Identification and characterization of microRNAs and their putative target genes in *Anopheles funestus s.s*

Mushal Allam Mohamed Alhaj Ali

South African National Bioinformatics Institute

University of the Western Cape

Supervisor: Prof. Alan Christoffels

A thesis submitted in fulfillment of the the requirements for the degree of Doctor of Philosophy at the South African National Bioinfomatics Institute, University of the Western Cape

March 1, 2013

Dedication

I dedicate this thesis to my loving parents

Mona A. M. Homody and Allam M. A. Ali

who laid the foundation to make this accomplishment possible.

Acknowledgements

I begin in the name of Allah, the Most Compassionate, the Most Merciful. Firstly, I would like to thank God, without whom I would never have come so far in life.

My most fervent appreciation is extended to my supervisor Prof. Alan Christoffels for his guidance, intellectual insight, and overwhelming support. Special thanks go to Prof. Lizette Koekemoer for hosting me in her laboratory, and Dr. Hiba Abdallah for her kind assistance during the laboratory work and all the staff and students in the Vector Control Reference Unit at the National Institute for Communicable Diseases, Johannesburg, South Africa. A heartfelt thanks goes to all SANBI staff and students who have engraved their memory in my heart. A very special thanks goes to my younger brother Mohamed, my extended family and friends, for always being so proud of my achievements.

Lastly, I would like to thank the National Research Foundation, South Africa, for funding during this project.

Abstract

The discovery of microRNAs (miRNAs) is one of the most exciting scientific breakthroughs in the last decade. miRNAs are short RNA molecules that do not encode proteins but instead, regulate gene expression. Over the past several years, thousands of miRNAs have been identified in various insect genomes through cloning and sequencing, and even by computational prediction. However, information concerning possible roles of miRNAs in mosquitoes is limited. Within this context, we report here the first systematic analysis of these tiny RNAs and their target mRNAs in one of the principal African malaria vectors, *Anopheles funestus s.s*.

Firstly, to extend the known repertoire of miRNAs expressed in this insect, the small RNAs from the four developmental stages (egg, larvae, pupae and the adult females), were sequenced using next generation sequencing technology. A total of 98 miRNAs were identified, which included 65 known *Anopheles* miRNAs, 25 miRNAs conserved in other insects and 8 novel miRNAs that had not been reported in any species. We further characterized new variants for *miR-2* and *miR-927* and stem-loop precursors for *miR-286* and *miR-2944*. The analysis showed that many miRNAs have stage-specific expression, and co-transcribed and co-regulated during development.

Secondly, for a better understanding of the molecular details of the miRNAs function, we identified the target genes for the *Anopheles* miRNAs using a novel approach that identifies overlap genes among three target prediction tools followed by filtering genes based on functional enrichment of GO terms and KEGG pathways. We found that most of the miRNAs are metabolic regulators. Moreover, the results suggest implication of some miRNAs not only in the development but also in insect-parasite interaction.

Finally, we developed the InsecTar database (http://insectar.sanbi.ac.za) for miRNA targets in the three mosquito species; *Anopheles gambiae*, *Aedes aegypti*, and *Culex quinquefasciatus*, which incorporates prediction and the functional analysis of these target genes. The proposed database will undoubtedly assist to explore the roles of these regulatory molecules in insects.

This type of analysis is a key step towards improving our understanding of the complexity and regulation mode of miRNAs in mosquitoes. Moreover, this study opens the door for exploration of miRNA in regulation of critical physiological functions specific to vector arthropods which may lead to novel approaches to combat mosquito-borne infectious diseases.

Keywords

MicroRNA

Non-Coding RNA

MicroRNA Target

InsecTar

Database

Anopheles funestus

Anopheles gambiae

Culex quinquefasciatus

Aedes aegypti

Mosquitoes

Vectors

Insect

Malaria

Publications

Mushal Allam, Hiba Abdallah, Lizette K Koekemoer, and Alan Christoffels. **Identification and characterization of microRNAs expressed in malaria vector *Anopheles funestus* developmental stages by high throughput sequencing**. *In preparation for submission*.

Mushal Allam, Saleem Adam, Peter Van Heusden, Musa Nur Gabere, and Alan Christoffels. **InsecTar: a database for microRNA target genes in insects**. *In preparation for submission*.

Declaration

I declare that **Identification and characterization of microRNAs and their putative target genes in *Anopheles funestus s.s*** is my own work, that it has not been submitted for any degree or examination in any other university, and that all the resources I have or quoted have been indicated and acknowledged by complete references.

Mushal A. M. A. Ali March 1, 2013

Signed

_____

Abbreviations

A	Adenine
C	Cytosine
G	Guanine
T	Thymine
U	Uracil
DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
ncRNA	non-coding RNA
miRNA	microRNA
mRNA	messenger RNA
rRNA	ribosomal RNA
tRNA	transfer RNA
snRNAs	small nuclear RNAs
snoRNAs	small nucleolar RNAs
mir	miRNA gene
cDNA	complementary DNA
RISC	RNA-induced silencing complex
3′	three prime end
5′	five prime end
bp	base pair
3′UTR	3′ untranslated region
5′UTR	5′ untranslated region
IAP	Inhibitor of Apoptosis Proteins
MAPK	Mitogen-Activated Protein Kinase
An. funestus	*Anopheles funestus*
An. funestus s.s	*Anopheles funestus sensu stricto*
An. gambiae	*Anopheles gambiae*
An. gambiae s.s	*Anopheles gambiae sensu stricto*
An. stephensi	*Anopheles stephensi*
An. arabiensis	*Anopheles arabiensis*
An. merus	*Anopheles merus*
An. melas	*Anopheles melas*
An. bwambae	*Anopheles bwambae*
An. quadriannulatus	*Anopheles quadriannulatus*
An. vaneedeni	*Anopheles vaneedeni*
An. parensis	*Anopheles parensis*
An. aruni	*Anopheles aruni*
An. leesoni	*Anopheles leesoni*
An. confusus	*Anopheles confusus*
An. rivulorum	*Anopheles rivulorum*
An. brucei	*Anopheles brucei*

An. fuscivenosus	*Anopheles fuscivenosus*
Ae. aegypti	*Aedes aegypti*
Ae. albopictus	*Aedes albopictus*
C. quinquefasciatus	*Culex quinquefasciatus*
D. melanogaster	*Drosophila melanogaster*
D. pseudoobscura	*Drosophila pseudoobscura*
Ap. mellifera	*Apis mellifera*
B. mori	*Bombyx mori*
C. elegans	*Caenorhabditis elegans*
P. falciparum	*Plasodium falciparum*
P. vivax	*Plasodium vivax*
P. ovale	*Plasodium ovale*
P. malariae	*Plasodium malariae*
P. knowlesi	*Plasodium knowlesi*
GO	Gene Ontology
KEGG	Kyoto Encyclopedia of Genes and Genomes
WHO	World Health Organization
CDC	Centers for Disease Control and Prevention
kcal/mol	kilocalorie per mole
EST	Expressed Sequence Tag
PCR	Polymerase Chain Reaction
HMM	Hidden Markov Model
SVM	Support Vector Machine
pSILAC	pulsed SILAC analysis
NGS	Next-generation sequencing
rpm	read per million
SQL	Structured Query Language
API	Application Programme Interface
DBI	Database Interface Module
EVD	Extreme Value Distribution
CGI	Common Gateway Interface
miRBase	The miRNA database
UCSC	University of California Santa Cruz
NICD	National Institute for Communicable Diseases
SANBI	South African National Bioinformatics Institute
UWC	University of the Western Cape

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Literature Review

1.1 Malaria

1.1.1 The disease

Malaria is a deadly mosquito-borne disease that affects millions of people each year in Africa and around the world. According to the World Health Organization (WHO), there were 216 million cases of malaria and an estimated 655,000 deaths in 2010 (WHO, 2011). Most deaths occur among pregnant women and children living in Africa where a child dies every minute. In total, the disease accounts for approximately 22% of all childhood deaths. There are four parasite species that associated with human malaria: *Plasmodium falciparum*, *P. vivax*, *P. ovale*, and *P. malariae*. Malaria due to *P. falciparum* is the most deadly, and it predominates in Africa. *P. vivax* is less dangerous but more widespread, while the remaining two species occur less frequently. Recently human malaria cases caused by *P. knowlesi*, a species that causes malaria among monkeys and occurs in certain forested areas of South-East Asia (Jongwutiwes et al., 2004; Cox-Singh et al., 2008).

1.1.2 The malaria parasites

Briefly, the malaria parasite exhibits a life cycle with typical apicomplexan features (Figure 1.1). There are three distinct invasive stages: sporozoite, merozoite, and ookinete. All are characterized by apical organelles and can invade or pass through host cells. Two distinct types of merogony are observed. The first, called exoerythrocytic schizogony, occurs in the liver and is initiated by the sporozoite. The resulting merozoites then invade erythrocytes and undergo repeated rounds of merogony called erythrocytic schizogony. Some of the merozoites produced from the erythro-

cytic schizogony will undergo gamogony. *Plasmodium* gamogony is described in two phases: gametocytogenesis occuring in the bloodstream of the vertebrate host, and gametogenesis taking place in the mosquito gut. The gametes fuse to become a zygote which first develops into an ookinete and then becomes an oocyst where sporogony takes place (Gilles and Warrell, 1993).

1.1.3 The malaria vectors

1.1.3.1 General information

Malaria is transmitted to humans by a bite of an infected female mosquito belonging to one of 30 anopheline species (WHO, 2011). Female mosquitoes require blood meals to complete egg production. The successful development of the malaria parasite in the mosquito (from the gametocyte stage to the sporozoite stage) depends on several factors. The most important factors are ambient temperature and humidity (higher temperatures accelerate the parasite growth in the mosquito) and whether the mosquito survives long enough to allow the parasite to complete its cycle in the mosquito host. In contrast to the human host, the mosquito host does not suffer noticeably from the presence of the parasites (CDC, 2010).

1.1.3.2 Geographic distribution

There are approximately 4500 species of mosquitoes grouped into 41 genera in the family *Culicidae*, order *Diptera*, class *Insecta* and phylum *Arthropoda*. Of the approximately 430 *Anopheles* species, only 30-40 transmit malaria. Anophelines are found worldwide except in Antarctica (Figure 1.2). Malaria is transmitted by different *Anopheles* species, depending on the region and environment. Anophelines that

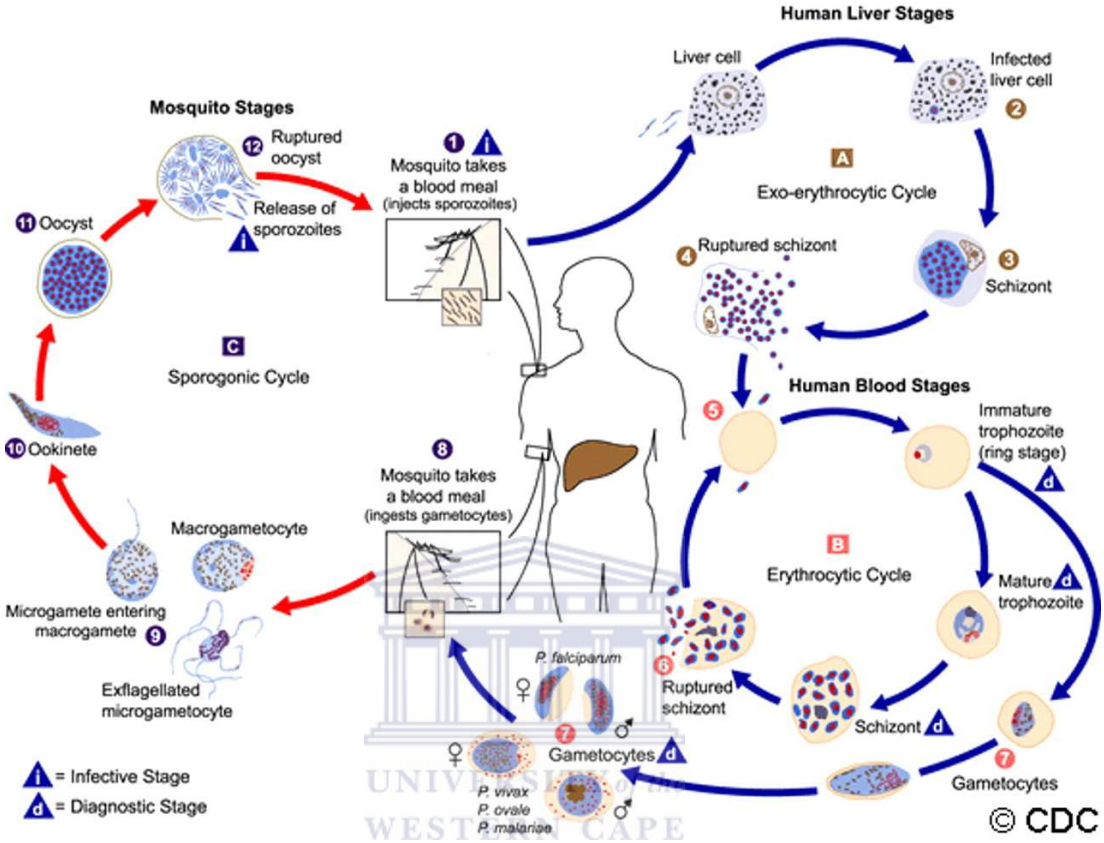


Figure 1.1: Malaria parasites life cycle. During a blood meal, a malaria-infected female *Anopheles* mosquito inoculates sporozoites into the human host (1). Sporozoites infect liver cells (2) and mature into schizonts (3), which rupture and release merozoites (4). (Of note, in *P. vivax* and *P. ovale* a dormant stage (hypnozoites) can persist in the liver and cause relapses by invading the bloodstream weeks, or even years later.) After this initial replication in the liver (exo-erythrocytic schizogony (A)), the parasites undergo asexual multiplication in the erythrocytes (erythrocytic schizogony (B)). Merozoites infect red blood cells (5). The ring stage trophozoites mature into schizonts, which rupture releasing merozoites (6). Some parasites differentiate into sexual erythrocytic stages (gametocytes) (7). The gametocytes, male (microgametocytes) and female (macrogametocytes), are ingested by an *Anopheles* mosquito during a blood meal (8). The parasites multiplication in the mosquito is known as the sporogonic cycle (C). While in the mosquito's stomach, the microgametes penetrate the macrogametes generating zygotes (9). The zygotes in turn become motile and elongated (ookinetes) (10) which invade the midgut wall of the mosquito where they develop into oocysts (11). The oocysts grow, rupture, and release sporozoites (12), which make their way to the mosquito's salivary glands. Inoculation of the sporozoites (1) into a new human host perpetuates the malaria life cycle (CDC, 2010).

can transmit malaria are found not only in malaria-endemic areas, but also in areas where malaria has been eliminated. The latter areas are thus constantly at risk of re-introduction of the disease (Kiszewski et al., 2004).

1.1.3.3 Life stages

Like all mosquitoes, anophelines go through four stages in their life cycle: the egg, the larva, the pupa, and the adult (Figure 1.3). The first three stages are aquatic and last 5-14 days, depending on the species and the ambient temperature. Adult females can live up to a month or more in captivity, but most do not live more than 1-2 weeks in nature (WHO, 2005; CDC, 2010).

1.1.3.3.1 Eggs

Adult mosquito females lay between 50-200 eggs per oviposition. Eggs with floats on either side, are laid singly directly on water. Eggs hatch within 2-3 days, however it may take up to 2-3 weeks in colder climates.

1.1.3.3.2 Larvae

Mosquito larvae have a well-developed head with mouth brushes used for feeding, a large thorax, and a segmented abdomen. They have no legs. In contrast to other mosquitoes, anopheline larvae lack a respiratory siphon and therefore position parallel to the surface of the water. Larvae breathe through spiracles located on the 8th abdominal segment and therefore must come to the surface frequently. The larvae spend most of their time feeding on algae, bacteria, and other micro-organisms in the surface micro-layer. They dive below the surface only when disturbed. Larvae swim either by jerky movements of the entire body or through propulsion with their mouth brushes.

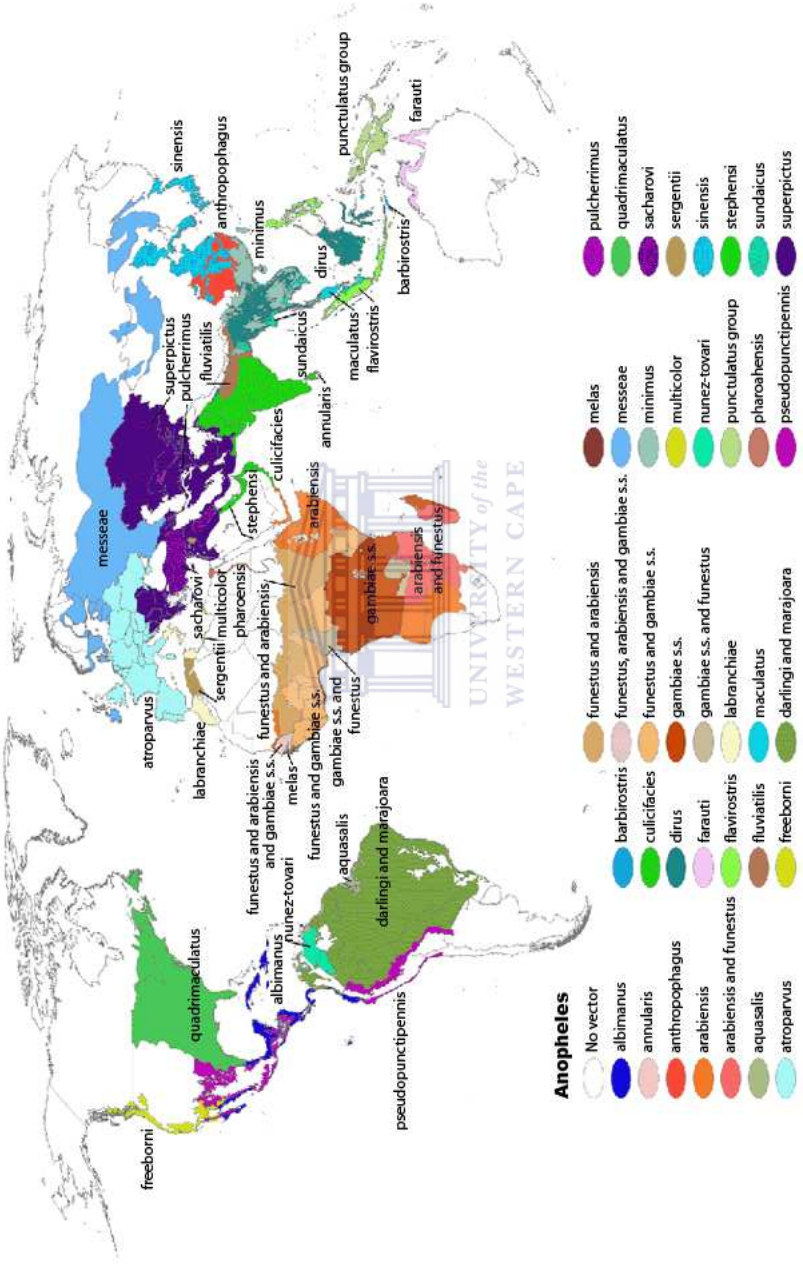


Figure 1.2: Global distribution of malaria vectors. Dominant malaria vectors were designated in each endemic or potentially endemic region. The 260 regions that identified are infested by a total of 34 dominant vector *Anopheles* (Kiszewski et al., 2004).

Larvae develop through four stages, or instars, after which they metamorphose into pupae. At the end of each instar, the larvae molt, shedding their exoskeleton to allow for further growth. Larvae occur in a wide range of habitats but most species prefer clean, unpolluted water. Larvae of *Anopheles* mosquitoes have been found in fresh or salt water marshes, mangrove swamps, rice fields, grassy ditches, the edges of streams and rivers, and small temporary rain pools.

1.1.3.3.3 Pupae

The pupa is comma-shaped when viewed from the side. The head and thorax are merged into a cephalothorax with the abdomen curving around underneath. As with the larvae, the pupae frequently come to the surface to breathe. After a few days as a pupa, the dorsal surface of the cephalothorax splits and the adult mosquito emerges. The duration from egg to adult varies considerably among species and is strongly influenced by ambient temperature. Mosquitoes can develop from egg to adult in as little as five days but usually take 10-14 days in tropical conditions.

1.1.3.3.4 Adults

Like all mosquitoes, adult anophelines have slender bodies with 3 sections: the head, the thorax and the abdomen. The head is modified to acquire sensory information and for feeding. The head contains the eyes and a pair of long many-segmented antennae. The antennae are important for detecting host odors as well as odors of breeding sites. The head also has an elongate forward-projecting proboscis used for feeding as well as sensory palps. The thorax is designed for locomotion. Three pairs of legs and a pair of wings are attached to the thorax. The abdomen is adapted to digest food, and for the development of eggs. This segmented body part expands considerably when a female

takes a blood meal. The blood is digested over time, serving as a source of protein for the production of eggs which gradually fill the abdomen.

Anopheles mosquitoes are distinguished from other mosquitoes by the palps, which are as long as the proboscis, and by the presence of discrete blocks of black and white scales on their wings. Adult *Anopheles* can also be identified by their typical resting position: males and females rest with their abdomens pointed upwards rather than parallel to the surface on which they are resting. Adult mosquitoes usually mate within a few days after emerging from the pupal stage. In most species of *Anopheles*, the males form large swarms, usually around dusk, and the females fly into the swarms to mate. Males live for about a week, feeding on nectar and other sources of sugar. Females also feed on sugar sources for energy but usually require a blood meal for the development of eggs. After obtaining a full blood meal, the female rests for a few days while the blood is digested and the eggs are developed. This process depends on the temperature, but in tropical conditions usually takes 2-3 days. Once the eggs are fully developed, the female lays them and resumes host-seeking. The cycle repeats itself until the female dies. Females can survive up to a month, however their chances of survival depend on the temperature and their ability to successfully obtain a blood meal while avoiding host defenses (WHO, 2005; CDC, 2010).

1.1.3.4 The African vectors

The principal malaria vector species in sub-Saharan Africa are *An. gambiae sensu stricto* (*An. gambiae s.s*) and *An. funestus sensu stricto* (*An. funestus s.s*) (Gillies and De Meillon, 1968, 1987; Coetzee et al., 2000; Coetzee and Fontenille, 2004; Hay et al., 2005), which belong to the same subgenus *Cellia* and diverged from a common ancestor approximately five million years ago (Sharakhov et al., 2002). Indeed, despite

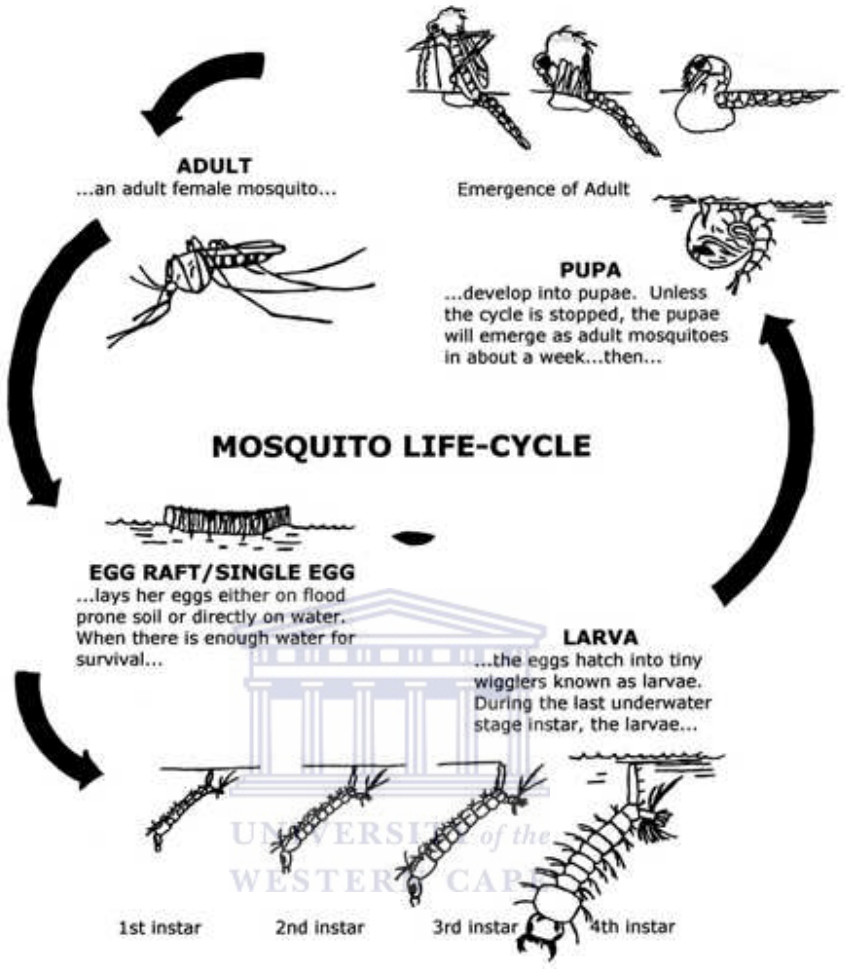


Figure 1.3: The vector life cycle. Females lay eggs on, or near, standing water bodies. Under suitable conditions, the eggs hatch within a few days into tiny larvae. Most air-breathing larvae typically remain just below the surface with the spiracle at the end of the syphon open to the air (there are some species which obtain air by boring into the interstitial cavities of emergent plant stems). The larvae undergo four larval stages averaging ten days in total, depending on the water temperature and the availability of resources. The final instar is usually the longest, lasting several days under normal conditions, after which pupation occurs. The comma-shaped pupa swims efficiently and breathes air, but unlike the larva, it is unable to feed. After several days, the adult mosquito emerges from the pupal case, but due to differences in tissue and organ development, the male mosquitoes usually develop faster than the females and are able to emerge sooner. Mating usually follows soon after the first flight and most females then immediately begin the search for a blood meal. After ingestion of sufficient protein, the eggs develop in the female and she searches for a suitable egg-laying site. Most female mosquitoes have a life span from 2-3 weeks. Although it is commonly thought that males die soon after mating, they too may live for many weeks. source: http://www.wumcd.org.

differences in morphology, breeding site preferences, mating behavior, and relative seasonal abundance, both species coexist geographically in many parts of sub-Saharan Africa and both are highly anthropophilic and endophilic (Gillies and De Meillon, 1968, 1987). In many places in Africa, *An. funestus* is the major vector responsible for malaria transmission and sometimes for malaria epidemics (Fontenille et al., 1990; Hargreaves et al., 2003). Parasite rates of 22% (De Meillon, 1933) and 27% (Swellengrebel et al., 1931) have been recorded in South Africa and more recently 11% in Tanzania (Shiff et al., 1995). In West Africa, rates of 2.6% and 3.3% were observed in Senegal (Fontenille et al., 1997; Dia et al., 2003), between 2.8% and 14.6% in Burkina Faso (Costantini et al., 1999) and around 5% in Cameroon (Antonio-Nkondjio et al., 2002). In 1991, a study in Burkina Faso, recorded up to 50% of *An. funestus* in one village positive for *P. falciparum* circumsporozoite protein (Costantini et al., 1999). Despite its obvious importance as a vector, *An. funestus* has been neglected for almost half a century, with most of the research focusing on members of the *An. gambiae* complex. This has largely been due to the adaptability of the *An. gambiae* complex to laboratory conditions and the ease with which species in the group can be colonized (Coetzee and Fontenille, 2004).

1.1.3.4.1 *Anopheles gambiae* complex

This complex consists of seven species of mosquitoes that are morphologically similar but differ in behaviour, feeding preferences, and breeding requirements (Gillies and De Meillon, 1968, 1987; Hunt et al., 1998). The species complex consists of the major vectors (*An. gambiae s.s* and *An. arabiensis*), the minor vectors (*An. merus*, *An. melas*, and *An. bwambae*), and the non-vector (*An. quadriannulatus*) (White, 1972; Hunt et al., 1998).

1.1.3.4.2 *Anopheles funestus* group

This group comprise nine species that are morphologically alike in the adult stage. Four species, *An. funestus*, *An. vaneedeni*, *An. parensis*, and *An. aruni*, have identical morphology at all life stages and are known as the *Funestus* sub-group (Gillies and De Meillon, 1968, 1987). Of the other species in the group, *An. leesoni* is the most distinct at both egg and larval stage, while *An. confusus* is easily identified on larval characteristics. *An. rivulorum* and *An. brucei* also have distinctive larvae although these two species are virtually indistinguishable from each other. The ninth species, *An. fuscivenosus*, is known only from the adult stage and by chromosomal banding arrangements that distinguish it from the other members of the group (Gillies and De Meillon, 1968; Green, 1982). The morphological identification of members of the *An. funestus* group entails obtaining egg batches from wild females and rearing the progeny through to adults, so that fourth instar larvae and adults can be used in the identification process (Gillies and Coetzee, 1987). Laboratory rearing of larvae is both difficult and time-consuming, taking 4 weeks or more to obtain the necessary specimens. The biology and vectorial capacity of the members of the *An. funestus* group are very different. Apart from *An. funestus*, which is highly anthropophilic, the rest of the group are mainly zoophilic.

An. funestus is a very efficient vector of human *Plasmodium* throughout its distribution. Historical evidence show that in order to conduct an efficient vector control program, precise identification of species is necessary to avoid misidentification of non-vector (Hargreaves et al., 2000, 2003). In South Africa and Tanzania, indoor spraying was implemented to eliminate *An. funestus* (Mouatcho et al., 2007; Shiff et al., 1995). However, some specimens remained, suggesting the failure of the control program. Subsequent careful identification revealed that these mosquitoes were in fact

An. parensis, *An. rivulorum* or *An. vaneedeni*, none of which are involved to any great extent in the transmission of human *Plasmodium* (Gillies and De Meillon, 1968; De Meillon et al., 1977). More recently, *An. parensis* was the most common member of the *An. funestus* group found resting inside human dwellings in a village in Kenya, but was not implicated in malaria transmission (Kamau et al., 2003).

1.1.3.5 Vector-Parasite interactions

In the mosquito, the *Plasmodium* parasite has to go through a series of complex developmental transitions within the vector before it can be transmitted to the human host (Dong et al., 2009). The major bottleneck for parasite development occurs during ookinete invasion of the midgut epithelium, where the parasite is attacked by the mosquito's innate immune system (Vlachou et al., 2005). Tens of thousands of gametocytes can be ingested into the mosquito blood meal, but normally just 50-100 ookinetes are produced: of these, typically fewer than five survive to produce oocysts on the midgut wall (Sinden and Billingsley, 2001). Similarly, of the 50,000 sporozoites produced in these oocysts, only 15-80 might be inoculated into a host by the bite of each infectious mosquito (Sinden and Billingsley, 2001). These reductions largely result from difficulties experienced by the parasite when invading and surviving within appropriate target tissues, indicating an existance of mosquito factors that regulate the development of the parasite (Sinden and Billingsley, 2001).

A class of small non-coding RNAs (ncRNA) molecules, known as microRNAs (miRNAs), was recently altered during *Anopheles* mosquito response to malaria parasites infection, suggesting that miRNAs are the likely candidates for serving as regulators of defence responses to parasites, either by being part of the sensing network detecting the presence of the parasite or by modulating expression levels of defence genes

(Winter et al., 2007; Skalsky et al., 2010).

1.2 MicroRNAs

Ribonucleic acid or RNA molecules display considerable functional diversity, ranging from genetic data storage to gene regulation (Khorana, 1965; Schweet and Heintz, 1966; Geiduschek and Haselkorn, 1969; Rich and RajBhandary, 1976). In addition to the three major classes of RNAs (messenger RNA (mRNA), ribosomal RNA (rRNA) and transfer RNA (tRNA)) responsible for information flow from DNA to protein, there are many small ncRNAs expressed in eukaryotic cells. These types of RNA play important catalytic, structural, and regulatory roles in the cell (Gesteland et al., 2006). These ncRNAs include small nuclear RNAs (snRNAs) that are involved in a variety of important processes such as RNA splicing, regulation of transcription factors and maintaining the telomeres, small nucleolar RNAs (snoRNAs) that play an essential role in RNA biogenesis and guide chemical modifications of rRNAs and other RNA genes, and many other RNAs involved in dosage compensation, imprinting, modulating RNA polymerase activity, and stress responses (Ma et al., 2003). The last few years have seen a continuous stream of novel RNA genes being reported, the most striking of which were miRNAs, which regulate essential cellular events (Brennecke et al., 2003).

1.2.1 Discovery of microRNAs

The first miRNA was reported in 1993 as the result of an effort to clone the *lin-4* gene. This gene caused developmental timing defects in *Caenorhabditis elegans* (*C. elegans*) worms when it mutated (Lee et al., 1993). More than three years of hard work established that *lin-4* dose not encode a protein but rather 21 RNA nucleotides (Lee

et al., 2004). Fortunately, a target of *lin-4*, *lin-14*, was already known. Researchers soon realized that the 3′ untranslated region (3′UTR) of *lin-14* mRNA contained seven sequence elements partially complementary to *lin-4*, giving credence to the notion that the small RNA was functional (Lee et al., 1993; Wightman et al., 1993; Ruvkun et al., 2004). RNA:RNA hybridization showed that the *lin-4* ncRNA inhibited *lin-14* mRNA translation (Wightman et al., 1993). In 2000, almost seven years after the identification of *lin-4*, a second miRNA, *let-7*, was discovered again from *C. elegans* (Pasquinelli et al., 2000; Reinhart et al., 2000). Unlike *lin-4*, *let-7* homologs were readily identified in the genomes of other organisms, including mammals. This suggested that these small RNAs were not developmental oddities in worms. Since then, this discovery triggered a revolution in the research of a new class small ncRNAs, called miRNAs.

1.2.2 Biogenesis of microRNAs

The genes that encode the miRNAs are called *mir* genes, and have been identified in many organisms (Griffiths-Jones et al., 2006). These genes are frequently expressed individually, or in clusters of 2-7 genes with small intervening sequences (Mendes et al., 2009). Experimental studies suggest that they are expressed co-transcriptionally indicating that they are controlled by common regulatory sequences (Baskerville and Bartel, 2005). It is also possible to find miRNA genes in the introns of protein-coding genes (Lai et al., 2003; Lim et al., 2003b), introns and exons of non-coding genes (Rodriguez et al., 2004), in the 3′UTR of protein coding gene (Cai et al., 2004), and in the repetitive regions as is the case in mammals (Smalheiser and Torvik, 2005).

The miRNA biogenesis starts by the transcription of a miRNA gene. The primary miRNA precursor (pri-miRNA) is processed in the nucleus by a multiprotein complex

(microprocessor) containing an enzyme called *Drosha*. *Drosha* cuts both strands of the pri-miRNA near the stem-loop to generate 60-70 nucleotide stem-loop miRNA precursors (pre-miRNA) (Lee et al., 2002) (Figure 1.4). The precursor is transported to the cytoplasm by an export receptor known as *Exportin-5* (Yi et al., 2003; Lund et al., 2004). In the cytoplasm, the pre-miRNA is further cleaved by another enzyme known as *Dicer* to about 21 nucleotides long miRNA star (the sequence of mature miRNA at the arm of the secondary structure) duplex and then mature miRNAs are released for regulating targeted gene expression (Hutvágner and Zamore, 2002; Bartel, 2004). The mature miRNA binds perfectly in plant, but imperfectly in animals to the 3′UTR of the target mRNA (Reinhart et al., 2002; Carthew and Sontheimer, 2009). This pairing guides a ribonucleoprotein complex known as the RNA-induced silencing complex (RISC) to inhibit mRNA translation.

1.2.3 Function of microRNAs

The main function of miRNAs is to control protein synthesis either by binding and guiding the RISC to cause degradation of the targeted mRNA or by inhibiting the translation (Bartel, 2009). The cleavage or translation inhibition of the mRNA depends on miRNA:mRNA complementarity. If the complementarity between the miRNA and its mRNA target is high, the RISC will cleave the mRNA. However, if the complementarity is not sufficient for cleaving but still fitting, translation will be repressed. In addition to mRNAs repression, recent studies indicate that miRNAs may have a positive regulatory effect (Vasudevan et al., 2007; Place et al., 2008). Most miRNAs bind in the 3′UTR of their target mRNAs. Nevertheless, binding can also occur in the 5′ untranslated region (5′UTR) in some rare cases (Lytle et al., 2007). Furthermore, an

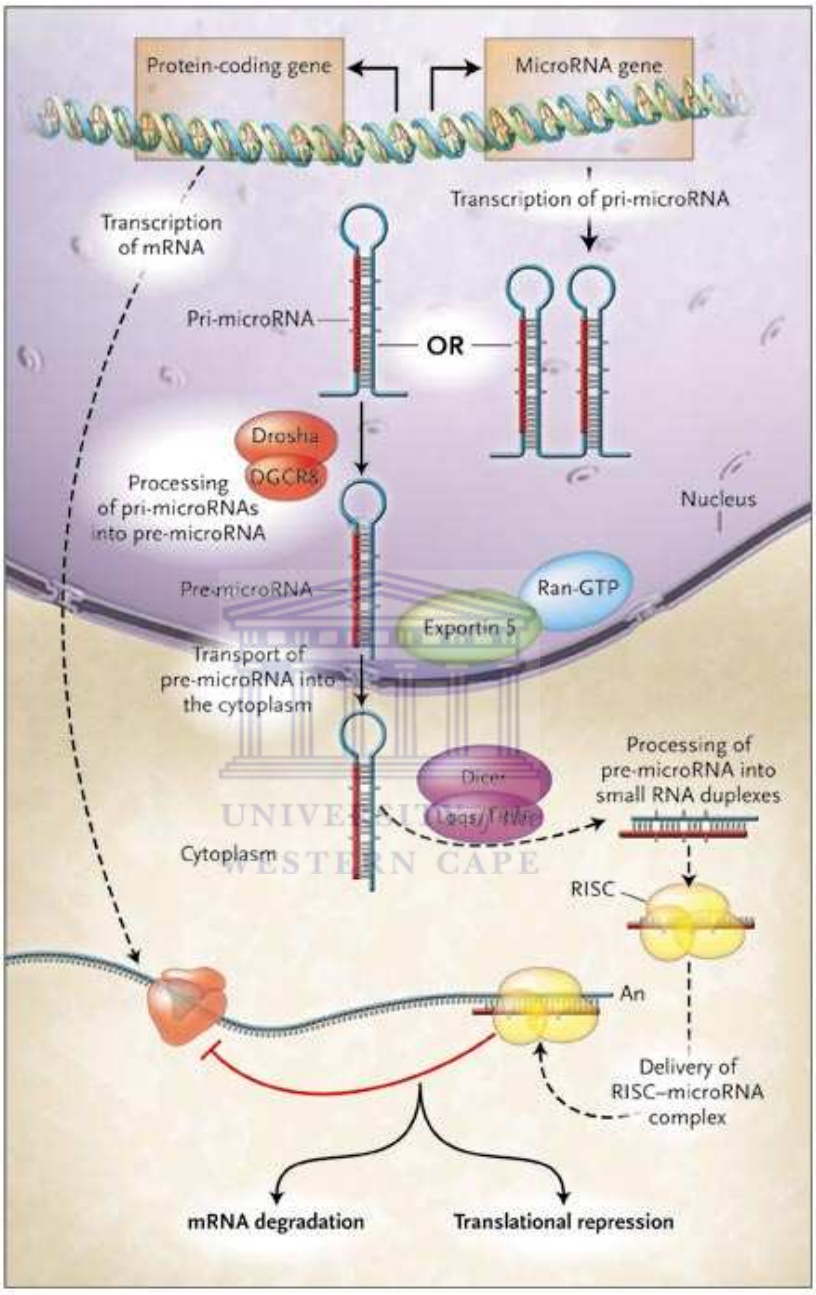


Figure 1.4: Biogenesis of miRNAs. Mature miRNAs are generated from long primary microRNA (pri-miRNA) transcripts. Firstly, the pri-miRNAs are processed in the nucleus into stem-loop precursors (pre-microRNA) by *Drosha*. Secondly, the pre-miRNAs are then actively transported into the cytoplasm by *Exportin-5* and further processed into small RNA duplexes of approximately 22 nucleotides by the *Dicer* enzyme. The functional strand of the microRNA duplex is then loaded into the RNA-induced silencing complex (RISC). Finally, the microRNA guides the RISC to the cognate mRNA target for translational repression or degradation of mRNA (Chen, 2005).

mRNA can contain multiple sites for the same or different miRNAs. Consequently several different miRNAs can act together to repress the same gene.

1.2.4 General characteristics of microRNAs

Although the lengths of miRNA genes and pre-miRNAs vary, ranging from tens to several hundreds of nucleotides (Zhang et al., 2006c), mature miRNAs are only 18-24 nucleotides llong (Ambros, 2001). All pre-miRNAs have a hairpin-shaped stem-loop secondary structures. The uracil nucleotide is dominant at the first position at the 5′ end. The majority of mature miRNAs are located either at the 3′ site or the 5′ site of the stems. There may be several mismatches (usually less than four nucleotides) between mature miRNAs and miRNA star sequences, but no gap or loop present in the matched region. The sizes of stems and loops differ due to the different length of pre-miRNAs; the secondary structures of pre-miRNAs are also slightly different from each other (Zhang et al., 2006b). Although forming specific hairpin stem-loop structures is one of the most important characteristics of pre-miRNAs, it is not unique to pre-miRNAs. Lots of other coding or ncRNAs such as rRNAs, tRNAs and mRNAs, also have the similar hairpin structures (Zhang et al., 2006a).

Pre-miRNAs have a high minimal folding free energy index (Zhang et al., 2006a). Several studies observed that miRNA precursors have low folding free energy. Low folding free energy is considered one of the important characteristics of miRNAs (Ambros et al., 2003a; Bonnet et al., 2004b). However, minimal folding free energy depends on the length of RNAs (Seffens and Digby, 1999). The length of miRNA precursors vary significantly, for example, the lengths of plant miRNA precursors range from 60 to more than 400 nucleotides (Zhang et al., 2006a). Thus, it is impossible to argue

that low folding free energy is the only criterion to distinguish miRNAs from other RNAs and to compare miRNA precursors with each other (Zhang et al., 2006a). To avoid this pitfall, the length of RNAs must be considered as well (Adai et al., 2005). However, there is a study which demonstrates that the adjusted minimal folding free energy (a combination of folding free energy and miRNA length) of more than 50% of tRNAs still fall into the range of miRNAs although the average adjusted minimal folding free energy (-45.93 $\pm$ 9.43 kcal/mol) of 513 known pre-miRNAs was significantly lower than other RNAs, including tRNAs, rRNAs and mRNAs (Zhang et al., 2006a). To better distinguish miRNAs from other RNAs, Zhang et al. (2006c) combined several parameters to form a new criterion called the minimal folding free energy index. They found that the average minimal folding free energy index of miRNA precursors was 0.97 in previously known plant pre-miRNAs. This value is significantly higher than that for tRNAs (0.64), rRNAs (0.59), and mRNAs (0.62-0.66). More importantly, more than 90% of miRNA precursors had a minimal folding free energy index greater than 0.85, and no other RNAs had a minimal folding free energy index higher than 0.85. This suggests that the minimal folding free energy index is useful to distinguish miRNAs from other coding and ncRNAs. Their results suggest that the RNA sequences with a minimal folding free energy index larger than 0.85 are most likely to be miRNAs. This finding provides a more precise criterion to predict miRNAs using computational or experimental approaches.

Many mature miRNAs are believed to be evolutionarily conserved from species to species in both animal and plant kingdoms although their pre-miRNAs are less conserved. In plants, homology search using expressed sequence tags (ESTs) identified a total of 481 miRNA homologs in 71 closely and distantly related plant species (Zhang et al., 2005). This result suggests that several miRNA families predated the divergence

between vascular plants and mosses. Moreover, this finding also suggests that miRNAs are highly conserved across phylogenetic distances (Floyd and Bowman, 2004; Zhang et al., 2006a,b).

The miRNA star duplex is relatively consistent among different species, but the size of fold-back and the extent of base pairing outside the duplex are highly variable even in closely related species (Jones-Rhoades and Bartel, 2004; Zhang et al., 2006a). In addition to the conserved miRNAs, there are lots of non-conserved miRNAs (species-specific miRNAs) that may control the specific characteristics that are unique to those species. Many studies demonstrate that this class of non-conserved miRNAs exist in plants and animals (Bentwich et al., 2005; Lindow and Krogh, 2005; Zhang et al., 2006a).

The majority of known miRNAs have high complementary site(s) at their targeted mRNAs, and this complementarity is conserved evolutionarily (Rhoades et al., 2002; Jones-Rhoades and Bartel, 2004; Robins et al., 2005). In plants this site is either perfect or nearly perfect (Llave et al., 2002; Reinhart et al., 2002; Rhoades et al., 2002). However, it is imperfect in animals and this complementarity only exists in the seed regions (Reinhart et al., 2002; Carthew and Sontheimer, 2009). In both, no gaps were allowed in the matched regions between miRNAs and their targeted mRNAs (Zhang et al., 2006b; Nielsen et al., 2007).

Of all of these characteristics, the hairpin secondary structure and conservation are the two most important characteristics. Although miRNAs regulate gene expression by complementarity to their targeted mRNAs, siRNA also have a similar gene regulatory mechanism (Coburn and Cullen, 2003). To avoid inaccurately classifying siRNA fragments as miRNAs, two studies built a constricted set of criteria for considering miRNAs, which include both biogenesis and expression aspects (Ambros et al., 2003a,b).

1.2.5 Identification of microRNA genes

Approaches used to identify miRNA genes include biochemical methods based on the purification of RNAs after size fractionation (Lau et al., 2001) and computational approaches centering on the conservation of DNA region between two clearly related species (Lai et al., 2003; Lim et al., 2003b). Since the discovery of the first miRNA in 1993 hundreds of miRNAs from many organisms have been cloned. However, only abundant miRNA genes can be easily detected by methods such as polymerase chain reaction (PCR) or northern blot due to limitations of these techniques (Lim et al., 2003a; Chen et al., 2005). Computational prediction, on the other hand , provides a more efficient strategy to find those low-expression or tissue-specific miRNA (Bartel, 2004). Currently, several computational approaches have been reported to identify miRNAs (Grad et al., 2003; Lim et al., 2003b). Most of these methods are based on the major characteristic features of miRNAs: the hairpin-shaped stem-loop secondary structure (Lagos-Quintana et al., 2001; Lau et al., 2001; Lee and Ambros, 2001), high evolutionary conservation from species to species (Grad et al., 2003; Lai et al., 2003; Axtell and Bartel, 2005; Zhang et al., 2006a), and the high minimal folding free energy index (Zhang et al., 2006c). The computational approaches can be classified into five major categories namely; the homology based approach, the gene finding approach, the neighbor stem-loop search, algorithms based on comparative genomics and phylogenetic shadowing based approach (Zhang et al., 2006b).

1.2.5.1 Homology based approach

This approach was used to reveal orthologs and paralogs of known miRNAs (Pasquinelli et al., 2000; Lau et al., 2001; Lee and Ambros, 2001; Weber, 2005). Since the begin-

ning of miRNA identification, it was generally accepted that miRNAs are evolutionarily conserved in plants and animals (Lagos-Quintana et al., 2001; Axtell and Bartel, 2005; Zhang et al., 2006a). This approach is effective enough to identify miRNA orthologues or homologues by searching publicly available DNA databases against known miRNAs that are experimentally identified in model species. Using this approach, many more miRNAs were identified (Weber, 2005; Zhang et al., 2005), and produced evidence that miRNAs are conserved in different species (Pasquinelli et al., 2000; Lagos-Quintana et al., 2001; Lee and Ambros, 2001; Floyd and Bowman, 2004; Jones-Rhoades and Bartel, 2004). But, sequence alignment alone may fail to detect distant homologs that diverge in sequence but conserve their structure (Legendre et al., 2005). Thus, based on conserved sequences and secondary structures, more miRNAs were predicted in human, mosquito and plant genomes (Nam et al., 2005; Wang et al., 2005; Dezulian et al., 2006). Homology searches can be classified as genome-based searches or EST-based searches. EST-based search has proven to be an economically feasible alternative for gene discovery in species lacking a sequenced genome (Matukumalli et al., 2004). Many important genes have been found through EST analysis (Ohlrogge and Benning, 2000; Graham et al., 2004). Therefore, an EST based search is a powerful approach to identify miRNA genes in various species, especially in species whose genome sequences are not available (Zhang et al., 2006b).

1.2.5.2 Gene finding approach

Gene finding approaches do not depend on homology or proximity to previously known miRNAs, and can be used for an entire genome search (Ambros et al., 2003b; Brennecke et al., 2003; Grad et al., 2003; Lai et al., 2003; Lim et al., 2003b; Bartel, 2004; Brown and Sanseau, 2005). This approach identifies conserved genomic regions be-

tween different organisms (Xie et al., 2005). Conserved regions are then placed into a window that can hold about 110 nucleotides. The window is folded with a secondary structure prediction program, such as Mfold (Zuker, 2003) or RNAfold (Hofacker et al., 1994), which score these hairpin-shaped stem-loops for potential miRNA candidates. Two computer programs have been developed (miRseeker (Lai et al., 2003) and miRscan (Lim et al., 2003b)) and successfully predicted animal miRNA genes. miRseeker computationally identified miRNAs of *Drosophila melanogaster* and *D. pseudoobscura* by analyzing the completed euchromatic sequences for conserved sequences that adopt an extended stem-loop secondary structure and display a pattern of nucleotide divergence characteristic of previously known miRNAs (Lai et al., 2003). miRscan is another computational program used to specifically identify miRNAs based on common characteristics (such as base pairing and nucleotide bias) of previously known miRNAs and their conservation in two genomes (Lim et al., 2003b). This program was initially used to predict miRNA genes in *C. elegans* (Lim et al., 2003b), and then in humans (Lim et al., 2003a). miRseeker and miRscan have identified dozens of new miRNA genes in both invertebrates and vertebrates.These have been further confirmed by experimental approaches. However, due to the window size limit in these two programs, it is difficult to employ them in the prediction of new miRNAs which have different lengths (John et al., 2004; Rodriguez et al., 2004; Zhang et al., 2006c).

1.2.5.3 Neighbor stem-loop search

The neighbor stem-loop search approach is based on miRNA clusters and secondary structures. Many miRNA genes occur as tandem arrays within a cluster like operons (Seitz et al., 2004; Tanzer and Stadler, 2004; Altuvia et al., 2005). This cluster characteristic suggests a useful approach to identify new miRNA genes by searching the

neighbors of known miRNA genes for other hairpin shaped stem-loops that may be additional miRNA genes of a genomic miRNA cluster. This approach has been used to predict many human and animal miRNA genes (Ohler et al., 2004; Altuvia et al., 2005). Nevertheless, this approach cannot be used in plants due to the fact that very few plant miRNA clusters have been observed (Jones-Rhoades and Bartel, 2004; Zhang et al., 2006a).

1.2.5.4 Algorithms based on comparative genomics

Comparative genomics uses sequence comparisons between species to identify different genes and regulatory elements (Loots et al., 2000; Boffelli et al., 2003). Recently, it has become a powerful approach to predict miRNA genes in animals and plants through the comparison of two known genomes (Tagle et al., 1988; Boffelli et al., 2003; Bonnet et al., 2004a; Jones-Rhoades and Bartel, 2004; Berezikov et al., 2005; Brown and Sanseau, 2005). Although these methods are widely used in computational identification of miRNA genes, their application is somewhat limited by the lack of genome sequences for the majority of animal and plant species.

1.2.5.5 Phylogenetic shadowing based approach

Phylogenetic footprinting (cross species sequence comparison) is one approach to identify functional genetic elements (Tagle et al., 1988). Since its introduction, phylogenetic footprinting has been employed to predict many gene structures and regulatory elements (Berezikov et al., 2005). However, the sensitivity of this method decreases as phylogenetic distance increase. Additionally, species-specific elements may be missed by this approach (Boffelli et al., 2003), especially for short sequences such as miRNAs and other small RNAs. To overcome these limitations, Boffelli et al. (2003) devel-

oped a variant method of the phylogenetic footprinting method termed phylogenetic shadowing. Their modified method not only examines sequences from closely related species, but also takes into account the phylogenetic relationship between the target species. Thus, it allows unambiguous sequence alignments and accurate conservation determination at single nucleotide resolution levels.

1.2.6 The miRNAs database

Experimental and computational approaches have been used to predict and identify miRNA genes in various species. Accordingly, more than 18,000 hairpin precursor miRNAs, expressing about 21,000 mature miRNA products belonging to 153 different species have been characterized and deposited in the miRNA database, miRBase (Griffiths-Jones et al., 2006; Griffiths-Jones, 2006; Griffiths-Jones et al., 2008; Kozomara and Griffiths-Jones, 2011). The miRBase database, initially called the microRNA registry, contains published mature miRNA sequences, along with their predicted source hairpin precursors and annotations relating to their discovery, structure, and function. Recently, it has been updated to include predicted miRNA targets across many species (Enright et al., 2003; Griffiths-Jones et al., 2006).

1.3 MicroRNA targets

With the increase in the number of miRNA genes which have been identified in viruses, plants and animals, the targets for a majority of these miRNAs have not been identified due to the fact that large-scale experimental detection of targets is not currently available (Griffiths-Jones et al., 2006). Identifying miRNA targets in plants is much easier than animals, as animal miRNAs have limited sequence complementarity to

their gene targets (Reinhart et al., 2002; Wang and El Naqa, 2008; Carthew and Sontheimer, 2009). This allows the prediction of animal miRNA targets by computational approaches (Mazière and Enright, 2007; Wang and El Naqa, 2008). Several studies indicate that computational approaches play an important role not only in the discovery of miRNA genes but also in the identification of miRNA targeted genes (Zhang et al., 2006b). The majority of these approaches rely on either one or combination of miRNA target prediction features such as miRNA:mRNA pairing, site location, conservation, site accessibility, multiple sites and expression profiles (Saito and Saetrom, 2010).

1.3.1 MicroRNA target prediction features

1.3.1.1 MicroRNA and microRNA target pairing

The target mRNA must have at least one region that has the Watson-Crick pairing to the 5′ region of miRNA. This part, located at positions 2-7 from the 5′ end of miRNA, is known as the ′seed′, and RISC uses these positions as a nucleation signal for recognizing target mRNAs (Lewis et al., 2003; Stark et al., 2003; Rajewsky and Socci, 2004). The corresponding sites in mRNA are referred to as ′seed sites′. A stringent seed site has perfect Watson-Crick pairing and can be divided into four seed types; 8mer, 7mer-m8, 7mer-A1 and 6mer depending on the combination of the nucleotide of position 1 and pairing at position 8 (Figure 1.5). 8mer has both an adenine at position 1 of the target site and base pairing at position 8. 7mer-A1 has an adenine at position 1, while 7mer-m8 has base pairing at position 8. 6mer has neither an adenine at position 1 nor base pairing at position 8 (Lewis et al., 2005; Grimson et al., 2007; Nielsen et al., 2007).

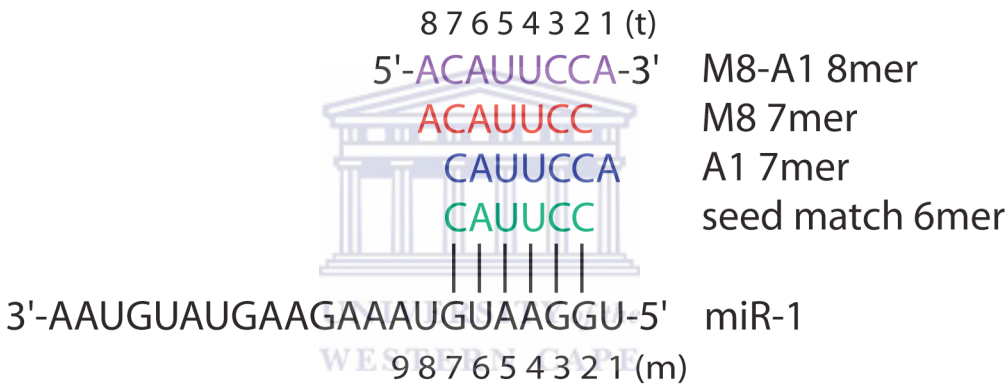


Figure 1.5: MicroRNA seed match types. Seed match types and numbering system illustrated for *miR-1*. Positions in the miRNA are numbered 5′-3′. (Seed match 6 mer) Watson-Crick inverse complement of miRNA bases 2-7; (A1) presence of adenosine opposite miRNA base 1; (M8) Watson-Crick match to miRNA base 8 (Nielsen et al., 2007).

1.3.1.2 Target site location

Most miRNA target sites can be found in the 3′UTR segment of the mRNA target genes (Lai, 2002; Zhao et al., 2008; Neilson and Sharp, 2008), even though miRNA-loaded RISC can theoretically bind any segment of mRNA (Wang et al., 2009a). Target genes tend to have longer 3′UTR, whereas ubiquitously expressed genes, such as house keeping genes, have shorter 3′UTRs, and so potentially avoiding regulation by miRNAs (Stark et al., 2005). Target sites are not evenly distributed within 3′UTR, but are located near both ends when the length of 3′UTR is more than 2000 nucleotides (Mazière and Enright, 2007; Fang and Rajewsky, 2011). For shorter 3′UTRs, sites tend to be near the stop codon (Gaidatzis et al., 2007). However, in longer 3′UTRs sites are located 15-20 nucleotides away from the stop codon (Grimson et al., 2007). In addition, some genes have alternative splicing in their 3′UTR segments, especially genes with long 3′UTRs (Hughes, 2006). These genes might therefore have different potential target sites for alternatively spliced 3′UTRs (Majoros and Ohler, 2007). Although functional miRNA sites are preferentially located in the 3′UTR, seed sites in the coding sequence (CDS) and 5′UTR regions can be targets for the miRNAs (Kloosterman et al., 2004; Lytle et al., 2007).

1.3.1.3 Conservation

miRNA families comprised of miRNAs that have the same seed site, and are well conserved among related species (Stark et al., 2003; Brennecke et al., 2005; Krek et al., 2005; Friedman et al., 2009). In addition, miRNA families have targets that are conserved among related species (Friedman et al., 2009). However, there are also species-specific miRNAs and targets. One study shows that about 30% of the experimentally

validated target genes might not be well conserved (Sethupathy et al., 2006). siRNA off-target effects occur whether or not the site is conserved (Burchard et al., 2009), therefore searching for all potential target sequences without considering conservation might increase siRNA off-target detection efficacy. Applying a filter that requires predicted target sites to be conserved can decrease the false positive rate, but such a filter is effective only for conserved miRNAs. It is important to identify targets both with and without conservation especially when species-specific miRNAs or siRNA off-targets are of interest (Lekprasert et al., 2011).

1.3.1.4 Target site accessibility

The mRNA secondary structure is very important for miRNA targeting. An effective miRNA:mRNA interaction needs an open structure on the target site to begin the hybridization reaction (Ragan et al., 2011; Mückstein et al., 2006). After binding, RISC can disrupt the secondary structure (Kertesz et al., 2007; Long et al., 2007). Minimum free energy is usually used to estimate the secondary structure and RNA hybridization, but the amount of A:Us surrounding the site can also be used to estimate the site accessibility (Robins et al., 2005; Long et al., 2007; Kertesz et al., 2007; Marín and Vanícek, 2011; Marín and Vaníček, 2012). Effective target sites often have A:U rich context in approximately 30 nucleotides upstream and downstream from the seed matching region of the target site (Grimson et al., 2007). Calculating the minimum free energy of accessibility and hybridization with the mRNA secondary structure requires the analysis of different mRNA folding patterns. This process requires enormous amounts of computing power as finding the most stable RNA structure is a computational problem that scales with the cube of the length of the RNA sequence (Eddy, 2004). Hence, finding hybridization sites in long 3′UTRs tends to be time-consuming. Moreover,

the current thermodynamic models used in RNA secondary structure prediction algorithms are only 90-95% accurate, which results in the algorithms only having 50-70% of the base pairs correct (Eddy, 2004). Thus, despite being theoretically sound, calculating site accessibility has limited practical value when predicting miRNA target sites. Heuristics that are easy to compute, such as local A:U context perform similarly.

1.3.1.5 Multiple target sites

Strong miRNA targets tend to have multiple target sites instead of one single site in the same 3′UTR (John et al., 2004). The use of predicted binding sites conserved across orthologous 3′UTRs in multiple species are considered more likely to reduce the number of false positives (Enright et al., 2003; Lewis et al., 2003; Stark et al., 2003; Mazière and Enright, 2007). However, recently evolved miRNAs, such as *miR-430* in zebrafish, may not have conserved targets in the scope of the currently available fish genomes (Giraldez et al., 2006).

1.3.1.6 MicroRNA and microRNA target expression profile

One miRNA can potentially regulate many genes. Therefore, expression profiles of mRNAs vary substantially depending on the miRNA expression levels (Ritchie et al., 2009). Many miRNA are also expressed differently in different tissues (Winter et al., 2007). Consequently, if negatively correlated expression levels of a miRNA:mRNA pair are detected across different tissue profiles, the mRNA of the pair is probably targeted by the miRNA (Lim et al., 2005; Rajewsky, 2006). Filtering putative targets based on expression profile correlations is an effective approach to reduce the false positive rate. Although the majority of miRNA targets appear to be regulated both at the mRNA and protein level, some targets only show an effect at the protein level.

1.3.2 MicroRNA target prediction tools

Currently, several freely available web-based and non web-based approaches are used to predict miRNA targets (Table 1.1). Most tools seek conserved 3′UTR sites with favourable thermodynamic hybridization energies, and use the detection of seed matches as a primary sieve (Mendes et al., 2009). Other approaches resort to machine learning techniques in an attempt to grasp the rules of target site recognition from the small set of confirmed targets. Although the success of many algorithms in predicting target sites for several miRNAs, the number of false positives remains high (Rajewsky and Socci, 2004; Lindow and Gorodkin, 2007; Mazière and Enright, 2007; Brodersen and Voinnet, 2009; Mendes et al., 2009). Two approaches have been proposed to try to achieve better specificity: the use of mRNA expression data, and the incorporation of mRNA secondary structure in the thermodynamic hybridization model (Saito and Saetrom, 2010).

1.3.2.1 Seed-based approaches

An early approach searched for targets of *Drosophila* miRNAs by preparing a database of 3′UTR sites conserved across two drosophilid species (Stark et al., 2003; Mendes et al., 2009). A distinguishable pattern of better conservation at the region that matches the 5′end of the miRNA was observed, showing better complementarity with the miRNA. Few mismatches or G:U pairs in the first 8 nucleotides were noticed. This observation prompted the search for conserved sequences that matched the first eight positions of the miRNA. Subsequently, the duplexes obtained were ranked by free folding energy. The statistical significance of the hybridization energies was evaluated against a background of 10,000 randomly selected target sites. Several instances of multiple target

Table 1.1: MicroRNA target prediction tools

Tool Name	Prediction Method	Organism	Website	Reference
TargetScan	-seed pairing -site location -conservation -site accessibility -multiple site	-Vertebrates -Worm -Fly	http://www.targetscan.org	(Lewis et al., 2005)
miRanda	-seed pairing -conservation -site accessibility -multiple site	-Human -Mouse -Rat -Worm -Fly	http://www.microrna.org	(Enright et al., 2003)
PicTar	-seed pairing -conservation -site accessibility -multiple site	-Vertebrates -Worm -Fly	http://pictar.mdc-berlin.de	(Krek et al., 2005)
EMBL	-seed pairing -conservation -site accessibility -multiple site	-Fly	http://.russelllab.org/	(Stark et al., 2003)
DIANA-microT	-seed pairing -conservation -site accessibility	-Human -Mouse	http://diana.cslab.ece.ntua.gr/microT/	(Maragkakis et al., 2009)
RNAhybrid	-seed pairing -free energy -conservation	-Human -Worm -Fly	http://bibiserv.techfak.uni-bielefeld.de	(Rehmsmeier et al., 2004)
MicroCosm Targets	-seed pairing -conservation -site accessibility -multiple site	-Vertebrates -Worm -Fly	http://www.ebi.ac.uk/enright-srv/microcosm	(Griffiths-Jones et al., 2006)
MovingTargets	-seed pairing -site location -conservation -site accessibility -multiple site	-Fly	-	(Burgler and Macdonald, 2005)
TargetBoost	-seed pairing	-Worm	https://demo1.interagon.com/targetboost	(Saetrom et al., 2005)
miTarget	-seed pairing -site accessibility	-	http://cbit.snu.ac.kr/∼miTarget	(Kim et al., 2006)
PITA	-seed pairing -conservation -site accessibility -multiple site	-Human -Mouse -Worm -Fly	http://genie.weizmann.ac.il	(Kertesz et al., 2007)
mirWIP	-seed pairing -site location -conservation -site accessibility -multiple site -expression profile	-Worm	http://146.189.76.171/query.php	(Hammell et al., 2008)
MicroInspector	-seed pairing -site accessibility	-Vertebrate -Invertebrate	http://bioinfo.uni-plovdiv.bg/microinspector/	(Rusinov et al., 2005)
MicroTar	-seed pairing -site accessibility	-Vertebrate -Invertebrate	http://tiger.dbs.nus.edu.sg/microtar/	(Thadani and Tammi, 2006)
miRDB	-seed pairing -site location -conservation -site accessibility	-Human -Mouse -Rat -Dog -Chicken	http://mirdb.org/miRDB/	(Wang, 2008)
EIMMo	-seed pairing -conservation -multiple site	-Human -Mouse -Fly -Fish	http://www.mirz.unibas.ch/ElMMo2/	(Gaidatzis et al., 2007)
NbmiRTar	-seed pairing -conservation -site accessibility	-Vertebrate -Invertebrate	http://wotan.wistar.upenn.edu/NBmiRTar/	(Yousef et al., 2007)
RNA22	-seed pairing -site location -conservation -site accessibility -multiple site	-Vertebrate -Invertebrate	http://cbcsrv.watson.ibm.com/rna22.html	(Miranda et al., 2006)
TargetRank	-seed pairing -conservation -site accessibility -multiple site	-Human -Mouse	http://genes.mit.edu/targetrank/	(Nielsen et al., 2007)
HOCTAR	-seed pairing -conservation -site accessibility -expression profile	-Human	http://hoctar.tigem.it/	(Gennarino et al., 2011)

sites per mRNA were observed and while single hits were generally not statistically significant; the combined score of the binding sites per mRNA generally was, which led to the idea that several regulatory sites were required for efficient regulation. This approach predicted several targets, including five previously validated sites. Three targets from the novel predictions were experimentally verified (Stark et al., 2003).

1.3.2.1.1 miRanda

The miRanda method (Enright et al., 2003) considers all the known miRNAs of *D. melanogaster*. It encompasses three phases. In the first phase, the miRNAs are matched against the 3′UTR regions of all possible targets allowing for G:U pairs as well as indels of moderate size. The method does not rely on seed matches directly. Instead, it privileges complementarity at the 5′ end of the miRNA by using a scaling factor for scores computed in this region based upon the incorporation of some position-specific empirical rules. The second phase consists of computing the thermodynamic stability of the miRNA:target duplex. The third, and final phase is an assessment of the evolutionary conservation of miRNA target associations across two additional species. Finally, using a randomization procedure, Enright and colleagues estimated the false positive rate and showed that it is reduced if one considers only mRNAs with multiple target sites. This approach was later used to predict targets in humans and other vertebrates (John et al., 2004).

1.3.2.1.2 TargetScan

The first method to explicitly use the concept of seed matches was TargetScan (Lewis et al., 2003). The algorithm takes miRNAs conserved across a group of organisms and scans a set of orthologous 3′UTR sequences from these organisms. They defined the

seed match as small segments of seven nucleotides has perfect complementarity to the bases in positions 2-8 of the miRNA. These matches are then extended to target sites including the entire miRNA, allowing for G:U pairs, and using a folding algorithm to predict the secondary structure of the heteroduplex. To each putative target, a folding free energy value is assigned, and a Z-score is calculated based on the number of matches predicted in the same target transcript and respective free energies. The candidate transcripts for each organism are ranked by Z-score, and the process is repeated for each organism in the set. Cut-off values for rank and Z-score are given, and the final candidate set is composed of targets that respect the established limits for all orthologous transcripts. The same authors would later add more organisms to their working set (Lewis et al., 2005), which enabled them to relax both the rank and cut-off scores and rely exclusively on seed matches consisting of a segment of only 6 nucleotides while still improving the signal/noise ratio. More improvements were attained by analysing the sequences flanking the 6 nucleotides seeds, which would show a bias towards the presence of certain nucleotides in key positions, particularly an adenosine at the 3′ end of the target site.

1.3.2.1.3 DIANA-microT

The tool DIANA-microT was used to predict targets in humans (Maragkakis et al., 2009). The tool searched for targets of 10 miRNAs which were conserved in *Mus musculus* and in the set of all repeats-masked human 3′UTR sequences. The search method considered two hypotheses about miRNA:mRNA regulatory associations: (1) they should be conserved high-affinity interactions; (2) they should be structurally restrained due to the enzymology of the miRISC complex. The first observation resulted in an algorithm to compute the thermodynamic stability of imperfect miRNA:mRNA

pairings. The second hypothesis led to the speculation that the structural restraints might be reduced to a set of general rules. In order to identify these rules, the study performed a series of experiments whereby some putative target site sequences were cloned onto a reporter construct. These rules were then used to filter the initial set of candidates. The results obtained with these experiments once again underlined the importance of near-perfect complementarity on the first few nucleotides at the 5′ end of the miRNA.

1.3.2.1.4 RNAhybrid

The first tool that used on a single-genome was RNA-hybrid (Rehmsmeier et al., 2004). This method consists of a dynamic programming algorithm that calculates the energetically most favourable hybridization of a miRNA and its target mRNA. It also allows the user to specify a portion of the miRNA that should form a perfect helix, corresponding to the seed site. The statistical significance of the predicted targets is determined using extreme value statistics for minimum free energies normalized for target length, and a Poisson distribution is used to model multiple binding sites of a miRNA for the same target. The statistical treatment is extended with a comparative analysis of conserved binding sites in orthologous targets of related species.

1.3.2.1.5 PicTar

A popular tool, called PicTar, is a combined method that identifies individual miRNA target sites by searching near-perfect seeds defined as a stretch of seven nucleotides starting at position 1 or 2 from the 5′ end of the miRNA (Krek et al., 2005). These target sites are then filtered with respect to the minimum free energy of the heteroduplexes, and to whether these sites fall into overlapping positions across the aligned

orthologous sequences. The target sites that pass both these filters are called anchors. Sequences that show a user-defined minimum number of anchors are then ranked using an Hidden Markov Model (HMM) maximum likelihood score. This score is computed considering all segmentations of the target sequence into target sites and background. This accounts for the synergistic effect of multiple binding sites for a single miRNA or several miRNAs co-regulating the same transcript.

1.3.2.1.6 MovingTargets

The MovingTargets tool (Burgler and Macdonald, 2005), relies on a database of potential miRNA targets obtained through the identification of highly conserved segments of not more than 50 nucleotides on orthologous 3′UTR regions of two closely-related species. Target sites for a given set of miRNAs are sought on this database according to five user-adjustable criteria: (1) number of target sites in the mRNA; (2) stability of the miRNA:mRNA hybridization as measured by the minimum free energies; (3) number of consecutive base pairs in the heteroduplex involving the 5′ end of the miRNA; (4) total number of paired nucleotides in the 5′ end of miRNA; (5) number of G:U base pairs in the 5′ region of the miRNA.

1.3.2.1.7 Network-level conservation

A more recent approach was entirely based on network-level conservation of seed matches (Chan et al., 2005). The method began by exhaustively enumerating all the k-mers of lengths 7, 8, and 9, conserved across orthologous 3′UTRs of worm and fly genomes. A conservation score measures the overlap between the sets of orthologous regions containing at least one copy of a given k-mer. These scores are then compared with those obtained with a control assay done over randomized 3′UTRs. The results

show that high scoring k-mers score much higher in the real data than in the control. Some of these k-mers are known to be involved in post-transcriptional regulation and many of them are complementary to the 5′ ends of known miRNAs, more often than what would be expected by chance. Most high-scoring k-mers identified in worms were also conserved in flies and vice-versa. Candidate targets were identified as those genes whose 3′UTRs and that of its orthologues contain a high-scoring k-mer which is also complementary to the 5′ end of a miRNA.

1.3.2.2 Machine learning approaches

TargetBoost is a machine learning method that combines genetic programming with boosting (Saetrom et al., 2005). Instead of relying on criteria based on sequence complementarity, thermodynamic stability, or evolutionary conservation, TargetBoost tries to learn the hidden rules of miRNA:target site hybridization. The genetic programming component consists of spawning and evolving a series of pattern sequences which try to describe the general properties of miRNA target sites, namely the existence of a nucleus of consecutive paired bases or a bulge of unpaired nucleotides. Each of these pattern sequences act as a classifier. The pattern sequences are all combined using a boosting technique that gives each classifier a weight depending on its performance on the training set. Additional filters can be added to this procedure: for example the verification of evolutionary conservation, or the existence of multiple target sites in the same 3′UTR. Other machine learning approaches using the popular support vector machine (SVM) framework have been proposed. These approaches try to generalize from a modest set of experimentally verified positive and negative examples. An example is miTarget which uses an SVM considering structural features of the 5′ and 3′ half of the hybridization site, thermodynamic features and positional features (Kim et al., 2006).

However, for SVMs to work well, they normally require a large negative training set, which is not currently available for miRNA targets (Mazière and Enright, 2007).

1.3.2.3 Integration of target gene expression data

Initially it was thought that animal miRNAs, by interacting with multiple sites on 3′UTRs, would inhibit the accumulation of protein products of the targeted messages without affecting the level of expression of the corresponding mRNAs (Bartel, 2004). However, it is clear now that, in many cases, there is a direct impact on the concentration of mRNA transcripts (Lim et al., 2005). But, there are still many documented miRNAs which have no impact on target mRNA levels, and whose influence cannot explain by itself the observed decrease in protein accumulation, or that are more plausibly independently down regulated at the level of transcription (Pillai, 2005; Stark et al., 2005). Nevertheless, several target prediction methods incorporating putative target expression levels have been developed and have proved to be a valuable approach to target identification (Cheng and Li, 2008; Wang and El Naqa, 2008).

1.3.2.4 Integration of target secondary structure

Some earlier studies tried to implicitly incorporate target secondary structure as a measure of site accessibility in their prediction methods (Robins et al., 2005; Thadani and Tammi, 2006). However, major progress was made in understanding miRNA:target recognition mechanisms with the development of a thermodynamic model that incorporates measures of accessibility of target sites (Kertesz et al., 2007). According to this model, a crucial determinant of effective binding is the change of free energy between the unbound 3′UTR, with its pre-existing secondary structure, and the hybridized state. This model permitted the development of a new target prediction method called PITA.

More recently, another method takes advantage of a large dataset of experimentally verified miRNA:mRNA associations to derive a scoring scheme that combines site conservation, 5′ seed pairing, structural accessibility and hybridization energy criteria, illustrating the need to combine several features in order to accurately identify new targets (Hammell et al., 2008).

1.3.3 Comparison of miRNA target prediction tools

Many computational approaches for miRNA target prediction have been published (see sub-section 1.3.2). However, only a few surveys have independently compared some of the tools. But a recent survey analyzed 7 different tools. In this survey, they compared the predicted targets of the programs TargetScanS, PicTar, rna22, PITA, miRBase, miRanda and DIANA-microT 3.0 with the results of a pulsed SILAC analysis (pSILAC is method to directly compare protein translation rates between 2 samples) (Selbach et al., 2008). More specially they investigated the predicted mRNAs for five different miRNAs (*miR-1*, *miR-16*, *miR-30a*, *miR-155*, *let7b*) from each tool. They then compared the number of mRNAs predicted by a specific tool measured with pSILAC against the fraction of mRNAs that showed a log2(fold-change) lower than -0.1. For the comparison of the results with PITA only the 600 top-ranked predictions for each miRNA were considered. 27% of a completely random selection from the mRNAs considered in the pSILAC data show down-regulation. This accuracy was topped by all prediction programs. Considering only seed matches for prediction reveals a hit rate of 44%. This is clearly exceeded by TargetScanS, PicTar and DIANA-microT 3.0, which can be explained by a more rigid conservation filter (Selbach et al., 2008).
These results are supported by an older survey by Sethupathy and colleagues (2006).

Based on the experimental supported miRNA target interactions provided by TarBase, Sethupathy and colleagues (2006) estimated the performance of each program by determining sensitivity. They revealed a relatively low sensitivity for TargetScan and DIANA-microT. miRanda, TargetScanS, and PicTar showed almost the same sensitivity of about 65%. However, miRanda predicted a considerably higher number of miRNA-targets interaction compared to the number of individual predictions of TargetScanS and PicTar respectively. This indicates a lower sensitivity for miRanda. This shows that the union of the results by some prediction programs yield higher sensitivity but also result in a higher number of total target prediction, whereas the intersection of the programs lead to a lower number of predictions but also drops the sensitivity. These results indicates a great diversity of the accuracy of predictions between the programs (Leitner, 2009).

1.4 Thesis rationale

Generally, most of what we know about insect miRNAs comes from studies in *Drosophila* miRNAs (Winter et al., 2007; Skalsky et al., 2010). Currently, 426 mature miRNAs have been identified for *D. melanogaster*, the majority of which have orthologous sequences in other insects (Griffiths-Jones, 2006; Kozomara and Griffiths-Jones, 2011). In non-drosophiloid insects like mosquitoes, only 65 miRNA genes in the malaria vector (*An. gambiae* and *An. stephensi*) (Winter et al., 2007; Mead and Tu, 2008), 124 in the yellow fever mosquito (*Aedes aegypti* and *Ae. albopictus*) (Li et al., 2009b; Skalsky et al., 2010) and 93 in West Nile virus vector (*Culex quinquefasciatus*) (Skalsky et al., 2010) have been identified by experimental and/or by computational methods. The challenge remains to fully identify all miRNAs (especially very low-abundance and species-specific miRNAs) and to determine their functions. Sequencing-based applications for identifying and profiling miRNAs have been hindered by laborious cloning techniques and the expense of capillary DNA sequencing (Pfeffer et al., 2005; Cummins et al., 2006). Nevertheless, direct small RNA sequencing has several advantages over hybridization-based methodologies. Discovery of novel miRNAs need not rely on querying candidate regions of the genome but rather can be achieved by direct observation and validation of the folding potential of flanking genomic sequence (Berezikov et al., 2006; Cummins et al., 2006). Direct sequencing also offers the potential to detect variation in mature miRNA length. Recently next generation sequencing (NGS) technologies offer inexpensive increases in throughput, thereby providing a more complete view of the miRNA transcriptome. With the added depth of sequencing now possible, we have an opportunity to identify low-abundance miRNAs or those exhibiting modest expression differences between samples, which may not be detected by hybridization-

based methods.

miRNAs have a wide variety of expression patterns. In *C. elegans* and *Drosophila* some miRNAs are differentially expressed in time during development, whereas others seem to be more ubiquitously expressed (Pasquinelli et al., 2000; Aravin et al., 2003). To date, there are limited studies of miRNA expression in mosquitoes. In 2007, a study showed a significant change in expression profile of four *An. gambiae* miRNAs in response to *Plasmodium* invasion (Winter et al., 2007). Furthermore, when they knockdown two genes, *Dicer* and *Argonaute*, which are important in miRNAs maturation, they observed better survival of the parasite suggesting involvement of these tiny ncRNAs as a new layer in the regulation of *Anopheles* defence mechanism against malaria parasite. Therefore, identification of miRNAs controlling key genes required for mosquitoes to complete their life-cycle will not only help to better understand the vector biology, but it will also create potential tools for transgenesis and paratransgenesis.

1.5 Thesis objectives

The overall objectives of this thesis are:

1- Identification and characterization of miRNA genes in *An. funestus s.s*:

Although there are over 30 species of *Anopheles* which transmit malaria in the world, miRNAs have so far only been identified in two malaria vectors, the African vector (*An. gambiae*) and in the Asian vector (*An. stephensi*), using direct cloning and/or computational methods. This study will attempt to experimentally identify and computationally characterize the miRNAs expressed in the different developmental stages of other African vector *An. funestus s.s.*

2- Prediction of *Anopheles* miRNAs putative targets:

Understanding the functional significance of miRNAs required an analysis of their targets. This study will identify targets for the *Anopheles* miRNAs by using a hybrid approach of three existing methodologies. Given the false positive rate of target prediction tools, the predicted targets will be filtered using functional enrichment analysis. This analysis will be implement a statistical framework that uses Gene Ontology (GO) function and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway to assign a regulatory function for each miRNA.

3- Development of insect miRNA targets database:

Development of an online database for miRNA target prediction and functional annotations in the three disease vectors; *Ae. aegypti*, *An. gambiae* and *C. quinquefasciatus*. The web-based tool will allow updated miRNAs to be screened for target as described in the previous objective. This automated framework will allow the scientific community to have easy access to up-to-date miRNA target information as soon as new miRNAs are added to the system.

Chapter 2

microRNAs expressed in *Anopheles funestus s.s* developmental stages

Abstract

The females of many *Anopheles* species of mosquitoes are the principal vectors of human malaria, a disease with an enormous impact on public health worldwide. The parasites of mammalian malaria (genus *Plasmodium*) infect relatively few mosquito species of the genus *Anopheles*, in which they undergo complex growth and differentiation events essential for disease transmission. Understanding the interactions between the parasite and the mosquito vector has been the objective of extensive investigations aimed at identifying novel and efficient ways to disrupt or reduce pathogen transmission. Accumulating evidence demonstrates differential expression of host miRNAs in response to infection by various micro-organisms and the involvement of microorganism-encoded miRNAs in host manipulation. Within this context, to comprehensively investigate the roles of miRNAs in vectorial capacity and provide basic information for further understanding of the miRNA-mediated post-transcriptional regulation during anopheles development, we constructed four separate libraries derived from the different developmental stages of the African malaria vector *Anopheles funestus s.s*, representing the egg, larva, pupa and adult female stages. High throughput sequencing in combination with computational analysis enabled us to identify 65 miRNAs previously identified in anopheline species and an additional 33 novel miRNAs that had

not been reported in this genus. Moreover, several previously reported *Aedes* and *Culex* specific miRNAs, including *miR-2940*, *miR-2942*, *miR-2943* and *miR-2945*, were detected. These miRNAs appear to be restricted to mosquitoes. Furthermore, we identified new stem-loop precursors for *miR-286* and *miR-2944*, and a diverse population of mature miRNA variants for *miR-2* and *miR-927*. The expression profiles of the characterized miRNAs were analyzed in the four libraries. The insect-specific miRNA, *miR-263* was the most abundant miRNA in egg, larva, and pupa libraries, while *miR-8* was the most abundant miRNA in the adult library. Finally, we assessed changes in miRNAs expression during the mosquito development and observed the co-regulation of seven pairs of co-localized miRNAs (*miR-1174* and *miR-1175*, *miR-278* and *miR-307*, *miR-305* and *miR-275*, *miR-210* and *miR-927*, *miR-309* and *miR-286*, *miR-306* and *miR-375*, and *miR-34* and *miR-317*) and co-transcribed (*let-7* and *miR-92* families) between stages. This is, the first profiling study of miRNA associated with the maturation of mosquitoes. Understanding the functions of these mosquito's tiny RNA will undoubtedly contribute to a better understanding of mosquito biology including longevity, reproduction, and mosquito-pathogen interactions, which are important to understanding the transmission of malaria.

2.1 Introduction

miRNAs comprise a large family of endogenous, evolutionarily conserved, ncRNA that post-transcriptionally regulate mRNAs and influence fundamental cellular processes and gene expression programs in metazoan animals, plants and protozoa (Bartel, 2004; Krol et al., 2010). They modulate gene expression by binding to their target mRNAs (Lim et al., 2005; Baek et al., 2008). Because of its versatility, miRNAs have evolved as a major class of gene-regulatory molecules critical for diverse biological processes such as cell proliferation, differentiation, apoptosis, stress and immune response (Macdonald and Struhl, 1986; Hutvágner and Zamore, 2002; Brennecke et al., 2003; Teleman et al., 2006; Winter et al., 2007; Zhang et al., 2009; Hilgers et al., 2010; Vallejo et al., 2011; Marco et al., 2012; Choi and Hyun, 2012). Furthermore, several studies have shown differential expression of host miRNAs following infection (Winter et al., 2007; Hussain and Asgari, 2010; Skalsky et al., 2010; Zeiner et al., 2010; Dkhil et al., 2011; Hussain et al., 2011). This could be due to intrinsic or extrinsic factors produced during infection (Asgari, 2011). Among the extrinsic factors are regulatory small RNAs or proteins produced by micro-organisms that may influence the host miRNA profile by targeting host miRNA genes or interfering with the host RNA silencing pathways. Intrinsic factors are signals produced by the host following recognition of a foreign invader or replication of a microorganism. Comparing miRNA expression profiles in infected and uninfected tissues provides an overall picture of cellular miRNAs that change following infection with a microorganism. Therefore, identifying and probing miRNAs to understand their physiological and pathological roles have become popular research topics. After the discovery of the first miRNA in worm, thousands of miRNAs have been identified (Griffiths-Jones et al., 2006; Griffiths-Jones,

2006; Griffiths-Jones et al., 2008; Kozomara and Griffiths-Jones, 2011). The major approaches to identifying miRNAs include genetic screening, direct cloning, bioinformatic analysis, and deep sequencing (Wang et al., 2009b; Metzker, 2010). The majority of known miRNAs have been identified through traditional direct cloning, which is both time-consuming and labour intensive. Bioinformatic analysis is limited because the majority of computational programs are based on complete genome sequences. Computational programs are available only for a limited number of model species, limiting the application of these programs for the numerous non-model species. Next generation sequencing (NGS) has provided an innovative tool to look into the genome with an unprecedented depth of coverage. It allows for a comprehensive coverage of miRNAs of any species because miRNAs can be detected in any organism without prior sequence or secondary structure information (Creighton et al., 2010; Jung et al., 2010; Hackl et al., 2011; Havecker, 2011; Keller et al., 2011; Kong, 2011; Persson et al., 2011; Gébelin et al., 2012; Gunaratne et al., 2012; Pérez-Quintero et al., 2012). As a consequence, this new approach has opened the door to functional genomic analyses of non-model species. It is widely used for profiling miRNAs in populations in various developmental stages, in either normal and diseased states (Wei et al., 2009; Buchold et al., 2010; Nikopoulos et al., 2010; Borges et al., 2011; Krawetz et al., 2011; Mohorianu et al., 2011; Wu et al., 2011b; Boeri et al., 2012; Cutting et al., 2012; Gilabert-Estelles et al., 2012; Kang et al., 2012; Leidner et al., 2012; Liu et al., 2012b; Nikaki et al., 2012; Shamimuzzaman and Vodkin, 2012; Wei et al., 2012; Yao et al., 2012).

2.1.1 Next generation sequencing of miRNAs

NGS technologies enable the sequencing of the complete set of miRNAs present in an RNA sample (Motameny et al., 2010). The millions of short sequence reads generated by NGS, like the SOLiD (AppliedBiosystems) and Illumina genome analyzer, are particularly useful for small RNA transcription profiling. They provide miRNA expression profiling at an unprecedented sensitivity and resolution. Compared to available miRNA microarray platforms, the NGS systems are not limited by a predefined number of features, probe design, probe cross-hybridization or array background problems (Buermans et al., 2010). NGS systems, moreover, directly count the number of transcripts found as a measure of expression abundance, have high multiplexing potential, are species-independent, show a high sensitivity towards low abundant transcripts, and display excellent reproducibility (’t Hoen et al., 2008; Buermans et al., 2010).

2.1.1.1 The Illumina sequencing of miRNAs

miRNA sequencing using the Illumina genome analyzer comprises several steps as outlined in Figure 2.1 . Firstly, the total RNA is extracted from the sample. RNA preparation either spins a column-based kit like miRVana (Ambion) or miRNeasy (Qiagen), which allow for the preparation of the small RNA fraction can be used. TRIzol (Invitrogen, USA) preparation following ethanol precipitation can also be performed. Standard column-based RNA preparation kits common for mRNA preparation should be avoided because this normally leads to the loss of smaller RNA molecules (Motameny et al., 2010). Furthermore, it is useful to not only isolate miRNAs because in this case the rRNA fraction cannot be used for an RNA integrity analysis (Motameny et al., 2010). It is essential to confirm the quality of the RNA before sequencing to make

sure that biologically relevant oligonucleotides are sequenced and degradation products do not influence the results. A size selection is performed to extract the miRNA fragments from the total RNA. Then, the total RNA is run on an agarose gel and the band corresponding to the size of miRNAs is cut out for further processing. This procedure excludes all bigger fragments, including all mRNAs and also rRNAs from the samples. In the next step, the sequencing adapters are ligated to the size-selected RNA molecules. This is followed by reverse transcription to complementary DNA (cDNA). The obtained cDNA library is run on an agarose gel again and the band (approximately 65-70 bp) corresponding to the miRNA fragments with ligated adapters is cut out for subsequent sequencing. To avoid contamination with adapter dimers in this step, an additional control sample, containing adapter dimers only and a size ladder allowing the isolation of cDNA species larger than adapter dimers, is run in parallel to the cDNA library (Motameny et al., 2010).

2.1.1.2 Sequence quality filtering

The output of a next generation miRNA sequencing experiment typically contains millions of short reads. The reads are commonly provided as FASTQ format. Such files contain four lines per read; the first line contains the '@' symbol followed by a (unique) read identifier; the second line contains the read's nucleotide sequence; the third line again contains the read identifier, this time preceded by the '+' symbol, and the fourth line contains quality scores that specify the probability that the nucleotide call is wrong for each nucleotide in the read sequence (Cock et al., 2010). Missing nucleotides in the sequence are usually denoted by the 'N' character. Based on the quality scores, an initial filtering step can be performed to exclude reads of low quality as well as reads that contain too many missing nucleotides.

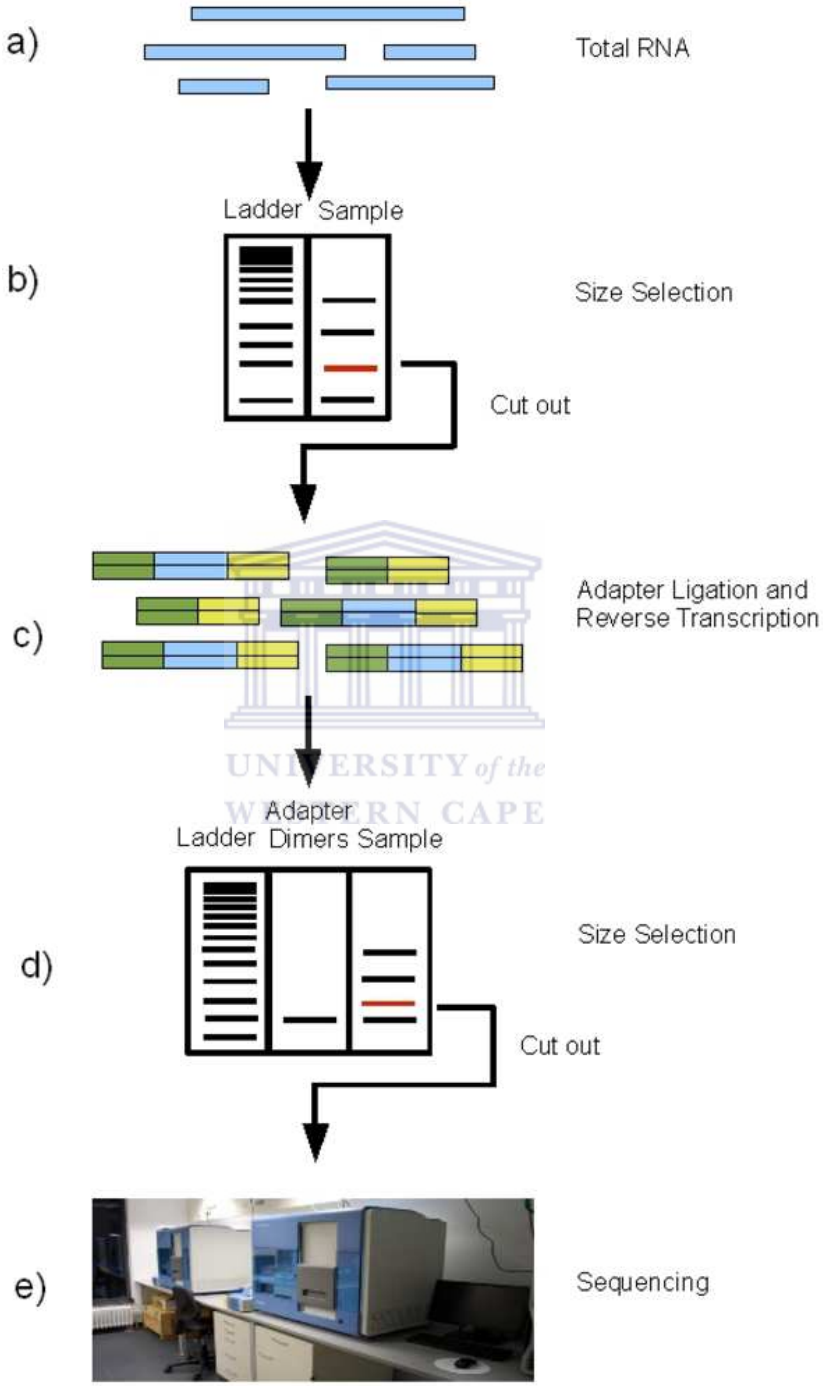


Figure 2.1: Sequencing procedure for miRNA on the Illumina genome analyzer. a) Extraction of total RNA, b) Size selection, c) Adapter ligation and reverse transcription, d) Size selection, e) Sequencing (Mendes et al., 2009).

2.1.1.3 Trimming sequencing adapters

As mature miRNAs are normally 24-25 bp in length, the reads will contain part of the 3′ adapter sequence that has to be removed. Tools for this purpose include the trimLRPatterns function contained in the Bioconductor Biostrings package for the R programming language and a BioPerl script. Softwares such as Novoalign (http://novocraft.com), FASTX-Toolkit (Pearson et al., 1997), miRDeep (Friedländer et al., 2008) and miRNAkey (Ronen et al., 2010) have an option to automatically trim adapter sequences from the reads. It is also convenient to align all reads against the adapter sequence using a flexible aligner like LASTZ or Megablast. The alignment information trims the adapter sequence from the reads. This strategy is preferable particularly if there are reads in the data that were sequenced through the complete 3′ and carry additional tails of unknown or highly error-prone sequence (typically occurring when longer reads of 50 bp and more are produced). In this case, the available tools for adapter trimming frequently miss the adapter sequence whereas aligners will reliably detect it (Motameny et al., 2010).

2.1.1.4 Alignment of reads to the reference genome

A minimum requirement for each sequence is that the read should originate from the genome of the sequenced organism. Therefore, as an initial step, all reads are aligned to the reference genome of the sequenced organism. The reads whose first part (15-17 bp) perfectly match the reference, are kept as potential miRNA reads (Morin et al., 2008). The remaining reads are discarded from further analysis. Short read aligners such as maq (Li et al., 2008a), SOAP (Li et al., 2008b, 2009a; Liu et al., 2012a), Eland (part of the Illumina pipeline software), or bwa (Pokrzywa, 2008) are preferred over BLAST

because they readily identify perfect matches at a much higher speed (Altschul et al., 1990). By default, short read-aligners consider reads that match the reference at several loci to be repetitive sequences and do not report such alignments. Modification to these default parameters are required to avoid the loss of relevant reads. To accurately identify known miRNAs, sequences can be aligned with those of annotated miRNAs as provided by provided by miRBase (Griffiths-Jones et al., 2008).

2.1.1.5 Filtering other small RNAs

In addition to mature miRNA reads, the sequencing data will most probably also contain reads from various other RNA species, including other small ncRNA species and RNA degradation products. It is reasonable to filter reads that align against such sequences prior to further downstream analysis in order to simplify the interpretation of the results. The sequences of such small RNAs like rRNA, small cytoplasmic RNA, small nuclear RNA, small nucleolar RNA, tRNA and protein coding regions, can be found in the RNA families database (Rfam) (Gardner et al., 2011) and on the University of California Santa Cruz Genome Browser (UCSC) (http://genome.ucsc.edu/index.html).

2.1.1.6 Prediction of known and novel miRNAs

A number of tools are available that predict miRNAs from NGS data (Wang et al., 2009b; Pantano et al., 2010; Zhu et al., 2010; Fasold et al., 2011; Hackenberg et al., 2011; Zhao et al., 2011; Friedländer et al., 2012; Zhang et al., 2012c). The predictions made by these tools are generally based on the current biological knowledge of the miRNA processing mechanism in living cells. One of the most commonly used prediction tools is miRDeep2 (Friedländer et al., 2012). The package was developed to discover known or novel miRNAs from deep sequencing data. It looks for the pat-

tern that the miRNA processing machinery leaves in the sequencing data. The most important pattern that miRDeep2 considers are clusters of reads that align along the reference genome that is compatible with the mature miRNA sequence, the loop sequence, and the star sequence structure of the miRNA precursor molecule. If such a pattern is found, miRDeep2 cuts out the potential miRNA precursor sequence from the reference genome and utilizes an RNA folding algorithm (randfold) from the Vienna package (Hofacker et al., 1994) to assess if the sequence can be folded into a hairpin structure. Furthermore, the prediction software searches for potential cleavage sites of *Drosha* and *Dicer*. The phylogenetical conservation and filtering of other known small ncRNA species can be optionally used to improve the predictions. More details about the miRDeep2 modules implementation in this project are described in the Methods sub-section 2.2.

2.1.1.7 Identification of miRNA isoforms

Data obtained from a miRNA sequencing experiment contains many sequences that are identical for all but a few nucleotides. These sequences represent different isoforms or variants of the same miRNA. Different types of miRNA isoforms have been described before, including isoforms that may arise from variability of *Drosha* and *Dicer* cleavage positions within the pre-miRNA and isoforms showing single-nucleotide 3′ extensions leading to mismatches with the reference genome (Morin et al., 2008). The origin and function of such isoforms are poorly understood but their presence suggests as yet unknown cellular mechanisms of miRNA processing. When analyzing miRNA sequencing experiments, isoforms can complicate the analysis process as well as the interpretation of the results. In expression analysis, for example, it is not immediately clear which of the different isoforms should be used in the expression profile compari-

son, especially if the expression changes of different isoforms show contrary behavior. Currently, it is common practice to only consider the isoform with the highest read count which seems to be a reasonable strategy for now but the handling of isoforms deserve greater attention in the future (Motameny et al., 2010; Pelaez et al., 2012).

2.1.1.8 miRNA expression patterns

miRNA expression levels are computed based on the read counts in each sequenced sample. For each unique sequence among the reads, the number of times it occurs among all the reads of the sample is computed and normalized against the total number of reads that were produced for the sample. Specifically, read counts are computed using per million value (rpm) of each sequence occurring in the sample according to the following formula: normalized expression = actual miRNA count / total count of clean reads $*10^6$. In theory, these normalized read counts should be a direct measure of the amount of fragments of the respective sequence in the sample, and therefore, its expression level. However, Linsen and colleagues (2009) showed that the fragment composition of a sample is significantly altered depending on the methods used for RNA extraction and library preparation (Linsen et al., 2009). The absolute normalized read counts are therefore not representative of expression levels. As in microarray analysis, the analysis is limited to relative comparisons of normalized read counts between samples to detect expression changes. Prior to performing expression analysis, sequencing errors have to be removed. For example, on the Illumina genome analyzer, single base substitution errors are the main concern. Assuming that the errors occur at random positions of the sequence and the substituted nucleotide is also selected randomly, sequences containing errors are expected to have low read counts. There is a big proportion of reads that are detected with less than 10 copies (Koh et al., 2010). Fil-

tering all sequences with read counts less than a threshold in each sample is therefore a common strategy to eliminate sequencing errors (Motameny et al., 2010). Usually, the threshold that is used for this filtering step is chosen arbitrarily. Many studies suggest a statistical method to determine the threshold automatically by iteratively comparing the cumulative distribution functions of read frequencies between replicate samples for different thresholds until the similarity between the distributions is satisfyingly high (Koh et al., 2010). The established methods from the analysis of microarray data are then used on the filtered sequences to identify differentially expressed sequences. These include the computation of fold changes if the experiment contains only two samples, the two-sample *t-test* if the experiment contains two groups of samples, or ANOVA if three or more groups of samples are used. In order to approximate the normality assumption that underlies most of the statistical methods mentioned above, the *log* normalized read counts should be used for these analyses.

The main objective of this chapter is to identify miRNAs expressed in the four different life stages of the mosquito *An. funestus s.s* using high throughput sequencing technology in conjunction with bioinformatics approaches. Determining the biological functions of these regulatory molecules in this mosquito will uncover novel approaches to control this vector-borne disease by investigating the potential roles of miRNAs in processes that are intimately linked to mosquito vectorial capacity.

2.2 Materials and Methods

The analysis pipeline that was developed and implemented for the identification and characterization of known and new miRNAs in *An. funestus s.s* is summarized in

Figure 2.2.

2.2.1 Mosquito strain and rearing condition

The experimental work was performed using a colony of *An. funestus s.s* (FUMOZ) that originates from Southern Mozambique (Hunt et al., 2005). The mosquitoes were reared in the insectary of the Vector Control Reference Unit at the National Institute for Communicable Diseases (NICD), Johannesburg, South Africa since 2000. The insectary is kept at 25°C, 80% relative humidity with a 12-h day/night lighting regime including 45-min dusk/dawn cycles.

2.2.2 RNA extraction

Total RNA was isolated separately from the four different life-cycle stages of *An. funestus s.s* (eggs, larvae, pupae, and adult females not fed on blood) using TRIzol reagent (Invitrogen, USA) according to the manufacturer's protocol. In order to obtain a large and broad miRNA transcriptome data set, RNA was extracted from 5 egg patches, 100 larvae, 100 pupae, and 100 unfed-females. The quantity of the RNA was assessed using a spectrophotometer (Nanodrop Technologies), while quality was determined by a 2100 Bioanalyzer (Agilent Technologies).

2.2.3 Small RNA sequencing

The RNA extracts from the four developmental stages were sent to Macrogen Inc (Seoul, South Korea) for small RNA sequencing. Sequencing was performed on the Illumina HiSeq^{TM} 2000 machine.

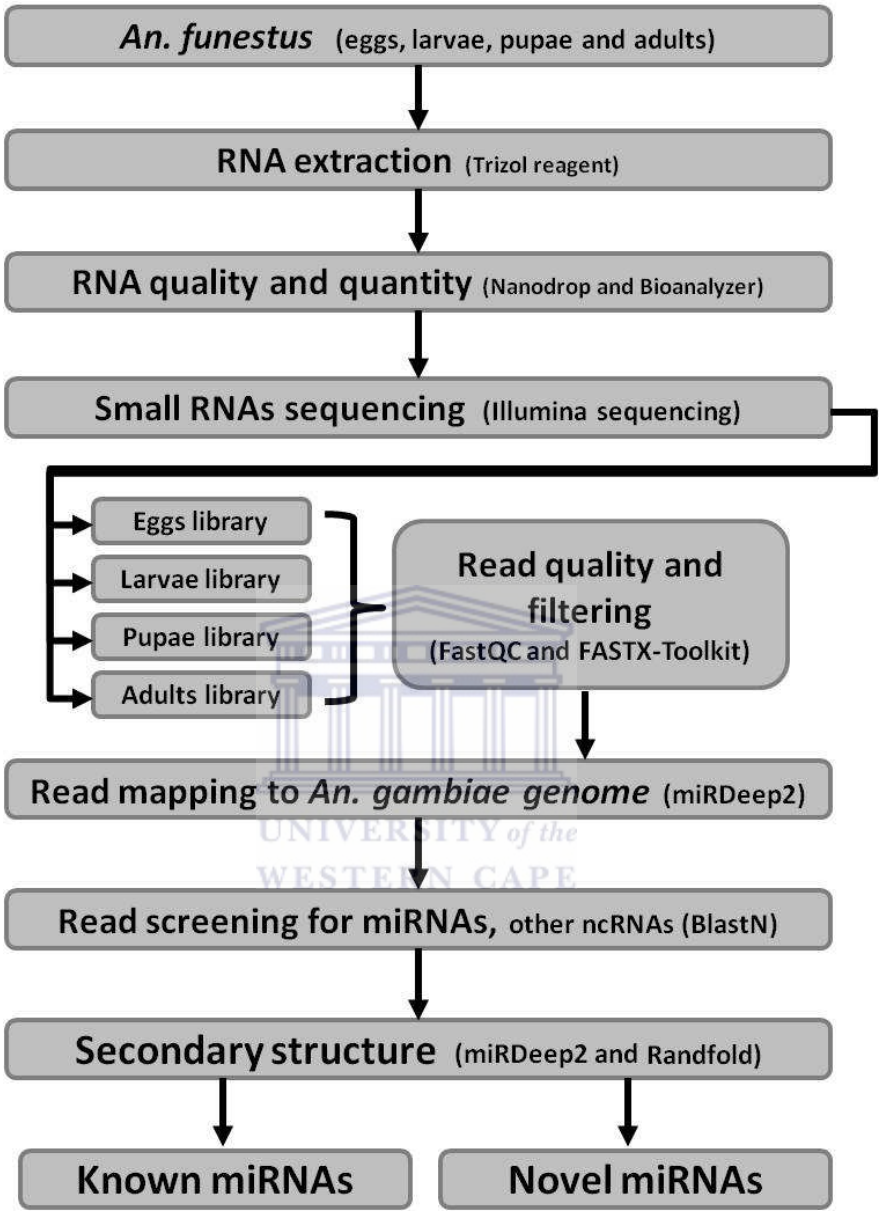


Figure 2.2: Schematic overview of analysis pipeline for identification and characterization of *An. funstus* miRNAs. The experimental work (mosquito rearing, RNA extraction, quality, and quantity) was conducted in the Vector Control Unit, National Institute for Communicable Diseases, South Africa. Small RNA sequencing was done by Macrogen Inc, South Korea and the computational analysis (read quality, filtering mapping, and identification of know and novel miRNAs) was performed at South African National Bioinformatics Institute, University of the Western Cape, South Africa.

2.2.4 Sequence data processing and analysis

2.2.4.1 Reads quality check and filtering

Following sequencing, the quality for the four sequenced libraries were checked using Phred (Ewing et al., 1998) and FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) (Appendix A, Figure 1,2,3 and 4). All sequencing reads with low quality tags and shorter than 10 nucleotides were removed using the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/).

2.2.4.2 Mapping the reads to the reference genome

In the absence of the *An. funestus* genome we mapped all obtained reads to the *An. gambiae* genome which was downloaded from Ensembl (http://metazoa.ensembl.org, assembly AgamP3, database version 63). Sequence reads were mapped to the genome using miRDeep2 (Friedländer et al., 2012). Briefly, the mapper module was used first to test the format of the input files (the FASTQ files from each library and the reference genome file). The sequencing file was then converted into a FASTA format. This was followed by clipping 3′ adapters. The reads were searched for matches to the adapter sequence. Once a match was found, the match to the adapter sequence and all nucleotides downstream were clipped from the read, and the next read was searched. Reads that have no matches were retained, but not clipped. All reads with identical sequences were collapsed to remove redundancy. A digit in the new FASTA identifiers showed how many times the corresponding sequence was present in the data set. Reads were mapped to the reference genome (*An. gambiae* genome) with bowtie, using the following options: bowtie -f -n 0 -e 80 -l 18 -a -m 5 –best –strata. Option '-n 0' kept only alignments with 0 mismatches in the seed region of a read mapped to the genome.

The seed region was defined by option '-l 18' which corresponded to the first 18 nucleotides of a read sequence. When using option '-n' it is possible to allow mismatches occurring after the seed region of a read in an alignment. This is determined by option '-e 80' and is the maximum sum of quality values at each mismatch position. The default quality value for each position in a sequence file was set to 40 which means that up to two mismatches were allowed in the region of a read after its seed region. Option '-m 5' keeps only reads that did not map more than five times to the genome. Option '–best –strata' ordered the mappings from best to worse alignments according to the strata definition of bowtie. If mappings with zero mismatches occurred then mappings with one or two mismatches were not reported. All the processed reads and the mappings to the genome were outputted in miRDeep2 format.

2.2.4.3 Small ncRNAs detection

In order to annotate the small ncRNAs present in the libraries, all reads were aligned to the Rfam database (Gardner et al., 2011) using the BLASTn algorithm which allows for a two-nucleotide mismatch and an *e-value* lower than 0.01.

2.2.4.4 Identification of known and novel miRNAs

For identification of known and novel miRNAs present in the four datasets, the miRDeep2.pl algorithm was used. The input files for miRDeep2 script were as following: (i) a FASTA file with deep sequencing reads from each stage; (ii) a FASTA file of the reference genome which is a file of mapped reads to the genome in miRDeep2 format, (iii) a FASTA file with known miRNAs of the *An. gambiae*; (iv) and, a FASTA file of known miRNAs in all species. After testing the format of all input files, a fast quantification of known miRNAs was done. Potential miRNA precursors were excised from

the genome using the genomic coordinates of the mapped miRNAs as guidelines. The genomic coordinates of the mapped miRNAs were first parsed such that only perfect mappings (no mismatches) of at least 18 nucleotides were retained. Furthermore, the genomic coordinates of the mapped miRNAs from reads that map perfectly in more than 5 loci in the genome were discarded. The two genome strands of each genome contig were scanned separately, from the 5′ to 3′ end. Excision was initiated when a stack of reads (height one or more) was encountered. However, if there was a higher read stack within 70 nucleotides downstream of the current read stack, then this was chosen instead. This downstream search was iterated until no higher read stack was found within 70 nucleotides. In this way, the highest local read stack was identified. The sequence covered by the highest local read stack was excised twice, once including 70 nucleotides upstream and 20 nucleotides downstream flanking sequence, and once including 20 nucleotides upstream and 70 nucleotides downstream flanking sequence. Subsequently, the genome scanning continue from the position one nucleotide downstream of the last excised sequence. If the total number of potential precursor sequences excised was less than 50000 (two precursors per genomic locus), then this set was output to the downstream analysis. If there were more sequences, then the entire excision step was repeated, with the height of the read stack necessary for initiating excision increased by one. The bowtie-build tool was used with default options to build a Burrows-Wheeler transform index of the excised potential precursors. The set of sequencing reads were mapped to the index, using bowtie with the following options: bowtie -f -v 1 -a –best –strata –norc. Option '-f' designates a FASTA file as input, option '-v 1' reports the genomic coordinates of the mapped miRNAs with up to one mismatch to the precursors, option '-a' lead to the report of all valid alignments, options '–best –strata' ordered the mappings from best to worse alignments according

to the strata definition of bowtie. If reads map perfectly to the precursors then mappings of the same read with one mismatch were not reported. Option '–norc' specifies that no reads must be mapped to the reverse complement of the precursor sequences in the bowtie index. The set of known mature miRNAs for the reference species was also mapped to the index, with the following options: bowtie -f -v 0 -a –best –strata –norc. Here the module did not allow any mismatches for the mappings because the mature miRNA sequence and the potential precursor sequences have not been subject to any source of noise. The two mapping files were concatenated and all lines were sorted according to the potential precursor ID's. The next step was to predict RNA secondary structures of the potential precursors. This was done with randfold with default options. Optionally, the randfold *P-values* for a subset of the potential precursors were calculated. This was done by selecting the potential precursors that (i) fold into an unbifurcated hairpin, (ii) can be partitioned into candidate mature, loop and star part based on the the genomic coordinates of the mapped miRNAs to it, and (iii) have minimum of 60% of the nucleotides in the candidate mature part base paired. The randfold *P-values* were calculated for the subset of potential precursors with these options: randfold -s 99. In the next step the potential precursors were individually scored or discarded by the miRDeep2 core algorithm. The core algorithm was run with these options: -s -v -50 -y. Option '-s' designates the reference mature miRNAs file in FASTA format as input, option '-v -50' keeps all precursors that have a miRDeep2 score above -50 and option '-y' supplies an additional file with randfold values. Furthermore, 100 rounds of permuted controls were performed as previously described, with the same options as the genuine run. The third step surveyed the score distributions of the genuine run and the control runs. The performance statistics were calculated for all score cut-offs from -10 to 10. The number of known miRNAs present

in the data was estimated as the number of known mature miRNAs that map perfectly to one or more excised potential precursors. The number of known miRNAs that were recovered was estimated as the number of known mature miRNAs that map perfectly to one or more hairpins that exceed the given score cut-off. The sensitivity of the run was estimated at se = (known miRNAs recovered)/(known miRNAs in data). The number of false-positives for a given score cut-off was estimated by the permuted controls. The fraction of true miRNAs reported was estimated by t = (novel miRNAs - estimated false-positive novel miRNAs)/novel miRNAs. The signal-to-noise ratio was estimated as n = total miRNAs/estimated total false-positive novel miRNAs (total miRNAs = novel miRNAs + known miRNAs) (Friedländer et al., 2012).

Finally, the miRDeep2 module integrated all these results in a .html file as well as a corresponding tab-separated file. The tab-separated file contained detailed information on every known and novel miRNA in the data. In the top of the .html file was a survey of miRDeep2 performance for varying score cut-offs. For each score cut-off the sensitivity and number of true positive novel miRNAs was estimated.

2.2.4.5 Differential expression of known miRNAs

Differentially expressed miRNAs can be naively determined by *log2(fold-change)* for the normalized reads. However, the two sample *t-test* was used for more effective determination of the miRNAs differentially expressed between two sequential life-stages (egg-larva, larva-pupa, pupa-adult). The Bonferroni error correction method was used as a control for multiple testing and the resultant adjusted *P-values* were assessed for its significance at alpha 0.05. All of the above mentioned analyses were computed using *gtools* and *stats* modules from the R programing language.

2.3 Results

2.3.1 Preprocessing of short reads

2.3.1.1 Sequence quality of the four libraries

Using deep sequencing technology, a total of 43.9, 43.6, 41.4 and 38.5 million raw reads were obtained from eggs, larvae, pupae, and adults, respectively (Table 2.1). More than 97% of the reads have Phred quality values (PQV) of 20. PQV has previously been reported to be an indicator of base call accuracy and therefore sequence quality (Ewing et al., 1998). Thus, a PQV of 20 indicates that there is a one in 100 chance that the base call is incorrect. The length distribution of the reads was significantly greater between 20-30 nucleotides (Figure 2.3). After filtering the impurities and reads of length smaller than 10 nucleotides, a total of 26.94 million, 23.20 million, 33.58 million, and 29.06 million, high-quality reads were retained in eggs, larvae, pupae, and adults, respectively (Table 2.1).

2.3.1.2 Mapping reads from the four libraries

The first step in the characterization of miRNAs in NGS data required the alignment of the high-quality reads to a reference genome. Mapping reads over the unmasked genome represents an unbiased option, allowing the detection of known and still undiscovered miRNAs. The total number of mapped reads constitutes only 13.24%, 31.04%, 12.26% and 17.11% of the total high-quality reads from eggs, larvae, pupae, and adults libraries, respectively (Table 2.1).

Table 2.1: Summary of small RNA sequencing data analysis for the four developmental stage libraries of *An. funestus s.s*

	Eggs	Larvae	Pupae	Adults
Total raw reads	43,925,802	43,620,467	41,445,526	38,526,050
PQV of 20 (%)	43,016,538 (97.93)	42,678,265 (97.84)	40,691,217 (98.18)	37,797,908 (98.11)
High-quality reads (%)	26,945,478 (61.34)	23,207,493 (53.20)	33,589,461 (81.04)	29,060,141 (75.42)
Mapped to genome(%)	5,819,985 (13.24)	7,618,770 (31.04)	4,060,802 (12.26)	4,685,610 (17.11)

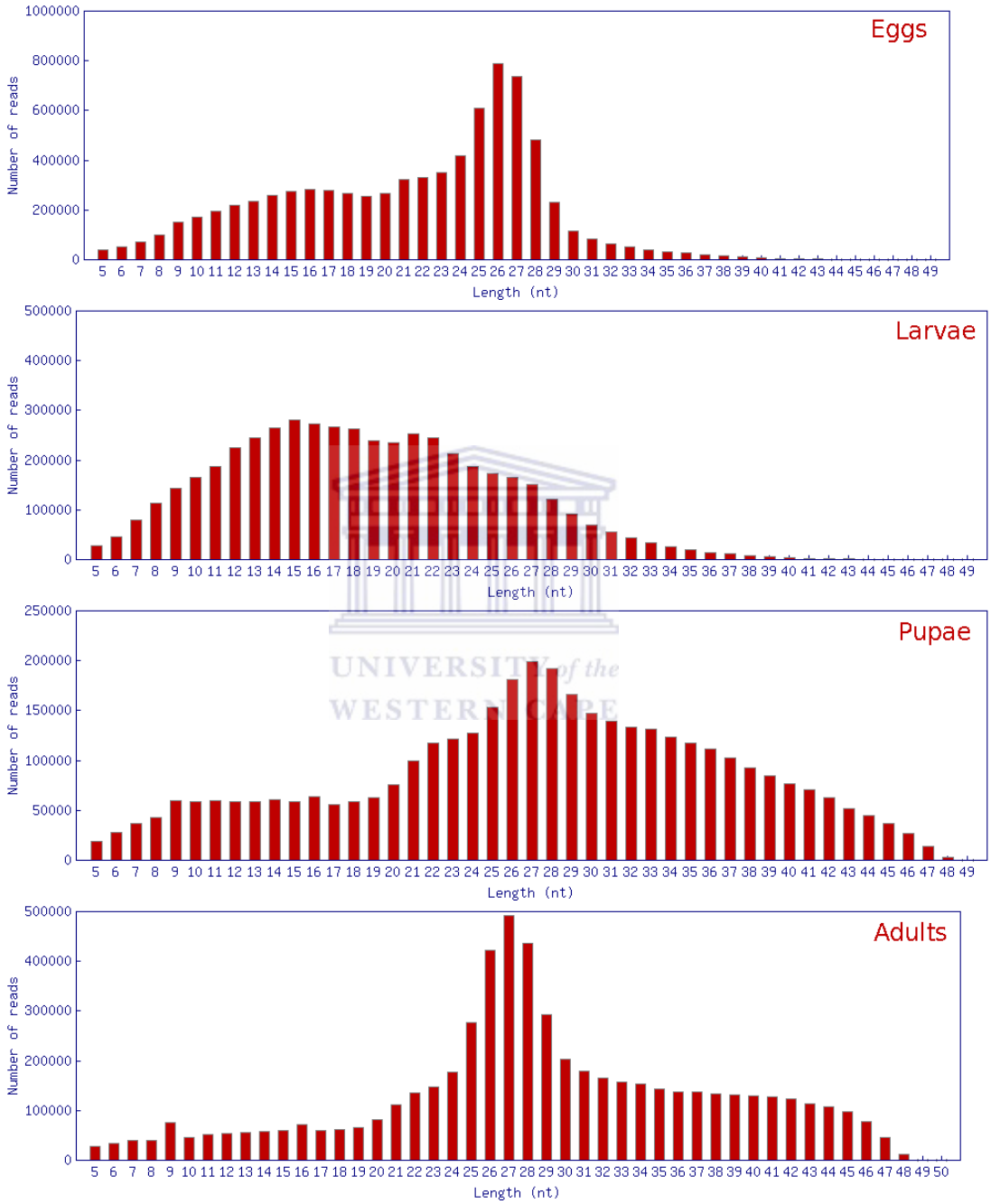


Figure 2.3: Length distribution of the raw reads from the four developmental stage libraries of *An. funestus s.s*. Among these reads, the number of 20-30 nucleotides reads was significantly greater than the other reads.

2.3.1.3 Annotation of small ncRNAs in the four libraries

In order to annotate other small ncRNAs in the four libraries, all clean reads were aligned against Rfam (v10.1). As shown in Figure 2.4, the most abundant class of small ncRNAs in the egg library were miRNAs. However, rRNAs were most abundant in the larvae and the pupae, and the tRNAs in the adult library.

2.3.2 Identification of known miRNAs in the four libraries

The 65 known mosquito miRNAs in miRBase were detected in the sequenced short reads (Table 2.2). For all characterized miRNAs, the full precursor structure (mature, loop and star sequence) or parts of it were detected.

2.3.2.1 Detection of miRNA isoforms

miRNAs frequently exhibit sequence differences from their reference mature sequence, generating multiple variations that are known as isoforms. With the aim of characterizing these variants, all sequences were aligned against miRBase. Two new isoform were identified (Table 2.3). The new isoform of the *miR-2* family (Figure 2.5, highlighted in grey) was highly expressed in the four libraries (eggs: 10143 rpm, larvae: 6893 rpm, pupae: 4376 rpm, and adults: 5142 rpm). Similarly, *Ae. aegypti* observed four member of the *miR-2* family (*aae-miR-2b*, *aae-miR-2a*, *aga-miR-13b* and *aae-miR-2c*). In contrast, this family consist of three members (*aga-miR-2-1*, *aga-miR-13b* and *aga-miR-2-2*) in *An. gambiae*. The new *miR-927* isoform had low expression levels in the four libraries (about 20 rpm per stage) and represent the first report of isoforms of the *miR-927* family in all mosquitoes genera.

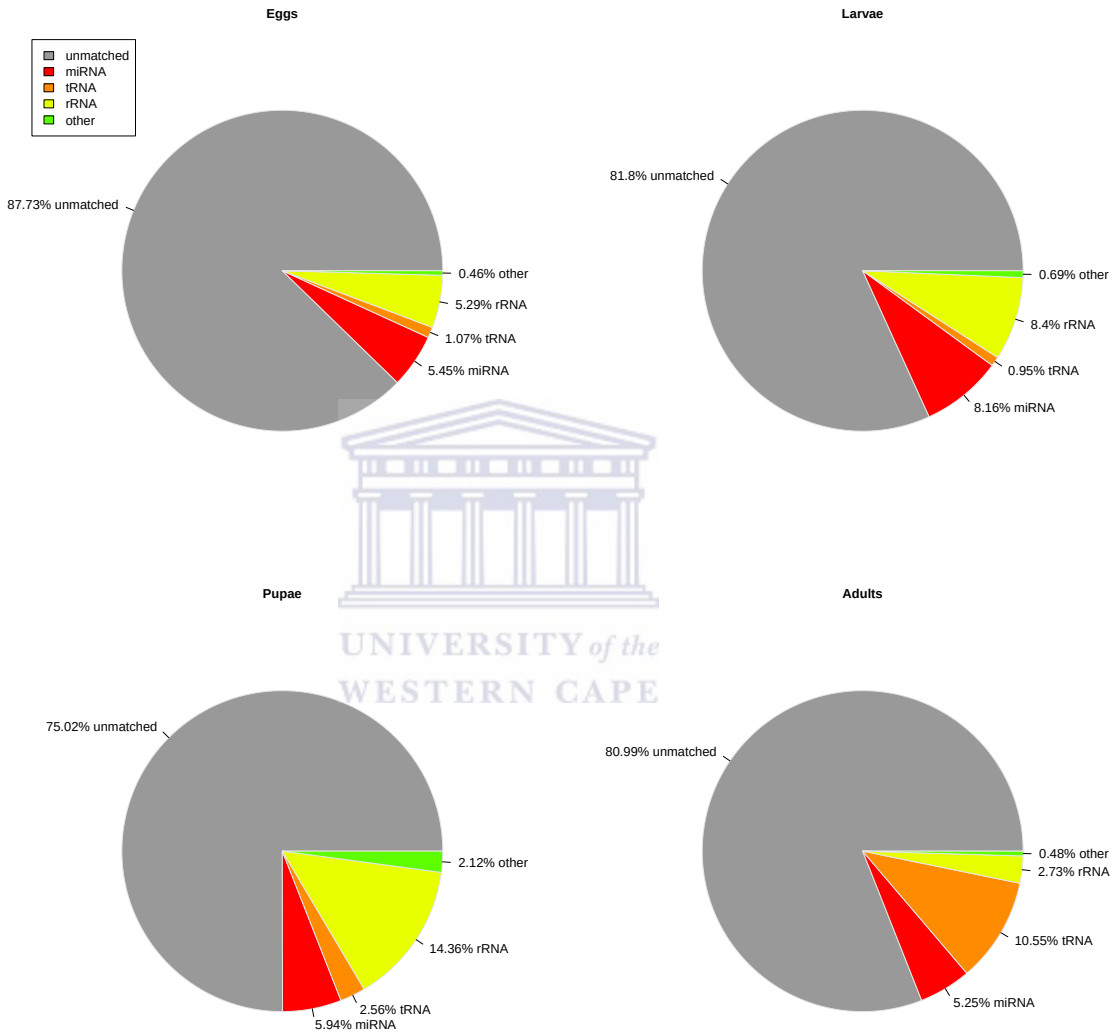


Figure 2.4: Small ncRNAs annotated from the four developmental stage libraries of *An. funestus s.s*. From aligning the clean reads to the Rfam database, the total reads can be divided into five categories; miRNAs, tRNA, rRNA, other and unmatched. The 'other' and 'unmatched' referred to the other class of small ncRNAs and remaining unannotated reads respectively.

2. MICRORNAS EXPRESSED IN *ANOPHELES FUNESTUS S.S* DEVELOPMENTAL STAGES

Table 2.2: The known miRNAs identified in the four developmental stage libraries of *An. funestus s.s*

miRNA name	Mature sequence	Length (nt)	Region in *An. gambiae*	Egg	Larva	Pupa	Adult
afu-bantam	UGAGAUCACUUUGAAAGCUGA	22	2R	√×√	√×√	√×√	√×√
afu-let-7	UGAGGUAGUUGGUUGUAUAG	21	3R	√×√	√×√	√×√	√×√
afu-miR-1	CCAUGCUUCCUUGCAUUCAAUA	23	3R	√×√	√××	√×√	√×√
afu-miR-10	ACCCUGUAGAUCCGAAUUUGU	22	2R	√×√	√×√	√×√	√×√
afu-miR-100	AACCCGUAGAUCCGAACUUG	21	3R	√×√	√√√	√×√	√×√
afu-miR-1000	AUAUUGUCCUGUCACAGCAGUA	23	2R	√×√	√×√	√×√	√×√
afu-miR-11	CAUCACAGUCUGAGUUCUUGC	22	2R	√××	√××	√××	√××
afu-miR-1174	UCAGAUCUACUUCAUACCCAUG	22	2R	√××	√××	√××	√××
afu-miR-1175	AAGUGGAGUAGUGGUCUCAUCGC	24	2R	√×√	√×√	√××	√×√
afu-miR-12	UGAGUAUUACAUCAGGUACUGGU	23	2R	√××	√××	√××	√××
afu-miR-124	UAAGGCACGCGGUGAAUGCCAA	23	3R	√××	√××	√××	√××
afu-miR-125	UCCCUGAGACCCUAACUUGUGA	23	3R	√×√	√×√	√√√	√×√
afu-miR-133	UUGGUCCCCUUCAACCAGC	20	3R	√××	√×√	√×√	√√√
afu-miR-137	UUAUUGCUUGAGAAUACACG	21	2L	√××	√×√	√×√	√×√
afu-miR-13b	UAUCACAGCCAUUUUGACGA	21	2L	√×√	√×√	√×√	√×√
afu-miR-14	UCAGUCUUUUUCUCUCUCCUA	22	3R	√×√	√×√	√×√	√×√
afu-miR-184	UGGACGGAGAACUGAUAAGGGC	23	3R	√×√	√×√	√×√	√×√
afu-miR-1889	ACACAUUACAGAUUGGGAUUA	21	2R	√××	√××	√××	√××
afu-miR-1890	UGAAAUCUUUGAUUAGGUC	20	3R	√××	√××	√××	√××
afu-miR-1891	UGAGGAGUUAAUUUGCGUGUU	22	3R	√×√	√×√	√×√	√×√
afu-miR-190	AGAUAUGUUUGAUAUUCUUGGU	23	2L	√×√	√×√	√×√	√×√
afu-miR-2	UAUCACAGCCAGCUUUGAUGAGC	24	2L	√√√	√√√	√√√	√×√
afu-miR-210	UUGUGCGUGUGACAACGGC	20	X	√××	√××	√××	√×√
afu-miR-219	AGAGUUGUGACUGGACAUCCG	21	2L	×××	√×√	√×√	√×√
afu-miR-263	AAUGGCACUGGAAGAAUUCACGGG	25	3R	√×√	√×√	√×√	√×√
afu-miR-263b	CUUGGCACUGGGAGAAUUCAC	21	2L	√××	√×√	√×√	√×√
afu-miR-275	UCAGGUACCUGAAGUAGCGCGC	23	3R	√×√	√×√	√×√	√×√
afu-miR-276-3p	AGCGAGGUAUAGAGUUCCUA	20	2L	√××	√××	√××	√××
afu-miR-276-5p	UAGGAACUUCAUACCGUGCUC	22	2L	√×√	√×√	√×√	√×√
afu-miR-277	UAAAUGCACUAUCUGGUACGACA	24	2R	√×√	√×√	√×√	√×√
afu-miR-278	ACGGACGAUAGUCUUCAACGACC	24	3L	√×√	√√√	√×√	√×√
afu-miR-279	UGACUAGAUCCACACUCAUUAA	23	2R	√××	√××	√××	√××
afu-miR-281	AAGAGAGCUAUCCGUCGACAG	22	2L	√√√	√×√	√×√	√√√
afu-miR-282	UAGCCUCUUCUAGGCUUUGUC	22	2L	√√√	√×√	√×√	√××
afu-miR-283	AAAUAUCAGCUGGUAAUUCUAGG	24	2R	√××	√×√	√×√	√×√
afu-miR-286	UGACUAGACCGAACACUCGCGUCC	25	3R	√√√	√×√	√×√	√×√
afu-miR-305	AUUGUACUUCAUCAGGUGCUC	22	3R	√×√	√×√	√×√	√×√
afu-miR-306	UCAGGUACUGGAUGACUCUCA	22	3R	√××	√××	√××	√××
afu-miR-307	UCACAACCUCCUUGAGUGAGCGA	24	3L	√××	√×√	√×√	√××
afu-miR-308	AAUCACAGGAGUAUACUG	19	3L	√××	√××	√××	√××
afu-miR-309	UCACUGGGCAAAGUUUGUCGCA	23	3R	√×√	×××	×××	√××
afu-miR-315	UUUUGAUUGUUGCUCAGAAAGCC	24	2R	√×√	√×√	√×√	√×√
afu-miR-317	UGAACACAUCUGGUGGUAUCUCAG	25	2R	√√√	√√√	√√√	√√√
afu-miR-34	UGGCAGUGUGGUUAGCUGGUUG	23	2R	√×√	√×√	√×√	√×√
afu-miR-375	UUUGUUCGUUUGGCUCGAGUUA	23	3R	√××	√×√	√×√	√××
afu-miR-7	UGGAAGACUAGUGAUUUUGUUGU	24	2L	√×√	√×√	√×√	√×√
afu-miR-79	AUAAAGCUAGAUUACCAAAGCA	23	3R	√×√	√×√	√×√	√×√
afu-miR-8	UAAUACUGUCAGGUAAAGAUGUC	24	3L	√√√	√√√	√√√	√√√
afu-miR-87	GUGAGCAAAUAUUCAGGUGUG	22	X	√××	√×√	√×√	√×√
afu-miR-927	UUUAGAAUUCCUACGCUUUACC	23	X	√×√	√×√	√×√	√√√
afu-miR-929	AAAUUGACUCUAGUAGGGAG	21	2R	×××	√×√	√×√	√×√
afu-miR-92a	UAUUGCACUUGUCCCGGCCUA	22	2R	√××	√××	√××	√××
afu-miR-92b	AAUUGCACUUGUCCCGGCCUGC	23	2R	√×√	√×√	√××	√××
afu-miR-957	GAAGCUCGUUUCUAUAGAGGUAUC	24	2R	√×√	√×√	√×√	√×√
afu-miR-965	UGAAACCGUCCAAAACUGAGGC	23	3L	√××	√××	√××	√××
afu-miR-970	UAAGCGUAUAGCUUUUCCCAUU	22	3R	√×√	√×√	√×√	√×√
afu-miR-981	UCAUAAGACACACGCGGCUA	21	X	√××	√××	√××	√××
afu-miR-988	UUCGUUGUCGACGAAACCUGCA	23	X	√××	√××	√××	√××
afu-miR-989	CCCCUUGUUGCAAACCUCACGC	22	2L	√××	√××	√××	√××
afu-miR-993	UGUGAUGUGACGUAGUGGUAC	21	3L	√××	√×√	√×√	√×√
afu-miR-996	UGACUAGAUUACAUGCUCGUC	22	2R	√××	√××	√××	√××
afu-miR-9a	UCUUUGGUUAUCUAGCUGUAUGA	24	2L	√×√	√√√	√×√	√×√
afu-miR-9b	ACUUUGGUGAUUUUAGCUGUAUG	23	3R	×××	×××	√××	√××
afu-miR-9c	UCUUUGGUAUUCUAGCUGUAGA	23	3R	√√√	√√√	√×√	√×√
afu-miR-iab-4	ACGUAUACUGAAUGUAUCCUGA	23	2R	√×√	√×√	√××	√××

The three columns for each development stage correspond to detection of the full precursor structure namely mature, loop and star sequence.
√√√ indicates that the full precursor structure was detected in the library.
××× indicates that the miRNA was not detected in the library.

Figure 2.5: Genomic organization of *miR-2* family in *An. funestus s.s*, *An. gambiae* and *Ae. aegypti*. Like *Ae. aegypti*, four members of this family were detected in *An. funestus s.s* (the new member highlighted in grey). *afu-miR-2-1* is orthologous to *aga-miR-2-1* and *aae-miR-2b*, *afu-miR-iso* is orthologous to *aae-miR-2a*, *afu-miR-13b* is orthologous to *aga-miR-13b* and *aae-miR-13b* and *afu-miR-2-2* is orthologous to *aga-miR-2-2* and *aae-miR-2c*. In *An. gambiae*, this family consist of only three members (*aga-miR-2-1*, *aga-miR-13b*, *aga-miR-2-2*). Numbers below miRNAs represent genomic coordinates. With the absence of an *An. funestus* genome, we used *An. gambiae* genome to determine the genomic organization of *miR-2* family.

2.3.3 Analysis of novel miRNAs

To identify novel miRNAs, we employed a combination of the miRDeep2 score and the randfold *P-value* to investigate if non-annotated sequences mapping to the *An. gambiae* genome demonstrated folding properties of pre-miRNA hairpins. A total of 33 candidates for new miRNA precursors were identified (Table 2.3). These novel candidates displayed a length distribution between 18-25 nucleotides with a peak at 22 nucleotides. Fifteen (46%) of these miRNAs were detected in all the stage libraries. There were 24 (73%) of the candidate precursors included reads for a mature miRNA and a putative miRNA star. The other 9 (27%) candidate precursors had reads only consistent with the predicted mature miRNA. There are 17 new miRNAs in the *Anopheles* genus and four are novel in mosquitoes. We also detected a new stem-loop for *miR-2944* and *mir-286*. Furthermore, our analysis uncovered eight new miRNAs that have not been described before in any species.

All identified miRNAs (novel or known) were named according to their most similar miRBase match. Finally, sequence data have been prepared for submission to the miRBase miRNA sequence repository.

2.3.4 miRNA expression profiles

The expression patterns of miRNAs provides clues of their functions (Yao et al., 2007). To obtain insight into possible stage-dependent roles of miRNAs in *An. funestus s.s*, the expression patterns of miRNAs in different developmental stages including egg, larva, pupa, and adult, were examined based on the number of reads obtained. The heatmaps (Figure 2.6) summarizes the expression of the known and novel miRNAs in

Table 2.3: Novel miRNAs identified in the four developmental stage libraries of *An. funestus s.s*.

miRNA name	miRNA with same seed	Length (nt)	Mismatch (nt)	Mature sequence	Region	Egg	Larva	Pupa	Adult	Description
afu-miR-2779	*bmo-miR-2779*	18	1	AUCCGGUUCGAAGGACCA	2L:25725096-25725149(-)	√	×	√	×	novel in mosquitoes
afu-miR-2944a-1	*aae-miR-2944a*	21	0	GAAGGAACUUCUGCUGUGAUC	2L:3537079-3537136(-)	×	√	√	√	novel in *Anopheles*
afu-miR-2944b-1	*aae-miR-2944b*	22	0	UAUCACAGCAGUAGUUACCUAC	2L:3537282-3537351(-)	×	√	×	√	novel in *Anopheles*
afu-miR-286b	*aga-miR-286b*	23	0	UGACUAGACCGAACACUCGUAUC	2L:3537482-3537547(-)	√	√	×	√	stem-loop
afu-miR-2-1a	*aga-miR-2-1*	24	0	UAUCACAGCCAGCUUUGAAGAGCG	2L:37757408-37757491(-)	√	√	√	√	isoform
aae-miR-71-3p	*aae-miR-71*	20	0	UCUCACUACCUUGUCUUUCA	2L:37759437-37759498(-)	×	√	√	×	novel in *Anopheles*
aae-miR-71-5p	*aae-miR-71*	20	0	UCUCACUACCUUGUCUUUCA	2L:37759478-37759568(-)	√	√	√	√	novel in *Anopheles*
afu-miR-2943	*aae-miR-2943*	22	0	UUAAGUAGGCACUUGCAGGCAA	2L:5667730-5667792(+)	√	×	×	×	novel in *Anopheles*
afu-miR-998	*cqu-miR-998*	21	0	UAGCACCAUGAGAUUCAGCUC	2R:13042402-13042473(+)	√	√	√	×	novel in *Anopheles*
afu-miR-a	-	18	-	GGCAUCCGGUCGUACGAC	2R:13285397-13285440(+)	√	√	×	×	novel
afu-miR-932	*cqu-miR-932*	22	0	UCAAUUCCGUAGUGCAUUGCAG	2R:15199832-15199899(+)	√	√	√	√	novel in *Anopheles*
afu-miR-971	*ame-miR-971*	23	2	UUGGUGUUAUAUCUUACAGUGAG	2R:15821042-15821118(-)	√	√	√	√	novel in mosquitoes
afu-miR-980	*cqu-miR-980*	22	0	UAGCUGCCUAGUGAAGGGCAAC	2R:21985895-21985955(-)	√	√	√	√	novel in *Anopheles*
afu-miR-2796	*ame-miR-2796*	23	0	GUAGGCCGGCGGAAACUACUUGC	2R:22427284-22427344(+)	√	√	√	√	novel in mosquitoes
afu-miR-252	*cqu-miR-252*	22	0	CUAAGUACUAGUGCCGCAGGAG	2R:52693316-52693371(-)	√	√	√	√	novel in *Anopheles*
afu-miR-3731	*ame-miR-3731*	23	3	CGAGAAUCUUUCGUCGAUGUCCG	2R:57041950-57042011(+)	×	√	×	×	novel in mosquitoes
afu-miR-999	*cqu-miR-999*	21	0	UGUUAACUGUAAGACUGUGUC	2R:8265180-8265240(-)	√	√	√	√	novel in *Anopheles*
afu-miR-b	-	22	-	AAAAGUUUUCAUAUUCUUGCGG	3L:17216221-17216289(+)	×	√	√	√	novel
afu-miR-c	-	18	-	GCCCGGUCGUACGCGGCA	3L:37467743-37467784(-)	√	×	×	√	novel
afu-miR-285	*cqu-miR-285*	22	0	UAGCACCAUUCGAAAUCAGUAC	3L:39242666-39242730(+)	√	√	√	√	novel in *Anopheles*
afu-miR-2942	*aae-miR-2942*	23	1	UAUUCGAGACCUUCACGAGUUAA	3L:39686859-39686931(+)	×	√	√	×	novel in *Anopheles*
afu-miR-2945	*aae-miR-2945*	20	0	UGACUAGAGGCAGACUCGUU	3R:10314333-10314397(-)	√	√	√	√	novel in *Anopheles*
afu-miR-d	-	20	-	CCUCCGGGCGCAUCUGACGG	3R:24095611-24095693(+)	√	×	×	√	novel
afu-miR-e	-	25	-	UGUCAUGCCAACGCUGCGCCAGUGC	3R:25647331-25647389(-)	√	×	×	√	novel
afu-miR-2944a-2	*aae-miR-2944a*	21	0	GAAGGAACUUCUGCUGUGAUC	3R:43008878-43008934(-)	√	√	√	√	stem-loop
afu-miR-2944b-2	*aae-miR-2944b*	24	1	UAUCACAGCGGUAGUUACCUGAUA	3R:43009063-43009127(-)	√	×	×	×	stem-loop
afu-miR-f	-	18	-	UAGUCGGCAGCCGGAACC	3R:43267769-43267843(-)	√	×	×	√	novel
afu-miR-g	-	21	-	UCCCCCGUAGCAAGGACUGAC	3R:46659716-46659758(-)	√	√	×	×	novel
afu-miR-h	-	25	-	UGUCCAAGUAGUCGUCCACGUAAUG	UNKN:2073561-2073635(-)	√	√	×	√	novel
afu-miR-33	*aae-miR-33*	22	3	CAGUACUUCUGCAAUGCAACCC	X:1256013-1256078(-)	√	√	√	√	novel in *Anopheles*
afu-miR-31	*cqu-miR-31*	23	1	UGGCAAGAUGUUGGCAUAGCUAA	X:14375923-14375980(+)	√	√	√	√	novel in *Anopheles*
afu-miR-927-5p	*aga-miR-927*	22	0	UAAAGCGUAGGAAUUCUAAAAC	X:18737380-18737443(-)	√	√	√	√	isoform
afu-miR-2940	*aae-miR-2940*	21	1	GUCGACAGAGAGAUAAAUCAC	X:2840460-2840525(-)	√	√	√	√	novel in *Anopheles*

√ indicates that the miRNA was detected in the library.
× indicates that the miRNA was not detected in the library.
novel in *Anopheles*= not described before in *Anopheles* species.
novel in mosquitoes= not described before in mosquito species.
novel= not described before in any species.
stem-loop= new miRNA stem-loop precursor.
isoform= new miRNA isoform or variant.

the four developmental stage libraries. The majority of miRNAs were sequenced between 1-10^6 times. Altogether, 64 known miRNAs were present at more than five rpm for at least one developmental stage. *miR-263* was the most frequent known miRNA in the eggs, the larvae, and the pupae libraries (egg $>$ 2e+05 rpm), and *miR-8* was the dominant known miRNA in the adults library ($>$ 1.5e+05 reads). Nevertheless, *miR-929* had the lower expression profile in the four libraries ($<$ 5 rpm). The expression profiles of miRNAs varied from highly specific to ubiquitous during the four developmental stages. Seven miRNAs were detected ubiquitously in all libraries with comparable expression (e.g.*miR-10*), while 10 miRNAs were expressed in only one (e.g. *miR-286*), four miRNAs in two libraries (e.g. *miR-92a*) and nine in three libraries (e.g. *miR-1890*). For the novel miRNAs, *miR-286b*, *miR-2944a-2*, and *miR-2944b-2* were the most expressed miRNAs respectively.

To analyze the changes in the expression of miRNAs during the mosquito development, all read counts within a dataset were normalized, then *log2(fold-changes)* and adjusted *t-test P-value* were calculated between the two stages (egg-larva, larva-pupa, and pupa-adult) (Appendix A, Figure 5). The results of the pairwise comparisons (egg-larva, larva-pupa and pupa-adult) of the expression level of all the known miRNAs in the four libraries are shown in Figure 2.7. From the egg to the larva, 19 miRNAs were up-regulated and 10 other miRNA were down-regulated. Three miRNAs were up-regulated and another nine down-regulated between the larva and pupa stage. However, between the pupa and adult stage, six miRNAs were up-regulated and three down-regulated.

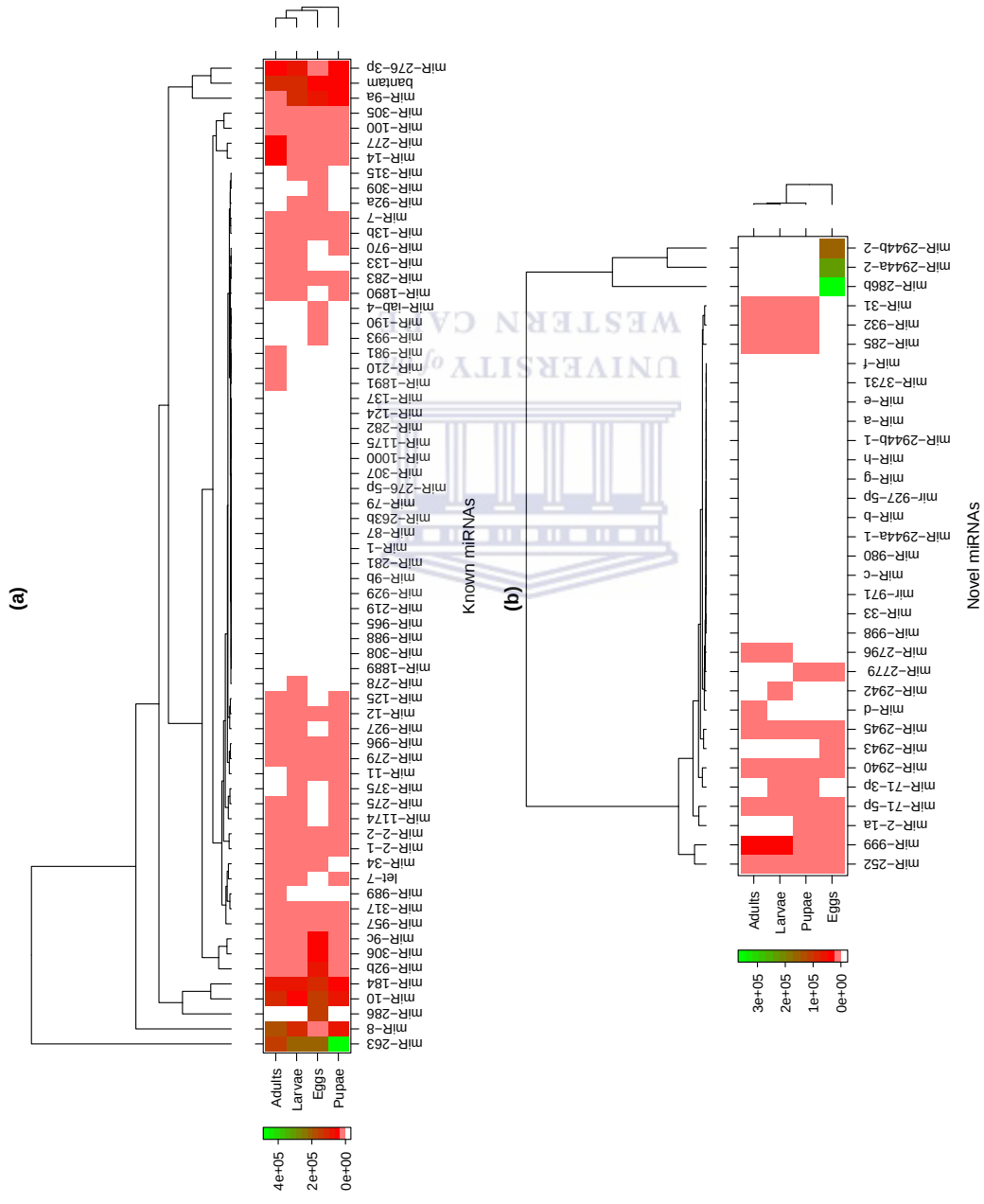


Figure 2.6: Heatmaps clustering of miRNAs expressed in the four developmental stage libraries of *An. funestus s.s*. The clustering was performed on all known (panel a) and novel (panel b) miRNAs based on a raw read copy number (sequencing frequency) and the four development stage samples. Each row represents a stage and each column represents one miRNA. The miRNA clustering is shown on top. The colour scale (shown on the left) illustrates the number of the reads of a miRNA across the developmental stages.

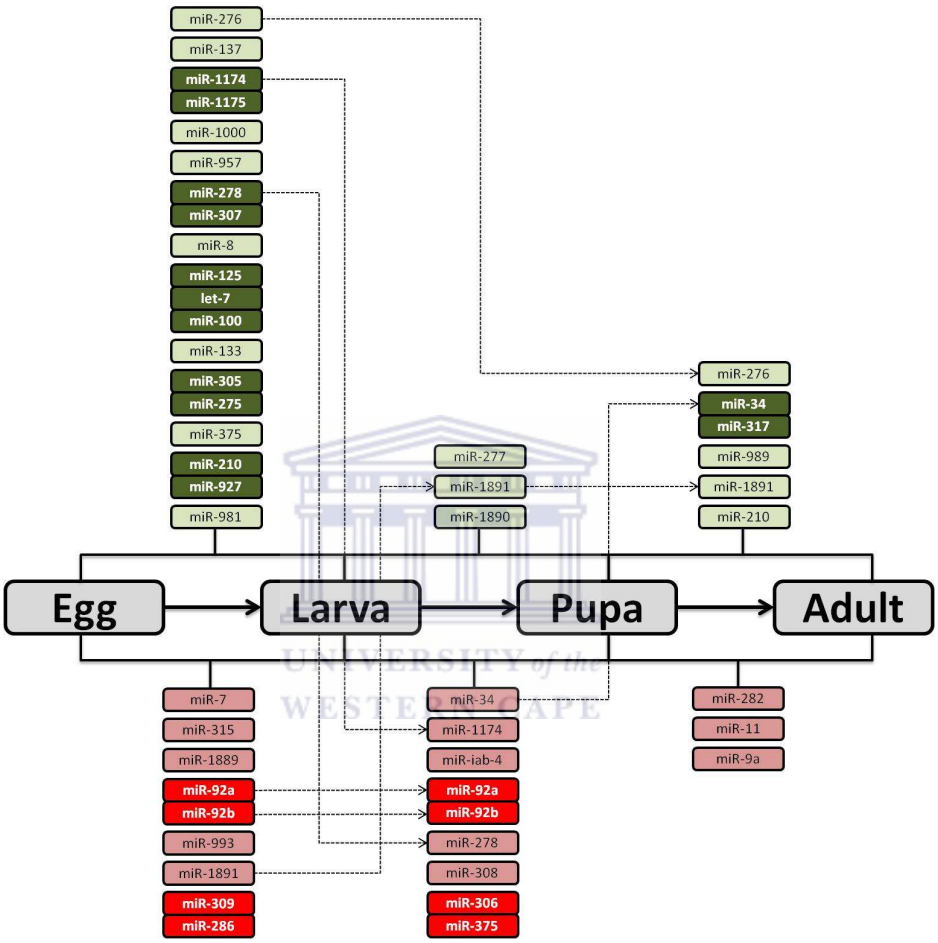


Figure 2.7: The dynamic changes in the 65 known miRNA expression profiles during the development of *An. funestus s.s*. The miRNA with significant expression patterns (*log2(fold-changes)* >2) across the four developmental stages (egg, larva, pupa and adult) are shown in boxes. Some miRNAs are significantly up-regulated (light green boxes) and/or down-regulated (light red boxes) between one, two, three or all four stages (indicated by the arrow). The miRNAs which are co-transcibed and co-expressed are highlighted in dark green or red. For example, *miR-1174* and *miR-1175* are up-regulated in the egg stage and *miR-92a* and *miR-92b* are down-regulated in the larvae and pupae stages.

2.4 Discussion

Although thousands of small RNAs have been identified in the recent past (Griffiths-Jones, 2006; Griffiths-Jones et al., 2006, 2008; Kozomara and Griffiths-Jones, 2011), the challenge remains to fully identify all small nuclear RNAs, especially very low abundant miRNAs and to determine their individual functions. The majority of known miRNAs have been identified through the traditional cloning method (Tian et al., 2010), which is however, both time-consuming and labour-intensive. The advantages of NGS technologies provides an innovative tool to look into the genome with an unprecedented depth of coverage. The Illumina sequencing approach is one of these high throughput technologies by which miRNAs in any organism can be detected without prior sequence or secondary structure information. This technology has been used to identify and detect miRNAs in multiple species (Burnside et al., 2008; Koh et al., 2010; Li et al., 2012; Inukai et al., 2012; Guo et al., 2012; Hao et al., 2012; Ji et al., 2012; Kang et al., 2012; Avesson et al., 2012). Currently, there are 65 *Anopheles* miRNAs predicted in the current miRBase release (Griffiths-Jones et al., 2008). Here, we confirmed the presence of the 65 known miRNAs, and 33 novel miRNAs using a high throughput sequencing approach. To our knowledge, this is the first profiling of mosquito miRNAs using deep sequencing approach. This approach is more powerful than other conventional technologies previously used in mosquitoes (Winter et al., 2007; Mead and Tu, 2008; Skalsky et al., 2010), as it is able to identify new miRNAs which are beyond the capabilities of older traditional methods.

In this study, approximately 20 million high-quality reads from each development stage were obtained by deep sequencing. The size distribution of sequenced reads peaked

between 20-30 nucleotides. The average miRNA length is 22 nucleotides in animals (Bartel, 2004). Using similar deep sequencing techniques, other studies on insects show small RNA sequence size distributions with a peak at 22 nucleotides (Li et al., 2009b; Skalsky et al., 2010; Liu et al., 2010b; Wei et al., 2009).

All the high-quality reads we obtained were mapped on to the *An. gambiae* genome with extremely low percentages. Such a mapping bias has been reported in several studies (Kozomara and Griffiths-Jones, 2011; Cordero et al., 2012; Oshlack and Wakefield, 2009; Pelaez et al., 2012). This approach has the weakness that it might favor alignment ambiguities due to the limited alignment specificity given by the small length of mature miRNAs (18-25 nucleotides) detected by the short reads, and to the size and high complexity of an unmasked reference genome (Cordero et al., 2012). Further studies are needed to understand to what extent the 24 nucleotides class representing siRNA populations (usually the most abundant and diverse class of small ncRNAs sequenced in small RNA libraries masks miRNA populations). Coverage analyses with fully sequenced genomes are needed to elucidate sequenced sample proportions of small ncRNAs such as tRNA, rRNA, snoRNA or snRNA (Cordero et al., 2012).

To further assess the efficiency of deep-sequencing in identifying small ncRNAs, all of the sequences were annotated by aligning the reads with rRNA, tRNA, miRNA, and other small RNA, deposited in Rfam databases. The results show clearly that there is a rich small RNA world in mosquitoes. Moreover, the changes in the expression profile of small ncRNAs demonstrate their roles to regulate the development, cell growth, and reproduction, in this insect.

2. MICRORNAS EXPRESSED IN *ANOPHELES FUNESTUS S.S* DEVELOPMENTAL STAGES

As expected, most of the known miRNAs identified in *An. funestus s.s* were highly conserved across diverse genera, suggesting that the ancient regulatory pathways mediated by evolutionary conserved miRNAs are present in mosquitoes. For some conserved miRNAs, the full stem-loop precursor structure (mature, loop, and star sequence) was detected in the four libraries (*miR-8* and *miR-317*) or in some libraries (*miR-2-1*, *miR-9a*, *miR-9a*, *miR-100*, *miR-125*, *miR-133*, *miR-278*, *miR-281*, *miR-282*, *miR-286* and *miR-927*). In vast majority of cases, mature miRNAs are more abundant than loop and star sequences. Additionally, *miR-219*, *miR-929* and *miR-9b* were not detected in the eggs library, *miR-309* was not found in larvae and pupae libraries. These result indicate that the expression of some miRNAs is probably tissue or developmental stage specific. It can be speculated that a miRNA may be involved in regulation of function and dysfunction, differentiation, growth and development of a specific stage (Tang et al., 2011).

The analysis of read numbers also revealed differential accumulation of *miR-2* and *miR-927* miRNA isoforms. The *miR-2* family is widespread in invertebrates, and it is the largest family of miRNAs in the model species *D. melanogaster* (15 members), *Bombyx mori* (*B. mori*) (9 members), *Apis mellifera* (*Ap. mellifera*) (5 members) and *Ae. aegypti* (4 members) (Griffiths-Jones et al., 2008). However, this the first detection of an isoform in *miR-927* family, the miRNA family only found in insect (Griffiths-Jones et al., 2008).

The identification of novel miRNAs is an imminent and challenging problem for the understanding of post-transcriptional gene regulation. The characteristic hairpin struc-

ture of miRNA precursors can be used to predict novel miRNAs. With this feature and using miRDeep2, we predicted novel miRNAs by exploring the secondary structure, the Dicer cleavage site, and the minimum free energy of the unannotated small RNA tags that could be mapped to the *An. gambiae* genome. Based on our analysis, 33 sequences were obtained and were regarded as novel miRNA candidates. These novel candidates displayed a concentrated length distribution between 18 nucleotides and 25 nucleotides, with a peak at 22 nucleotides. We identified the precursor sequences for some of the novel miRNAs. We believe that the detection of the miRNA star is a strong clue, albeit not infallible, for the formation of precursor hairpin structures. This adds weight to the authenticity of the predicted candidates (Fahlgren et al., 2007; Sunkar et al., 2008). However, the evolution and function of the star miRNAs remains unclear. Two studies proposed that these miRNAs might differ from their sense partners by acting on different mRNA targets (Zhang et al., 2008; Chi et al., 2011).

We searched the miRBase databases for homologs to determine whether these novel miRNAs are conserved among other animal species. This search indicated that most of the candidates (22 (67%)) are conserved in other insect species but not the *Anopheles* genus, suggesting that these are insect-specific miRNAs. Among these new *Anopheles* miRNAs, we detected the four *Aedes* mosquito specific miRNAs (*miR-2940*, *miR-2942*, *miR-2943* and *miR-2945*) (Li et al., 2009b). This is the first description of these four miRNAs in the anopheline species.

Secondary structure prediction analysis of some reads resulted in the identification of new stem-loop precursors. The miRNAs identified with a precursor: *miR-286* and *miR-2944*. The new predicted stem-loop precursor for *miR-286* is found in a different chromosomal location and produced different mature miRNA (2 mismatches). These

results are congruent with the existence of a multi stem-loop precursor for this miRNA in other insects such as *Drosophila* and *Aedes* (Griffiths-Jones et al., 2006). Moreover, we characterize four miRNA precursors producing two forms of mature *miR-2944* (*miR-2944a* and *miR-2944b*). Surprisingly, the precursors for the two miRNAs (*miR-286* and *miR-2944*) are located very close to each other (about 131 bp). The remaining new miRNAs lacks a seed homology to any known miRNAs in insect or other animal species.

Counting redundant miRNA reads revealed that expression varies significantly among different miRNAs. In the three pre-adult stages, insect-specific miRNA (*miR-263*) was found to be the most abundant miRNA. However, *miR-8* was the common miRNA in the adult library. Furthermore, the four libraries shared five out of the top ten most frequently occurring miRNAs: *miR-263, miR-10, miR-184, bantam* and *miR-281*. In *Drosophila*, miRNA *miR-263* confers robustness during development by protecting nascent sense organs from apoptosis (Hilgers et al., 2010). Moreover, in a very recent study on one of the rice pests (*Nilaparvata lugens*), *miR-263* was found as high abundant miRNAs in the last instar female nymph females (Chen et al., 2012). Both *miR-8* and *miR-184* were reported in the embryos of the *Drosophila* (Li et al., 2011; Iovino et al., 2009), mosquitoes (Li et al., 2009b; Zheng et al., 2010), silk worm (Liu et al., 2009), the Japanese chistosoma(Huang et al., 2009), fish (Flynt et al., 2009; Xia et al., 2011), mouse (Wenguang et al., 2007; Juhila et al., 2011; Wu et al., 2011a) and humans (Hyun et al., 2009; Vallejo et al., 2011). This suggest a conservative developmental function for these two miRNAs across different animal.

The expression of miRNAs varies in different developmental stages (Xu et al., 2006;

Liu et al., 2009). Among the up-regulated miRNA between the egg and larva stage is *miR-133* and *miR-278*. In *D. melanogaster* embryos, *miR-133* plays a key role in controlling alternative splicing during muscle formation, and defining the properties of differentiated muscle cells (Boutz et al., 2007). However, over-expression of *miR-278* promotes tissue growth in the eye and the wings in the *Drosophila* (Teleman et al., 2006). In the gregarious phase of locust, canonical miRNAs were expressed at levels between 1.5 and 2-fold higher than in the solitary phase. The most prominent differences were found in *miR-276*, *miR-125*, and *miR-315* (Wei et al., 2009). Interestingly, we observed the same change in these miRNAs between the egg and the larva stage in *An. funestus s.s*. As is known, the genes that encode for miRNA are distributed across the chromosome either individually, or in clusters in which two or more miRNA genes are located within a short distance on the same segment of a chromosome (Altuvia et al., 2005; Zhang et al., 2012b). Therefore, it is assumed that miRNA genes located in a gene cluster are first transcribed as a single primary transcript that is subsequently processed to generate the individual miRNAs (Winter et al., 2007). In the larvae sample, we observed very high relative expression levels of the *let-7*-complex locus (*let-7*, *miR-100*, and *miR-125*), the cluster of the two mosquito-specifc miRNAs; *miR-1174* and *miR-1175*, *miR-278* and *miR-307*, *miR-305* and *miR-275*, *miR-210* and *miR-927*. We also observed the down-regulation of the *miR-92* cluster (*miR-92a* and *miR-92a*) and *miR-309* cluster include *miR-309* and *miR-286*, respectively. Like *Drosophila*, *miR-309* and *miR-286* (which was processed from a single 1.5 kb primary transcript) displayed a dynamic pattern of expression in the early embryo.

In the pupae sample, only *miR-1891, miR-277* and *miR-1890* up-regulated. With some other miRNAs, *miR-278* and *miR-1174* down-regulated in the pupa after increasing in the larva stage. Again, we noticed another down-regulation of the *miR-92a* cluster.

The expression of this cluster members has been related to the embryonic development in *Ae. aegypti* (Li et al., 2009b) and *B. mori* (Liu et al., 2010b).

In the adult sample, more of *miR-1891* were expressed. Interestingly, this mosquito-specific miRNA was the most miRNA that displayed changes in its expression levels in this study, suggesting a significant role for this miRNA during its development beside its function in the host response to infection (Hussain et al., 2011). In our adult female library, we noticed up-regulation of *miR-34* coming after a significant down-regulation in the pupae stage, and a major change in *miR-989* expression for the first time. The expression of these miRNAs has been studied in different mosquito species, and results have shown that *miR-989* expression is restricted to females, and predominantly to the ovary in *An. stephensi* and *Ae. aegypti* (Mead and Tu, 2008; Li et al., 2009b), although it was later detected in the midgut of *An. gambiae* (Winter et al., 2007). In addition to all this, both miRNAs (*miR-34* and *miR-989*) with two other miRNAs displayed changes in the expression levels during the parasite invasion (Winter et al., 2007). On the other hand, and for unknown reasons, some miRNAs such as the arthropod-specific *mir-929* showed low read counts. Further studies are needed to reveal the function of such miRNAs.

The predicted novel miRNAs exhibited much lower expression levels, consistent with the evidence that non-conserved miRNAs are often expressed at a lower level than conserved miRNAs (Ling et al., 2011; Inukai et al., 2012; Zhang et al., 2012a). The low abundance of novel miRNAs might suggest a specific role for these miRNAs under various growth conditions or during developmental stages.

2.5 Conclusion

Overall, these results present the first direct experimental evidence of miRNAs in *An. funestus s. s*. Our data revealed 33 new miRNAs in this species. This discovery expanded the list of *Anopheles* mosquitoes miRNAs from 65 to 98. This identification and characterization of these miRNAs expands the current repertoire of *Anopheles* mosquito miRNAs. In addition, the expression profiles of miRNAs suggest stage-specific functions as well as functions related to embryonic development. These results indicate that miRNAs play important roles in growth and development in this species. Silencing such molecules in a specific developmental stage could decrease the vector population and therefore interrupt malaria transmission. Moreover, further studies on their target genes and biological functions may offer new insights in mosquito biology and may lead to novel approaches to the development of insecticides.

Chapter 3

InsecTar: a database for microRNA target genes in insects

Abstract

Many studies showed that miRNAs play an essential roles in gene regulatory networks by controlling the expression of genes involved in important biological processes. In insects, thousands of miRNA genes have been identified, but the function of most of these miRNAs remain unknown due to the lack of experimental and computational approaches to predict their exact target mRNAs. We introduce an integrated database for miRNAs targets in insects, InsecTar. The InsecTar database architecture comprise a MySQL relational database backend and a web front, developed using Perl-CGI. The database incorporates target prediction and functional analysis. The miRNA function is inferred from the overlap target genes predicted by two or three different prediction programs. Statistical enrichment tests for gene ontology and biological pathways of target genes is performed. The database provide an automated framework to update targets for new miRNAs as they are sequenced. InsecTar was used for prediction of known *Anopheles* and novel miRNA targets and their functional annotation. Our data suggests that mosquito miRNAs may play an important role by regulating: cell proliferation (*miR-2* and *miR-13b*), cell death (*miR-bantam*), fertilization (*miR-275*), growth (*let-7*), metabolism (*miR-277*), hospitality (*miR-2490*), and immunity (*miR-989*). These findings improve our understanding of how miRNAs function in this mosquito.

3.1 Introduction

The discovery of miRNAs dramatically changed our perspective of gene expression regulation. The most miRNAs function by repressing the translation of their target mRNA, by promoting mRNA transcript cleavage and degradation (Bartel, 2009). Many experimental and computational approaches have been developed to predict miRNA target genes (Zhang et al., 2006b). However, one major issue facing miRNA research today is the lack of an efficient method to biologically identify relevant miRNA targets. Gene-specific experimental validation with the well-established techniques of quantitative reverse transcriptase PCR (qRT-PCR), luciferase reporter assays, and western blot are commonly used to indicate individual miRNA:mRNA interactions (Kuhn et al., 2008; Thomson et al., 2011). Generally, the downstream effects of differential miRNA expression are observed at the protein level by western blot, and, at the mRNA level by qRT-PCR, although these measures cannot distinguish between direct and secondary miRNA targets (Thomson et al., 2011). Reporter assays have been employed extensively to demonstrate a direct link whereby expression of a luciferase reporter-3′UTR construct will be altered through the manipulation of a regulatory miRNA (Thomson et al., 2011). Direct miRNA effects are demonstrated by the loss of regulation in constructs with mutated miRNA target sites. The disadvantages of reporter assays are that they are labour intensive, dependent upon the region chosen for cloning, and can be sensitive to variances in protocol such as the method of transfection or promoter identity (Lytle et al., 2007; Kong et al., 2008; Hendrickson et al., 2009). Experimental studies alone are too slow and limited in scope to be relied on as the only source of miRNA target identification. In order to facilitate a deeper investigation into miRNA function, numerous computational algorithms have been developed for target

prediction. These span a wide range of approaches and techniques. They include miRanda (Enright et al., 2003), DIANA-microT (Kiriakidou et al., 2004), RNAhybrid (Rehmsmeier et al., 2004), MovingTargets (Burgler and Macdonald, 2005), PicTar (Krek et al., 2005), TargetBoost (Saetrom et al., 2005), miTarget (Kim et al., 2006), rna22 (Miranda et al., 2006), NBmiRTar (Yousef et al., 2007), PITA (Kertesz et al., 2007), TargetScan (Grimson et al., 2007), miRtif (Yang et al., 2008), MirTarget (Wang and El Naqa, 2008), MTar (Chandra et al., 2010), miR-TRAP (Baigude et al., 2012). All these algorithms make predictions based mainly on specific features of miRNA-target nucleotide sequence interaction. Although different algorithms utilize different sets of features, a few important features including seed region complementary, binding free energy and sequence conservation are among the most common ones (none of the existing prediction tools has been able to incorporate all currently known features) (Saito and Saetrom, 2010). A central goal of various algorithms concerns the selection of the most discriminative features that can lead to better prediction accuracy, since using different features will result in different prediction performance (Liu et al., 2010a). Despite the success of the above-mentioned methods in predicting functional target sites, the number of false positives remains high (Brodersen and Voinnet, 2009; Lindow and Gorodkin, 2007; Mendes et al., 2009). Most of the miRNA target databases available in the public domain are simple collections of miRNA-related information such as miRNA itself and target binding. Analyzing properties of miRNA targets is a promising approach to predicting miRNA functions (Ulitsky et al., 2010). As the number of validated targets is currently limited, methods for target-based inference of a miRNA function must rely on these predictions. If the targets of a specific miRNA are enriched with genes annotated with some biological process or pathway, it is reasonable to infer that the miRNA is involved in the same process.

3. INSECTAR: A DATABASE FOR MICRORNA TARGET GENES IN INSECTS

Insects are among the largest and most diverse groups of animals on the planet and include more than one million described species. They represent more than half of all known living organisms. Insects are also important human disease vectors and agriculture pests (Doane, 1910). Despite their importance, miRNA research in insects lag behind mammals and nematodes (Zhang et al., 2009). The current version of the miRNA database (miRBase) contains over 3000 miRNA genes from various insect species including fruit flies (Lai et al., 2003; Stark et al., 2007), the three mosquito vectors; *Anopheles gambiae*, *Aedes aegypti* and *Culex quinquefasciatus* (Li et al., 2009b; Skalsky et al., 2010; Winter et al., 2007), Western honey bee (*Apis mellifera*) (Weaver et al., 2007), silkworm (*Bombyx mori*) (Cao et al., 2008; He et al., 2008; Liu et al., 2010b), *Acyrthosiphon pisum*, *Locusta migratoria* and many more insects (Kozomara and Griffiths-Jones, 2011). Unfortunately, the biological functions of most of these miRNAs remain unknown. A rather mechanistic understanding of miRNA functions has relied on the properties of a single key target gene in a regulatory or signaling pathway (Nam et al., 2008). However, since a miRNA can target several hundred genes on average (Lewis et al., 2005), it is possible that individual miRNAs modulate the expression of gene clusters that often have related functions (Carè et al., 2007; van Rooij et al., 2008). Even though it remains to be determined whether or not simultaneous targeting of related genes is the norm and provides a coordinated control mechanism as is often seen in transcriptional regulation, the identification of functional relationships among target genes certainly provides valuable information in exploring the functional significance of each miRNA.

In this chapter, we present an integrated database for the functional annotation of miRNA targets in insects. Insect miRNA Targets (InsecTar) provides a tool for the

statistical enrichment test of miRNA targets in a number of functional annotation categories such as the GO terms and KEGG pathway. Moreover, The database provide information for the miRNA target genes that were identified in chapter 2.

3.2 Material and Methods

3.2.1 InsecTar pipeline

The analysis pipeline is implemented and automatized using Perl Programming Language (Figure 3.1). In the first phase, each miRNA is scanned for targets using three existing algorithms (miRanda, RNAhybrid, and MicroTar). In the second phase, the targets for each miRNA are tested for enriched GO terms and KEGG pathways.

3.2.1.1 Identification of miRNA targets

In order to predict the putative target genes, all known mature miRNA sequences were retrieved from miRBase release 18.0 and stored in a Structured Query Language (SQL). To obtain the 3′UTR sequences and relative appropriate information, the Ensembl Core database (species-specific database) for each insect was downloaded from Ensembl (http://metazoa.ensembl.org) and installed locally. Perl programming language with BioPerl modules and Application Programme Interface (API) provides efficient access to tables within the Ensembl Core database. Therefore, a script was written using a set of code from Ensembl Core API documentation (Appdendix B, script:1). In the first step, the script connects to the Ensembl local database using the database interface module (DBI). Since the Ensembl database API allows for the manipulation of the database data through objects, the three objects; Slice, Gene, and

Transcript, and their methods were used to get the 3′UTR sequence for each gene in the genome. Finally, both miRNA sequence and 3′UTR FASTA format files were provided as the input to three algorithms; miRanda (Enright et al., 2003), RNAhybrid (Rehmsmeier et al., 2004) and MicroTAR (Thadani and Tammi, 2006) to predict target mRNA. For this purpose, in-house Perl scripts were used to predict and store the prediction results (Appdendix B, script:2 and script:3).

miRanda identifies the potential target sites using dynamic programming alignment which is carried out between the miRNA and the 3′UTR sequences. This alignment procedure generates scores based on sequence complementarity and not sequence identity at the 5′ end of the miRNA. The second phase of the algorithm takes high-scoring alignments detected from the first phase and estimates the thermodynamic stability of RNA duplexes based on these alignments using the Vienna RNA folding package. Target genes predicted by miRanda had a cut-off score of 80 and minimum energy $\leq$ -14 kcal/mol (Enright et al., 2003).

RNAHybrid identifies regions in the 3′UTR that have the potential to form a thermodynamically favourable duplex with a specific miRNA. The maximum free energy of a miRNA is calculated for every 3′UTR of a set of shuffled 3′UTR sequences with maintained dinucleotide frequencies. Normalisation for both 3′UTR and miRNA length using S_{norm} = log(S/mn) is applied to these energies. Random energies derived in this manner should exhibit an extreme value distribution (EVD). Subsequently, the parameters of the EVD that best describe the data for a given miRNA are empirically calculated using the derived distribution from shuffled sequences. Each hit to any 3′UTR for this miRNA is then assigned a *P-value* calculated directly from these pa-

rameters. Hence, at the scanning stage, miRNAs are scanned against a database of 3′UTRs, and each hit is compared with the expected distribution and assigned a *P-value*. The statistical model implemented in RNAHybrid takes into account multiple sites and conserved sites by combining individual *P-values* using Poisson statistics and calculating conservative *P-values* for conserved sites. A statistical approach that fits, corrects highly conserved 3′UTRs by evaluating the overall conservation in the group of sequences compared with the conservation at the site. The resulting statistics cover individual site quality, quantity of sites, whether they are conserved and how significant this conservation is, given the input sequences.

MicroTAR uses predicted free energies of unbound mRNA and putative mRNA-miRNA heterodimers, implicitly addressing the accessibility of the mRNA 3′UTR. The algorithm does not rely on evolutionary conservation to discern functional targets, and is able to predict both conserved and non-conserved targets.

For each target predicted by miRanda, the hit score, thermodynamic hybridization energy, hit position in the 3′UTR, and seed match were captured. For RNAhybrid, the *P-value*, free energy and hit position were captured. For MicroTAR, the dimer-monomer difference, negative normalized free energy and hit position were captured. The predicted targets for each algorithms were then uploaded into a SQL database.

3.2.1.2 Functional enrichment analysis

Functional enrichment analysis of the target genes was performed by a package written in the R programming language called CORNA (Wu and Watson, 2009). CORNA uses a hypergeometric test to search for significant enrichment of GO terms or KEGG

Figure 3.1: Schematic overview of the InsecTar system. Both the miRNA sequence and 3′UTR FASTA formated files are scanned for targets using three algorithms; miRanda, RNAhybrid, and MicroTAR. A number of parameters are captured in the SQL database: for genes predicted by miRanda, the complementarity score, thermodynamic hybridization energy, hit position in the 3′UTR and the seed match. For RNAhybrid, the *P-value* for the prediction parameters, free energy and hit position. For MicroTAR, dimer-monomer difference, negative normalized free energy, and hit position. All target genes are enriched for functional association using GO terms and KEGG pathways. Finally, using InsecTar (http://insectar.sanbi.ac.za), the user can explore all of the above mentioned information.

pathways. To perform the test, one needs to define a gene universe and select a list of specific genes from the universe. With CORNA, all genes in the target genome that have at least one miRNA association, are used as the population. Thus, when using predicted miRNA targets to test for enrichment of pathways or GO terms, only those genes with at least one pathway, or at least one GO term, are used as the population. We used cut-off *P-value* of $\leq$0.05 for each enrichment. Both R scripts for the enrichment analysis using GO terms or KEGG pathways can be found in Appendix B.

3.3 Results

3.3.1 miRNA targets prediction

For the three vector (*A. aegypti*, *An. gambiae*, and *C. quinquefasciatus*), we scanned all miRNAs against the 3′UTRs which is available in Ensembl using the three algorithms (Table 3.1). More than 5,000 putative targets were identified for the 65 *Anopheles* miRNA by each algorithm. All of these targets have passed the prediction filters. All results are available online (http://insectar.sanbi.ac.za).

3.3.2 InsecTar: user interface

We developed a website (http://insectar.sanbi.ac.za) that uses Perl DBI and Common Gateway Interface (CGI) to make the database available for public use. Since all calculations are pre-computed and stored in a MySQL database, the Perl-CGI script communicates with the database and then retrieves and displays data relevant to the user's query. The user first selects a species, then the search is initiated either with a miRNA

Table 3.1: Summary of miRNA targets analysis

Insect	No. of miRNAs	No. of genes	No. of 3′UTRs	miRanda targets	RNAhybrid targets	MicroTAR targets
An. gambiae	65	13320	5504	5018	5239	5482
C. quinquefasciatus	93	19555	2392	2134	2317	2245
Ae. aegypti	125	14471	6893	6346	6757	7237

name or with a gene ID (Ensembl ID). A snapshot of the graphical user interface to the database is provided in Figure 3.2. The website is designed to allow the user to query the database for miRNA targets or miRNAs which may target a gene (Figure 3.3).

3.3.2.1 Search using a miRNA name

A query search of the database using a miRNA name returns a page display with five tabs. The first tab shows a table that contains information about the miRNA itself such as the miRNA accession number in miRBase, genomic location, and the mature sequence. The next three tabs contains the top 500 targets predicted by miRanda, RNAhybrid, and MicroTAR respectively. For a more effective prediction, we also added a tab detailing target genes predicted by two or all three algorithms.

Each tab consists of three tables: the ”targets table”, and two ”enrichment tables”. The target table shows details about predicted target genes. In the case of miRanda, the genes are ranked based on the score that was generated from the alignment between the miRNA sequence and the 3′UTR sequence. Targets predicted by RNAhybrid are ranked based on the *P-value* that were calculated from the prediction parameters. In MicroTAR, the targets are ranked according to the normalized free energy. In addition, the table displays the genomic location of the target in the genome, the free energy, seed match, and the hit position in the 3′UTR sequence. And the targets with multiple binding sites are highlighted in yellow. Clicking on each target ID leads to a new page which displays Ensembl information of the target such as the Ensembl gene ID, Ensembl transcripts ID, description, genomic location, biotype, InterPro ID, Enterz gene ID, GO ID, KEGG pathway, WikiGene ID, ortholog genes and the 3′UTR sequence for each transcript.

The enrichment tables show the significant pathways or GO term calculated from

Figure 3.2: InsecTar home page. The home page is accessible via the InsecTar website (http://insectar.sanbi.ac.za) and it displays information about the database and the main menu. The main menu consists of the **Home** page; the **Search** page where the user can search the database; the **Help** page which gives documentation about how to use the database; the **Links** page which shows the most popular miRNA targets prediction tools; and, the **Team** page where the collaborators and sponsors of the database are listed.

Figure 3.3: InsecTar search page. The search page contains tabs for each insect database. Each tab has two search options: a search using miRNA name (drop-down menu) or typing Ensembl gene ID.

CORNA for predicted target genes. The targets genes which shared the same KEGG pathway or GO term were grouped together in the same row. Also the pathway name, the sub pathway name, the main pathway name, and the hypergeometric test *P-value* is reported in case of enrichment using a KEGG pathway, and the GO term, GO description and the hypergeometric test *P-value* in case of enrichment using a GO term. Furthermore, by clicking on the pathway or the GO ID, the user can view the mapping of the predicted target(s) in the KEGG pathway or the GO term details respectively. In addition, the user can visualize the graph for enriched GO terms using the AmiGO visualization tool by clicking on the link following the GO enrichment table.

At present, InsecTar contains prediction and functional information for: 65 *An. gambiae*, 125 *Ae. aegypti* and 93 *C. quinquefasciatus* miRNAs respectively.

3.3.2.2 Search using an Ensembl gene ID

If the gene has a 3′UTR sequence annotated by Ensembl, the user can query the database using a Ensemble gene ID. The results page will show two tabs. The first tab displays information in tabular format. It describes the gene, including the Ensembl gene ID, Ensembl transcripts ID, description, genomic location, biotype, InterPro ID, Enterz gene ID, GO ID, KEGG pathway, WikiGene ID, ortholog genes, and the 3′UTR sequence for each transcript. The second tab shows a list of miRNAs that are predicted to target this gene by two or three algorithms.

3.3.3 Functional characterization of *Anopheles* miRNA targets

To understand the biological functions of the miRNAs in the *Anopheles* mosquito, we searched for putative target genes using the InsecTar system. Generally, the putative target genes appeared to be involved in a broad range of biological pathways (Figure 3.4). The heatmap represents the analysis of enrichment using KEGG pathways for targets predicted by two or three methods. Colour-coded values correspond to absolute *P-values* from the hypergeometric test. The KEGG pathways enrichment showed that many of the miRNAs target genes are involved in several metabolic networks including lipid metabolism, amino acid metabolism, carbohydrate metabolism, energy metabolism, nitrogen metabolism, and signal transduction. GO molecular function analysis demonstrated that miRNA targets are involved in 371 different molecular functions. Most of the significantly over-represented genes have binding or catalytic activities (Figure 3.5).

3.3.3.1 *let-7*: The moulting miRNA

The *lethal-7* (*let-7*) gene was first discovered in worms (Reinhart et al., 2000). This miRNA gene was found to be perfectly conserved throughout bilaterian phylogeny (Pasquinelli et al., 2000). In *C. elegans*, the *let-7* family consists of genes encoding nine miRNAs sharing the same seed sequence (Lim et al., 2003b). Among them, *let-7*, *miR-84*, *miR-48* and *miR-241* are involved in the *C. elegans* heterochronic pathway, sequentially controlling the developmental timing of larva transitions (Moss, 2007). Furthermore, the expression of *let-7* has the same rhythmic pattern with the hormone pulse before each cuticular moult in *Drosophila* (Thummel, 2001). Moulting is the manner in which an animal routinely sheds a part of its body (often, but not always,

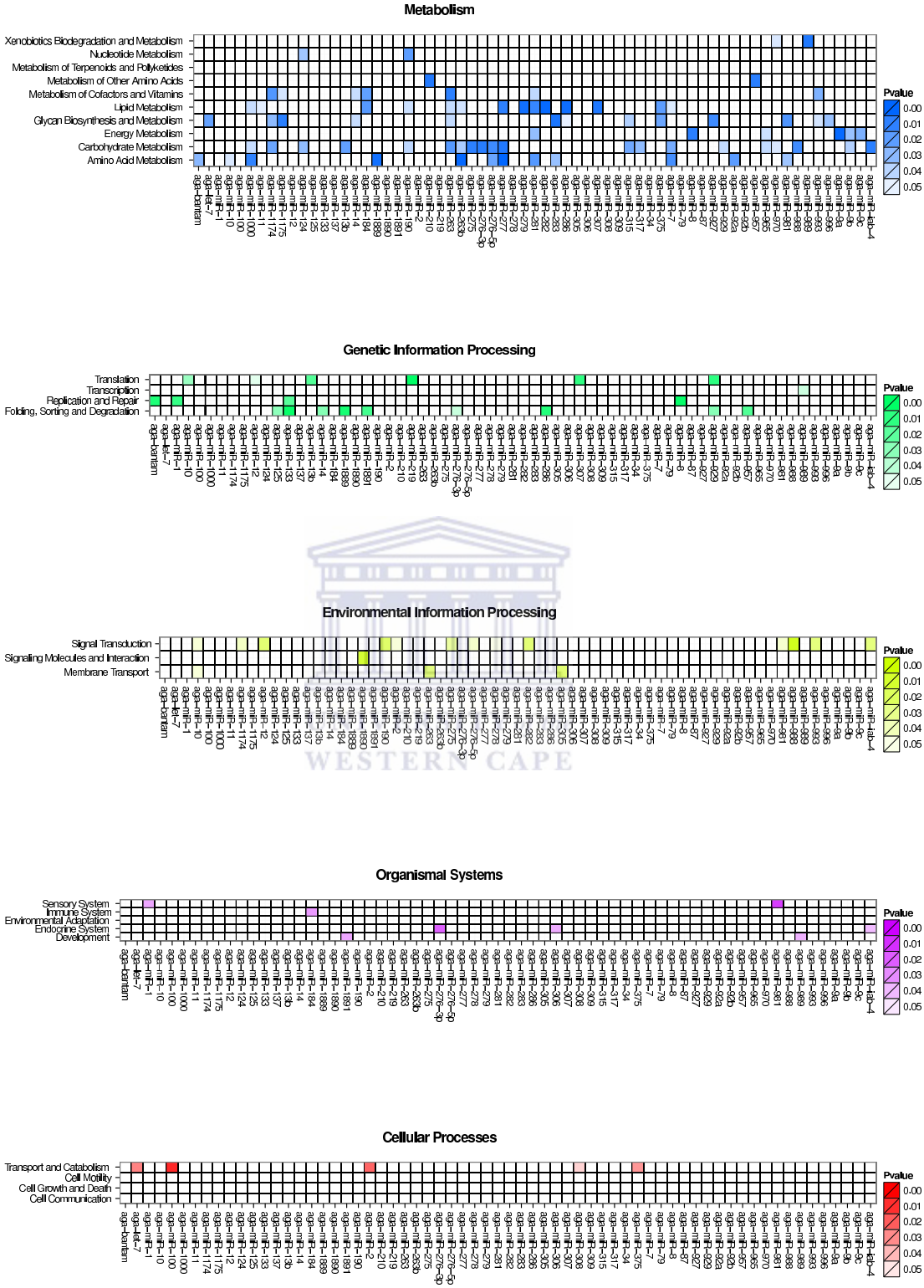


Figure 3.4: Functional map of *Anopheles* miRNAs and their targets genes. The heatmap representing the analysis of the enrichment using KEGG pathways for targets predicted by two or three methods (colour-coded values correspond to absolute *P-values* from the hypergeometric test.)

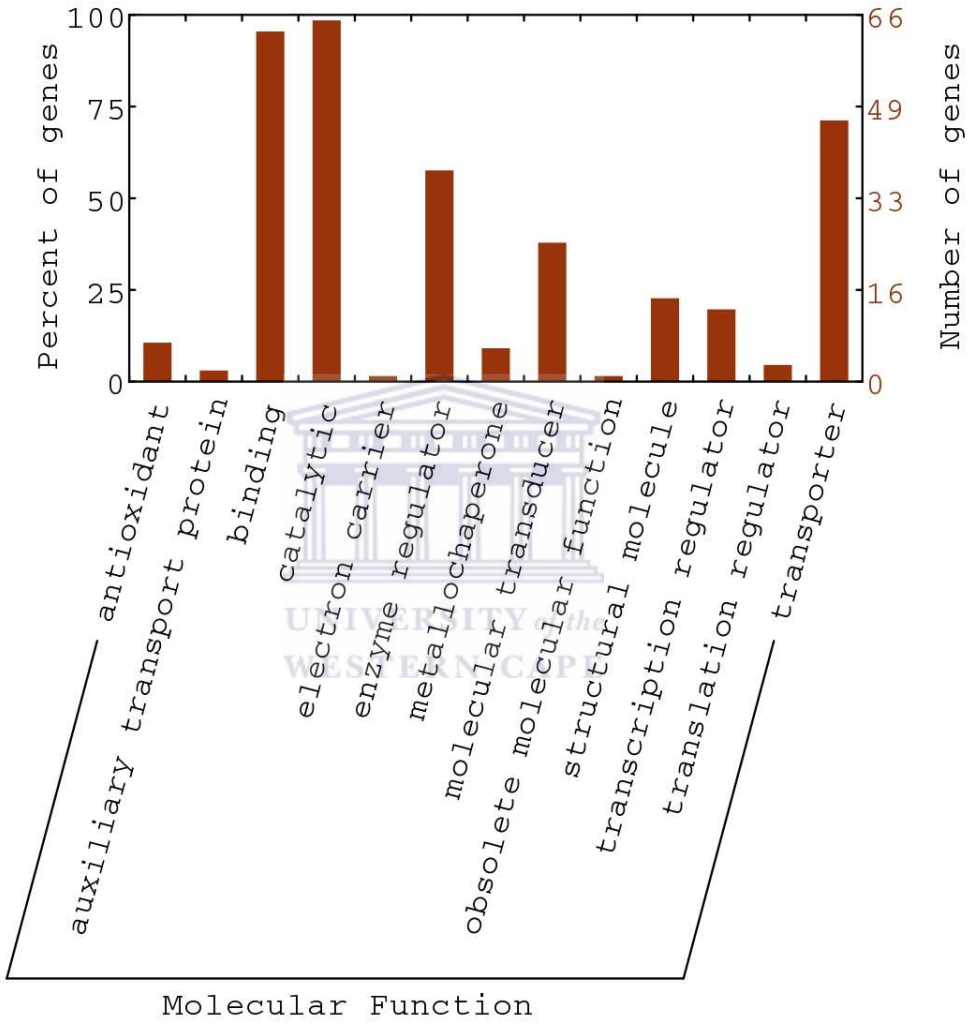


Figure 3.5: Enriched GO terms of all *An.gambiae* miRNA target genes predicted by two or three methods. The over-represented GO classification from the molecular function (X-axis). Percentage of target genes with the enriched GO term (Y-axis). All genes with statistically over-represented GO annotation were included (*P-value* of $\leq$0.05).

an outer layer or covering), either at specific times of year, or at specific points in its life-cycle. In some insects, environmental factors such as temperature and food availability control moluting, while in others, the number of moults is fixed and is controlled by hormones. To understand how moulting occurs, it helps to know the parts of the insect's exoskeleton also called the integument or skin. The insect's exoskeleton consists of both living and non-living layers. The outermost layer is called the cuticle and is non-living. The cuticle protects the insect against physical injury and water loss and provides rigidity for muscle attachment. It's the layer that sheds during a moult. Our results showed *let-7* targets six *Anopheles* genes sharing exactly the same GO term (GO:0042302: structural constituent of cuticle). Moreover, in *Aedes* and *Culex* mosquitoes more than 15 genes are enriched for the same GO term (Figure 3.6).

In mosquitoes, moulting occurs four times between the larva stage and the pupa stage only. After the fourth moult, mosquito larvae change into pupae. interestingly, we noticed a significant up-regulation of this miRNA cluster (*let-7*, *miR-125* and *miR-100*) in our larvae sample (Chapter 2, Figure 2.6).

3.3.3.2 *bantam*: The apoptotic miRNA

In *Drosophila*, cell death, or apoptosis, is inhibited by an inhibitor of apoptosis proteins (IAP), and induced by three IAP antagonistic genes: Head involution defective (*hid*), *reaper* and *grim* (Yoo et al., 2002). The proapoptotic gene (*hid* has been identified as a direct target for regulation by *bantam* miRNA, providing an explanation for *bantam's* anti-apoptotic activity in *Drosophila* (Brennecke et al., 2003). However, in *Anopheles* mosquitoes the absence of these antagonistic genes left a significant lacuna in our comprehension of cell death regulation (Zhou et al., 2005), thereby in the understanding of the function of *bantam* in these vetor insects. Our predictions showed that two mem-

Functional associations for aae-let-7 targets using GO terms:

Targets	Observation	GO ID	GO Term	P-value
AAEL012746-RA AAEL013852-RA AAEL000042-RA AAEL007728-RA AAEL009109-RA AAEL003366-RA AAEL002337-RB AAEL011708-RA	8	GO:0051082	unfolded protein binding	0.00137066862279913
AAEL004492-RA AAEL005863-RA	2	GO:0016538	cyclin-dependent protein kinase regulator activity	0.00242466542303361
AAEL008776-RA AAEL008783-RA AAEL001109-RA	3	GO:0015035	protein disulfide oxidoreductase activity	0.00243335946177077
AAEL012676-RC AAEL012676-RA AAEL005527-RA	3	GO:0008408	3'-5' exonuclease activity	0.00327490903439804
AAEL005752-RA AAEL003110-RA AAEL004361-RA AAEL002969-RA AAEL003066-RA AAEL009782-RA AAEL005460-RA	7	GO:0004553	hydrolase activity, hydrolyzing O-glycosyl compounds	0.00571366320692643
AAEL009796-RA AAEL003232-RA AAEL007192-RA AAEL004951-RA AAEL004758-RA AAEL012883-RA AAEL015363-RA AAEL003235-RA AAEL013517-RA AAEL011444-RA AAEL011032-RA AAEL008764-RA AAEL013515-RA AAEL003232-RB AAEL009002-RA	15	GO:0042302	structural constituent of cuticle	0.0066232944533992

Functional associations for aga-let-7 targets using GO terms:

Targets	Observation	GO ID	GO Term	P-value
AGAP000801-RA AGAP006026-RA	2	GO:0005234	extracellular-glutamate-gated ion channel activity	0.00218843041903027
AGAP000801-RA AGAP006026-RA	2	GO:0004970	ionotropic glutamate receptor activity	0.00218843041903027
AGAP006095-RA AGAP006003-RA AGAP005456-RA AGAP000047-RA AGAP006261-RA AGAP003390-RA	6	GO:0042302	structural constituent of cuticle	0.0232386963617736
AGAP000470-RA AGAP003490-RA AGAP009161-RA	3	GO:0016829	lyase activity	0.0251869585420433

Functional associations for cqu-let-7 targets using GO terms:

Targets	Observation	GO ID	GO Term	P-value
CPIJ008231-RA CPIJ006797-RA CPIJ009321-RA CPIJ004287-RA CPIJ018582-RA CPIJ013769-RA CPIJ016326-RA CPIJ001835-RA CPIJ016715-RA CPIJ001840-RA CPIJ017876-RA CPIJ001830-RA CPIJ013782-RA CPIJ016316-RA	14	GO:0042302	structural constituent of cuticle	4.85044104129877e-06
CPIJ017217-RA CPIJ019847-RA CPIJ006225-RA CPIJ006030-RA	4	GO:0008121	ubiquinol-cytochrome-c reductase activity	5.92213385055414e-06
CPIJ012341-RA CPIJ009262-RA CPIJ008366-RA CPIJ002859-RA	4	GO:0016616	oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor	0.000281868223075662
CPIJ010291-RA CPIJ019861-RA	2	GO:0046912	transferase activity, transferring acyl groups, acyl groups converted into alkyl on transfer	0.00176361984088044

Figure 3.6: Functional association for *let-7* target genes in *Ae. aegypti*, *An. gambiae.* and *C. quinquefasciatus*. *let-7* target genes contribute to the structural integrity of the cuticle (GO:0042302) in the three vector mosquitoes.

bers of the non-homologous end joining DNA repair pathways (Ku heterodimer 70 (KU70) and DNA Ligase IV (Ligase 4)) are *bantam* targets. This pathway eliminates DNA double-strand breaks by direct ligation (Harper and Elledge, 2007). This process includes binding the KU heterodimer (KU70/80) to double-stranded DNA ends, recruiting DNA-dependent protein kinase catalytic sub-unit (DNA-PKcs,) processing ends, and recruiting the DNA ligase IV (LIG4) complex, which brings about ligation. If we assume that *bantam* acts as a negative regulator of target genes (KU70 and Ligase 4), it suggests that enhances apoptosis by blocking the DNA repair pathway in the *Anopheles* mosquitoes rather than suppressing cell death, as in *Drosophila*. In addition, several studies report that apoptosis is an important defence mechanism against the parasite ookinete invasion of the midgut epithelial cells of the mosquitoes (Abraham and Jacobs-Lorena, 2004; Christophides et al., 2002; Han et al., 2000; Kumar et al., 2004; Zhou et al., 2005; Zieler and Dvorak, 2000).

3.3.3.3 *miR-2*: The translation inhibitor miRNA

One of the largest conserved insect microRNA family is the *miR-2* family (Griffiths-Jones et al., 2006; Griffiths-Jones, 2006; Griffiths-Jones et al., 2008; Kozomara and Griffiths-Jones, 2011). This family represents the miRNA precursors *miR-2* and *miR-13*. *miR-2* family widely represented in invertebrates (the invertebrate-specific family) and the copy number and genomic distribution varies greatly between species (Marco et al., 2012). Deep sequencing data reveal that all *miR-2* family members produce their dominant mature miRNAs from the 3′ arm, whose sequence is highly conserved (Leaman et al., 2005). Furthermore, most *miR-2* precursors have the same Dicer cleavage site, thus producing functional mature *miR-2* sequences with the same seed region and predicted targets. According to the available deep sequencing data, most *miR-2* loci

within the same species produce redundant products. This family also contains the *miR-13* precursor which, despite giving rise to a distinct mature sequence, appears to be related in sequence. In *Drosophila*, both family members (*miR-2* and *miR-13*) regulate cell survival by translational repression of proapoptotic genes (*rpr, hid*, and *grim*) (Leaman et al., 2005). In the three mosquitoes we found that these family members target a large number of genes involved in protein translation and ribosomal activity (Figure 3.7). In *Anopheles*, two miRNAs target more than eight genes in the ribosome biogenesis pathway. Additionally, those genes significantly enriched for a GO term related to the ribosomal RNA (rRNA) binding (GO:0019843). In *Aedes* and *Culex*, these two miRNA target more genes in the same pathway, suggesting a conservative function for this family in insects.

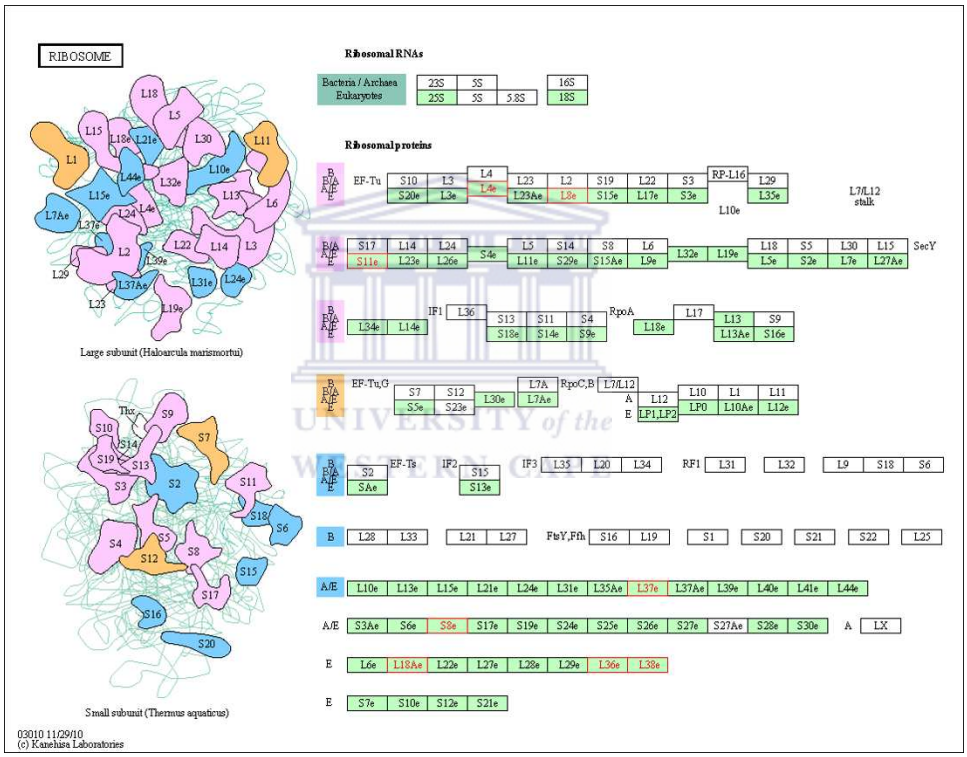
3.3.3.4 *miR-277*: The energy regulator miRNA

For the three mosquitoes species, we found *miR-277* regulates the pathway for valine, leucine, and isoleucine degradation by targeting a huge number of its enzyme members (Figure 3.8). These results are consistent with two studies that identified the role of this miRNA in *D. melanogaster* (Stark et al., 2003) and another seven *Drosophila* species (Grün et al., 2005). The three amino acids in this pathway; valine, leucine, and isoleucine, commonly referred to as branched-chain amino acids (BCAAs) because of their branched carbon skeletons . It has been extensively demonstrated that these amino acids play important roles in many aspects of animal growth and development (Nair and Short, 2005; Harris et al., 2005; Brosnan and Brosnan, 2006; Axtell and Bartel, 2005). For example, leucine may stimulate the phosphorylation of the mammalian target of the rapamycin (mTOR) pathway, in part, by serving both as a mitochondrial fuel through oxidative carboxylation and, as an allosteric activator of glutamate dehy-

Functional associations for aga-miR-2 targets using KEGG Pathway:

Targets	Observation	Pathway	Sub Pathway	Main Pathway	P-value
AGAP001778 AGAP006264 AGAP007520 AGAP002283 AGAP000928	5	Peroxisome	Transport and Catabolism	Cellular Processes	0.0134531006463264
AGAP010163 AGAP009998 AGAP005802 AGAP002437 AGAP002306 AGAP000952 AGAP002921 AGAP012284	8	Ribosome	Translation	Genetic Information Processing	0.0327944880454849

(a)



(b)

Functional associations for aga-miR-2 targets using GO terms:

Targets	Observation	GO ID	GO Term	P-value
AGAP005802-RA AGAP009998-RA AGAP000952-RA AGAP003217-RA	4	GO:0019843	rRNA binding	0.000352062521919048

(c)

Figure 3.7: Target genes of *miR-2*. *miR-2* target eight genes involved in ribosomal biogenesis pathway (a) and (b) (target genes boxed in red). In addition, four of these gene are enriched for a GO term which is related to the rRNA binding (GO:0019843) (c).

drogenase (Tokunaga et al., 2004; Chotechuang et al., 2009). In addition, BCAAs can affect the synthesis and compartmentalization of the excitatory neurotransmitter glutamate (Hutson et al., 2001; Fernstrom, 2005). The possible roles of BCAAs in energy metabolism in different species reveals the role of *miR-277* as a energy regulator.

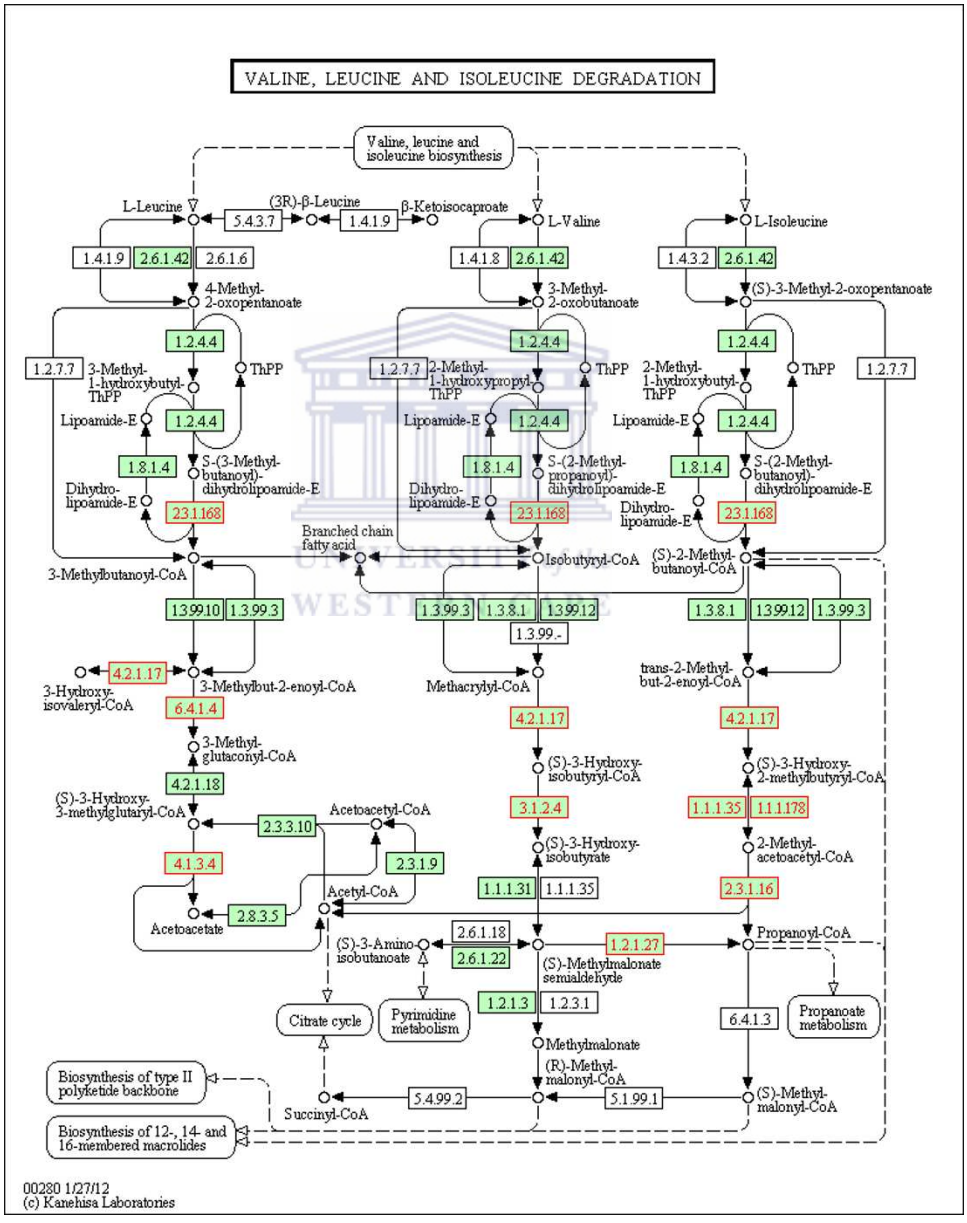
3.3.3.5 *miR-275*: The embryogenesis miRNA

Among the top targets for *miR-275*, we found two maternal genes; *torso* (*tor*) and *torso-like* (*tsl*) the triggers of the mitogen-activated protein kinase (MAPK) signaling pathway (Figure 3.9). In *Drosophila*, the *tor* receptor is distributed evenly at the embryonic surface at the blastoderm stage. *tor* becomes activated specifically at the anterior and posterior poles by a ligand which diffuses locally from a source near the poles. The restricted activation of *tor* critically depends on the presence of the product of the *tsl* gene in a subpopulation of follicle cells that overlay each end of the maturing oocyte. In the absence of *tor*, the *tor* receptor is not activated, while ubiquitous expression of *tsl* during oogenesis leads to the general activation of the receptor along the entire embryo. Upon activation, *tor* initiates the MAPK pathway, which will ultimately lead to the expression of the zygotic genes *tailless* (*tll*) and *huckebein* (*hkb*) which initiate the developmental programs giving rise to the anterior and posterior terminal regions of the embryo (Furriols and Casanova, 2003; Casanova, 2005). Nevertheless, these maternal RNAs and proteins must be properly localized or translated in order to direct the rapid mitotic divisions, patterning, and morphogenesis, of the early embryo (Sackton et al., 2007). This localization is accomplished through the degradation of certain maternal RNAs (Tadros et al., 2003) and the translation of others (Macdonald and Struhl, 1986). Moreover, changes in MAPK activity have been reported in the egg to embryo transition in several insect and other non-insect animals

Functional associations for aga-miR-277 targets using KEGG Pathways:

Targets	Observation	Pathway	Sub Pathway	Main Pathway	P-value
AGAP002761 AGAP000549 AGAP010228 AGAP002499 AGAP011833 AGAP004097 AGAP003414 AGAP008717 AGAP006821	9	Valine, leucine and isoleucine degradation	Amino Acid Metabolism	Metabolism	8.64987755506456e-07

(a)



(b)

Figure 3.8: Target genes of *miR-277*. *aga-miR-277* targets in Valine, leucine, and isoleucine, degradation pathway (a), *miR-277* regulates this pathway by targeting nine of its enzymes and thus acts as a metabolic switch (b) (target genes boxed in red).

(Sackton et al., 2007). However, unlike invertebrates and vertebrates (Tachibana et al., 2000; Tunquist and Maller, 2003; Zhang et al., 2006d), in insects this decrease is independent of fertilization (Sackton et al., 2007). This proves that *miR-275* inhibits the expression of the MAPK pathway trigger *tor* and its ligand gene *tsl*. Further evidence of it's role was derived from a study of *Ae. aegypti* (Bryant et al., 2010). The depletion of this miRNA after injecting its specific antagomir, resulted in severe defects linked to the inability to digest blood, excrete excessive fluids, and properly develop eggs in adult female mosquitoes.

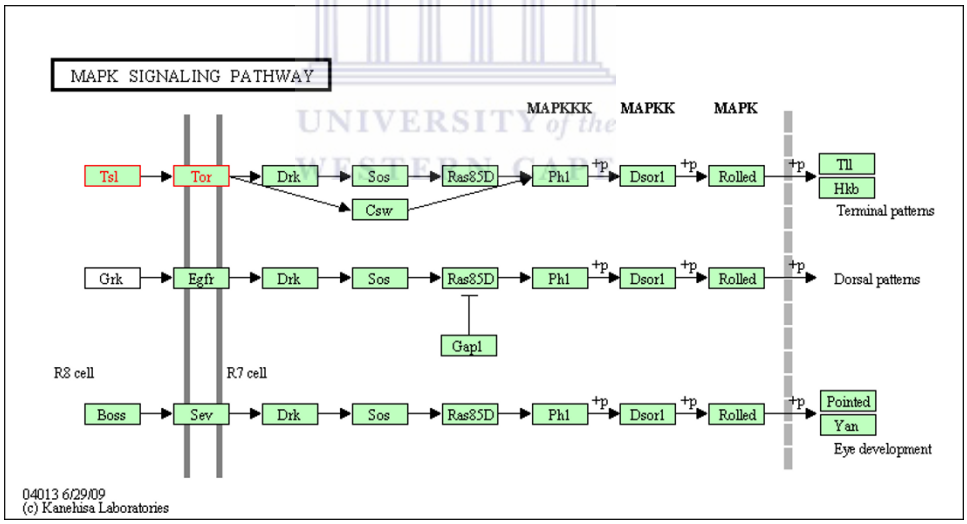
3.3.3.6 *miR-989*: The detoxification miRNA

Our prediction demonstrates that *miR-989* is targeting four detoxification enzymes involved in the metabolism of foreign chemical compounds (xenobiotic). Such compounds pose a constant challenge to the survival of all living organisms. Animals defend themselves against these toxins through an elaborate three-phase detoxification system, metabolizing xenobiotics into less harmful substances and facilitating their excretion (Xu et al., 2005). The detoxification enzymes of Phase I represent the most abundant class of xenobiotic metabolizing enzymes. They consist of cytochrome P450 monooxygenases (P450s) which decrease the biological activity of a broad range of substrates (or, less often, increase their toxicity). The enzymes of Phase II act on the toxic by-products of the Phase I response and include glutathione S-transferases (GSTs), UDP-glucuronosyltransferases (UGTs), and carboxylesterases. GSTs and UGTs add bulky side groups onto toxic compounds to increase their hydrophilicity, facilitating their excretion from the organism, while carboxylesterases catalyze the hydrolysis of ester-containing xenobiotics, leading to their detoxification. The Phase III system consists of the ATP-binding cassette (ABC) and other transmembrane trans-

Functional associations for aga-miR-275 targets using KEGG Pathways:

Targets	Observation	Pathway	Sub Pathway	Main Pathway	P-value
AGAP011353 AGAP010769 AGAP001205 AGAP004391	4	Amino sugar and nucleotide sugar metabolism	Carbohydrate Metabolism	Metabolism	0.00968039167248271
AGAP002282 AGAP005763	2	MAPK signaling pathway - fly	Signal Transduction	Environmental Information Processing	0.0315457163773455

(a)



(b)

Figure 3.9: Target genes of *miR-275*. *miR-275* bind to genes in the mitogen-activated protein kinase (MAPK) signaling pathway. Our results shows that *miR-275* target the MAPK pathway trigger *tor* (a) and its ligand gene *tsl* (b) (target genes boxed in red).

porters that actively export the conjugated toxins out of the cell (Misra et al., 2011). In the malaria vector, the expression of several detoxification enzymes increases in the midgut and fat body after a blood meal. Furthermore, the expression of several of these enzymes increases to even higher levels when mosquitoes are fed a *Plasmodium berghei* infected meal. This indicates that the oxidative stress after a blood meal is exacerbated by parasite infection (Molina-Cruz et al., 2008; Kumar et al., 2003). The suppression of the midgut catalase expression is a specific response to ookinete midgut invasion and is expected to lead to higher local levels of hydrogen peroxide. The reduction of detoxification in the midgut limits *Plasmodium* infection through a lytic mechanism (Molina-Cruz et al., 2008). Evidently, *miR-989* represses the activity of some of these enzymes involved in the xenobiotic elimination. The expression of this miRNA is affected by the presence of *Plasmodium* and flavivirus in *Anopheles* and *Culex* mosquitoes, respectively (Winter et al., 2007; Skalsky et al., 2010). Interestingly, the homologs of this miRNA in the other two mosquitoes (*Aedes* and *Culex*) is also enriched for the same pathway.

3.3.3.7 *miR-2490*: The endosymbiont miRNA

miR-2490 is a mosquito-specific miRNA. This miRNA was only found in the *Aedes* mosquitoes (Skalsky et al., 2010). In a recent study, Hussain and his colleagues observed the induction of *aae-miR-2940*, *aae-miR-309a-2*, *aae-miR-2943-1*, *aae-miR-970*, *aae-miR-308*, and *aae-miR-2941-2* in *Wolbachia* bacteria-infected mosquitoes (Hussain et al., 2011). They also determined that the metalloprotease gene (a protease enzyme whose catalytic mechanism involves a metal: AAEL012278) is a target for *aae-miR-2940*, which was up-regulated during the infection. In addition, when they silenced this gene they noticed significant reductions in Wolbachia density *in vitro* and

in vivo. Based upon our prediction, we found that this gene can be targeted by more than 20 miRNAs, including the ones reported by Hussain and his colleagues. However, the enrichment analysis for these miRNA targets (using miRanda) showed that this miRNA regulates the expression of subsets of autophagy pathway genes. Autophagy is an innate immune defense mechanism that acts against various intercellular bacterial (Huang and Brumell, 2009). These results indicate that *miR-2940* may facilitate the bacterial endosymbiont by inhibiting the autophagy machinery.

For the other mosquito-specific miRNAs like *miR-1889*, *miR-1890* and *miR-1891*, the prediction results indicate a major role for these miRNAs in the regulation of the oxidative phosphorylation pathway, the metabolic pathway that uses energy released by the oxidation of nutrients to produce adenosine triphosphate (ATP).

3.4 Discussion

The role of small RNAs as key regulators of mRNA turnover and translation has been well established. Recent advances indicate that miRNAs play important roles in animal development and physiology. Cellular activities such as proliferation, morphogenesis, apoptosis, and differentiation, are regulated by miRNAs. The expression of various genes are regulated by miRNAs. Several miRNAs act in reciprocal negative feedback loops with protein factors to control cell fate decisions that are triggered by signal transduction activity. These observations implicate small RNAs as important mediators of gene regulation in response to cell-cell signaling. The mechanism by which miRNAs silence gene expression is post-transcriptional, possibly influencing the stability, compartmentalization, and translation, of mRNAs. This mechanism is an efficient means to regulate production of a diverse range of proteins.

Each individual miRNA is likely to down-regulate the abundance or translation of many mRNAs (Lim et al., 2005; Selbach et al., 2008; Guo et al., 2010). Compounding the complexity of miRNA control, multiple miRNAs can act together on individual mRNAs to produce additive or synergistic effects on the production of protein (Wu et al., 2010). Thus, miRNA research will increasingly focus upon miRNA-regulated networks (Peter, 2010), in addition to identifying individual miRNA:mRNA interactions. Multiple methodologies are now available to ascertain miRNA targeting, each with intrinsic strengths and weaknesses. Combining multiple strategies is a requirement to obtain a comprehensive high-confidence description of miRNA targeting networks (Thomson et al., 2011).

3. INSECTAR: A DATABASE FOR MICRORNA TARGET GENES IN INSECTS

In this study, we constructed InsecTar, a database which facilitates a comprehensive exploration of miRNA targets in insects. The genome-wide target sites of miRNAs were predicted, organized, and integrated with a number of publicly available databases, including Ensembl, KEGG and GO. By incorporating these databases together with functional annotation, InsecTar offers the most comprehensive and integrated view of these target sites. Currently, InsecTar includes the miRNA targets for only three insects (*An. gambia*, *Ae. aegypti* and *C. quinquefasciatus*). In the next update of the database sequenced insect genomes and associated miRNAs will be included.

Finally, functional characterization of *Anopheles* miRNA targets provides evidence that miRNAs have been found to be involved in a variety of pathways and biological processes such as embryogenesis (e.g. *miR-275*), metamorphosis (e.g. *let-7*), apoptosis (*bantam*), metabolism (e.g. *miR-277*), endosymbiont (e.g. *2490*), and immunity (e.g. *miR-989*). These results suggest that some of the miRNAs have roles not only in the development but also in insect-host and insect-parasite interactions.

3.5 Conclusion

InsecTar is a comprehensive database specifically developed to provide an open-access repository of information for miRNA targets in insects. The database incorporates target prediction and functional analysis of insect miRNA target genes. The proposed database is a useful resource for experimental miRNA researchers and computational biologist to study miRNA regulation in insects. The database can be freely accessed from http://insectar.sanbi.ac.za.

Chapter 4

Conclusion

After uncovering the first miRNA in nematodes, a remarkable diversity of miRNAs has been reported using computational and experimental methods (Griffiths-Jones, 2006). Computational methods based on the evolutionary conservation of genomic sequences and their ability to fold into stable hairpin structures have been applied to species with genomes sequenced, such as a number of arthropods, worms, and vertebrates (Lai et al., 2003; Grad et al., 2003; Sethupathy et al., 2006; Weaver et al., 2007; Pérez-Quintero et al., 2012). In addition, the development of novel techniques for directional cloning of small RNAs has led to the identification of many other miRNAs (Yao et al., 2007; Mead and Tu, 2008). But, the greatest progress came with the advent of high throughput sequencing technologies (Motameny et al., 2010). These technologies confirmed most of the miRNAs predicted *in-silico*, and enabled the discovery of new and unexpected miRNAs. Its also allowed for the discovery of new miRNAs in species whose genome has not sequenced (Kozomara and Griffiths-Jones, 2011).

In this thesis, using the above-mentioned technologies, we identified the first miRNAs in *An. funestus s.s* and described their expression in the four developing stages (egg, larva, pupa, and adults). Furthermore, this study offers a substantially expanded

list of miRNAs in *Anopheles* mosquitoes and we provided evidence of an additional 33 new miRNA in this species. These results significantly expand the set of miRNAs in the anopheline species to 98 miRNAs. Among theses miRNAs there are 65 miRNAs that match previously known miRNAs in *An. gambiae* and *An. stephensi*. Sixteen miRNAs has not be described in the *Anopheles* genus, but, previously reported in Aedes and Culex mosquitoes, four miRNAs new in mosquitoes, eight did not match any known miRNAs in any organism, three new miRNA stem-loop precursors, and two new miRNA isoforms or variants.

The expression profile analysis of *An. funestus s.s* miRNAs, including the new miRNAs, revealed distinct patterns of expression from early embryo to adult stages. Changes in the expression profiles were mirrored in all stages indicating a role for these miRNAs in the mosquito maturation. We found that many miRNAs were stage-specific, which suggests that miRNAs function in development. Therefore, knockdown or blocking the biogenesis pathway of one of these stage-specific miRNAs can limit the mosquito's development at a crucial stage which will lead to determine better ways to control this mosquito in the early development stages. For example, our results showed that the *miR-305*, *let-7* and *mir-277*, miRNA families are expressed at very high levels during the egg, larva, and pupa stage, respectively. Thus, silencing such miRNAs in these water stages could help control mosquitoes in the water before they have a chance to emerge.

Researchers embarked on miRNA target prediction long ago with the serendipitous findings that emerged from miRNA target recognition. The key principles were then applied to computational methods for miRNA target prediction. These methods soon

allowed for the prediction of hundreds of miRNA targets (Lewis et al., 2003; Stark et al., 2003; Adai et al., 2005). However, computational prediction of miRNA targets still relies only on the few principles defined more than 20 years ago, and, arguably, this will not help to unveil novel aspects of miRNA target mechanisms. As the number of validated targets is currently limited, methods for target-based inference of miRNA function must rely on these predictions. Thus, analyzing properties of miRNA targets would be valuable in order to identify miRNA function. If the targets of a specific miRNA are enriched with genes annotated with some biological process or pathway, it is reasonable to infer that the miRNA is involved in the same process.

In this manner, we have made a significant contribution to improved miRNA targets predication in insects. Combining different miRNA target prediction programs result in significant improvement of target prediction accuracy. Moreover, we tested the predicted target genes for statistically significant associations using a hypergeometric test to link miRNA targets to functional annotation using GO terms and KEGG pathways. Additionally, we developed an integrated database to query the miRNA target prediction data. The InsecTar database was designed not just to store the predicted targets of insect miRNAs but also to supply functional information about these target genes. Unlike other miRNA target databases, InsecTar provides a broad range of adjustable search options. Therefore, InsecTar will be a useful source of information for future studies in both basic and applied entomology. Currently, InsecTar provides information of miRNA targets for three insects (*Ae aegypti*, *An. gambiae* and *C. quinquefasciatus*) for which their genome data is available in Ensembl. In future, we plan to include data from more insect species to enhance the miRNA regulatory network in insects.

4. CONCLUSION

We believe that the phase of predicting miRNA putative targets by computational methods must be followed by experimental work to validate the predictions. miRNA silencing currently remains the most useful approach. The validation of targets will contribute to elucidating the role of miRNAs in the molecular network that regulates the biological processes.

Finally, the overall results provide a new understanding of insect miRNAs as we have identified functions for some mosquito miRNAs. We found that miRNA plays a critical role in the regulatory functions of the mosquito as a vector. Disrupting the mosquito at any stage of the life-cycle would reduce disease transmission in the next generation of mosquitoes by reducing survival and reproduction. We postulate that if the total number of mosquitoes are reduced, fewer infected mosquitoes would bite humans, and transmission of disease would drop. To this end, introducing a gene disruptor that reduces levels of a specific miRNA in the mosquito population could be used to inhibit transmission of vector-borne diseases such as malaria. Since the mosquito life-cycle lasts about 6-8 weeks, with just 1-2 weeks in the adult stage, any intervention at the various developmental stages would reduce the mosquito population with a notable effect on parasite transmission.

In conclusion, this type of analysis is a key step towards improving our understanding of the complexity and regulation mode of miRNAs in mosquitoes. Moreover, this study opened the door for exploration of miRNA in regulation of critical physiological functions specific to vector arthropods which may lead to novel approaches to combat mosquito-borne infectious diseases.

References

Abraham, E. G. and Jacobs-Lorena, M. (2004). Mosquito midgut barriers to malaria parasite development. *Insect Biochem Mol Biol*, 34(7):667–71. 103

Adai, A., Johnson, C., Mlotshwa, S., Archer-Evans, S., Manocha, V., Vance, V., and Sundaresan, V. (2005). Computational prediction of miRNAs in Arabidopsis thaliana. *Genome Res*, 15(1):78–91. 18, 117

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, 215(3):403–10. 52

Altuvia, Y., Landgraf, P., Lithwick, G., Elefant, N., Pfeffer, S., Aravin, A., Brownstein, M. J., Tuschl, T., and Margalit, H. (2005). Clustering and conservation patterns of human microRNAs. *Nucleic Acids Res*, 33(8):2697–706. 22, 23, 80

Ambros, V. (2001). MicroRNAs: tiny regulators with great potential. *Cell*, 107(7):823–6. 17

Ambros, V., Bartel, B., Bartel, D. P., Burge, C. B., Carrington, J. C., Chen, X., Dreyfuss, G., Eddy, S. R., Griffiths-Jones, S., Marshall, M., Matzke, M., Ruvkun, G., and Tuschl, T. (2003a). A uniform system for microRNA annotation. *RNA*, 9(3):277–9. 17, 19

Ambros, V., Lee, R. C., Lavanway, A., Williams, P. T., and Jewell, D. (2003b). MicroRNAs and other tiny endogenous RNAs in C. elegans. *Curr Biol*, 13(10):807–18. 19, 21

Antonio-Nkondjio, C., Awono-Ambene, P., Toto, J.-C., Meunier, J.-Y., Zebaze-Kemleu, S., Nyambam, R., Wondji, C. S., Tchuinkam, T., and Fontenill, D. (2002).

High malaria transmission intensity in a village close to Yaounde, the capital city of Cameroon. *J Med Entomol*, 39(2):350–5. 10

Aravin, A. A., Lagos-Quintana, M., Yalcin, A., Zavolan, M., Marks, D., Snyder, B., Gaasterland, T., Meyer, J., and Tuschl, T. (2003). The small RNA profile during Drosophila melanogaster development. *Dev Cell*, 5(2):337–50. 41

Asgari, S. (2011). Role of micrornas in insect host-microorganism interactions. *Front Physiol*, 2:48. 46

Avesson, L., Reimegård, J., Wagner, E. G. H., and Söderbom, F. (2012). Micrornas in amoebozoa: Deep sequencing of the small rna population in the social amoeba dictyostelium discoideum reveals developmentally regulated micrornas. *RNA*. 75

Axtell, M. J. and Bartel, D. P. (2005). Antiquity of microRNAs and their targets in land plants. *Plant Cell*, 17(6):1658–73. 20, 21, 104

Baek, D., Villén, J., Shin, C., Camargo, F. D., Gygi, S. P., and Bartel, D. P. (2008). The impact of microRNAs on protein output. *Nature*, 455(7209):64–71. 46

Baigude, H., Ahsanullah, Li, Z., Zhou, Y., and Rana, T. M. (2012). miR-TRAP: A benchtop chemical biology strategy to identify microRNA targets. *Angew Chem Int Ed Engl*. 86

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281–97. 15, 20, 21, 37, 46, 76

Bartel, D. P. (2009). MicroRNAs: target recognition and regulatory functions. *Cell*, 136(2):215–33. 15, 85

REFERENCES

Baskerville, S. and Bartel, D. P. (2005). Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA*, 11(3):241–7. 14

Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., Barad, O., Barzilai, A., Einat, P., Einav, U., Meiri, E., Sharon, E., Spector, Y., and Bentwich, Z. (2005). Identification of hundreds of conserved and nonconserved human microRNAs. *Nat Genet*, 37(7):766–70. 19

Berezikov, E., Cuppen, E., and Plasterk, R. H. A. (2006). Approaches to microrna discovery. *Nat Genet*, 38 Suppl:S2–7. 40

Berezikov, E., Guryev, V., van de Belt, J., Wienholds, E., Plasterk, R. H. A., and Cuppen, E. (2005). Phylogenetic shadowing and computational identification of human microRNA genes. *Cell*, 120(1):21–4. 23

Boeri, M., Pastorino, U., and Sozzi, G. (2012). Role of micrornas in lung cancer: microrna signatures in cancer prognosis. *Cancer J*, 18(3):268–74. 47

Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K. D., Ovcharenko, I., Pachter, L., and Rubin, E. M. (2003). Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, 299(5611):1391–4. 23

Bonnet, E., Wuyts, J., Rouzé, P., and Van de Peer, Y. (2004a). Detection of 91 potential conserved plant microRNAs in Arabidopsis thaliana and Oryza sativa identifies important target genes. *Proc Natl Acad Sci U S A*, 101(31):11511–6. 23

Bonnet, E., Wuyts, J., Rouzé, P., and Van de Peer, Y. (2004b). Evidence that microRNA

precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics*, 20(17):2911–7. 17

Borges, F., Pereira, P. A., Slotkin, R. K., Martienssen, R. A., and Becker, J. D. (2011). Microrna activity in the arabidopsis male germline. *J Exp Bot*, 62(5):1611–20. 47

Boutz, P. L., Chawla, G., Stoilov, P., and Black, D. L. (2007). MicroRNAs regulate the expression of the alternative splicing factor nPTB during muscle development. *Genes Dev*, 21(1):71–84. 80

Brennecke, J., Hipfner, D. R., Stark, A., Russell, R. B., and Cohen, S. M. (2003). bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in Drosophila. *Cell*, 113(1):25–36. 13, 21, 46, 101

Brennecke, J., Stark, A., Russell, R. B., and Cohen, S. M. (2005). Principles of microrna-target recognition. *PLoS Biol*, 3(3):e85. 27

Brodersen, P. and Voinnet, O. (2009). Revisiting the principles of microRNA target recognition and mode of action. *Nat Rev Mol Cell Biol*, 10(2):141–8. 30, 86

Brosnan, J. T. and Brosnan, M. E. (2006). Branched-chain amino acids: enzyme and substrate regulation. *J Nutr*, 136(1 Suppl):207S–11S. 104

Brown, J. R. and Sanseau, P. (2005). A computational view of microRNAs and their targets. *Drug Discov Today*, 10(8):595–601. 21, 23

Bryant, B., Macdonald, W., and Raikhel, A. S. (2010). microRNA miR-275 is indispensable for blood digestion and egg development in the mosquito Aedes aegypti. *Proc Natl Acad Sci U S A*, 107(52):22391–8. 108

REFERENCES

Buchold, G. M., Coarfa, C., Kim, J., Milosavljevic, A., Gunaratne, P. H., and Matzuk, M. M. (2010). Analysis of microrna expression in the prepubertal testis. *PLoS One*, 5(12):e15317. 47

Buermans, H. P. J., Ariyurek, Y., van Ommen, G., den Dunnen, J. T., and 't Hoen, P. A. C. (2010). New methods for next generation sequencing based microRNA expression profiling. *BMC Genomics*, 11:716. 48

Burchard, J., Jackson, A. L., Malkov, V., Needham, R. H. V., Tan, Y., Bartz, S. R., Dai, H., Sachs, A. B., and Linsley, P. S. (2009). MicroRNA-like off-target transcript regulation by siRNAs is species specific. *RNA*, 15(2):308–15. 28

Burgler, C. and Macdonald, P. M. (2005). Prediction and verification of microRNA targets by MovingTargets, a highly adaptable prediction method. *BMC Genomics*, 6:88. 31, 35, 86

Burnside, J., Ouyang, M., Anderson, A., Bernberg, E., Lu, C., Meyers, B. C., Green, P. J., Markis, M., Isaacs, G., Huang, E., and Morgan, R. W. (2008). Deep sequencing of chicken micrornas. *BMC Genomics*, 9:185. 75

Cai, X., Hagedorn, C. H., and Cullen, B. R. (2004). Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA*, 10(12):1957–66. 14

Cao, J., Tong, C., Wu, X., Lv, J., Yang, Z., and Jin, Y. (2008). Identification of conserved microRNAs in Bombyx mori (silkworm) and regulation of fibroin L chain production by microRNAs in heterologous system. *Insect Biochem Mol Biol*, 38(12):1066–71. 87

REFERENCES

Carè, A., Catalucci, D., Felicetti, F., Bonci, D., Addario, A., Gallo, P., Bang, M.-L., Segnalini, P., Gu, Y., Dalton, N. D., Elia, L., Latronico, M. V. G., Høydal, M., Autore, C., Russo, M. A., Dorn, 2nd, G. W., Ellingsen, O., Ruiz-Lozano, P., Peterson, K. L., Croce, C. M., Peschle, C., and Condorelli, G. (2007). Microrna-133 controls cardiac hypertrophy. *Nat Med*, 13(5):613–8. 87

Carthew, R. W. and Sontheimer, E. J. (2009). Origins and mechanisms of miRNAs and siRNAs. *Cell*, 136(4):642–55. 15, 19, 25

Casanova, J. (2005). Developmental evolution: torso–a story with different ends? *Curr Biol*, 15(23):R968–70. 106

CDC (2010). About malaria. *Centers for Disease Control and Prevention, USA, http://www.cdc.gov/malaria/about/index.html*. 3, 4, 5, 8

Chan, C. S., Elemento, O., and Tavazoie, S. (2005). Revealing posttranscriptional regulatory elements through network-level conservation. *PLoS Comput Biol*, 1(7):e69. 35

Chandra, V., Girijadevi, R., Nair, A. S., Pillai, S. S., and Pillai, R. M. (2010). MTar: a computational microRNA target prediction architecture for human transcriptome. *BMC Bioinformatics*, 11 Suppl 1:S2. 86

Chen, C., Ridzon, D. A., Broomer, A. J., Zhou, Z., Lee, D. H., Nguyen, J. T., Barbisin, M., Xu, N. L., Mahuvakar, V. R., Andersen, M. R., Lao, K. Q., Livak, K. J., and Guegler, K. J. (2005). Real-time quantification of microrna by stem-loop rt-pcr. *Nucleic Acids Res*, 33(20):e179. 20

Chen, C.-Z. (2005). MicroRNAs as oncogenes and tumor suppressors. *N Engl J Med*, 353(17):1768–71. 16

Chen, Q., Lu, L., Hua, H., Zhou, F., Lu, L., and Lin, Y. (2012). Characterization and comparative analysis of small RNAs in three small RNA libraries of the brown planthopper (Nilaparvata lugens). *PLoS One*, 7(3):e32860. 79

Cheng, C. and Li, L. M. (2008). Inferring microRNA activities by combining gene expression with microRNA target prediction. *PLoS One*, 3(4):e1989. 37

Chi, X., Yang, Q., Chen, X., Wang, J., Pan, L., Chen, M., Yang, Z., He, Y., Liang, X., and Yu, S. (2011). Identification and characterization of microRNAs from peanut (Arachis hypogaea L.) by high-throughput sequencing. *PLoS One*, 6(11):e27530. 78

Choi, I. K. and Hyun, S. (2012). Conserved microRNA miR-8 in fat body regulates innate immune homeostasis in Drosophila. *Dev Comp Immunol*, 37(1):50–4. 46

Chotechuang, N., Azzout-Marniche, D., Bos, C., Chaumontet, C., Gausserès, N., Steiler, T., Gaudichon, C., and Tomé, D. (2009). mTOR, AMPK, and GCN2 coordinate the adaptation of hepatic energy metabolic pathways in response to protein intake in the rat. *Am J Physiol Endocrinol Metab*, 297(6):E1313–23. 106

Christophides, G. K., Zdobnov, E., Barillas-Mury, C., Birney, E., Blandin, S., Blass, C., Brey, P. T., Collins, F. H., Danielli, A., Dimopoulos, G., Hetru, C., Hoa, N. T., Hoffmann, J. A., Kanzok, S. M., Letunic, I., Levashina, E. A., Loukeris, T. G., Lycett, G., Meister, S., Michel, K., Moita, L. F., Müller, H.-M., Osta, M. A., Paskewitz, S. M., Reichhart, J.-M., Rzhetsky, A., Troxler, L., Vernick, K. D., Vlachou, D., Volz, J., von Mering, C., Xu, J., Zheng, L., Bork, P., and Kafatos, F. C.

(2002). Immunity-related genes and gene families in Anopheles gambiae. *Science*, 298(5591):159–65. 103

Coburn, G. A. and Cullen, B. R. (2003). siRNAs: a new wave of RNA-based therapeutics. *J Antimicrob Chemother*, 51(4):753–6. 19

Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., and Rice, P. M. (2010). The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants. *Nucleic Acids Res*, 38(6):1767–71. 49

Coetzee, M., Craig, M., and le Sueur, D. (2000). Distribution of African malaria mosquitoes belonging to the Anopheles gambiae complexfrican malaria mosquitoes belonging to the Anopheles gambiae complex. *Parasitol Today*, 16(2):74–7. 8

Coetzee, M. and Fontenille, D. (2004). Advances in the study of Anopheles funestus, a major vector of malaria in Africa. *Insect Biochem Mol Biol*, 34(7):599–605. 8, 10

Cordero, F., Beccuti, M., Arigoni, M., Donatelli, S., and Calogero, R. A. (2012). Optimizing a massive parallel sequencing workflow for quantitative miRNA expression analysis. *PLoS One*, 7(2):e31630. 76

Costantini, C., Sagnon, N., Ilboudo-Sanogo, E., Coluzzi, M., and Boccolini, D. (1999). Chromosomal and bionomic heterogeneities suggest incipient speciation in Anopheles funestus from Burkina Faso. *Parassitologia*, 41(4):595–611. 10

Cox-Singh, J., Davis, T. M. E., Lee, K.-S., Shamsul, S. S. G., Matusop, A., Ratnam, S., Rahman, H. A., Conway, D. J., and Singh, B. (2008). Plasmodium knowlesi malaria in humans is widely distributed and potentially life threatening. *Clin Infect Dis*, 46(2):165–71. 2

REFERENCES

Creighton, C. J., Benham, A. L., Zhu, H., Khan, M. F., Reid, J. G., Nagaraja, A. K., Fountain, M. D., Dziadek, O., Han, D., Ma, L., Kim, J., Hawkins, S. M., Anderson, M. L., Matzuk, M. M., and Gunaratne, P. H. (2010). Discovery of novel microrna in female reproductive tract using next generation sequencing. *PLoS One*, 5(3):e9637. 47

Cummins, J. M., He, Y., Leary, R. J., Pagliarini, R., Diaz, Jr, L. A., Sjoblom, T., Barad, O., Bentwich, Z., Szafranska, A. E., Labourier, E., Raymond, C. K., Roberts, B. S., Juhl, H., Kinzler, K. W., Vogelstein, B., and Velculescu, V. E. (2006). The colorectal microrna ome. *Proc Natl Acad Sci U S A*, 103(10):3687–92. 40

Cutting, A. D., Bannister, S. C., Doran, T. J., Sinclair, A. H., Tizard, M. V. L., and Smith, C. A. (2012). The potential role of microrna s in regulating gonadal sex differentiation in the chicken embryo. *Chromosome Res*, 20(1):201–13. 47

De Meillon, B. (1933). On Anopheles funestus and its allies in the Transvaal. *Annals of Tropical Medicine and Parasitology*, 27:83–97. 10

De Meillon, B., Van Eeden, G., Coetzee, L., Coetzee, M., Meiswinkel, R., Du Toit, C., and Hansford, C. (1977). Observations on a species of the Anopheles funestus subgroup, a suspected exophilic vector of malaria parasites in North-Eastern Transvaal, South Africa. *Mosquito News*, 37:657–661. 12

Dezulian, T., Remmert, M., Palatnik, J. F., Weigel, D., and Huson, D. H. (2006). Identification of plant microRNA homologs. *Bioinformatics*, 22(3):359–60. 21

Dia, I., Diop, T., Rakotoarivony, I., Kengne, P., and Fontenille, D. (2003). Bionomics of Anopheles gambiae Giles, An. arabiensis Patton, An. funestus Giles and An. nili

(Theobald) (Diptera: Culicidae) and transmission of Plasmodium falciparum in a Sudano-Guinean zone (Ngari, Senegal). *J Med Entomol*, 40(3):279–83. 10

Dkhil, M., Abdel-Baki, A. A., Delić, D., Wunderlich, F., Sies, H., and Al-Quraishy, S. (2011). Eimeria papillata: upregulation of specific mirna-species in the mouse jejunum. *Exp Parasitol*, 127(2):581–6. 46

Doane, R. W. (1910). *Insects and Diseases*. The Quinn and Boden Co. Press. 87

Dong, Y., Manfredini, F., and Dimopoulos, G. (2009). Implication of the mosquito midgut microbiota in the defense against malaria parasites. *PLoS Pathog*, 5(5):e1000423. 12

Eddy, S. R. (2004). How do RNA folding algorithms work? *Nat Biotechnol*, 22(11):1457–8. 28, 29

Enright, A. J., John, B., Gaul, U., Tuschl, T., Sander, C., and Marks, D. S. (2003). MicroRNA targets in Drosophila. *Genome Biol*, 5(1):R1. 24, 29, 31, 32, 86, 89

Ewing, B., Hillier, L., Wendl, M. C., and Green, P. (1998). Base-calling of automated sequencer traces using phred. i. accuracy assessment. *Genome Res*, 8(3):175–85. 58, 63

Fahlgren, N., Howell, M. D., Kasschau, K. D., Chapman, E. J., Sullivan, C. M., Cumbie, J. S., Givan, S. A., Law, T. F., Grant, S. R., Dangl, J. L., and Carrington, J. C. (2007). High-throughput sequencing of Arabidopsis microRNAs: evidence for frequent birth and death of miRNA genes. *PLoS One*, 2(2):e219. 78

Fang, Z. and Rajewsky, N. (2011). The impact of mirna target sites in coding sequences and in 3'utrs. *PLoS One*, 6(3):e18067. 27

REFERENCES

Fasold, M., Langenberger, D., Binder, H., Stadler, P. F., and Hoffmann, S. (2011). Dario: a ncrna detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Res*, 39(Web Server issue):W112–7. 52

Fernstrom, J. D. (2005). Branched-chain amino acids and brain function. *J Nutr*, 135(6 Suppl):1539S–46S. 106

Floyd, S. K. and Bowman, J. L. (2004). Gene regulation: ancient microRNA target sequences in plants. *Nature*, 428(6982):485–6. 19, 21

Flynt, A. S., Thatcher, E. J., Burkewitz, K., Li, N., Liu, Y., and Patton, J. G. (2009). miR-8 microRNAs regulate the response to osmotic stress in zebrafish embryos. *J Cell Biol*, 185(1):115–27. 79

Fontenille, D., Lepers, J. P., Campbell, G. H., Coluzzi, M., Rakotoarivony, I., and Coulanges, P. (1990). Malaria transmission and vector biology in Manarintsoa, high plateaux of Madagascar. *Am J Trop Med Hyg*, 43(2):107–15. 10

Fontenille, D., Lochouarn, L., Diagne, N., Sokhna, C., Lemasson, J. J., Diatta, M., Konate, L., Faye, F., Rogier, C., and Trape, J. F. (1997). High annual and seasonal variations in malaria transmission by anophelines and vector species composition in Dielmo, a holoendemic area in Senegal. *Am J Trop Med Hyg*, 56(3):247–53. 10

Friedländer, M. R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S., and Rajewsky, N. (2008). Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotechnol*, 26(4):407–15. 51

Friedländer, M. R., Mackowiak, S. D., Li, N., Chen, W., and Rajewsky, N. (2012).

miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res*, 40(1):37–52. 52, 58, 62

Friedman, R. C., Farh, K. K.-H., Burge, C. B., and Bartel, D. P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res*, 19(1):92–105. 27

Furriols, M. and Casanova, J. (2003). In and out of torso rtk signalling. *EMBO J*, 22(9):1947–52. 106

Gaidatzis, D., van Nimwegen, E., Hausser, J., and Zavolan, M. (2007). Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics*, 8:69. 27, 31

Gardner, P. P., Daub, J., Tate, J., Moore, B. L., Osuch, I. H., Griffiths-Jones, S., Finn, R. D., Nawrocki, E. P., Kolbe, D. L., Eddy, S. R., and Bateman, A. (2011). Rfam: Wikipedia, clans and the "decimal" release. *Nucleic Acids Res*, 39(Database issue):D141–5. 52, 59

Gébelin, V., Argout, X., Engchuan, W., Pitollat, B., Duan, C., Montoro, P., and Leclercq, J. (2012). Identification of novel microrna s in hevea brasiliensis and computational prediction of their targets. *BMC Plant Biol*, 12:18. 47

Geiduschek, E. P. and Haselkorn, R. (1969). Messenger rna. *Annu Rev Biochem*, 38:647–76. 13

Gennarino, V. A., Sardiello, M., Mutarelli, M., Dharmalingam, G., Maselli, V., Lago, G., and Banfi, S. (2011). Hoctar database: a unique resource for microRNA target prediction. *Gene*, 480(1-2):51–8. 31

REFERENCES

Gesteland, R., Cech, T., and Atkins, J. (2006). The RNA world. *Cold Spring Harbor Laboratory Press*. 13

Gilabert-Estelles, J., Braza-Boils, A., Ramon, L. A., Zorio, E., Medina, P., Espana, F., and Estelles, A. (2012). Role of micrornas in gynecological pathology. *Curr Med Chem*, 19(15):2406–13. 47

Gilles, H. M. and Warrell, D. A. (1993). *Bruce-Chwatt's Essential Malariology*. Hodder Arnold Publishers, UK, third edition. 3

Gillies, M. T. and De Meillon, B. (1968). The Anophelinae of Africa South of the Sahara. *South African Institute for Medical Research*, 54:127–150. 8, 10, 11, 12

Gillies, M. T. and De Meillon, B. (1987). Supplement to the Anophelinae of Africa South of the Sahara. *South African Institute for Medical Research*, 55. 8, 10, 11

Giraldez, A. J., Mishima, Y., Rihel, J., Grocock, R. J., Van Dongen, S., Inoue, K., Enright, A. J., and Schier, A. F. (2006). Zebrafish MiR-430 promotes deadenylation and clearance of materanl mRNAs. *Science*, 312(5770):75–9. 29

Grad, Y., Aach, J., Hayes, G. D., Reinhart, B. J., Church, G. M., Ruvkun, G., and Kim, J. (2003). Computational and experimental identification of C. elegans microRNAs. *Mol Cell*, 11(5):1253–63. 20, 21, 115

Graham, M. A., Silverstein, K. A. T., Cannon, S. B., and VandenBosch, K. A. (2004). Computational identification and characterization of novel genes from Legumes. *Plant Physiol*, 135(3):1179–97. 21

Green, C. A. (1982). Cladistic analysis of mosquito chromosome data (Anopheles (Cellia) Myzomyia. *J Hered*, 73(1):2–11. 11

REFERENCES

Griffiths-Jones, S. (2006). miRBase: the microRNA sequence database. *Methods Mol Biol*, 342:129–38. 24, 40, 46, 75, 103, 115

Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A., and Enright, A. J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res*, 34(Database issue):D140–4. 14, 24, 31, 46, 75, 79, 103

Griffiths-Jones, S., Saini, H. K., van Dongen, S., and Enright, A. J. (2008). miRBase: tools for microRNA genomics. *Nucleic Acids Res*, 36(Database issue):D154–8. 24, 47, 52, 75, 77, 103

Grimson, A., Farh, K. K.-H., Johnston, W. K., Garrett-Engele, P., Lim, L. P., and Bartel, D. P. (2007). MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell*, 27(1):91–105. 25, 27, 28, 86

Grün, D., Wang, Y.-L., Langenberger, D., Gunsalus, K. C., and Rajewsky, N. (2005). microRNA target predictions across seven Drosophila species and comparison to mammalian targets. *PLoS Comput Biol*, 1(1):e13. 104

Gunaratne, P. H., Coarfa, C., Soibam, B., and Tandon, A. (2012). mirna data analysis: next-gen sequencing. *Methods Mol Biol*, 822:273–88. 47

Guo, H., Ingolia, N. T., Weissman, J. S., and Bartel, D. P. (2010). Mammalian micrornas predominantly act to decrease target mrna levels. *Nature*, 466(7308):835–40. 112

Guo, Y., Liu, H., Yang, Z., Chen, J., Sun, Y., and Ren, X. (2012). Identification and characterization of mirnaome in tobacco (nicotiana tabacum) by deep sequencing combined with microarray. *Gene*, 501(1):24–32. 75

Hackenberg, M., Rodríguez-Ezpeleta, N., and Aransay, A. M. (2011). miranalyzer: an update on the detection and analysis of micrornas in high-throughput sequencing experiments. *Nucleic Acids Res*, 39(Web Server issue):W132–8. 52

Hackl, M., Jakobi, T., Blom, J., Doppmeier, D., Brinkrolf, K., Szczepanowski, R., Bernhart, S. H., Höner Zu Siederdissen, C., Bort, J. A. H., Wieser, M., Kunert, R., Jeffs, S., Hofacker, I. L., Goesmann, A., Pühler, A., Borth, N., and Grillari, J. (2011). Next-generation sequencing of the chinese hamster ovary microrna transcriptome: Identification, annotation and profiling of micrornas as targets for cellular engineering. *J Biotechnol*, 153(1-2):62–75. 47

Hammell, M., Long, D., Zhang, L., Lee, A., Carmack, C. S., Han, M., Ding, Y., and Ambros, V. (2008). mirWIP: microRNA target prediction based on microRNA-containing ribonucleoprotein-enriched transcripts. *Nat Methods*, 5(9):813–9. 31, 38

Han, Y. S., Thompson, J., Kafatos, F. C., and Barillas-Mury, C. (2000). Molecular interactions between Anopheles stephensi midgut cells and Plasmodium berghei: the time bomb theory of ookinete invasion of mosquitoes. *EMBO J*, 19(22):6030–40. 103

Hao, D.-C., Yang, L., Xiao, P.-G., and Liu, M. (2012). Identification of taxus micrornas and their targets with high-throughput sequencing and degradome analysis. *Physiol Plant*, 9999(9999). 75

Hargreaves, K., Hunt, R. H., Brooke, B. D., Mthembu, J., Weeto, M. M., Awolola, T. S., and Coetzee, M. (2003). Anopheles arabiensis and An. quadriannulatus resistance to DDT in South Africa. *Med Vet Entomol*, 17(4):417–22. 10, 11

Hargreaves, K., Koekemoer, L. L., Brooke, B. D., Hunt, R. H., Mthembu, J., and Coetzee, M. (2000). Anopheles funestus resistant to pyrethroid insecticides in South Africa. *Med Vet Entomol*, 14(2):181–9. 11

Harper, J. W. and Elledge, S. J. (2007). The dna damage response: ten years after. *Mol Cell*, 28(5):739–45. 103

Harris, R. A., Joshi, M., Jeoung, N. H., and Obayashi, M. (2005). Overview of the molecular and biochemical basis of branched-chain amino acid catabolism. *J Nutr*, 135(6 Suppl):1527S–30S. 104

Havecker, E. R. (2011). Detection of small rnas and micrornas using deep sequencing technology. *Methods Mol Biol*, 732:55–68. 47

Hay, S. I., Guerra, C. A., Tatem, A. J., Atkinson, P. M., and Snow, R. W. (2005). Urbanization, malaria transmission and disease burden in Africa. *Nat Rev Microbiol*, 3(1):81–90. 8

He, P.-a., Nie, Z., Chen, J., Chen, J., Lv, Z., Sheng, Q., Zhou, S., Gao, X., Kong, L., Wu, X., Jin, Y., and Zhang, Y. (2008). Identification and characteristics of microRNAs from Bombyx mori. *BMC Genomics*, 9:248. 87

Hendrickson, D. G., Hogan, D. J., McCullough, H. L., Myers, J. W., Herschlag, D., Ferrell, J. E., and Brown, P. O. (2009). Concordant regulation of translation and mRNA abundance for hundreds of targets of a human microRNA. *PLoS Biol*, 7(11):e1000238. 85

Hilgers, V., Bushati, N., and Cohen, S. M. (2010). Drosophila microRNAs 263a/b con-

fer robustness during development by protecting nascent sense organs from apoptosis. *PLoS Biol*, 8(6):e1000396. 46, 79

Hofacker, I., Fontana, W., Stadler, P., Bonhoeffer, S., Tacker, M., and Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Monatshefte f. Chemie*, 125:167–188. 22, 53

Huang, J. and Brumell, J. H. (2009). Autophagy in immunity against intracellular bacteria. *Curr Top Microbiol Immunol*, 335:189–215. 111

Huang, J., Hao, P., Chen, H., Hu, W., Yan, Q., Liu, F., and Han, Z.-G. (2009). Genome-wide identification of Schistosoma japonicum microRNAs using a deep-sequencing approach. *PLoS One*, 4(12):e8206. 79

Hughes, T. A. (2006). Regulation of gene expression by alternative untranslated regions. *Trends Genet*, 22(3):119–22. 27

Hunt, R. H., Brooke, B. D., Pillay, C., Koekemoer, L. L., and Coetzee, M. (2005). Laboratory selection for and characteristics of pyrethroid resistance in the malaria vector Anopheles funestus. *Med Vet Entomol*, 19(3):271–5. 56

Hunt, R. H., Coetzee, M., and Fettene, M. (1998). The Anopheles gambiae complex: a new species from Ethiopia. *Trans R Soc Trop Med Hyg*, 92(2):231–5. 10

Hussain, M. and Asgari, S. (2010). Functional analysis of a cellular microrna in insect host-ascovirus interaction. *J Virol*, 84(1):612–20. 46

Hussain, M., Frentiu, F. D., Moreira, L. A., O'Neill, S. L., and Asgari, S. (2011). Wolbachia uses host microRNAs to manipulate host gene expression and facili-

tate colonization of the dengue vector Aedes aegypti. *Proc Natl Acad Sci U S A*, 108(22):9250–5. 46, 81, 110

Hutson, S. M., Lieth, E., and LaNoue, K. F. (2001). Function of leucine in excitatory neurotransmitter metabolism in the central nervous system. *J Nutr*, 131(3):846S–850S. 106

Hutvágner, G. and Zamore, P. D. (2002). A microRNA in a multiple-turnover RNAi enzyme complex. *Science*, 297(5589):2056–60. 15, 46

Hyun, S., Lee, J. H., Jin, H., Nam, J., Namkoong, B., Lee, G., Chung, J., and Kim, V. N. (2009). Conserved microRNA miR-8/miR-200 and its target USH/FOG2 control growth by regulating PI3K. *Cell*, 139(6):1096–108. 79

Inukai, S., de Lencastre, A., Turner, M., and Slack, F. (2012). Novel micrornas differentially expressed during aging in the mouse brain. *PLoS One*, 7(7):e40028. 75, 81

Iovino, N., Pane, A., and Gaul, U. (2009). miR-184 has multiple roles in Drosophila female germline development. *Dev Cell*, 17(1):123–33. 79

Ji, Z., Wang, G., Xie, Z., Zhang, C., and Wang, J. (2012). Identification and characterization of microrna in the dairy goat (capra hircus) mammary gland by solexa deep-sequencing technology. *Mol Biol Rep*. 75

John, B., Enright, A. J., Aravin, A., Tuschl, T., Sander, C., and Marks, D. S. (2004). Human microRNA targets. *PLoS Biol*, 2(11):e363. 22, 29, 32

Jones-Rhoades, M. W. and Bartel, D. P. (2004). Computational identification of

plant microRNAs and their targets, including a stress-induced miRNA. *Mol Cell*, 14(6):787–99. 19, 21, 23

Jongwutiwes, S., Putaporntip, C., Iwasaki, T., Sata, T., and Kanbara, H. (2004). Naturally acquired Plasmodium knowlesi malaria in human, thailand. *Emerg Infect Dis*, 10(12):2211–3. 2

Juhila, J., Sipilä, T., Icay, K., Nicorici, D., Ellonen, P., Kallio, A., Korpelainen, E., Greco, D., and Hovatta, I. (2011). MicroRNA expression profiling reveals miRNA families regulating specific biological pathways in mouse frontal cortex and hippocampus. *PLoS One*, 6(6):e21495. 79

Jung, C.-H., Hansen, M. A., Makunin, I. V., Korbie, D. J., and Mattick, J. S. (2010). Identification of novel non-coding rnas using profiles of short sequence reads from next generation sequencing data. *BMC Genomics*, 11:77. 47

Kamau, L., Koekemoer, L. L., Hunt, R. H., and Coetzee, M. (2003). Anopheles parensis: the main member of the Anopheles funestus species group found resting inside human dwellings in Mwea area of central Kenya toward the end of the rainy season. *J Am Mosq Control Assoc*, 19(2):130–3. 12

Kang, M., Zhao, Q., Zhu, D., and Yu, J. (2012). Characterization of micrornas expression during maize seed development. *BMC Genomics*, 13(1):360. 47, 75

Keller, A., Backes, C., Leidinger, P., Kefer, N., Boisguerin, V., Barbacioru, C., Vogel, B., Matzas, M., Huwer, H., Katus, H. A., Stähler, C., Meder, B., and Meese, E. (2011). Next-generation sequencing identifies novel micrornas in peripheral blood of lung cancer patients. *Mol Biosyst*, 7(12):3187–99. 47

Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., and Segal, E. (2007). The role of site accessibility in microRNA target recognition. *Nat Genet*, 39(10):1278–84. 28, 31, 37, 86

Khorana, H. G. (1965). Polynucleotide synthesis and the genetic code. *Fed Proc*, 24(6):1473–87. 13

Kim, S.-K., Nam, J.-W., Rhee, J.-K., Lee, W.-J., and Zhang, B.-T. (2006). miTarget: microRNA target gene prediction using a support vector machine. *BMC Bioinformatics*, 7:411. 31, 36, 86

Kiriakidou, M., Nelson, P. T., Kouranov, A., Fitziev, P., Bouyioukos, C., Mourelatos, Z., and Hatzigeorgiou, A. (2004). A combined computational-experimental approach predicts human microRNA targets. *Genes Dev*, 18(10):1165–78. 86

Kiszewski, A., Mellinger, A., Spielman, A., Malaney, P., Sachs, S. E., and Sachs, J. (2004). A global index representing the stability of malaria transmission. *Am J Trop Med Hyg*, 70(5):486–98. 5, 6

Kloosterman, W. P., Wienholds, E., Ketting, R. F., and Plasterk, R. H. A. (2004). Substrate requirements for let-7 function in the developing zebrafish embryo. *Nucleic Acids Res*, 32(21):6284–91. 27

Koh, W., Sheng, C. T., Tan, B., Lee, Q. Y., Kuznetsov, V., Kiang, L. S., and Tanavde, V. (2010). Analysis of deep sequencing microRNA expression profile from human embryonic stem cells derived mesenchymal stem cells reveals possible role of let-7 microRNA family in downstream targeting of hepatic nuclear factor 4 alpha. *BMC Genomics*, 11 Suppl 1:S6. 54, 55, 75

Kong, B.-W. (2011). Identification of virus encoding micrornas using 454 flx sequencing platform. *Methods Mol Biol*, 733:81–91. 47

Kong, Y. W., Cannell, I. G., de Moor, C. H., Hill, K., Garside, P. G., Hamilton, T. L., Meijer, H. A., Dobbyn, H. C., Stoneley, M., Spriggs, K. A., Willis, A. E., and Bushell, M. (2008). The mechanism of micro-RNA-mediated translation repression is determined by the promoter of the target gene. *Proc Natl Acad Sci U S A*, 105(26):8866–71. 85

Kozomara, A. and Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res*, 39(Database issue):D152–7. 24, 40, 47, 75, 76, 87, 103, 115

Krawetz, S. A., Kruger, A., Lalancette, C., Tagett, R., Anton, E., Draghici, S., and Diamond, M. P. (2011). A survey of small rnas in human sperm. *Hum Reprod*, 26(12):3401–12. 47

Krek, A., Grün, D., Poy, M. N., Wolf, R., Rosenberg, L., Epstein, E. J., MacMenamin, P., da Piedade, I., Gunsalus, K. C., Stoffel, M., and Rajewsky, N. (2005). Combinatorial microRNA target predictions. *Nat Genet*, 37(5):495–500. 27, 31, 34, 86

Krol, J., Loedige, I., and Filipowicz, W. (2010). The widespread regulation of microrna biogenesis, function and decay. *Nat Rev Genet*, 11(9):597–610. 46

Kuhn, D. E., Martin, M. M., Feldman, D. S., Terry, Jr, A. V., Nuovo, G. J., and Elton, T. S. (2008). Experimental validation of miRNA targets. *Methods*, 44(1):47–54. 85

Kumar, S., Christophides, G. K., Cantera, R., Charles, B., Han, Y. S., Meister, S., Dimopoulos, G., Kafatos, F. C., and Barillas-Mury, C. (2003). The role of reac-

tive oxygen species on Plasmodium melanotic encapsulation in Anopheles gambiae. *Proc Natl Acad Sci U S A*, 100(24):14139–44. 110

Kumar, S., Gupta, L., Han, Y. S., and Barillas-Mury, C. (2004). Inducible peroxidases mediate nitration of Anopheles midgut cells undergoing apoptosis in response to Plasmodium invasion. *J Biol Chem*, 279(51):53475–82. 103

Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. *Science*, 294(5543):853–8. 20, 21

Lai, E. C. (2002). Micro rnas are complementary to 3' utr sequence motifs that mediate negative post-transcriptional regulation. *Nat Genet*, 30(4):363–4. 27

Lai, E. C., Tomancak, P., Williams, R. W., and Rubin, G. M. (2003). Computational identification of Drosophila microRNA genes. *Genome Biol*, 4(7):R42. 14, 20, 21, 22, 87, 115

Lau, N. C., Lim, L. P., Weinstein, E. G., and Bartel, D. P. (2001). An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science*, 294(5543):858–62. 20

Leaman, D., Chen, P. Y., Fak, J., Yalcin, A., Pearce, M., Unnerstall, U., Marks, D. S., Sander, C., Tuschl, T., and Gaul, U. (2005). Antisense-mediated depletion reveals essential and specific functions of microRNAs in Drosophila development. *Cell*, 121(7):1097–108. 103, 104

Lee, R. C. and Ambros, V. (2001). An extensive class of small RNAs in Caenorhabditis elegans. *Science*, 294(5543):862–4. 20, 21

REFERENCES

Lee, R. C., Feinbaum, R. L., and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):843–54. 13, 14

Lee, Y., Jeon, K., Lee, J.-T., Kim, S., and Kim, V. N. (2002). MicroRNA maturation: stepwise processing and subcellular localization. *EMBO J*, 21(17):4663–70. 15

Lee, Y., Kim, M., Han, J., Yeom, K.-H., Lee, S., Baek, S. H., and Kim, V. N. (2004). MicroRNA genes are transcribed by RNA polymerase II. *EMBO J*, 23(20):4051–60. 13

Legendre, M., Lambert, A., and Gautheret, D. (2005). Profile-based detection of microRNA precursors in animal genomes. *Bioinformatics*, 21(7):841–5. 21

Leidner, R. S., Ravi, L., Leahy, P., Chen, Y., Bednarchik, B., Streppel, M., Canto, M., Wang, J. S., Maitra, A., Willis, J., Markowitz, S. D., Barnholtz-Sloan, J., Adams, M. D., Chak, A., and Guda, K. (2012). The microrna, mir-31 and mir-375, as candidate markers in barrett's esophageal carcinogenesis. *Genes Chromosomes Cancer*, 51(5):473–9. 47

Leitner, A. (2009). MicroRNA target prediction. Master's thesis, Graz University of Technology, Graz, Austria. 39

Lekprasert, P., Mayhew, M., and Ohler, U. (2011). Assessing the utility of thermodynamic features for microrna target prediction under relaxed seed and no conservation requirements. *PLoS One*, 6(6):e20622. 28

Lewis, B. P., Burge, C. B., and Bartel, D. P. (2005). Conserved seed pairing, often

flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1):15–20. 25, 31, 33, 87

Lewis, B. P., Shih, I.-h., Jones-Rhoades, M. W., Bartel, D. P., and Burge, C. B. (2003). Prediction of mammalian microRNA targets. *Cell*, 115(7):787–98. 25, 29, 32, 117

Li, H., Ruan, J., and Durbin, R. (2008a). Mapping short dna sequencing reads and calling variants using mapping quality scores. *Genome Res*, 18(11):1851–8. 51

Li, P., Peng, J., Hu, J., Xu, Z., Xie, W., and Yuan, L. (2011). Localized expression pattern of miR-184 in Drosophila. *Mol Biol Rep*, 38(1):355–8. 79

Li, R., Li, Y., Kristiansen, K., and Wang, J. (2008b). Soap: short oligonucleotide alignment program. *Bioinformatics*, 24(5):713–4. 51

Li, R., Yu, C., Li, Y., Lam, T.-W., Yiu, S.-M., Kristiansen, K., and Wang, J. (2009a). Soap2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15):1966–7. 51

Li, S., Mead, E. A., Liang, S., and Tu, Z. (2009b). Direct sequencing and expression analysis of a large number of miRNAs in Aedes aegypti and a multi-species survey of novel mosquito miRNAs. *BMC Genomics*, 10:581. 40, 76, 78, 79, 81, 87

Li, X., Wang, X., Zhang, S., Liu, D., Duan, Y., and Dong, W. (2012). Identification of soybean micrornas involved in soybean cyst nematode infection by deep sequencing. *PLoS One*, 7(6):e39650. 75

Lim, L. P., Glasner, M. E., Yekta, S., Burge, C. B., and Bartel, D. P. (2003a). Vertebrate microRNA genes. *Science*, 299(5612):1540. 20, 22

Lim, L. P., Lau, N. C., Garrett-Engele, P., Grimson, A., Schelter, J. M., Castle, J., Bartel, D. P., Linsley, P. S., and Johnson, J. M. (2005). Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*, 433(7027):769–73. 29, 37, 46, 112

Lim, L. P., Lau, N. C., Weinstein, E. G., Abdelhakim, A., Yekta, S., Rhoades, M. W., Burge, C. B., and Bartel, D. P. (2003b). The microRNAs of Caenorhabditis elegans. *Genes Dev*, 17(8):991–1008. 14, 20, 21, 22, 98

Lindow, M. and Gorodkin, J. (2007). Principles and limitations of computational microRNA gene and target finding. *DNA Cell Biol*, 26(5):339–51. 30, 86

Lindow, M. and Krogh, A. (2005). Computational evidence for hundreds of non-conserved plant microRNAs. *BMC Genomics*, 6:119. 19

Ling, K.-H., Brautigan, P. J., Hahn, C. N., Daish, T., Rayner, J. R., Cheah, P.-S., Raison, J. M., Piltz, S., Mann, J. R., Mattiske, D. M., Thomas, P. Q., Adelson, D. L., and Scott, H. S. (2011). Deep sequencing analysis of the developing mouse brain reveals a novel microrna. *BMC Genomics*, 12:176. 81

Linsen, S. E. V., de Wit, E., Janssens, G., Heater, S., Chapman, L., Parkin, R. K., Fritz, B., Wyman, S. K., de Bruijn, E., Voest, E. E., Kuersten, S., Tewari, M., and Cuppen, E. (2009). Limitations and possibilities of small RNA digital gene expression profiling. *Nat Methods*, 6(7):474–6. 54

Liu, C.-M., Wong, T., Wu, E., Luo, R., Yiu, S.-M., Li, Y., Wang, B., Yu, C., Chu, X., Zhao, K., Li, R., and Lam, T.-W. (2012a). Soap3: Ultra-fast gpu-based parallel alignment tool for short reads. *Bioinformatics*. 51

Liu, F., Peng, W., Li, Z., Li, W., Li, L., Pan, J., Zhang, S., Miao, Y., Chen, S., and Su, S. (2012b). Next-generation small rna sequencing for micrornas profiling in apis mellifera: comparison between nurses and foragers. *Insect Mol Biol*, 21(3):297–303. 47

Liu, H., Yue, D., Chen, Y., Gao, S.-J., and Huang, Y. (2010a). Improving performance of mammalian microRNA target prediction. *BMC Bioinformatics*, 11:476. 86

Liu, S., Li, D., Li, Q., Zhao, P., Xiang, Z., and Xia, Q. (2010b). MicroRNAs of Bombyx mori identified by Solexa sequencing. *BMC Genomics*, 11:148. 76, 81, 87

Liu, S., Zhang, L., Li, Q., Zhao, P., Duan, J., Cheng, D., Xiang, Z., and Xia, Q. (2009). MicroRNA expression profiling during the life cycle of the silkworm (Bombyx mori). *BMC Genomics*, 10:455. 79, 80

Llave, C., Kasschau, K. D., Rector, M. A., and Carrington, J. C. (2002). Endogenous and silencing-associated small rnas in plants. *Plant Cell*, 14(7):1605–19. 19

Long, D., Lee, R., Williams, P., Chan, C. Y., Ambros, V., and Ding, Y. (2007). Potent effect of target structure on microRNA function. *Nat Struct Mol Biol*, 14(4):287–94. 28

Loots, G. G., Locksley, R. M., Blankespoor, C. M., Wang, Z. E., Miller, W., Rubin, E. M., and Frazer, K. A. (2000). Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science*, 288(5463):136–40. 23

Lund, E., Güttinger, S., Calado, A., Dahlberg, J. E., and Kutay, U. (2004). Nuclear export of microRNA precursors. *Science*, 303(5654):95–8. 15

Lytle, J. R., Yario, T. A., and Steitz, J. A. (2007). Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5'UTR as in the 3'UTR. *Proc Natl Acad Sci U S A*, 104(23):9667–72. 15, 27, 85

Ma, Z.-L., Yang, H.-Y., and Tien, P. (2003). Progress of miRNA and its functions in eukaryotes. *Yi Chuan Xue Bao*, 30(7):693–6. 13

Macdonald, P. M. and Struhl, G. (1986). A molecular gradient in early Drosophila embryos and its role in specifying the body pattern. *Nature*, 324(6097):537–45. 46, 106

Majoros, W. H. and Ohler, U. (2007). Spatial preferences of microRNA targets in 3' untranslated regions. *BMC Genomics*, 8:152. 27

Maragkakis, M., Reczko, M., Simossis, V. A., Alexiou, P., Papadopoulos, G. L., Dalamagas, T., Giannopoulos, G., Goumas, G., Koukis, E., Kourtis, K., Vergoulis, T., Koziris, N., Sellis, T., Tsanakas, P., and Hatzigeorgiou, A. G. (2009). Diana-microt web server: elucidating microRNA functions through target prediction. *Nucleic Acids Res*, 37(Web Server issue):W273–6. 31, 33

Marco, A., Hooks, K., and Griffiths-Jones, S. (2012). Evolution and function of the extended miR-2 microRNA family. *RNA Biol*, 9(3). 46, 103

Marín, R. M. and Vanícek, J. (2011). Efficient use of accessibility in microrna target prediction. *Nucleic Acids Res*, 39(1):19–29. 28

Marín, R. M. and Vaníček, J. (2012). Optimal use of conservation and accessibility filters in microrna target prediction. *PLoS One*, 7(2):e32208. 28

Matukumalli, L. K., Grefenstette, J. J., Sonstegard, T. S., and Van Tassell, C. P. (2004). Est-page–managing and analyzing est data. *Bioinformatics*, 20(2):286–8. 21

Mazière, P. and Enright, A. J. (2007). Prediction of microRNA targets. *Drug Discov Today*, 12(11-12):452–8. 25, 27, 29, 30, 37

Mead, E. A. and Tu, Z. (2008). Cloning, characterization, and expression of microRNAs from the Asian malaria mosquito, Anopheles stephensi. *BMC Genomics*, 9:244. 40, 75, 81, 115

Mendes, N. D., Freitas, A. T., and Sagot, M.-F. (2009). Current tools for the identification of miRNA genes and their targets. *Nucleic Acids Res*, 37(8):2419–33. 14, 30, 50, 86

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nat Rev Genet*, 11(1):31–46. 47

Miranda, K. C., Huynh, T., Tay, Y., Ang, Y.-S., Tam, W.-L., Thomson, A. M., Lim, B., and Rigoutsos, I. (2006). A pattern-based method for the identification of microRNA binding sites and their corresponding heteroduplexes. *Cell*, 126(6):1203–17. 31, 86

Misra, J. R., Horner, M. A., Lam, G., and Thummel, C. S. (2011). Transcriptional regulation of xenobiotic detoxification in Drosophila. *Genes Dev*, 25(17):1796–806. 110

Mohorianu, I., Schwach, F., Jing, R., Lopez-Gomollon, S., Moxon, S., Szittya, G., Sorefan, K., Moulton, V., and Dalmay, T. (2011). Profiling of short rnas during fleshy fruit development reveals stage-specific srnaome expression patterns. *Plant J*, 67(2):232–46. 47

Molina-Cruz, A., DeJong, R. J., Charles, B., Gupta, L., Kumar, S., Jaramillo-Gutierrez, G., and Barillas-Mury, C. (2008). Reactive oxygen species modulate Anopheles gambiae immunity against bacteria and Plasmodium. *J Biol Chem*, 283(6):3217–23. 110

Morin, R. D., O'Connor, M. D., Griffith, M., Kuchenbauer, F., Delaney, A., Prabhu, A.-L., Zhao, Y., McDonald, H., Zeng, T., Hirst, M., Eaves, C. J., and Marra, M. A. (2008). Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res*, 18(4):610–21. 51, 53

Moss, E. G. (2007). Heterochronic genes and the nature of developmental time. *Curr Biol*, 17(11):R425–34. 98

Motameny, S., Wolters, S., Nürnberg, P., and Schumacher, B. (2010). Next Generation Sequencing of miRNAs - Strategies, Resources and Methods. *Genes*, 1(1):70–84. 48, 49, 51, 54, 55, 115

Mouatcho, J. C., Hargreaves, K., Koekemoer, L. L., Brooke, B. D., Oliver, S. V., Hunt, R. H., and Coetzee, M. (2007). Indoor collections of the anopheles funestus group (diptera: Culicidae) in sprayed houses in northern kwazulu-natal, south africa. *Malar J*, 6:30. 11

Mückstein, U., Tafer, H., Hackermüller, J., Bernhart, S. H., Stadler, P. F., and Hofacker, I. L. (2006). Thermodynamics of rna-rna binding. *Bioinformatics*, 22(10):1177–82. 28

Nair, K. S. and Short, K. R. (2005). Hormonal and signaling role of branched-chain amino acids. *J Nutr*, 135(6 Suppl):1547S–52S. 104

Nam, J.-W., Shin, K.-R., Han, J., Lee, Y., Kim, V. N., and Zhang, B.-T. (2005). Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic Acids Res*, 33(11):3570–81. 21

Nam, S., Kim, B., Shin, S., and Lee, S. (2008). miRGator: an integrated system for functional annotation of microRNAs. *Nucleic Acids Res*, 36(Database issue):D159–64. 87

Neilson, J. R. and Sharp, P. A. (2008). Small rna regulators of gene expression. *Cell*, 134(6):899–902. 27

Nielsen, C. B., Shomron, N., Sandberg, R., Hornstein, E., Kitzman, J., and Burge, C. B. (2007). Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. *RNA*, 13(11):1894–910. 19, 25, 26, 31

Nikaki, A., Piperi, C., and Papavassiliou, A. G. (2012). Role of micrornas in glioma-genesis: targeting mirnas in glioblastoma multiforme therapy. *Expert Opin Investig Drugs*. 47

Nikopoulos, K., Gilissen, C., Hoischen, A., van Nouhuys, C. E., Boonstra, F. N., Blok-land, E. A. W., Arts, P., Wieskamp, N., Strom, T. M., Ayuso, C., Tilanus, M. A. D., Bouwhuis, S., Mukhopadhyay, A., Scheffer, H., Hoefsloot, L. H., Veltman, J. A., Cremers, F. P. M., and Collin, R. W. J. (2010). Next-generation sequencing of a 40 mb linkage interval reveals tspan12 mutations in patients with familial exudative vitreoretinopathy. *Am J Hum Genet*, 86(2):240–7. 47

Ohler, U., Yekta, S., Lim, L. P., Bartel, D. P., and Burge, C. B. (2004). Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. *RNA*, 10(9):1309–22. 23

Ohlrogge, J. and Benning, C. (2000). Unraveling plant metabolism by EST analysis. *Curr Opin Plant Biol*, 3(3):224–8. 21

Oshlack, A. and Wakefield, M. J. (2009). Transcript length bias in RNA-seq data confounds systems biology. *Biol Direct*, 4:14. 76

Pantano, L., Estivill, X., and Martí, E. (2010). Seqbuster, a bioinformatic tool for the processing and analysis of small rnas datasets, reveals ubiquitous mirna modifications in human embryonic cells. *Nucleic Acids Res*, 38(5):e34. 52

Pasquinelli, A. E., Reinhart, B. J., Slack, F., Martindale, M. Q., Kuroda, M. I., Maller, B., Hayward, D. C., Ball, E. E., Degnan, B., Müller, P., Spring, J., Srinivasan, A., Fishman, M., Finnerty, J., Corbo, J., Levine, M., Leahy, P., Davidson, E., and Ruvkun, G. (2000). Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, 408(6808):86–9. 14, 20, 21, 41, 98

Pearson, W. R., Wood, T., Zhang, Z., and Miller, W. (1997). Comparison of dna sequences with protein sequences. *Genomics*, 46(1):24–36. 51

Pelaez, P., Trejo, M. S., Iniguez, L. P., Estrada-Navarrete, G., Covarrubias, A. A., Reyes, J. L., and Sanchez, F. (2012). Identification and characterization of microRNAs in Phaseolus vulgaris by high-throughput sequencing. *BMC Genomics*, 13(1):83. 54, 76

Pérez-Quintero, Á. L., Quintero, A., Urrego, O., Vanegas, P., and López, C. (2012). Bioinformatic identification of cassava mirnas differentially expressed in response to infection by xanthomonas axonopodis pv. manihotis. *BMC Plant Biol*, 12:29. 47, 115

Persson, H., Kvist, A., Rego, N., Staaf, J., Vallon-Christersson, J., Luts, L., Loman, N., Jonsson, G., Naya, H., Hoglund, M., Borg, A., and Rovira, C. (2011). Identification of new microrna in paired normal and tumor breast tissue suggests a dual role for the erbb2/her2 gene. *Cancer Res*, 71(1):78–86. 47

Peter, M. E. (2010). Targeting of mrnas by multiple mirnas: the next step. *Oncogene*, 29(15):2161–4. 112

Pfeffer, S., Sewer, A., Lagos-Quintana, M., Sheridan, R., Sander, C., Grässer, F. A., van Dyk, L. F., Ho, C. K., Shuman, S., Chien, M., Russo, J. J., Ju, J., Randall, G., Lindenbach, B. D., Rice, C. M., Simon, V., Ho, D. D., Zavolan, M., and Tuschl, T. (2005). Identification of microrna of the herpesvirus family. *Nat Methods*, 2(4):269–76. 40

Pillai, R. S. (2005). MicroRNA function: multiple mechanisms for a tiny RNA? *RNA*, 11(12):1753–61. 37

Place, R. F., Li, L.-C., Pookot, D., Noonan, E. J., and Dahiya, R. (2008). MicroRNA-373 induces expression of genes with complementary promoter sequences. *Proc Natl Acad Sci U S A*, 105(5):1608–13. 15

Pokrzywa, R. (2008). New method for yeast identification using burrows-wheeler transform. *J Bioinform Comput Biol*, 6(2):403–13. 51

Ragan, C., Zuker, M., and Ragan, M. A. (2011). Quantitative prediction of mirna-mrna interaction based on equilibrium concentrations. *PLoS Comput Biol*, 7(2):e1001090. 28

REFERENCES

Rajewsky, N. (2006). MicroRNA target predictions in animals. *Nat Genet*, 38 Suppl:S8–13. 29

Rajewsky, N. and Socci, N. D. (2004). Computational identification of microRNA targets. *Dev Biol*, 267(2):529–35. 25, 30

Rehmsmeier, M., Steffen, P., Hochsmann, M., and Giegerich, R. (2004). Fast and effective prediction of microRNA/target duplexes. *RNA*, 10(10):1507–17. 31, 34, 86, 89

Reinhart, B. J., Slack, F. J., Basson, M., Pasquinelli, A. E., Bettinger, J. C., Rougvie, A. E., Horvitz, H. R., and Ruvkun, G. (2000). The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature*, 403(6772):901–6. 14, 98

Reinhart, B. J., Weinstein, E. G., Rhoades, M. W., Bartel, B., and Bartel, D. P. (2002). MicroRNAs in plants. *Genes Dev*, 16(13):1616–26. 15, 19, 25

Rhoades, M. W., Reinhart, B. J., Lim, L. P., Burge, C. B., Bartel, B., and Bartel, D. P. (2002). Prediction of plant microRNA targets. *Cell*, 110(4):513–20. 19

Rich, A. and RajBhandary, U. L. (1976). Transfer rna: molecular structure, sequence, and properties. *Annu Rev Biochem*, 45:805–60. 13

Ritchie, W., Rajasekhar, M., Flamant, S., and Rasko, J. E. J. (2009). Conserved expression patterns predict microrna targets. *PLoS Comput Biol*, 5(9):e1000513. 29

Robins, H., Li, Y., and Padgett, R. W. (2005). Incorporating structure to predict microRNA targets. *Proc Natl Acad Sci U S A*, 102(11):4006–9. 19, 28, 37

Rodriguez, A., Griffiths-Jones, S., Ashurst, J. L., and Bradley, A. (2004). Identification of mammalian microRNA host genes and transcription units. *Genome Res*, 14(10A):1902–10. 14, 22

Ronen, R., Gan, I., Modai, S., Sukacheov, A., Dror, G., Halperin, E., and Shomron, N. (2010). miRNAkey: a software for microRNA deep sequencing analysis. *Bioinformatics*, 26(20):2615–6. 51

Rusinov, V., Baev, V., Minkov, I. N., and Tabler, M. (2005). Microinspector: a web tool for detection of miRNA binding sites in an RNA sequence. *Nucleic Acids Res*, 33(Web Server issue):W696–700. 31

Ruvkun, G., Wightman, B., and Ha, I. (2004). The 20 years it took to recognize the importance of tiny RNAs. *Cell*, 116(2 Suppl):S93–6, 2 p following S96. 14

Sackton, K. L., Buehner, N. A., and Wolfner, M. F. (2007). Modulation of MAPK activities during egg activation in Drosophila. *Fly (Austin)*, 1(4):222–7. 106, 108

Saetrom, O., Snøve, Jr, O., and Saetrom, P. (2005). Weighted sequence motifs as an improved seeding step in microRNA target prediction algorithms. *RNA*, 11(7):995–1003. 31, 36, 86

Saito, T. and Saetrom, P. (2010). MicroRNAs-targeting and target prediction. *N Biotechnol*, 27(3):243–9. 25, 30, 86

Schweet, R. and Heintz, R. (1966). Protein synthesis. *Annu Rev Biochem*, 35:723–58. 13

Seffens, W. and Digby, D. (1999). mRNAs have greater negative folding free en-

ergies than shuffled or codon choice randomized sequences. *Nucleic Acids Res*, 27(7):1578–84. 17

Seitz, H., Royo, H., Bortolin, M.-L., Lin, S.-P., Ferguson-Smith, A. C., and Cavaillé, J. (2004). A large imprinted microRNA gene cluster at the mouse Dlk1-Gtl2 domain. *Genome Res*, 14(9):1741–8. 22

Selbach, M., Schwanhäusser, B., Thierfelder, N., Fang, Z., Khanin, R., and Rajewsky, N. (2008). Widespread changes in protein synthesis induced by microRNAs. *Nature*, 455(7209):58–63. 38, 112

Sethupathy, P., Megraw, M., and Hatzigeorgiou, A. G. (2006). A guide through present computational approaches for the identification of mammalian microRNA targets. *Nat Methods*, 3(11):881–6. 28, 115

Shamimuzzaman, M. and Vodkin, L. (2012). Identification of soybean seed developmental stage-specific and tissue-specific mirna targets by degradome sequencing. *BMC Genomics*, 13:310. 47

Sharakhov, I. V., Serazin, A. C., Grushko, O. G., Dana, A., Lobo, N., Hillenmeyer, M. E., Westerman, R., Romero-Severson, J., Costantini, C., Sagnon, N., Collins, F. H., and Besansky, N. J. (2002). Inversions and gene order shuffling in anopheles gambiae and a. funestus. *Science*, 298(5591):182–5. 8

Shiff, C. J., Minjas, J. N., Hall, T., Hunt, R. H., Lyimo, S., and Davis, J. R. (1995). Malaria infection potential of anopheline mosquitoes sampled by light trapping indoors in coastal Tanzanian villages. *Med Vet Entomol*, 9(3):256–62. 10, 11

Sinden, R. E. and Billingsley, P. F. (2001). Plasmodium invasion of mosquito cells: hawk or dove? *Trends Parasitol*, 17(5):209–12. 12

Skalsky, R. L., Vanlandingham, D. L., Scholle, F., Higgs, S., and Cullen, B. R. (2010). Identification of microRNAs expressed in two mosquito vectors, Aedes albopictus and Culex quinquefasciatus. *BMC Genomics*, 11:119. 13, 40, 46, 75, 76, 87, 110

Smalheiser, N. R. and Torvik, V. I. (2005). Mammalian microRNAs derived from genomic repeats. *Trends Genet*, 21(6):322–6. 14

Stark, A., Brennecke, J., Bushati, N., Russell, R. B., and Cohen, S. M. (2005). Animal microRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. *Cell*, 123(6):1133–46. 27, 37

Stark, A., Brennecke, J., Russell, R. B., and Cohen, S. M. (2003). Identification of Drosophila microRNA targets. *PLoS Biol*, 1(3):E60. 25, 27, 29, 30, 31, 32, 104, 117

Stark, A., Kheradpour, P., Parts, L., Brennecke, J., Hodges, E., Hannon, G. J., and Kellis, M. (2007). Systematic discovery and characterization of fly microRNAs using 12 Drosophila genomes. *Genome Res*, 17(12):1865–79. 87

Sunkar, R., Zhou, X., Zheng, Y., Zhang, W., and Zhu, J.-K. (2008). Identification of novel and candidate miRNAs in rice by high throughput sequencing. *BMC Plant Biol*, 8:25. 78

Swellengrebel, N., Annecke, S., and De Meillon, B. (1931). Malaria investigations in some parts of the Transvaal and Zululand. *South African Institute for Medical Research*, 4:245–274. 10

REFERENCES

't Hoen, P. A. C., Ariyurek, Y., Thygesen, H. H., Vreugdenhil, E., Vossen, R. H. A. M., de Menezes, R. X., Boer, J. M., van Ommen, G.-J. B., and den Dunnen, J. T. (2008). Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Res*, 36(21):e141. 48

Tachibana, K., Tanaka, D., Isobe, T., and Kishimoto, T. (2000). c-Mos forces the mitotic cell cycle to undergo meiosis ii to produce haploid gametes. *Proc Natl Acad Sci U S A*, 97(26):14301–6. 108

Tadros, W., Houston, S. A., Bashirullah, A., Cooperstock, R. L., Semotok, J. L., Reed, B. H., and Lipshitz, H. D. (2003). Regulation of maternal transcript destabilization during egg activation in Drosophila. *Genetics*, 164(3):989–1001. 106

Tagle, D. A., Koop, B. F., Goodman, M., Slightom, J. L., Hess, D. L., and Jones, R. T. (1988). Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J Mol Biol*, 203(2):439–55. 23

Tang, Y., Liu, D., Zhang, L., Ingvarsson, S., and Chen, H. (2011). Quantitative analysis of mirna expression in seven human foetal and adult organs. *PLoS One*, 6(12):e28730. 77

Tanzer, A. and Stadler, P. F. (2004). Molecular evolution of a microRNA cluster. *J Mol Biol*, 339(2):327–35. 22

Teleman, A. A., Maitra, S., and Cohen, S. M. (2006). Drosophila lacking microRNA miR-278 are defective in energy homeostasis. *Genes Dev*, 20(4):417–22. 46, 80

Thadani, R. and Tammi, M. T. (2006). MicroTar: predicting microRNA targets from RNA duplexes. *BMC Bioinformatics*, 7 Suppl 5:S20. 31, 37, 89

Thomson, D. W., Bracken, C. P., and Goodall, G. J. (2011). Experimental strategies for microRNA target identification. *Nucleic Acids Res*, 39(16):6845–53. 85, 112

Thummel, C. S. (2001). Molecular mechanisms of developmental timing in C. elegans and Drosophila. *Dev Cell*, 1(4):453–65. 98

Tian, G., Yin, X., Luo, H., Xu, X., Bolund, L., Zhang, X., Gan, S.-Q., and Li, N. (2010). Sequencing bias: comparison of different protocols of microrna library construction. *BMC Biotechnol*, 10:64. 75

Tokunaga, C., Yoshino, K.-i., and Yonezawa, K. (2004). mtor integrates amino acid- and energy-sensing pathways. *Biochem Biophys Res Commun*, 313(2):443–6. 106

Tunquist, B. J. and Maller, J. L. (2003). Under arrest: cytostatic factor (csf)-mediated metaphase arrest in vertebrate eggs. *Genes Dev*, 17(6):683–710. 108

Ulitsky, I., Laurent, L. C., and Shamir, R. (2010). Towards computational prediction of microrna function and activity. *Nucleic Acids Res*, 38(15):e160. 86

Vallejo, D. M., Caparros, E., and Dominguez, M. (2011). Targeting Notch signalling by the conserved miR-8/200 microRNA family in development and cancer cells. *EMBO J*, 30(4):756–69. 46, 79

van Rooij, E., Marshall, W. S., and Olson, E. N. (2008). Toward microrna-based therapeutics for heart disease: the sense in antisense. *Circ Res*, 103(9):919–28. 87

Vasudevan, S., Tong, Y., and Steitz, J. A. (2007). Switching from repression to activation: microRNAs can up-regulate translation. *Science*, 318(5858):1931–4. 15

Vlachou, D., Schlegelmilch, T., Christophides, G. K., and Kafatos, F. C. (2005). Functional genomic analysis of midgut epithelial responses in anopheles during plasmodium invasion. *Curr Biol*, 15(13):1185–95. 12

Wang, H.-W., Noland, C., Siridechadilok, B., Taylor, D. W., Ma, E., Felderer, K., Doudna, J. A., and Nogales, E. (2009a). Structural insights into rna processing by the human risc-loading complex. *Nat Struct Mol Biol*, 16(11):1148–53. 27

Wang, W.-C., Lin, F.-M., Chang, W.-C., Lin, K.-Y., Huang, H.-D., and Lin, N.-S. (2009b). mirexpress: analyzing high-throughput sequencing data for profiling microrna expression. *BMC Bioinformatics*, 10:328. 47, 52

Wang, X. (2008). mirDB: a microRNA target prediction and functional annotation database with a wiki interface. *RNA*, 14(6):1012–7. 31

Wang, X. and El Naqa, I. M. (2008). Prediction of both conserved and nonconserved microRNA targets in animals. *Bioinformatics*, 24(3):325–32. 25, 37, 86

Wang, X., Zhang, J., Li, F., Gu, J., He, T., Zhang, X., and Li, Y. (2005). MicroRNA identification based on sequence and structure alignment. *Bioinformatics*, 21(18):3610–4. 21

Weaver, D. B., Anzola, J. M., Evans, J. D., Reid, J. G., Reese, J. T., Childs, K. L., Zdobnov, E. M., Samanta, M. P., Miller, J., and Elsik, C. G. (2007). Computational and transcriptional evidence for microRNAs in the honey bee genome. *Genome Biol*, 8(6):R97. 87, 115

Weber, M. J. (2005). New human and mouse microRNA genes found by homology search. *FEBS J*, 272(1):59–73. 20, 21

REFERENCES

Wei, C., Salichos, L., Wittgrove, C. M., Rokas, A., and Patton, J. G. (2012). Transcriptome-wide analysis of small rna expression in early zebrafish development. *RNA*, 18(5):915–29. 47

Wei, Y., Chen, S., Yang, P., Ma, Z., and Kang, L. (2009). Characterization and comparative profiling of the small RNA transcriptomes in two phases of locust. *Genome Biol*, 10(1):R6. 47, 76, 80

Wenguang, Z., Jianghong, W., Jinquan, L., and Yashizawa, M. (2007). A subset of skin-expressed microRNAs with possible roles in goat and sheep hair growth based on expression profiling of mammalian microRNAs. *OMICS*, 11(4):385–96. 79

White, G. (1972). Confirmation that Anopheles longipalpis (theobald) and Anopheles confusus Evans and Leeson occur in Ethiopia.nopheles confusus evans and leeson occur in ethiopia. *Mosquito systematics*, 4:131–132. Author(s) - White, G. B. 10

WHO (2005). *Malaria Control in Complex Emergencies - An Inter-Agency Field Handbook*. World Health Organization, Geneva, Switzerland. 5, 8

WHO (2011). World malaria report 2011. *Geneva, Switzerland: World Health Organization*. 2, 3

Wightman, B., Ha, I., and Ruvkun, G. (1993). Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in c. elegans. *Cell*, 75(5):855–62. 14

Winter, F., Edaye, S., Hüttenhofer, A., and Brunel, C. (2007). Anopheles gambiae miRNAs as actors of defence reaction against Plasmodium invasion. *Nucleic Acids Res*, 35(20):6953–62. 13, 29, 40, 41, 46, 75, 80, 81, 87, 110

Wu, J., Bao, J., Wang, L., Hu, Y., and Xu, C. (2011a). MicroRNA-184 downregulates nuclear receptor corepressor 2 in mouse spermatogenesis. *BMC Dev Biol*, 11:64. 79

Wu, Q., Lu, Z., Li, H., Lu, J., Guo, L., and Ge, Q. (2011b). Next-generation sequencing of micrornas for breast cancer detection. *J Biomed Biotechnol*, 2011:597145. 47

Wu, S., Huang, S., Ding, J., Zhao, Y., Liang, L., Liu, T., Zhan, R., and He, X. (2010). Multiple micrornas modulate p21cip1/waf1 expression by directly targeting its 3' untranslated region. *Oncogene*, 29(15):2302–8. 112

Wu, X. and Watson, M. (2009). CoRNA: testing gene lists for regulation by microRNAs. *Bioinformatics*, 25(6):832–3. 90

Xia, J. H., He, X. P., Bai, Z. Y., and Yue, G. H. (2011). Identification and characterization of 63 microRNAs in the Asian seabass Lates calcarifer. *PLoS One*, 6(3):e17537. 79

Xie, Z., Allen, E., Fahlgren, N., Calamar, A., Givan, S. A., and Carrington, J. C. (2005). Expression of arabidopsis miRNA genes. *Plant Physiol*, 138(4):2145–54. 22

Xu, C., Li, C. Y.-T., and Kong, A.-N. T. (2005). Induction of phase i, ii and iii drug metabolism/transport by xenobiotics. *Arch Pharm Res*, 28(3):249–68. 108

Xu, H., Wang, X., Du, Z., and Li, N. (2006). Identification of micrornas from different tissues of chicken embryo and adult chicken. *FEBS Lett*, 580(15):3610–6. 79

Yang, Y., Wang, Y.-P., and Li, K.-B. (2008). MiRTif: a support vector machine-based microRNA target interaction filter. *BMC Bioinformatics*, 9 Suppl 12:S4. 86

Yao, M.-J., Chen, G., Zhao, P.-P., Lu, M.-H., Jian, J., Liu, M.-F., and Yuan, X.-B. (2012). Transcriptome analysis of micrornas in developing cerebral cortex of rat. *BMC Genomics*, 13(1):232. 47

Yao, Y., Guo, G., Ni, Z., Sunkar, R., Du, J., Zhu, J.-K., and Sun, Q. (2007). Cloning and characterization of microRNAs from wheat (Triticum aestivum L.). *Genome Biol*, 8(6):R96. 70, 115

Yi, R., Qin, Y., Macara, I. G., and Cullen, B. R. (2003). Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes Dev*, 17(24):3011–6. 15

Yoo, S. J., Huh, J. R., Muro, I., Yu, H., Wang, L., Wang, S. L., Feldman, R. M. R., Clem, R. J., Müller, H.-A. J., and Hay, B. A. (2002). Hid, rpr and grim negatively regulate diap1 levels through distinct mechanisms. *Nat Cell Biol*, 4(6):416–24. 101

Yousef, M., Jung, S., Kossenkov, A. V., Showe, L. C., and Showe, M. K. (2007). Naïve Bayes for microRNA target predictions–machine learning for microRNA targets. *Bioinformatics*, 23(22):2987–92. 31, 86

Zeiner, G. M., Norman, K. L., Thomson, J. M., Hammond, S. M., and Boothroyd, J. C. (2010). Toxoplasma gondii infection specifically increases the levels of key host microrna s. *PLoS One*, 5(1):e8742. 46

Zhang, B., Pan, X., Cannon, C. H., Cobb, G. P., and Anderson, T. A. (2006a). Conservation and divergence of plant microRNA genes. *Plant J*, 46(2):243–59. 17, 18, 19, 20, 21, 23

Zhang, B., Pan, X., and Stellwag, E. J. (2008). Identification of soybean microRNAs and their targets. *Planta*, 229(1):161–82. 78

Zhang, B., Pan, X., Wang, Q., Cobb, G. P., and Anderson, T. A. (2006b). Computational identification of microRNAs and their targets. *Comput Biol Chem*, 30(6):395–407. 17, 19, 20, 21, 25, 85

Zhang, B. H., Pan, X. P., Cox, S. B., Cobb, G. P., and Anderson, T. A. (2006c). Evidence that miRNAs are different from other RNAs. *Cell Mol Life Sci*, 63(2):246–54. 17, 20, 22

Zhang, B. H., Pan, X. P., Wang, Q. L., Cobb, G. P., and Anderson, T. A. (2005). Identification and characterization of new plant microRNAs using EST analysis. *Cell Res*, 15(5):336–60. 18, 21

Zhang, J.-Z., Ai, X.-Y., Guo, W.-W., Peng, S.-A., Deng, X.-X., and Hu, C.-G. (2012a). Identification of mirnas and their target genes using deep sequencing and degradome analysis in trifoliate orange [poncirus trifoliate (l.) raf]. *Mol Biotechnol*, 51(1):44–57. 81

Zhang, W. L., Huitorel, P., Geneviere, A.-M., Chiri, S., and Ciapa, B. (2006d). Inactivation of MAPK in mature oocytes triggers progression into mitosis via a Ca2+-dependent pathway but without completion of S phase. *J Cell Sci*, 119(Pt 17):3491–501. 108

Zhang, X., Azhar, G., and Wei, J. Y. (2012b). The expression of microRNA and microRNA clusters in the aging heart. *PLoS One*, 7(4):e34688. 80

Zhang, Y., Xu, B., Yang, Y., Ban, R., Zhang, H., Jiang, X., Cooke, H. J., Xue, Y., and

Shi, Q. (2012c). Cpss: a computational platform for the analysis of small rna deep sequencing data. *Bioinformatics*, 28(14):1925–1927. 52

Zhang, Y., Zhou, X., Ge, X., Jiang, J., Li, M., Jia, S., Yang, X., Kan, Y., Miao, X., Zhao, G., Li, F., and Huang, Y. (2009). Insect-Specific microRNA involved in the development of the silkworm Bombyx mori. *PLoS One*, 4(3):e4677. 46, 87

Zhao, W., Liu, W., Tian, D., Tang, B., Wang, Y., Yu, C., Li, R., Ling, Y., Wu, J., Song, S., and Hu, S. (2011). waprna: a web-based application for the processing of rna sequences. *Bioinformatics*, 27(21):3076–7. 52

Zhao, Y., He, S., Liu, C., Ru, S., Zhao, H., Yang, Z., Yang, P., Yuan, X., Sun, S., Bu, D., Huang, J., Skogerbø, G., and Chen, R. (2008). Microrna regulation of messenger-like noncoding rnas: a network of mutual microrna control. *Trends Genet*, 24(7):323–7. 27

Zheng, P.-m., Wu, J.-y., Gu, J.-b., Tu, Z.-j., and Chen, X.-g. (2010). [isolation, identification and analysis of the expression profile of miRNAs in Aedes albopictus]. *Nan Fang Yi Ke Da Xue Xue Bao*, 30(4):677–80. 79

Zhou, L., Jiang, G., Chan, G., Santos, C. P., Severson, D. W., and Xiao, L. (2005). Michelob x is the missing inhibitor of apoptosis protein antagonist in mosquito genomes. *EMBO Rep*, 6(8):769–74. 101, 103

Zhu, E., Zhao, F., Xu, G., Hou, H., Zhou, L., Li, X., Sun, Z., and Wu, J. (2010). mirtools: microrna profiling and discovery based on high-throughput sequencing. *Nucleic Acids Res*, 38(Web Server issue):W392–7. 52

Zieler, H. and Dvorak, J. A. (2000). Invasion in vitro of mosquito midgut cells by

the malaria parasite proceeds by a conserved mechanism and results in death of the invaded midgut cells. *Proc Natl Acad Sci U S A*, 97(21):11516–21. 103

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*, 31(13):3406–15. 22

Appendix A

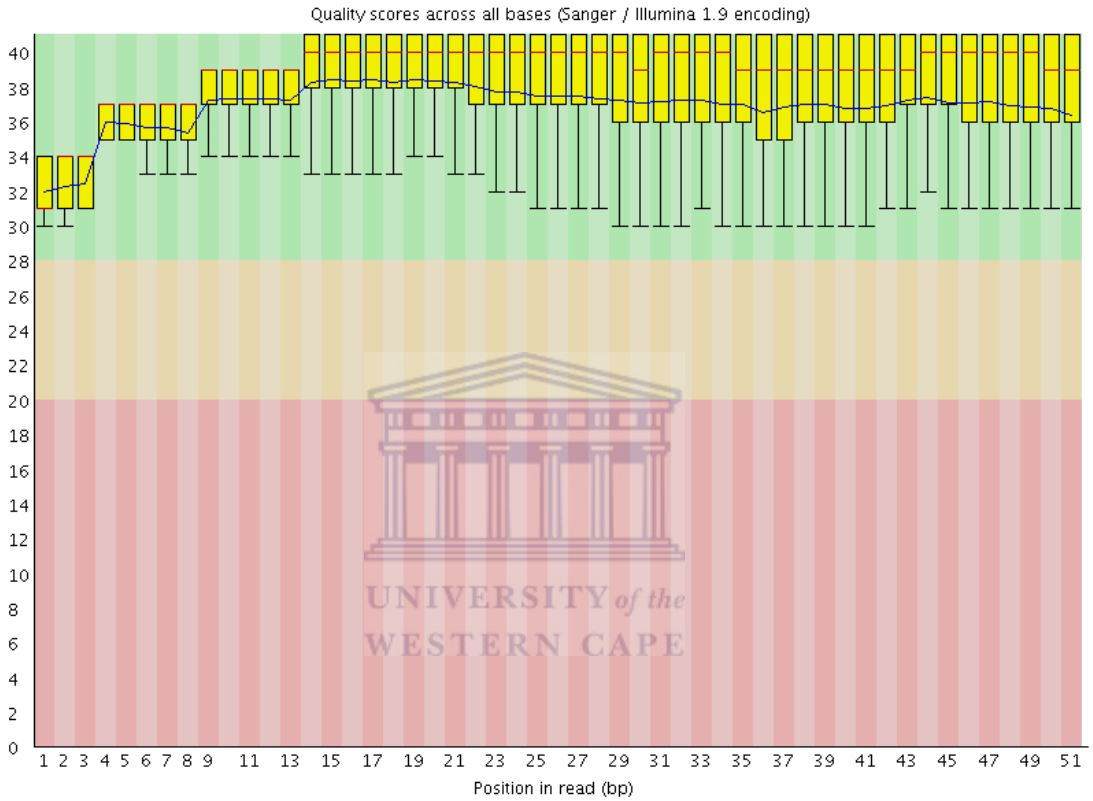


Figure A.1: An overview of the range of quality values across all bases at wach postion in the Fastq file for eggs library. For each position a BoxWhisker type plot is drawn. The elements of the plot are as follows; the central red line is the median value, the yellow box represents the inter-quartile range (25-75%), the upper and lower whiskers represent the 10% and 90% points, the blue line represents the mean quality. The y-axis on the graph shows the quality scores. The higher the score the better the base call. The background of the graph divides the y axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red). The quality of calls on most platforms will degrade as the run progresses, so it is common to see base calls falling into the orange area towards the end of a read.

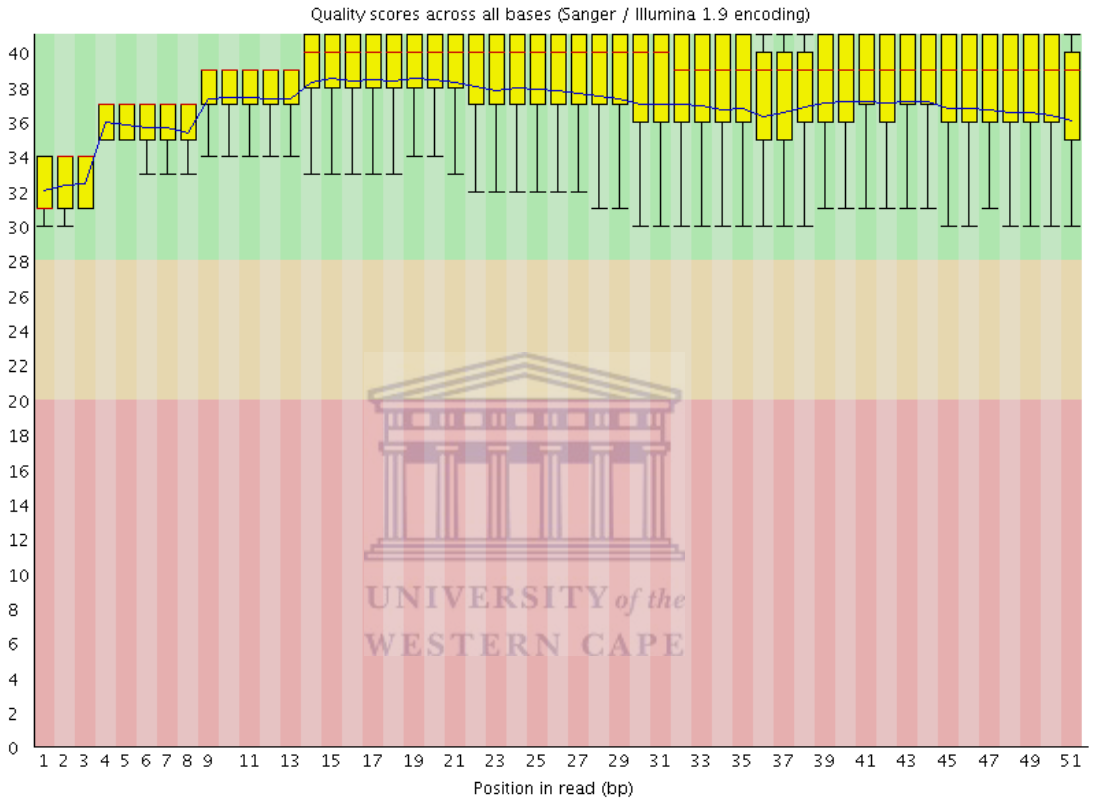


Figure A.2: An overview of the range of quality values across all bases at wach postion in the Fastq file for larvae library. For each position a BoxWhisker type plot is drawn. The elements of the plot are as follows; the central red line is the median value, the yellow box represents the inter-quartile range (25-75%), the upper and lower whiskers represent the 10% and 90% points, the blue line represents the mean quality. The y-axis on the graph shows the quality scores. The higher the score the better the base call. The background of the graph divides the y axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red). The quality of calls on most platforms will degrade as the run progresses, so it is common to see base calls falling into the orange area towards the end of a read.

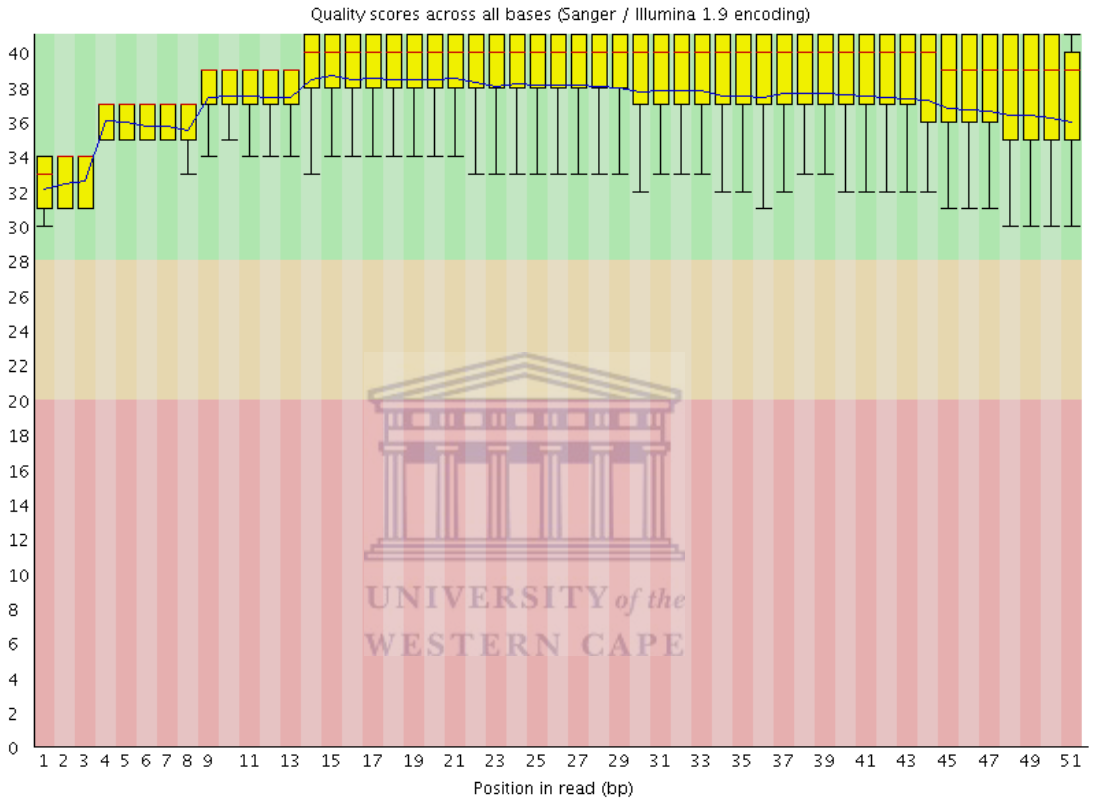


Figure A.3: An overview of the range of quality values across all bases at wach postion in the Fastq file for each pupae library. For each position a BoxWhisker type plot is drawn. The elements of the plot are as follows; the central red line is the median value, the yellow box represents the inter-quartile range (25-75%), the upper and lower whiskers represent the 10% and 90% points, the blue line represents the mean quality. The y-axis on the graph shows the quality scores. The higher the score the better the base call. The background of the graph divides the y axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red). The quality of calls on most platforms will degrade as the run progresses, so it is common to see base calls falling into the orange area towards the end of a read.

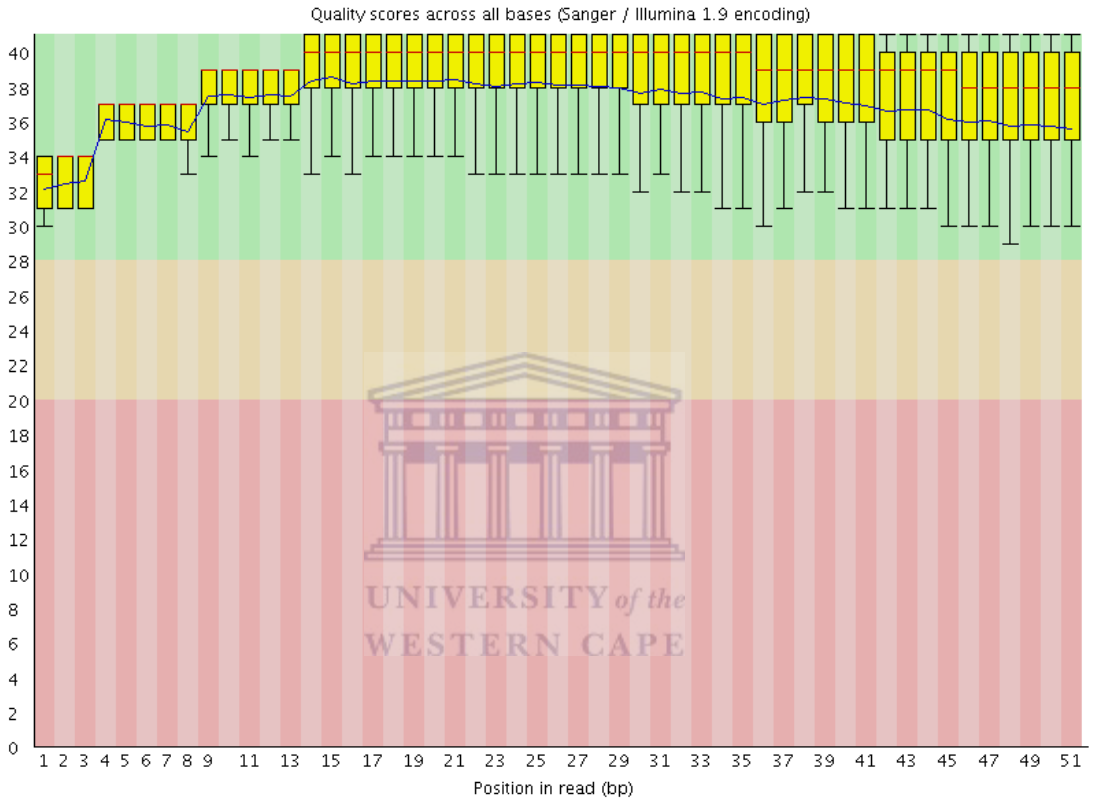


Figure A.4: An overview of the range of quality values across all bases at wach postion in the Fastq file for each adults library. For each position a BoxWhisker type plot is drawn. The elements of the plot are as follows; the central red line is the median value, the yellow box represents the inter-quartile range (25-75%), the upper and lower whiskers represent the 10% and 90% points, the blue line represents the mean quality. The y-axis on the graph shows the quality scores. The higher the score the better the base call. The background of the graph divides the y axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red). The quality of calls on most platforms will degrade as the run progresses, so it is common to see base calls falling into the orange area towards the end of a read.

Figure A.5: Fold-change of the known miRNAs during the development of *An. funestus s.s.* The fold-changes were calculated for each miRNA (x-axis) using normalized reads +1 (y-axis, bars).

Appendix B

. APPENDIX B

script:1 get_3'UTR_sequences.pl

```
#!/usr/bin/perl
#Connect to the local Ensembl Core database:
use Bio::EnsEMBL::DBSQL::DBAdaptor;
my $host='localhost';
my $user='root';
my $dbname='ag_63';
my $db=new Bio::EnsEMBL::DBSQL::DBAdaptor(
        -host=>$host,
        -user=>$user,
        -dbname=>$dbname
);
# obtain 3'UTR sequences
$slice_adaptor=$db->get_SliceAdaptor();
my @slices=@{$slice_adaptor->fetch_all('chromosome')};
        foreach $slice(@slices)
        {
        $slice->seq_region_name();
        my $genes=$slice->get_all_Genes();
                        while($gene=shift@{$genes})
                {
                my $gstring=feature2string($gene);
                my $transcripts=$gene->get_all_Transcripts();
                        while ($transcript=shift @{$transcripts})
                        {
                        my $tstring=feature2string($transcript);
                        my $thr_utr=$transcript->three_prime_utr();
                                if (defined $thr_utr==1)
                                {
                                print ">$gstring_",$thr_utr->seq(),"\n",$thr_utr->seq(),"\n";
                                }
                        }
                }
        }
sub feature2string
{
my $feature=shift;
my $stable_id=$feature->stable_id();
return sprintf($stable_id);
}
```

. APPENDIX B

script:2 miRanda_and_RNAhybrid_targets_prediction.pl

```
#!/usr/bin/perl
use DBI;
use DBD::mysql;
use Bio::EnsEMBL::DBSQL::DBAdaptor;
use Bio::EnsEMBL::Compara::DBSQL::DBAdaptor;
my $hostname="hostname";
my $port="0000";
my $username="xxx";
my $password='yyy';
my $mirnas_db="mirnas";
my $ag_db="ag_63";
my $ae_db="ae_63";
my $cq_db="cq_63";
my $dm_db="dm_63";
my $compara_db ="compara_58";
my $results_db ="insectar";
my $ag_core_connection=new Bio::EnsEMBL::DBSQL::DBAdaptor(
        -host=>$hostname,-user=>$username,
        -pass=> $password,-port=>$port,
        -dbname=>$ag_db
        );
my $ae_core_connection=new Bio::EnsEMBL::DBSQL::DBAdaptor(
        -host=>$hostname,-user=>$username,
        -pass=> $password,-port=>$port,
        -dbname=>$ae_db
        );
my $cq_core_connection=new Bio::EnsEMBL::DBSQL::DBAdaptor(
        -host=>$hostname,-user=>$username,
        -pass=> $password,-port=>$port,
        -dbname=>$cq_db
        );
my $dm_core_connection=new Bio::EnsEMBL::DBSQL::DBAdaptor(
        -host=>$hostname,-user=>$username,
        -pass=> $password,-port=>$port,
        -dbname=>$dm_db
        );
my $compara_connection=new Bio::EnsEMBL::Compara::DBSQL::DBAdaptor(
        -host=>$hostname,
        -user=>$username,
        -pass=>$password,
        -port=>$port,
        -dbname=>$compara_db
        );
my $mirnas_db_connection = "DBI:mysql:database=$mirnas_db;host=$hostname;port=$port";
my $dbh_mirna = DBI->connect($mirnas_db_connection, $username, $password)
my $results_db_connection = "DBI:mysql:database=$results_db;host=$hostname;port=$port";
my $dbh_results = DBI->connect($results_db_connection, $username, $password)
################################################################################
open (list,"mirnas.list")
```

. APPENDIX B

```
my @mirna_names=<list>;
close list;
        foreach my $mirna_ID(@mirna_names)
        {
        chomp $mirna_ID;
        my $mirna_sequence_file=mirna_sequence_file($mirna_ID);
        push @mysql,$mirna_ID;
        my $slice_adaptor = $ag_core_connection->get_SliceAdaptor();
        my @slices = @{$slice_adaptor->fetch_all('chromosome')};
                foreach $slice (@slices)
                {
                my $genes = $slice->get_all_Genes();
                        while ($gene = shift @{$genes})
                        {
                        my $transcripts = $gene->get_all_Transcripts();
                                foreach $transcript (@{$transcripts})
                                {
                                my $thr_utr = $transcript->three_prime_utr;
                                        if (defined $thr_utr==1)
                                        {
                                        my $gene_ID = stable_id($gene);
                                        my $gene_transcript_ID = stable_id($transcript);
                                        open (utr_file, ">utrs.fasta");
                                        print utr_file ">$gene_ID","\n",$thr_utr->seq(),"\n\n";
################################################################################
my $mirna_file = "$mirna_ID.fasta";
my $utr_file="utrs.fasta";
my $miranda_parameters="-sc 100 -en -14 -keyval -out miranda_file";
my $miranda = "/usr/local/bin/miranda $mirna_file $utr_file $miranda_parameters";
system($miranda);
open (miranda_file,"miranda_file");
my @miranda_file=<miranda_file>;
        foreach my $line(@miranda_file)
        {
        chomp $line;
                if ($line =~/\/\/hit_info/)
                {
                chomp $line;
                my @miranda_analysis=miranda_analysis($line);
                my @miranda_result;
                push @miranda_result,$mirna_ID,$gene_ID,$gene_transcript_ID,@miranda_analysis;
        my $sth=$dbh_results->prepare('insert into ag_miranda values(?,?,?,?,?,?,?,?,?,?,?,?,?,?,?)');
                $sth->execute(@miranda_result);
                }
        }
################################################################################
my @ortholog_info=ortholog($gene_ID);
my $mirna_file = "$mirna_ID.fasta";
my $orthologs_utrs_file="orthologs_utrs_file.fasta";
my $miranda_parameters="-sc 80 -en -14 -keyval -out miranda_ortho_file";
```

. APPENDIX B

```
100 my $miranda = "/usr/local/bin/miranda $mirna_file $orthologs_utrs_file $miranda_parameters";
101 system($miranda);
102 open (miranda_file, "miranda_ortho_file");
103 my @miranda_file=<miranda_file>;
104         foreach my $line(@miranda_file)
105         {
106         chomp $line;
107                 if ($line =~/\/\/hit_info/)
108                 {
109                 chomp $line;
110                 my @miranda_analysis=miranda_analysis($line);;
111                 my @miranda_result;
112                 push @miranda_result,$mirna_ID,@ortholog_info,$gene_ID,$gene_transcript_ID,@miranda_analysis;
113         my $sth = $dbh_results->prepare('insert into ag_miranda_ortho values (?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?)');
114                 $sth->execute(@miranda_result);
115                 }
116         }
117 ##############################################################################
118 my $rnahybrid = "/usr/local/bin/RNAhybrid -s 3utr_fly -q $mirna_file -t $utr_file -c > rnahybrid_file";
119 system($rnahybrid);
120 open (rnahybrid_file,"rnahybrid_file");
121 my @rnahybrid_file=<rnahybrid_file>;
122         foreach my $line(@rnahybrid_file)
123         {
124         chomp $line;
125                 if ($line =~/$gene_ID/)
126                 {
127                 chomp $line;
128                 my @rnahybrid_analysis=rnahybrid_analysis($line);
129                 my @rnahybrid_result;
130                 push @rnahybrid_result,$mirna_ID,$gene_ID,$gene_transcript_ID,@rnahybrid_analysis;
131         my $sth = $dbh_results->prepare('insert into ag_rnahybrid values (?,?,?,?,?,?,?,?,?,?,?)');
132                 $sth->execute(@rnahybrid_result);
133                 }
134         }
135 ##############################################################################
136 }
137 }
138 }
139 }
140 }
141 ##############################################################################
142 sub mirna_sequence_file
143 {
144 my $mirna_ID=shift;
145 my $sth = $dbh_mirna->prepare("select * from ag_mirnas where mirna_id = '$mirna_ID'");
146 my $ret = $sth->execute;
147         while ( my @row = $sth->fetchrow_array)
148         {
149          open (mirna_file, ">$mirna_ID.fasta");
```

. APPENDIX B

```
150         print mirna_file ">$row[0]\n$row[3]\n";
151         }
152 }
153 ################################################################################
154 sub stable_id
155 {
156 my $feature = shift;
157 my $stable_id  = $feature->stable_id();
158 return ($stable_id);
159 }
160 ################################################################################
161 sub miranda_analysis
162 {
163 my $line=shift;
164 my @analysis;
165 my @hit_line=split("\t",$line);
166 $hit_line[0]=~s/.*=//;
167 $hit_line[1]=~s/.*=//;
168 $hit_line[2]=~s/.*=//;
169 $hit_line[3]=~s/.*=//;
170 $hit_line[4]=~s/.*=//;
171 $hit_line[5]=~s/.*=//;
172 $hit_line[6]=~s/.*=//;
173 $hit_line[7]=~s/.*=//;
174 $hit_line[8]=~s/.*=//;
175 $hit_line[9]=~s/.*=//;
176 $hit_line[10]=~s/.*=//;
177 $hit_line[11]=~s/.*=//;
178 $hit_line[12]=~s/.*=//;
179 $hit_line[13]=~s/.*=//;
180 $hit_line[14]=~s/.*=//;
181 $hit_line[12]=~tr/a-z/A-Z/;
182 $hit_line[14]=~tr/a-z/A-Z/;
183 my @mirna_alignment=seed($hit_line[12]);
184 my @utr_alignment=seed($hit_line[14]);
185 my @map_alignment=seed($hit_line[13]);
186 my $a=join('',@map_alignment);
187 $a=~s/\s/s/g;
188 $a=~s/\|/m/g;
189 $a=~s/\:/s/g;
190         if (($a eq "mmmmmmms") and ($utr_alignment[7] eq "A"))
191         {
192         push @analysis,'8mer';
193         }
194         elsif ($a eq "mmmmmmms")
195         {
196         push @analysis,'7mer.M8';
197         }
198         elsif (($a eq "smmmmmms") and ($utr_alignment[7] eq "A"))
199         {
```

. APPENDIX B

```
200         push @analysis, '7mer.A1';
201         }
202         elsif (($a eq "smmmmmms") or ($a eq "mmmmmmss"))
203         {
204         push @analysis, '6mer';
205         }
206         else
207         {
208         push @analysis,'no_seed_mer';
209         }
210 push@analysis,
211         "$hit_line[3]",
212         "$hit_line[4]",
213         "$hit_line[5]",
214         "$hit_line[6]",
215         "$hit_line[7]",
216         "$hit_line[8]",
217         "$hit_line[9]",
218         "$hit_line[10]",
219         "$hit_line[11]",
220         "$hit_line[12]",
221         "$hit_line[13]",
222         "$hit_line[14]";
223 return (@analysis);
224 }
225 ##############################################################################################
226 sub seed
227 {
228 my $seed=shift;
229 my $end=length($seed);
230 my $start=$end-8;
231 my @array=split('',$seed);
232 my@seed;
233 my $i=0;
234 my $n;
235         for(@array)
236         {
237         my $n=$array[$start+$i];
238         push @seed, $n;
239         $i++;
240         }
241         return @seed;
242 }
243 ##############################################################################################
244 sub ortholog
245 {
246 open (orthologs_utrs_file, ">orthologs_utrs_file.fasta");
247 my $gene = shift;
248 my $member_adaptor= $compara_connection->get_adaptor("Member");
249 my $member= $member_adaptor->fetch_by_source_stable_id("ENSEMBLGENE","$gene");
```

. APPENDIX B

```
my $homology_adaptor=$compara_connection->get_adaptor("Homology");
        if (defined $member)
        {
        my @species=("Drosophila_melanogaster","Aedes_aegypti","Culex_quinquefasciatus");
                foreach $sp(@species)
                {
                my $homologies = $homology_adaptor->fetch_all_by_Member_paired_species($member,"$sp");
                        foreach my $homology (@$homologies)
                        {
                        my @paired_species;
                                foreach my $member_attribute (@{$homology->get_all_Member_Attribute})
                                {
                                my ($member, $attribute) = @{$member_attribute};
                                push @paired_species, $member;
                                }
                        my @paired_species = map ($_->stable_id, @paired_species);
                                foreach $ortholog(@paired_species)
                                {
                                chomp $ortholog;
                                        if ($ortholog=~/FB/)
                                        {
                                        my $gene_adaptor = $dm_core_connection->get_GeneAdaptor;
                                        my $gene = $gene_adaptor->fetch_by_stable_id("$ortholog");
                                        my $transcripts = $gene->get_all_Transcripts();
                                                while ($transcript = shift @{$transcripts})
                                                {
                                                my $thr_utr = $transcript->three_prime_utr;
                                                        if (defined $thr_utr==1)
                                                        {
                                        print orthologs_utrs_file ">$ortholog","\n",$thr_utr->seq(),"\n\n";
                                                        my $ortholog_gene_transcript_ID = stable_id($transcript);
                                                        return ($ortholog,$ortholog_gene_transcript_ID);
                                                        }
                                                }
                                        }
                                        if ($ortholog=~/CP/)
                                        {
                                        my $gene_adaptor = $cq_core_connection->get_GeneAdaptor;
                                        my $gene = $gene_adaptor->fetch_by_stable_id("$ortholog");
                                        my $transcripts = $gene->get_all_Transcripts();
                                                while ($transcript = shift @{$transcripts})
                                                {
                                                my $thr_utr = $transcript->three_prime_utr;
                                                        if (defined $thr_utr==1)
                                                        {
                                        print orthologs_utrs_file ">$ortholog","\n",$thr_utr->seq(),"\n\n";
                                                        my $ortholog_gene_transcript_ID = stable_id($transcript);
                                                        return ($ortholog,$ortholog_gene_transcript_ID);
                                                        }
                                                }
```

. APPENDIX B

```
                                    }
                                    if ($ortholog=~/AA/)
                                    {
                                    my $gene_adaptor = $ae_core_connection->get_GeneAdaptor;
                                    my $gene = $gene_adaptor->fetch_by_stable_id("$ortholog");
                                    my $transcripts = $gene->get_all_Transcripts();
                                            while ($transcript = shift @{$transcripts})
                                            {
                                            my $thr_utr = $transcript->three_prime_utr;
                                                    if (defined $thr_utr==1)
                                                    {
                                    print orthologs_utrs_file ">$ortholog","\n",$thr_utr->seq(),"\n\n";
                                                    my $ortholog_gene_transcript_ID = stable_id($transcript);
                                                    return ($ortholog,$ortholog_gene_transcript_ID);
                                                    }
                                            }
                                    }
                            }
                    }
            }
    }
}
##############################################################################################
sub rnahybrid_analysis
{
my $line=shift;
my @analysis;
my @hit_line=split(":",$line);
$hit_line[0]=~s/.*=//;
$hit_line[1]=~s/.*=//;
$hit_line[2]=~s/.*=//;
$hit_line[3]=~s/.*=//;
$hit_line[4]=~s/.*=//;
$hit_line[5]=~s/.*=//;
$hit_line[6]=~s/.*=//;
$hit_line[7]=~s/.*=//;
$hit_line[8]=~s/.*=//;
$hit_line[9]=~s/.*=//;
$hit_line[10]=~s/.*=//;
$hit_line[7]=~ tr/a-z/A-Z/;
$hit_line[7]=~s/\s/-/g;
$hit_line[8]=~ tr/a-z/A-Z/;
$hit_line[8]=~s/\s/-/g;
$hit_line[9]=~ tr/a-z/A-Z/;
$hit_line[9]=~s/\s/-/g;
$hit_line[10]=~ tr/a-z/A-Z/;
$hit_line[10]=~s/\s/-/g;
push @analysis,
"$hit_line[1]",
"$hit_line[3]",
```

```
350 "$hit_line[4]",
351 "$hit_line[5]",
352 "$hit_line[6]",
353 "$hit_line[7]",
354 "$hit_line[8]",
355 "$hit_line[9]",
356 "$hit_line[10]";
357 return (@analysis);
358 }
359 ################################################################################
```

script:3 MicroTar_targets_prediction.sh

```
1 #!/bin/sh
2 microtar -t ag_utrs.fa -q ag_mirnas.fa -f ag_microtar_results.tsv
```