Identifying Conserved microRNAs in a Large Dataset of Wheat Small RNAs

by

Md Safiur Rahman Mahdi

A thesis submitted to
The Faculty of Graduate Studies of
The University of Manitoba
in partial fulfillment of the requirements
of the degree of

Master of Science

Department of Computer Science
The University of Manitoba
Winnipeg, Manitoba, Canada
June 2015

Thesis advisor

Michael Domaratzki

Author

Md Safiur Rahman Mahdi

Identifying Conserved microRNAs in a Large Dataset of Wheat Small RNAs

Abstract

MicroRNAs (miRNAs) play a vital role in regulating gene expression. Detecting conserved and novel miRNAs in very large genomic datasets generated using next generation sequencing platforms is a new research area in the field of gene regulation, but finding useful miRNA information from a large wheat genome is a challenging research project. We propose to design a toolchain that will identify conserved miRNAs using various software tools such as Basic Local Alignment Search Tool (BLAST), Bowtie 2, MAFFT and RNAfold. Our toolchain identified 36 wheat conserved miRNA families that matched with 232 experimental sequences. Moreover, we found 87 plant conserved miRNA families that matched between 613 experimental sequences and the miRBase dataset. In addition, we observed significant differential expression for the wheat exposed to the heat stress compared to those exposed to light and UV stresses or no stress (control).

Contents

List of Figures

Acknowledgments

First of all, I would like to thank my supervisor, Dr. Michael Domaratzki, for his help and advice for the entire duration of my masters program. I would not be able to complete my thesis without his continuous guidance and directions. I would also like to acknowledge the financial support I received from the department of Computer Science in the form of a Guaranteed Funding Package and would like to thank Dr. Michael for the assistance provided.

Secondly, I would like to thank Raja Ragupathy for his valuable times and efforts in my thesis through valuable suggestions and feedback.

Thirdly, I like to thank fellow lab-mates in the Bioinformatics lab who have always been very generous to support me in various ways. Mamun Sharif, Maryam Ayat, Graham Alvare, Justin Zhang for their ideas, comments, feedback and help. I also like to thank Md. Mahfuzur Rahman and Tanzeem Bin Noor for their help and advice.

Last but not least, I would like to acknowledge the unconditional love and support of my parents. I also like to thank my wife, Farhana Islam, for her support.

This thesis is dedicated to my parents and my lovely wife.

Chapter 1

Introduction

Wheat plays an important role in the Canadian economy and is an 11 billion dollar industry in Canada [NRCC, 2015]. Thus, research on wheat breeding has great significance. To improve the breeding of wheat, researchers need to produce better varieties that are resistant to stresses such as heat, excess light and excess ultraviolet (UV) light. These stresses can be seen as the result of climate change.

To improve the breeding of wheat in response to climate change, we need to research wheat genes, and microRNAs (miRNAs), which act as gene regulators. The process of converting DNA to protein is known as gene expression and gene regulators control the expression of other genes (the definition of genes is given in Section 2.1.1). Moreover, we must explore what influence different miRNAs have on wheat during stresses to improve wheat varieties' ability to cope with climate change. In particular, we can examine the relationship between miRNA expression and stress by selecting varieties of wheat that can react to stress.

MiRNAs act as gene regulators by influencing which proteins are created. Proteins

form the fundamental molecules of all the machinery of a cell. The flow of information from Deoxyribonucleic Acid (DNA) to messenger Ribonucleic Acid (mRNA) (transcription) and mRNA to protein synthesis (translation) is known as the central dogma of molecular biology [Crick, 1970]. The whole process is described in Section 2.1.2. During the transcription process, besides mRNA synthesis, portions of DNA are also transcribed into non-coding RNA (ncRNA), which are not translated into proteins. MiRNAs are one category of ncRNAs that act as post-transcriptional regulatory molecules by preventing protein creation. The detailed process is described in Section 2.1.4. For example, wheat genes Ta.5303 and Ta.39646 are the predicted targets of miRNA 504 and miRNA 519 [Yao et al., 2007]. When miRNA 504 is expressed, the gene Ta.5303 is predicted to be down regulated. Thus miRNAs act as gene regulators controlling transcript accumulation [Chen and Rajewsky, 2007].

Genes can be differentially expressed in humans, animals or plants based on various internal cues as per the growth and developmental blueprint of the organism and as per conditions such as biotic (fungi, bacteria, etc.) or abiotic (temperature, light, ultraviolet ray, etc.) stresses or states. For example, genes may be differentially expressed in plants exposed to heat stress or affected by diseases. As miRNA expression can be changed rapidly in response to stress, miRNAs act as gene regulators. In particular, we can examine how miRNA expression can change in response to stress by selecting varieties of wheat that can react to stress.

To improve the breeding of wheat in response to climate change, we need to produce better varieties of wheat that are resistant to stresses. We obtained experimental data for bread wheat (*Triticum aestivum*) exposed to three stresses and for

plants grown under control condition to help determine which miRNAs are expressed differently under different stresses due to climate change.

We used a wheat small RNA dataset of approximately 21GB, generated from leaf samples collected from wheat plants subjected to: heat, light or UV stresses or no stress (control). Ninety-six *Triticum aestivum cv* Glenlea plants were grown in a growth cabinet (Conviron Technologies, Winnipeg, Canada) under long day conditions (16 hours light at 18 °C and 8 hours darkness at 16 °C). Growth cabinets provide a variety of temperatures and lighting patterns that are essential for plant growth research. At advanced boot leaf stage, three batches of 24 plants each were exposed to three different stresses under controlled growth conditions: continuous light for three days; heat stress (at 37 °C for 72 hours) and UV stress (2 minutes of exposure to UV light for 3 consecutive days). Twenty-four plants were grown as control. Leaf tissues were collected at six time points after the end of the stress period: day 0, 1, 2, 3, 7 and 10.

With this data set, we identified conserved miRNAs to help determining which miRNAs are expressed differently under different stresses. We used a toolchain including various software such as Basic Local Alignment Search Tool (BLAST) [Altschul et al., 1990], Bowtie 2 [Langmead and Salzberg, 2012], MAFFT [Katoh et al., 2002] and RNAfold [Zuker and Stiegler, 1981]. The software are described in the Related work section and details are described in the Solution methodology section.

We identified 36 wheat conserved miRNA families that matched 232 experimental sequences and datasets from two recent papers by Mayer et al. [2014] and Sun et al. [2014]. Moreover, we found 87 plant conserved miRNA families that matched 613

experimental sequences in the miRBase [Kozomara and Griffiths-Jones, 2014] dataset.

In addition, we observed significant differential expression for the wheat exposed to heat stress compared to those exposed to light and UV stresses or no stress (control). Thirty-four conserved miRNA families were differentially expressed for the heat stress whereas only 8 conserved miRNA families were differentially expressed for light and only 7 conserved miRNA families were differentially expressed for UV stress. We also found that increasing number of days post treatment affect the number of conserved miRNA families differentially expressed for control versus stresses. Again, different conserved miRNA families expressed differently based upon stress or day. We found that miRNA 395 and 398 were strongly suppressed whereas miRNA 5064, 5175, 2020, and 1439 were expressed with heat stress at all post-stress time points. MiRNA 395 was suppressed in all stresses samples regardless of the stress or time point of the stress.

Chapter 2

Background

To understand various terms and concepts of my thesis, we separate the background chapter in two different sections: biology and bioinformatics.

2.1 Biology

In this section, we describe basic concepts of biology: DNA, DNA to protein, miRNA, and the process of prevents protein creation by decreasing transcription process by miRNA.

2.1.1 DNA

All living organisms contain DNA as hereditary material. DNA is a linear molecule consisting of a sequence of four nucleotides or bases: adenine (A), cytosine (C), guanine (G) and thymine (T). In DNA, these nucleotides form base pairs (bp) by making bonds with each other: A pairs with T and C pairs with G. DNA is double-

stranded and forms a helix structure. This structure of DNA was first described by James Watson and Francis Crick [Watson and Crick, 1953]. Figure 2.1 shows a figure representing the structure of a DNA helix.

Genes are the portions of DNA coding sequences that hold the physical and functional unit of heredity [Pearson, 2006]. A gene codes for a polypeptide which is described in the following Section. Genes can span a few hundred DNA bases to more than 2 million bases [Medicine, 2015]. The complete set of an organism's DNA is called the genome. In general, genes occupy a very small portion of the genome. For example, the human genome contains 30000 genes, which is only 1.5% of the entire genome [Annenberg, 2015]. The rest of the genome consists of the repeated sequences and DNA that regulates genes.

The wheat genome is ∼17000 million base pair (Mb) in size which is very large compared to the genome size of the model plant species of *Arabidopsis thaliana* (∼130 Mb) and the human genome which is ∼2800 Mb in size.

2.1.2 Gene expression: DNA to protein

Gene expression is the process of converting DNA to RNA (using transcription process), and then RNA to protein (using translation process). Protein is a combination of 20 different amino acids that forms the fundamental molecules of the machinery of a cell. At first, DNA is converted into mRNA, and this process is called *transcription*. Besides mRNA synthesis, portions of DNA are also transcribed into ncRNA such as miRNAs, transfer RNAs (tRNAs), and ribosomal RNAs (rRNAs), which are not translated into proteins.

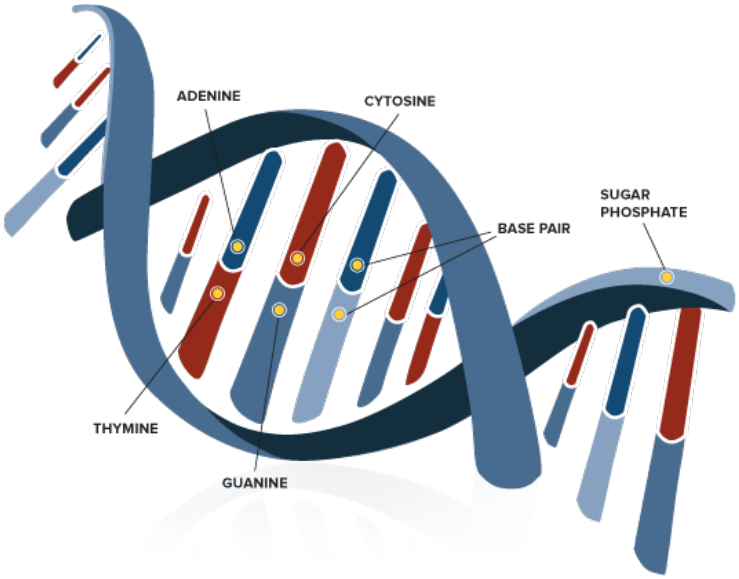

Figure 2.1: DNA structure [Geographic, 2015].

A region of DNA that initiates transcription of a particular gene is known as a promoter region. Promoter regions are recognized by transcription factor (TF). As opposed to TF, repressors are DNA or RNA binding proteins that block transcription or translation.

In the transcription process, double stranded DNA is copied into single stranded mRNA where thymine (T) is replaced by uracil (U). Next, mRNA is translated into polypeptide. A combination of three consecutive nucleotides is called a codon and each codon in mRNA is translated into one amino acid. The process of converting mRNA to protein is known as translation. Translation starts with an initiation codon named start codon and ends with a terminator codon named stop codon.

Figure 2.2 shows how protein is created from DNA. At first, by the transcription process, DNA is converted to mRNA and mRNA is converted to polypeptide by

the translation process. As an example, in the Figure, at first DNA (ACCGTG...) is converted to mRNA (ACCGUG...) and then the codon ACC is translated to the amino acid Threonine (T), the codon GUG is translated to the amino acid Valine (V), and the same process is applicable for the rest of the DNA sequence. The nucleotide triplets of DNA and RNA molecules that correspond amino acids is known as the genetic code.

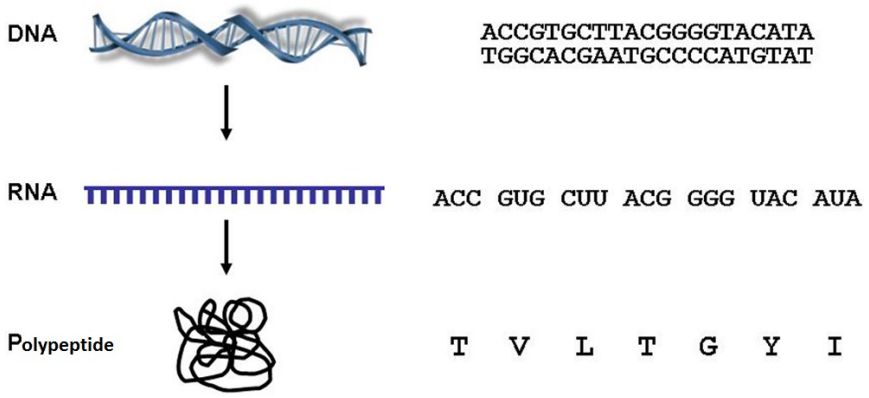

Figure 2.2: From DNA to protein [Burdine and Sheldon, 2015].

2.1.3 MiRNA

In general, mature miRNAs are 17-24 nucleotide (nt) long single-stranded sequences resulting from longer (around 1000 nt) primary transcripts (pri-miRNAs) [Kusenda et al., 2006], which are one category of ncRNAs generated from the transcription process. After transcription, with the help of the enzyme Dicer in plant, pri-miRNAs form approximately 70-600 nt precursor miRNAs (pre-miRNAs) which are folded into hairpin-shaped structures [Kim, 2005]. Then the pre-miRNAs are cleaved into two complementary stranded sequences of mature miRNAs (17-24 nt)

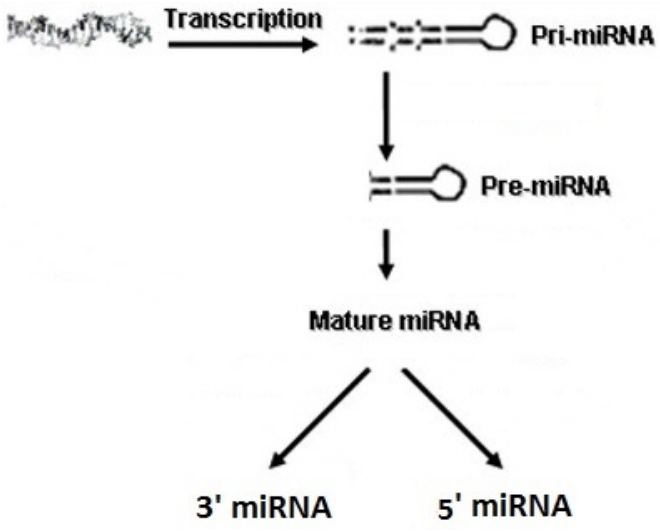

Figure 2.3: From pri-mRNA to mature miRNA [Papagiannakopoulos and Kosik, 2008].

where one strand is the 3-prime mature miRNA and the other strand is the 5-prime mature miRNA [Krol et al., 2004]. Figure 2.3 shows an illustration of the process.

The two strands of mature miRNA are the 3-prime mature miRNA and 5-prime mature miRNA. Between the 3-prime mature miRNA and 5-prime mature miRNA, one of them is called the mature or guided strand and the other is known as the miRNA star strand. Generally, mature and star strands are complementary to each other. However, mismatches may occur between the mature and star strand. Moreover, there is also a two base pair overhang between the mature and star strand. See Figure 2.4 for more details.

In Figure 2.4, the whole sequence represents a partial structure of pre-miRNA that contains the sequence of two complementary strands. The mature strand is represented with the red colored sequence and the green colored sequence represents the miRNA star strand. There are two-bp overhangs between the mature and star

strand at both ends of the pre-miRNA. In this thesis, to predict the star strand, we allowed up to 4 mismatches between the mature and star strand [Michael Axtell, personal communication].

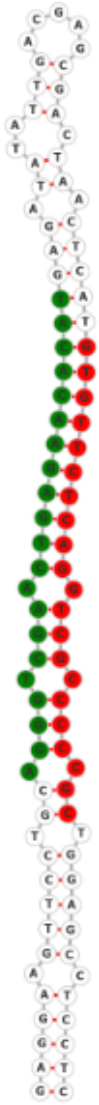

Figure 2.4: Structure of a pre-miRNA containing a mature miRNA sequence and a miRNA star sequence. The green coloured sequence represents the mature miRNA strand and the miRNA star strand is represented with the red coloured sequence.

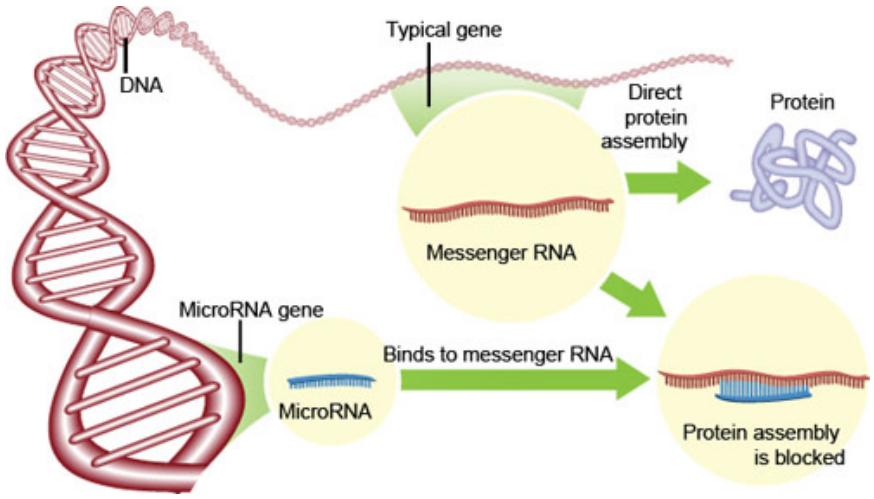

Figure 2.5: Impact of miRNA on protein creation [Papagiannakopoulos and Kosik, 2008].

2.1.4 MiRNA: blocking protein creation

As described in Chapter 1, in addition to mRNA, some ncRNAs such as miRNAs, tRNAs and rRNAs are also generated by the transcription process. These ncRNAs are not translated into proteins. MiRNAs are one of the ncRNAs that act as post-transcriptional regulatory molecules by cleaving and attaching to complementary regions of mRNAs. Similar to mRNA, miRNA is also a single stranded sequence. As shown in Figure 2.5, miRNA searches for the complementary regions of mRNA [Ruvkun, 2001] and binds to it. Thus it prevents protein creation.

2.1.5 Assembly, read and contig

A genome assembly is the genome sequence made after chromosomes have been split [Ensembl, 2015]. Chromosomes are located inside the nucleus, mitochondria or chloroplast of the cells and contain most of the DNA of a living organism. The

chromosomes are split into smaller parts as we can only read short sequences. In my thesis, we used a Low Copy-number Genome (LCG) assembly of wheat [Brenchley et al., 2012]. In a genome assembly, the same sequences can be repeated several times; the number of times the same sequence is repeated is called copy number. The genome can be subdivided into 3 constituents : i) low copy regions, ii) moderately repetitive regions, and iii) high copy regions (for example centromere, where a single repeat will be present thousands of time). LCG assembly is made of low copy regions like genes.

Sequence reads represents a series of nucleotides (A, T, C, G) obtained by a sequencer. The length of a read can vary according to the technology used to derive it. For example, Illumina short read are 50 bp to 250 bp while Pac-bio read can be 10Kb nowadays.

A sequence of DNA representing overlapping molecules is known as a contig [Information, 2015]. In my thesis, we used a wheat genome sequence [Science, 2014] provided by Mayer et al. [2014] where the genome sequences are represented as contigs. Moreover, the contig sequences are categorized by chromosome arms. Each chromosome has a binding point called the centromere, which divides the chromosome into two sections, or arms: the short arm (denoted by S) and the long arm (denoted by L) [Medicine, 2014].

2.2 From biology to informatics

In this section, we will describe the important terms and concepts that are necessary to understand the following chapters.

2.2.1 Sequence alignment

Sequence alignment is the process of comparing two or more sequences (DNA, RNA, or protein) based upon sequence similarity either locally or globally. The best match (locally or globally) between two sequences is known as pairwise sequence alignment where alignment of three or more sequences is known as multiple sequence alignment (MSA). The goal of sequence alignment is to find the homologous regions between the sequences by maximizing their similarity. Local alignments find matches with sub-strings or portions whereas global alignment matches end to end between two sequences. Figure 2.6 shows an example of global and local alignments. In the figure, the two sequences TCCCAGTTATGTCAGGGGACACGAGCATGCAGAGAC and AATTGCCGCCGTCGTTTTCAGCAGTTATGTCAGATC are located over each other horizontally to align both locally and globally where the vertical bars (|) between the sequences represent matches and the hyphens (-) represent gaps. In the global alignment the two sequences are matched end to end making many gaps but the local alignment finds the sub string CAGTTATGTCAG, which is the highest matching portions between the two sequences.

Figure 2.6: Local and global sequence alignment [Rosalind, 2015].

Between two sequences, there may be exact matches, some mismatches, or insertions or deletions (indels). This is due to mutation caused by evolution. Comparing

more than two sequences is known as multiple sequence alignment.

Sequence alignments provide a score that represents the similarity of the sequences in the alignment based on the mismatches, insertions or deletions between two sequences. Figure 2.7 shows examples of sequence alignment with perfect matches, mismatches and indels. For example, if match score is +2 and mismatch/indel penalty -1, then the score of the left example of the Figure is: 8*2 = 16 as there are 8 matches and the score of the right example of the Figure is: (5*2) + (3*-1) = 10-3 = 7 as there are 5 matches, 1 mismatch and 2 indels.

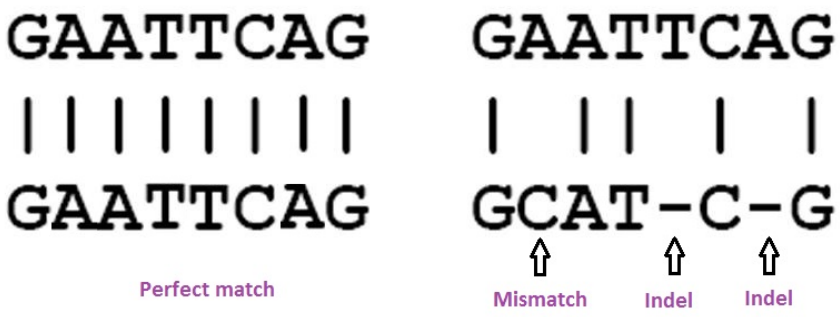

Figure 2.7: Sequence alignment: perfect match (left), mismatch and indels (right) [Stanford, 2015].

2.2.2 BLAST

BLAST [Altschul et al., 1990] is a software tool that allows a user to find similar sub-sequences between a query and a subject sequence. The sequence that we want to align is the query sequence and the sequence that the query is compared with is the subject sequence. BLAST identifies subject sequences that match with one of the query sequences up to a certain threshold or certain limit of similarity. The minimum threshold is also called the expect (e) value. The default e value is 10, which means

that 10 matches between the query and the subject sequences of the selected database of similar quality are expected to be found randomly [NCBI, 2015] even without any biological relatedness. Depending on the database size, e value cut off is determined. Lower e values produce more stringent results.

BLAST uses a heuristic method to find similar sequences. To acquire a perfect match between a query and a subject sequence, there is an option to configure match and mismatch scores between query and subject sequences in BLAST.

BLAST creates a k-letter word list for query sequences that contains sub-strings (words) of length k. In general, for protein query sequences, k is 3 and for DNA query sequences k is 11. BLAST categorizes the high-scoring words into an efficient search tree. These high-scoring words of the search tree are then scanned in the target sequences of the designated or selected database. Then, BLAST extends the exact matches to high-scoring segment pairs (HSP). The HSP contains the best matches between the query and subject sequences and the BLAST output is the results of the descending order of the HSPs. Figure 4.2 shows an example of how an HSP works.

As an example, in my thesis, we used nucleotide-nucleotide BLAST (blastn), where both the query and the subject sequences are nucleotide sequences. Other versions of BLAST are nucleotide-protein BLAST (blastx) and protein-protein BLAST (blastp). We can not use pairwise sequence alignment for this purpose as pairwise sequence alignment is too slow for processing the large volume of query (15158 species) and subject sequences. We used an Rfam database as the subject sequences where all miRNAs were removed from the Rfam-database [Burge et al., 2013]. Thus, the Rfam-database contains all the ncRNA sequences except miRNAs. Detailed process of

removing ncRNA sequences is described in Section 4.3. To identify the conserved miRNAs, we also used the miRBase database, release 20 (June, 2013), which contains the known conserved miRNAs of plants (see Section 2.2.3). A detailed description of this process is described in the Section 4.5.

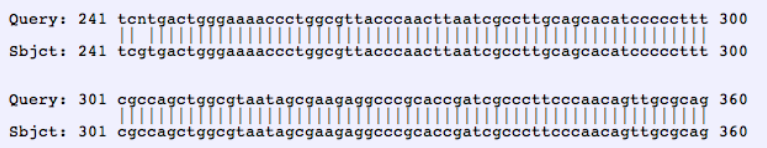

Figure 2.8: Sample partial BLAST output where the vertical lines between query and subject sequences represent matches [BLAST, 2015].

Figure 2.8 shows a sample partial output of BLAST where "|" between the two sequences represents a match and space (" ") between the sequences represents a mismatch between corresponding positions of the query and the target sequences.

2.2.3 MiRBase

MiRBase [Kozomara and Griffiths-Jones, 2014] is the biological database of all known conserved miRNAs along with their precursors. The conserved miRNAs are also known as conserved miRNA families. Mature miRNAs having significant similarity to entries in miRBase are called conserved miRNAs and undiscovered miRNAs are called novel miRNAs. The latest version of miRBase contains 30424 miRNAs from 206 species [Kozomara and Griffiths-Jones, 2014]. Figure 2.9 shows a sample miRBase entry [Garcia, 2015], where the whole sequence represents a pre-miRNA (Pre-miR-146b) that contains the sequence of two complementary strands. The upper arm (5' end) contains the 5-prime mature miRNA (miR-146b-5p), UGAGAACUGAAUUCCAUAGGCU and the lower arm (3' end) contains the 3-prime mature miRNA

(miR-146b-3p), UGCCCUGUGGACUCAGUUCUGG.

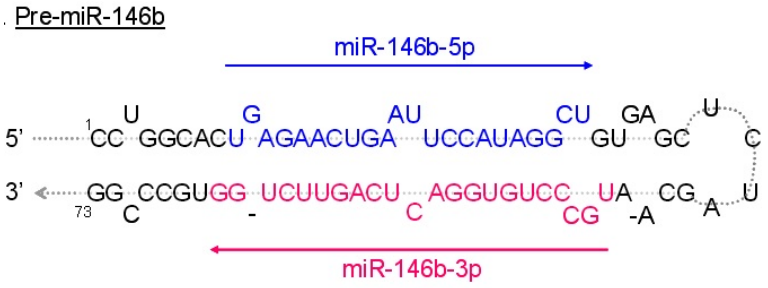

Figure 2.9: MiRNAs (hsa-miR-146b-3p and hsa-miR-146b-5p) and their precursor [Garcia, 2015].

2.2.4 Bowtie 2

Bowtie 2 [Langmead and Salzberg, 2012] is a very fast and memory-efficient sequence alignment algorithm. It is normally used to align short sequences against large reference genomes. It is used when a user has a set of query sequences and the corresponding genome sequence, to position the query sequences onto the reference genome sequence. Thus, Bowtie 2 takes a genome and a set of reads as input and outputs a list of alignments. Figure 2.10 shows an example of how Bowtie 2 works.

Bowtie 2 can access multiple processors at a time to achieve faster alignment speed [Langmead and Cole, 2015]. For example, it can align 35-base-pair reads to the human genome at a rate of 25 million reads per hour [Langmead and Cole, 2015]. We can not use BLAST for this purpose as our wheat genome sequence is very large in size and query sequences are small (around 20-24 bp).

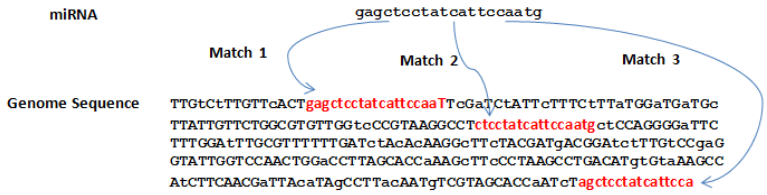

Figure 2.10: Alignment of miRNA and genome sequence using Bowtie 2.

2.2.5 MAFFT

Katoh et al. [2002] developed MAFFT, a multiple sequence alignment program using Fast Fourier Transformation (FFT). It works faster than other multiple sequence alignment program such as CLUSTALW and T-COFFEE. The authors used two novel techniques to make MAFFT extremely fast: FFT and a simplified scoring system where they designed the scoring matrix and gap penalty efficiently. Figure 2.11 shows a sample output of MAFFT.

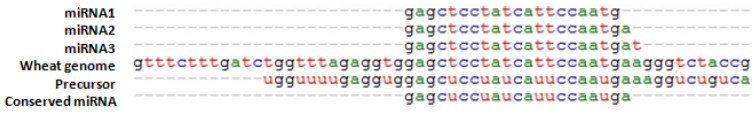

Figure 2.11: Multiple sequence alignment among query, conserved miRNA, genome and precursor sequences using MAFFT. These miRNA1, miRNA2 and miRNA3 come from the experimental dataset and the precursor and conserved miRNA comes from miRBase.

2.2.6 RNAfold

Single-stranded RNA molecules can fold and form base pair matches with itself to form secondary structure. The most common shape of secondary structure is either helix or loop, and for RNA secondary structure, one of the shapes looks like a stem-

loop shaped hairpin. For example, transfer RNA (tRNA) and precursors of miRNAs can form stem-loop shaped hairpin structures. Figure 2.12 shows the structure of the stem-loop. Figure 2.4 and 2.9 are also examples of miRNA precursor's stem-loop shaped hairpin structure.

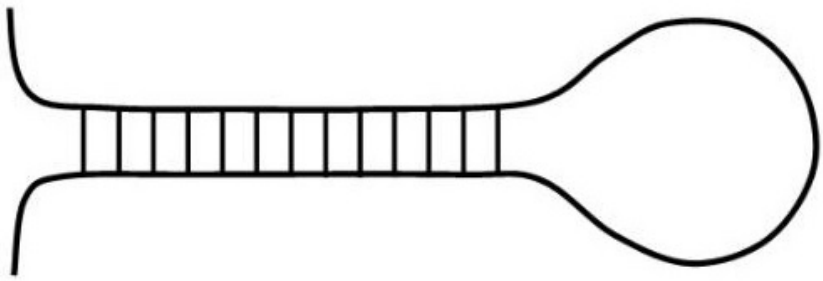

Figure 2.12: Secondary structure: stem-loop [Cronodon, 2015].

RNAfold predicts secondary structure of RNA [Lorenz et al., 2011]. It uses the Minimum Free Energy (MFE) method to predict secondary structure. Finding an energetically stable structure of RNA using the sequence known as the MFE method [Zuker, 1989]. RNAfold takes an RNA sequence as input and predicts a MFE structure of RNA, similar to Figure 2.4, as output. We used RNAfold to predict RNA secondary structure, specifically the stem-loop structure of conserved and novel miRNAs. The structure is predicted with the help of a loop-based energy model and the dynamic programming algorithm introduced by Zuker and Stiegler [1981].

2.2.7 Dot-bracket notation

Besides the secondary structure, RNAfold also produces a dot-bracket notation. The goal of the dot-bracket notation is to represent RNA secondary structure in a convenient way. Dot-bracket notation consists of dots ".", opening "(" and closing ")" parentheses. Each character represents a base. A dot "." represents an unpaired base, whereas open parenthesis "(" represents a base that is paired (5' end) to another base ahead of it (3' end) and closed parenthesis ")" represents a base that is paired (3' end) to another base behind it (5' end). Figure 2.13 shows an example of how a secondary structure can also be presented with dot-bracket notation.

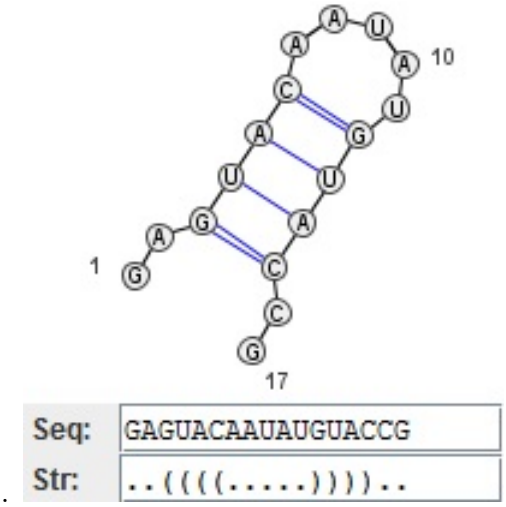

Figure 2.13: Dot-bracket notation and secondary structure [Darty et al., 2009]. At the 5' end we have nucleotides "GUAC", dot-bracket notation: "((((", that matches with "GUAC" at 3' end, dot-bracket notation: "))))", the rest of the nucleotides are unpaired.

Chapter 3

Related Work

To date, there has not been much research to identify conserved and novel microRNAs in plants. Yao et al. [2007] constructed an RNA database to identify conserved and novel miRNAs in wheat. At first they performed BLASTN with the Rfam [Burge et al., 2013] database to remove other ncRNA sequences such as ribosomal RNA (rRNA) and tRNA. This is similar to our technique in Section 4.3. Then, the authors used BLASTN to query the wheat Expressed Sequence Tag (EST) database from the National Center for Biotechnology Information (NCBI) to identify 58 miRNAs. This is not sufficient to predict miRNA as the miRNAs need precursor and star miRNA support. An EST database is a collection of short single-read complementary DNA (cDNA) that comes from the reverse transcription of mRNAs. Among the 58 miRNAs, 35 miRNAs were reported to be conserved and 23 were novel.

Yin and Shen [2010] studied 42 conserved miRNA families in plants. They also used BLASTN with the miRBase database and found 34 conserved miRNAs in wheat.

Additionally, Kurtoglu et al. [2014] identified 52 conserved miRNAs and 7 novel

miRNAs in wheat using the whole genome sequence. They performed BLAST with miRBase to identify miRNAs by sequence similarity. This is similar to our technique in Section 4.5. For miRNA prediction, the authors implemented a two step procedure: homology search to known plant miRNAs and consistency of pre-miRNAs. They also used LCG assembly of wheat.

Lei and Sun [2014] built miR-PREFeR, a plant miRNA prediction tool using small RNA-seq data that uses RNAfold to predict miRNA correctly by examining stem-loop structures.

For conserved miRNA identification, Mayer et al. [2014] analysed wheat contigs as query using the BLASTN algorithm with all plants' mature miRNA sequences from the miRBase database (release 18, November 2011) as subject with the following parameters: E-value 10, word size 7 and match reward 2. All hits were filtered using threshold criteria of maximum 4 mismatches between contig and miRBase sequence. In addition, matched sequences were analyzed for the intervening distance and two hits falling within a distance between 3 nucleotides to 239 nucleotides were retained. The two hits were a direct hit representing a potential mature miRNA and a reverse complementary hit representing a star sequence and vice versa. The intervening distance between this mature miRNA and its cognate star sequence could vary from 3 to 239 nucleotides in length according to Kadri et al. [2009]. Hence, the the loop region separating the mature miRNA and its star sequence could be varying from 3 bp to 239 bp to form an ideal secondary hairpin structure. From the retained sequences, potential precursor sequences were extracted with flanking sequences by cutting at positions which are 13 nucleotides upstream of the 5 hit and 13 nt downstream of the

3 hit, respectively. These sequences were folded using software NOVOMIR [Teune and Steger, 2010] for generating secondary hairpin structures with embedded mature miRNAs and satisfying other requirements like Minimum free energy. This is similar to our technique in Section 4.7.1. NOVOMIR uses RNAfold to predict the secondary structure of genomic sequence [Teune and Steger, 2010].

Mayer et al. identified 98,068 putative miRNA precursor sequences encoding 270 different mature miRNAs indicating multiple putative precursors for some individual mature microRNAs. Only 1,668 precursor sequences out of 98,068 (1.7%) aligned with wheat expressed sequence tags (ESTs) and reads from all RNA molecules that uses deep-sequencing technologies. Only 52 of the 270 mature miRNA sequences were encoded by precursors originating from non-repeat regions.

Mayer et al. also identified the potential targets of the mature miRNAs using PsRobot v1.2 [Wu et al., 2012] against a set of 133,090 ESTs. Out of 270 identified mature miRNAs, at least one target was identified for 257 miRNAs. In total, 68,641 target protein coding genes were identified for all 270 mature miRNAs.

On the other hand, Sun et al. [2014] sequenced small RNAs from 11 tissues from wheat namely: dry grain, embryo from germinating seed, shoot, seedling root, seedling leaf, culm, 5-mm long inflorescence, 10 to 15 mm long spike, flag leaf, grain collected 8 days after pollination (DAP), and grains collected 15 DAP. They identified miRNAs representing 276 families.

From 118,301,178 reads (18 to 30 nucleotides), Sun et al. obtained a total of 36,235,609 unique sequences for all 11 libraries. Using Bowtie, the unique sequences were mapped against the Rfam database [Burge et al., 2013] and a plant repetitive

sequence database [Ouyang and Buell, 2004] to filter rRNAs, tRNAs and repeats. The tags were mapped to the following wheat reference genome and EST sequences: (i) wheat genome [Brenchley et al., 2012], (ii) illumina sequences from individual chromosome arms generated by International Wheat Genome Sequencing Consortium which helps to identify the chromosomal location of miRNA precursors [Mayer et al., 2014] and (iii) wheat ESTs from the database of genetic resources, from Japan and NCBI. In our experiment, we also filtered ncRNA sequences (rRNAs, tRNAs, etc.) in Section 4.3, performing a BLASTN search with the Rfam-database [Burge et al., 2013], which contains all the ncRNA sequences except miRNAs.

Reads with zero mismatch against the contigs were taken up by Sun et al. for miRNA identification using the miReap algorithm [miReap, 2015]. Using a threshold of a minimum of 20 tags in one library, miRNAs, miRNA star sequences and precursors were predicted using the following criteria. Firstly, the miRNA and its star should be embedded in the opposite strand of the hairpin with two bp overhang. Secondly, there should be only a maximum of four mismatches between the miRNA and miRNA star sequence. Thirdly, the candidate miRNA tag must span at least 70% of all reads mapping to the precursor at miRNA start site with flanking regions of 20 nucleotides on each side. If a set of unique reads mapping to a genomic region has 70% identity, then the remaining 30% of the length of each of the read should originate within 20bp, either upstream or downstream to the primary Dicer clipping nucleotide site.

The predicted candidate miRNAs identified by Sun et al. were compared against the known mature miRNAs in the plant miRNA database (PMRD) [Zhang et al.,

2010], and conserved and novel miRNAs were identified. Using bowtie, all identified mature miRNAs were also mapped against reference contigs from other species such as Arabidopsis, rice, maize, Brachypodium, barley and sorghum for the presence of their precursors, by extracting 200bp each of flanking regions from miRNA matching site.

The presence of 366 perfectly matching known mature miRNAs belonging to 260 miRNA families was identified in all 11 libraries. Highly conserved miRNA families among land plants, namely miR159, miR160, miR167, miR169, miR171, miR172, miR393, miR396 and miR398, were also identified from the datasets.

The targets of miRNAs were identified from ESTs. A total of 524 targets for 124 miRNA families were identified. Compared to the wheat genome paper Mayer et al. where co-ordinates of 98,068 precursors were reported for 270 miRNA families, Sun et al. reported mature miRNA sequences and the corresponding star sequences, which is a very useful resource for discovering genuine miRNAs in a new dataset.

Chapter 4

Solution Methodology

To identify conserved miRNAs from the wheat dataset, we developed a toolchain. We used the python programming language in the Ubuntu Linux system to implement the toolchain. During the first step of the toolchain, we removed all non-miRNAs from the experimental sequences (Section 4.3) using the Rfam database. Next, all conserved miRNAs were detected by comparing the experimental sequences to the miRBase database. The detailed process of identifying conserved miRNAs is described in Section 4.5.

To identify conserved miRNAs, besides the miRBase database, we also used the supplementary materials [Science, 2014] of Mayer et al. [2014], which contain the precursors from the wheat genome. We aligned them with our experimental samples and only considered those sequences that have an exact match between the experimental samples and the precursors. The detailed process and rules to predict the most suitable precursors and star sequences are described in Section 4.7.

We also matched our experimental sequences with the conserved miRNA and

miRNA star sequences provided by the supplementary materials [BMC, 2014] of Sun et al. [2014]. The supplementary materials also contain the wheat precursors along with the dot-bracket notation of the precursors. The detailed process of matching our experimental sequences with Sun et al.'s mature miRNA and miRNA star sequences are described in Section 4.8.

Figure 4.1 shows a detailed overview of the methodology of my thesis. Each of the steps of Figure 4.1 is described below.

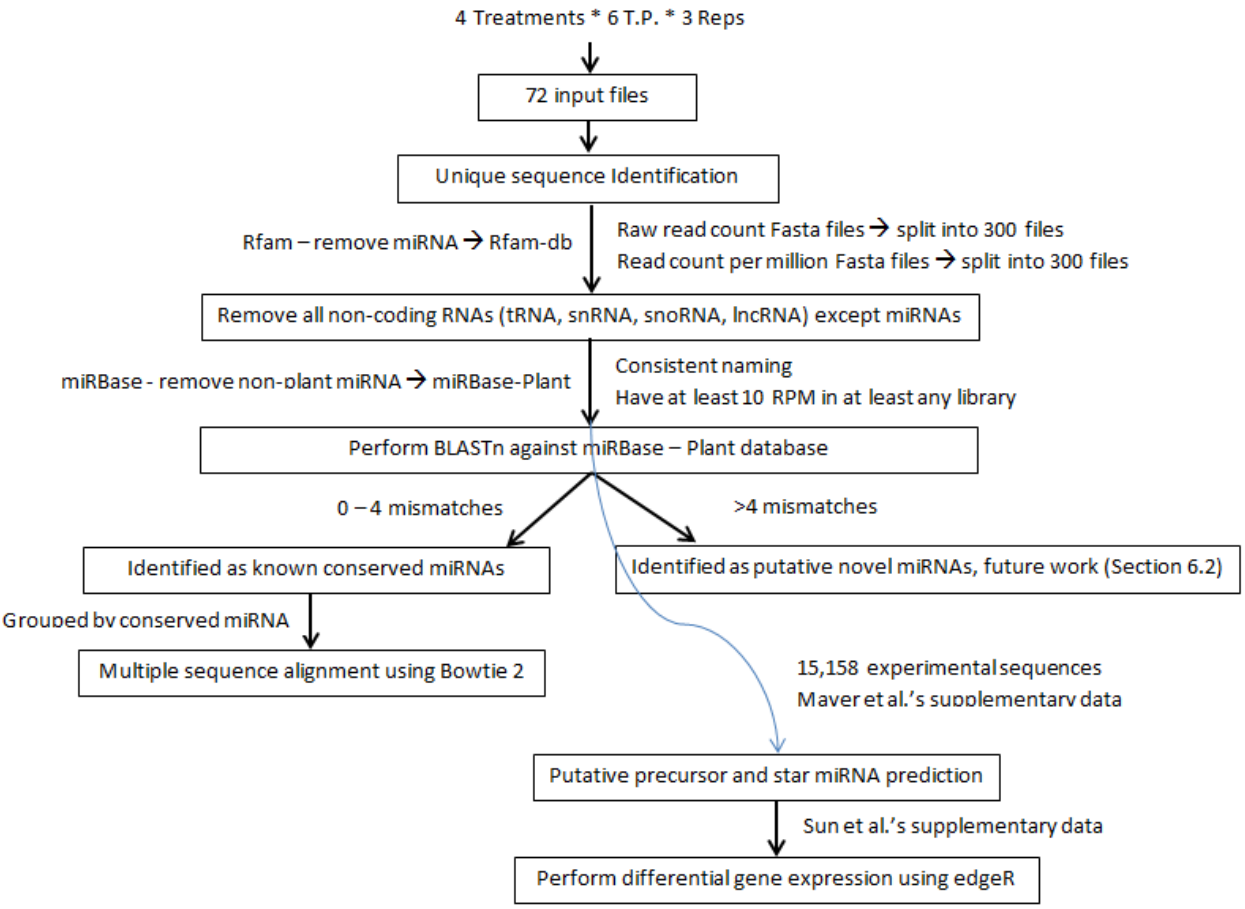

Figure 4.1: Detailed flow chart of solution methodology.

4.1 Input files: description and organization

We used a wheat (*Triticum aestivum*) small RNA dataset, consisting of sequences generated from leaf samples collected from wheat plants subjected to heat, light or UV stresses or no stress (control). Leaf tissues were collected at six time points after the end of the constant stress period: day 0, 1, 2, 3, 7 and 10. Ninety-six *Triticum aestivum cv* Glenlea plants were grown in a growth cabinet (Conviron Technologies, Winnipeg, Canada) under long day conditions (16 hours light at 18 °C and 8 hours darkness at 16 °C). Three batches of 24 plants each represents 3 replicates of eight plants were exposed to three different stresses under controlled growth conditions: continuous light for three days; heat stress (at 37 °C for 72 hours) and UV stress (2 minutes of exposure to UV light for 3 consecutive days). The final set of twenty-four plants were grown as control conditions (16 hours light at 18 °C and 8 hours darkness at 16 °C).

A total of 72 fasta files corresponding to the 72 small RNA libraries and containing a total of $\sim$ 523 million reads constituted the input files. Fasta is a text-based format for representing nucleotide sequences. The sequences in this format begins with single-line description, followed by sequence data. The description or header line starts with the "$>$" symbol to distinguish between the description line and the sequence. Each fasta file was 250 to 450 MB in size. Input data was organized into several stress conditions, replicates and sampling time points (T.P.) after stress. Each T.P. contained 3 replicates and each replicate contained 4 fasta files (1 fasta file for each conditions: control, heat, light and UV). Thus, 72 fasta files were organized into 6 T.P. * 3 replicates * 4 conditions.

4.2 Unique sequence identification

Initially, the same sequence could be present several times in a sample file as the same RNA can be expressed as several copies in the same replicate/condition/time point. So, we first identified all unique sequences and their total number of occurrences in a particular sample file. Then we also generated a summary file which contains the total number of reads for each unique sequence and the normalized read counts or reads per million (RPM) using the formula: (1000000 * read count) / total number of sequences. Next, we split each unique fasta file into 300 files (total 72*300 = 21600 files) because single fasta files were too large.

4.3 Removal of ncRNA sequences

After identifying the unique sequences and splitting the input files, our next goal was to remove all contaminating ncRNA sequences from the input dataset other than miRNAs, such as rRNA, tRNA, small nuclear RNA (snRNA), small nucleolar RNA (snoRNA) and long non-coding RNA (lncRNA). We removed those ncRNA sequences as our goal is to find miRNA sequences only. For this, at first we removed all the miRNAs from the Rfam database by discarding those sequences that has "MIR" in the header of the Rfam fasta file. Then, we performed a BLASTN search, for each sequence in the 300 unique, split files, against the Rfam-database [Burge et al., 2013], which contains all the ncRNA sequences except miRNAs. This step is similar to Yao et al. [2007] and Sun et al. [2014], described in Chapter 3. After removing all the contaminating ncRNA sequences, we then aggregated the updated

300 unique split files that contain no contaminating ncRNA sequences. As we needed to process all 72 input files, we used the grid-based infrastructure of Westgrid to do the high performance computing. We used the Hermes server of Westgrid that contains 2112 cores [Westgrid, 2015]. Using Hermes, we executed the entire process as a series of programs in the Unix bash shell scripting language that allows execution of commands.

4.4 Filtering the data

4.4.1 Consistent naming

The same sequence can be present in different different replicates, conditions or time points because the same RNA can be expressed in all these samples. So, in next step, we consistently named the sequences so that every experimental sequence has the same identity across the whole experimental dataset. All sequences were named with the format: species_[no]_sample_[no]_of72_[length]. For example, the very first sequence was named as species1_sample1of72_21bp.

4.4.2 Consistent named sequences having 10 RPM in any experimental sample

We considered only those sequences having at least ten RPM in any of the experimental samples [Montes et al., 2014] and these sequences were used as the input sequence in Section 4.5 and Section 4.7. There were a total 15,158 sequences having at least ten RPM in at least one of the experimental samples. Sequences having less

than 10 RPM were used for predicting star sequences, which is described in Section 4.7.2.

4.5 Conserved miRNAs identification

After removing all other ncRNAs except miRNAs, we needed to identify and characterize the known or conserved miRNAs from wheat or other plant species. As miRBase contains all known conserved miRNAs, in this step we compared the filtered data from Section 4.4.2 with the miRBase database to identify the known conserved miRNAs and their precursors. At first, from the miRBase database, we discarded all the miRNAs except those from plant species to create miRBase-Plant. Then, similar to Yao et al. [2007], Yin and Shen [2010], and Mayer et al. [2014], described in Chapter 3, we also performed a BLASTN search using the miRBase database, release 20 (June, 2013). We tuned the parameters and chose e-value 10, word size 11, match and mismatch score 4 and -5 respectively to obtain the HSP with the best score between experimental and miRBase sequence. For conserved miRNA identification, we considered only those matches where we found 0 to 4 mismatches between the experimental sequence and a sequence from miRBase [Ragupathy, personal communication]. The remaining sequences (with greater than 4 mismatches between the experimental and the miRBase sequence) were considered as the input for the identification of novel miRNAs in Section 4.6. Figure 4.2 shows acceptable alignments between experimental and miRBase sequences as a partial output of BLAST, to identify known conserved miRNAs.

Moreover, we grouped together the experimental sequences that matched the

```
a)  HSP:
    Expect value 8.1e-04, alignment length 21
    Experimental Sequence:       TTTGGATTGAAGGGAGCTCTG
                                 |||||||||||||||||||||
    miRBase sequence:            TTTGGATTGAAGGGAGCTCTG
    Mismatches_in_HSP: 0
------------------------------------------------------------
b)  HSP:
    Expect value 1.2e-02, alignment length 18
    Experimental Sequence:       TTGGATTGAAGGGAGCTC
                                 ||||||||||||||||||
    miRBase sequence:           CTTGGATTGAAGGGAGCTCCC
    Mismatches_in_HSP: 0
------------------------------------------------------------
c)  HSP:
    Expect value 6.2e-03, alignment length 21
    Experimental Sequence:     AGGGTCGAACTGAGAACACATGAG
                                ||| ||||||||||||||||
    miRBase sequence:           GGGGCGAACTGAGAACACATG
    Mismatches_in_HSP: 1
------------------------------------------------------------
d)  HSP:
    Expect value 2.5e-01, alignment length 19
    Experimental Sequence:       GGGTCGAACTGGGAACACA
                                 ||| || ||||||||||||
    miRBase sequence:            GGGGCGGACTGGGAACACATG
    Mismatches_in_HSP: 2
------------------------------------------------------------
e)  HSP:
    Expect value 2.2e+00, alignment length 21
    Experimental Sequence:       TTCGTATTTCAGGGAGCTCTT
                                 || | |||  ||||||||||||
    miRBase sequence:            TTTGGATTGAAGGGAGCTCTT
    Mismatches_in_HSP: 4
```

Figure 4.2: Known conserved miRNA identification, sample BLAST output using 0 to 4 mismatches between the miRBase and experimental sequence: a) exact match , b) exact match where the miRBase sequence is longer than the experimental sequence, c) 1 mismatch where the experimental sequence is longer than the miRBase sequence, d) 2 mismatches where the miRBase sequence is longer than the experimental sequence and e) 4 mismatches between the miRBase sequence and the experimental sequence of the same length.

same miRNA from miRBase with 0 to 4 mismatches. Figure 4.3 shows how experimental sequences were grouped together that matched with the same miRNA from miRBase. In the figure, for each row, the first column represents the miRNA

from miRBase and the remaining columns represent the experimental sequences that matched with that miRNA. For example, miRNA tae-miR1135 matched with the experimental sequence species3971_sample68of72_21bp only whereas bdi-miR159-3p miRNA matched with the experimental sequence: species21_sample68of72_21bp, species34_sample68of72_20bp, species932_sample68of72_21bp, etc.

tae-miR1135	species3971_sample68of72_21bp			
sbi-miR1432	species7320_sample68of72_21bp			
osa-miR1432-5p	species2723_sample68of72_21bp	species12056_sample68of72_20bp		
ama-miR156	species427_sample68of72_20bp	species3541_sample68of72_21bp		
bdi-miR156h-3p	species981_sample68of72_22bp	species3855_sample68of72_23bp		
bdi-miR156i-3p	species345_sample68of72_21bp			
smo-miR159	species213_sample68of72_19bp	species1401_sample68of72_18bp	species3860_sample68of72_20bp	
bdi-miR159-3p	species21_sample68of72_21bp	species34_sample68of72_20bp	species932_sample68of72_21bp	....
htu-miR159a	species235_sample68of72_21bp	species3339_sample68of72_22bp		
lus-miR159b	species35_sample68of72_20bp	species96_sample68of72_19bp	species149_sample68of72_20bp	....
tae-miR159b	species1_sample68of72_21bp	species36_sample68of72_22bp	species53_sample68of72_20bp	....
aly-miR159b-3p	species189_sample68of72_21bp	species536_sample68of72_22bp	species961_sample68of72_23bp	....
bdi-miR159b-3p.3	species866_sample68of72_21bp			
bdi-miR159b-5p.1	species204_sample68of72_20bp	species350_sample68of72_21bp	species2506_sample68of72_22bp	
bdi-miR159b-5p.3	species6193_sample68of72_21bp			
zma-miR159h-3p	species565_sample68of72_21bp			
tae-miR160	species130_sample68of72_21bp	species1378_sample68of72_20bp		
bdi-miR160e-5p	species678_sample68of72_21bp			
ppt-miR160i	species8016_sample68of72_21bp			
bdi-miR166b-3p	species68_sample68of72_21bp	species117_sample68of72_21bp	species4634_sample68of72_21bp	....
bdi-miR166b-5p	species136_sample68of72_21bp	species1708_sample68of72_20bp		
bdi-miR166e-3p	species8351_sample68of72_21bp			
....				

Figure 4.3: Experimental sequences matched with the miRBase conserved miRNAs.

We also produced a fasta file that contains all the miRNAs from miRBase and the matched experimental sequences. In the fasta file, miRNA "tae-miR1135" begins with the ">>" sign and the matched experimental sequences begin with the ">" sign. Figure 4.4 shows an example of the fasta file.

Next, for each grouped experimental sequences (the experimental sequences that

```
>>tae-miR1135 MIMAT0005370 Triticum aestivum miR1135
CUGCGACAAGUAAUUCCGAACGGA
>species3971_sample68of72_21bp
cgacaagtaattccgaacgga
            .....
>>osa-miR1432-5p MIMAT0005966 Oryza sativa miR1432-5p
AUCAGGAGAGAUGACACCGAC
>species2723_sample68of72_21bp
atcaggagagatgacaccgac
>species12056_sample68of72_20bp
atcaggagagatgacaccga
            .....
>>bdi-miR159-3p MIMAT0020692 Brachypodium distachyon miR159-3p
CUUGGAUUGAAGGGAGCUCU
>species21_sample68of72_21bp
cttggattgaagggagctctg
>species34_sample68of72_20bp
cttggattgaagggagctct
>species932_sample68of72_21bp
cttggattgaagggagctctc
>species3100_sample68of72_21bp
cttggattgaagggagctctt
>species3249_sample68of72_22bp
cttggattgaagggagctctgt
            .....
```

Figure 4.4: Fasta file of conserved miRNAs matched with experimental sequences.

matched with the same miRNA from miRBase), we used Bowtie 2 against LCG assembly to identify putative regions of the wheat genome that matched with the grouped experimental sequences. This helped us aligning multiple sequences in the next step. Figure 2.10 shows a sample of this step.

We then extracted the precursors (pre-miRNAs) of the conserved miRNAs. For this we used the fasta file hairpin.fa from miRBase website [miRBase, 2015], which contains a list of all precursor sequences for miRBase sequences. We generated all the precursor sequences of the conserved miRNAs with the help of the hairpin.fa file. For example, if the conserved miRNA name in the miRBase miRNA database is "tae-miR159b Triticum aestivum miR159b", we also identified the particular pre-

cursor from hairpin.fa file searching with same name such as "tae-MIR159b Triticum aestivum miR159b stem-loop" and used that particular precursor sequence.

We now have the grouped experimental sequences that matched with conserved miRNAs, along with conserved miRNA sequence from miRBase and related precursors (from wheat or other plants) from miRBase, and the wheat genome regions matching the grouped experimental sequences. We then performed an MSA using the default MAFFT values, with the grouped experimental sequences, matched conserved miRNA sequences from miRBase and related precursors from miRBase, and the matched wheat genome regions from the wheat LCG assembly. We did not use the wheat genome assembly by Mayer et al. as this experiment had been done before Mayer et al.'s genome assembly was published.

With these MSAs, we were able to determine the sequence similarity between them. Higher sequence similarity among the grouped experimental sequences, the matched conserved miRNA, the related precursor and the genome sequence represents the identification of conserved miRNAs with high accuracy. Figure 2.11 shows a sample output of this step. It shows the similarity among the grouped experimental sequences (miRNA1, miRNA2, and miRNA3) from the experimental dataset, conserved miRNA and precursor from miRBase and wheat genome.

4.6 Identification of novel miRNAs

Mature miRNAs having significant similarity to entries in miRBase are called conserved miRNAs and undiscovered miRNAs are called novel miRNAs. For conserved miRNA identification, we considered only those matches where we found 0

to 4 mismatches between the experimental sequence and a sequence from miRBase. The remaining sequences (with greater than 4 mismatches between the experimental and the miRBase sequence) became the input for the identification of novel miRNAs. The detailed process of identifying conserved miRNAs is discussed in Section 4.5.

For identifying the novel miRNAs, we first removed all sequences that matched with conserved miRNAs from miRBase in Section 4.5 from the filtered experimental sequences obtained from Section 4.4.2. So the resultant sequences contained more than 4 mismatches between the input sequences and any sequence in the miRBase database, and these resultant sequences were used as input sequences for novel identification in this step. To claim an input sequence as a novel miRNA, the sequence must have a hairpin shaped precursor with <-0.2 Kilocalorie/mole/nt folding free energy [Kozomara and Griffiths-Jones, 2014], and have a star sequence in the sample dataset. Free energy provides a measure of thermodynamic stability for possible secondary structures that a molecule or molecules could form [Zuker et al., 1999].

As we did not finish the whole procedure of identification of novel miRNAs, the rest of the tasks are described as a future work in Section 6.2.

4.7 Identification of conserved miRNAs using supplementary materials of Mayer et al.

Mayer et al. identified 98,068 putative miRNA precursor sequences encoding 270 different mature miRNAs indicating multiple putative precursors for some individual mature microRNAs. We predicted putative miRNAs by matching these 98,068

putative miRNA precursor sequences with our experimental sequences.

In our experiment, we used miRBase database that contains the known conserved miRNAs and precursors of all plants including wheat. As we experimented with wheat sequences, to identify the conserved miRNAs for our experimental samples, besides miRBase, we also used the supplementary materials [Science, 2014] provided by Mayer et al. [2014], which contains predicted precursor sequences only for wheat. The procedure was divided into two steps: candidate precursor prediction and prediction of star sequence.

4.7.1 Candidate precursor prediction

The supplementary materials [Science, 2014] provided by Mayer et al. [2014] contained a list of predicted precursors' ID of wheat, putative miRNAs, chromosome arm, contig ID, start and end location of the precursors in the contigs. Mayer et al. [2014] also provided all contig sequences representing each chromosome arm. In this step, we used the 15,158 sequences (at least 10 RPM in at least one library, generated in Section 4.4.2) as input. This contains sequences with at least 10 RPM in any experimental sample. Figure 4.5 shows an example of the data file and the contig sequences. The fourth column of the Figure 4.5.a contains the contig IDs and the right side of the Figure contains the sequences of those contigs. The position of the precursors in the contigs are also given by the start and end location (fifth and sixth columns of the Figure).

We generated a fasta database of all precursors. As the files containing contig sequences were named based on the chromosome arm, using the chromosome arm

and contig ID, we obtained the particular contig sequence. Then, we extracted the precursor sequence from the contig sequence using the start and end locations.

After that, we aligned the input sequences (15,158 sequences having at least 10 RPM in at least one library, generated in Section 4.4.2) with this precursor database using Bowtie 2 and retained only the exact matches. We discarded the input sequences that did not exactly match with the precursor database. Thus, we obtained the experimental samples that exactly matched with the wheat precursors and putative miRNAs (from the second column of Figure 4.5.a) provided by Mayer et al.

Kurtoglu et al. [2014] provided a list of 52 genuine microRNAs identified in wheat with high quality confidence annotation of precursors. We compared our predicted putative miRNAs with these 52 miRNAs and discarded any putative miRNA that were not part of the 52 miRNAs identified in wheat.

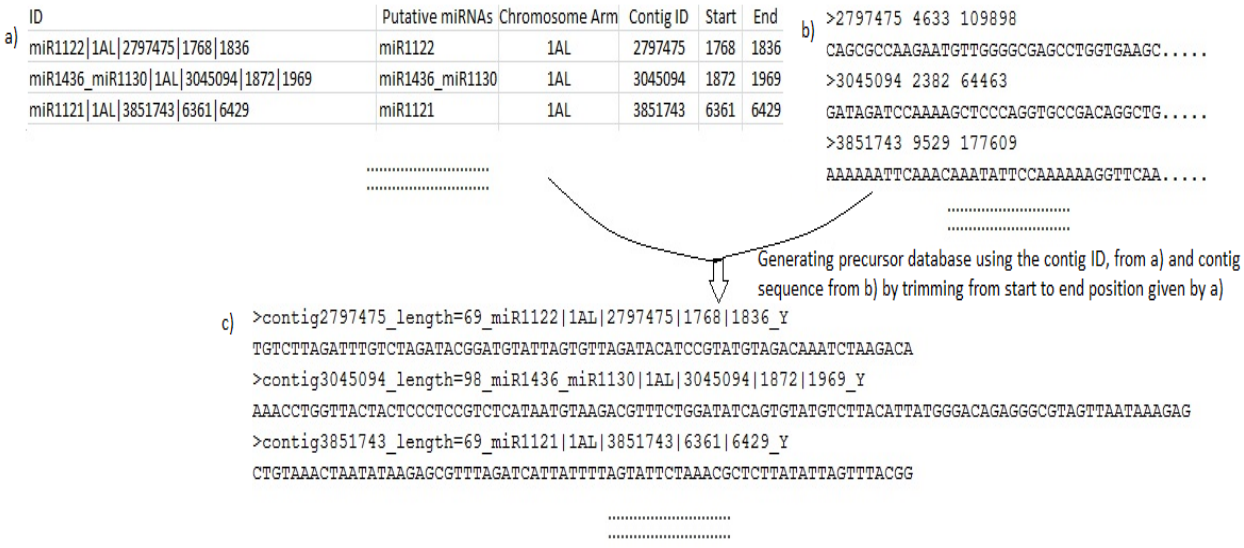

Figure 4.5: Precursor database (c) creation using two supplementary files by Mayer et al. [2014]: a) precursor table, b)contig sequences.

We thus generated the experimental samples, related precursors and putative miR-

NAs. Next, we predicted the star sequence and hairpin shaped structure. For predicting the star sequences, we required the dot-bracket notations of the precursors. We predicted the dot-bracket notations of the precursor sequences using RNAfold [Lorenz et al., 2011]. Besides the dot-bracket notations, RNAfold also provided the hairpin shaped structures and minimum free energy. We considered only those sequences which have <-0.2 Kilocalorie/mole/nt free energy.

Figure 2.4 shows an example of hairpin shaped structures and Figure 4.6 shows one of the dot-bracket notation outputs produced by RNAfold that represents: precursor sequence, MFE and dot-bracket notation. The process of star sequence prediction using dot-bracket notation is described in the next section.

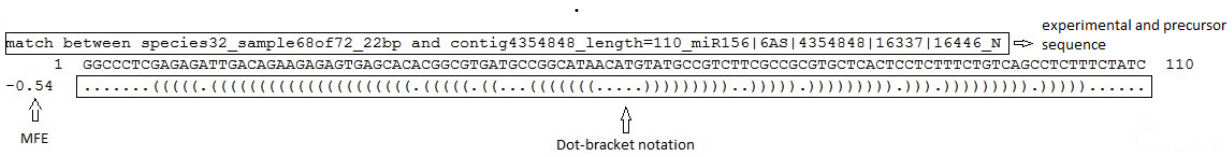

Figure 4.6: Example of dot-bracket notation of precursor sequence by RNAfold.

4.7.2 Prediction of star sequence

The two strands of mature miRNA are the 3-prime mature miRNA and 5-prime mature miRNA. Between the 3-prime mature miRNA and 5-prime mature miRNA, one of them is called the mature or guided strand and the other is known as the miRNA star strand. For identifying a conserved or novel miRNA, an miRNA sequence must have miRNA-star support, that is, a star sequence should be observed in the experimental data. Predicting the miRNA star sequence is a challenging task as there is no prior knowledge of the location of the star sequence. Moreover, there may be

some length variability between miRNA and miRNA star sequences.

In this step, we also used the filtered data (with at least 10 RPM in any experimental sample) generated in Section 4.4.2 as the input sequences. From Section 4.7.1, we obtained the related precursor sequence or contig sequences and dot-bracket notation for the input sequences. We assumed the star sequence has the same length as the sequence and that there is a two bp overhang [Kozomara and Griffiths-Jones, 2014] between the input and star sequences as described in Figure 2.4. The rules of predicting star sequences is slightly different depending on whether the input sequence is matched with the precursor sequence either on 5' or 3' side. We describe the two sets of rules in the following three subsections.

Star sequence prediction when input and precursor sequence matched on 5' side

Figure 4.7 shows the process of predicting a star sequence at the 5' end. First, we aligned the experimental sequence with the contig sequence of precursor predicted by Mayer et al.. The alignment is shown between positions X and Y in the Figure. This alignment is guaranteed to have zero mismatches because all sequences with mismatches were filtered in Section 4.7.1. Next, we found the matching brackets of X and Y on the 3' side: these are positions x and y. We shifted x and y by two positions to the right. This addition is necessary to create the two bp overhang between the input and star sequences. These shifted positions are w and z in the figure. The region from w to z is the potential miRNA-star sequence we were looking for.

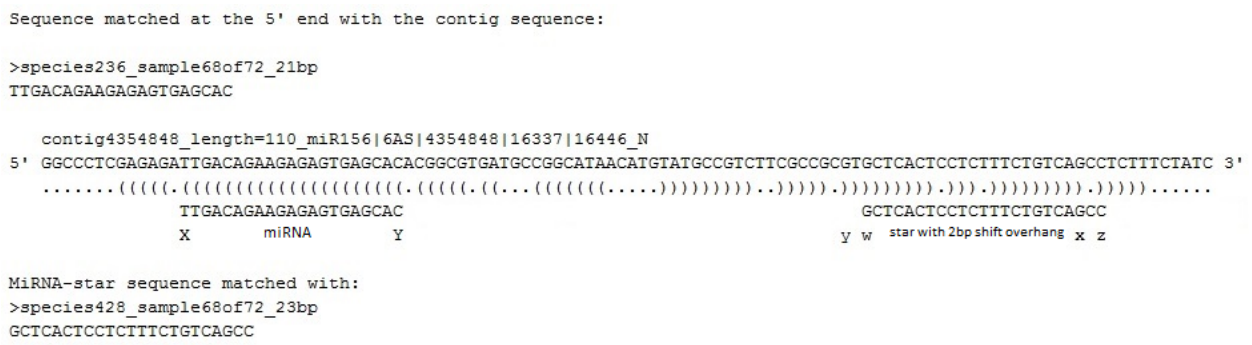

Figure 4.7: MiRNA-star prediction: 5' end.

Star sequence prediction when input and precursor sequence matched on 3' side

Figure 4.8 shows the process of predicting a star sequence at the 3' end. At first, we aligned the species sequence with the contig sequence. The alignment is shown between positions X and Y in the Figure. Next, we shifted X and Y by two positions to the left for creating the two bp overhang between input and star sequence. These shifted positions are x and y in the figure. We found the matching brackets of x and y on the 5' side. These positions are w and z in the figure. Thus, the region from w to z is the potential miRNA-star sequence we were looking for.

Moreover, we also found some cases where the aligned dot-bracket notation started and/or ended with a dot (.) while aligning input sequence either on 5' side or 3' side (after shifting 2 positions right) with the contig sequence. For this, we traversed right until we get a bracket notation: "(" on 5' side or ")" on 3' side. Similarly, for predicting star sequence we traversed same distance in the same direction after matching bracket ")" on 3' side or "(" on 5' side. Figure 4.9 shows an example of predicting star sequence where input sequence aligned with the contig sequence on

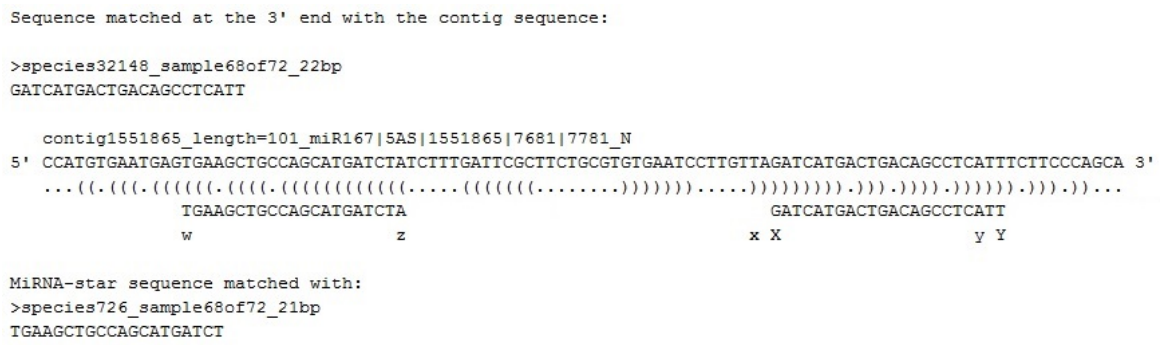

Figure 4.8: MiRNA-star prediction: 3' end.

the 5' side and dot-bracket notation started with a dot (.). The alignment is shown between positions X and Y in the Figure where position X is a dot (.). So we traversed right until we get a bracket notation: "(", which is position X'. Next, we found the matching brackets of X' and Y on the 3' side: these are positions x and y. Now, we also needed to shift position x right by 1, which gives position x'. Then we shifted x' and y by two positions to the right. This addition is necessary to create the two bp overhang between the input and star sequences. These shifted positions are w and z in the figure. The region from w to z is the potential miRNA-star sequence we were looking for.

Exceptional cases

Besides the cases stated above, there are also some other cases where we discarded the input and star sequences. We discarded any input or star sequence if there are at least 4 unpaired base "." in the sequence except in the 2 bp overhanging positions.

Again, to create the two bp overhang between input and star sequence, we shifted the aligned 3' star sequence by two positions to the right (Figure 4.7), but we could

```
Sequence matched at the 5' end with the contig sequence where dot-bracket notation starts with a dot (.):

>species1881_sample68of72_21bp
CAGCCAAGGATGACTTGCCGG

   contig4354848_length=110_miR156|6AS|4354848|16337|16446_N
5' GCAGAGAACGGGATGCAGCCAAGGATGACTTGCCGGCTCCTGGTGCTGGGAGTTCGTTGAGCTTTAGAGTTAGCCGGCAAGTTGTCTTTGGCTACACTTAGTTCTCTTC 3'
   ..((((((((((.((.(((((((((((((((((((((((.((((.(((.((......)).))))....)))))))))))))))))))))).)))))).)))))))..
                 CAGCCAAGGATGACTTGCCGG                                    GGCAAGTTGTCTTTGGCTACA
                 XX'                 Y                                  y w               xx'z

MiRNA-star sequence matched with:
>species17356_sample68of72_21bp
GGCAAGTTGTCTTTGGCTACA
```

Figure 4.9: Sequence matched at the 5' end with the contig sequence where dot-bracket notation starts with a dot (.).

not shift right by two positions if there is no nucleotide remaining on the right side, as in position z in the Figure 4.10.

```
MiRNA-star matched at the very end of 3' end with the contig sequence:

   contig4354848_length=110_miR156|6AS|4354848|16337|16446_N
5' GGCCCTCGAGAGATTGACAGAAGAGAGTGAGCACACGGCGTGATGCCGGCATAACATGTATGCCGTCTTCGCCGCGTGCTCACTCCTCTTTCTGTCAGCCTCTTTCTATC 3'
   .......(((((.(((((((((((((((((((.(((((.((...(((((((.....)))))))))..)))))..)))))))))).)))).)))))).....
                                                                                  CTTTCTGTCAGCCTCTTTCTATC
                                                                                  y w                  x z
```

Figure 4.10: Discarded input sequence where we could not create two bp overhang.

Also, we found the input and star sequences overlap for some cases but physically its not possible to form an hairpin structure if the overlap occurs. So, we also discarded those input and star sequences where there were overlaps between input and star sequences or if the input or the star sequence were a part of the hairpin loop. Figure 4.11 shows an example of these discarded cases where the input and star sequence overlapped, and also the input sequence is part of the hairpin loop. Note that the middle unpaired region (between the last open parenthesis "(" and the first close parenthesis ")") of the dot-bracket notation represents the hairpin loop structure.

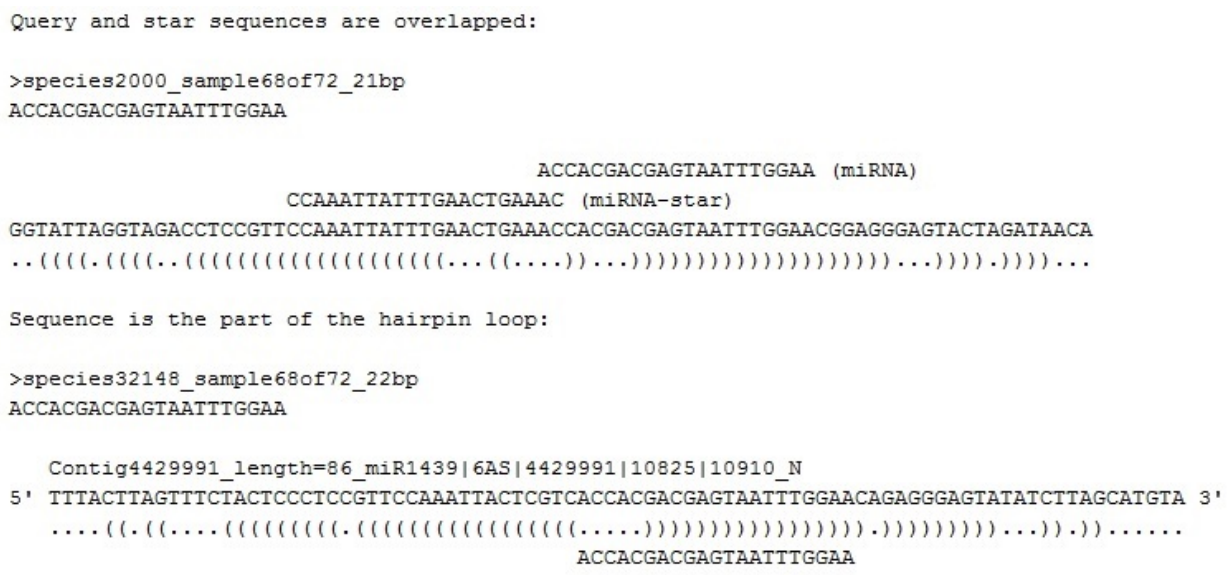

Figure 4.11: Discarded input and star sequences where input and star sequences overlapped (top) and where input sequence is a part of the hairpin loop (bottom).

4.8 Identification of conserved miRNAs using supplementary materials of Sun et al.

We also used the supplementary materials [BMC, 2014] of Sun et al. [2014]. These supplementary materials contains conserved miRNA, their star sequences, along with wheat precursors and the dot-bracket notation of the precursors. As conserved miRNAs with star sequence, along with precursors and dot-bracket notation of the precursors are already given, we neither needed to predict the candidate precursor nor the star sequence. Using sequence comparison, we exactly matched the conserved miRNA sequences given by Sun et al. with our filtered data (with at least 10 RPM in any experimental sample) generated in Section 4.4.2. To match the star miRNA sequences given by Sun et al., we used our consistently named data generated in Section 4.4.1, which is not restricted to at least 10 RPM in any experimental sample

(see section 4.4.1 for details). We discarded those sequences having no match with either the conserved or star miRNA sequences of our experimental sequences.

4.9 Differential gene expression analysis

Differential gene expression helps us determine what changes occur in the wheat genome upon exposure to abiotic stresses. Abiotic stresses can cause miRNAs to be up-regulated (increased in expression) or down-regulated (decreased in expression).

We analyzed differential gene expression for identifying conserved miRNAs expressed under different abiotic stresses. For this, we used edgeR [Robinson et al., 2010], a bio-conductor package of R language with the count values calculated in Section 4.2 and with the experimental sequences of Section 4.7 and 4.8 which we identified as potential conserved miRNAs.

In total, we identified 232 experimental sequences matched with either Mayer et al. or Sun et al. dataset. Among them, we found 205 and 12 experimental sequences, as an outcome of comparing our experimental sequences (filtered data with at least 10 RPM in any experimental sample generated in Section 4.4.2), with Mayer et al. dataset in Section 4.7 and with Sun et al. dataset in Section 4.8 respectively. Fifteen experimental sequences were common in both the Mayer et al. and Sun et al. analyses. In total, there were 325 experimental sequences as some experimental sequences matched with multiple conserved miRNA families. For example, experimental sequence species24 matched with both conserved miRNA family miR2020 and miR5064.

Input files were divided into control versus 3 stresses: heat, light and UV. So for

each T.P. there were 3 files: control versus (vs) heat, control vs light and control vs UV. We had 6 T.P.: day 0, 1, 2, 3, 7 and 10. Thus we had total of 18 files as input to edgeR (6 days * 3 control vs stress files). Each of these 18 files contained 6 columns where the first 3 columns contained the count data of the 3 replicates of control and other 3 columns contained the count data of the 3 replicates for stress (either heat, light, or UV).

For each of the control vs stress file, we categorized 325 experimental sequences into 36 unique conserved miRNA families by grouping experimental sequences that matched with same conserved miRNA family and summing up the count data (Figure 4.12). Thus, each of the 18 input files contains 36 rows of conserved miRNA family where each row contained count data of control and stress treatment. As each treatment contained count data of 3 replicates, each row contained 6 columns of count data that divided into 3 columns of count data for control and 3 columns of count data for stress (either heat, light, or UV).

Figure 4.12 shows an example of day 0, control vs heat input file processing for edgeR. Figure 4.12.a is an example where we had 325 rows where each row represented unique experimental sequence that matched with a conserved miRNA family with either Mayer et al. or Sun et al. Column B to G contained the count values calculated in Section 4.2 where the first 3 columns and the last 3 columns contained the count data for control and heat treatment respectively. Then, in the Figure 4.12.b, the same conserved miRNA families were grouped together. For example, miR156(Sum_8) (column A, row 21 in the Figure) shows that there were 8 experimental sequences that matched with the conserved miRNA family 156 (species 32, 236, 345, 427, 428,

1322, 3541, and 20602 from Figure 4.12.a). Also, the count values were summed up. For example, 14666 (column B, row 21 in the Figure) is the summation of all the count values of those 8 experimental sequences for day0, Replication 1, control treatment (column B, rows 2 to 9). We used the grouped miRNA families file such as 4.12.b as input to edgeR. The detailed results of differential gene expression analysis are discussed in Section 5.3.

	A	B	C	D	E	F	G
1	a) **family_species_number**	**1_0_DAT_Rep_1_Control**	**5_0_DAT_Rep_2_Control**	**9_0_DAT_Rep_3_Control**	**2_0_DAT_Rep_1_Heat**	**6_0_DAT_Rep_2_Heat**	**10_0_DAT_Rep_3_Heat**
2	miR156_species32_sample68of72_22bp	11124	9568	4439	150	362	307
3	miR156_species236_sample68of72_21bp	1169	552	728	52	104	208
4	miR156_species345_sample68of72_21bp	809	583	444	67	76	76
5	miR156_species427_sample68of72_20bp	635	312	509	23	42	78
6	miR156_species428_sample68of72_23bp	634	540	299	8	16	14
7	miR156_species1322_sample68of72_23bp	209	162	81	31	28	41
8	miR156_species3541_sample68of72_21bp	73	57	52	8	8	7
9	miR156_species20602_sample68of72_21bp	13	13	31	24	1	6
10							
11	miR159_species1003_sample68of72_20bp	267	264	151	90	96	143
12	miR159_species2000_sample68of72_21bp	133	48	38	16	11	24
13	miR159_species34_sample68of72_20bp	10430	4826	3425	1932	1924	2212
14	miR159_species53_sample68of72_20bp	6061	4376	1950	1059	1407	1254
15							
16	Tae-miR394b_species2679_sample68of72_20bp	97	98	128	57	71	120
17	.....	.....	.....	.....	.....	.....	.....
18							
19							
20	b) **majorFamily (Sum_No. of species matched)**	**1_0_DAT_Rep_1_Control**	**5_0_DAT_Rep_2_Control**	**9_0_DAT_Rep_3_Control**	**2_0_DAT_Rep_1_Heat**	**6_0_DAT_Rep_2_Heat**	**10_0_DAT_Rep_3_Heat**
21	miR156 (Sum_8)	14666	11787	6583	363	637	737
22	miR159 (Sum_4)	16891	9514	5564	3097	3438	3633
23	Tae-miR394b (Sum_1)	97	98	128	57	71	120
24	.....	.....	.....	.....	.....	.....	.....

Figure 4.12: Processing of input file for determining differential expression by edgeR.

Chapter 5

Results

In this chapter, we will describe the results and findings of my thesis. The sections are divided into: i) results of conserved miRNA identification using the miRBase database, ii) conserved miRNAs identification using the supplementary materials of Mayer et al. and Sun et al., and iii) differential gene expression.

5.1 Conserved miRNA identification using miRBase database

From Section 4.1 to Section 4.3, we described how we identified unique sequences, removed ncRNA sequences and filtered the data. Figure 5.1 shows the read counts after unique sequence identification, removal of ncRNAs and consistent naming. In the figure, we see that, the numbers are gradually decreasing in each step. After all steps, there were a total 15,158 sequences having at least ten RPM in at least one of the experimental samples.

		Control				Heat				Light				UV			
	Replication	Processed reads (18-24bp)	Unique Sequence	After ncRNA removal	After consistent naming	Processed reads (18-24bp)	Unique Sequence	After ncRNA removal	After consistent naming	Processed reads (18-24bp)	Unique Sequence	After ncRNA removal	After consistent naming	Processed reads (18-24bp)	Unique Sequence	After ncRNA removal	After consistent naming
0 DPT	R1	7,954,051	1862962	1813309	1813309	4,544,251	1,036,511	986,968	986,968	7,367,439	2,089,019	2,042,192	2,042.192	7,498,533	1,882,176	1,823,459	1,823,459
	R2	6,834,706	1721867	1680564	1681482	5,783,423	1,112,982	1,049,420	1,050,012	6,807,080	1,674,862	1,626,553	1,627.480	7,713,193	1,801,813	1,733,162	1,734,039
	R3	5,125,089	1552617	1513958	1513958	6,064,609	1,254,552	1,191,423	1,191,423	7,324,748	1,950,603	1,901,258	1,901.258	7,401,219	1,681,419	1,624,811	1,624,811
1 DPT	R1	4,062,658	1118928	1077126	1077126	5,084,398	1,043,633	995,513	995,513	5,452,030	1,443,266	1,403,229	1,403.229	4,978,405	1,119,864	1,082,387	1,082,387
	R2	7,822,886	1790917	1734260	1734260	6,090,675	1,251,622	1,199,259	1,199,259	7,351,130	1,726,653	1,669,238	1,669.238	7,152,749	1,861,112	1,807,534	1,807,534
	R3	6,380,378	1801870	1748948	1748948	4,613,125	995,059	954,066	954,066	3,866,319	953,658	920,791	920,791	5,723,798	1,576,116	1,529,744	1,529,744
2 DPT	R1	8,572,489	2309743	2259436	2259436	5,872,789	1,463,322	1,409,302	1,409,302	7,522,036	1,866,087	1,813,730	1,813.730	7,165,963	1,821,007	1,764,080	1,764,080
	R2	7,996,889	1998530	1939588	1939588	4,988,186	1,268,978	1,218,711	1,218,711	6,671,291	1,673,909	1,627,646	1,627.646	5,113,993	1,214,715	1,176,640	1,176,640
	R3	6,599,840	1705219	1656449	1656449	4,764,107	1,271,939	1,229,440	1,229,440	6,773,196	1,681,601	1,637,318	1,637.318	6,656,062	1,684,686	1,633,614	1,633,614
3 DPT	R1	7,967,083	1931237	1868280	1869289	6,069,544	1,536,225	1,477,738	1,478,649	6,383,381	1,674,521	1,634,684	1,635.569	6,922,110	1,832,705	1,780,038	1,781,024
	R2	7,516,858	2017209	1965731	1965731	6,831,434	1,855,547	1,803,066	1,803,066	6,515,602	1,574,651	1,531,851	1,531.851	8,073,855	2,251,912	2,192,923	2,192,923
	R3	8,070,714	2239014	2184827	2184827	7,488,938	1,932,095	1,865,384	1,865,384	7,299,915	1,729,866	1,673,787	1,673.787	7,948,992	2,145,919	2,088,329	2,088,329
7 DPT	R1	10,527,034	2303592	2232434	2232434	7,497,557	1,726,070	1,661,225	1,661,225	10,252,126	2,324,072	2,262,223	2,262.223	8,490,916	2,030,865	1,968,954	1,968,954
	R2	7,414,837	1967699	1910652	1910652	6,886,546	1,792,722	1,740,009	1,740,009	11,116,456	2,486,979	2,420,351	2,420.351	8,532,772	1,980,148	1,921,571	1,921,571
	R3	10,082,155	2330790	2266509	2266509	7,475,776	1,991,955	1,926,531	1,926,531	7,699,841	1,967,116	1,919,834	1,919.834	7,191,818	1,633,590	1,580,098	1,580,098
10 DPT	R1	9,929,272	2363429	2302364	2302364	8,788,873	1,977,247	1,892,145	1,892,145	8,149,714	2,159,360	2,102,233	2,102.233	7,672,946	1,777,350	1,714,941	1,714,941
	R2	9,761,978	2423642	2367327	2367327	8,247,600	2,125,387	2,070,517	2,070,517	9,177,379	2,276,314	2,215,159	2,215.159	6,961,867	1,684,655	1,625,316	1,625,316
	R3	9,052,566	2172642	2115103	2115103	8,236,354	2,036,309	1,976,852	1,976,852	9,657,322	2,335,165	2,264,356	2,264.356	10,384,851	2,403,684	2,318,637	2,318,637

Figure 5.1: Processed reads and read counts after unique sequence identification, removal of ncRNAs and consistent naming for different days post treatment (DPT).

For conserved miRNA identification, in total we found 87 conserved miRNA families from the miRBase database that matched with 613 sequences from experiment (with at least 10 RPM in any experimental sample, see Section 4.4.2 for details on experimental sequences) with 0 to 4 mismatches. Among the 87 conserved miRNA families, many miRNA families matched multiple experimental sequences. For example, conserved miRNA family tae-miR159b matched 150 experimental sequences, which is the highest number of experimental sequences matched with a particular conserved miRNA family.

5.2 Conserved miRNAs identification using supplementary materials of Mayer et al. and Sun et al.

We identified a total of 232 experimental sequences as a result of comparing our experimental sequences (filtered data with at least 10 RPM in any experimental sample, generated in Section 4.4.2) with the Mayer et al. and Sun et al. datasets. Among these 232 sequences, we found 205 sequences from the Mayer et al. dataset and 12 sequences from the Sun et al. dataset. Fifteen experimental sequences were common in both the Mayer et al. and Sun et al. analyses. For each of these matched experimental sequences, we recorded the conserved miRNA family, the sequence and the dot-bracket notation of the species, the total number of contigs matched with the sequence, and precursor information (name, sequence and dot-bracket notation).

For example, we found that the conserved miRNA family miR398 expressed as down-regulated for heat stress in all days (see Section 5.3 for more details on differential expression). Figure 5.2 shows the structure of the precursor of miR398, matched experimental sequence (green) and predicted miRNA star sequence (red). See Section 4.7 for more details on precursor prediction.

If we compare our experiment with Mayer et al. [2014] experiment, in our experiment 31,512 precursor sequences out of 98,068 (32.1%) aligned with the experimental sequences (filtered data with at least 10 RPM in any experimental sample, generated in Section 4.4.2), whereas 1,668 precursor sequences out of 98,068 (1.7%) aligned with wheat expressed sequence tags (ESTs) and reads from RNA-seq studies in Mayer et

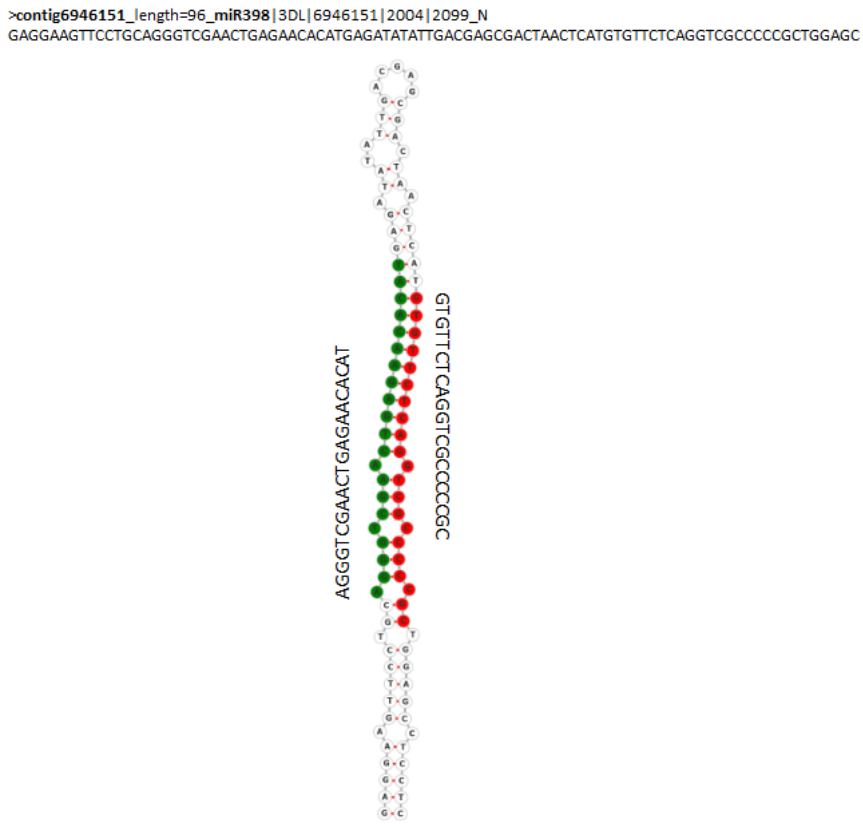

Figure 5.2: Conserved miRNA family miR398: structure of the precursor, matched experimental sequence (species6_sample68of72_21bp, coloured as green), and predicted miRNA star sequence (species134_sample68of72_21bp, coloured as red).

al. experiment.

We also grouped together the same conserved miRNA families that matched with different experimental sequences. Figure 4.12 shows an example of this. In total we identified 36 conserved miRNA families that matched with the Mayer et al. and Sun et al. datasets.

5.3 Differential gene expression

This is a preliminary analysis that did not take into consideration the library-based and model-based normalizations required to fit the binomial distribution of these data.

We observed that abiotic stresses (heat, light and UV) can cause different miRNA families to be up-regulated (increased in expression) or down-regulated (decreased in expression) between treatment and control groups. For this, we used edgeR [Robinson et al., 2010], a bio-conductor package of R language with the count values calculated in Section 4.2 and with the experimental sequences of Section 4.7 and 4.8 which we identified as potential conserved miRNAs.

Our null hypothesis (H0) is that there is no difference between control and treatments (heat, UV and continuous light) in the expression of miRNAs. Our alternative hypothesis (H1) is that there is differential expression of miRNAs between control and treatments.

P-value is a test statistic returned in null hypothesis statistical significance testing. A P-value of 0.05 means that there is a 5% chance of observation of test statistic value purely by chance, even without treatment effects. In simple words, there is a probability for 5% false positives in our results. A P-value threshold of 0.05 is used to reject the null hypothesis. On the other hand, false discovery rate (FDR) adjusted P-value is used for multiple hypothesis testing like differential expression of many miRNAs. For instance, if we measure differential expression of 2500 miRNA genes, then a simple P value of 5% means 125 genes. Hence, FDR adjusted p-value of 0.05 mean 5% of the result (125 genes) will be false positives.

We used <0.05 as the value of false discovery rate (FDR) to determine which miRNA families were up-regulated or down-regulated. FDR is a statistical method to correct multiple comparisons [Benjamini and Hochberg, 1995]. The Benjamini Hochberg FDR reduces the FDR at P ¡ 0.05. Up-regulated and down-regulated miRNA families were determined by positive log fold change and negative log fold change respectively. Figure 5.3 shows the differential expression of each stress for all 6 days. For example, on day 7 with the heat stress, among 36 conserved miRNA families, 2 families were down-regulated, 25 families were up-regulated and 9 families were not differentially expressed.

For our data, we observed more differential expression for the heat stress compared to light and UV stresses. Thirty-four conserved miRNA families were differentially expressed for the heat stress whereas only 8 conserved miRNA families were differentially expressed for light and only 7 conserved miRNA families were differentially expressed for UV stress. Figure 5.4 shows the number of conserved miRNA families differentially expressed for control versus stresses for all days.

Figure 5.5 shows the number of conserved miRNA families differentially expressed for control versus stresses on different days.

Again, different conserved miRNA families expressed differently based upon stress or day. We found that miRNA 395 and 398 were strongly suppressed whereas miRNA 5064, 5175, 2020, and 1439 were expressed with heat stress for all days. Figure 5.6 shows an example of this using logarithm fold change (logFC) values that represents the differential expression: zero represents no change, positive value represents up-regulated and negative value represents down-regulated. Also, we found that miRNA

Day/Stress		Down-regulated	Non-differentially expressed	Up-regulated
0	Heat	4	26	6
	Light	0	36	0
	UV	0	36	0
1	Heat	7	21	8
	Light	0	36	0
	UV	0	36	0
2	Heat	8	19	9
	Light	1	35	0
	UV	2	31	3
3	Heat	7	19	9
	Light	2	34	0
	UV	0	36	0
7	Heat	2	9	25
	Light	5	31	0
	UV	2	34	0
10	Heat	2	26	8
	Light	0	36	0
	UV	0	36	2

Figure 5.3: Differential gene expression of 36 conserved miRNA families at 6 post-time points with heat, light and UV stresses.

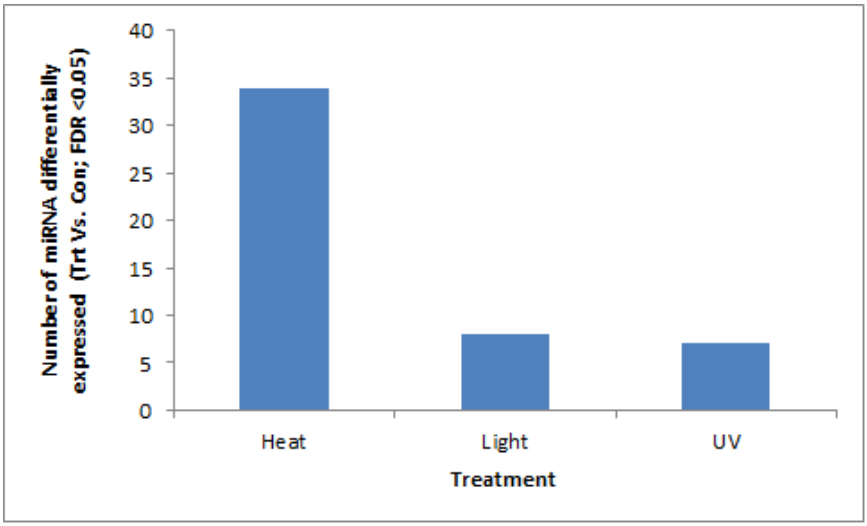

Figure 5.4: Total number of conserved miRNA families differentially expressed for control versus heat, light and UV stresses.

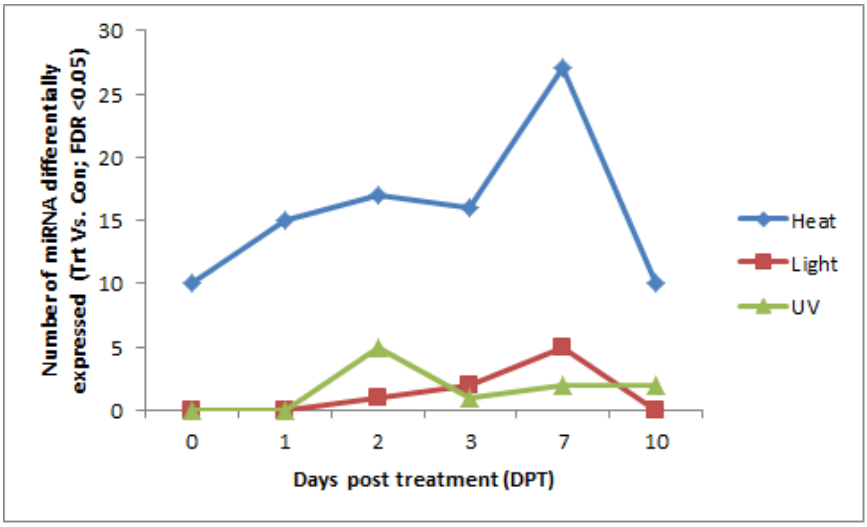

Figure 5.5: Number of conserved miRNA families differentially expressed for control versus heat, light and UV stresses in each day.

395 was suppressed with all stresses for all days.

Moreover, based on the differential expressions found each day, we generated two Venn diagrams: Figure 5.7.a shows a Venn diagram with the total number of conserved miRNA families differentially expressed for each stress of heat, light and UV,

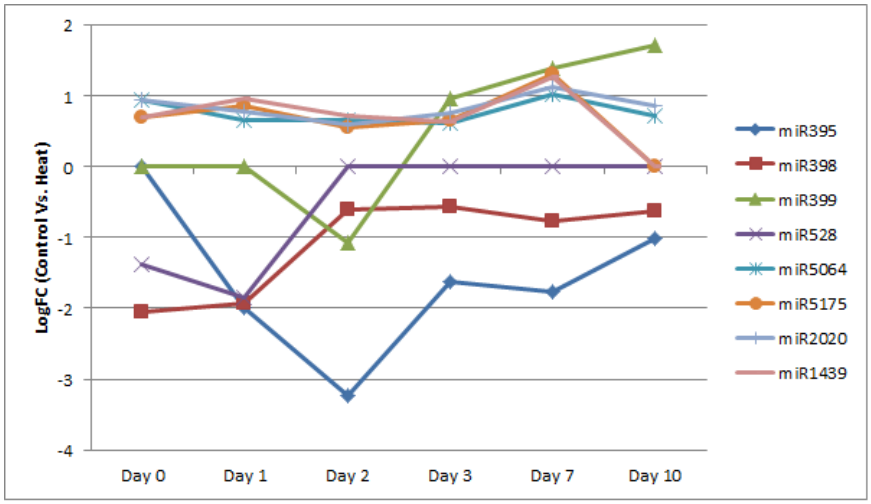

Figure 5.6: Differential expression for the conserved miRNA families 398, 398, 399, 528, 5064, 5175, 2020, and 1439, with heat stress in each day.

and Figure 5.7.b shows a Venn diagram with the name of the conserved miRNA families differentially expressed for each stress of heat, light and UV. These figures give the differentially expressed conserved miRNA families (both up-regulated and down-regulated) for day 7 only. The Venn diagrams represent how many conserved miRNAs were differentially expressed each day with different stresses and which conserved miRNAs are commonly expressed among the stresses for each day . For example, in the figure, for day 7, conserved miRNA family Tae-miR2020b and miR5064 are expressed in heat and light stress.

In addition, we also generated Venn diagrams for heat, light and UV stresses for each day to observe how increasing number of days affected the expression of conserved miRNA families with the stresses. Figure 5.8.a shows the differentially expressed conserved miRNA families for heat stress in days 1, 2, 3, 7 and 10. Figure 5.8.b shows the differentially expressed conserved miRNA families for heat stress for each day combining day 0 with Figure 5.8.a (due to the limitation of the tools for

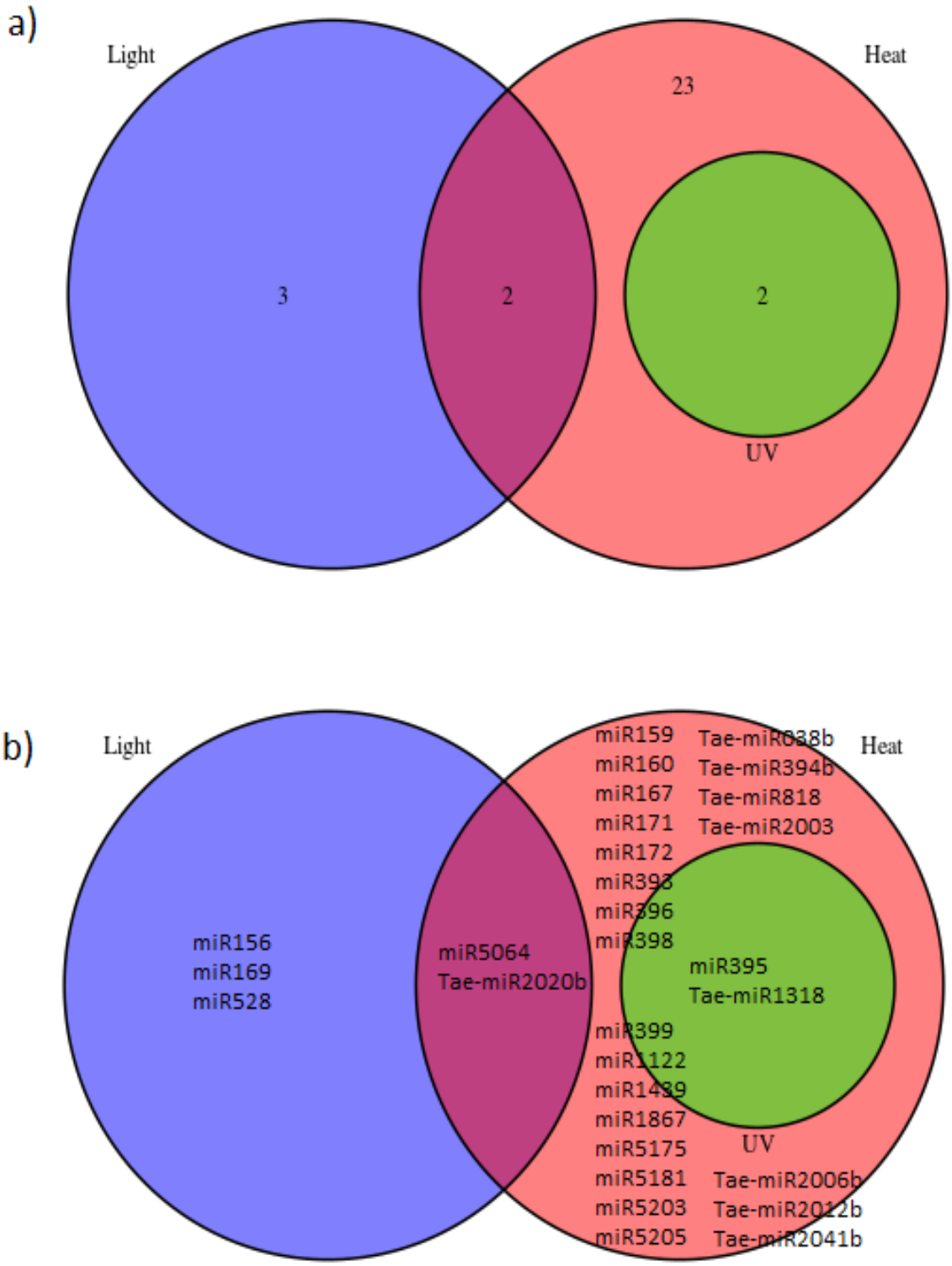

Figure 5.7: Venn diagrams of differentially expressed conserved miRNA families for day 7 with heat, light and UV stresses: a) with count values and b) with elements.

creating Venn diagram that does not support more than 5 sets). For example, in the figure, 3 conserved miRNA families expressed in all days for the heat stress.

5.4 Comparison with *Brassica rapa* dataset

To evaluate the toolchain, we executed it using the dataset provided by Bilichak et al. [2015]. Then we compared Bilichak et al. [2015]'s results and our results for *Triticum aestivum*.

5.4.1 Comparison with miRBase database

The supplementary dataset of Bilichak et al. [2015] contains different tissues of *Brassica rapa* such as leaf, pollen, embryo, endosperm and progeny. The experiment was completed with control (no stress) and heat stress. As the experiment used the wheat (*Triticum aestivum*) leaf tissue, we also applied *Brassica rapa*'s leaf tissue to my toolchain. At first, we removed the adapter sequences, provided by the authors, using "cutadapt" command [Martin, 2011]. Adapter sequences are short length of known DNA sequence that were added at the ends of the cDNA sequences. Then, we trimmed the sequences to 18 to 24 bp. We executed the same steps of Sections 4.2, 4.3, 4.4 and 4.5. In Section 4.4.2, we identified 15,158 sequences with the *Triticum aestivum* dataset. On the other hand, we identified 7,065 sequences in the *Brassica rapa* dataset.

For the *Triticum aestivum* dataset, in total we found 87 conserved miRNA families from the miRBase database that matched with 613 sequences. On the other hand, for the *Brassica rapa* dataset, we found 71 conserved miRNA families from the miRBase

database that matched with 146 sequences.

We identified 613 sequences out of 15,158 *Triticum aestivum* sequences (4.04%) whereas we identified 146 sequences out of 7,065 *Brassica rapa* sequences (2.07%), that matched with the 87 and 71 conserved miRNA families from miRBase, respectively. Figure 5.9 shows a comparison of this. One reason for obtaining a higher percentage of matched sequences with miRBase for *Triticum aestivum*, may be that researchers are doing more research concerning miRNAs in *Triticum aestivum* or species related to it than *Brassica rapa*, which causes more entries in miRBase similar to the *Triticum aestivum* miRNAs than the *Brassica rapa* miRNAs.

5.4.2 Comparison with differential expression

Bilichak et al. [2015] applied the DESeq bioconductor package and reported that miRNA family 168 is differentially expressed in the endosperm tissue of heat-stressed plants. To confirm the similar result with our toolchain, we applied edgeR bioconductor package, similar to Section 4.9, with the sequences from the endosperm tissue [Bilichak et al., 2015] that matched with the miRBase conserved miRNA families database, processed in the same way as the data in Section 5.4.1. We also identified miRNA family 168 as being differentially expressed in the endosperm tissue of heat-stressed plants. Thus we identified the same differentially expressed miRNA family 168. Bilichak et al. [2015] identified miRNA bra-miR168 family with 6.48 log2 fold change whereas we identified miRNA cca-miR168 family with 5.83 log2 fold change. Our log fold change may be lower because we mapped the *Brassica rapa* dataset to miRBase with 0 to 4 mismatches, which may possibly have excluded some *Brassica*

rapa miRNAs.

We could not use the toolchain portions from Section 4.7 and 4.8 as Bilichak et al. [2015] did not use any *Triticum aestivum* dataset or precursors.

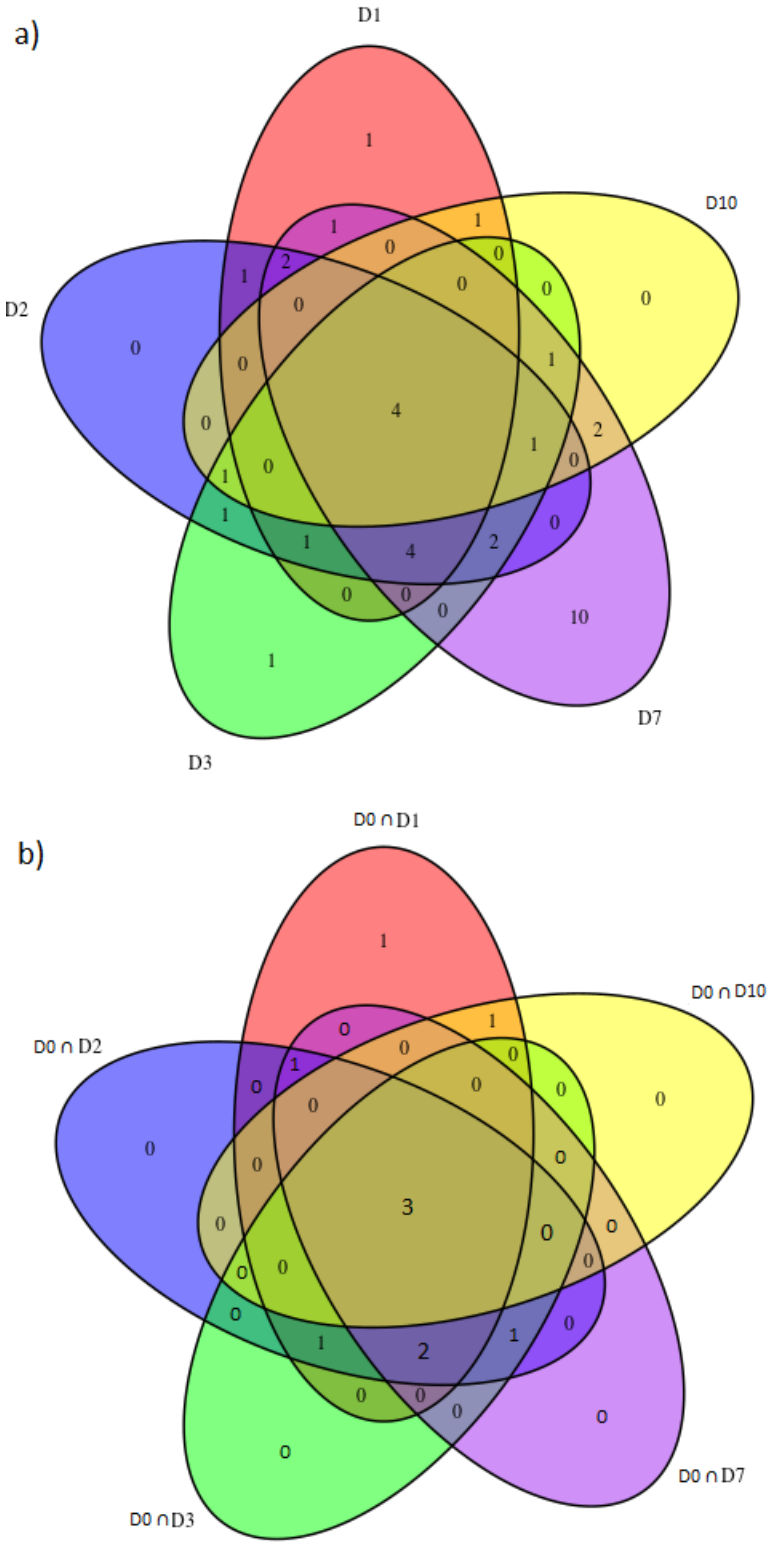

Figure 5.8: Venn diagram of differentially expressed conserved miRNA families for heat stress for each day: a) day 1 to day 7, and b) intersection of the previous Venn diagram and day 0.

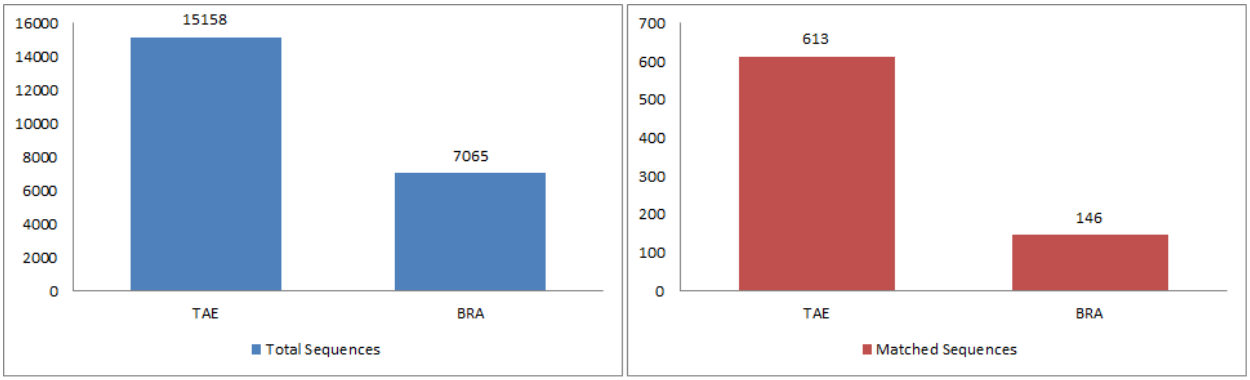

Figure 5.9: Comparison of total experimental sequences and matched miRBase sequences between *Triticum aestivum* (TAE) and *Brassica rapa* (BRA).

Chapter 6

Conclusion

To identify conserved miRNAs from the wheat dataset, we designed a toolchain. We examined $\sim$ 523 million reads and filtered it down to 15,158 experimental sequences having at least ten RPM in any of the 72 experimental samples. Using the toolchain, we identified 36 wheat conserved miRNA families that matched between 232 experimental sequences and datasets from two recent papers by the Mayer et al. [2014] and Sun et al. [2014]. Moreover, we found 87 plant conserved miRNA families that matched between 613 experimental sequences and the miRBase [Kozomara and Griffiths-Jones, 2014] dataset.

In addition, we observed significant differential expression for the wheat exposed to the heat stress compared to those exposed to light and UV stresses or no stress (control). Thirty-four conserved miRNA families were differentially expressed for the heat stress whereas only 8 conserved miRNA families were differentially expressed for light and only 7 conserved miRNA families were differentially expressed for UV stress. We also found that increasing number of days post treatment affected the number of

conserved miRNA families differentially expressed for control versus stresses. Again, different conserved miRNA families expressed differently based upon stress or day. We found that miRNA 395 and 398 were strongly suppressed whereas miRNA 5064, 5175, 2020, and 1439 were expressed with heat stress at all post-stress time points. MiRNA 395 was suppressed in all stresses samples regardless of the stress or time point of the stress.

6.1 Importance of our thesis

Researchers can use the obtained conserved miRNAs and matched experimental sequences for the heat, light and UV stresses and find out the target genes using the Plant Small RNA Target Analysis Server (psRNATarget) [Dai and Zhao, 2011]. Then, the researchers can analyze how those target genes affect wheat phenotypes under different environmental conditions.

Moreover, from our experiment, breeders may want to focus more on heat stress than light or UV stress as we found significant differential expression for the wheat exposed to the heat stress compared to those exposed to light and UV stresses.

6.2 Future work

In Section 4.6, we partially implemented novel miRNA identification. In particular, to detect novel miRNA, the unfinished tasks are future work due to time restriction. A sequence alignment between the input sequence and the wheat genome sequence can be performed using Bowtie 2. After finding the matched portion in

genome sequence with the input sequence, the matched portion of sequence can be extended on both left and right side. This sequence can be extracted to be tested as a possible precursor.

Then, if this potential precursor produces a stem-loop structure having <-0.2 Kilocalorie/mole/nt folding free energy [Kozomara and Griffiths-Jones, 2014], only then we can conclude that the input sequence is a novel miRNA. Thus, after precursor identification the secondary structure of the putative precursors can be predicted using RNAfold.

The extension length of the matched portion can be varied depending on the putative precursor. Besides the stem-loop structure of the putative precursor, RNAfold also produces the dot-bracket notation of the putative precursor. We will only retain those predicted hairpin structure having <-0.2 Kilocalorie/mole/nt folding free energy [Kozomara and Griffiths-Jones, 2014] and the remaining precursor sequences can be discarded. Then, the star sequences can be predicted for the putative novel miRNAs by the process described in Section 4.7.2 using the predicted putative precursor and dot-bracket notation.

Bibliography

S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

Annenberg. Human genome : Genes and transcription. `http://www.learner.org/interactives/dna/project6.html`, 2015. Online; accessed 30-March-2015.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

A. Bilichak, Y. Ilnytskyy, R. Wóycicki, N. Kepeshchuk, D. Fogen, and I. Kovalchuk. The elucidation of stress memory inheritance in *Brassica rapa* plants. *Frontiers in plant science*, 6, 2015.

BLAST. Blast output. `http://parts.igem.org/File:VF_Blast_Example.png`, 2015. Online; accessed 19-January-2015.

BMC. Supplementary materials of whole-genome discovery of miRNAs and their targets in wheat (*Triticum aestivum* l.). `http://www.biomedcentral.com.proxy1.`

lib.umanitoba.ca/1471-2229/14/142/suppl/S2, 2014. Online; accessed 03-December-2014.

R. Brenchley, M. Spannagl, M. Pfeifer, G. L. Barker, R. DAmore, A. M. Allen, N. McKenzie, M. Kramer, A. Kerhornou, D. Bolser, et al. Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature*, 491(7426):705–710, 2012.

R. D. Burdine and E. Sheldon. From DNA to protein. http://www.cureangelman.org/what-testing101.html, 2015. Online; accessed 19-January-2015.

S. W. Burge, J. Daub, R. Eberhardt, J. Tate, L. Barquist, E. P. Nawrocki, S. R. Eddy, P. P. Gardner, and A. Bateman. Rfam 11.0: 10 years of RNA families. *Nucleic Acids Research*, 41(D1):D226–D232, 2013.

K. Chen and N. Rajewsky. The evolution of gene regulation by transcription factors and microRNAs. *Nature Reviews Genetics*, 8(2):93–103, 2007.

F. Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

Cronodon. Stem-loop structure. http://cronodon.com/BioTech/Ribosomes.html, 2015. Online; accessed 22-January-2015.

X. Dai and P. X. Zhao. psRNATarget: a plant small RNA target analysis server. *Nucleic Acids Research*, 39(suppl 2):W155–W159, 2011.

K. Darty, A. Denise, and Y. Ponty. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, 25(15):1974, 2009.

Ensembl. Genome assembly. `http://uswest.ensembl.org/Help/Faq?id=216`, 2015. Online; accessed 27-January-2015.

S. Garcia, A ; Mazoyer. microRNA 146b. `http://atlasgeneticsoncology.org/Genes/MIR146BID50855ch10q24.html`, 2015. Online; accessed 17-January-2015.

N. Geographic. Genetics overview. `https://genographic.nationalgeographic.com/science-behind/genetics-overview`, 2015. Online; accessed 19-January-2015.

H. G. P. Information. Contig. `http://web.ornl.gov/sci/techresources/Human_Genome/glossary.shtml`, 2015. Online; accessed 27-January-2015.

S. Kadri, V. Hinman, and P. V. Benos. HHMMiR: efficient de novo prediction of microRNAs using hierarchical hidden markov models. *BMC Bioinformatics*, 10 (Suppl 1):S35, 2009.

K. Katoh, K. Misawa, K.-i. Kuma, and T. Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Research*, 30(14):3059–3066, 2002.

V. N. Kim. MicroRNA biogenesis: coordinated cropping and dicing. *Nature reviews Molecular cell biology*, 6(5):376–385, 2005.

A. Kozomara and S. Griffiths-Jones. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research*, 42:D68–D73, 2014.

J. Krol, K. Sobczak, U. Wilczynska, M. Drath, A. Jasinska, D. Kaczynska, and W. J. Krzyzosiak. Structural features of microRNA (miRNA) precursors and their rele-

vance to miRNA biogenesis and small interfering RNA/short hairpin RNA design. *Journal of Biological Chemistry*, 279(40):42230–42239, 2004.

K. Y. Kurtoglu, M. Kantar, and H. Budak. New wheat microRNA using whole-genome sequence. *Functional & Integrative Genomics*, 14(2):363–3791, 2014.

B. Kusenda, M. Mraz, J. Mayer, S. Pospisilova, et al. MicroRNA biogenesis, functionality and cancer relevance. *Biomedical Papers*, 150(2):205–215, 2006.

B. Langmead and T. Cole. Bowtie 2. `http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml`, 2015. Online; accessed 13-January-2015.

B. Langmead and S. L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, 2012.

J. Lei and Y. Sun. miR-PREFeR: an accurate, fast and easy-to-use plant miRNA prediction tool using small RNA-seq data. *Bioinformatics*, pages 1–3, 2014.

R. Lorenz, S. H. Bernhart, C. H. Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, I. L. Hofacker, et al. ViennaRNA package 2.0. *Algorithms for Molecular Biology*, 6(1):26, 2011.

M. Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1):pp–10, 2011.

K. F. Mayer, J. Rogers, J. Doležel, C. Pozniak, K. Eversole, C. Feuillet, B. Gill, B. Friebe, A. J. Lukaszewski, P. Sourdille, et al. A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science*, 345 (6194):1251788, 2014.

U. N. L. Medicine. Chromosome. `http://ghr.nlm.nih.gov/handbook/basics/chromosome`, 2014. Online; accessed 26-October-2014.

U. N. L. Medicine. Genes. `http://ghr.nlm.nih.gov/handbook/basics/gene`, 2015. Online; accessed 27-March-2015.

miRBase. Precursor database. `http://www.mirbase.org/ftp.shtml`, 2015. Online; accessed 05-April-2015.

miReap. miReap. `http://mireap.sourceforge.net`, 2015. Online; accessed 12-April-2015.

R. A. C. Montes, E. De Paoli, M. Accerbi, L. A. Rymarquis, G. Mahalingam, N. Marsch-Martínez, B. C. Meyers, P. J. Green, S. de Folter, et al. Sample sequencing of vascular plants demonstrates widespread conservation and divergence of microRNAs. *Nature Communications*, 5, 2014.

NCBI. BLAST. `http://www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml`, 2015. Online; accessed 13-January-2015.

NRCC. Canadian wheat alliance. `http://www.nrc-cnrc.gc.ca/eng/news/releases/2013/wheat_nrc_factsheet.html`, 2015. Online; accessed 13-January-2015.

S. Ouyang and C. R. Buell. The TIGR plant repeat databases: a collective resource for the identification of repetitive sequences in plants. *Nucleic Acids Research*, 32 (suppl 1):D360–D363, 2004.

T. Papagiannakopoulos and K. S. Kosik. MicroRNAs: regulators of oncogenesis and stemness. *BMC Medicine*, 6(1):15, 2008.

H. Pearson. Genetics: what is a gene? *Nature*, 441(7092):398–401, 2006.

M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

Rosalind. Global and local alignment. `http://rosalind.info/problems/swat/`, 2015. Online; accessed 13-January-2015.

G. Ruvkun. Glimpses of a tiny RNA world. *Science*, 294(5543):797–799, 2001.

Science. Supplementary materials of a chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. `http://www.sciencemag.org/content/345/6194/1251788/suppl/DC1`, 2014. Online; accessed 13-December-2014.

Stanford. Sequence alignment. `http://statweb.stanford.edu/~nzhang/345_web/`, 2015. Online; accessed 13-January-2015.

F. Sun, G. Guo, J. Du, W. Guo, H. Peng, Z. Ni, Q. Sun, and Y. Yao. Whole-genome discovery of miRNAs and their targets in wheat (*Triticum aestivum* l.). *BMC Plant Biology*, 14(1):142, 2014.

J.-H. Teune and G. Steger. NOVOMIR: de novo prediction of microRNA-coding regions in a single plant-genome. *Journal of Nucleic Acids*, 2010, 2010.

J. D. Watson and F. Crick. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171:737–738, 1953.

Westgrid. Hermes. `https://www.westgrid.ca/support/systems/hermesnestor`, 2015. Online; accessed 28-February-2015.

H.-J. Wu, Y.-K. Ma, T. Chen, M. Wang, and X.-J. Wang. PsRobot: a web-based plant small RNA meta-analysis toolbox. *Nucleic Acids Research*, pages W22–W28, 2012.

Y. Yao, G. Guo, Z. Ni, R. Sunkar, J. Du, J.-K. Zhu, and Q. Sun. Cloning and characterization of microRNAs from wheat (*Triticum aestivum* l.). *Genome Biology*, 8 (6):R96, 2007.

Z. Yin and F. Shen. Identification and characterization of conserved microRNAs and their target genes in wheat (*Triticum aestivum*). *Genetics and Molecular Research*, 9(2):1186–1196, 2010.

Z. Zhang, J. Yu, D. Li, Z. Zhang, F. Liu, X. Zhou, T. Wang, Y. Ling, and Z. Su. PMRD: plant microRNA database. *Nucleic Acids Research*, 38(suppl 1):D806–D813, 2010.

M. Zuker. Computer prediction of RNA structure. *Methods in Enzymology*, 180: 262–288, 1989.

M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9(1):133–148, 1981.

M. Zuker, D. H. Mathews, and D. H. Turner. Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide. In *RNA Biochemistry and Biotechnology*, pages 11–43. Springer, 1999.