CHAPTER 1: INTRODUCTION

Small RNAs (sRNAs) are short non-coding RNAs which plays important regulatory roles in gene expression. The sRNAs categorized into two classes, microRNAs (miRNAs) and small interfering RNAs (siRNAs). The miRNAs downregulate gene expression by attaching themselves to messenger RNAs (mRNAs) and preventing them from being translated into proteins. The siRNAs mediate the phenomenon of RNA interference (RNAi) (Buckingham, 2003).

In plants, due to the complementarity between the miRNA and its target is very high, sRNA-mediated post-transcriptional gene regulation generally leads to messenger RNA (mRNA) cleavage and degradation. The cleavage of the target RNA is specific and typically occurs at the center of the paired region. The cleavage of mRNA occurs between positions 10th and 11th nucleotide of the bound miRNA (Folkes *et al.,* 2012). This perfect complementarity widely used to predict miRNA targets in plants (Zheng *et al.,* 2012).

The sRNA and degradome data can be used to identify interactions between sRNAs and their target mRNAs. Degraded mRNA fragments provide support for the interaction between sRNAs and their complementary mRNA targets that lead to cleavage and degradation of the mRNA. 'Parallel Analysis of RNA Ends' (PARE) profiles the mRNA cleavage products induced by sRNA on a large-scale. PARE sequences the 5'-ends of uncapped mRNAs including all transcripts targeted by sRNAs and subjected to endonucleolytic cleavage (Folkes *et al.,* 2012). PARE matches 5′ end sequences of sRNA cleavage products back to their corresponding cDNA sequences to detect mRNA cleavage sites. This method enables to identify matches to known or new potential miRNAs that could direct their cleavage (German *et al.,* 2008).

Currently, there are many tools to discover targets of miRNAs in plants based on predictions programs using pre-defined rules but CleaveLand, SeqTar and PARESnip are the only available tools specifically to analyse degradome data in plants.

CleaveLand is a publicly available computational method for identifying plant miRNA targets from degradome data. CleaveLand scores sRNA complementary sites based on a mismatch-based scoring scheme (Zheng *et al.*, 2012).

SEQuencing-based sRNA TARget prediction (SeqTar) uses a modified Smith-Waterman algorithm to align sRNA to a target sequence. SeqTar is designed to allow more mismatches in order to identify targets with weak complementarities from degradome data (Zheng *et al.*, 2012). A study by Zheng *et al.* on *Arabidopsis* and rice degradome data sets using SeqTar has predicted a solid number of novel miRNA targets in addition to previously reported targets and 12 novel miRNA targets were experimentally verified (2012).

PARESnip is constructed within the UEA sRNA Workbench to predict sRNA target interactions with degradome sequencing. PARESnip operates by Rule-Based Complementarity Search algorithm traversing the partitioned 4-way tree (Folkes *et al.*, 2012).

In this study, degradome data from the inflorescence of the African Oil Palm (*Elaeis guineensis Jacq.*) was analysed. The objective of this study was to identify miRNA/siRNA cleavage products and SNPs in miRNA targets. Degradome analysis tools CleaveLand, SeqTar and PARESnip were used to analyse degradome data towards the objective of identifying miRNA target pairs that may be important in oil palm flowering and hence yield of this major crop.

CHAPTER 2: LITERATURE REVIEW

2.1 Importance of African Oil Palm (*Elaeis guineensis Jacq.*)

The African Oil Palm (*Elaeis guineensis Jacq.)*, a monocotyledon plant belongs to Arecaceae family, is an economically important species. Palm oil extracted from mesocarp is used in food products and palm kernel oil produced by seed or kernel is used widely by oleochemical industry for making soap, cosmetics, detergent and toiletries.

Malaysia is a global leader in the palm oil and basic oleochemicals industry. Tenth Malaysian Plan targets gross domestic product (GDP) of the palm oil industry to achieve RM21.9 billion with export earnings of RM69.3 billion. Palm Oil Industrial Clusters are developed to promote sources for biofuel, oleochemicals, biomass products, biofertilisers, nutraceuticals, pharmaceuticals and specialty food products (Tenth Malaysian Plan, 2010). Biofuel, an alternative green renewable for diesel and capsulated palm oil as dietary supplement due to rich natural source of vitamin E are examples of new uses for palm oil. A high demand for palm oil is expected in near future due to its broadened uses (Basiron, 2007).

Elaeis guineensis is a tall perennial and solitary palm typically reaching, 8.5-30 m tall. The trunk is stout, erect and ringed covered by the persistent leaf-bases above, bare below, dark grey-brown. The African Oil Palm is monoecious with male and female flowers in separate clusters but on the same tree. Alternatively, separate male and female inflorescences are produced on the same palm although mixed sex inflorescences are occasionally observed. The male inflorescence bears individual staminate flowers; the female inflorescence produces floral triads consist of a pistillate flower strengthen by two staminate flowers. The inflorescence of African Oil Palm is large, with head-like with

spiny tipped branches held near the trunk among the leaves. It is dense with numerous stout spikes tipped by short spines and produces trimerous flowers (Adam *et al.,* 2007).

In the oil palm, the inflorescence buds in the axil of each leaf from an early stage right after germination. This development continues throughout the lifetime of the plant. Meanwhile, the development of the inflorescence takes over 2 years while the organ is completely enclosed at the base of the subtending leaf for most of this time (Adam *et al.,* 2005). This makes the species allogamous, which enables cross fertilization. Thus, seed progenies of oil palm show significant levels of genetic variability. Due to its long life cycle, there is a high demand for clonally breeded palms to be distributed in a shorter time frame as an advanced planting material (Beulé *et al.,* 2011).

2.2 Small RNA (sRNA) in Plants

The sRNAs are riboregulators involved in a variety of silencing processes in plants. The sRNAs consist of 20-27 nucleotides which include two predominant groups, microRNA (miRNA) and small interfering RNA (siRNA) (Vaucheret., 2006).

The miRNAs are about 21 nucleotides in length and are involved in post-transcriptional regulation of gene expression (German *et al*., 2008). This is done through base-pairing to the target mRNA causing translational repression or silencing. The siRNAs are around 22 nucleotides in length and are involved in RNA interference (RNAi). The siRNAs bind and label to specific mRNAs for destruction by endonucleases enzymes (Buckingham, 2003). By triggering DNA methylation and histone modifications, siRNAs are able to direct post-transcriptional gene silencing through mRNA degradation or transcriptional gene silencing (Sunkar *et al.,* 2007). The miRNAs are chemically and functionally similar to siRNAs in mediating RNA interference (RNAi), post-transcriptional gene silencing (PTGS) and transcriptional gene silencing

(TGS). The miRNAs and siRNAs are created in the same way which is processed by the Dicer RNaseIII family of enzymes. The siRNAs differs from miRNAs because processed from long and double-stranded RNAs instead of deriving from local stem-loop structures (Jones-Rhoades *et al.*, 2006).

The Argonaute protein is a part of a ribonucleoprotein complex known as RNA-induced silencing complex (RISC) and essential in all known sRNA-directed regulatory pathways (Vaucheret, 2008). In order to guide repression of target genes both miRNAs and siRNAs are incorporated into silencing complexes that contain Argonaute (AGO) proteins (Jones-Rhoades *et al.*, 2006). AGO-complexes are guided by sRNAs to regulate target RNA sequences based on Watson-Crick base pairing. AGO-catalyzed target cleavage enables plant miRNAs to have perfect or near perfect complementarity to their target mRNAs which occurs exactly between the 10th and 11th nucleotide of complementarity relative to the 5'-end sRNA (Addo-Quaye *et al.*, 2008). Fast and confident bioinformatics identification of plant miRNA targets are possible due to plant miRNAs ability to shows high degree of complementary with conserved target mRNAs (Jones-Rhoades *et al.*, 2006).

2.3 Degradome Data and Parallel Analysis of RNA Ends (PARE)

Degradome sequencing (Degradome-Seq) is known as parallel analysis of RNA ends (PARE). In this method, the 5'-ends of polyadenylated products of sRNA-mediated mRNA decay are sequenced and subsequently aligned to the cDNA sequences to detect mRNA cleavage sites and quantify the abundance of cleavage products to determine the effects of sRNA-guided gene expression regulation. PARE method is cost effective compared to experimental validation with modified 5'- RACE assay (Stocks *et al.*, 2012).

Degradome data can be examined closely to find evidence of cleaved sRNA targets. It is important to use degradome data because degradome sequencing provides a

comprehensive means of analyzing patterns of RNA degradation. It is difficult to differentiate between miRNAs and siRNAs due to their structural and biochemical characteristics. Most genuine miRNA precursors have typical cleavage patterns whereas siRNA precursors are cleaved randomly. Degradome patterns can help to recognize and differentiate miRNAs from other small RNA-producing loci (Li *et al.*, 2010). Degradome sequencing is an effective method for confirming small RNA targets in plants due to its ability to identify additional conserved and nonconserved targets for miRNAs (Li *et al.*, 2010).

2.4 MicroRNAs and Expressed Sequence Tags (EST)

The sequences of the 5' and or 3' ends of randomly selected cDNA clones serves as markers or tags for transcripts. ESTs (Ho *et al.*, 2007) are useful for the discovery of novel genes, investigation of genes of unknown function, recognition of exon/intron boundaries and comparative genomic study. Lack of sequence information (Ho *et al.*, 2007) limits the progress of gene discovery and characterisation, global transcript profiling, probe design for development of gene arrays, and generation of molecular markers for oil palm.

Study by Nasaruddin *et al.*, 2007 shows five potential miRNA encoding sequences by a combined homology and structural analysis approach were found in oil palm EST sequences expressed in apical meristem, immature and mature flowers. Regulation of the auxin response, floral development and the regulation of basal transcription are the potential roles of these miRNAs in oil palm. Useful data for the functional analysis of these sequences provided by the annotation of oil palm EST as miR gene candidates is important in the developmental biology of African Oil Palm (Nasaruddin *et al.*, 2007).

2.5 Bioinformatic Tools for Degradome Analysis

In this study CleaveLand, SeqTar and PARESnip were used to analyse the degradome data from African Oil Palm (*Elaeis guineensis Jacq.)* inflorescence.

2.5.1 CleaveLand

CleaveLand (Addo-Quaye *et al.*, 2009) detects cleaved miRNA targets from degradome data provided degradome sequences, sRNAs and an mRNA database and outputs sRNA targets. CleaveLand uses mismatch-based scoring scheme to score sRNA complementary sites. Scores are given accordable to the defined rules.

Rules	Score
Mismatch in sRNA complementary sites	1
G:U pair in sRNA complementary sites	0.5
A mismatch in the core region from 2 to 13 nucleotide	2
G:U pairs in the core region from 2 to 13 nucleotides	2
Mismatch at positions 10 and 11 in a complementary site	0
G:U pair at positions 10 and 11 in a complementary site	0

Table 2.5.1 Mismatch-based scoring scheme

According to mismatch-based scoring scheme, sRNA complementary sites with scores of ≤ 4 were used in identifying miRNA targets (Zheng *et al.*, 2012).

2.5.2 SeqTar (SEQuencing-based sRNA TARget prediction)

SeqTar (Zheng *et al.*, 2012) is a program that aligns sRNA to target sequences using a modified Smith–Waterman algorithm. SeqTar finds complementary nucleotides in alignments instead of performing alignments with matched nucleotides (A-A and C-C).

The affine gap penalty is used for gap opening and gap extension which is the penalty increasing linearly with the length of gap after the initial gap opening penalty.

Rules	Score
Complementary nucleotides G-C	+6
Complementary nucleotides A-U	+4
Complementary nucleotides G-U (Wobble pairs)	+2
Gap opening	-8
Gap extension	-4
Known mismatch	-3
Mismatch of unspecified nucleotides ('N')	-1

Table 2.5.2 Rules used in modified Smith–Waterman algorithm

2.5.3 PARESnip

PAREsnip uses the 5-category system which is also used in CleaveLand.

Category 0	>1 raw read at the position, abundance at position is equal to the maximum on the on the transcript, and there is only one maximum.
Category 1	Same as Category 0 in all aspects except that more than one maximum is found on the transcript. This implies that there are two or more signals on the transcript with the same strength
Category 2	>1 raw read at the position, abundance at position in is less than the maximum, but greater than the median abundance for the transcript
Category 3	>1 raw read at the position and the abundance at that position is less than or equal to the median value for that transcript
Category 4	One raw read at the position.

Table 2.5.3 5-category system in PAREsnip

Raw abundance is calculated by dividing the abundance of a degradome fragment (tag) by the number of positions across all transcripts to which the tag has aligned. The strongest signals are categories 0, 1 and 2 which shows the strongest empirical evidence for true cleavage products (Addo-Quaye *et al.*, 2008).

Small RNAs are encoded into a 4-way tree. The tree is partitioned based on the nucleotides at positions 10 and 11 in the pattern sequence to be searched for. As the tree is searched, sRNA target binding rules are applied accordingly. The tree is followed towards the root performing Watson and Crick base pairing. At each transversal, the binding rules are checked. If the root is reached successfully the algorithm jumps back to entry and begin pre-order walk down the tree, if the rules are broken the traversals of branch stops. If a terminator node is reached a successful alignment has been made and sRNA target interaction discovered (Folkes *et al.*, 2012).

sRNA target binding rules applied in Rule-Based Complementarity Search algorithm in traversing the partitioned 4-way tree.

Rule 1	≤ 4.5 mismatches between sRNA and target (A mismatch score 1, G-U wobble pair score 0.5, gap in alignment score 1)
Rule 2	≤ 2 adjacent mismatches in the sRNA target
Rule 3	No adjacent mismatches in the position 2-12 of the sRNA target from 5' end of the sRNA
Rule 4	No mismatches in position 10-11 of the sRNA target
Rule 5	≤ 2.5 mismatches in position 1-12 of the sRNA target from 5' of sRNA

Table 2.5.4 sRNA target binding rules applied in traversing the partitioned 4-way tree.

CHAPTER 3: METHODOLOGY

3.1 Degradome and Sequence Data Sets Used

Two degradome data sets derived from mature female flower (F1) and immature female flower (F4) were used in this study. Known miRNA sequences are orthologs which were annotated based on the miRNA Database (mirBase) matches. Novel miRNA are those with no match with mirBase but have predicted precursor miRNA. Novel and known miRNA sequences of mature and immature female flower in oil palm were used in the analysis. This study is focused on miRNA expressed during early and late flowering stages in oil palm. Non-redundant miRNA sequences were obtained by removing redundancy in the sequences with a compiled script (APPENDIX A).

3.2 The CleaveLand Pipeline

The pipeline requires input of three datasets in Fasta format which are degradome sequences trimmed to the 5′-most 20 nucleotides, a set of query small RNAs and a target database consisting mRNAs. FASTA headers for the degradome and small RNA files require special formatting which has very short headers with no whitespace or metacharacters. An example:

>t0000001 257932

CAACATAGGTCATCGAAAGG

P-values in CleaveLand were estimated using the likelihood of observing a degradome peak at the 10th nucleotide of a sRNA/mRNA alignment and related to the number of nucleotides in the transcriptome which have degradome data in a given library and the complementarity level of the sRNA/mRNA alignment itself. In CleaveLand, this chance is modelled as a binomial experiment.

Installation

CleaveLand Version 3.01 which is freely available under a GNU license at http://www.bio.psu.edu/people/faculty/Axtell/AxtellLab/Software.html was used in this study. CleaveLand3 is designed to process degradome data, small RNA target predictions and to output concise results depicted sliced small RNA target RNAs. CleaveLand was installed to a system with Intel Core i7 processor, 8GB ram and Linux 64 bit operating system. The modules Getopt::Std and File::Temp was ensured in the Perl path after installation of Perl 5.14.2. Module Math::CDF downloaded and installed from CPAN. The 'targetfinder.pl' version 1.6 was downloaded from http://carringtonlab.org/resources/targetfinder and installed according to installation notes provided. The fasta35 was downloaded and installed from the source http://fasta.bioch.virginia.edu/fasta_www2/fasta_list2.shtml. Finally R from http://www.r-project.org/ was installed.

Workflow

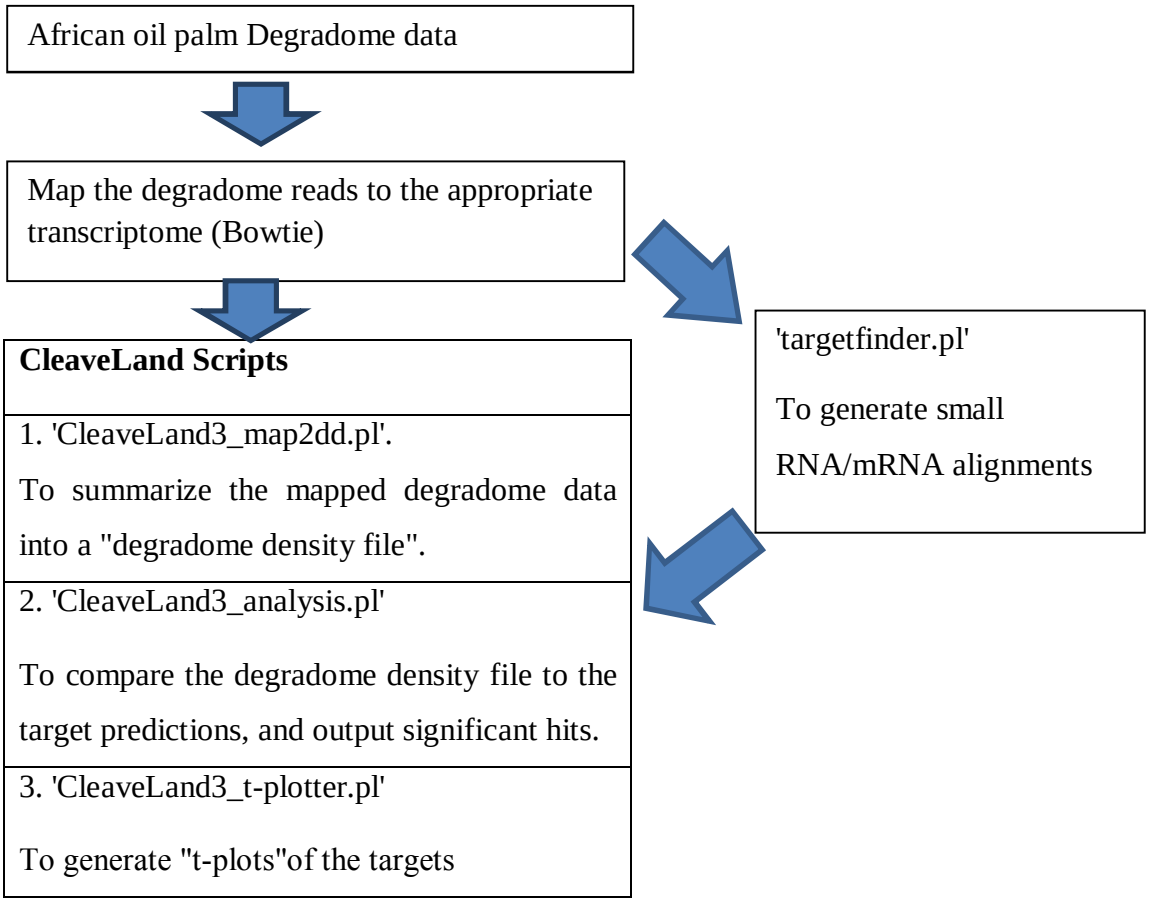


Figure 3.2 Workflow of CleaveLand

The output of CleaveLand (Addo-Quaye *et al.*, 2009) were a list of detected mRNA targets and the corresponding alignments for the small RNA–mRNA pairs, complete information on the degradome profile of each target mRNA and signal-to-noise information. The output files were analysed and the data summarized into Supplementary Table S1.

3.3 The SeqTar pipeline

SeqTar is developed in consideration of the large amount of degradome data for identifying miRNA targets. More mismatches are allowed and the false positive predictions reduced by using two P-values to control the qualities of its predictions (Zheng *et al.*, 2012). The number of mismatches in sRNA complementary site is assigned mismatched p-value based on the shuffled sRNA sequences against randomly chosen target sequences. The reads mapped to the 9-11th nucleotide are named as valid reads. The number of reads accumulated at the central region of the sRNA complementary site, the 9-11th nucleotide from the 5'-end of miRNA is valid p-value given by a Binomial test.

Installation

SeqTar 2010-2011 was provided by courtesy of Dr. Yun Zheng. SeqTar was installed to a system with processor Intel Core i7 and 8GB ram with Linux 64 bit operating system. Java version 1.6.0 was downloaded and installed according to provided manual from http://www.java.com/en/download/help/linux_install.xml. BLAST obtained from ftp://ftp.ncbi.nlm.nih.gov/blast/executables/release/LATEST/.

Workflow

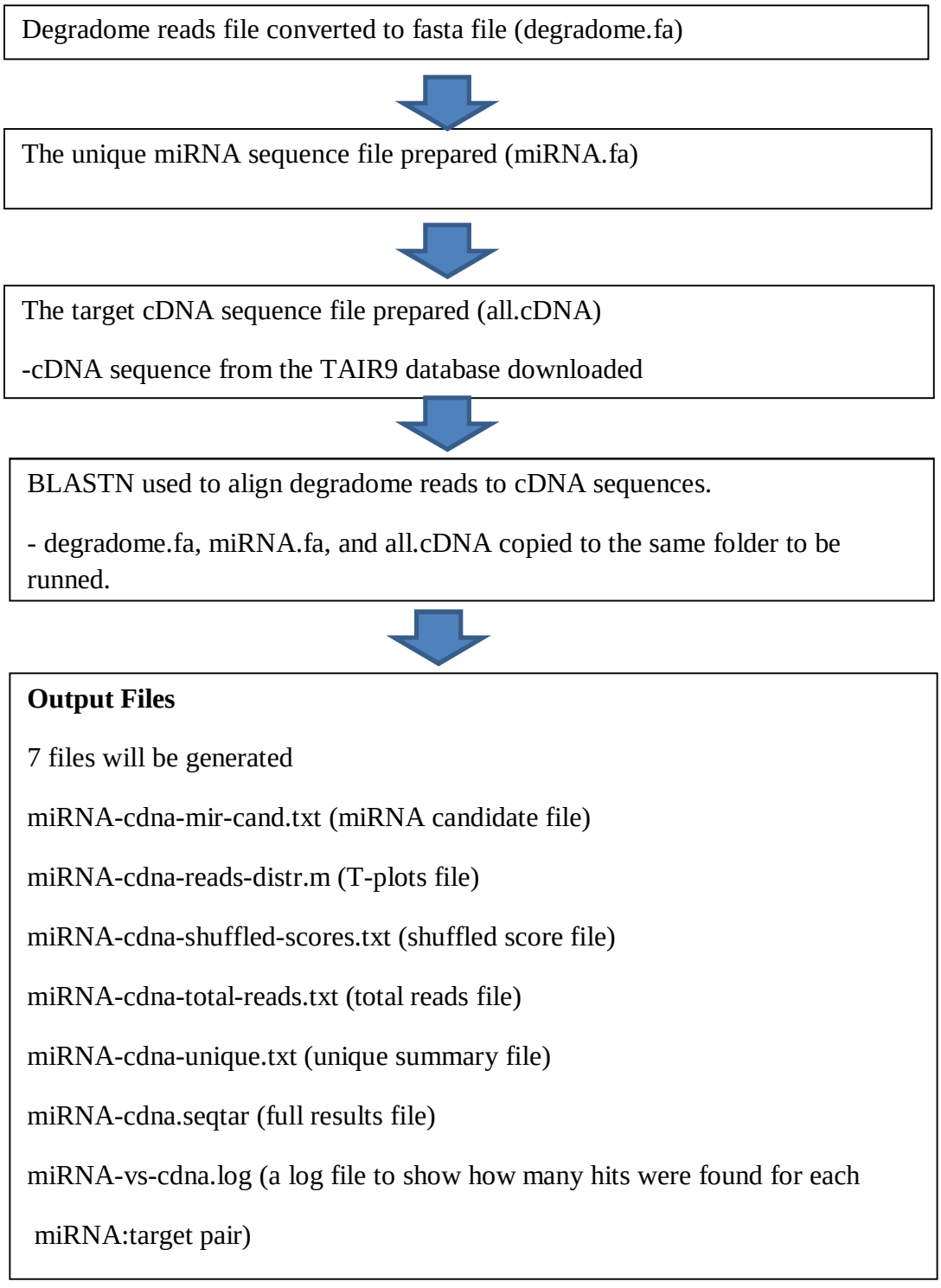


Figure 3.3 Workflow of SeqTar

3.4 PARESnip

PARESnip was installed to a system with Intel Core i7 processor, 8GB ram and windows operating system. sRNA Workbench in GUI mode launched by downloading all the files from zip archive to a new directory followed by launching the sRNAWorkbenchStartup.jar. System was scanned for available memory and a portion of it allocated to the Workbench. PARESnip was launched by clicking Tools->PAREsnip.

The inputs for PAREsnip are mRNA dataset (transcriptome), transcript degradation fragments obtained from a PARE experiment (degradome) and small RNA dataset (sRNAome). All of the inputs should be in FASTA format.

P-value is calculated for each sRNA target reported by PARESnip. The p-value score indicates how likely the reported duplex occurred by chance.

Workflow

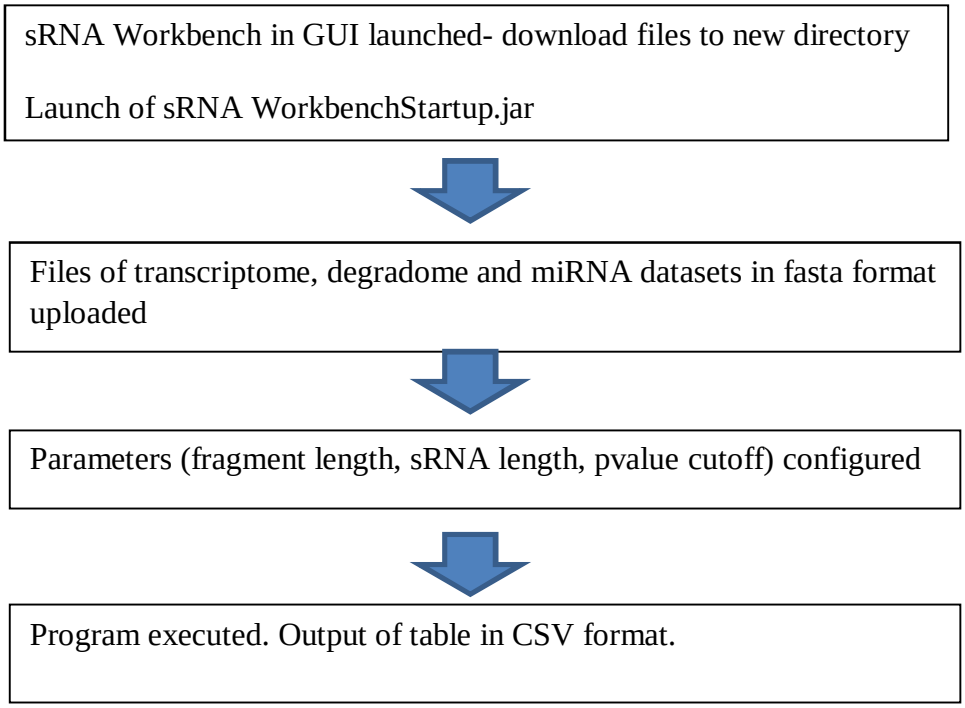


Figure 3.4 Workflow of PARESnip

Parameters used in PARESnip

min_sRNA_abundance: 1	max_fragment_length: 21
subsequences_are_secondary_hits: no	min_sRNA_length: 19
output_secondary_hits_to_file: no	max_sRNA_length: 24
use_weighted_fragments_abundance: yes	allow_single_nt_gap: yes
category_0: yes	allow_mismatch_position_11: yes
category_1: yes	allow_adjacent_mismatches: yes
category_2: yes	max_mismatches: 4.5
category_3: yes	calculate_pvalues: yes
category_4: yes	number_of_shuffles: 100
discard_tr_rna: yes	pvalue_cutoff: 0.05
discard_low_complexity_srnas: yes	do_not_include_if_greater_than_cutoff: yes
discard_low_complexity_candidates: yes	number_of_threads: 7
min_fragment_length: 20	auto_output_tplot_pdf: no

Figure 3.5 Parameters used in PARESnip

The output files which were in CSV format were analysed and saved as Supplementary Table S3.

CHAPTER 4: RESULTS

4.1 Predicted miRNA targets by CleaveLand, SeqTar and PARESnip

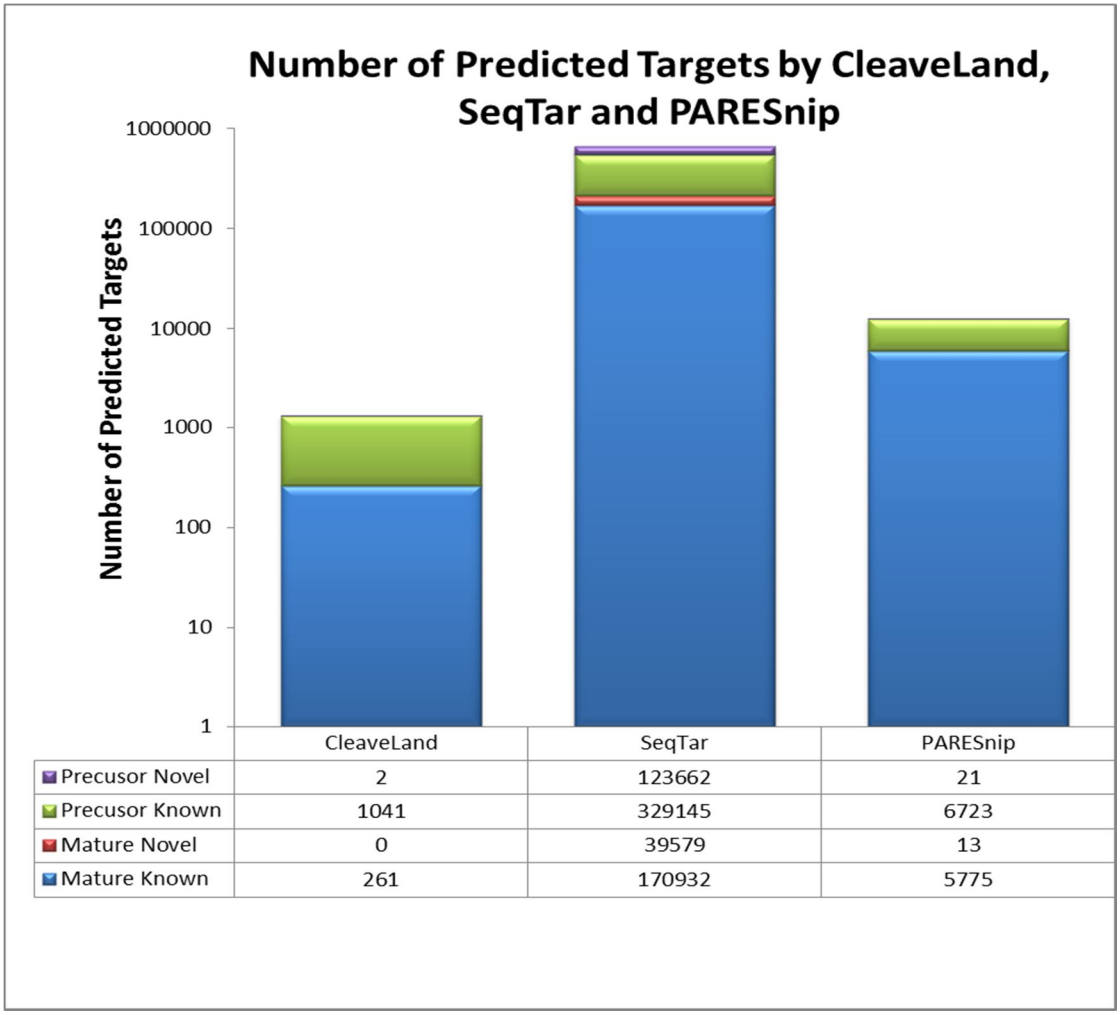
The predicted miRNA target (t0474181_3_miR172: CL1Contig4708) is from mature known miRNA category. The duplex diagram of predicted miRNA target by CleaveLand, SeqTar and PARESnip is shown in Table 4.1.1.

Tool	Predicted miRNA target
CleaveLand	query=t0474181_3_miR172, target=CL1Contig4708, **target 5' CUGCAGCAUCAuCAGGAUUCU 3'** ::::: :::::::.:::::: **query 3' UACGUCAUAGUAGUUCUAAGA 5'**
SeqTar	t0474181_3_miR172 CL1Contig4708 **Query: 3' tACGTCATAGTAGTTCTAAGA 5'** ||||| |||||||:|||||| **Ref.: 5' cUGCAGCAUCAUCAGGAUUCU 3'**
PARESnip	>t0474181_3_miR172 CL1Contig4708 **5' AGAATCTTGATGATACTGCAT 3'** ||||||o||||||| ||||| **3' CCACTCTTAGGACTACTACGACGTCGACACCA 5'**

Table 4.1.1 Example of a predicted miRNA target by CleaveLand, SeqTar and PARESnip

Datasets of non-redundant sequences were used for analysis by each of the tools to avoid repetition of predicted targets.

The number of Predicted targets by CleaveLand, SeqTar and PARESnip represented by a bar chart.

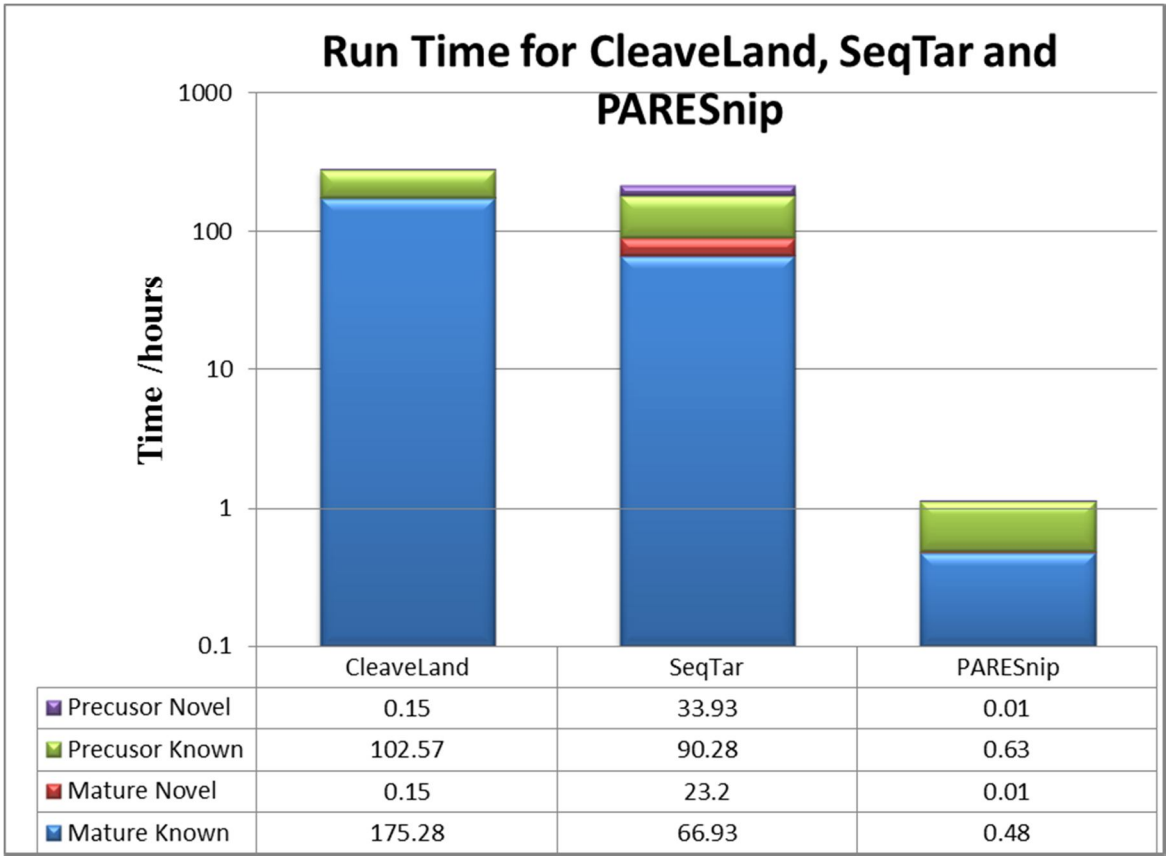


	CleaveLand	SeqTar	PARESnip
Precusor Novel	2	123662	21
Precusor Known	1041	329145	6723
Mature Novel	0	39579	13
Mature Known	261	170932	5775

Figure 4.1.1Number of Predicted targets by CleaveLand, SeqTar and PARESnip

4.2 Run Time of Tools

Run time of CleaveLand, SeqTar, PAREsnip were about 278 hours, 214 hours and 1 hour respectively. Time to complete processing the data by each tool according to miRNA category is represented by a bar chart.



	CleaveLand	SeqTar	PARESnip
Precusor Novel	0.15	33.93	0.01
Precusor Known	102.57	90.28	0.63
Mature Novel	0.15	23.2	0.01
Mature Known	175.28	66.93	0.48

Figure 4.2.1 Run time for CleaveLand, SeqTar and PARESnip

4.3 Prediction Comparisons

Predicted targets by CleaveLand, SeqTar and PARESnip were compared to each other and the intersection of predictions were analysed. Similar prediction of miRNA targets by CleaveLand, SeqTar and PARESnip were listed in Appendix C.

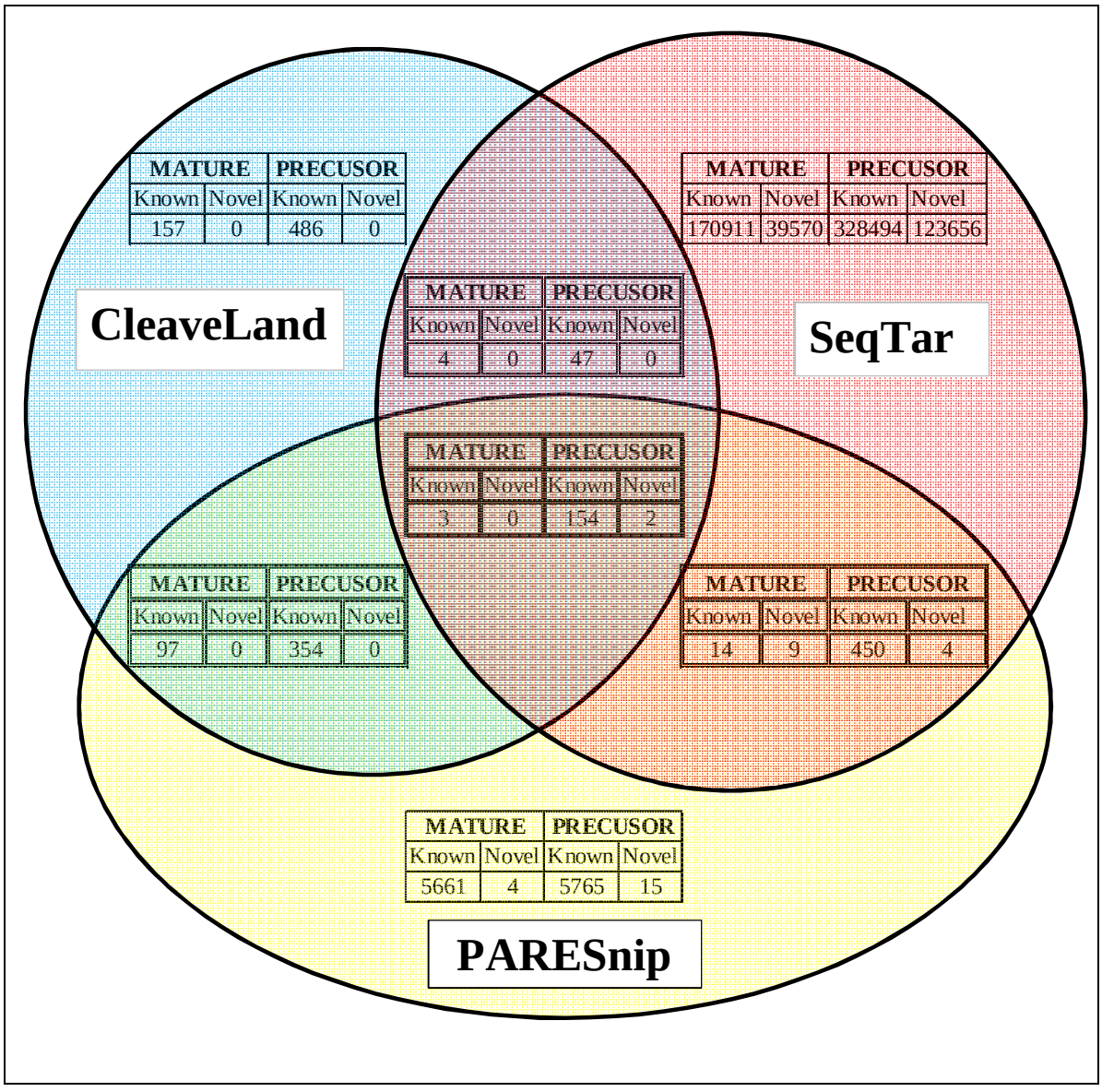


Figure 4.3.1 Venn diagram showing the comparison of miRNA target predictions by CleaveLand, SeqTar and PARESnip. The Venn diagram shows the intersection of predictions made by CleaveLand, SeqTar and PARESnip and is a summary of results within supplementary Tables S1, S2 and S3.

4.4 P-Values

P-value is calculated to assess whether the predicted target likely to have occurred simply through chance. Predicted targets are statistically significant to a designated limit of $p < 0.05$. Smaller p-values, $p < 0.01$ are classified as highly significant (Davies & Crombie, 2009).

P-value in CleaveLand

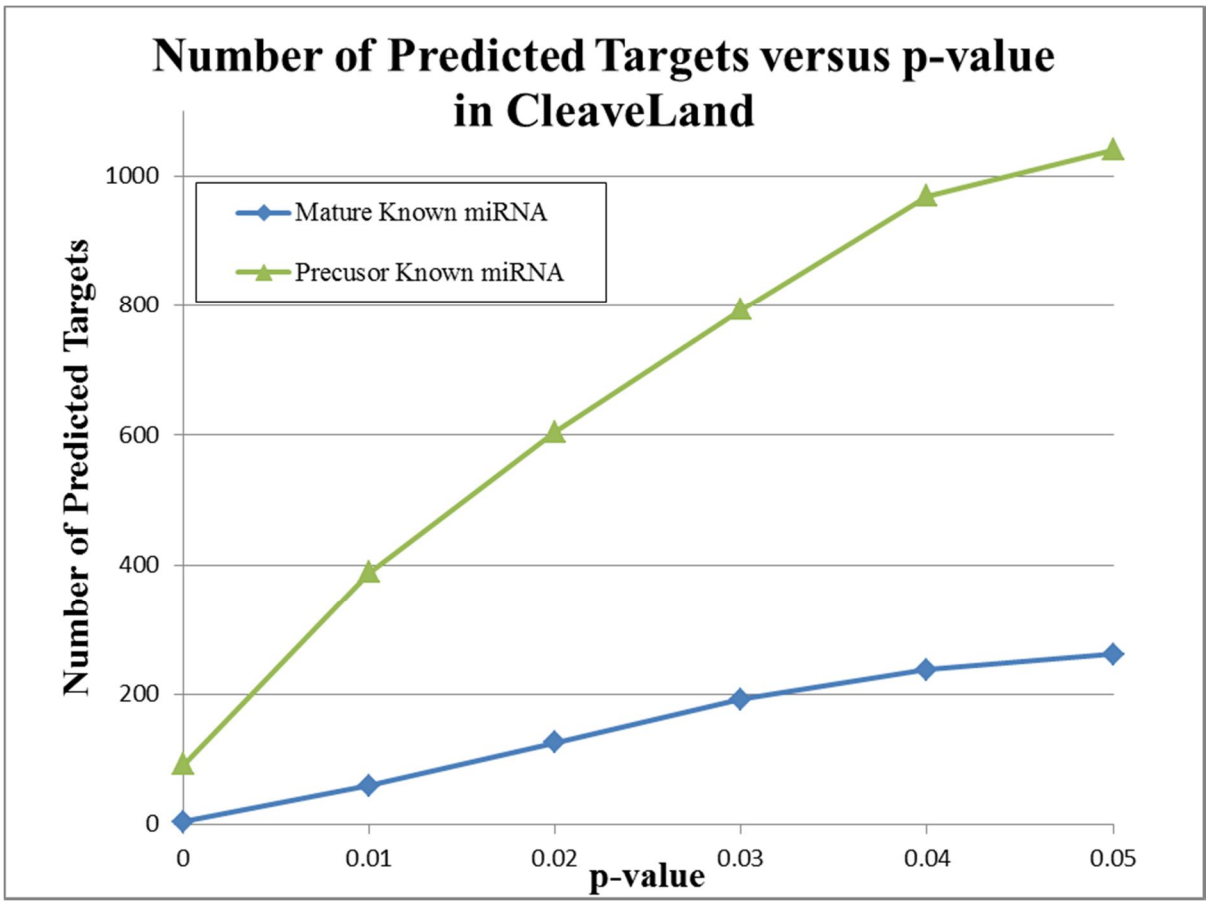


Figure 4.4.1 Cumulative Plot showing number of predicted targets versus p-value reported by CleaveLand

CleaveLand predicted 4 mature known miRNA targets and 92 precusor known miRNA targets with smaller p-values, $p < 0.01$ which are classified as highly significant. A total of 96 known miRNA targets classified as highly significant were predicted by CleaveLand.

P-values in SeqTar

Two p-values, mismatch p-value (P_m) and valid p-value (P_v) were introduced in the methodology to reduce the false positive prediction as more mismatches were allowed.

Mismatch p-value

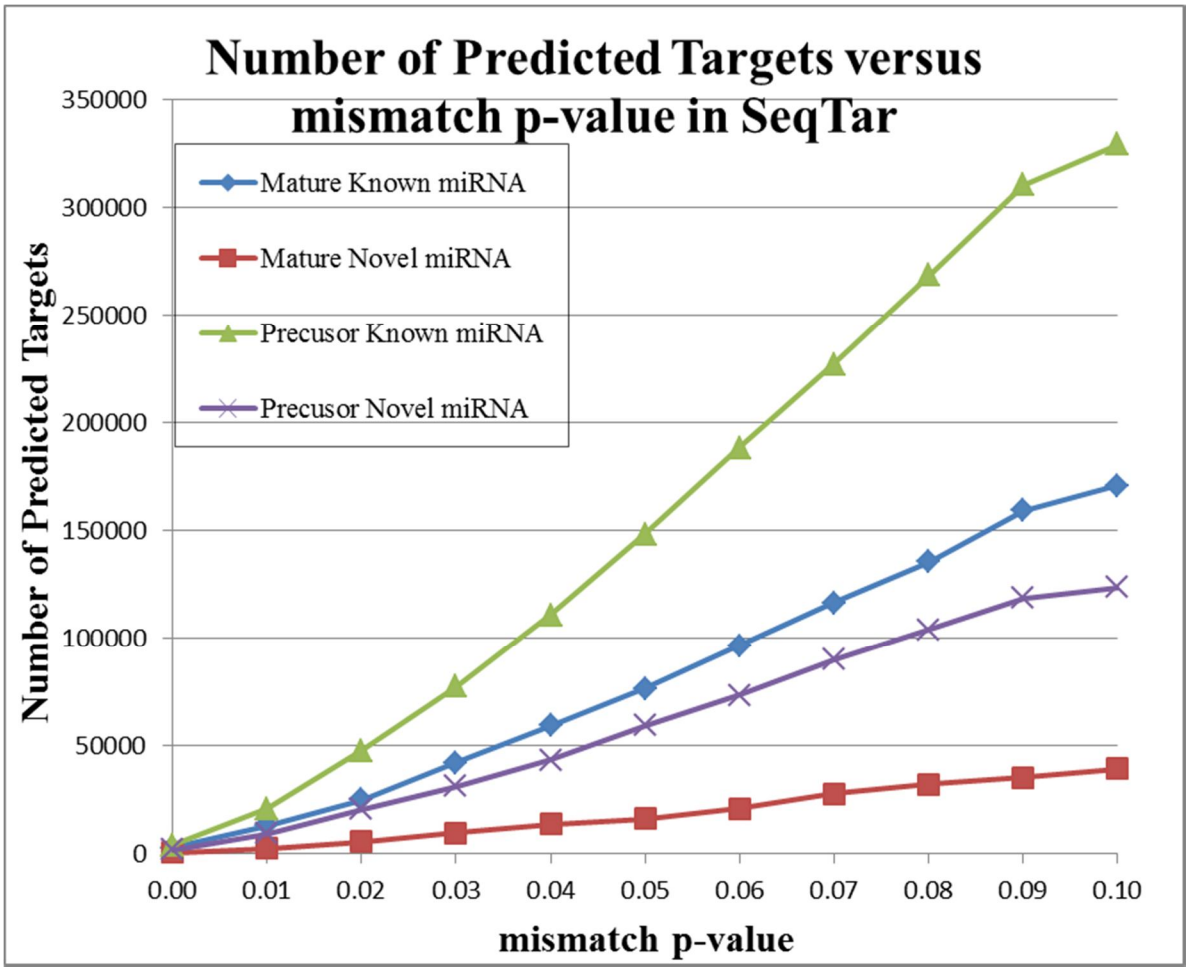


Figure 4.4.2 Cumulative Plot showing number of predicted targets versus mismatch p-value reported by SeqTar

Valid p-value

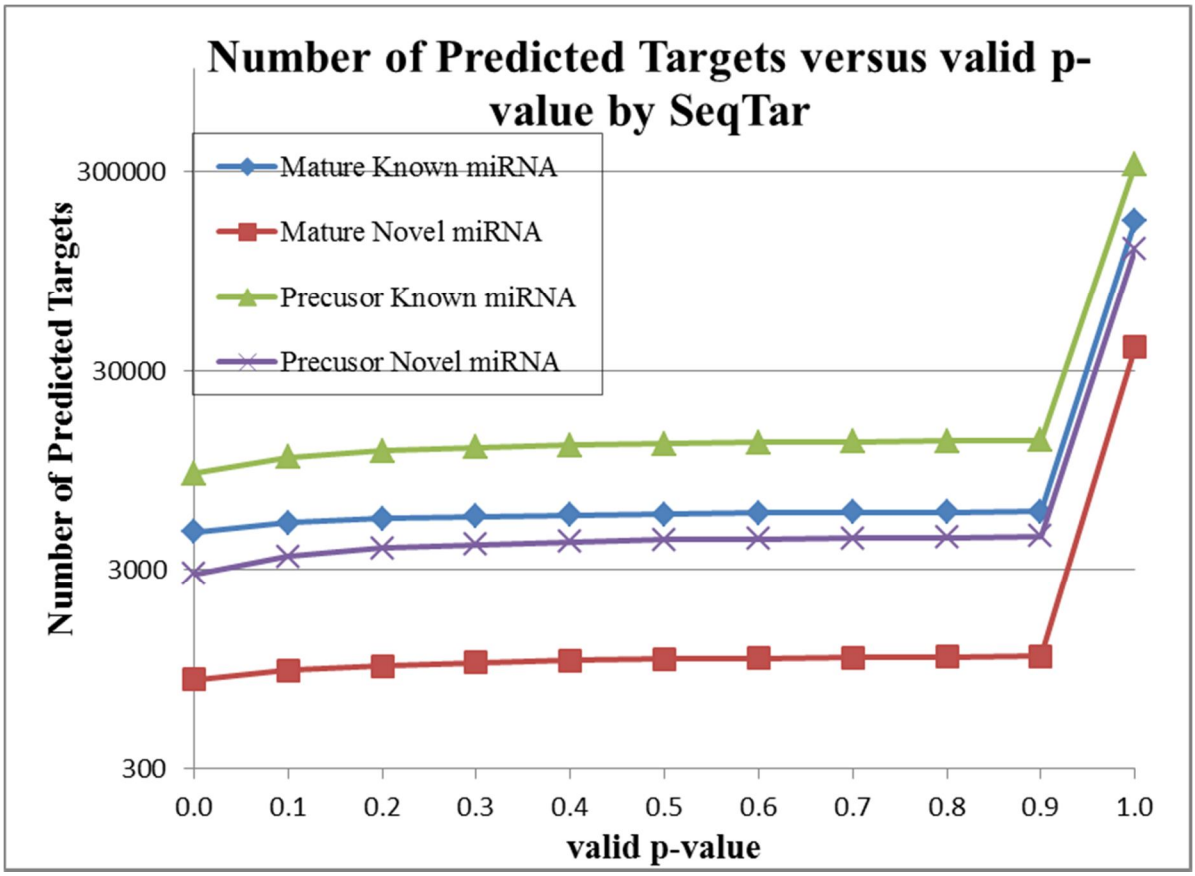


Figure 4.4.3 Cumulative Plot showing number of predicted targets versus valid p-value reported by SeqTar

The predicted miRNA target pairs by SeqTar can be grouped into four categories. A highly significant target falls in Category I with valid p-value, $P_v < 10^{-5}$ and mismatch p-value, $P_m < 0.1$, considering also valid reads, $v \geq 5$ and mismatch, $m \leq 3$. This category is considered reliable because both satisfactory complementary sites and enriched valid reads are present (Zheng *et al.*, 2012).

Predicted miRNA target pairs are classified into Category I($P_m < 0.1$ and $P_v < 10^{-5}$). These numbers of predicted targets were further reduced after considering valid reads, $v \geq 5$ and mismatch, $m \leq 3$.

The classification into Category I and reducing the number of predicted targets with valid reads, $v \geq 5$ and mismatch, $m \leq 3$ is summarized into Figure 4.5.4.

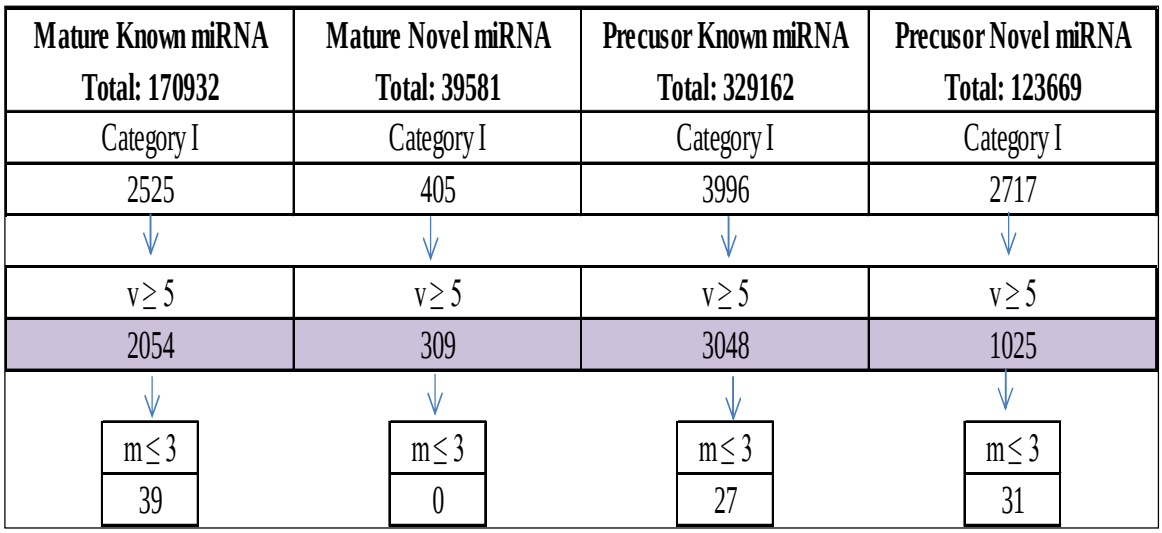


Figure 4.4.4 Category I miRNA target pairs classified by their P_v and P_m-values. Category I miRNA target pairs are listed in APPENDIX D.

P-value in PARESnip

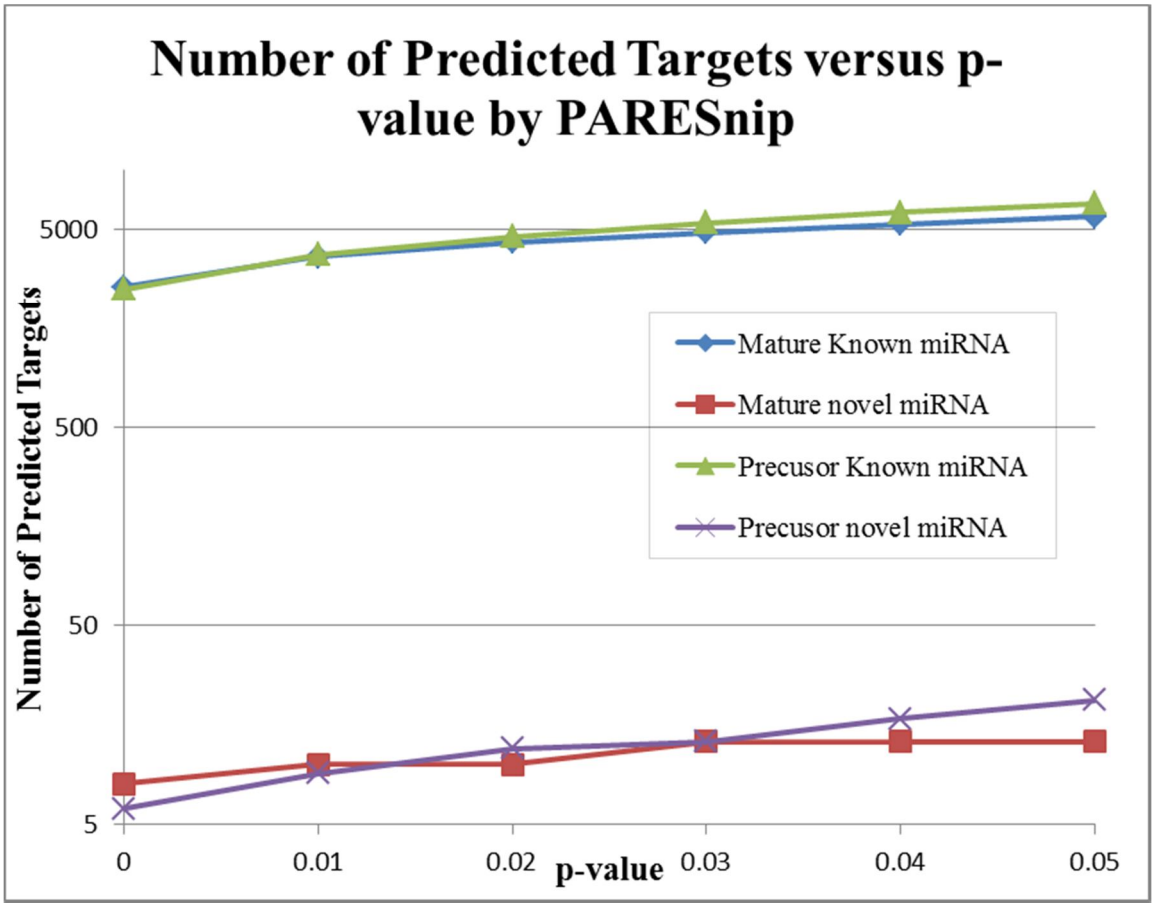


Figure 4.4.5 Cumulative Plot showing number of predicted targets versus p-value reported by PARESnip

PARESnip predicted 2564 mature known miRNA targets, 8 mature novel miRNA targets, 2495 precusor known miRNA targets and 6 precusor novel miRNA targets with smaller p-values, $p < 0.01$ which are classified as highly significant. A total of 5073 highly significant targets consist of 5059 known miRNA targets and 14 novel miRNA targets were predicted by PARESnip.

Number of predicted miRNA targets by CleaveLand, SeqTar and PARESnip which were classified as highly significant according to small p-value performed by Figure 4.4.6.

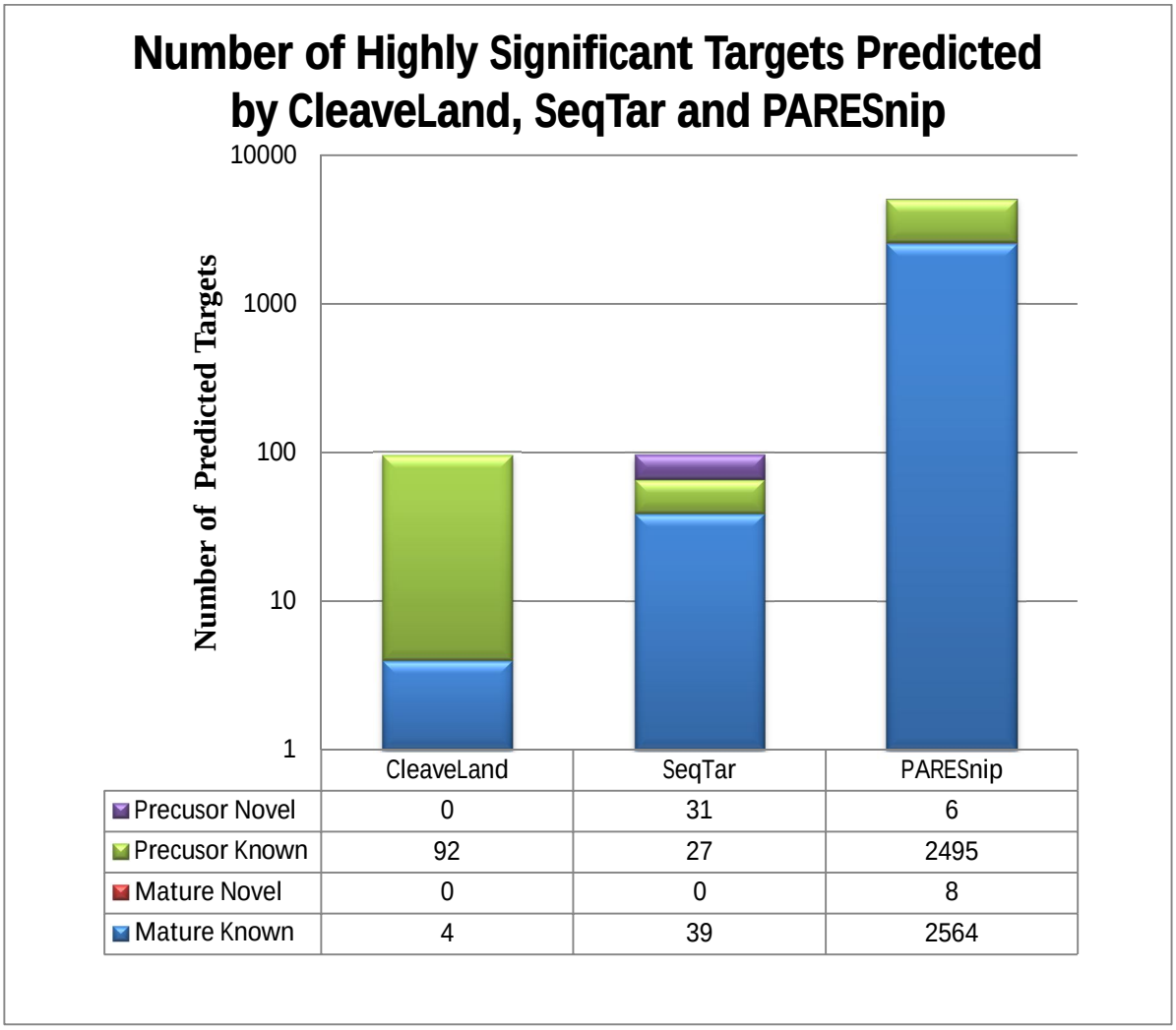


Figure 4.4.6 Number of highly significant miRNA targets according to small p-value predicted by CleaveLand, SeqTar and PARESnip.

CHAPTER 5: DISCUSSION

5.1 Run Time of Tools

PARESnip is the fastest tool to complete the analysis followed by SeqTar and CleaveLand. PARESnip encodes the data into the 4-way tree which enables fast mapping of the sequences. CleaveLand takes the longest time to complete the analysis because of the complexity in mismatch-based scoring scheme which generates randomized sRNA-target alignments. The alignments rules implemented in CleaveLand increases the run time of the tool. Even tough, SeqTar uses a modified algorithm; it is unable to compete with PARESnip in speed.

5.2 Predicted Target Comparisons

According to the Figure 4.3.1 Venn diagram, predicted targets by CleaveLand, SeqTar and PARESnip were found to overlap; different targets were also predicted. Differences in target predictions between CleaveLand and PARESnip occurs because algorithm in PARESnip do not accept a mismatch at position 10, multiple gaps within a duplex and adjacent mismatches within a duplex as accepted by CleaveLand (Folkes *et al*., 2012).

SeqTar predicted total of 500 077 of known miRNAs targets and 163 241 of novel miRNA targets in oil palm degradome data which is the highest number of targets predicted. SeqTar is suitable to be used to predict a wide range of targets without missing any possibilities of the predicted targets.

Figure 4.4.6 shows miRNA targets classified as highly significant using small p-values. PARESnip predicts 5093 highly significant miRNA targets which is the highest prediction among the three tools. Flexibility to customize search in PARESnip allows in

parameter optimization for searching degradome data sets, it is difficult to estimate an accurate false positive rate since the use of binding rules and p-value filtering provides a strong set of sRNA target (Folkes *et al.,* 2012). PARESnip (Folkes *et al.,* 2012) can be used to search more deeply for individual miRNA targets by relaxing the stringency of the binding rule.

According to Figure 4.4.6, SeqTar predicted most numbers of highly significant precursor novel miRNA targets. According to False Discovery Rate of SeqTar's results using Storey and Tibshirani method, FDR values suggested that the results of SeqTar were reliable and had a very low ratio of false positive if both Pm and Pv were set 0.05, or even Pm < 0.1 in all cases and Pv < 0.1 in most cases (Zheng *et al*., 2012). In this study, Category I with Pm < 0.1 and Pv < 10^{-5} falls in the range suggested by FDR values. CleaveLand is not suitable to predict novel miRNA targets because only two novel miRNA targets are predicted and the novel targets did not fall in highly significant targets. CleaveLand can be used to predict known miRNA targets but it is not suitable to predict novel miRNA targets in oil palm degradome data. The criteria adopted in CleaveLand are too stringent which omit many genuine targets and relaxation of SeqTar can identify additional novel targets for miRNAs from the degradomes (Zheng *et al*., 2012).

5.3 User Friendly Tool

CleaveLand and SeqTar are command-line-based applications used in Linux. PARESnip is available as a GUI-based and also in command-line-based for Windows, Linux and MacOS operating system. SeqTar is not available in a publicly downloadable package as Cleaveland and PARESnip. PARESnip is more user friendly, CleaveLand and SeqTar require knowledge in Linux. PARESnip (Folkes *et al.*, 2012) is user-configurable

enables users to make more liberal and stringent searches through customize search parameters and binding rules.

5.4 Validation of Results

The numbers of false positive predictions for each of the tools was not determined since experimental validation is not available for oil palm degradome data at this time. However, experimental validation available for targets from *Arabidopsis* and rice degradome data sets predicted by SeqTar. A total of 12 novel miRNA targets, 6 each in *Arabdidopsis* and rice degradome datasets were experimentally verified from a solid number of novel miRNA targets predicted by SeqTar (Zheng *et al*., 2012).

CHAPTER 6: CONCLUSIONS

Total targets predicted by CleaveLand, SeqTar, PARESnip were 1304, 663 318 and 12 532 respectively. SeqTar is suitable to be used to predict a wide range of targets without missing any possibilities of the predicted target in oil palm degradome data. SeqTar is also suitable to predict precursor novel miRNA targets in oil palm degradome data. CleaveLand can only be used to predict known miRNA targets and it is not suitable to predict novel miRNA targets in oil palm degradome data.

Of the three currently available tools for degradome analysis, PARESnip was found to be the fastest and most convenient tool. PARESnip is user friendly and user configurable tool. PARESnip also shows the highest number of prediction of highly significant miRNA targets compared to CleaveLand and SeqTar. Thus, PARESnip can be considered as a more reliable tool to analyse oil palm degradome data.

The numbers of false positive predictions for each of the tools was not determined since experimental validation is not available for oil palm data at this time. Computational prediction methods have been successfully employed to find candidate miRNA targets in African Oil Palm inflorescence degradome data.