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 Narcı** 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 Oğul
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

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

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 enrole 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 MİLRNALARI 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

Mikro RNA'lar (miRNA) bir hücrenin transkripsiyon makinesini kontrol eden küçük kodlanmamış RNA molekülleridir. miRNA'lar 20-22 nükleotit uzunluğunda RNA'lar olup hedeflendikleri mRNA'ları susturarak kontrol ederler. miRNA dizilerinin ve işlevlerinin benzerliği gözlemlenmiştir. Bunlardan bazıları da aynı anda sentezlenerek aynı fonksiyonel olaylarda rol alırlar. Son yillarda yapılan miRNA gruplama analizlerinde bunların sadece ifade paternleri kullanılmaktadır. Bu tür analizler miRNA dizileri arasında işlevlerinin saptanması hala araştırmaların odak noktasıdır. Bununla beraber, bazı işlevsel miRNA grupları arasında önemli sekans benzerlikleri ve hatta miRNA dizileri, hedef mRNA RISC (RNA-indüklenmiş susturma kompleksi) bağlanma bölgelerinin algılamasında oldukça önemli olduğu için, sekans benzerliğine bakılarak benzer mRNA bölgelerine hedeflenmiş yeni miRNA grupları bulunabilir. miRNA dizilim benzerliğinden bağımsız olarak dizilimin ölçütlenmesi ile miRNA sekansları arasında önemli benzerlikler bulunabilir. Bu yöntem ile olgun miRNA'lara ait dizilim ölçütleri uygun matrislere yerleştirilerek farklı kümeleme analizlerine tabi tutulabilir. Bu çalışmada işlevsel olarak anlamlı miRNA gruplarını bulmaya yönelik olarak sadece olgun miRNA sekanslarına dayalı olan geliştirdiğimiz yöntem sunulmaktadır. Bu amaçla, birbirinden farklı bazı makine öğrenimi algoritmaları değişik sekans gösterim şemalarıyla kullanılmıştır. Bununla birlikte, geliştirdiğimiz bu yöntem kullanilarak laboratuvarmızda deneysel olarak saptanmış iki bitkisel patojenin ve bir simbiyotik fungusun küçük RNA diziliminden elde edilen veriler analiz edilmiştir. Analizler yeni tanımlanan miRNA dizilerinin işlevsel benzerliklerine dair gruplar bulunmasini

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 brialliant 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; Bayantes Dagvadorj, Adnan Yaramış, Bahtiyar Yılmaz, and Sait Erdoğan for their frendship. I am also very thankful to my dearest friends and labmates Çağlar Özketen and Ayşe Andaç for prepearing 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 frıendship 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

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

miRNA: micro RNA

UTR: Untranslated Region

RISC: RNA-induced silencing complex

MilRNA: miRNA-like small RNA

RNAi: RNA interference

NcRNA:Non-Coding RNA

NcRNA-Seq: Next Generation Sequencing

NextGen-Seq: Next Generation Sequencing

SNP : Single Nucleotide Polymorhpism

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 (Zymañski, 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 polycistronicly 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 machinary is realized, scientists are directed to discover many small RNA seqeuences. The process for the expolaration 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 seqeunces 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 (milRNA) sequences of the plant pathogenic fungi; *Puccinia striiformis* f. sp. *tritici* and *Blumeria graminis* f. sp. *hordei*, and the symbiotic fungus *Piriformospora indica*. The pipeline established was applied into the candidate milRNA sets. Additionally, plant mRNAs of fungal milRNAs 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 milRNA 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 biomoleculer 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).

$$F(i,j) = max \begin{cases} F(i-1,j-1) + s(x_i, y_j), \\ F(i-1,j) - d, \\ F(i,j-1) - d. \end{cases}$$

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).

$$F(i,j) = max \begin{cases} 0, \\ F(i-1,j-1) + s(x_i, y_j), \\ F(i-1,j) - d, \\ F(i,j-1) - d. \end{cases}$$

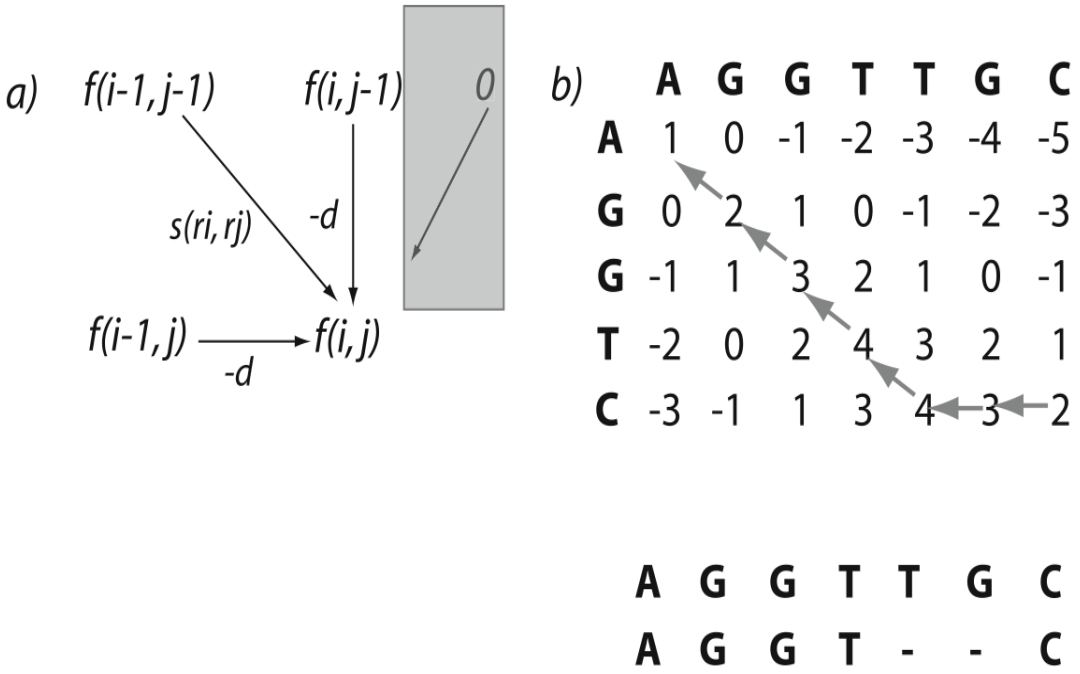


Figure 1. Dynamic programming representation. **a)** Recursively calculating each entities of the matrix by Needleman-Wunsch algorithm. –d represents the score for a gap, s(ri,rj) stands for the score match or mismatch corresponds to the entity. The grey box is additional requirement for Smith-Waterman extension (Equation 1 and 2). **b)** An illustration of alignment matrix by using Needleman-Wunsch algorithm, back tracking of the matrix and the best alignment results from the matrix. The final cell is where the back-track starts, and it ends with the first cell. The arrows show the direction for best-scoring alignment. Figure is directly taken from (http://www.hiv.lanl.gov/content/sequence/HIV/REVIEWS/2006_7/ABECASIS/abecasis.html).

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, Δ and the threshold for environmental and symmetric scores, clustered amount of data points is arranged. Parameter Δ 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.

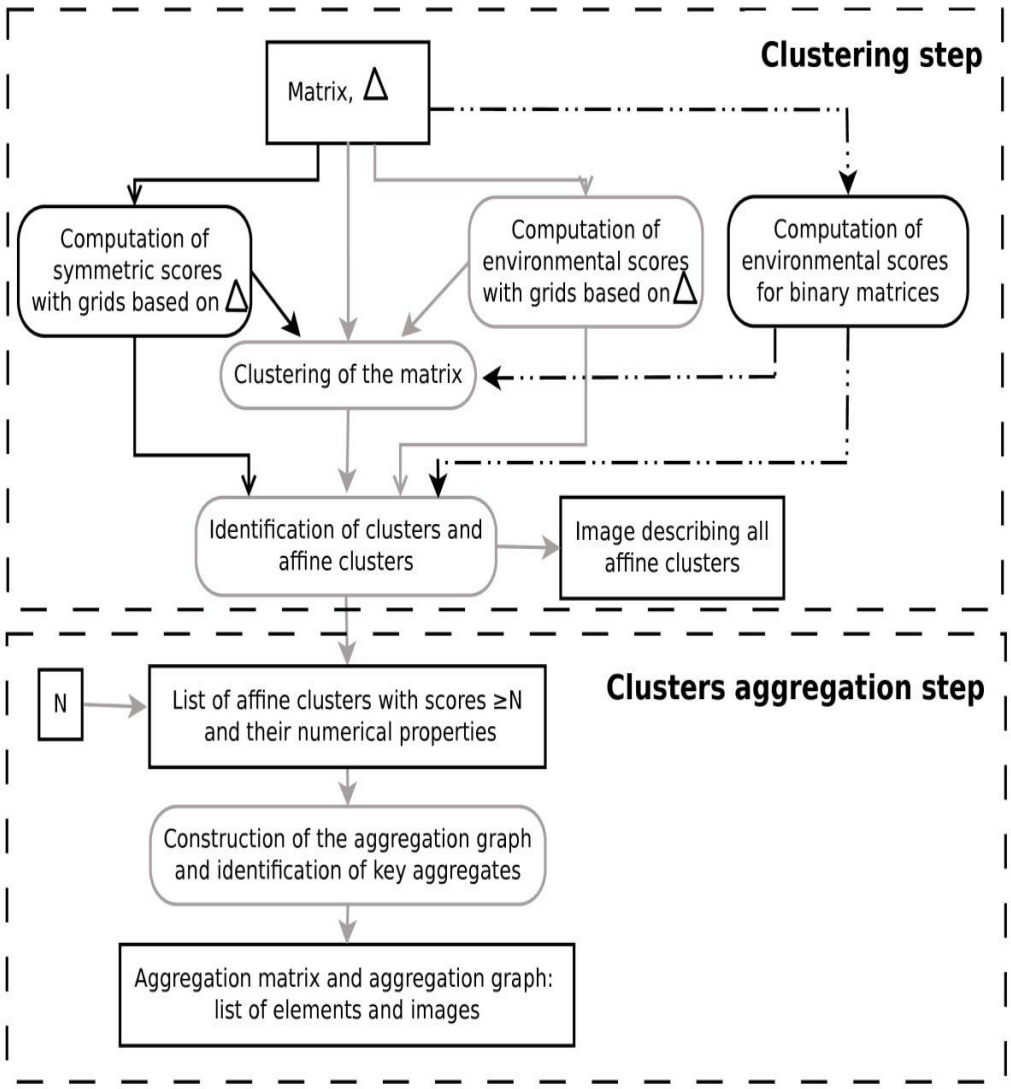


Figure 2. CLAG flowchart. Representation of the different stages of the algorithm is shown. CLAG has three inputs from the user; Matrix, Δ and the scores threshold N. Environmental score differs between the feature matrix (solid line) and symmetric matrix. Directly taken from (Dib & Carbone, 2012).

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 Stijin 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 matric 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 matric 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 matric 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, mammalians, 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).

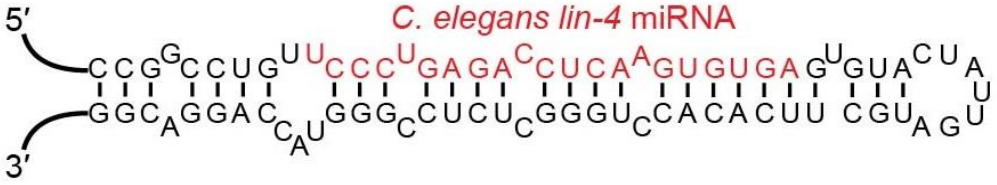


Figure 3. Stem loop of *C. elegans* lin-4 miRNA. Hairpin structure of miRNA lin-4 is represented with 5`and 3`overhanges. The red sequence represents mature miRNA. Adapted from (D. P. Bartel, 2004).

The other functionally important miRNA is *let-7,* which was found controling differentiation in *C. elegans*. The *let-7* was 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).

let-7a-1	UGAGGUAGUAGGUUGUAUAGUU
let-7a-2	UGAGGUAGUAGGUUGUAUAGUU
let-7a-3	UGAGGUAGUAGGUUGUAUAGUU
let-7b	UGAGGUAGUAGGUUGUGUGGUU
let-7c	UGAGGUAGUAGGUUGUAUGGUU
let-7d	AGAGGUAGUAGGUUGCAUAGU-
let-7e	UGAGUUAGGAGGUUGUAUAGU-
let-7f-1	UGAGUUAGUAGAUUGUAUAGUU
let-7f-2	UGAGUUAGUAGAUUGUAUAGUU
let-7g	UGAGUUAGUAGUUUGUACAGU-
let-7i	UGAGUUAGUAGUUUGUGCUGU-

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).

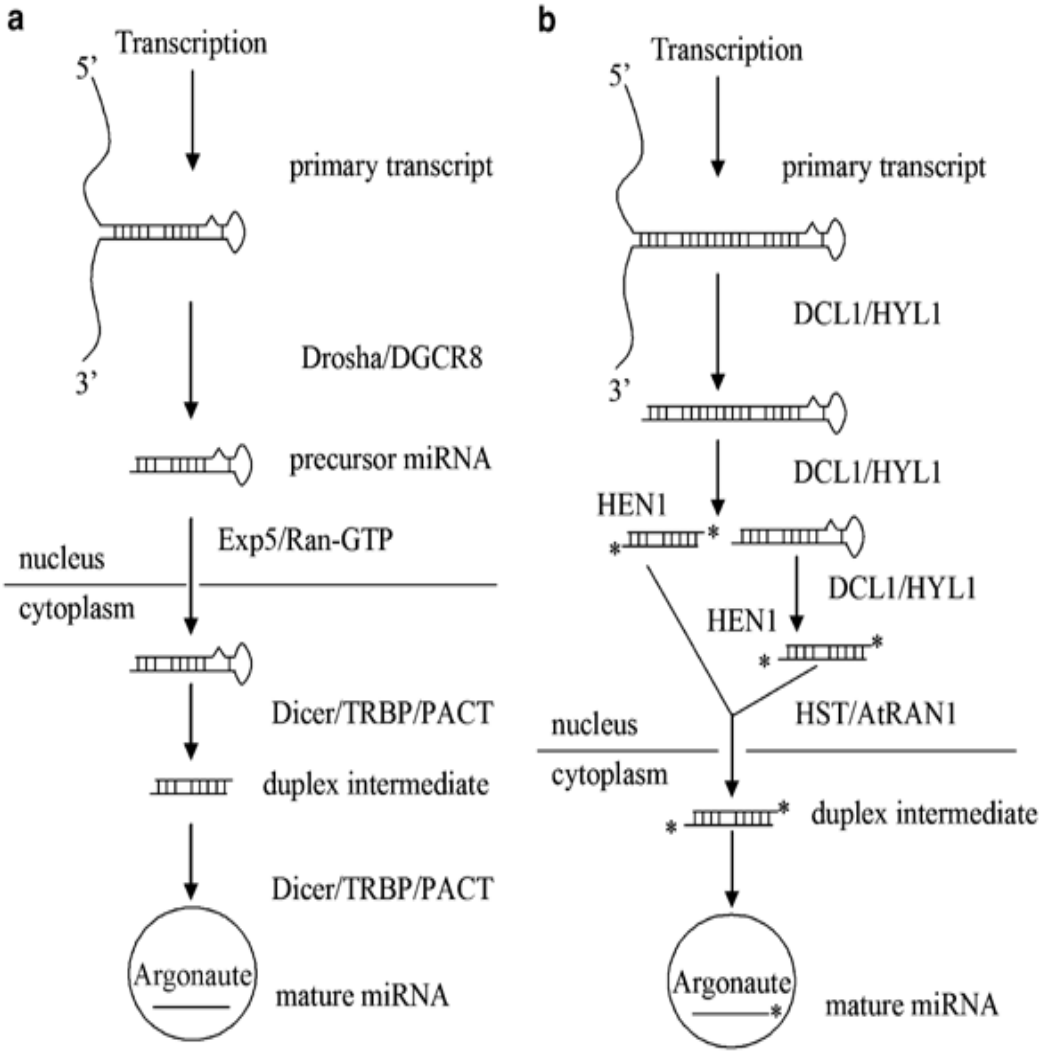


Figure 5. Plant and animal miRNA biogenesis pathways. **a)** miRNA biogenesis pathway in animals **b)** in plants. MIR gene is transcribed by PolII and pri-miR which is attached by cap proteins and polA tail is formed in the nucleus. Pri-miR is transformed into pre-miRNA after 1st processing and miRNA-miRNA* duplex is formed after 2nd 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 varios species are being discovered, a system of nomenclature was required in a systematic manner thus improved nomenclaturing is progressively established to coverall 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-1which 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 choromosmal 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, offen, 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 ofen 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) machinary by interfeering 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 corressponding genes in the genome in relatively inexpensive way. However, microarray analysis offen 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 filteration. The output is further mapped into the refarance 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 estalished 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 quantitiy 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 enhanged 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).

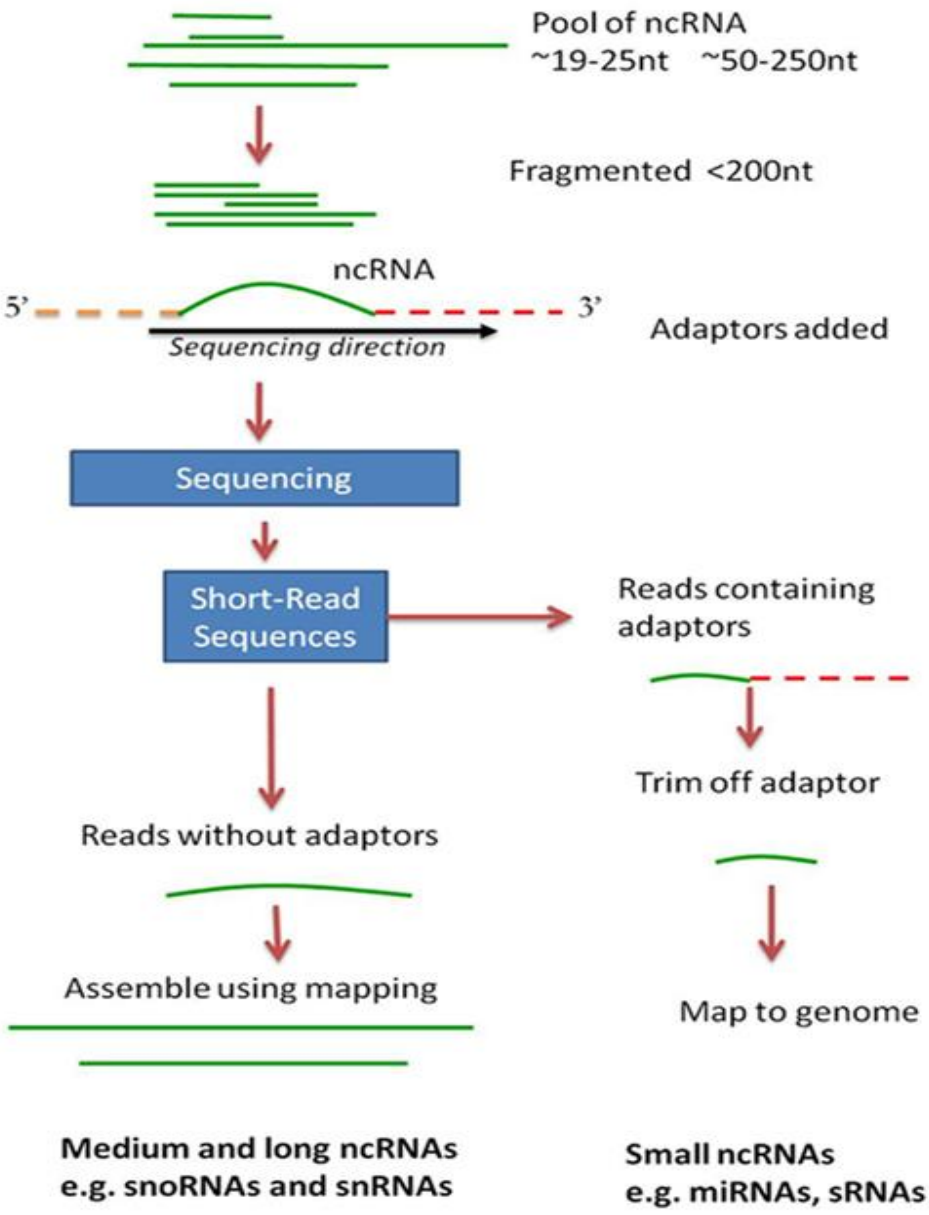


Figure 6. A general approach for using high-throughput sequencing data in search of small RNAs. Adaptor sequences are added into the pool of total RNA containing ncRNA sequences. Sequencing data will produce ncRNA sequnces with adaptor sequence, and long RNA sequences without adaptor. When adaptor sequences 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 varios 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 paracitisim. 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 biotropic 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 (*Triticulum aestivum L.*), but it can affect some cultivated barleys (*Hordeum vulgare L.*) without any serious epidemic (Chen et al., 2000).

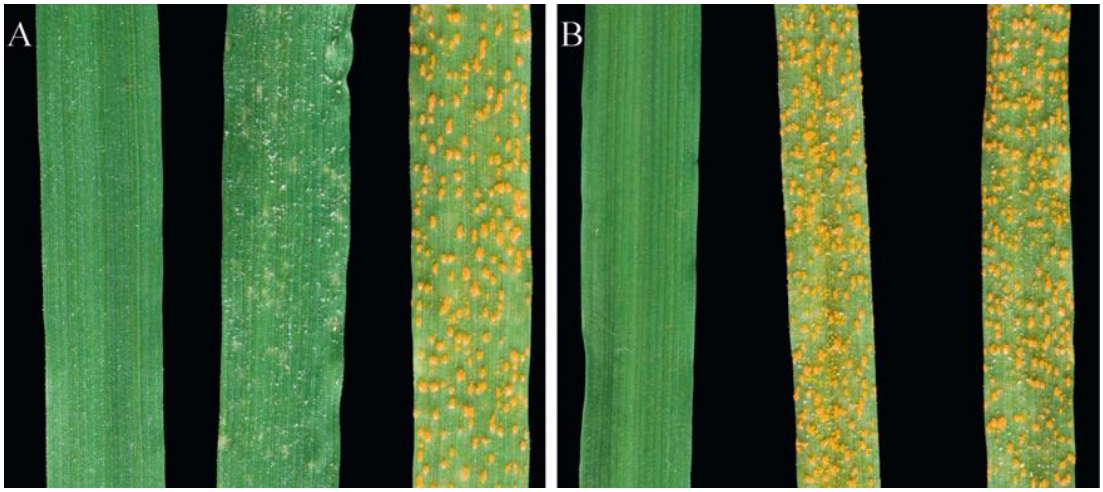


Figure 7. Wheat yellow rust. Figure is directly taken from (Bozkurt, Mcgrann, Maccormack, Boyd, & Akkaya, 2010)**.**

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).

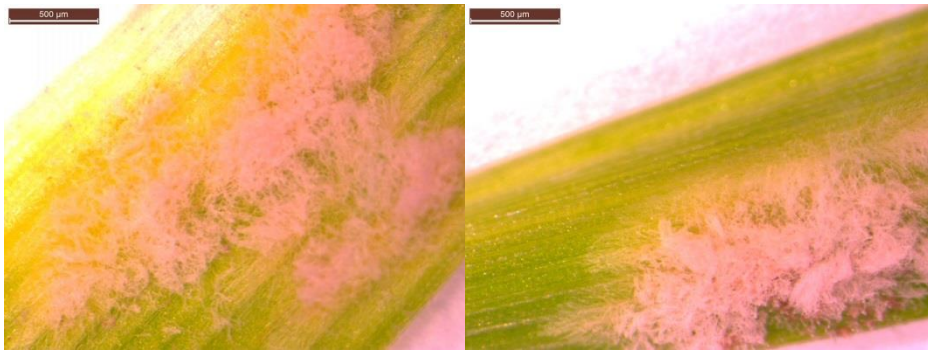


Figure 8. Powdery Mildew. Photos are taken by A. Çağlar Özketen, with Leica microscope systems.

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 milRNAs.

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 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 Bejing 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 anotate 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 1. Fungi referance genome descriptions used in sequencing

Fungi	Reference genome description
Blumeria graminis f. sp. hordei	Blumeria graminis f. sp. hordei DH14, whole genome sequence, sequenced by Blumeria Genome Consortium (Bioproject : PRJNA28821)
Piriformospora indica	Piriformospora indica DSM 11827, whole genome sequence, (Bioproject: PRJEA76339) .Sequencing was performed by MWG (Germany)
Puccinia striiformis f. sp. tritici	2K41-Yr9 isolate, race PST-78, a Great Plains isolate from 2000 (Bioproject : PRJNA123765) sequenced by Broad Institute

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 Gap=-1, Mismatch=-1, and 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 constract 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 matrixes were generated. For that purpose Perl rand function is used to assign each of the matrixes. 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 matric 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 4k, 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 4k 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 constract 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 algoritms

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 4k 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.

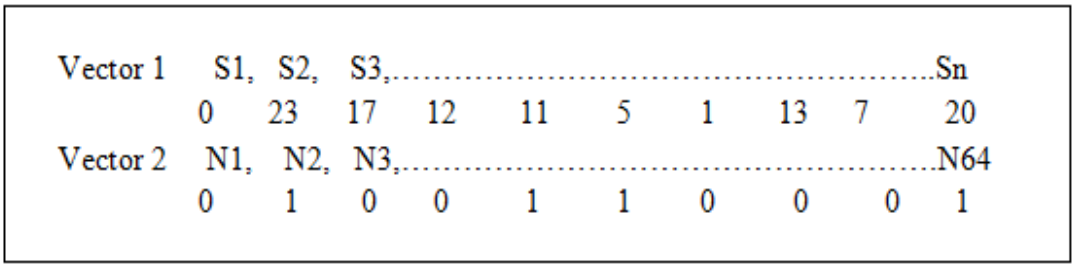


Figure 9. Vector representations of a miRNA sequence. Vector 1 is the object definition for a similarity alignment, all-to-all type of matrix. Vector 2 is a vector defined for k-mer distribution matrix.

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.

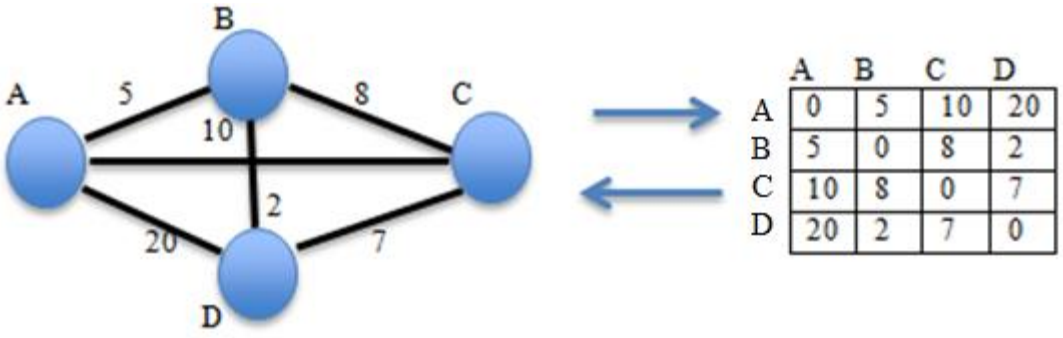


Figure 10. Transition of a graph to a matrix by MCL algorithm. Example of a distance graph for four objects (A-D), circles represents nodes (objects) and lines represent distances. Weighted transition matrix is shown generated from the graph.

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.

Enrichment analysis results

Text file of results

Term	Count	Percent	Fold	P-value
Category: Cluster (2 Items)				
hsa-mir-106b cluster	2	0.67	33.2593	1.06e-3
hsa-mir-17 cluster	2	0.33	16.6296	5.15e-3
Category: Family (3 Items)				
mir-17 family	2	0.25	12.4722	9.41e-3
mir-19 family	1	0.5	24.9444	0.0397
mir-25 family	1	0.33	16.6296	0.0591
Category: Function (15 Items)				
Akt pathway	2	0.12	5.8693	0.0416
Angiogenesis	2	0.08	4.1574	0.0784
Apoptosis	2	0.05	2.2677	0.2174
Bone regeneration	3	0.09	4.402	0.0244
Cell cycle related	3	0.05	2.2677	0.1325
Cell proliferation	1	0.04	1.7817	0.4428
Chemosensitivity of tumor cells	1	0.25	12.4722	0.0781
HIV latency	4	0.19	9.5026	3.86e-4
Hormones regulation	5	0.08	4.0233	3.49e-3

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 showen in Appendix E.

2.2.6 milRNA 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 milRNAs are targeting plant genomes; we have decided to use this server to find targets of milRNAs 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 releasesis 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.

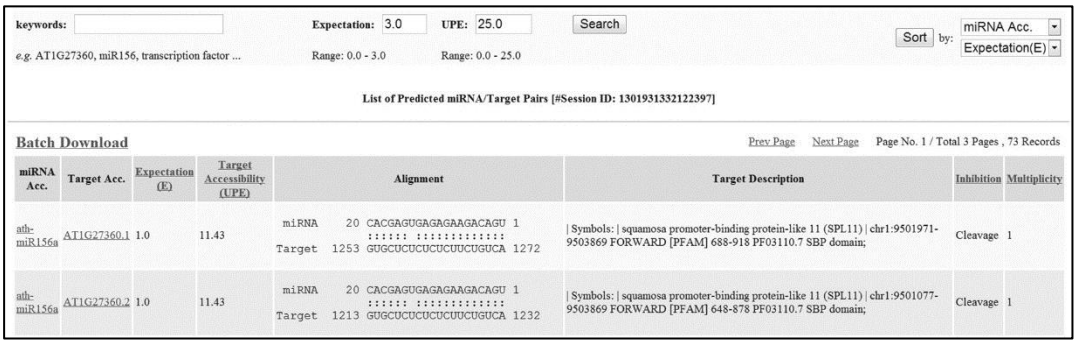
Table 2. psRNATarget tool functions.

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

Table 3. Libraries for the organisms selected from psRNATarget catalogue

Fungi	Target Organism; Libraries selected
Bgh	*Triticum aestivum* (wheat) by DFCI Gene Index (TAGI)
	Hordeum vulgare (barley) by DFCI Gene Index (HVGI)
Pst	*Triticum aestivum* (wheat) by DFCI Gene Index (TAGI)
	Hordeum vulgare (barley) by DFCI Gene Index (HVGI)
Pi	All of the library except Human dataset*

*psRNATarget library list is presented in Appendix G



keywords: Expectation: 3.0 UPE: 25.0 Search Sort by: miRNA Acc. / Expectation(E)

e.g. AT1G27360, miR156, transcription factor ... Range: 0.0 - 3.0 Range: 0.0 - 25.0

List of Predicted miRNA/Target Pairs [#Session ID: 1301931332122397]

Batch Download Prev Page Next Page Page No. 1 / Total 3 Pages , 73 Records

miRNA Acc.	Target Acc.	Expectation (E)	Target Accessibility (UPE)	Alignment	Target Description	Inhibition	Multiplicity
ath-miR156a	AT1G27360.1	1.0	11.43	miRNA 20 CACGAGUGAGAGAAGACAGU 1 :::::: :::::::::::::: Target 1253 GUGCUCUCUCUCUUCUGUCA 1272	| Symbols: | squamosa promoter-binding protein-like 11 (SPL11) | chr1:9501971-9503869 FORWARD [PFAM] 688-918 PF03110.7 SBP domain;	Cleavage	1
ath-miR156a	AT1G27360.2	1.0	11.43	miRNA 20 CACGAGUGAGAGAAGACAGU 1 :::::: :::::::::::::: Target 1213 GUGCUCUCUCUCUUCUGUCA 1232	| Symbols: | squamosa promoter-binding protein-like 11 (SPL11) | chr1:9501077-9503869 FORWARD [PFAM] 648-878 PF03110.7 SBP domain;	Cleavage	1

Figure 13. psRNATarget tool result page illustration. Example execution of the psRNATarget is shown. Figure is adapted from Dai & Zhao, 2011.

2.2.6.2 miRBase database

MiRBase is the largest database for the glossary of published miRNA sequences. Currently, there are thousands of miRNA entiries 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 milRNAs by the pipeline

Fungal (*Puccinia striiformis* f. sp. *tritici* (Pst), *Blumeria graminis* f. sp. *hordei* (Bgh) and *Piriformospora indica* (Pi)) milRNA 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 milRNA 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 analaysed in our labrotory. 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 milRNA 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 4. Total reads, short reads, adaptors and clean reads for fungi.

Fungi	Total Reads	< 18 nt*	3' adaptor	5' adaptor	Clean Reads	Unique clean reads
Bgh	26317125	349950	15002	16424	25767253	2539594
Pst	32490108	205066	81350	242498	31802410	4863345
Pi	11744418	289475	19992	19548	11299452	2095999

*Sequence reads smaller than 18 nucleotides

The length distributions of the unique clean reads for the fungi are shown in Figure 14. For each organisms 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.

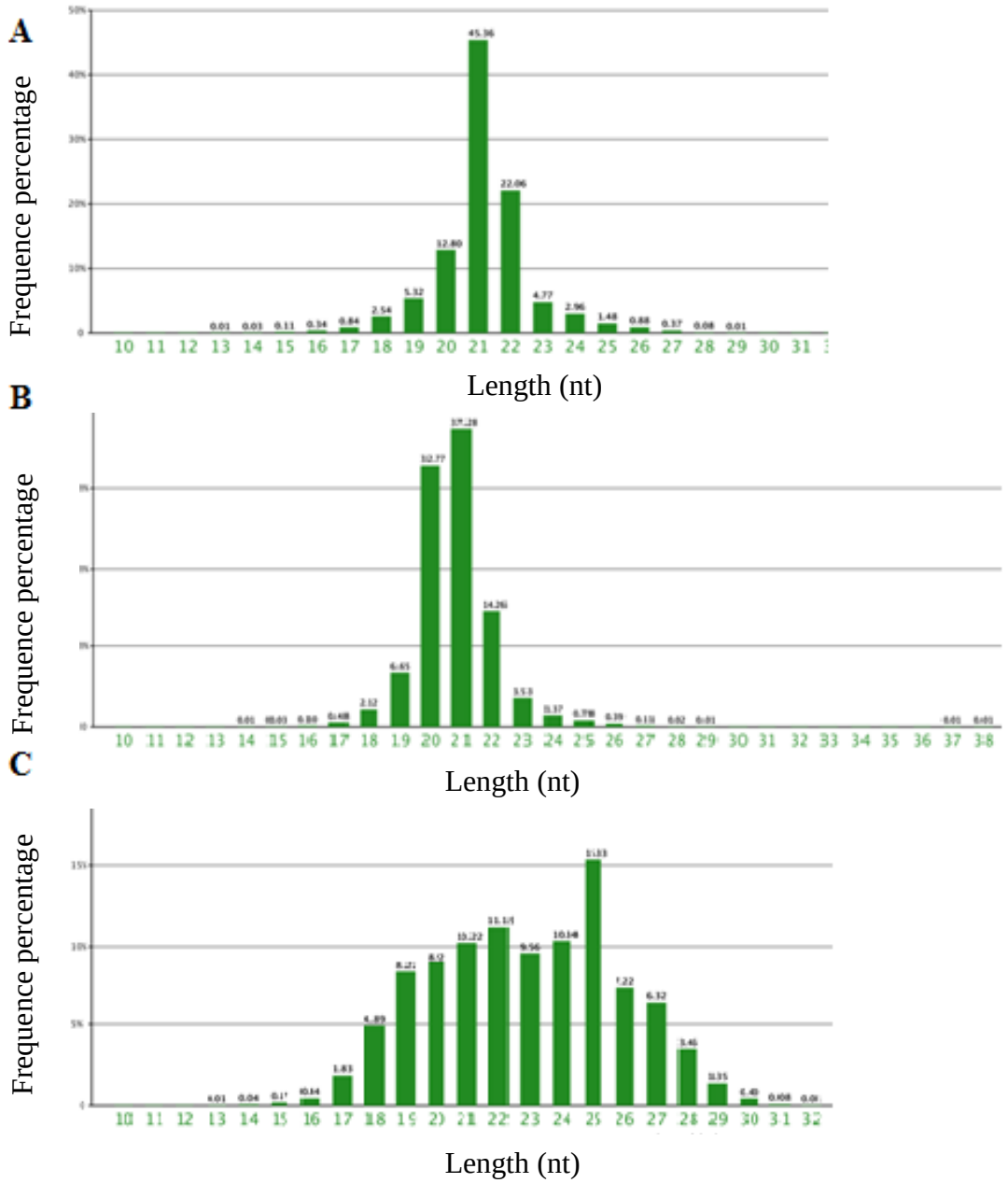


Figure 14. Graph demonstrating length distribution of read small RNAs of **A)** Bgh, **B)** Pst, and **C)** Pi.

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 5. Small RNA annotations for each organism

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

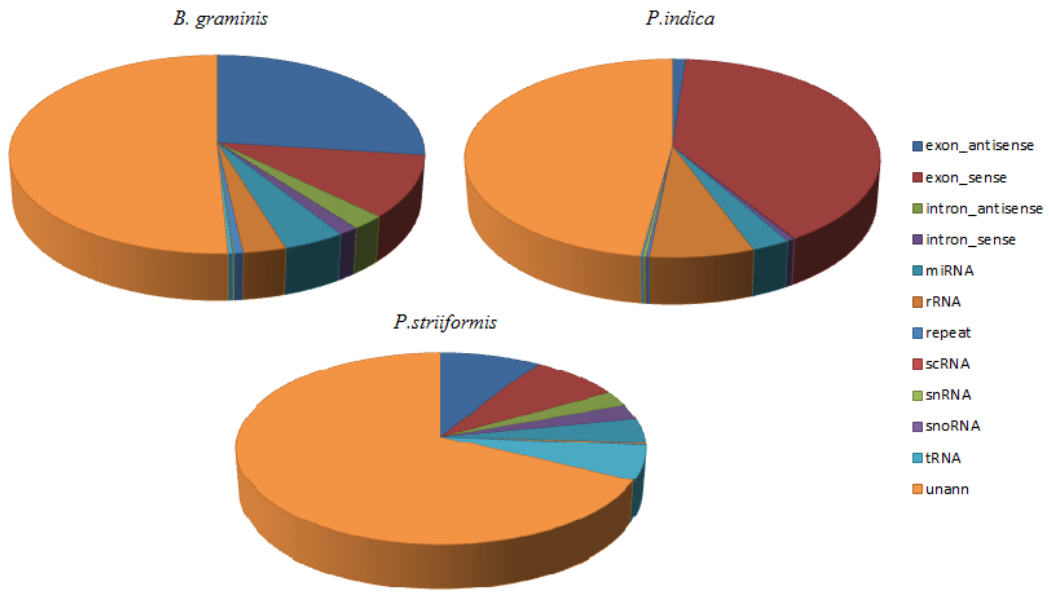


Figure 15. Pie charts representing small RNA annotations for Bgh (*B. graminis*), Pst (*P. striiformis*), and Pi (*P. indica*).

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 6. Clean reads, predicted miRNA and hairpin structure numbers per fungi.

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

*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 lenght 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).

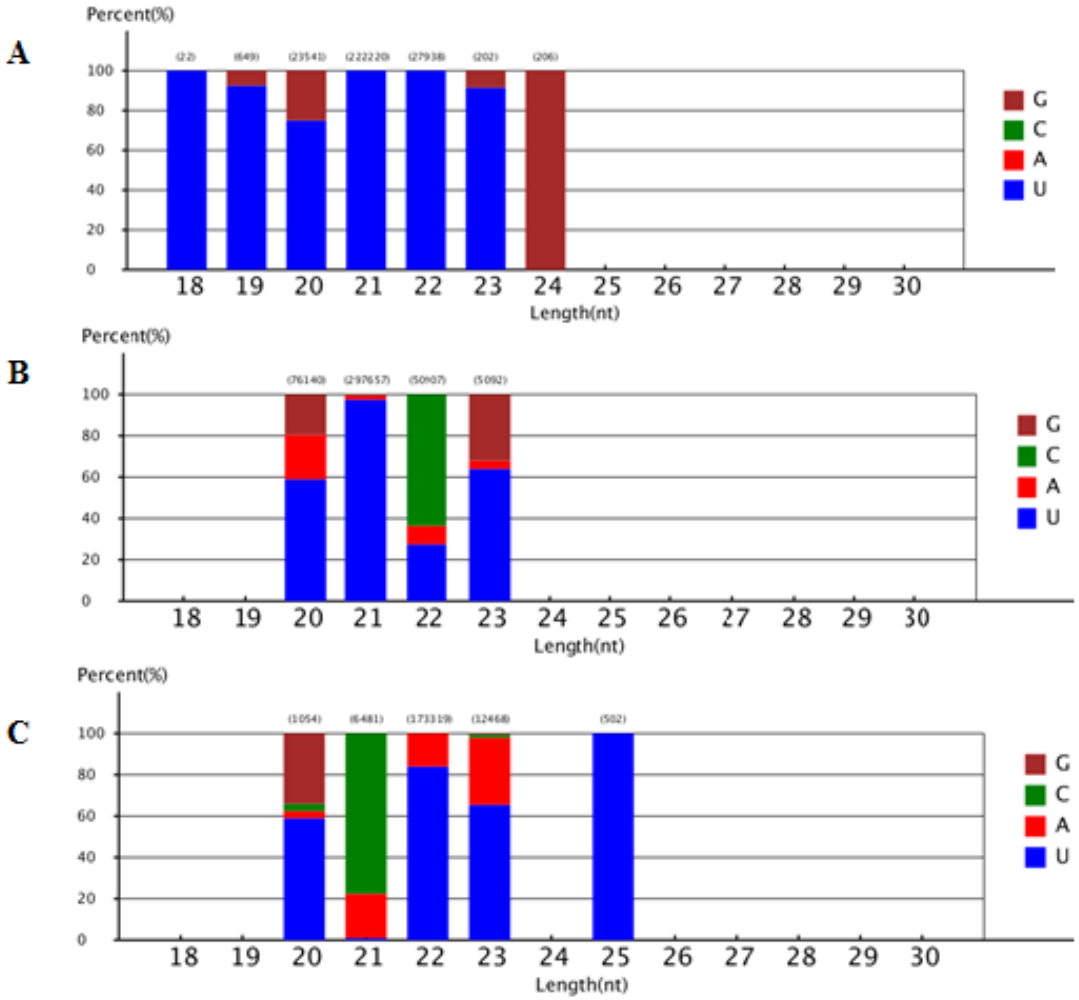


Figure 16. Graph representing first nucleotide bias for novel miRNA predictions of **A)** Bgh**, B)** Pst, and **C)** Pi.

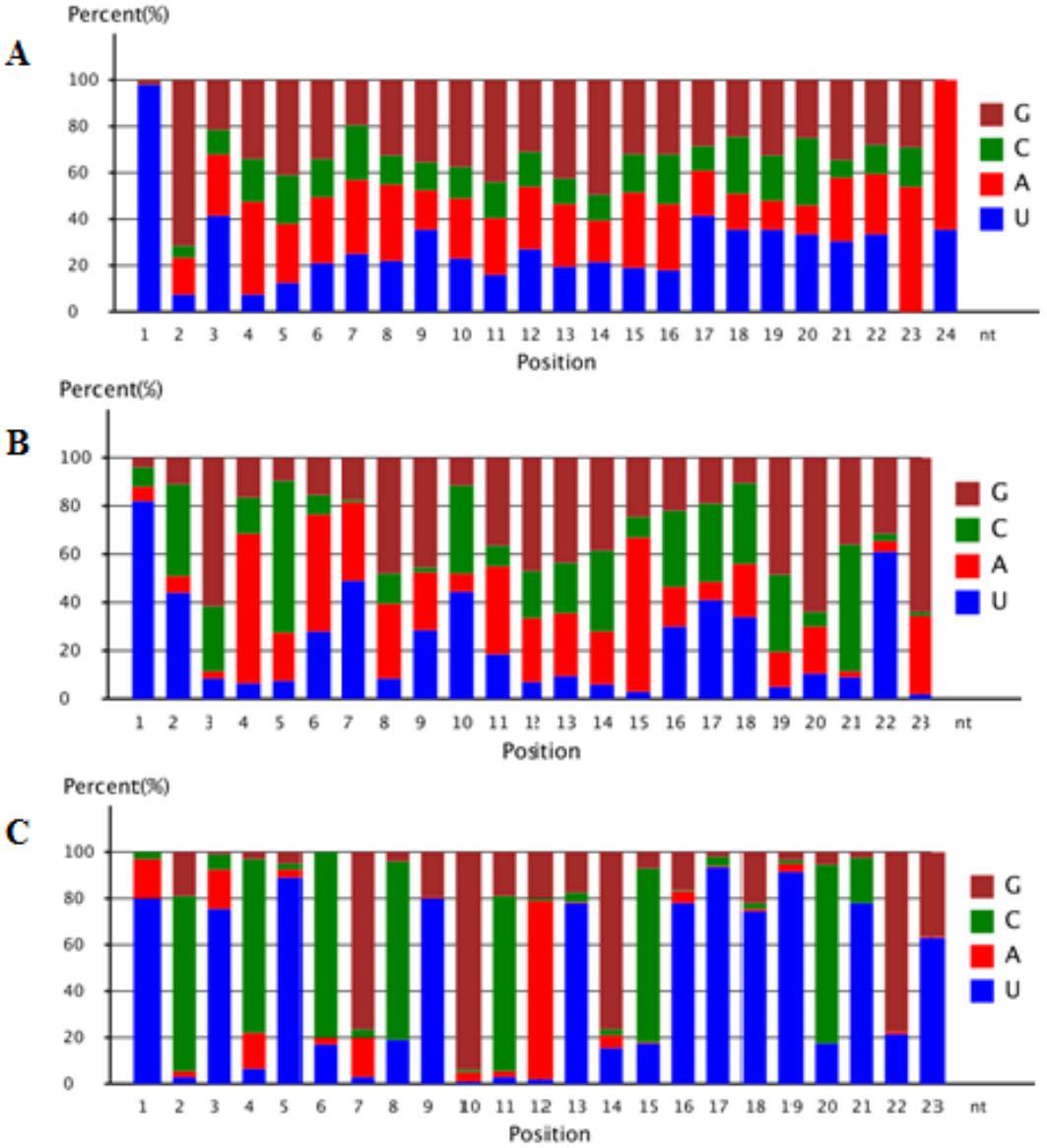


Figure 17. Graph for position base bias of novel miRNA predictions of **A)** Bgh**, B)** Pst, and **C)** Pi.

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.

matrix	miR-153-3p	miR-502-3p	miR-519c-3p	miR-21-3p	miR-216b-3p	miR-520g-5p
miR-153-3p	0	4	3	1	2	1
miR-502-3p	4	0	4	1	6	1
miR-519c-3p	3	4	0	1	2	3
miR-21-3p	1	1	1	0	3	2
miR-216b-3p	2	6	2	3	0	1
miR-520g-5p	1	1	3	2	1	0

Table 8. MCL matrix by Smith-Waterman algorithm.

matrix	miR-153-3p	miR-502-3p	miR-519c-3p	miR-21-3p	miR-216b-3p	miR-520g-5p
miR-153-3p	22	4	3	1	2	1
miR-502-3p	4	22	4	1	6	1
miR-519c-3p	3	4	22	1	2	3
miR-21-3p	1	1	1	21	3	2
miR-216b-3p	2	6	2	3	24	1
miR-520g-5p	1	1	3	2	1	23

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

matrix	miR-153-3p	miR-502-3p	miR-519c-3p	miR-21-3p	miR-216b-3p	miR-520g-5p
miR-153-3p	0	18	19	21	20	21
miR-502-3p	18	0	18	21	16	21
miR-519c-3p	19	18	0	21	20	19
miR-21-3p	20	20	20	0	18	19
miR-216b-3p	22	18	22	21	0	23
miR-520g-5p	22	22	20	21	22	0

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

matrix	miR-153-3p	miR-502-3p	miR-519c-3p	miR-21-3p	miR-216b-3p	miR-520g-5p
miR-153-3p	22	1	-1	-2	-2	-2
miR-502-3p	1	22	3	0	4	-4
miR-519c-3p	-1	3	22	-3	-1	-1
miR-21-3p	-2	0	-3	21	2	-2
miR-216b-3p	-2	4	-1	2	24	-4
miR-520g-5p	-2	-4	-1	-2	-4	23

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

matrix	miR-153-3p	miR-502-3p	miR-519c-3p	miR-21-3p	miR-216b-3p	miR-520g-5p
miR-153-3p	0	21	23	24	24	24
miR-502-3p	21	0	19	22	18	26
miR-519c-3p	23	19	0	25	23	23
miR-21-3p	23	21	24	0	19	23
miR-216b-3p	26	20	25	22	0	28
miR-520g-5p	25	27	24	25	27	0

Table 12. Randomly Filled matrix.

matrix	miR-153-3p	miR-502-3p	miR-519c-3p	miR-21-3p	miR-216b-3p	miR-520g-5p
miR-153-3p	0	21	13	24	8	6
miR-502-3p	6	0	18	1	17	4
miR-519c-3p	19	17	0	12	10	1
miR-21-3p	18	22	8	0	11	2
miR-216b-3p	15	14	4	10	0	15
miR-520g-5p	25	24	5	19	13	0

Table 13. 3-mer distribution datrix (K-mer)

matrix	AAA	AAU	AAC	AAG	AUA	AUU	AUC	AUG	ACA
miR-153-3p	1	0	0	1	1	0	1	0	1
miR-502-3p	0	1	0	1	0	1	0	1	0
miR-519c-3p	1	0	0	1	0	0	1	0	0
miR-21-3p	0	0	1	0	0	0	0	1	1
miR-216b-3p	0	0	0	0	0	1	0	0	1
miR-520g-5p	0	0	0	1	0	0	0	0	0

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.

	Matrix*	Cluster Number	Data Coverage (%)
K-means	K-mer	47	99.85
	NW-Similarity	46	85.44
	NW-Distance	46	82.73
	SW-Similarity	38	98.95
	SW-Distance	37	96.55
	Random Matrix	47	100.00
CLAG	K-mer	29	9.16
	NW-Similarity	30	10.96
	NW-Distance	31	11.26
	SW-Similarity	50	18.62
	SW-Distance	24	8.56
	Random Matrix	104	97.60
SOTA	K-mer	30	100.00
	NW-Similarity	30	100.00
	NW-Distance	30	100.00
	SW-Similarity	30	100.00
	SW-Distance	30	100.00
	Random Matrix	30	100.00
MCL	A	15	86.04
	B	18	73.12
	C	17	63.81
	D	56	75.96
	E	46	58.41
	F	46	52.70

*A, B and C are the 2nd; D, E, and F are the 4th power of the originaL Smith Waterman applied MCL matric. 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.

	Matrix	Dunn index
K-means	K-mer	0.3511
	NW-Similarity	0.2641
	NW-Distance	0.2297
	SW-Similarity	0.4539
	SW-Distance	0.3507
	Random Matrix	0.7920
CLAG	K-mer	0.7454
	NW-Similarity	0.6257
	NW-Distance	0.4498
	SW-Similarity	0.4867
	SW-Distance	0.4789
	Random Matrix	0.8311
SOTA	K-mer	0.2970
	NW-Similarity	0.2430
	NW-Distance	0.2043
	SW-Similarity	0.2766
	SW-Distance	0.2369
	Random Matrix	0.8396

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 matric. Various optimizations are needed to made 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

	Matrix	Clusters	Function	Family	HMDD	Tissue	All*
K-means	K-mer	44.68	12.76	72.34	44.68	10.63	80.85
	NW-Similarity	57.78	22.22	82.22	40.00	11.11	88.89
	NW-Distance	52.17	19.56	76.09	45.65	6.52	84.78
	SW-Similarity	63.16	23.68	78.94	55.26	15.79	92.11
	SW-Distance	56.76	32.43	70.27	40.54	10.81	81.08
	Ramdom Matrix	8.51	6.38	8.51	14.89	4.25	34.04
CLAG	K-mer	24.13	13.79	75.86	27.59	13.79	75.86
	NW-Similarity	30.00	16.67	80.00	36.67	13.33	80.00
	NW-Distance	22.58	16.13	77.42	41.94	9.68	77.42
	SW-Similarity	20.00	14.00	70.00	24.00	10.00	70.00
	SW-Distance	16.67	20.83	79.17	45.83	4.17	79.17
	Random matrix	4.81	2.88	4.81	8.63	0	16.35
MCL	A	60.00	26.67	86.66	33.33	13.33	86.66
	B	38.89	22.22	66.67	33.33	16.67	72.22
	C	41.17	23.53	58.82	29.41	17.65	70.59
	D	32.14	14.29	51.79	25.00	7.14	66.07
	E	39.13	13.04	47.03	21.74	8.70	60.87
	F	39.13	15.22	47.83	21.74	10.87	60.87
SOTA	K-mer	50.00	33.33	63.33	36.67	10.00	80.00
	NW-Similarity	40.00	23.33	73.33	26.67	6.67	80.00
	NW-Distance	36.67	13.33	70.00	50.00	6.67	83.33
	SW-Similarity	50.00	30.00	70.00	43.33	10.00	83.33
	SW-Distance	43.33	23.33	60.00	50.00	10.00	73.33
	Ramdom Matrix	10.00	10.00	10.00	20.00	3.33	43.33
Random Clusters	10 groups	10.00	13.33	6.67	33.33	0	46.67
	30 groups	14.44	10.00	11.11	22.22	0	40.00
	150 groups	3.16	4.28	2.48	9.93	0.68	16.7

* 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 prior 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.

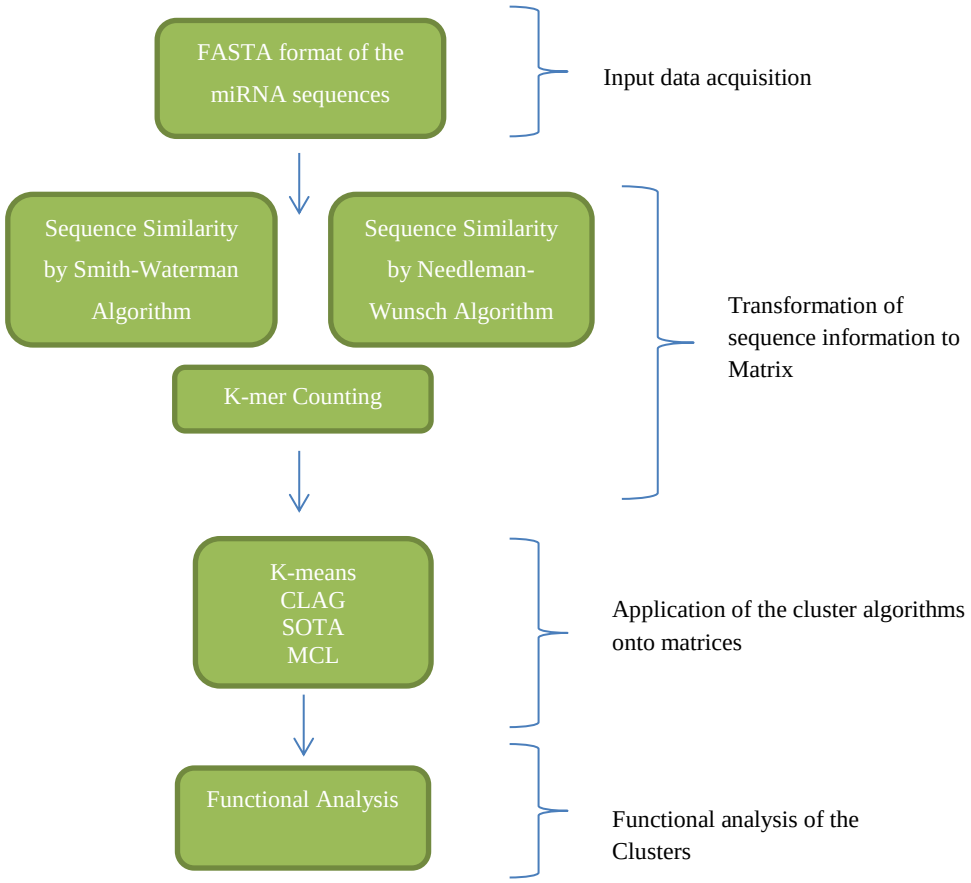


Figure 18. Workflow of miRNA sequence based clustering.

3.10 Analysis of the milRNAs 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 measing 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 Supplementray Material 3 Data folder and Appendix B includes explanations of data organisation. Table 17 signifies the analysis results.

Table 17. milRNA cluster results.

Fungus	Novel miRNA Count	Known miRNA Count	Total Input Number	Cluster Number	Data Covarage (%)
Bgh	2593	180	2773	25	76.5
Pst	492	301	793	2	0.9
PI	61	54	115	11	75.7

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 milRNA 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 milRNA by more than 500 count number.

Detected milRNAs	Bgh	Pst	Pi
Novel miRNA	111	34	12
Known miRNA	41	54	18
Total	152	88	30

Count number distributions of the selected mirlRNAs 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.

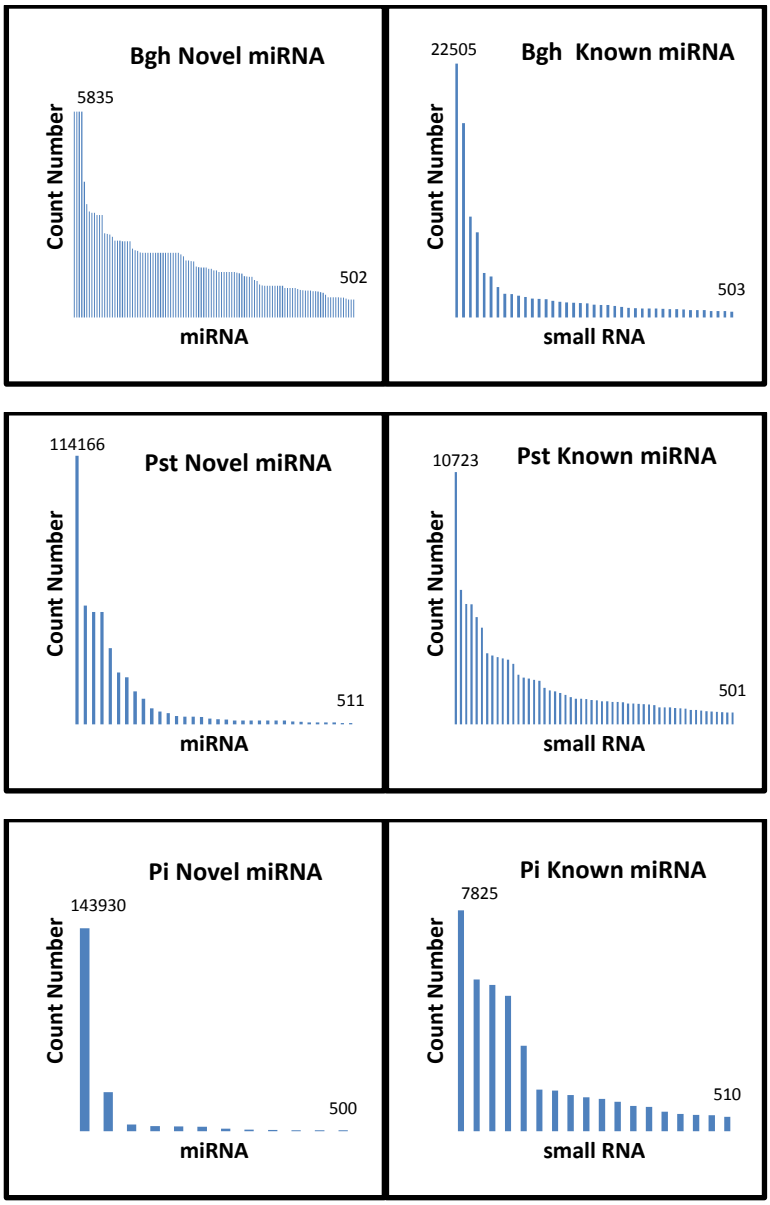


Figure 19. Novel miRNA and known miRNA distributions in fungi.

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 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.

Hit Number	Bgh	Pst	Pi
Novel miRNA	991	295	2130
Known miRNA	528	536	3949
Total	1519	831	6079

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

Fungi	Analyzed milRNA	Number of Hit	Unique milRNA	Unique Target Entry
Novel miRNA				
Bgh	111	114	59	47
Pst	34	24	10	16
Pi	12	111	12	111
Known miRNA				
Bgh	41	35	20	26
Pst	54	43	18	38
Pi	18	203	16	156

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 milRNA sequences are selected and shown. Novel milRNA 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 milRNA sequences.

miRNA name	Sequences	MFE kcal/mol	L*	Count**
P.indica-m0005_3p	TTTGAATTTCTTGACTGATGCA	-52.30	22	1146
P.indica-m0038_3p	AGAATTATGGGGGTTGTGTTCT	-94.40	22	27648
P.indica-m0004_5p	TCTCTCGCTGCATGCTTTTCTG	-26.80	22	143930
B1202A-m0614_5p	TGTGTTGTGGACGTACGGGC	-20.10	20	570
B1202A-m0367_5p	TGGCCGGTAATGAATGAGGCA	-24.90	21	561
PSTmix-m0195_3p	TGCAGGGGAGGGGTAGTCGA	-164.80	20	736
PSTmix-m0459_3p	TTGCAAGGACTCCGGAGAGGC	-120.30	21	50514

*Length of the miRNA,** Number of reads of the miRNA

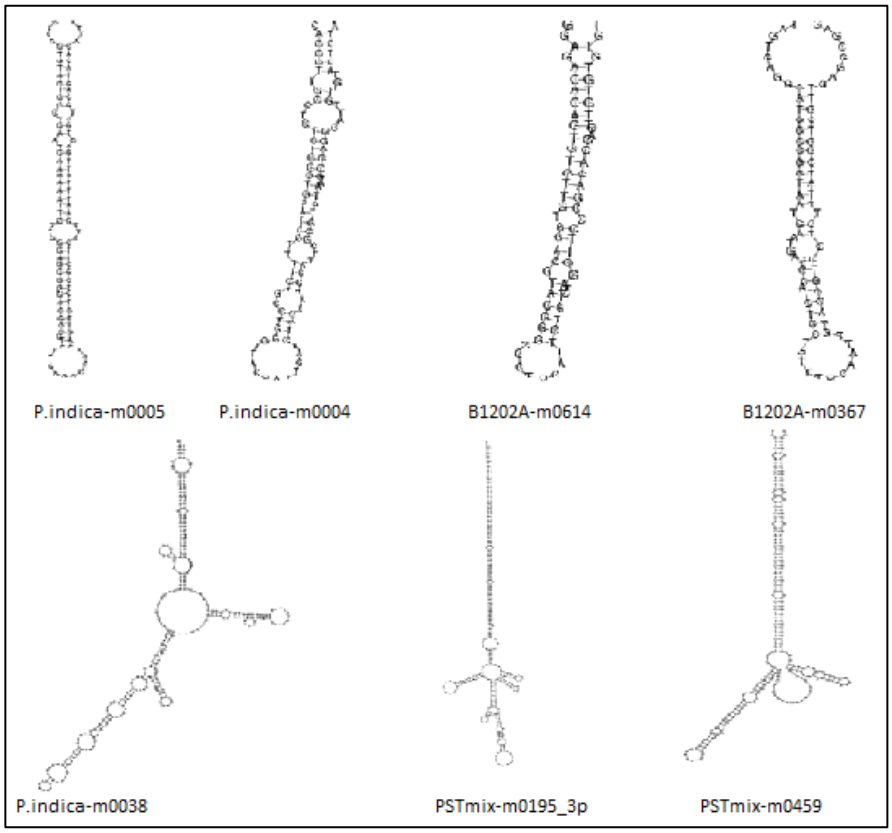


Figure 20. Novel miRNA precursors shown in Table 21.

Table 22. Selected known miRNA sequences and their best homologs.

Small RNA ID	L*	Count**	Sequence	Best homolog (mirBASE)***	L*	Family
Bgh-t0005282	21	566	Small RNA TGAAGGATTGAGTTGATGGAA Homolog TGGAATGATTGAGCTTGATGGA	cel-miR-1819-3p	22	undef
Bgh-t0006075	22	503	Small RNA TTGGAAGACTTGTGTATTATTT Homolog TGGAAGACTTGTGATTTTGTT	dre-miR-7b	21	mir-7
PI-t0000089	22	7825	Small RNA AGGAATGTAAAGAAGTATGTAT Homolog TGGAATGTAAAGAAGTATGTAT	mmu-miR-1a-3p	22	mir-1
PI-t0001651	22	611	Small RNA TGAGGTAGTAGGTTGTATAGTT Homolog TGAGGTAGTAGGTTGTATAGTT	cel-let-7-5p	22	let-7
PI-t0000286	25	3028	Small RNA GCAGATCTTGGTGGTAGTAGCAAAT Homolog AGATCTTGGTGGTAGTAGCAAATAT	sha-miR-716b	25	mir-716
Pst-t0000930	21	2740	Small RNA TTGGCAAAGTGGTTGGAATGT Homolog TGGGCAAAGTTGGTTGAGAAGT	cel-miR-4814-3p	22	Undef
Pst-t0002068	21	1444	Small RNA TGGGGGAAGAAGATAGAAAGG Homolog TGGGAGAAAAGATAGAATGTG	mtr-miR5239	21	Undef

* 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 milRNAs. For more detailed information raw results Appendix should be seen. The selected milRNA to target regions are sorted by Target accecibily (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 milRNAs' predicted target regions in plant genomes.

miRNA name	Accession**	Exp*	UPE	milRNA/Target Alignment	Target Description	Target Organism
P.indica-m0005_3p	GE484550	2.0	14.835	miRNA 20 GUAGUCAGUUCUUUAAGUUU 1 ::: : :::::::::::::: Target 230 CAUGACUCAAGAAAUUCAAA 249	Proline-rich protein	Helianthus annuus
P.indica-m0038_3p	LOC_Os03g42020.1|12003.m34716|cDNA	1.5	20.066	miRNA 22 UCUUGUGUUGGGGGUAUUAAGA 1 : .::::::::::..:::::::: Target 567 AAGACACAACCCUUAUAAUUCU 588	Calcium-transporting ATPase 2	Oryza sativa
P.indica-m0004_5p	TC143937	2.0	23.293	miRNA 21 UCUUUUCGUACGUCGCUCUCU 1 .:::.::::::::: .:::::: Target 505 GGAAGAGCAUGCAUUGAGAGA 525	Pyruvate kinase	Populus trichocarpa
B1202A-m0614_5p	CD058210	1.0	17.178	miRNA 20 CGGGCAUGCAGGUGUUGUGU 1 :::::::::::::::::: Target 123 GCCCGUACGUCCACAACACU 142	RND efflux system outer membrane lipoprotein NodT	Hordeum vulgare
B1202A-m0367_5p	CA682639	2.0	12.132	miRNA 21 ACGGAGUAAGUAAUGGCCGGU 1 :::::::::.::::::: ::: Target 82 UGCCUCAUUUAUUACCGACCA 102	A-kinase anchor protein 11	Tritium aestivum
PSTmix-m0195_3p	TC241290	2.5	16.720	miRNA 20 AGCUGAUGGGGAGGGGACGU 1 ::::: :::::.: ::::::: Target 454 UCGACAACCCUUGCCCUGCA 473	Serine carboxypeptidase family protein	Hordeum vulgare
PSTmix-m0459_3p	TC403010	3.0	18.670	miRNA 20 GGAGAGGCCUCAGGAACGUU 1 .::::::::: ::.:::::::. Target 299 CUUCUCUGGAUUCUUUGCAC 318	Protein kinase	Tritium aestivum

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

Table 24. Known milRNAs' predicted target regions in plant genomes.

miRNA name	Accession**	Exp*	UPE	milRNA/Target Alignment	Target Description	Target Organism
PI-t0000089	TC12570	2.0	16.779	miRNA 21 AUGUAUGAAGAAAUGUAAGGA 1 ::.::: ::::::::::::.: Target 802 UAUAUAAUUCUUUACAUUCUU 822	Clp protease proteolytic subunit	Phaseolus coccineus
PI-t0001651	BQ848540	2.0	4.681	miRNA 21 UGAUAUGUUGGAUGAUGGAGU 1 :.::::::. :::::::::::: Target 51 AUUAUACGUCCUACUACCUCA 71	ATPase subunit 7	Lactuca sativa
PI-t0000286	TC400366	0	14.232	miRNA 25 UAAACGAUGAUGGUGGUUCUAGACG 1 ::::::::::::::::::::::::: Target 560 AUUUGCUACUACCACCAAGAUCUGC 584	Pspzf zinc finger protein-like	Arabidopsis thaliana
Bgh-t0005282	TC251661	3.0	13.112	miRNA 21 AAGGUAGUUGAGUUAGGAAGU 1 :. ::::::::.:::::::.: Target 3197 UUGUAUCAGCUUGAUCCUUCA 3217	NB-ARC domain containing protein	Hordeum vulgare
Bgh-t0006075	CA698277	2.5	8.53	miRNA 20 UAUUAUGUGUUCAGAAGGUU 1 : :::::::::::::::: :: Target 473 ACAAUACACAAGUCUUCNAA 492	Pyruvate kinase	Tritium aestivum
Pst-t0000930	AJ476982	1.5	10.593	miRNA 21 UGUAAGGUUGGUGAAACGGUU 1 :.:::::::::: ::::::::: Target 137 AUAUUCCAACCUCUUUGCCAA 157	Protein kinase like protein	Hordeum vulgare
Pst-t0002068	CJ641074	3.0	19.941	miRNA 21 UGGAAAGAUAGAAGAAGGGGG 1 ::::::: ::::::::::.::: Target 274 CCUUUCUAUCUUCUUCUUCGC 254	Phosphatidylinositol 4-kinase	Tritium aestivum

*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 *Triticulum 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 *Triticulum 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 logarithmsof 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) covarge 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 Eucledian metrices. The DI results for other algorithms can be viewed through Table 15. DI resuls 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 marices. 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 enrichmentresults 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 milRNA application to the pipeline

The workflow generated through our clustering analysis was shown in Section 3.10, Figure 18. Novel and known candidate milRNA 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 milRNA 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 mutualisticly 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 mılRNA generation mechanism in Pi. A different mechanism can be considered likey since Bgh and Pst are obligate parasites as part of baciosides, Pi on the other hand is a mutualisticly 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 experiemental analysis in laboratories can be costy through several steps of the experiemets. 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) milRNAs 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 messanger 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 milRNA 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 milRNA character on target region is unknown. The tool has some important functions while identifying the target locations. Target accecibilty 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 milRNAs targeting to 174 unique regions, and 54 known milRNAs 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 milRNA candidates with their low Maximum Expectation and high UPE (target accecibility) values.

The selected milRNA target regions found to be very valuable in determination of function of milRNAs 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 milRNA 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 milRNA, 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 milRNA target genes; such as Pyruvate kinase in *Triticum aestium* and protein kinase like protein in *Hordeum vulgare* by milRNA Pst-t0000930 *Puccinia striiformis* f. sp. *tritici,* milRNA Bgh-t0006075 of *Blumeria graminis* f.sp. *hordei*, repectively (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, milRNAs were hunt down by obtaining small non-coding RNA sequencing dataset in our laboratory, and their target regions were identified, therefore candidate milRNAs 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 familes 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 asignment 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 seeked 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 milRNAs 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 knock-down assay. Silencing the miRNA in the cell through mutation on miRNA gene itself can be one approach.

This master thesis supported by TÜBİTAK 2210 BIDEB thesis grant and research grand of Akkaya project 113Z038 and 110T445.

REFERENCES

Abbott, A. L., Alvarez-Saavedra, E., Miska, E. a, Lau, N. C., Bartel, D. P., Horvitz, H. R., & Ambros, V. (2005). The let-7 MicroRNA family members mir-48, mir-84, and mir-241 function together to regulate developmental timing in Caenorhabditis elegans. *Developmental Cell*, *9*(3), 403–14. doi:10.1016/j.devcel.2005.07.009

Altuvia, Y., Landgraf, P., Lithwick, G., Elefant, N., Pfeffer, S., Aravin, A., … Margalit, H. (2005). Clustering and conservation patterns of human microRNAs. *Nucleic Acids Research*, *33*(8), 2697–706. doi:10.1093/nar/gki567

An, J., Lai, J., Lehman, M. L., & Nelson, C. C. (2013). miRDeep*: an integrated application tool for miRNA identification from RNA sequencing data. *Nucleic Acids Research*, *41*(2), 727–37. doi:10.1093/nar/gks1187

Asgari, S. (2011). Role of MicroRNAs in Insect Host-Microorganism Interactions. *Frontiers in Physiology*, *2*(August), 48. doi:10.3389/fphys.2011.00048

Bartel, B., & Bartel, D. P. (2003). Update on Small RNAs MicroRNAs : At the Root of Plant Development ? 1. *Plant Physiology*, *132*(June), 709–717. doi:10.1104/pp.103.023630.predicted

Bartel, D. P. (2004). MicroRNAs : Genomics , Biogenesis , Mechanism , and Function Genomics : The miRNA Genes. *Cell*, *116*, 281–297.

Bartel, D. P. (2013). Micro RNA Target Recognition and Regulatory Functions. *Cell*, *136*(2), 215–233. doi:10.1016/j.cell.2009.01.002.MicroRNA

BGI. (n.d.). MIREAP-MicroRNA discovery by deep sequencing. Retrieved from https://sourceforge.net/projects/mireap/

Bindschedler, L. V, Burgis, T. a, Mills, D. J. S., Ho, J. T. C., Cramer, R., & Spanu, P. D. (2009). In planta proteomics and proteogenomics of the biotrophic barley fungal pathogen Blumeria graminis f. sp. hordei. *Molecular & Cellular Proteomics : MCP*, *8*(10), 2368–81. doi:10.1074/mcp.M900188-MCP200

Bonnet, E., He, Y., Billiau, K., & Van de Peer, Y. (2010). TAPIR, a web server for the prediction of plant microRNA targets, including target mimics. *Bioinformatics (Oxford, England)*, *26*(12), 1566–8. doi:10.1093/bioinformatics/btq233

Borchert, G. M., Lanier, W., & Davidson, B. L. (2006). RNA polymerase III transcribes human microRNAs. *Nature Structural & Molecular Biology*, *13*(12), 1097–101. doi:10.1038/nsmb1167

Bozkurt, T. O., Mcgrann, G. R. D., Maccormack, R., Boyd, L. A., & Akkaya, M. S. (2010). Cellular and transcriptional responses of wheat during compatible and incompatible race-specific interactions with Puccinia striiformis f . sp . tritici. *Molecular Plant Pathology*, *11*(5), 625–640. doi:10.1111/J.1364-3703.2010.00633.X

Brock, G., Pihur, V., Susmita, D., & Somnath, D. (2008). clValid : An R Package for Cluster Validation. *Journal of Statistical Software*, *25*(4).

Burge, S. W., Daub, J., Eberhardt, R., Tate, J., Barquist, L., Nawrocki, E. P., … Bateman, A. (2013). Rfam 11.0: 10 years of RNA families. *Nucleic Acids Research*, *41*(Database issue), D226–32. doi:10.1093/nar/gks1005

Cai, X., Hagedorn, C. H., & Cullen, B. R. (2004). Human microRNAs are processed from capped , polyadenylated transcripts that can also function as mRNAs. *RNA Biology*, *10*(12), 1957–1966. doi:10.1261/rna.7135204.miRNAs

Chen, C., Ridzon, D. a, Broomer, A. J., Zhou, Z., Lee, D. H., Nguyen, J. T., … Guegler, K. J. (2005). Real-time quantification of microRNAs by stem-loop RT-PCR. *Nucleic Acids Research*, *33*(20), e179. doi:10.1093/nar/gni178

Chen, X., Genetics, W., Long, D. L., Paul, S., Line, R. F., Pathology, P., & Marshall, D. (2000). Wheat Stripe Rust Epidemics and Races of Puccinia striiformis f . sp . tritici in the United States in 2000, 28–30.

Collins, L. J. (2011). Characterizing ncRNAs in Human Pathogenic Protists Using High-Throughput Sequencing Technology. *Frontiers in Genetics*, *2*(December), 96. doi:10.3389/fgene.2011.00096

Consorium, T. gene O. (2000). Gene Ontology : tool for the unification of biology, *25*(may), 25–29.

Dai, X., & Zhao, P. X. (2011). psRNATarget: a plant small RNA target analysis server. *Nucleic Acids Research*, *39*(Web Server issue), W155–9. doi:10.1093/nar/gkr319

Deshmukh, S., Hückelhoven, R., Schäfer, P., Imani, J., Sharma, M., Weiss, M., … Kogel, K.-H. (2006). The root endophytic fungus Piriformospora indica requires host cell death for proliferation during mutualistic symbiosis with barley.

Proceedings of the National Academy of Sciences of the United States of America, *103*(49), 18450–7. doi:10.1073/pnas.0605697103

Dib, L., & Carbone, A. (2012). Open Access CLAG : an unsupervised non hierarchical clustering algorithm handling biological data.

Dopazo, J., Wang, H. C., de la Fraga, L. G., Zhu, Y. P., & Carazo, J. M. (1997). Self-organizing tree-growing network for the classification of protein sequences. *Protein Science : A Publication of the Protein Society*, *7*(12), 2613–22. doi:10.1002/pro.5560071215

Duggleby, R. G., & Dennis, D. T. (1973). Pyruvate kinase, a possible regulatory enzyme in higher plants. *Plant Physiology*, *52*(4), 312–7. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=366494&tool=pmcentrez&rendertype=abstract

Dunn, J. C. (1973). A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics*, *3*(3), 32–57. doi:10.1080/01969727308546046

Durbin, R., Eddy, S., Krogh, A., & Mitchison, G. (1998). *Biological sequence analysis*. Cambridge University Press.

Elmore, J. M., & Coaker, G. (2011). The role of the plasma membrane H+-ATPase in plant-microbe interactions. *Molecular Plant*, *4*(3), 416–27. doi:10.1093/mp/ssq083

Enright, a J., Van Dongen, S., & Ouzounis, C. a. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, *30*(7), 1575–84. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=101833&tool=pmcentrez&rendertype=abstract

Glawe, D. a. (2008). The powdery mildews: a review of the world's most familiar (yet poorly known) plant pathogens. *Annual Review of Phytopathology*, *46*, 27–51. doi:10.1146/annurev.phyto.46.081407.104740

Haas, B. J., & Zody, M. C. (2010). Advancing RNA-Seq analysis. *Nature Biotechnology*, *28*(5), 421–3. doi:10.1038/nbt0510-421

He, L., & Hannon, G. J. (2004). MicroRNAs: small RNAs with a big role in gene regulation. *Nature Reviews. Genetics*, *5*(7), 522–31. doi:10.1038/nrg1379

Herrero, J., Diaz-Uriarte, R., & Dopazo, J. (2003). Gene expression data preprocessing. *Bioinformatics*, *19*(5), 655–656. doi:10.1093/bioinformatics/btg040

Herrero, J., Valencia, A., & Joaquin, D. (2001). network for clustering gene expression patterns, *17*(2), 126–136.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. doi:10.1016/j.patrec.2009.09.011

Jansson, M. D., & Lund, A. H. (2012). MicroRNA and cancer. *Molecular Oncology*, *6*(6), 590–610. doi:10.1016/j.molonc.2012.09.006

Jiang, M., Sang, X., & Hong, Z. (2012). Beyond nutrients: food-derived microRNAs provide cross-kingdom regulation. *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology*, *34*(4), 280–4. doi:10.1002/bies.201100181

Jiang, P., Wu, H., Wang, W., Ma, W., Sun, X., & Lu, Z. (2007). MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Research*, *35*(Web Server issue), W339–44. doi:10.1093/nar/gkm368

Kang, K., Zhong, J., Jiang, L., Liu, G., Gou, C. Y., Wu, Q., … Gou, D. (2013). Identification of microRNA-Like RNAs in the filamentous fungus Trichoderma reesei by solexa sequencing. *PloS One*, *8*(10), e76288. doi:10.1371/journal.pone.0076288

Koutsoulidou, A., Mastroyiannopoulos, N. P., Furling, D., Uney, J. B., & Phylactou, L. a. (2011). Expression of miR-1, miR-133a, miR-133b and miR-206 increases during development of human skeletal muscle. *BMC Developmental Biology*, *11*(1), 34. doi:10.1186/1471-213X-11-34

Kozomara, A., & Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Research*, *39*(Database issue), D152–7. doi:10.1093/nar/gkq1027

Kuhn, D. E., Martin, M. M., Feldman, D. S., Terry, A. V, Nuovo, G. J., & Elton, T. S. (2008). Experimental validation of miRNA targets. *Methods (San Diego, Calif.)*, *44*(1), 47–54. doi:10.1016/j.ymeth.2007.09.005

Lagos-Quintana, M., Rauhut, R., Lendeckel, W., & Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. *Science (New York, N.Y.)*, *294*(5543), 853–8. doi:10.1126/science.1064921

Larkin, M. a, Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. a, McWilliam, H., … Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics (Oxford, England)*, *23*(21), 2947–8. doi:10.1093/bioinformatics/btm404

Lee, H., Li, L., Gu, W., Xue, Z., Crosthwaite, S. K., Pertsemlidis, A., … Blvd, H. (2011). Diverse pathways generate microRNA-like RNAs and Dicer-independent small interfering RNAs in fungi. *NIH Public Access*, *38*(6), 803–814. doi:10.1016/j.molcel.2010.04.005.Diverse

Lee, R. C., & Ambros, V. (2001). An extensive class of small RNAs in Caenorhabditis elegans. *Science (New York, N.Y.)*, *294*(5543), 862–4. doi:10.1126/science.1065329

Lee, R. C., Feinbaum, R. L., & Ambros, V. (1993). The C . elegans Heterochronic Gene lin-4 Encodes Small RNAs with Antisense Complementarity to & II-14, *75*, 843–854.

Lee, Y., Jeon, K., Lee, J.-T., Kim, S., & Kim, V. N. (2002). MicroRNA maturation: stepwise processing and subcellular localization. *The EMBO Journal*, *21*(17), 4663–70. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=126204&tool=pmcentrez&rendertype=abstract

Lee, Y., Kim, M., Han, J., Yeom, K.-H., Lee, S., Baek, S. H., & Kim, V. N. (2004). MicroRNA genes are transcribed by RNA polymerase II. *The EMBO Journal*, *23*(20), 4051–60. doi:10.1038/sj.emboj.7600385

Lelandais-Brière, C., Sorin, C., Declerck, M., Benslimane, A., Crespi, M., & Hartmann, C. (2010). Small RNA diversity in plants and its impact in development. *Current Genomics*, *11*(1), 14–23. doi:10.2174/138920210790217918

Li, L., Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research*, *13*(9), 2178–89. doi:10.1101/gr.1224503

Li, R., Li, Y., Kristiansen, K., & Wang, J. (2008). SOAP: short oligonucleotide alignment program. *Bioinformatics (Oxford, England)*, *24*(5), 713–4. doi:10.1093/bioinformatics/btn025

Lu, M., Shi, B., Wang, J., Cao, Q., & Cui, Q. (2010). TAM: a method for enrichment and depletion analysis of a microRNA category in a list of microRNAs. *BMC Bioinformatics*, *11*, 419. doi:10.1186/1471-2105-11-419

Lu, M., Zhang, Q., Deng, M., Miao, J., Guo, Y., Gao, W., & Cui, Q. (2008). An analysis of human microRNA and disease associations. *PloS One*, *3*(10), e3420. doi:10.1371/journal.pone.0003420

Lund, E., & Dahlberg, J. E. (2006). Substrate selectivity of exportin 5 and Dicer in the biogenesis of microRNAs. *Cold Spring Harbor Symposia on Quantitative Biology*, *71*, 59–66. doi:10.1101/sqb.2006.71.050

Macqueen, J. (1967). Some Methods For Classification and Analysis of Multivariate Observation. In *Berkeley Symposium on Matematical Statistic and Probablity* (Vol. 233, pp. 281–297). University of California Press.

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, *48*(3), 443–453. doi:10.1016/0022-2836(70)90057-4

Oğul, H., & Mumcuoğlu, E. U. (2007). A discriminative method for remote homology detection based on n-peptide compositions with reduced amino acid alphabets. *Bio Systems*, *87*(1), 75–81. doi:10.1016/j.biosystems.2006.03.006

Pascarella, S., & Argos, P. (1992). Analysis of insertions/deletions in protein structures. *Journal of Molecular Biology*, *224*(2), 461–471. doi:10.1016/0022-2836(92)91008-D

Pratt, A. J., & MacRae, I. J. (2009). The RNA-induced silencing complex: a versatile gene-silencing machine. *The Journal of Biological Chemistry*, *284*(27), 17897–901. doi:10.1074/jbc.R900012200

Raabe, C. a, Tang, T.-H., Brosius, J., & Rozhdestvensky, T. S. (2014). Biases in small RNA deep sequencing data. *Nucleic Acids Research*, *42*(3), 1414–26. doi:10.1093/nar/gkt1021

Rawlins, T., Lewis, A., Hettenhausen, J., & Mirjalili, S. (2012). Interactive k-means clustering for investigation of optimisation solution data, *0*, 1–2.

Reddy, S. D. N., Ohshiro, K., Rayala, S. K., & Kumar, R. (2008). MicroRNA-7, a homeobox D10 target, inhibits p21-activated kinase 1 and regulates its functions. *Cancer Research*, *68*(20), 8195–200. doi:10.1158/0008-5472.CAN-08-2103

Rhoades, M. W., Reinhart, B. J., Lim, L. P., Burge, C. B., Bartel, B., & Bartel, D. P. (2002). Prediction of plant microRNA targets. *Cell*, *110*(4), 513–20. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20430753

Sisodia, D. (2012). Clustering Techniques : A Brief Survey of Different Clustering Algorithms, *1*(3), 82–87.

Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, *147*(1), 195–197. doi:10.1016/0022-2836(81)90087-5

Srivastava, P. K., Moturu, T. R., Pandey, P., Baldwin, I. T., & Pandey, S. P. (2014). A comparison of performance of plant miRNA target prediction tools and the characterization of features for genome-wide target prediction. *BMC Genomics*, *15*(1), 348. doi:10.1186/1471-2164-15-348

Tempel, S., & Tahi, F. (2012). A fast ab-initio method for predicting miRNA precursors in genomes. *Nucleic Acids Research*, *40*(11), e80. doi:10.1093/nar/gks146

Varma, a, Savita, V., Sudha, Sahay, N., Butehorn, B., & Franken, P. (1999). Piriformospora indica, a cultivable plant-growth-promoting root endophyte. *Applied and Environmental Microbiology*, *65*(6), 2741–4. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=91405&tool=pmcentrez&rendertype=abstract

Wang, Z., Gerstein, M., & Snyder, M. (2010). RNA-Seq : a revolutionary tool for transcriptomics. *Nat Rev Genet*, *10*(1), 57–63. doi:10.1038/nrg2484.RNA-Seq

Weiberg, A., Wang, M., Lin, F.-M., Zhao, H., Zhang, Z., Kaloshian, I., … Jin, H. (2013). Fungal small RNAs suppress plant immunity by hijacking host RNA interference pathways. *Science (New York, N.Y.)*, *342*(6154), 118–23. doi:10.1126/science.1239705

Williamson, M. P. (1994). The structure and function of proline-rich regions in proteins. *The Biochemical Journal*, *297 (Pt 2*, 249–60. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1137821&tool=pmcentrez&rendertype=abstract

Xu, P., Lloyd, C. W., Staiger, C. J., & Drobak, B. K. (1992). Association of Phosphatidylinositol 4-Kinase with the Plant Cytoskeleton. *The Plant Cell*, *4*(8), 941–951. doi:10.1105/tpc.4.8.941

Zeng, Y. (2006). Principles of micro-RNA production and maturation. *Oncogene*, *25*(46), 6156–62. doi:10.1038/sj.onc.1209908

Zhang, B., Stellwag, E. J., & Pan, X. (2009). Large-scale genome analysis reveals unique features of microRNAs. *Gene*, *443*(1-2), 100–109. doi:10.1111/mono.12118

Zymañski, M. S., Arciszewska, M. Z. B., & Ywicki, M. (2002). Noncoding RNA transcripts. *Journal of Applied Genetics*, *44*(1), 1–19.

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 matrcing to exon or intron regions.
    1.3. Match_genome: Chromosome distributions of small RNA reads
    1.4. Match_ncgb: Statistical analysis for small RNAs matching to NCGB (National Crop Gene Bank).
    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.
    1.8. Prediction: Position and first nucleotide bias analysis for novel miRNA predictions, and novel miRNA prediction results.
    1.9. Primary: Lenght 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_intr_sense.txt: Small RNAs matching intron sense strand.
   overlap_intr_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 statisti

APPENDIX B: CLUSTER ANALYSIS OF MIRNAS

Supplementary Material 3 folder includes three excel files containing cluster analysis results of Bgh, Pst aand 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 thirth 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 milRNA 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 milRNA selections for Pst, Bgh and PI. All of the results are organized in excel files.

   miRNA_Acc: miRNA name
   Target_Acc: Accesion ID of target gene
   Expectation: Score for complementing miRNA and its target. Score is calculated by algortihm of miRU.
   UPE: It is the maximum energy (delta G) to open hairpin structure of the miRNA. Less energy means more possiblity 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 te 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 milRNAs 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 algortihm of miRU.
   UPE: It is the maximum energy (delta G) to open hairpin structure of the miRNA. Less energy means more possiblity to target the region.
   miRNA aligned fragment: miRNA sequence fragment aligning to the targe region
   Target aligned fragment: Target region aligning to the miRNA sequence
   Target Descreption: Target gene annotations
   Target Organism: Target gene Organism

APPENDIX D: PLANT DEFENCE PROTEIN KEYWORDS

Plant defence proteins as keywords used in target gene analysis.

Ribosomal protein
Malate dehydrogenase
Folylpolyglutamate 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
LysR family protein
E3 ubiquitin ligase
Fatty acid metabolism
Oxidase
Protease
Cer5
Reverse transcriptase
YibE family protein precursor
Avra10
TE1a
Calcium
Senescence
Ig lambda chain
Jasmonate
Condensin subunit ScpA
Glycosyl transferase
HOX family protein
Fibroin
Porphobilinogen deaminase
Protogenin precursor

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

1. Perl Script to construct All-vs-All Matrix by Needleman-Wunsch 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/3.1+Global+Alignment+Needleman-Wunsch/.

```
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
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;
```

```
      #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
$"="\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" ;
      }
      print OUTPUT "\n";
}
print "$#blastmatrix\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; $j <= length($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} + $GAP;
                  if ($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-back
      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
$"="\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" ;
      }
```

```
      print OUTPUT "\n";
}
print "$#blastmatrix\n";
```

3. Perl Script to count 3-mers and build a feature matrix from FASTA files of miRNA sequences. The code is self-written.

```
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;
      }
      else
      {
            $kmer_matrix[$i][$j]{score}=0;
      }
}
}
 no warnings 'uninitialized';
#write the distance/blast matrix on outfile
$"="\t";
print OUTPUT "matrix\t@kmer_array\n";
for(my $i = 0; $i < $#headers+1; $i++)
{
```

```
   print OUTPUT "$headers[$i]\t";
   for(my $j = 0; $j < 64 ; $j++)
   {
      print OUTPUT "$kmer_matrix[$i][$j]{score}\t" ;
   }
   print OUTPUT "\n";
}
```

4. Perl Script to calculate TAM percentage enrichments of the groupings. The code is self-written.

```
use strict;
use warnings;
# parameters
my $pvalue_Control = 0.005;
my $percentage_Control = 0.2;
my $SumClust=0;
my $SumFunc=0;
my $SumFam=0;
my $SumHM=0;
my $SumTissue=0;
my $SumAll =0;
my $output_file="output.txt";
open(OUTPUT,">".$output_file);
my $directory = "C:/strawberry/perl/bin/4txtdir/";
opendir DIR, $directory ;
my @files = grep{ /\.txt/} readdir (DIR) ;
closedir DIR;
my $groupnumber=$#files+1;
foreach my $file (@files)
{
      my (@fields,@mydata);

    open (IN, $directory.$file);
      my $headline= <IN>;
    while (<IN>)

    {
        chomp;
        @fields = split (/\,/,$_) ;

            if ($#fields==7)
            {
                  push(@mydata,@fields);
            }
            else
            {
                  my $val2 = $#fields - 7;
                  splice @fields,1, $val2;
            }
    }
    close(IN);
      my (@category, @term , @count, @percent, @fold ,@pvalue,
@bonferroni, @FDR);
```

```
      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 /\n/,$text ;
close (IN);
open(OUT, '>', "outfile.txt") or die "Couldn't open: $!";
foreach (@a)
{
      if       ($_       =~/Ribosomal       protein/       |/Malate
dehydrogenase/|/Folylpolyglutamate             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/
```

```
|
/LysR family protein/ |/E3 ubiquitin ligase/ |/Fatty acid
metabolism/ |/oxidase/ |/protease/ |/Cer5/ |
/Reverse transcriptase/ |/YibE family protein precursor/
|/Avra10/|/TE1a/ |/Calcium/|/Senescence/ |/Ig lambda chain/|
/jasmonate/ |/Condensin subunit ScpA/ |/Glycosyl transferase/
|/HOX family protein/ |/Fibroin/|/Porphobilinogen deaminase/ |
/Protogenin precursor/ |/Beta-fructofuranosidase/ |/Auxin/ |/Trev/
|/peroxidase/ |/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 NodT/ |/Caleosin/|
/ACC synthase/|/Crossover junction endodeoxyribonuclease RuvC/
|/TE4/ |/PHD domain containing/|/peptidase U7 family precursor/ |
/Phosphoribosylformylglycinamidine/) #####Defense Related
KeyWords
      {
            print OUT $_ . "\n";
      }
}
close (OUT);
```

6. The perl code to read MCL clustering results. The code is self-written.

```
open (my $inFile, '<', 'SW.txt') or die $!;
while (<$inFile>) {
  push(@array,split /\n+/);
  #print "@array\n";
}
close ($inFile);
open OUT, ">MCL-groups.txt";
$group==0;
my @headers = split('\s', $array[0]);
for ($n=1;$n<=666;$n++){
      @gec=();
      @gec=split('\s',$array[$n]);
      $group++;
      foreach $el (0..$#gec)
            {
                  if ($gec[$el]==1)
                  {
                        print OUT "$headers[$el]\t $group\n";
                  }
                  else{}
            }
}
```

APPENDIX F: R CODES

1. R code for K-means clustering. The code is adapted from K-means functions in stats library of R package and with help of the site Quick R http://www.statmethods.net/advstats/cluster.html.

```
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.

```
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.

```
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,sotaC1$clust,method = "euclidean")
sink("C:\\ws\\sota-clusters.txt")
lapply(sotaC2$clust,print)
sink()
```

4. R code for MCL clustering. MCL functions are directly taken from the site http://www.bigre.ulb.ac.be/Users/jvanheld/BMOL-F-501/practicals/r_scripts/mcl.R written by Slyvain Brohee.

```
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);
}
collect.mcl.clusters <- function (M      # Matrix (mcl result)
) {
  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.

Allium_cepa (Onion), unigene, DFCI Gene Index (ONGI), version 2,released on 2008_07_17
Arabidopsis thaliana, transcript, removed miRNA gene, TAIR, version 10,released on 2010_12_14
Arabidopsis thaliana, unigene, DFCI Gene Index (AGI), version 15,released on 2010_04_08
Arabidopsis thaliana, genomic DNA, 3.4K segments from strand with 0.4K overlapped region, TAIR, released on 2004_01_22
Aquilegia (columbine), unigene, DFCI Gene Index (AQGI), version 2.1,released on 2008_06_06
Beta vulgaris (beet), unigene, DFCI Gene Index (BVGI), version 4,released on 2011_03_17
Brachypodium distachyon (purple false brome), transcript, JGI genomic project, Phytozome, phytozome v8.0,internal number 142
Brachypodium distachyon (purple false brome), unigene, DFCI Gene Index (BDGI), version 1,released on 2010_05_26
Brassica napus (rape), unigene, DFCI Gene Index (BNGI), version 5,released on 2011_03_18
Brassica rapa (turnip, turnip rape, fast plants, field mustard, or turnip mustard), cds, de novo scaffolds assembly, version 1.1, released on 2011_08_30
Capsicum annuum, 454 unigene, 454 transcripts assembly, Capsicum Transcriptome DB, unknown version
Capsicum annuum (Pepper), unigene, DFCI Gene Index (CAGI), version 4,released on 2009_05_21
Carica papaya (Papaya), unigene, DFCI Gene Index (CAPAGI), version 1,released on 2010_05_27
Carica papaya (Papaya), transcript, JGI genomic project, Phytozome, phytozome v8.0,internal number 113
Chlamydomonas reinhardtii, unigene, DFCI Gene Index (CHRGI), version 8,released on 2011_03_21
Chlamydomonas reinhardtii, transcript, JGI genomic project, Phytozome, phytozome v9.0,internal number 236
Chlamydomonas reinhardtii, mRNA, augustus gene model prediction, JGI-Assembly v4, augustus u9
Cicer arietinum (chickpea), transcriptome sequence, nipgr
Cicer arietinum (chickpea), cds, ICGGC draft genome sequencing project, version 1
Citrus clementina (Clementine), unigene, DFCI Gene Index (CICLGI), version 2,released on 2009_05_22
Citrus sinensis (Orange), unigene, DFCI Gene Index (CSGI), version 1,released on 2008_06_25
Coffea canephora (coffee), unigene, SGN unigene, SGN, version 3
Coffea canephora (coffee), unigene, DFCI Gene Index (COCAGI), version 3,released on 2011_03_18
Cucumis sativus (cucumber), cds, Cucumber genome sequencing project, version 2
Cuscuta pentagona, unigene, Cuscuta pentagona, de-novo assembly of NGS RNA-seq reads, Sinha Lab,Department ofPlant Biology, Life Sciences Addition,UC Davis
Ectocarpus siliculosus (Brown algae), unigene, DFCI Gene Index (BAGI), version 1,released on 2010_05_27
Eucalyptus grandis (Flooded gum or Rose gum), transcript, JGI genomic project, Phytozome, phytozome v8.0,internal number 201
Eucalyptus grandis (Flooded gum or Rose gum), unigene, EST contig,
Euphorbia esula (Leafy spurge), unigene, DFCI Gene Index (EUESGI), version 1,released on 2008_06_30
Festuca arundinacea (tall fescue), unigene, DFCI Gene Index (FAGI), version 3,released on 2010_04_07
Festuca pratensis (Meadow ryegrass), unigene, DFCI Gene Index(MRGI), version 1,released on

2010_05_28
fusarium oxysporum (fusarium oxysporum), transcript, Broad institute, 4287
Glycine max (soybean), unigene, DFCI Gene Index (GMGI), version 16,released on 2011_03_31
Glycine max (soybean), transcript, JGI genomic project, Phytozome, phytozome v8.0,internal number 189
Gossypium (cotton), unigene, DFCI Gene Index (CGI), version 11,released on 2011_03_22
Gossypium raimondii (cotton_raimondii), unigene, DFCI Gene Index (GORAGI), version 1,released on 2008_07_02
Gossypium raimondii (cotton_raimondii), transcript, JGI genomic project, Phytozome, phytozome v8.0,internal number 211
Gossypium raimondii (cotton_raimondii), cds, Chinese Academy of Agricultural Sciences,Cotton Research Institute
Helianthus annuus (Sunflower), unigene, DFCI Gene Index (HAGI), version 6,released on 2009_05_24
Homo sapiens (human), transcript, Human genomic sequencing project,
Hordeum vulgare (barley), unigene, DFCI Gene Index (HVGI), version 12,released on 2011_03_19
Ipomoea nil (Morning glory), unigene, DFCI Gene Index (IPNIGI), version 1,released on 2008_06_30
Lactuca sativa (Lettuce), unigene, DFCI Gene Index (LSGI), version 3,released on 2008_07_02
Lactuca serriola (Prickly lettuce), unigene, DFCI Gene Index (LASEGI), version 1,released on 2008_06_28
Linum usitatissimum (flax or linseed), unigene, Unigene Library,
Linum usitatissimum (flax or linseed), transcript, JGI genomic project, Phytozome, phytozome v7.0,internal number 200
Lotus japonicus (lotus), unigene, DFCI Gene Index (LJGI), version 6,released on 2010_05_18
Magnaporthe oryzae (rice blast fungus), transcript, Broad institute, Magnaporthe grisea Database, Magnaporthegrisea Assembly 6
Malus x domestica (apple), cds, predicted consensus gene set CDS sequences, , version 1
Malus_x_domestica (apple), unigene, DFCI Gene Index (MDGI), version 3,released on 2010_04_08
Manihot esculenta (Cassava), unigene, DFCI Gene Index (MAESGI), version 1,released on 2010_05_28
Medicago truncatula (Barrel Medic), transcript, Mt3.5v4 splice transcripts
Medicago truncatula (Barrel Medic), transcript, Mt4.0v1 spliced transcripts, IMGAG, Mt4.0V1
Mesembryanthemum crystallinum (ice plant), unigene, DFCI Gene Index (MCGI), version 5,released on 2008_06_19
Mimulus guttatus (Spotted monkey flower), unigene, DFCI Gene Index (MIGUGI), version 1,released on 2010_05_27
Morchella esculenta (common morel), transcript, JGI genomic project, Phytozome, phytozome v9.0,internal number 147
Musa acuminata (banana), cds, all gene coding sequences, Banana Genome Hub, version 1
Nicotiana benthamiana, unigene, DFCI Gene Index (NBGI), version 4,released on 2010_04_09
Nicotiana benthamiana, transcript, draft genome 0.4.4, ,
Nicotiana tabacum (tobacco), unigene, DFCI Gene Index (NTGI), version 7,released on 2011_04_01
Nicotiana tabacum (tobacco), unigene, SGN unigene,
Oryza sativa (rice), transcript, TIGR genome cDNA OSA1 Release 5(OSA1R5), version 5
Oryza sativa (rice), unigene, DFCI Gene index (OSGI), version 19,released on 2011_03_28
Oryza sativa (rice), genomic DNA, 3.4K segments from strand with 0.4K overlapped region, ,
Oryza sativa (rice), transcript, MSU Rice Genome Annotation, version 7
Oryza sativa (rice), transcript, JGI genomic project, Phytozome, phytozome v7.0,internal number 193
Oryza sativa (rice), transcript, cDNA library, Ensembl,
Panicum virgatum (switchgrass), unigene, DFCI Gene Index (PAVIGI), version 2,released on 2011_03_24

Petunia hybrida (Petunia), unigene, DFCI Gene Index (PHGI), version 3,released on 2011_03_17
Phaseolus coccineus (Scarlet bean), unigene, DFCI Gene Index (PCGI), version 1,released on 2009_05_27
Phaseolus vulgaris (bean), unigene, DFCI Gene index (PHVGI), version 4,released on 2011_03_24
Phoenix dactylifera (date palm), transcript, Date Palm Genome Draft Assembly, Version 3
Phyllostachys heterocycla (Moso Bamboo), cds, National Center for Gene Research of Chinese Academy of Sciences, Version 1
Physcomitrella patens (moss), transcript, JGI genomic project, Phytozome, phytozome v7.0,internal number 152
Physcomitrella patens (moss), unigene, DFCI Gene Index (PPSPGI), version 1,released on 2009_05_26
Picea (Spruce), unigene, DFCI Gene Index (SGI), version 5,released on 2011_03_30
Pinus (pine), unigene, DFGI Gene Index (PGI), version 9,released on 2011_03_26
Populus trichocarpa (poplar), unigene, DFCI Gene Index (PPLGI), version 5,released on 2010_04_16
Populus trichocarpa (poplar), genomic DNA, Genome Assembly R1.0, 3.4K segments from strand with 0.4K overlapped regions
Populus trichocarpa (poplar), transcript, JGI genomic project, Phytozome, phytozome v9.0, internal number 218
Populus trichocarpa (poplar), transcript, JGI genomic project, Phytozome, phytozome v8.0, genome V3.0, internanal number 210
Populus trichocarpa (poplar), transcript, JGI genomic project, Phytozome, phytozome v7.0,internal number 156
Prunus persica (peach), unigene, DFCI Gene Index (PRPEGI), version 2,released on 2009_05_26
Quercus (oak), unigene, DFCI Gene Index (OGI), version 2,released on 2011_03_23
Raphanus sativus (Radish), unigene, DFCI Gene Index (RSGI), version 1,released on 2010_05_27
Saccharum officinarum (sugarcane), unigene, DFCI Gene Index (SOGI), version 3,released on 2010_04_09
Secale cereale (rye), unigene, DFCI Gene Index (RYEGI), version 4,released on 2008_07_03
Selaginella moellendorffii (club mosses), unigene, DFCI Gene Index (CMGI), version 1,released on 2010_05_26
Setaria italica (foxtail millet), cds, BGI sequencing project,
Setaria italica (foxtail millet), cds, JGI genomic project, phytozome v9.0,internal number 164
Solanum Lycopersicon (tomato Lycopersicon), unigene, Lycopersicon Combined (Tomato) Unigenes, SGN Unigene, unknown version
Solanum Lycopersicon (tomato Lycopersicon subgenus), transcript, SGN mRNA sequences, SGN genomic sequencing project, released on 2011_05_08
Solanum lycopersicum (tomato), transcript, cDNA library, version 2.3
Solanum lycopersicum (tomato), unigene, SGN unigene,
Solanum lycopersicum (tomato), unigene, DFCI Gene Index (LGI), version 13,released on 2010_04_13
Solanum melongena (eggplant), unigene, DFCI Gene Index (SOMEGI), version1,released on 2010_05_28
Solanum tuberosum (potato), unigene, DFCI Gene Index (STGI), version 13,released on 2010_04_16
Solanum tuberosum (potato), transcript, Group Phureja DM1-3 516R44 (CIP801092) Genome 3.4 transcripts, ,
Sorghum bicolor (Sorghum), unigene, DFCI Gene Index (SBGI), version 9,released on 2008_07_25
Sorghum bicolor (Sorghum), transcript, JGI genomic project, Phytozome, phytozome v9.0,internal number 79
Striga hermonthica (Purple witchweed), unigene, DFCI Gene Index (SHGI), version 1,released on 2010_05_27
Thellungiella halophila (salt cress), transcript, JGI genomic project, Phytozome, phytozome v9.0,internal number 173
Thellungiella halophila (salt cress), unigene, ESTContig on PlantGDB, PlangGDB unigene,

released on 2009_01_16
Theobroma cacao (cocoa), unigene, DFCI Gene Index (TCAGI), version 3,released on 2009_05_21
Triphysaria, unigene, DFCI Gene Index (TRIPHGI), version 1,released on 2008_08_15
Triphysaria versicolor, unigene, DFCI Gene Index (TVERGI), version 2,released on 2008_06_28
Triticum aestivum (wheat), unigene, DFCI Gene Index (TAGI), version 12,released on 2010_04_18
Vigna unguiculata (cowpea), unigene, DFCI Gene Index (VUGI), version 1,released on 2010_05_27
vitis vinifera (wine grape), transcript, Wine grape genomic sequencing project, V1
Vitis vinifera (grape), unigene, DFCI Gene Index (VVGI), version 8,released on 2011_04_02
Vitis vinifera (grape), transcript, JGI genomic project, Phytozome, phytozome v9, internal number 145
Volvox carteri (green algae), unigene, DFCI Gene Index (VCGI), version 1, released on 2010_05_28
Zea mays (maize), unigene, DFCI Gene Index (ZMGI), version 19,released on 2009_06_05
Zea mays (maize), cds, PlangGDB genomic project,
Zea mays (maize), transcript, NSF-funded Maize Genome Sequencing Project, Release 5a , filtered set