>
<td>v8.0</td>
</tr>
<tr>
<td>2011_03_28</td>
<td>Gossypium raimondii (cotton_raimondii)</td>
<td>Transcript</td>
<td>JGI genomic project</td>
<td>v8.0</td>
</tr>
<tr>
<td>2008_07_02</td>
<td>Gossypium raimondii (cotton_raimondii)</td>
<td>Transcript</td>
<td>JGI genomic project</td>
<td>v8.0</td>
</tr>
<tr>
<td>2008_06_30</td>
<td>Ipomoea nil (Morning glory)</td>
<td>Unigene</td>
<td>DFCI Gene Index (IPNIGI)</td>
<td>1</td>
</tr>
<tr>
<td>2008_07_02</td>
<td>Lactuca serriola (Prickly lettuce)</td>
<td>Unigene</td>
<td>DFCI Gene Index (LASEGI)</td>
<td>1</td>
</tr>
<tr>
<td>2008_06_28</td>
<td>Lactuca serriola (Prickly lettuce)</td>
<td>Transcript</td>
<td>JGI genomic project</td>
<td>v7.0</td>
</tr>
<tr>
<td>2010_05_24</td>
<td>Helianthus annuus (Sunflower)</td>
<td>Unigene</td>
<td>DFCI Gene Index (HAGI)</td>
<td>6</td>
</tr>
<tr>
<td>2011_03_19</td>
<td>Hordeum vulgare (barley)</td>
<td>Unigene</td>
<td>DFCI Gene Index (HVGI)</td>
<td>12</td>
</tr>
<tr>
<td>2008_06_30</td>
<td>Ipomoea nil (Morning glory)</td>
<td>Unigene</td>
<td>DFCI Gene Index (IPNIGI)</td>
<td>1</td>
</tr>
<tr>
<td>2008_07_02</td>
<td>Lactuca serriola (Prickly lettuce)</td>
<td>Unigene</td>
<td>DFCI Gene Index (LASEGI)</td>
<td>1</td>
</tr>
<tr>
<td>2008_07_28</td>
<td>Linum usitatissimum (flax or linseed)</td>
<td>Unigene</td>
<td>DFCI Gene Index (LUSGI)</td>
<td>200</td>
</tr>
<tr>
<td>2010_05_18</td>
<td>Magnaporthe oryzae (rice blast fungus)</td>
<td>Transcript</td>
<td>Broad institute, Magnaporthe grisea Database</td>
<td></td>
</tr>
<tr>
<td>2010_04_08</td>
<td>Malus x domestica (apple)</td>
<td>CDS</td>
<td>Predicted consensus gene set</td>
<td>200</td>
</tr>
<tr>
<td>2010_04_08</td>
<td>Malus x domestica (apple)</td>
<td>Transcript</td>
<td>JGI genomic project</td>
<td>v7.0</td>
</tr>
<tr>
<td>2010_05_28</td>
<td>Manihot esculenta (Cassava)</td>
<td>Unigene</td>
<td>DFCI Gene Index (MAESGI)</td>
<td>1</td>
</tr>
<tr>
<td>2010_05_28</td>
<td>Medicago truncatula (Barrel Medic)</td>
<td>Transcript</td>
<td>Mt3.5v4 splice transcripts</td>
<td></td>
</tr>
<tr>
<td>2010_05_28</td>
<td>Medicago truncatula (Barrel Medic)</td>
<td>Transcript</td>
<td>Mt4.0v1 splice transcripts</td>
<td>IMGAG, Mt4.0V1</td>
</tr>
<tr>
<td>2010_05_28</td>
<td>Mesembryanthemum crystallinum (ice plant)</td>
<td>Unigene</td>
<td>DFCI Gene Index (MCGI)</td>
<td>5</td>
</tr>
<tr>
<td>2010_05_28</td>
<td>Musa acuminata (banana)</td>
<td>Unigene</td>
<td>DFCI Gene Index (NTGI)</td>
<td>7</td>
</tr>
<tr>
<td>2010_05_28</td>
<td>Nicotiana tabacum (tobacco)</td>
<td>Transcript</td>
<td>JGI genomic project</td>
<td>v9.0</td>
</tr>
<tr>
<td>2010_05_28</td>
<td>Oryza sativa (rice)</td>
<td>Unigene</td>
<td>TIGR genome cDNA OSA1 Release 5(OSA1R5)</td>
<td>5</td>
</tr>
<tr>
<td>2011_03_28</td>
<td>Oryza sativa (rice)</td>
<td>Unigene</td>
<td>DFCI Gene Index (OSGI)</td>
<td>19</td>
</tr>
<tr>
<td>2011_03_28</td>
<td>Oryza sativa (rice)</td>
<td>Genomic DNA</td>
<td>3.4K segments from strand with 0.4K overlapped region</td>
<td></td>
</tr>
<tr>
<td>2011_03_24</td>
<td>Oryza sativa (rice)</td>
<td>Transcript</td>
<td>JGI genomic project</td>
<td>v7.0</td>
</tr>
<tr>
<td>2011_03_24</td>
<td>Oryza sativa (rice)</td>
<td>Transcript</td>
<td>JGI genomic project</td>
<td>v7.0</td>
</tr>
<tr>
<td>2011_03_24</td>
<td>Panicum virgatum (switchgrass)</td>
<td>Unigene</td>
<td>DFCI Gene Index (PAVIGI)</td>
<td>2</td>
</tr>
</tbody>
</table>

92
<table>
<thead>
<tr>
<th>Species</th>
<th>Type</th>
<th>Source</th>
<th>Version</th>
<th>Release Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Petunia hybrida (Petunia)</td>
<td>unigene</td>
<td>DFCI Gene Index (PHGI)</td>
<td>3</td>
<td>2011_03_17</td>
</tr>
<tr>
<td>Phaseolus coccineus (Scarlet bean)</td>
<td>unigene</td>
<td>DFCI Gene Index (PCGI)</td>
<td>1</td>
<td>2009_05_27</td>
</tr>
<tr>
<td>Phaseolus vulgaris (bean)</td>
<td>unigene</td>
<td>DFCI Gene Index (PHVGI)</td>
<td>4</td>
<td>2011_03_24</td>
</tr>
<tr>
<td>Phoenix dactylifera (date palm)</td>
<td>transcript</td>
<td>Date Palm Genome Draft Assembly, Version 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phyllostachys heterocycla (Moso Bamboo)</td>
<td>cds</td>
<td>National Center for Gene Research of Chinese Academy of Sciences, Version 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physcomitrella patens (moss)</td>
<td>transcript</td>
<td>JGI genomic project, Phytozome, phytozome v7.0, internal number 152</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physcomitrella patens (moss)</td>
<td>unigene</td>
<td>DFCI Gene Index (PPSPGI)</td>
<td>1</td>
<td>2009_05_26</td>
</tr>
<tr>
<td>Picea (Spruce)</td>
<td>unigene</td>
<td>DFCI Gene Index (SGI)</td>
<td>5</td>
<td>2011_03_30</td>
</tr>
<tr>
<td>Pinus (pine)</td>
<td>unigene</td>
<td>DFGI Gene Index (PGI)</td>
<td>9</td>
<td>2011_03_26</td>
</tr>
<tr>
<td>Populus trichocarpa (poplar)</td>
<td>genomic DNA</td>
<td>Genome Assembly R1.0, 3.4K segments from strand with 0.4K overlapped regions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Populus trichocarpa (poplar)</td>
<td>transcript</td>
<td>JGI genomic project, Phytozome, phytozome v9.0, internal number 218</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Populus trichocarpa (poplar)</td>
<td>transcript</td>
<td>JGI genomic project, Phytozome, phytozome v8.0, genome V3.0, internal number 210</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Populus trichocarpa (poplar)</td>
<td>transcript</td>
<td>JGI genomic project, Phytozome, phytozome v7.0, internal number 156</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prunus persica (peach)</td>
<td>unigene</td>
<td>DFCI Gene Index (PRPEGI)</td>
<td>2</td>
<td>2009_05_26</td>
</tr>
<tr>
<td>Quercus (oak)</td>
<td>unigene</td>
<td>DFCI Gene Index (OGI)</td>
<td>2</td>
<td>2011_03_23</td>
</tr>
<tr>
<td>Raphanus sativus (Radish)</td>
<td>unigene</td>
<td>DFCI Gene Index (RSGI)</td>
<td>1</td>
<td>2010_05_27</td>
</tr>
<tr>
<td>Saccharum officinarum (sugarcane)</td>
<td>unigene</td>
<td>DFCI Gene Index (SOGI)</td>
<td>3</td>
<td>2010_04_09</td>
</tr>
<tr>
<td>Secale cereale (rye)</td>
<td>unigene</td>
<td>DFCI Gene Index (RYEGI)</td>
<td>4</td>
<td>2008_07_03</td>
</tr>
<tr>
<td>Selaginella moellendorfii (club mosses)</td>
<td>unigene</td>
<td>DFCI Gene Index (CMGI)</td>
<td>1</td>
<td>2010_05_26</td>
</tr>
<tr>
<td>Setaria italica (foxtail millet)</td>
<td>cds</td>
<td>BGI sequencing project,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Setaria italica (foxtail millet)</td>
<td>cds</td>
<td>JGI genomic project, phytozome v9.0, internal number 164</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Solanum Lycopersicon (tomato Lycopersicon)</td>
<td>unigene</td>
<td>Lycopersicon Combined (Tomato) Unigenes, SGN Unigene, unknown version</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Solanum Lycopersicon (tomato Lycopersicon subgenus)</td>
<td>transcript</td>
<td>SGN mRNA sequences, SGN genomic sequencing project, released on 2011_05_08</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Solanum melongena (eggplant)</td>
<td>unigene</td>
<td>DFCI Gene Index (SOMEVI)</td>
<td>1</td>
<td>2010_04_13</td>
</tr>
<tr>
<td>Solanum tuberosum (potato)</td>
<td>unigene</td>
<td>DFCI Gene Index (STGI)</td>
<td>13</td>
<td>2010_05_28</td>
</tr>
<tr>
<td>Sorghum bicolor (Sorghum)</td>
<td>unigene</td>
<td>DFCI Gene Index (SBGI)</td>
<td>9</td>
<td>2008_07_25</td>
</tr>
<tr>
<td>Striga hermonthica (Purple witchweed)</td>
<td>unigene</td>
<td>DFCI Gene Index (SHGI)</td>
<td>1</td>
<td>2010_05_27</td>
</tr>
<tr>
<td>Thellungiella halophila (salt cress)</td>
<td>transcript</td>
<td>JGI genomic project, Phytozome, phytozome v9.0, internal number 173</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thellungiella halophila (salt cress)</td>
<td>unigene</td>
<td>ESTContig on PlantGDB, PlangGDB unigene,</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

93
<table>
<thead>
<tr>
<th><strong>Species</strong></th>
<th><strong>Unigene</strong></th>
<th><strong>DFCI Gene Index</strong></th>
<th><strong>Version</strong></th>
<th><strong>Release Date</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Theobroma cacao (cocoa)</td>
<td>unigene</td>
<td>TCAGI</td>
<td>version 3</td>
<td>2009_05_21</td>
</tr>
<tr>
<td>Triphysaria</td>
<td>unigene</td>
<td>TRIPHGI</td>
<td>version 1</td>
<td>2008_08_15</td>
</tr>
<tr>
<td>Triphysaria versicolor</td>
<td>unigene</td>
<td>TVERGI</td>
<td>version 2</td>
<td>2008_06_28</td>
</tr>
<tr>
<td>Triticum aestivum (wheat)</td>
<td>unigene</td>
<td>TAGI</td>
<td>version 12</td>
<td>2010_04_18</td>
</tr>
<tr>
<td>Vigna unguiculata (cowpea)</td>
<td>unigene</td>
<td>VUGI</td>
<td>version 1</td>
<td>2010_05_27</td>
</tr>
<tr>
<td>Vitis vinifera (wine grape)</td>
<td>transcript</td>
<td>V1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vitis vinifera (grape)</td>
<td>unigene</td>
<td>VVGI</td>
<td>version 8</td>
<td>2011_04_02</td>
</tr>
<tr>
<td>Vitis vinifera (grape)</td>
<td>transcript</td>
<td>JGI</td>
<td></td>
<td>145</td>
</tr>
<tr>
<td>Volvox carteri (green algae)</td>
<td>unigene</td>
<td>VCGI</td>
<td>version 1</td>
<td>2010_05_28</td>
</tr>
<tr>
<td>Zea mays (maize)</td>
<td>unigene</td>
<td>ZMGI</td>
<td>version 19</td>
<td>2009_06_05</td>
</tr>
<tr>
<td>Zea mays (maize), cds</td>
<td>PlangGDB</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zea mays (maize), transcript</td>
<td>NSF-funded</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Maize Genome</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sequencing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Project</td>
<td></td>
<td></td>
<td>Release 5a</td>
</tr>
<tr>
<td></td>
<td>filtered set</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>