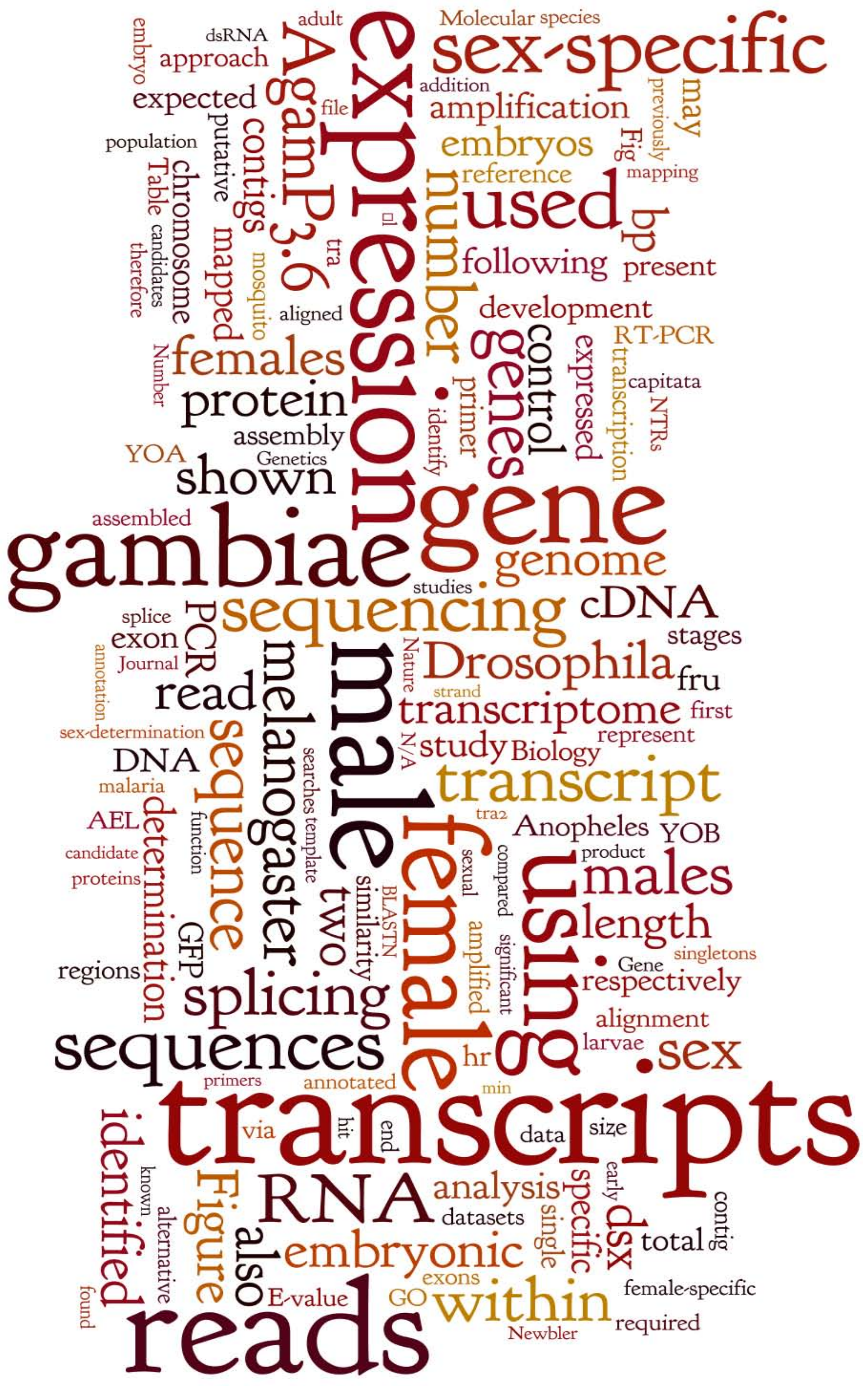


transcripts
gambiae
gene
male
expression
reads
female
using
sex-specific
AgamP3.6
used
number
genes
females
protein
shown
sequencing
cDNA
Drosophila
melanogaster
transcriptome
transcript
sequence
sequences
splicing
RNA
embryonic
within
identified
Figure
also
males
length
sex
read
determination
two
embryos
amplification
contigs
chromosome
mapped
following
development
genome
control
present
bp
may
expected
approach
putative
population
Table
candidates
therefore
mosquito
aligned
reference
mapping
previously
Fig
addition
file
adult
dsRNA
embryo
Molecular species
RT-PCR
expressed
transcription
primer
identify
capitata
NTRs
assembly
YOA
Genetics
assembled
studies
splice
exon
PCR
Journal
annotation
Nature
strand
fru
first
represent
sex-determination
study
Biology
N/A
DNA
malaria
searches
template
function
AEL
tra2
Anopheles
YOB
product
candidate
proteins
sexual
compared
BLASTN
similarity
GFP
amplified
significant
Gene
singletons
regions
respectively
alignment
larvae
hr
primers
annotated
min
via
hit
end
data
size
known
early
contig
analysis
datasets
single
specific
dsx
total
alternative
found
exons
GO
E-value
female-specific
Newbler
required
tra

Sex-specific expression during embryonic development of *Anopheles gambiae*

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy

by Nathan Joseph Dennison

April 2012

Decleration

This thesis is the result of my own work. The work contained within this thesis has not been presented, nor is currently being presented, either wholly or in part for any other degree or qualification.

The research work was carried out at the Liverpool School of Tropical Medicine (University of Liverpool). DNA Sanger sequencing was carried out at Biological Sciences with the assistance of Derek Daley (University of Liverpool). 454 Pyrosequencing was performed at the Centre for Genomic Research (CGR), University of Liverpool by Margaret Hughes. Custom Perl script length.pl was kindly provided by Anton Tsyganov-Bodounov.

ABSTRACT

Currently employed vector control strategies are experiencing resistance that threaten to limit their effectiveness in controlling malaria transmission. Therefore, development of new genetic vector control strategies may benefit efforts to curb the devastating effects of malaria. Sex-specific gene expression has not been studied in mosquito embryos thus far, but its understanding may lead to creation of new tools for vector control.

Here we analysed sex-specific transcription using firstly a targeted gene approach and secondly *de novo* transcriptional profiling of male and female embryonic transcriptomes. Two genes, *doublesex (dsx)* and *fruitless (fru)*, belonging to sex determination/differentiation pathway, have sex-specific transcripts described previously in adult *An. gambiae*, and here, we present evidence that both *dsx* and *fru* are also sex-specifically spliced in the embryonic stages. *dsx*, a final gene in the sex determination pathway and a key regulator of sexual differentiation, is maternally deposited as a female transcript, but establishes a persistent pattern of alternative sex-specific splicing eight hours after egg laying (AEL). *fru*, a modulator of adult male behaviour, has sex-specific transcription starting at least 16 h AEL.

The *An. gambiae* sex determination pathway was further investigated through isolation of a putative homologue of the *D. melanogaster* sex determination gene *transformer2* (*tra2*), which was knocked down via RNAi in an attempt to confirm its involvement in *Anopheles* sexual development. Using a transient reduction in transcript levels, an effect on sex determination phenotype was not observed. In addition, bioinformatics and degenerate PCR approaches identified a large number of candidates with sequence similarity to *transformer* (*tra*), though these candidates require additional characterisation.

Previous work on sex-specific expression has taken an *a priori* approach, using microarrays for gene expression analysis. Here, a *de novo* approach was undertaken and an approximate ~500,000 transcript fragments have been independently sampled using the 454 platform from each transcriptome of male and female embryos and assembled into 17,492 and 16,899 contigs from male and female reads, respectively. *In silico* subtraction of the two RNA-seq databases revealed a large number of putative male or female specific transcripts, of which 60 were tested using RT-PCR. Among those, two transcripts, displaying male-specific expression, represented two novel genes (YOA and YOB) located on the Y chromosome and sharing partially overlapping exons. YOA is expressed from late larvae through to adulthood. YOB is expressed from 4 hr AEL and continues across all stages through to adulthood. Their amino acid translation has no significant similarity to known proteins, or conserved domains. Furthermore, a functional role could not be assigned since gene silencing approaches did not reveal an obvious phenotype.

This study has also emphasised that the current AgamP3.6 gene build is not complete. Deep sequencing of the embryonic transcriptome has produced a large dataset containing 6,436 transcripts that map to previously unannotated genomic regions. These sequences likely represent extensions to untranslated regions, novel exons or unique gene sequences. Our study demonstrates that RNA-seq is a powerful tool to identify sex-specific expression in mosquito embryos and that sex-specific transcription starts very early during embryonic development. The remaining question is whether the data produced within this study can be exploited in the context of a genetic vector control strategy.

Acknowledgements

The writing of this thesis has been one of the most challenging aspects of my life so far and it represents three and a half years of solid toil, which would have not been possible without a great number of people. Firstly, my great thanks to Jarek and Gareth who agreed to take me on as a PhD student and for their continued support throughout the process. I have learned a great deal from them both that I hope will stand me in good stead for my future work.

Without my parents and siblings support, completing this PhD would not have been possible. They have continually encouraged me to follow whatever path is best for me and I thank them for all their help and support through the years of all my education. Without question, I owe a great deal to Urvashi, who has supported me constantly throughout the three years on my Despite us both pursuing PhDs at the same time and encountering the associated stresses simultaneously, she has always been there to get me through and I will be forever grateful.

A large number of the Vector Group has been instrumental to this thesis. Anton, Marco, Mauritio, Amy, and Marion have all helped me, be it with mosquitoes, in the lab or with bioinformatics. Thanks also to Ben who kindly lent me the FU III plasmid. Thanks also go to Derek Daley, Margaret Hughes and Richard Gregory at Biological Sciences who have provided great support for bioinformatics and sequencing aspects of this project. Also, anyone who I have played 5-aside football with from the school I am grateful to, for escaping the PhD for an hour a week to attempt to play football.

List of Figures

List of Tables

Abbreviations

aa	amino acid
AEL	after egg laying
APS	adenosine 5' phosphosulfate
ATP	adenosine triphosphate
BED	Browser Extensible Data
BLAST	basic local alignment search tool
bp	base pair
cDNA	complementary deoxyribonucleic acid
CGR	Centre for Genomic Research
ddNTPs	dideoxynucleoside triphosphate
DNA	deoxyribonucleic acid
dNTPs	deoxynucleoside triphosphate
dsRNA	double-stranded ibonucleic acid
dsx	doublesex
edgeR	empirical analysis of digital gene expression in R
EST	expressed sequence tag
fru	fruitless
GFP	green flourescent protein
$GMAP_1$	Global Malaria Action Plan
$GMAP_2$	A Genomic Mapping and Alignment Program for mRNA and EST Sequences
GTF	gene transfer format
IGV	Integrative Genomics Viewer
IPTG	isopropyl β-D-1 thiogalactopyranoside
LB	Luria-Bertani
LSTM	Liverpool School of Tropical Medicine
Medea	Maternal-effect-dominant embryonic arrest
MgCl2	Magnesium chloride
MIRA	Mimicking Intelligent Read Assembly
MOL	muscle of lawrence
mRNA	messenger ribonucleic acid
MZT	maternal to zygotic transition
NCBI	National Centre for Biotechnology Information
ncRNA	non-coding RNA
NGS	next generation sequencing
NTR	new transcribe region
ORF	open-reading frame
PCR	polymerase chain reaction
RACE	rapid extension of cDNA ends
RBM	Roll Back Malaria
RIDL	Release of Insects Carrying a Dominant Lethal
RNA	ribonucleic acid
RNAi	ribonucleic acid interference
RPKM	reads per kilobase of exon model per million mapped reads
RRM	ribonucleic acid recognition motif
RT-PCR	reverse-transcription polymerase chain reaction
rtTA	reverse tetracycline transactivator
SAGE	seriel analysis of gene expression
SAM	sequence alignment/map format

SIT	Sterile insect technique
SMARTer	Switching Mechanism at 5’ End of RNA Template
SMRT	Single Molecule Real Time
SOLiD	Sequencing by Oligonucleotide Ligation and Detection
sxl	sex-lethal
tetO	tetracycline responsive element
tetR	tetracycline repression
tra	transformer
tra2	transformer2
tTA	tetracycline transactivator
UTR	untranslated region
WHO	World Health Organisation
XGAL	bromo-chloro-indolyl-galactopyranoside
XSE	x-linked signal element

Chapter 1: Introduction

1.1 Malaria Prevalence

Over a century of research into malaria vectors began with the identification of 'brown mosquitoes' showing ability to harbour parasites from the blood of an infected individual (Ross 1897). Today, searching for the keyword 'malaria' identifies over 60,000 peer reviewed articles, a fact that highlights the extensive efforts in progress to curtail its devastating effects. Similarly, commitment to control and eradicate the disease is seen in funding increasing from less than $500 million in 2005 to $1.8 billion in 2010 and $2 billion in 2011, though this is expected to decrease from 2015 onwards (WHO 2011). Currently over 3.3 billion people are at risk of malaria worldwide and an estimated 216 million cases were reported in 2010 of which 80 % occurred in sub-Saharan Africa, with transmission also present in Asia, Latin America and the Middle East. Worldwide an estimated 650,000 malaria deaths were reported in 2010, a 13 % reduction since 2000 (WHO 2011). A recent report states that malaria mortality was actually over 1.2 million in 2010, though it also concludes that this has reduced by 32 % since a peak in 2004 (Murray *et al.* 2012). The targets of the 2008 Global Malaria Action Plan (GMAP) by the Roll Back Malaria (RBM) campaign were revised in 2011 to reduce malaria incidence by 75 % and eliminate mortality by 2015. Though prevalence of interventions such as insecticide treated nets (ITNs) have increased from 3 to 50 % in the last 10 years, the GMAP targets are far from being achieved (RBM 2008; RBM 2011). To compound the issue, there is a strong link between malaria burden and poverty, with causality suggested to be occurring in both directions (Sachs and Malaney 2002). The RMB report suggests that targets will not be achieved via 'business as usual', highlighting the need for additional control strategies to complement those currently utilised.

1.2 *An. gambiae* and disease transmission

One of the main vectors of malaria, females of the *Anopheles gambiae* (Diptera: Culicidae) complex display anthropophilic behaviour for transmission of *Plasmodium* parasites including *P. falciparum, P. vivax, P. ovale* and *P. malariae* (WHO 2011). Female *Anopheles* mosquitoes can lay up to 200 eggs following mating and blood meal ingestion. Females likely will only mate once and following coitus store sperm within the spermathecae, this will provide the paternal haploid genome to all eggs laid during the female's lifespan.

Hatched larvae undergo four stages of molting; stages between molts are termed instars. Fourth instar larvae then undergo metamorphosis into pupae followed by adult emergence. Adults have the potential to mate within four days post emergence and females seek a blood meal to facilitate egg development soon after (Clements 1992). Following ingestion of a *Plasmodium* infected blood meal; the female mosquito is capable of an infectious bite to another human host within 10 - 15 days. It is therefore important for control strategies to target the mosquito before this stage.

Plasmodium infection of a human host begins with injection of sporozoites along with saliva by an infected mosquito during biting. Sporozoites are transported to the liver where they invade hepatocytes and undergo asexual replication resulting in the formation of merozoites that are released into the bloodstream. Merozoites invade erythrocytes (red blood cells) and enlarge, developing into trophozoites following consumption of the cells cytoplasm and breakdown of haemoglobin into consumable amino acids. Trophozoites undergo cell division without cytokinesis, forming schizonts that then release merozoites following rupture of the erythrocyte, allowing further infection of additional erythrocytes. Lysis of the erythrocytes is responsible for the pathology associated with malaria infection. Most merozoites will continually invade new erythrocytes, but a fraction will develop into sexually reproductive forms within erythrocytes, known as micro or macro gametocytes. Following ingestion of blood containing gametocytes by a mosquito gametogenesis occurs, after which flagellated microgametes fertilize the macrogamete, leading to formation of a zygote in the midgut lumen. A motile ookinete develops from the zygote which is able to cross the midgut epithelium and develop into an oocyst within the midgut wall. The oocyst undergoes asexual reproduction forming sporozoites and rupture of the oocyst releases the sporozoites into the mosquito hemocoel. Migration of sporozoites and their invasion of the mosquito salivary glands completes the life cycle, making the mosquito able to infect another host (Miller and Hoffman 1998).

Anopheline mosquitoes are also able to transmit neglected tropical diseases including the alphavirus O'nyong-nyong virus (ONNV) and have caused large scale outbreaks of viral disease in several African countries (Williams *et al.* 1965; Lanciotti *et al.* 1998; Powers *et al.* 2000). Alphaviruses, part of the Arboviruses group (Arthropod Borne Viruses), have enveloped positive strand RNA genomes and are able to replicate within both arthropod hosts and invertebrate hosts (Strauss and Strauss 1994). Previously, ONNV infected

anopheline mosquitoes were responsible for an epidemic across East Africa, affecting over 2 million people (Williams *et al.* 1962). The virus re-emerged in Uganda after 35 years and it causes symptoms including fever, skin rash and inflammation of lymph nodes (Rwaguma *et al.* 1997). Filarial worms are also transmitted by anopheline mosquitoes, specifically *Wuchereria bancrofti* that cause over 90 % of the 120 million global infections of lymphatic filariasis in humans (WHO 1997). Lymphatic filariasis, commonly known as elephantiasis, occurs when a blood feeding mosquito introduces filarial worms onto the skin of a human, followed by penetration of the worm into the bite wound. Migration to the lymph nodes causes severe inflammation and can result in lymphoedema or elephantiasis in chronic cases. Mosquitoes become infected with *Wuchereria bancrofti* microfilariae following biting of an infected host, and the microfilariae develop through successive stages prior to migration to the proboscis. Microfilariae develop into adult worms in the human host where they can produce several million microfilariae throughout their lifespan. *Wuchereria bancrofti* infected *An. gambiae* mosquitoes are prevalent in both Nigeria and Ghana (Boakye *et al.* 2004; Anosike *et al.* 2005; Lenhart *et al.* 2007).

1.3 Vector control and reducing disease transmission

1.3.1 Insecticides

The use of insecticides has been widespread for the control of mosquito populations for over half a century (Fay *et al.* 1947). Insecticides are deployed using different methods including indoor residual spraying (IRS), insecticide treated nets (ITNs), long lasting insecticide treated nets (LLINs) microbial larvicides and space spraying/fogging, of which the first two are currently the WHO recommended method of control (WHO 2011). Integrated Vector Management (IVM), targeting both adult and immature larval stages has been proposed as an effective control strategy (Townson *et al.* 2005). Many of the currently deployed insecticides target the mosquito voltage gated sodium channel to induce mortality (Davies *et al.* 2007). Delayed closure of the sodium channel through insecticide binding causes repeated firing of the nervous system, leading to death (Corbett 1974). Four classes of insecticides are currently used for vector control: pyrethroids, organochlorines, organophosphates and carbamates (WHO 2006). Vector control strategies have proven to be successful in reducing malaria morbidity in several countries including Eritrea, Zambia and Kenya (ter Kuile *et al.* 2003; Mufunda *et al.* 2007; Chanda *et al.* 2008).

For example, deployment of permethrin treated ITNs in the Bondo and Siaya districts of western Kenya reduced prevalence of clinical malaria by 44 % (ter Kuile *et al.* 2003).

No new classes of insecticides have been deployed in malaria endemic regions for the last 30 years and most control efforts currently rely upon pyrethroids, they are the only class of insecticides currently recommended for ITNs (Zaim *et al.* 2000). This dependency on a single class of insecticides has increased the presence of resistance in malaria endemic countries (Santolamazza *et al.* 2008). Currently 78 countries are reported to be monitoring insecticide resistance and pyrethroid resistance has been documented in 27 sub-Saharan African countries (WHO 2011). Resistance to insecticides is well documented and is caused by both target site mutations in the sodium channel reducing sensitivity to the insecticide (Martinez-Torres *et al.* 1998; Ranson *et al.* 2000) and metabolic resistance caused by insecticide detoxification through elevated enzyme levels (Djouaka *et al.* 2008; Stevenson *et al.* 2011). Additional target site mutations, conferring elevated levels of resistance, have been identified within the *An. gambiae* neurotransmitter gamma-aminobutyric acid receptor (Du *et al.* 2005) and the enzyme acetylcholinesterase (Weill *et al.* 2004). Members of the P450 family of enzymes are able to metabolise insecticides in *An. gambiae* and elevated P450 levels have been associated with increased resistance in many studies (Vontas *et al.* 2005; Djouaka *et al.* 2008; Muller *et al.* 2008). Similarly, elevated levels of glutathione S-transferases (GSTs) have been strongly linked to resistance to currently used insecticides, including pyrethroids (David *et al.* 2005) and DDT (Ding *et al.* 2003).

In addition to ITNs and IRS, space spraying, or fogging, of insecticide can be utilised for control of insect vector populations. This process involves the heating of the insecticide, commonly a synthetic pyrethroid, to form droplets that are then sprayed to coincide with mosquito peak activity (WHO 2003; Raghavendra *et al.* 2011). Emerging resistance to commonly used insecticides has seen an increase in field trials of the microbial larvicide *Bacillus thuringiensis* var. *israelensis* for control of immature life stages (Fillinger and Lindsay 2006). A field trial recently demonstrated that integrated control using ITNs and microbial larvicides reduced malaria incidence two fold compared to ITNs alone (Fillinger *et al.* 2009). Recent studies have also suggested that adult mosquitoes could be utilised for dissemination of the larvicides to larvae breeding sites (Devine *et al.* 2009; Devine and Killeen 2010). What is evident is the need for additional control measures to reduce the current burden placed upon insecticides for vector control. Limiting exposure of wild

mosquitoes to insecticides through use of additional complementary control strategies will reduce the risk of resistance emergence.

1.3.2 Sterile Insect Technique

The concept of using sterile males to control insect populations was first postulated over 50 years ago (Bushland *et al.* 1955; Knipling 1955). It relies upon the mass release of chemically or irradiated sterile males into a target population. It is expected that sterile males will mate with wild females, resulting in non-viable offspring and therefore population reduction. The sterile insect technique has been applied to screwworm flies, *Cochliomyia hominivorax*, to eliminate screwworm populations in North and Central America (Krafsur *et al.* 1987). Up to 150 million screwworm flies, irradiated at the pupal stage, were released within a 158,000 km^2 'border zone' at the Mexico-USA border weekly, resulting in eradication of screwworm flies in USA and a substantial decrease in Mexico. SIT was also successfully utilised in Libya, to eradicate screwworm flies following their accidental introduction in 1988. (Vargas-Teran *et al.* 1994). It is interesting to note that the 1300 million irradiate flies used in the Libyan control strategy were delivered from the same rearing facility utilised in the North American eradication. Following a two year release program, cases of screwworm were completely absent, demonstrating SIT as a valuable control method, even when the jet-lag of millions of insects following transatlantic flights is taken into account (Vargas-Teran *et al.* 1994). Release of 8.5 million gamma sterilised male tsetse flies (*Glossina austeni*) across a period of three years resulted in a population crash of the indigenous Zanzibar fly population (Vreysen *et al.* 2000). SIT has also been used within an exclusion program in California, USA, to prevent establishment of Mediterranean fruitfly (*Ceratitis capitata*) populations (Dowell *et al.* 2000). Across a 4,000 km^2 area of Los Angeles, 96,500 sterile flies per week per km^2 were released and this reduced the capture of wild type *C. capitata* by 94 %, following two years of controlled releases.

With respect to mosquito SIT approaches, field trials have been undertaken to target *An. albimanus*, with over two million irradiated males released in El Salvador, which lead to a reduction of target population density up to 97 % (Lofgren *et al.* 1974). The release of sterilised males over a five month period also prevented the seasonal rapid rise of *An. albimanus* populations seen prior to the release. Chemically sterilised *Culex pipiens*

quinquefasciatus were released on the Seahorse Key island (Florida, USA), resulting in a 96 % reduction in the number of viable egg rafts collected (Patterson *et al.* 1970). A Kenyan *Ae. aegypti* population has also been targeted using males containing sex-linked translocations through irradiation treatment that results in sterility. Daily release of sterile males over 70 days resulted in a reduction in the number of wild mosquitoes found through landing-traps and oviposition monitoring (McDonald *et al.* 1977). Males released in the approach by McDonald *et al.* (1977) were shown to have reduced competitiveness and survival rates compared to wild males. Inducing sterility and laboratory colonisation has been shown to induce a fitness cost compared to wild type males, though the extent of negative fitness effects does vary within the literature (Sharma *et al.* 1978; Andreasen and Curtis 2005; Munhenga *et al.* 2011). For example, laboratory colonised *An. arabiensis* have been shown to have no negative fitness costs associated with mating competitiveness compared to a field population (Munhenga *et al.* 2011). Whereas other studies have demonstrated that irradiation of mosquitoes has a negative impact upon fertility and will therefore limit the effectiveness of the SIT program (Sharma *et al.* 1978).

Mosquito SIT approaches require the manual separation of males and females and studies report that manual separation was not 100 % effective, causing undesirable release of biting females (Lofgren *et al.* 1974). Use of genetic sexing strains, with translocation of an autosomal insecticide resistance gene to the male-specific Y chromosome, also resulted in the release of a small fraction of females (Seawright *et al.* 1978). Despite the apparent limitations in fitness of released males and inappropriate release of females, SIT has been shown to be effective in controlling mosquito populations.

More recently, supported by the International Atomic Energy Agency, an SIT program is being tested in field-cage trials in Sudan, with the aim of eventually targeting isolated *An. arabiensis* populations that may prevent wild population migration into the area (Helinski *et al.* 2008). Targeting of isolated populations has been shown previously to lead to successful eradication of a vector population (Patterson *et al.* 1970), whereas lack of geographical barriers have been shown to result in population recovery to normal densities following the culmination of the mosquito releases (Lofgren *et al.* 1974). Therefore prior study of vector distribution is essential to the success of SIT programs for both elimination and population control.

1.3.3 Post-genomic era

Efforts to develop novel control strategies were greatly enhanced when malaria research entered the post-genomic era, following genome sequencing of a vector, host and parasite genome. Firstly the human genome in 2001, shortly followed by those of *An. gambiae* and *P. falciparum* a year later (Venter *et al.* 2001; Gardner *et al.* 2002; Holt *et al.* 2002). The genomes of a PEST strain as well as the M and S forms of *An. gambiae* have been sequenced. The PEST strain was intended to provide a reference genome for genomic studies but unfortunately is a hybrid of a laboratory colonised strain from Nigeria (M form) and the offspring from Kenyan field caught individuals (S form) (Holt *et al.* 2002). As the PEST strain is a hybrid, this may have implications when characterising gene expression from the differene molecular forms. Genome sequencing of the individual M and S forms was performed recently, using laboratory colonised populations from Mali (Lawniczak *et al.* 2010). With regards to controlling transmission of malaria, the *An. gambiae* genome sequence has been utilised for development of transgenic *An. gambiae* refractory to parasite development. Through identification of immune genes and tissue-specific promoters, expression of anti-parasitic effectors has been driven in tissues the *Plasmodium* parasite develops within (Dong *et al.* 2011; Meredith *et al.* 2011).

1.3.4 Transgenic *An. gambiae* refractory to parasite development

Plasmodium ookinete invasion of the mosquito midgut represents a bottleneck in the parasite life cycle that can be targeted to reduce disease transmission (Dinglasan and Jacobs-Lorena 2008). Genetic manipulation of *An. gambiae* has been possible for over a decade (Grossman *et al.* 2001) and transgenic *An. gambiae* refractory to parasite development through midgut specific expression of anti-parasitic molecules that targets this bottleneck have been created (Grossman *et al.* 2001; Dong *et al.* 2011; Meredith *et al.* 2011). Subsequent release of refractory vectors into wild populations would be expected to reduce vectorial capacity and reduce disease transmission. To ensure the success of such an approach, the refractory phenotype must be driven into the population, as the refractory-gene frequency would not be expected to increase in frequency through Mendelian inheritance alone (James *et al.* 2006). Through the use of gene-drive mechanisms, that show non-Mendelian inheritance, a refractory gene could increase in frequency in wild populations (Sinkins and Gould 2006). A number of alternative gene-drive

mechanisms could be utilised, including, transposable elements, homing endonuclease genes and meiotic drive mechanisms, such as a synthetic *Medea*-like element, as reviewed by Sinkins et al. (2006). *Medea* (maternal-effect-dominant embryonic arrest) elements were first identified in *Tribolium castaneum* and are thought to comprise of two tightly linked genes, one encoding a maternal-toxin and the other, a maternal anti-dote that lead to lethality of non-*Medea* element bearing progeny (Beeman *et al.* 1992; Beeman and Friesen 1999). A synthetic *Medea* element has been created in transgenic *Drosophila melanogaster* and tight-linkage of a refractory gene to the element has been hypothesised to drive a parasite-refractory phenotype into a population (Chen *et al.* 2007).

Transgenic-refractory approaches are greatly facilitated by the isolation of tissue specific promoters. To date, tissue-specific expression *in vivo* has been documented in neurons, testes and imaginal tissues (Lycett *et al.* 2011), salivary gland (Lombardo *et al.* 2009), midgut (Meredith *et al.* 2011; Nolan *et al.* 2011) and germline tissues (Papathanos *et al.* 2009) of *An. gambiae*. Isolation of blood meal inducible genes and their promoters have also improved transgenic refractory approaches. In particular, promoters from the *carboxypeptidase* and *vitellogenin* genes have been isolated that provide upregulated expression following a bloodmeal (Edwards *et al.* 1997; Nirmala *et al.* 2005; Chen *et al.* 2007; Meredith *et al.* 2011). Temporal co-ordination of parasite arrival and expression of anti-parasitic molecules would increase the effectiveness of a refractory strategy.

Through sequence comparison, homologues of *D. melanogaster* immunity genes have been identified in *Anopheles* and subsequently been shown to affect *Plasmodium* parasite development (Christophides *et al.* 2002; Meister *et al.* 2005). The *An. gambiae* transcription factor REL2 activates transcription of downstream anti-microbial peptides and regulates prevalence of *Plasmodium* infections (Meister *et al.* 2005). Heterologous expression of *An. gambiae* REL2 from the *vitellogenin* and *carboxypepditase* promoters, in *An. stephensi* transgenic mosquitoes reduced *Plasmodium* prevalence (Dong *et al.* 2011). Creation of refractory transgenic mosquitoes requires transposon-mediated transgenesis and this can prove problematic due to random insertion into the genome, potentially into transcriptionally quiet heterochromatin regions (Lloyd *et al.* 2003; Condon *et al.* 2007). However, recently it has been demonstrated that site specific integration of transgenes can be performed in *An. gambiae*, a technique that overcomes potential positional effects following random transposon integration in the genome (Meredith *et al.* 2011).

1.3.5 Release of Insects Carrying a Dominant Lethal (RIDL)

An alternative transgenic system that drives population suppression rather than a parasite-refractory phenotype is the Release of Insects carrying a Dominant Lethal (RIDL) strategy. RIDL is a transgenic-control strategy for the targeted elimination of insect vectors (Alphey and Andreasen 2002; Coleman and Alphey 2004). Mass release of transgenic males engineered to carry a conditional female-specific lethal gene into a wild population enables the production of a male only population. Repression of lethality in laboratory conditions allows rearing of a continuous population. Following removal of the repressor, release of transgenic insects drives the lethal gene into the target population through successful mating with wild insects. In the absence of the repressor, female offspring from transgenic male-wild type female matings die, subsequently reducing vector population size.

For mosquito control, male-only releases are advantageous primarily due to female biting. In addition, males within a bisexual release have been shown to mate with released females, thereby acting as a distraction and reducing the incidence of sterile male wild type female matings (Rendon *et al.* 2004). Previously, morphological approaches to separate males for release has been shown to be less than 100 % accurate, resulting in release of biting females (Lofgren *et al.* 1974). To ensure that exclusively males are released, genetic sexing of individuals is an attractive option compared to the more laborious morphological approach.

An. gambiae genetic sexing strains have previously been developed using translocation of the dieldrin resistance gene to the male Y chromosomes, allowing separation of 1st instar larvae (Curtis *et al.* 1976). Treatment of 1st instar larvae with dieldrin resulted in female-specific mortality, allowing selection of males only (Curtis *et al.* 1976). Also, through use of the β-tubulin promoter driving the fluorescent reporter gene EGFP in a transgenic *An. gambiae* line, 3rd instar larvae males could be distinguished by expression of the reporter in the developing testes (Catteruccia *et al.* 2005; Windbichler *et al.* 2008). Sex-specific alternative splicing of the sex determination gene *doublesex* (*dsx*) has also been utilised to allow separation of 1st instar larvae. By integration of the coding sequence of the reporter GFP within the female-specific *dsx* exon, unexpectedly, males expressed selectively higher levels of GFP protein and could thus be separated from females (Magnusson *et al.* 2011).

As the GFP-containing female-specific exon would be skipped in males, it is surprising males expressed higher GFP levels. Identification of other genes with sex-specific expression or splicing would increase the feasibility of a mosquito directed RIDL strategy. Identification of such events in the embryonic stages would be of great benefit as separation at this early stage would reduce rearing costs associated with later sex-separation strategies.

RIDL requires a system for the conditional expression of a lethal gene. The development of the conditional expression systems, Tet-on and Tet-off in *An. stephensi* has also increased the available tools to generate a mosquito RIDL strategy (Lycett *et al.* 2004). The Tet-on and Tet-off systems are tetracycline-controlled transcriptional systems, able to drive expression of genes of interest from a specific regulatory promoter. In the Tet-on system, gene expression is activated in the presence of tetracycline. In the Tet-off system, expression is activated in the absence of tetracycline. The hybrid tetracycline transactivator protein (tTA) is a fusion of the herpes simplex virus VP16 transcriptional activator and the tetracycline repressor (tetR) of *Escherichia coli* that drives tetracycline repressible (Tet-off) expression from the *E. coli* tetracycline responsive element (tetO) (Takahashi *et al.* 1986; Gossen and Bujard 1992). In contrast, fusion of a mutant tetR to VP16 allows tetracycline inducible expression (Tet-on) via the reverse tetracycline transactivator protein rtTA (Gossen *et al.* 1995).

The Tet-off expression system is utilised in a number of RIDL transgenic strains and is summarised in Figure 1.1. Shown in Figure 1.1A and B is a one-component system. In this system, a specific promoter (e.g. female-specific), regulates transcription of tTA. In the presence of tetracycline (Fig1.1A), tTA is inhibited, but in the absence of tetracycline (Fig1.1B), tTA binds to tetO for transcription of a lethal gene. In Figure 1.1C and D, a positive-feedback loop one-component system is shown, that does not contain a specific promoter to initiate tTA transcription. In the presence of tetracycline (Fig. 1.1C), tTA is unable to bind to tetO and cannot activate transcription. Tetracycline is able to bind to tetR dimers, preventing tTA binding to tetO (Takahashi *et al.* 1986). In the absence of tetracycline (Fig1.1D), expression of tTA from the tetO regulatory element results in a positive feedback loop, driving tTA to toxic transcriptional squelching levels (Sadowski *et al.* 1988). Initiation of the positive feedback loop is thought to occur through basal transcription levels from tetO (Gong *et al.* 2005).

Prototype tTA lethal systems have been created in *D. melanogaster*, driving expression of a repressible female-specific lethal gene (*msl*-2^{NOPU}) via a non-sex-specific promoter or use of a female-specific promoter (*yp3*) to drive expression of a cytotoxic gene (*Ras64B*^{V12}) (Thomas *et al.* 2000). Conditional female-specific lethality driven from the *yp1* enhancer has also been demonstrated (Heinrich and Scott 2000). Tetracycline repression of embryonic mortality has been demonstrated using expression of tTA from the promoters of blastoderm cellularisation genes (*serendipidy* and *nullo*), driving the pro-apoptotic gene *hid* (Horn and Wimmer 2003). RIDL control strategies have also been developed to target *Ceratitis capitata*, an economically important agricultural pest. An early acting (pre-pupation) dominant lethal strain using tetracycline repressible tTA under the control of tetO to drive a positive feedback loop system described in Figure 1.1C, resulting in toxic tTA levels was created that demonstrated comparable fitness to wild type individuals (Gong *et al.* 2005).

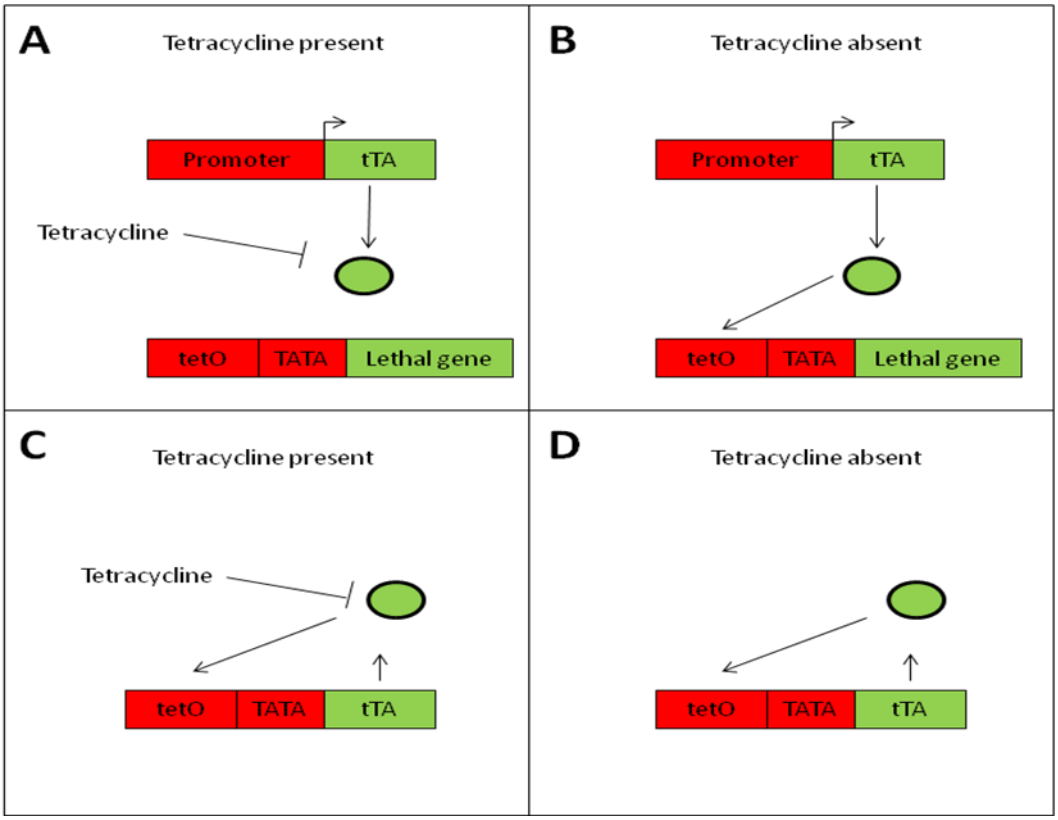


Figure 1.1. Tetracycline repressible (Tet-off) expression system. A schematic representation of one component Tet-off expression systems are shown. (A) A one component tet-off system with tTA under the control of an inducible promoter e.g. female or tissue specific. In the presence of tetracycline tTA is inhibited, preventing tTA binding to tetO and driving expression of a lethal gene. (B) In the absence of tetracycline, tTA drives expression of the lethal gene from tetO. (C) In a positive feedback loop one component system, tTA drives a positive feedback loop of its own expression. In the presence of tetracycline, tTA is prevented from driving tTA transcription. (D) In the absence of tetracycline, tTA is able to active its own expression in a positive feedback loop to toxic levels.

This positive feedback system has been utilised to induce lethality in *Ae. aegypti,* enabling tetracycline repression of male and female mortality of 4th instar larvae and pupae (Phuc *et al.* 2007). In the absence of tetracycline, successful mating with wild females results in mortality of progeny prior to adulthood. The same positive feedback loop has been made female-specific in *C. capitata* via insertion of a female-specific intron of the gene *transformer* into the tTAV (a tTA variant optimised for expression by removal of rare codons) coding sequence. The intron is spliced out in females only, leading to the production of toxic levels of functional tTAV and female-specific mortality (Fu *et al.* 2007).

Sex-specific RIDL approaches have been developed for mosquitoes, utilising the *actin-4* female-specific promoter to drive toxic tTA levels in the indirect flight muscles of *Ae. aegypti*, resulting in a flightless phenotype. As females are unable to fly, they are effectively rendered sterile, as mating requires a wing oscillation courtship song (Cator *et al.* 2009; Fu *et al.* 2010). Sex-specificity was further ensured via insertion of an additional start codon in a UTR of a male splice variant that introduces a premature stop codon. Female-specific flightless transgenic *Ae. aegypti* were able to eliminate target populations within 20 weeks in laboratory cage trials (Fu *et al.* 2010; de Valdez *et al.* 2011). Late acting sterility (effectively at the adult stage), increases the density-dependent effects of released transgenic mosquitoes with the wild population for resources and therefore increases the effect on population suppression (Phuc *et al.* 2007). A caveat of this approach is the requirement for additional rearing effort to later developmental stages that increases associated costs prior to release. Early acting sex-specific embryonic mortality is therefore beneficial in the context of a mass release programme.

Recently, it has been shown that targeting the germ cell development gene *zero population growth* (*zpg)* via RNAi renders males spermless and results in eggs that do not hatch (Thailayil *et al.* 2011). As with SIT and RIDL, the approach described by Thailayil *et al.* (2011) interferes with male fertility, aimed at reducing population densities through unsuccessful matings. It should be noted that the approach taken by Thailayil *et al.* (2011) transiently reduces transcript levels of *zpg*, thus scaling up the technique for mass release would prove problematic. Further investigation of male biology may identify additional candidates that affect reproductive capacity. In addition, development of RIDL strategies will also benefit from identification of further sex-specific genes that can donate regulatory regions to drive

conditional lethality and these genes are likely to play a role in sex determination or differentiation.

Sex determination cascade is initiated in early embryos and expression of its components continues through all life stages. Identification of the genes involved will be most likely achieved at the earliest stages of development (embryos), before the complexity of sex-specific transcripts is greatly increased by the progress of sexual differentiation (Graveley *et al.* 2011). Similarly, identification of sex-specific promoters in the embryonic stages would enable creating transgenic mosquitoes with very early acting mortality. Population suppression requires the release of millions of transgenic insects; therefore any reduction in rearing costs would be very beneficial when put into this mass release context.

1.4 Sex determination

Sexual dimorphism observed in insects is primarily due to sex chromosome composition of the individual. *An. gambiae* has two pairs of autosomes and a pair of sex chromosomes with females carrying a homomorphic karyotype (XX) and males heteromorphic (XY). Sex determination has been studied in a variety of insect orders, principally using the well studied *D. melanogaster* pathway as reference (shown in Figure 1.2). A defined hierarchy of genes (*sxl* > *tra*/*tra2* > *dsx*/*fru*) in *D. melanogaster* stems from sex chromosome composition (XX female and XY male) and is responsible for sexual dimorphism. The identification of mutant XX diploid males lead to the discovery of molecular control of sex determination and the loci containing *transformer* (*tra*), *transformer2* (*tra2*) and *doublesex* (*dsx*), were all subsequently identified as being part of the sex determination pathway (Sturtevant 1945; Baker and Ridge 1980). Insect sex determination pathways are believed to have undergone 'bottom up' evolution (Wilkins 1995; Shearman 2002). 'Bottom up' evolution refers to the conservation of the ancient final step gene in the sex determination pathway, whereas upstream regulatory genes have been recruited more recently. This is demonstrated in the remarkable conservation of *dsx* (the terminal sex determination transcription factor) structure and function in dipterans (Hediger *et al.* 2003; Scali *et al.* 2005), lepidopterans (Ohbayashi *et al.* 2001) and hymenopterans (Cho *et al.* 2007; Oliveira *et al.* 2009) and great evolutionary lability of genes regulating *dsx* expression (Schutt and Nothiger 2000; Verhulst *et al.* 2010; Salz 2011). Central to all insect sex determination

systems elucidated to date is sex-specific splicing of key genes that control sex determination and are first expressed in the embryonic stages.

1.4.1 Sex determination in *D. melanogaster*

1.4.2 *D. melanogaster: Sex-lethal*

In *D. melanogaster*, *sex-lethal* (*sxl*) is sex-specifically spliced at the summit of the sex determination cascade and encodes a RNA binding protein (Sanchez and Nothiger 1982; Bell *et al.* 1988). SXL is presumed only functional in females; presence of premature stop codons within a male-specific exon leads to truncated protein product. In females, S*xl* establishes a positive feedback loop of its own splicing to maintain female-specific splicing (Bell *et al.* 1988; Bell *et al.* 1991; Keyes *et al.* 1992; Horabin and Schedl 1993). To ensure female-specific S*xl* splicing, SXL binds to introns flanking the male-specific third exon, disrupting activity of general splicing factors and causing exon skipping (Nagengast *et al.* 2003; Salz and Erickson 2010).

Sxl expression is initially driven from an early (*SxlPe*) promoter and then from a maintenance (*SxlPm*) promoter that share a common enhancer and are activated via X-linked signal elements (XSEs) (Erickson and Cline 1993; Kramer *et al.* 1999). Females incorporate an extra nuclear division prior to dosage compensation (the process of X chromosome hyper transcription in heterogametic males) to allow accumulation of sufficient XSEs to initiate *Sxl* expression (Erickson and Quintero 2007; Gonzalez *et al.* 2008). This allows early SXL (*SxlPe*) to sex-specifically splice late *Sxl* transcripts (*SxlPm*) to establish auto-regulation not found in males. XSEs include *sisterless-A* and *sisterless-B* (Cline 1988), *runt* (Duffy and Gergen 1991) and the JAK/STAT pathway (Jinks TM *et al.* 2000). Though activation of early *Sxl* transcription is required to initiate splicing of late *Sxl* trasncripts, it does not do so alone, requiring general splicing factors also (Salz and Erickson 2010). Small nuclear ribonucleoproteins (snRNPs) are splicing factors that are essential components for the formation of the spliceosome, an RNA-protein complex that directs splicing of immature messenger RNA (Graveley 2000). Female SXL polypeptides interact with a number of general splicing factors including Sans-fille, Fl(2)d and splicing factor U1snRNP, preventing them from activating the male-specific splice site (Flickinger and Salz 1994; Nagengast *et al.* 2003; Penn *et al.* 2008).

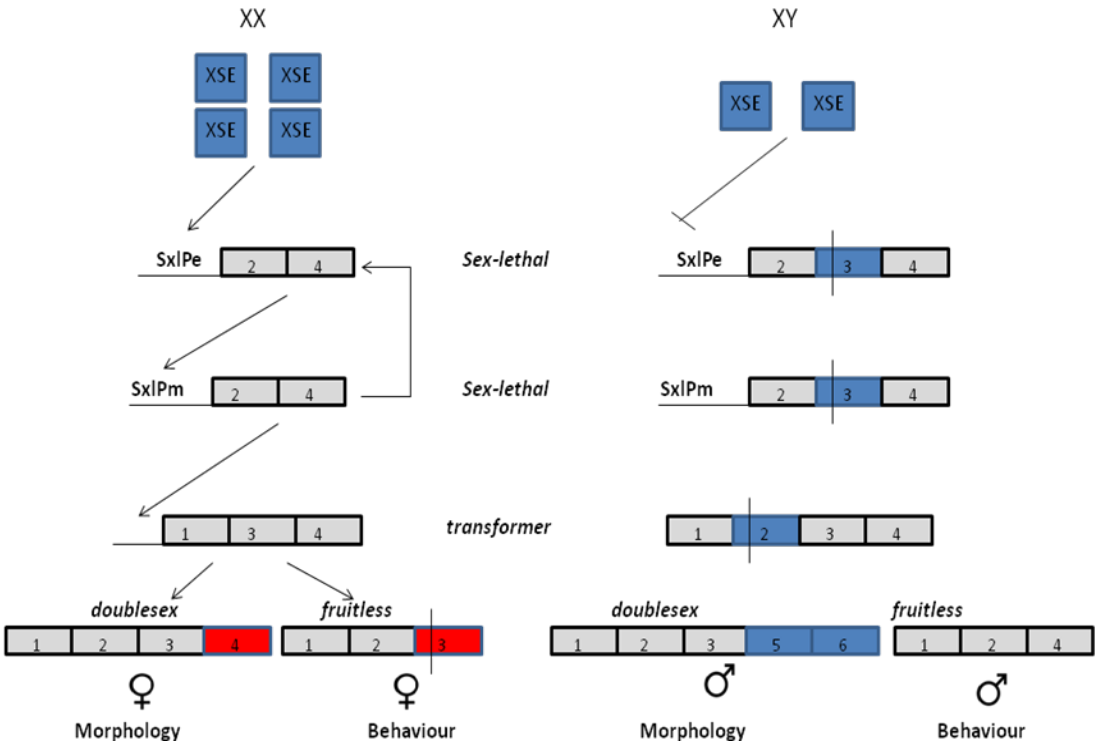


Figure 1.2. Schematic diagram of sex-specific alternative splicing events in the *D. melanogaster* sex determination pathway. Sex determination revolves around alternative splicing of key genes in the *D. melanogaster* pathway that is initiated by the level of X-linked signal elements (XSEs). Boxes represent exons; blue boxes represent male specific exons; red boxes represent female specific exons. Vertical lines indicate presence of premature stop codons. Gene names from which transcripts originate are indicated. In females, double the dose of XSEs is sufficient to drive expression from the early sxlPe promoter, skipping the male-specific exon 3 to produce a full length polypeptide that then regulates female-specific splicing of *sex-lethal* (*sxl*) transcripts from the late SxlPe promoter. Functional SXL establishes female-specific splicing of transformer (*tra)* by exon skipping of the male-specific exon 2. TRA regulation of *doublesex* (*dsx)* and *fruitless* (*fru)* splicing causes inclusion of a female-specific fourth *dsx* exon and a stop codon containing *fru* exon 3. The direct effects of these two transcription factors are responsible for observed sexual dimorphism. Not shown is the constitutively expressed non-sex-specific *transformer 2*, that regulates dsx and fru sex-specific splicing with functional female TRA. In males (XY), reduced levels of XSEs are insufficient to activate transcription of *sxl* from the early sxlPE promoter. Transcription from the late promoter (sxlPm) therefore follows a default splicing pattern and inclusion of a stop codon containing exon. Non-functional SXL is then unable to activate the female-specific 3' splice site of *tra* exon 3 and a stop codon containing exon 2 is included in the mature transcript. Non-functional TRA causes default male splicing of the downstream transcription factors *dsx* and *fru*. Male *dsx* transcripts have two male-specific exons (5 and 6) and *fru* transcripts splice out a female-specific exon 3.

1.4.3 *D. melanogaster: transformer and transformer2*

Functional SXL targets transcripts of the downstream gene *tra*. In females, SXL binds to the first intron of *tra* pre-mRNA, negatively regulating a non-sex-specific 3' splice site by preventing binding of splicing factors and causing inclusion of a female-specific exon (Inoue *et al.* 1990; Valcarcel *et al.* 1993). In males, absence of functional SXL leads to default inclusion of a male-specific exon that contains a premature stop codon and a non-functional TRA protein (Sosnowski *et al.* 1989; Inoue *et al.* 1990). TRA is a serine/arginine (SR) rich protein, containing a RNA recognition motif (RRM). SR proteins are splicing factors that interact with both RNA transcripts and other proteins via RNA recognition motif (RRM) and serine arginine (SR) rich regions respectively (Long and Caceres 2009). A second SR rich protein, TRA2, contains a RNA recognition motif and functions with TRA for control of downstream genes. There are three alternatively spliced *tra2* transcripts of which two function within the somatic tissues with TRA, and a single isoform functions in the male germline (Amrein *et al.* 1990; Mattox *et al.* 1996). Downstream of *sxl* and *tra/tra2*, there are two branches to the pathway that regulate the sex-specific splicing of two gene, *dsx* and *fru*.

1.4.4 *D. melanogaster: doublesex*

Functional TRA, along with the constitutively expressed, non-sex-specific RNA binding protein TRA2 are responsible for the female-specific splicing of the transcription factor *dsx* (Hedley and Maniatis 1991; Hoshijima *et al.* 1991). TRA and TRA2 bind to a splicing enhancer (dsxRE) in *dsx* transcripts and cause inclusion of a female-specific exon via a weak 3' splice site and recruitment of general splicing factors. The dsxRE consists of six copies of a 13 nt repeat sequence and a purine rich region to which serine/arginine (SR) rich proteins bind along with TRA and TRA2 to facilitate female-specific splicing (Lynch and Maniatis 1996). Absence of TRA in males leads to default *dsx* expression and a protein that differs at its carboxyl terminus, being 32 amino acids in length compared to 152 amino acids in females (Baker and Wolfner 1988; Burtis and Baker 1989). Translation of both male and female DSX proteins (DSX^M and DSX^F, respectively) begins in the second exon. Translation terminates in the female specific exon four and male specific exon five for the two different DSX proteins (Burtis and Baker 1989).

Few targets of the DSX transcription factor have been identified; those that have include yolk protein (*yp*) genes, *bric á brac1* and *fad2* (Burtis *et al.* 1991; Williams *et al.* 2008; Shirangi *et al.* 2009). Bioinformatics analysis of DSX binding sites suggests that a larger number of genes are targeted (Luo *et al.* 2011). The sex-specific forms of DSX both promote sex differentiation and repress differentiation of the opposite sex to determine sexual dimorphism (Baker and Ridge 1980; Waterbury *et al.* 1999). DSX^M protein binds to and represses transcription of the female-specific *yp* genes, where as DSX^F activates *yp* transcription (Coschigano and Wensink 1993; An and Wensink 1995). The ratio of male to female DSX protein has also been shown critical in determining formation of male-specific structures and for expression of female-specific *yp* genes (Waterbury *et al.* 1999).

Targets of DSX regulation are responsible for the formation of sexually dimorphic cuticular structures that arise from imaginal discs (Cohen 1993). Two non-sex-specific genes, *intersex* (*ix*) and *hermaphrodite* (*her*) act together with *dsx* to regulate formation of sexually dimorphic features (Pultz *et al.* 1994; Chase and Baker 1995). *her* mutants have an intersexual phenotype, affecting the developing female genitalia and sexually dimorphic abdomen pigmentation (Pultz *et al.* 1994; Li and Baker 1998). In addition to this, *her* is independently required for the sex-specific expression of *yp* genes. *her* activates transcription of *yp* genes in both sexes, but this is repressed by the male form DSX protein to ensure *yp* genes are female-specific (Li and Baker 1998). In *dsx* or *ix* mutants, removal of the other gene has no effect of sexual phenotype, demonstrating that they act together to regulate sexual dimorphism (Waterbury *et al.* 1999; Garrett-Engele *et al.* 2002). Female DSX protein physically interacts with IX for the control of sexually dimorphic cuticular structures, such as the vaginal teeth and abdominal pigmentation (Garrett-Engele *et al.* 2002).

The genital disc comprises of sexually dimorphic epithelia structures (primordia) that allow discrimination between males and female genitalia (Estrada *et al.* 2003). Two distinct primordial cell populations develop in the embryo that are capable of forming both male and female genitalia (Epper and Nothiger 1982). The development of male and female genitalia in *dsx* mutants suggested that *dsx* represses formation of primordia of the opposing sex (Nothiger *et al.* 1987). More recently, it has been demonstrated the *dsx* has an instructive, rather than repressive role in primordia formation (Keisman *et al.* 2001). Sex-specific *dsx* splicing within cells of the anterior/posterior (A/P) organiser (a group of

cells that direct growth and patterning of genital imaginal discs), controls the formation of the male and female genitalia (Chen and Baker 1997). Feminising male A/P cells to express female form *dsx* (through expression of full length *tra*) results in the formation of female genitalia. Similarly, masculinisation of female A/P cells to express male form *dsx* (through a *tra2* null mutant) results in formation of male genitalia (Keisman *et al.* 2001). Altering *dsx* sex-specific splicing in these cells affects expression of *wingless* (wg), *decapentaplegic (dpp) and dachshund* (*dac*), three genes with known involvement in formation of genitalia structures from the imaginal discs (Keisman and Baker 2001; Sanchez *et al.* 2001). The work by Keisman *et al.* (2001) and Sanchez *et al.* (2001) demonstrated the sex-specific splicing of *dsx* is responsible for aspects of somatic sexual dimorphism observed between males and females. An additional role of DSX^M protein is to function within the central nervous system (CNS) to establish sexually dimorphic behaviour. *dsx* null mutant males have an abnormal courtship song and altered courtship behaviour (McRobert and Tompkins 1985; Lee *et al.* 2002; Rideout *et al.* 2007).

Recently, it has been demonstrated that *dsx* is not expressed in all somatic cells, but restricted to specific tissues with known (e.g. genitalia) and, interestingly, those without (e.g. leg muscles) previously described sexually dimorphic features (Robinett *et al.* 2010). Spatial restriction of *dsx* transcripts has also been shown in the embryonic stages, with expression restricted to the developing somatic cells of the gonads (Tomancak *et al.* 2002). This reveals an additional level of *dsx* regulation; *dsx* transcripts must be present in a cell so that SXL and TRA can act upon them for female-specific splicing. This spatial and temporal regulation is important for DSX^F negative regulation of the *D. melanogaster* fibroblast growth factor *branchless* (*bnl*). *bnl* is required in the male genital disc for formation of male internal genitalia, yet is also required for tracheal development in both males and females (Sutherland *et al.* 1996; Ahmad and Baker 2002). Therefore, spatial regulation of DSX^F is important to negatively regulate *bnl*, to prevent formation of the male genital disc.

1.4.5 *D. melanogaster: fruitless*

The second branch of the sex determination cascade that is controlled by TRA and TRA2 is the sex-specific splicing of *fru* transcripts (Heinrichs *et al.* 1998). FRU is a BTB zinc-finger transcription factor and is responsible for aspects of male mating behaviour through expression in the CNS and the formation of the male-specific muscle, the muscle of

Lawrence (MOL) (Gailey *et al.* 1991; Ito *et al.* 1996; Vrontou *et al.* 2006). The MOL is a large bilateral muscle in the fifth abdominal segment of adult abdomens and has been shown to be non-essential for male fertility (Gailey *et al.* 1991). Function of *fru* in male mating behaviour has been shown by *fru* mutant males demonstrating abnormal copulation with females in addition to homosexual courtship (Ito *et al.* 1996).

Both sex and non-sex-specific transcripts of *fru* are driven from multiple promoters (P1-P4), of which only transcripts from a single promoter (P1) are sex-specifically spliced by the sex determination hierarchy (Anand *et al.* 2001). Binding of TRA and TRA2 to an exonic splicing enhancer (fruRE) in P1 *fru* transcripts activates a 5' splice site causing inclusion of a female-specific exon containing a premature stop codon (Heinrichs *et al.* 1998; Lee *et al.* 2000; Anand *et al.* 2001; Lam *et al.* 2003). Non-functional TRA in males means that the 5' splice site is not activated and default splicing leads to male-specific *fru* mRNA and protein. Sex-specific *fru* transcripts are only present in a subset of neurons within the CNS (Ryner *et al.* 1996). Anti-FRU antibodies only detect male protein within these neurons, suggesting that female transcripts are not translated and that sex-specific splicing is not the sole regulator of fru function (Lee *et al.* 2000; Usui-Aoki *et al.* 2000). Non-sex-specific *fru* transcripts function within the CNS to direct neuronal development in embryonic and adult stages (Anand *et al.* 2001; Song *et al.* 2002).

1.4.6 Sex determination in non-drosophilid species

The presence of *fru* at the bottom of the *D. melanogaster* sex determination cascade is conserved in other insect species. *fru* homologues have been identified in dipterans (*An. gambiae* and *C. capitata)* (Gailey *et al.* 2006; Salvemini *et al.* 2009), hymenopterans (*Nasonia vitripennis and A. mellilfera*) (Bertossa *et al.* 2009), Blattodea (*Blattella germanica)* (Clynen *et al.* 2011) and Orthoptera (*Schistocerca gregaria)* (Boerjan *et al.* 2011). Sex-specific *fru* expression is under the control of TRA and TRA2 in *C. capitata* (Salvemini *et al.* 2009) and TRA/TRA2 consensus binding sites have been identified in *An. gambiae fru* transcripts (Gailey *et al.* 2006). Conservation of function has been demonstrated by the rescue of *D. melanogaster fru* mutants by ectopic expression of the *An. gambiae fru* homologues (Gailey *et al.* 2006). Expression of *An. gambiae* male *fru* transcripts in *D. melanogaster fru* mutant males and wild type females resulted in the formation of the muscle of Lawrence (MOL) (Gailey *et al.* 2006).

Further up the cascade, *tra* shows a remarkable conservation of function and splicing regulation but a distinct lack of sequence similarity. All identified thus far *tra* orthologues, known from dipteran families Tephritidae (*Anastrepha* species, *Bactrocera oleae* and *C. capitata*), Muscidae (*Musca domestica*) and Calliphoridae (*Lucilia cuprina*) (Pane *et al.* 2002; Lagos *et al.* 2007; Ruiz *et al.* 2007; Concha and Scott 2009; Hediger *et al.* 2010) contain serine-arginine dipeptides and proline rich regions, a feature of all SR-proteins, to which *tra* belongs (Long and Caceres 2009). Conservation of TRA function has also been demonstrated in the hymenopteran *N. vitripennis* (Verhulst *et al.* 2010). *A. mellifera feminizer (fem)* is a likely homologue of *D. melanogaster tra* and is sex-specifically spliced based upon hemi or homozygosity of the upstream gene *complementary sex-determiner* (*csd*) (Hasselmann *et al.* 2008; Gempe *et al.* 2009). Similar to *Dmtra*, male-specific *fem* transcripts contain premature stop codons, where as females encode a full length polypeptide and as in *D. melanogaster*, female FEM regulates sex-specific splicing of the downstream gene *dsx* (Hasselmann *et al.* 2008; Gempe *et al.* 2009). This degree of functional conservation within different insect orders is perhaps surprising as low sequence similarity is found even within *Drosophila* species (O' Neil and Belote 1992).

Outside of *Drosophila* species, *tra* has developed an assumed autoregulatory role and in some instances performs the role of *Sxl* at the top of the sex determination hierarchy. Maternally deposited *M. domestica tra* (*Mdtra*) acts as a female determiner by establishing an auto-regulatory feedback loop to maintain female-specific splicing of zygotic *Mdtra*. In males, an M factor acts to prevent establishment of the feedback loop, leading to default non-functional *Mdtra* transcripts (Hediger *et al.* 2010). This system is demonstrated by arrhenogenic mutant females, that have no maternal contribution of *Mdtra* and therefore produce only male offspring as the feedback loop is not initiated (Schmidt *et al.* 1997). The molecular identity of the Y-linked M factor has yet to be established but it is interesting to note that translocations can cause the M factor to be present on either autosomes or X chromosomes in wild populations (Hiroyoshi 1964; Inoue and Hiroyoshi 1982; Denholm *et al.* 1983).

L. cuprina and *N. vitripennis* embryos also establish auto-regulation of *tra* sex-specific splicing, similar to *M. domestica* (Concha and Scott 2009; Verhulst *et al.* 2010). In fertilised female *N. vitripennis* eggs, *tra* (*Nvtra*) zygotic expression along with maternal Nv*tra*

deposition is sufficient to sustain a feedback loop to promote female development. Males that arise from unfertilised eggs do not reach this threshold due to lack of zygotic expression and therefore express default Nv*tra* transcripts containing premature stop codons (Verhulst *et al.* 2010). Presence of consensus TRA binding sites in *L. cuprina tra (Lctra)* and *dsx (Lcdsx)* transcripts, suggests that *Lctra* is responsible for its own auto-regulation and for regulation of *Lcdsx* sex-specific splicing (Concha and Scott 2009; Concha *et al.* 2010). Similarly, *C. capitata tra (Cctra)* regulates *Ccdsx* alternative splicing and also regulates sex-specific splicing of *Cctra* transcripts. Structurally conserved TRA binding sites found in *Ccdsx* transcripts are also found in *Cctra* pre-mRNA, indicating that TRA is able to bind and regulate splicing in both *Ccdsx* and *Cctra* transcripts (Pane *et al.* 2002). Though protein similarity is low (~18 %), CcTRA function is conserved in *D. melanogaster* when ectopically expressed (Pane *et al.* 2005). A Y chromosome derived M factor prevents Cc*tra* establishing autoregulation in males and is presumed to affect Cc*tra* splicing or protein activity. No difference in transcription levels is found for *Cctra* between males and females, therefore it is presumed that the M factor regulates sex-specific splicing or has a posttranscriptional effect on CcTRA protein activity (Willhoeft and Franz 1996; Gabrieli *et al.* 2010).

TRA proteins that exhibit autoregulation are termed *tra*^{ep} (epigenetic) and show limited similarity to *D. melanogaster* TRA, but share an additional region of homology termed TRA-CAM (*Ceratitis Anasterpha Musca*) (Hediger *et al.* 2010). It has been hypothesised that *DmTra* is derived from the ancestral *tra*^{ep} and that *Sxl* was recruited for *tra* regulation in *D. melanogaster* (Saccone *et al.* 2011). In accordance with this, it has been proposed that the study of sex determination in insects should adopt a less *Drosophila* centric outlook. Variations of the control of *dsx* sex-specific expression are summarised in Figure 1.3. TRA2 has also adopted a secondary function in other dipterans, both regulating *dsx* splicing as in *D. melanogaster* but also regulating *tra* splicing. Both *M. domestica tra2* (*Mdtra2*) and *C. capitata tra2* (*Cctra2*) regulate female-specific splicing of *dsx* and *tra* in conjunction with *tra* itself (Burghardt *et al.* 2005; Salvemini *et al.* 2009).

1.4.7 *An. gambiae* sex determination

With regards to *An. gambiae* sex determination, early experiments identified a 6 kb genomic fragment with 75 % sequence similarity to DmSXL. From this region several probes were designed that showed no sex-specific transcripts following Northern blots of poly(A)+ RNA from males and females. Within the same study, a fragment of *An. gambiae dsx* was amplified from a mixed sex cDNA library using degenerate primers, designed against the proteins sequences of *Drosophila* DSX proteins. Screening of the cDNA library with the amplified fragment isolated further clones that showed sex-specific amplification (Pannuti *et al.* 2000). More recently, knowledge has been extended to the identification of full length transcripts of two genes, *dsx* and *fru*, that function in *An. gambiae* sex determination (Scali *et al.* 2005; Gailey *et al.* 2006). Sex-specific splicing of *Aedes aegypti dsx* transcripts has also been demonstrated (Salvemini *et al.* 2011).

Sex-specific *An. gambiae dsx* transcripts were identified through sequence similarity searches to DmDSX against *An. gambiae* EST databases and screening of cDNA libraries (Scali *et al.* 2005). *An. gambiae* sex-specific splicing occurs due to female-specific exon inclusion by activation of a downstream 5' splice site. A male-specific exon is transcribed in females as part of the UTR and sex-specific splicing of *dsx* mRNA was shown to occur during adulthood (Scali *et al.* 2005). Recently, sex-specific splicing of *dsx* transcripts has been shown to occur as early as 1st instar larvae (Magnusson *et al.* 2011). Similarly, sequence similarity searches using male *Dmfru* sequence to query the *An. gambiae* genome identified a *fru* homologue (Gailey *et al.* 2006). It is interesting to note that male *fru* transcripts are found in whole adult females, whereas in the mosquito head, male and female transcripts are exclusive to their respective sex (Gailey *et al.* 2006). This is consistent with the sex-specific expression of male *Dmfru* transcripts in neuronal tissues that confer aspects of male sexual behaviour (Billeter and Goodwin 2004).

An. gambiae s*xl* is not sex-specifically spliced and presumably has no role in sex determination. Non-sex-specific transcription of s*xl* is conserved in other dipterans including *M. domestica*, *C. capitata* and *Megaselia. scalaris* (Meise *et al.* 1998; Saccone *et al.* 1998; Sievert *et al.* 2000; Traut *et al.* 2006). Presumed absence of a functional sex-determining SXL in *An. gambiae* means that upstream regulators of *dsx* and *fru* have yet to be identified. In other dipterans, a male determining factor has been shown to act

upstream of *dsx* and *fru* (Schmidt *et al.* 1997; Dubendorfer *et al.* 2002).. The identification of XXY triploid *Anopheles culicifacies* males has suggested that a male determining factor is present on the Y chromosome (Baker and Sakai 1979).

Though variations exist between insect sex determination pathways, central to all is sex-specific splicing and auto-regulation of a regulatory gene (*sxl* or *tra*), that controls sex-specific splicing of a downstream effector (*dsx* and *fru*). Identification of homologues of downstream sex determination genes within different insect families supports the theory of 'bottom-up' evolution for insect sex determination (Wilkins 1995). It is interesting to note that both *D. melanogaster* and *A. mellifera* employ an additional level of control at the summit of the pathway and that both arose from gene duplication events. In *D. melanogaster*, SXL and CG3056 are paralogues and it is presumed that SXL has adopted the function of a sex determination gene (Traut *et al.* 2006). Null phenotype CG3056 mutants, more recently termed *sister-of-Sex-lethal* (*ssx*), have been shown to have no effects upon sex determination or sex-specific functions (Cline *et al.* 2010). In *A. mellifera*, heterozygosity at the *csd* locus controls female-specific splicing of *fem*. Both *csd* and *fem* are located within the *A. mellifera* sex determination locus (SDL) and *csd* is hypothesised to originate from the ancestral *fem* (Hasselmann *et al.* 2008).

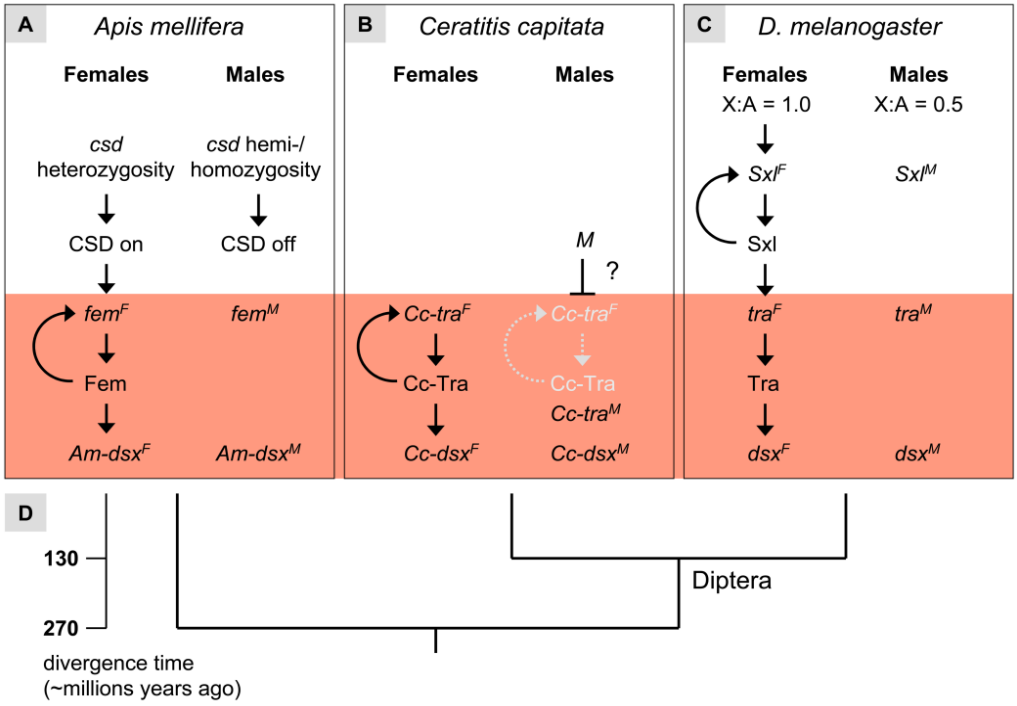


Figure 1.3. Three sex determination systems present in Insecta. Three variations, around a common theme of autoregulation, alternative splicing and *doublesex* (*dsx*) are present within Insecta. Despite an evolutionary distance of ~270 million years between the three species (A = *A. melilifera*; B = *C. capitata*; C = *D. melanogaster*), all three systems show remarkable similarity in the regulation of sex, depending upon establishing autoregulation leading to sex-specific splicing of a binary switch gene. Image taken from (Gempe *et al.* 2009). A) *Complementary sex determiner* (*csd*) heterozygosity determines female-specific splicing of the *transformer* (*tra*) homologue *feminizer* (*fem*) that undergoes autoregulation leading to sex-specific *dsx* splicing. Non-heterozygosity of *csd* prevents the feedback loop initiating. B) Zygotic expression of female *tra* transcripts establishes an autoregulatory feedback loop that then leads to female-specific splicing of *dsx*. In males, a male promoting M factor prevents formation of the feedback loop and default male *dsx* transcripts are expressed. C) Number of sex-chromosomes establishes female-specific sex-lethal splicing of *tra* leading to female development via *dsx*. Half the number of sex chromosomes in males is insufficient to establish *sxl* expression leading to default male expression of *dsx*.

1.4.8 Temporal control of sex determination by the maternal to zygotic transition

In addition to regulation by upstream genes, correct sexual differentiation also relies upon correct temporal expression of key genes. The maternal to zygotic transition (MZT) is a stage of embryonic development where maternally deposited transcripts are targeted for degradation and transcription from zygotic chromosomal DNA template begins. Following MZT, developmental responsibility for sex determination is transferred from maternally deposited transcripts to zygotic transcription. *D. melanogaster* embryos are competent for zygotic expression at cycle 10 (~100 min after fertilisation) though the majority of genes are not expressed until cycle 14 (Edgar and Schubiger 1986). Prior to cycle 14, mitosis occurs without cytoplasmic divisions (with the exception of the developing pole cells), forming a syncytial blastoderm. During interphase of cycle 14, nuclei undergo cellularisation and these cells contain progenitors of all somatic tissues (Foe and Alberts 1983). Prior to cellularisation, embryo development is reliant upon maternal provision of transcripts and proteins. Following cellularisation, this responsibility is transferred to zygotic transcription. This has been demonstrated through chromosome arm ablation in *D. melanogaster* embryos. Removal of autosome and X chromosome arms that serve as template for zygotic transcription leads to developmental abnormalities (Merrill *et al.* 1988; Wieschaus and Sweeton 1988).

The transcription factor *Zelda* is responsible for activation and temporal expression of zygotic transcription via binding to a consensus cis-regulatory motif termed TAGteam present in target genes, including those involved in sex determination (ten Bosch *et al.* 2006; Liang *et al.* 2008). Up to a third of maternally deposited transcripts are degraded by onset of MZT at cycle 14 and these share a common UTR motif directing them for degradation (De Renzis *et al.* 2007). Degradation is location specific in the embryo; for example transcripts of *hsp83* and *nanos* are degraded in the bulk cytoplasm but not within posterior pole cells (Wang and Lehmann 1991; Ding *et al.* 1993). There is both a maternal and zygotic contribution to degradation as maternal *Hsp83* takes longer to be degraded in unfertilised eggs (Bashirullah *et al.* 1999). Deadenylation is the main process leading to degradation. Smaug (SMG) and piwi-interacting RNA (piRNA) recruit the CCR4–NOT deadenylation complex to transcripts targeted for degradation (Semotok *et al.* 2005; Tadros *et al.* 2007; Rouget *et al.* 2010). SMG is a highly conserved RNA binding protein that acts as a posttrancriptional regulator to destabilise maternal mRNA by binding to a conserved SMG response element (SRE), in the 3' UTR region of maternal transcripts

(Smibert *et al.* 1999; Tadros *et al.* 2007). Microarray analysis has shown that two thirds of degraded transcripts are associated with functional SMG (Tadros *et al.* 2007). piRNAs are short 24 to 30 nucleotide RNAs that are present in early embryos and together with Piwi proteins, regulate the transcription of transposable elements (Brennecke *et al.* 2007; Brennecke *et al.* 2008). Recently, it has been shown that piRNAs and Piwi proteins also regulate stability of maternal transcripts through binding within 3' UTRs (Rouget *et al.* 2010).

Correct temporal regulation of the MZT is important to many developmental processes, including the sex determination pathway, as in *D. melanogaster*, Zelda regulates *Sxl* expression. Zelda activates transcription of XSEs as well as early *sxl* expression (sxlPE) through binding to TAGteam motif regulatory regions (ten Bosch *et al.* 2006; Nien *et al.* 2011). MZT is also important for sex determination in other insects. *N. vitripennis* relies upon maternal contribution of *Nvtra* to elicit a feminising effect upon zygotic *Nvtra* expression to maintain female-specific splicing. Therefore, correct timing of MZT is essential to ensure sufficient maternal female *Nvtra* transcripts are present con-currently with zygotic *Nvtra*, to establish an autoregulatory feedback loop (Verhulst *et al.* 2010). Premature degradation or delayed zygotic transcription would prevent the feeback loop being established, leading to default male-specific expression. Conversely, in *C. capitata* maternally provided *Cctra* is insufficient to establish *Cctra* autoregulation. Maternally deposited *dsx* is degraded and sexual identity effectively reset until accumulation of sufficient zygotically expressed *Cctra* at the onset of cellularisation (Gabrieli *et al.* 2010). Therefore *C. capitata* sex determination is not reliant on maternal *Cctra* contribution and highlights the differences in sex determination regulation. A possible dependence upon maternal contribution of *tra2* in *C. capitata* embryos is apparent as only mature *tra2* transcripts are identified in the embryo (Gabrieli *et al.* 2010).

Prior to the availability of genome and transcriptome sequences, homologues of sex determination genes were largely identified *in vitro*. Degenerate primers, a mixture of primers able to bind to the alternative potential codons of amino acids, can be designed from consensus sequences from alignments of homologous protein sequences (Mitsuhashi 1996) and used to amplify gene fragments from cDNA or DNA templates. Such approaches have been used followed by the Rapid amplification of cDNA ends (RACE) from cDNA libraries or genomic DNA for identification of full length *sxl* homologues (Sievert *et al.* 1997;

Meise *et al.* 1998) and *dsx* (Shearman and Frommer 1998; Hediger *et al.* 2003). More recently, large scale sequencing has enabled identification of sex determination genes through transcriptome sequencing using Next Generation Sequencing (NGS) platforms.

The development of NGS has provided an exciting tool for transcriptome sequencing that may identify putative *An. gambiae* sex determination genes. Massively parallel sequencing of embryonic tissue presents the opportunity to sample early zygotic transcription that may identify genes involved in sex determination that are known to be expressed at this time. Previously, large scale capillary electrophoresis sequencing of expressed sequences tags (ESTs) has been used for *de novo* identification of sex determination genes (Gomulski *et al.* 2008). The *C. capitata* homologue of *tra2* was identified through sequence similarity searches to *Dmtra2* within ESTs generated from an embryo cDNA library (Gomulski *et al.* 2008). Use of NGS for sequencing of ESTs has also identified homologues of *tra* and *dsx* in the stable fly *Stomoxys calcitrans* (Olafson *et al.* 2010) and *sxl* in the screwworm fly, *Cochliomyia hominivorax* (Guerrero *et al.* 2009).

1.5 Next generation sequencing (NGS)

For the last 30 years it has been possible to isolate the nucleotide sequence of DNA fragments of interest, originating from the work of Frederick Sanger (Sanger and Coulson 1975). Based upon the inhibitory action of ddNTPs, Sanger sequencing or chain-termination sequencing required the amplification and cloning of fragments prior to sequencing (Sanger and Coulson 1975; Sanger *et al.* 1977). The advent of NGS technology has opened the door to faster and cheaper access to large scale sequencing. Various technologies are routinely used that differ in the read (sequenced fragment) size, number of reads, run time and cost that can be applied to limitless number of experimental conditions (Table 1.1). The cost of sequencing per Mb of information produced has markedly decreased since the introduction of NGS (Figure 1.4). The number of base pairs able to be sequenced per dollar investment has doubled every five months since 2004. As the amount of sequence data produced has increased, the requirement for computer storage requirements has similarly increased.

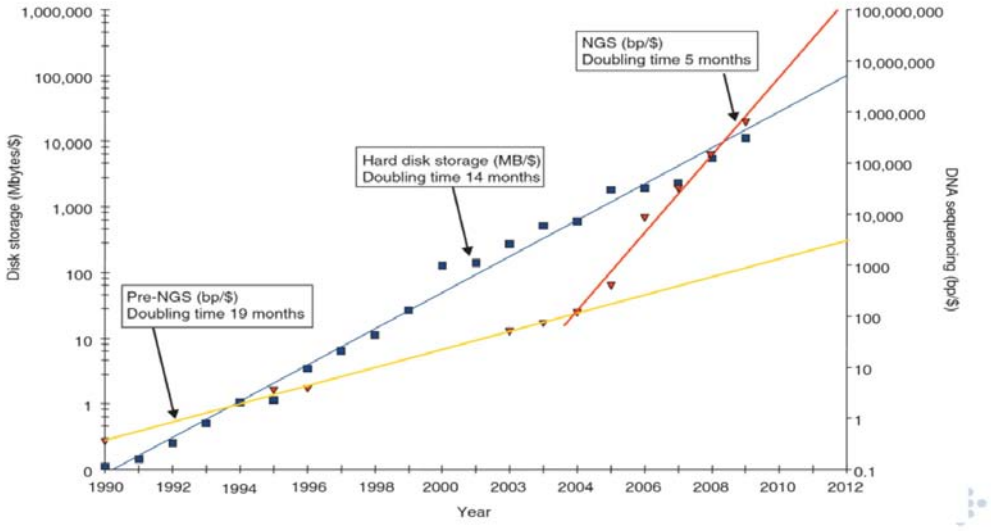


Figure 1.4. Historical changes in the storage space required and cost of DNA sequencing. In the last two decades, the cost of sequencing has dramatically reduced owing to the development of NGS. The number of base pairs sequenced per dollar is 10^9 greater today than by traditional Sanger sequencing methods. Conversely this has seen the amount of storage required for sequencing data to increase rapidly. The cost of sequencing a single human genome is expected to decrease to $1,000 in the next 5 years. Blue line = disk storage requirements. Yellow line = Sanger sequencing. Red line = NGS technology. Image taken from (Stein 2010).

1.5.1 Roche 454 Sequencing

The 454 sequence platform combines emulsion PCR with pyrophosphate sequencing (pyrosequencing) to currently produce ~ 1 million reads per sequencing run (Nyren *et al.* 1993; Ronaghi *et al.* 1996; Tawfik and Griffiths 1998). Pyrosequencing uses detection of emitted photons following nucleotide incorporation in sequencing by synthesis method. Randomly fragmented templates are adaptor ligated and attached to individual beads via limiting dilution and clonally amplified via emulsion PCR in water-oil emulsion micro-reactors. Enrichment is then performed for beads with amplified DNA strands and single beads are added to picoliter wells. Sequential flowing of sequencing reagents over the wells allows simultaneous template extension reactions to take place in each well. Following annealing of a sequencing primer, polymerase extension of templates results in release of inorganic pyrophosphates following nucleotide addition. Released pyrophosphates are converted to adenosine triphosphate (ATP) by sulfurylase in the presence of adenosine 5' phosphosulfate (APS). ATP excitation of luciferase generates

photons and intensity of the signal is proportional to the amount of ATP generated and therefore indicates the number of bases incorporated (Ronaghi *et al.* 1998; Margulies *et al.* 2005). The resulting flowgram image (signal intensity of released photons for each nucleotide flow) is then used for sequence analysis. Reads produced are up to 1,000 bp in length, with an average of 700 bp for the GS FLX Titanium platform.

1.5.2 Sequencing by Oligonucleotide Ligation and Detection (SOLiD)

SOLiD sequencing from Applied Biosystems/Life Technologies also employs emulsion PCR generated clonally amplified DNA-bead populations followed by ligation of adaptor fragments. These are hybridised to a glass slide and undergo sequencing by ligation. Di-base fluorescent probes ligate to the template and are detected via fluorophore emission and are then cleaved. This is repeated for a defined number of cycles (determines read length) and is then repeated at the n-1 position, for a succession of 5 primer ligations. This ensures sequence accuracy by binding of each base twice, allowing DNA sequences from the resulting fluorophore colour space to be determined (Mardis 2008). A single sequencing run generates millions of single or paired end reads. Paired-end reads are sequenced from fragmented DNA/cDNA templates and typically represent two 60 bp fragments of the termini of the template.

1.5.3 Illumina (Solexa) Sequencing

Illumina (previously Solexa) sequencing uses detection of fluorophore labelled nucleotides following incorporation into an extending template, a process termed sequencing via synthesis. Sheared DNA is adaptor ligated and hybridised onto oligonucleotide anchored flow cells and amplified via isothermal bridging followed by anchoring to a second oligonucleotide. Amplified double stranded DNA is denatured to single stranded DNA and extended a second time to form two bridges. Repeated denaturation and amplification creates a dense lawn of amplified templates that are then sequenced via single base polymerase extension using labelled nucleotides. The millions of sequences produced are detected via laser excitation of the fluorophore for a defined number of cycles to standardise read lengths (Bentley *et al.* 2008). Illumina sequencing can also offer paired-end reads where, typically, two 75 bp reads are sequenced from the termini of a 200-500 bp fragmented template.

1.5.4 Single Molecule Real Time (SMRT™) DNA Sequencing

SMRT sequencing is a real-time sequencing by synthesis method that combines high throughput of Illumina and SOLiD with the longer read lengths of Roche 454 via a modified polymerase enzyme. The process uses zero-mode waveguides (ZMWs), essentially holes (~10 nm diameter) that allows visualisation of single polymerase reactions that have been reduced in speed to ~3 nucleotides per second. Phosphate backbone fluorophore labelled nucleotides are sequentially added followed by cleavage of the fluorophore via the polymerase enzyme and fluorophore detection (Eid *et al.* 2009). Maximum read length extends into the thousands of base pairs, with average lengths currently around 1 Kb, a figure likely to increase as the technology develops.

1.5.5 Bench-top machines

In recent years bench-top NGS machines have been developed, aiming to offer routine access at cheaper prices. The IonTorrent platform from Life Technologies detects emission of photons following nucleotide inclusion, shifting the pH of the surrounding solution and resulting in a detectable voltage change. Sequential addition of nucleotides and template elongation produces reads up to 250 bp and in less than 2 hr (Rothberg *et al.* 2011). Development of the size of sequencing chips used on the platform is expected to increase the size of reads produced. The development of the Roche GS Junior System has also reduced the costs associated with NGS technology but at a cost of reduced read length. An estimated 10 fold reduction is also found when compared to the GS FLX platform. MiSeq ($125,000 to buy and ~$500 to run) from Illumina also offers a bench-top solution to NGS sequencing that is able to complete a run in a day and still produces millions of reads.

1.5.6 Comparison of technologies

The 454 platform currently offers the largest read size for the established technologies, improving assembly results and therefore useful for *de novo* sequencing. Though not currently in routine use, the Pacific Biosciences SMRT™ platform offers by far the longest average read length but this benefit is offset via an increased error rate (Metzker 2010). Both SOLiD and Illumina offer the greatest Mb/$ by far, compared to long read technologies as they offer 10 fold cheaper access per Mb of data produced. Though the cost per base pair is the lowest, cost of a single SOLiD or Illumina run is greater than for the 454 platform. At the time of the onset of this study, the 454 platform was locally available

and cheaper than the short-read platforms. The greater number of reads produced from the short read technologies make them ideal for expression or alternative splicing analysis but less reliable for *de novo* assembly, though development of longer read lengths is addressing this issue. Bench-top machines aim to offer the same advantages at a lower cost, though this saving is found at the expense of either read length or total read number. These technologies are particularly advantageous to research on organisms with small genomes.

1.5.7 Comparison of NGS to other high throughput approaches

Access to high throughput sequencing has opened the door to many uses of the technology. Application of NGS has been reviewed previously in the fields of whole genome sequencing (Pareek *et al.* 2011), transcriptome studies (Costa *et al.* 2011), SNP discovery (Nielsen *et al.* 2011), differential gene expression (Auer *et al.* 2011) and clinical diagnosis (Meldrum *et al.* 2011). As the throughput and accuracy of sequencing increases, its benefits continue to far outweigh those offered by alternative expression techniques for large scale projects.

For the past decade microarrays have been utilised for comparisons of gene expression in mosquitoes, dissecting changes in expression caused by numerous factors including blood meal digestion (Sanders *et al.* 2003), mosquito-parasite interactions (Xu *et al.* 2005) and insecticides resistance (David *et al.* 2005). Serial analysis of gene expression (SAGE) has also been utilised to correlate phenotypic changes to molecular causes (Rosinski-Chupin *et al.* 2007). Limitations in these technologies are circumvented through use NGS technology that also bears its own restrictions in high cost and limitation in accuracy of assembly. Microarrays require *a priori* knowledge of gene annotation and previous NGS studies have demonstrated that current gene builds are not exhaustive for mosquito species (Gibbons *et al.* 2009). This means that detection of unannotated genes is impossible using microarray approaches.

Detection of differential gene expression using the Illumina platform has been shown to compare well to using a microarray based approach (Marioni *et al.* 2008). In the study described by Marioni *et al.* (2008), Illuminia sequencing and an Affymetrix microarray (U133 Plus 2) were compared. Following analysis of the same cDNA sample, 81 % of

microarray identified differentially expressed genes were confirmed by Illumina sequencing. In addition, Marioni *et al.* (2008) demonstrated high reproducibility of gene expression levels between technical replicates. Sensitivity of detection of low-level expressed genes has also been shown to be improved through use of Illumina sequencing as opposed to microarrays (Su *et al.* 2011). Therefore NGS technology is more applicable to identification of low level gene expression, such as may be expected of sex determination genes. Sequencing using the 454 platform has also demonstrated high reproducibility of microarray differential expression results, demonstrating that it is a reliable technology for dissecting gene expression (Sugarbaker *et al.* 2008; Torres *et al.* 2008).

In addition, some, but not all microarray platforms contain multiple probes per gene transcript to allow detection of gene isoform abundance (Lapuk *et al.* 2010). A recent study of the *D. melanogaster* transcriptome has suggested that at least 60.7 % of genes have at least one alternative splicing event (Graveley *et al.* 2011). Therefore detection of gene expression using microarrays may miss transcript-isoform-specific expression. As NGS is not restricted by probe design, relative expression levels of transcript isoforms can be detected and compared between datasets (Toung *et al.* 2011). Through comparison of sex-specific NGS sequence data, it may be expected that differentially expressed isoforms (i.e. sex-specific splicing) could be identified.

Table 1.1. Summary of NGS platforms and their associated performance and cost statistics. The NGS platform and associated bench-top technology is shown in brackets. SMRT™ technology is still not routinely used and performance statistics except average read lengths are not yet available.

***average read size**

‡ http://www.nature.com/nrg/journal/v11/n1/fig_tab/nrg2626_T1.html

† http://my454.com/products/gs-flx-system/index.asp

+ http://my454.com/products/gs-junior-system/index.asp

Δ http://www.illumina.com/systems/sequencing.ilmn

•http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing/next-generation-systems.html

ˆ http://molbiol.ru/forums/index.php?act=Attach&type=post&id=112976

Platform	Sequencing chemistry	Total number of reads (x10^6)	Read size	Machine Cost (US $)	Run Time	Data Produce /run(Gb) ‡
454 GS FLX+ (Jnr Flx)	Pyrosequencing	1† (0.1) +	700* (400*)	500,000‡ (125,000)	23 hr† (10 hr) +	0.7† (0.35) +
SOLiD (Ion torrent)	Ligation based sequencing	700ˆ (0.1) ˆ	Fragment – 75• Paired – 2x60• (100)	595,000‡	7-14 days ‡ (2 hr) ˆ	30-50 ‡ (0.1) ˆ
Illumina HiSeq 2000 (miSEQ)	Sequencing via synthesis	300 single and 600 paired end Δ (3.4 single and 6.8 paired end Δ)	75 or 100 Δ (150) Δ	540,000‡ (125,000)	4-9 days‡ (26 hr)	600 Δ (1) Δ
SMRT™	Single molecule sequencing via synthesis	N/A	1000‡	N/A	N/A	N/A

1.6 Study aims and strategies

Current control strategies targeting mosquito populations are not expected to achieve global targets set by the RBM campaign. Therefore additional control strategies are required to reduce the burden currently placed upon insecticides. Recently, a proof of principle RIDL release strategy has demonstrated that genetic control strategies can be effective (Harris *et al.* 2011). Additional research into sex determination and sex-specific expression in *An. gambiae* may lead to identification of genetic elements necessary for development of transgenic strains with characteristics required in population suppression strategies. Previous microarray studies have shown success in identifying sex-biased gene expression in post-embryonic *An. gambiae* (Warr *et al.* 2007; Magnusson *et al.* 2011). Advances in next generation sequencing technology have enabled transcriptome assessment to an unprecedented level and may identify putative sex determination homologues and sex-specific expression.

The overall aims and objectives of this thesis were:

- Investigate the embryonic sex-specific expression profiles of two genes, *doublesex* and *fruitless*, with known involvement in *An. gambiae* sex determination.
- Isolate putative *An. gambiae* homologues of two genes, *transformer* and *transformer2*, with known involvement in insect sex determination, and investigate a putative role in *An. gambiae* sex determination.
- Isolate sufficient mRNA from male and female embryos separately to generate sufficient quantities of cDNA for transcriptome sequencing.
- Generate sex-specific embryonic transcript libraries using the Roche 454 platform to analyse both the embryo and sex-specific embryo transcript content.
- Perform in-depth bioinformatics analysis to characterise the embryonic transcriptome at the nucleotide level for the first time.
- Carry out comparative bioinformatics techniques to isolate sex-specific differences between male and female transcript libraries, including differential gene expression, alternative splicing and sex-specific transcript sequences.

Chapter 2: Materials and Methods

2.1 Mosquitoes

Mosquitoes (*An. gambiae* G3 strain, hereafter *An. gambiae*) were reared at 27±4 °C and 70±20% relative humidity with a 12/12 hr light/dark cycle including one hour gradual dimming to simulate dusk and dawn periods. Adults were reared within Bugdorm cages (MegaView Science Co., Taiwan). Adults were offered a 10 % sucrose-water solution *ad libitum*, and blood fed (Research Whole Blood, National Blood Service, Liverpool) three days prior to egg collection. Females (at least 4 days post emergence) were blood fed approximately 3 ml using a Hemotek artificial membrane system for at least 30 min at ~ 37 °C. Eggs were laid on wet filter papers overnight and transferred to distilled water trays (20 x 20 x 6.2 cm) the following day. Larvae were reared in distilled water and were fed a pinch of Tetramin® fish food once per day, gradually increasing the amount fed as larvae develop. Tetramin® fish food added was sufficient to cover the surface of the water by 3rd instar.

2.2 Embryo Collection

Embryos were collected via the force egg laying of mated females 3 days post blood feeding, similar to the method described by Lobo et al 2006 (Lobo *et al.* 2006). 10-12 females were introduced into an egg laying chamber (shown in Figure 4.4) containing a small amount of water at the base. Females were allowed to lay eggs for 1 hr and embryos were then allowed to develop in insectary conditions for the desired time.

2.3 PCR

PCR reactions were carried out in volumes of 20 µl and contained 1 x ThermoPol Buffer (NEB, Ipswich, MA), 0.8 mM $MgCl_2$, 0.5 mM dNTPs (each), 0.25 µM each primer and two units Taq DNA polymerase (NEB, Ipswich, MA). DNA samples were amplified using the following cycles: initial denaturation 93 °C 3 min, XX cycles of denaturation 93 °C 30 s, annealing XX-XX °C 30 s, extension 72 °C 45 s and a final extension at 72 °C for 5 min. Annealing temperature and cycle number was optimised for each primer pair based upon primer sequence and expected amplification product size respectively.

To amplify from embryonic genomic DNA, individual embryos were crushed using a dissecting needle and residual material used as DNA template. The needle was heat

sterilised after crushing each embryo to remove any remaining material. 50 ng adult male or female genomic DNA was used as positive control. Products were visualized on UV transiluminator following electrophoresis in a 1.5 % agarose gel. Annealing temperature and number of cycles was optimised for each primer pair and is shown in the appendix.

Where required, (RNAi experiments), band intensities were measure using GeneTools Software on a GeneGenius Bio Imaging System (Syngene, Cambridge, UK).

2.4 RNA extraction and DNase treatment

All samples for RNA extraction were homogenised in TRIzol® (Invitrogen, Paisley, UK) using RNAse free plastic pestles and stored at -80 °C prior to extraction.

RNA was extracted using TRIzol® according to the manufacturer's protocol, with small modification. Briefly, samples stored in TRIzol® were incubated at room temperature (RT) for 5 min. A 1:10 volume of 1-bromo-3-chloropropane to the volume of TRIzol® used was added and the samples shaken vigorously for 15 s. Following 10 min RT incubation, samples were centrifuged at 13,000 rpm for 15 min at 4 °C and the supernatant transferred to a clean 1.5 ml Eppendorf tube. An equal volume of isopropanol to initial TRIzol® volume was added to the supernatant and vortexed at moderate speed for 10 s. Samples were incubated at RT for 15 min, centrifuged at 13,000 rpm for 15 min at 4 °C and the supernatant removed. The RNA pellet was washed in 75 % ethanol, centrifuged at 13,000 rpm for 5 min at 4 °C and the supernatant removed. The pellet was briefly air dried and resuspended in 10 µl sterile water. RNA concentration and purity was assessed via NanoDrop™ 1000 spectrophotometer (Thermo Scientific, MA, USA) and then stored at -80 °C. Removal of contaminating DNA from RNA samples was undertaken via Turbo™ DNase treatment of samples. In a 1.5 ml Eppendorf tube, 1X DNase buffer and two units Turbo™ DNase (Ambion, Austin, TX, USA) were added to extracted RNA and incubated for 1 hr at 37 °C. To stop the reaction, 1x DNA inactivation reagent was added to the sample and incubated for 2 min at RT. Following centrifugation for 1 min at 13,000 rpm, supernatant containing RNA was transferred to another 1.5 ml Eppendorf tube and stored at -80 °C.

2.5 DNA extraction following Trizol RNA Extraction

For collection of pooled RNA for sequencing, DNA extraction from TRIzol® was also required to check for contaminating DNA of the opposite sex.

DNA was extracted following TRIzol® RNA extraction as previously described (http://changlab.stanford.edu/protocols.html). Following RNA extraction, the remaining TRIzol® was centrifuged at 14,000 rpm for 5 min at 4 °C and the remaining RNA aqueous phase removed. DNA precipitation was carried out by addition of 0.3 ml 100 % ethanol per 1 ml TRIzol® initially used, mixed by inversion and incubated at RT for 3 min. Samples were centrifuged at 3,500 rpm for 5 min at 4 °C, supernatant removed and the pellet washed with 0.1 M sodium citrate in 10 % ethanol and stored at RT for 30 min with occasional mixing. Samples were centrifuged at 3,500 rpm for 5 min at 4 °C and the previous wash and centrifuge steps repeated. The DNA pellet was washed in 1 ml 75 % ethanol, centrifuged at 3,500 rpm for 5 min at 4 °C, supernatant removed and briefly air dried. The pellet was then resusupended in 50 µl 8 mM NaOH.

2.6 Amplification from RNA template

Two methods were utilised for amplification of RNA templates. One step RT-PCR was used for amplification from sex-specific embryonic RNA due to small quantities of template. Where sufficient template was available, first strand cDNA was first synthesised followed by PCR amplification.

RT-PCR

RT-PCR reactions were carried out upon template RNA (50 ng/µl) using the SuperScript® III One-Step RT-PCR System with Platinum® Taq DNA Polymerase (Invitrogen, NY, USA) according to the manufacturer's guidelines.

First strand cDNA synthesis

cDNA was synthesised from 1 µg total RNA using the SuperScript® III RTS First-Strand cDNA Synthesis Kit (Invitrogen) in 25 µl volume according to the manufacturer's guidelines.

2.7 DNA cloning

Following gel electrophoresis, PCR fragments were gel extracted using QIAGEN Gel Extraction kit, as per manufacturer's guidelines. PCR fragments were ligated into the pGEM®-T Easy Vector System (Promega) in a 3:1 ratio of insert to vector. Ligations were left overnight at 4 °C, ethanol precipitated and resuspended in 3 µl water. 1 µl of resuspended DNA was added to 20 µl 1/10 diluted electrocompetent *E.coli* cells (Invitrogen, Max Efficiency® DH10B™ Competent Cells) and were added to a pre-chilled electoporation cuvette (Sigma 0.1 cm gap width). Cells were electroporated at 1.7 Kv/200 Ω and 0.75 ml LB broth added immediately. Bacteria were transferred to a 15 ml falcon tube and incubated at 37 °C, 230 rpm for 90 min. 50 µl of the bacteria suspension was plated onto LB-ampicillin (100 mg/ml), 0.5 mM IPTG and 80 µg/ml X-Gal plates and were incubated overnight at 37 °C. X-Gal and IPTG were added to plates to allow screening of white (positive) or blue *E. coli* (negative) colonies.

2.8 Sanger Sequencing

White *E. coli* colonies potentially containing inserts were selected and suspended in 20 µl TE buffer and used as template for PCR amplification using sequencing primers (UP/RP). Gel purified PCR fragments were used for sequencing reactions. Two 10 µl PCR reactions were set up for each template with 5x Sequencing Buffer, forward or reverse primer (UP or RP) (5 pmol), 0.5 µl BigDye® (APPLIED BIOSYSTEMS), 5-20 ng template and sterile water to 10 µl. Samples were amplified with the following cycle conditions: initial denaturation 96 °C 2 min, 30 cycles of denaturation 96 °C 15 s, 50 °C 10 s annealing and 60 °C 4 min extension. Amplified sequencing templates were precipitated by addition of 2 µl 750 mM sodium acetate, 125 mM EDTA and 40 µl 100 % ethanol, followed by mixing and incubation at -20 °C for 15 min. Samples were centrifuged at 13,000 rpm for 15 min at 4 °C, the supernatant discarded, dried in a Speed Vac® for 10 min and stored at -20 °C. Pellets were resuspended in 14 µl HiDi-Formimide and loaded onto a 96 well plate. Samples were denatured at 94 °C for 2min and immediately chilled on ice before sequencing via ABI Prism 3100 Genetic Analyzer.

2.9 dsRNA synthesis

T7 tagged gene specific primers were used to amplify dsRNA templates in a 50 µl PCR reaction (Milligan *et al.* 1987). Gel excised PCR fragments following electrophoresis were purified via Gel Extraction Kit (QIAGEN). 1 µg of purified DNA was used as template for transcription using Megascript RNAi kit (Ambion). In brief, ATP, CTP, GTP and UTP (7.5 mM each), 10x reaction buffer and T7 RNA polymerase (4 µl) were added to a final 40 µl reaction volume and incubated at 37 °C overnight. Samples (40 µl) were then DNase and RNase treated via addition of 10x digestion buffer, 2µl DNase I, 2µl RNase and water to a final volume of 50µl and incubation at 37 °C for 1 hr. dsRNA was purified using MEGAclear™ columns (Ambion) and eluted in 100 µl pre-heated elution buffer as per manufacturer's instructions. Double stranded green fluorescent protein (GFP) was used as a negative control with the cloning vector pEGFP-N1 (Clontech; GeneBank Accession number U55762) being used as a template for GFP dsRNA synthesis. Primer pair gfpF/gfpR was used to amplify a 653 bp dsRNA template. Double stranded β-lactamase (dsβ-lac) was also used as a negative control with the cloning vector pGEM®-T Easy (Promega) being used as template for dsβ-lac synthesis. Primer pair B-lacFT7/β-lacRT7 was used to amplify a 464 bp dsRNA template. Use of which dsRNA control is indicated within each results section.

Table 2.1: Primer sequences used for dsRNA synthesis. Underlined sequences indicate T7 region for dsRNA synthesis.

Primer	Sequence
B-lacFT7	TAATACGACTCACTATAGGGAACTTTATCCGCCTCCATCC
B-lacRT7	TAATACGACTCACTATAGGGGCTATGTGGCGCGGTATTAT
GFPF	TAATACGACTCACTATAGGGACGTAAACGGCCACAAGTTC
GFPR	TAATACGACTCACTATAGGGCTTGTACAGCTCGTCCATGCC
6798F	TAATACGACTCACTATAGGGCGGTGGCTCCTCGTCCAG
6798R	TAATACGACTCACTATAGGGACACCTTGCCGTAGTAG
Contig8484	TAATACGACTCACTATAGGGCGGCAGCTCACTAATTCCTC
Contig8484	TAATACGACTCACTATAGGGGGAAGTTGCCGATGTAAAGC

2.10 Microinjection:

Microinjection of embryos was carried out as previously described (Lobo *et al.* 2006). In brief, 10 gravid females were allowed to lay eggs for 20 min and aged until ~30 min old in insectary conditions. Embryos were aligned with posterior ends facing the same direction to a maximum of 50 along the edge of 25 mM NaCl moistened filer paper. Excess moisture was removed from the filter paper and embryos were allowed to dry at room temperature for less than 30 s. Embryos were transferred onto double-sided sticky tape (Scotch 3M, type 415) on a coverslip and covered with 25 mM NaCl. The coverslip was attached to a glass slide and mounted onto the injection stage. Injection needles (Sutter 7.5 mm length 1 mm diameter) were prepared using the Sutter P-2000 micropipette puller (settings: HEAT: 650, FIL: 4, VEL: 40, DEL: 150, PUL: 157) and loaded with 3 µl of dsRNA (1 µg/µl) and injected into the posterior end of embryos (~ 45 – 90 mins old). In some experiments, needles were also loaded with a GFP expression plasmid FUIII* (a kind gift from B. Tresman, LSTM) (300 ng/µl). After injection, coverslips with embryos were submerged in water and kept in insectary conditions until larval emergence. Larvae were then reared as previously described.

* The FUIII plasmid drives expression of the fluorescent reporter GFP from the *An. gambiae* polyubuiquitin (AGAP001971) promoter. Following microinjection of embryos, GFP expression is visible in the posterior fragments of hatched larvae abdomens under a fluorescent microscope.

2.11 Phylogenetic analysis

Sequences of TRA2 homologues and *An. gambiae* candidates were accessed from NCBI (http://www.ncbi.nlm.nih.gov/) using Genbank accession numbers and are indicated within the main text. The protein sequences were aligned using standalone ClustalW (http://www.clustal.org/clustal2/) and a maximum parsimony phylogenetic analysis was performed using standalone MEGA5 software, with banch support values analysed with 1000 bootstrap replciates (http://www.megasoftware.net/megamac.php).

2.12 Clustal W2 Protein Alignments

All protein sequences were accessed from NCBI (http://www.ncbi.nlm.nih.gov/) using GenBank accession numbers and were uploaded to EMBL-EBI ClustalW (http://www.ebi.ac.uk/Tools/msa/clustalw2/) and aligned using default parameters (Larkin *et al.* 2007).

2.13 Online BLAST Searches

All online BLAST seraches were carried out at http://blast.ncbi.nlm.nih.gov/. With specific regard to PHI-BLAST (Pattern Hit Initiated), this search utilises both alignment and matching of a regular expression (a sequence pattern) and therefore limits the number of false positives returned. For a significant hit to be returned, a sequence must show similarity to the query (defined by E-value) and contain the regular expression. Within this study, *M. domestica* TRA (ACY40709) was used as query with the following regular expressions to search the *An. gambiae* genome:

1. [FH]GEG[ST]KPLFXR[DL]D
2. PYFXD[EP]XRE[RQ]DR
3. PQII[TP]IPVPVP

Regular expressions were generated following alignment of TRA homologue protein sequences, shown in Figure 3.4. Regular expressions were then created to represent a consensus sequences within conserved regions. Within the regular expression, amino acid

symbols within square brackets means that either residue can match and X indicates that any residue can match.

2.14 Statistical Analysis

Student's t-test was performed using SPSS16 software (SPSS Inc. Chicago, Illinois, USA) to test for significant difference between experimental and control groups. A difference was deemed significant with a p-value <0.05.

2.15 Assessment of RNA quality via Agilent 2100 Bioanalyzer

Quality of RNA was analysed using the Agilent 2100 Bioanalyzer (Agilent, CA, USA) according to the manufacturer's guidelines (electropherograms for RNA samples are shown in Figure 4.5).

2.16 cDNA Synthesis

cDNA was synthesised using the Clontech (TakaraBio, CA, USA) SMARTer™ (Switching Mechanism at 5' End of RNA Template) cDNA synthesis kit. 3.5 µl of template mRNA was incubated with 1 µl of 3' SMART CDS Primer IIA (12 µM) for 72 °C for 3 min followed by 42 °C for 2 min. To initiate the first strand cDNA synthesis, the following reagents were added to the tube: 5x First strand buffer, 0.25 µl DTT(100 mM), 1 µl dNTP mix (10 mM) 1 µl SMARTer IIA Oligonucleotide (12 µM), 0.25 µl RNase inhibitor, 1 µl SMARTscribe Reverse Transcriptase (100 units/µl). Samples were incubated at 42 °C for 90 min, followed by 72 °C for 10 min in a thermocycler and then diluted by addition of 90 µl TE buffer. The cDNA synthesis process is summarised in Figure 2.1.

2.17 cDNA amplification of sequencing template

Clontech Advantage 2® PCR kit was used to amplify first strand cDNA. A 1/10 dilution of first strand cDNA was mixed with the following in a 100 µl reaction: 74 µl deionised H_2O, 10 µl 10x Advantage 2 PCR buffer, 2µl 50x dNTP mix (10 uM) 2 µl 5' PCR Primer II A (12 uM), 2

µl 50x Advantage 2 Polymerase mix. cDNA was amplified in a thermocycler under the following conditions: Initial denaturation 95 °C 1 min, (user defined) cycles of denaturation 95 °C 15 s, annealing 65 °C 30 s and extension 68 °C 6 min. The number of cycles used was identified empirically as follows: After 15 cycles of amplification, 30 µl of sample was removed from each reaction tube and labelled 'optimisation'. The remaining 70 µl were stored at 4 °C and labelled at 'experimental. 5 µl of the optimisation sample was removed and stored at 4 °C for subsequent gel electrophoresis; the remaining 25 µl was subjected to further PCR amplification. After every three cycles (from 18 to 30), 5 µl was removed and stored at 4 °C. Finally, each 5 µl aliquot was used for gel electrophoresis and following visualisation on a UV transiluminator, the optimum number of cycles was determined. Optimum number was defined as one less than the number resulting in a plateau in the intensity of the cDNA smear. The 15 cycle 70 µl experimental samples were then returned to the thermocycler and subjected to additional cycles to reach the optimum number. Amplified cDNA was purified using the Qiagen PCR purification kit, as per the manufacturer's guidelines.

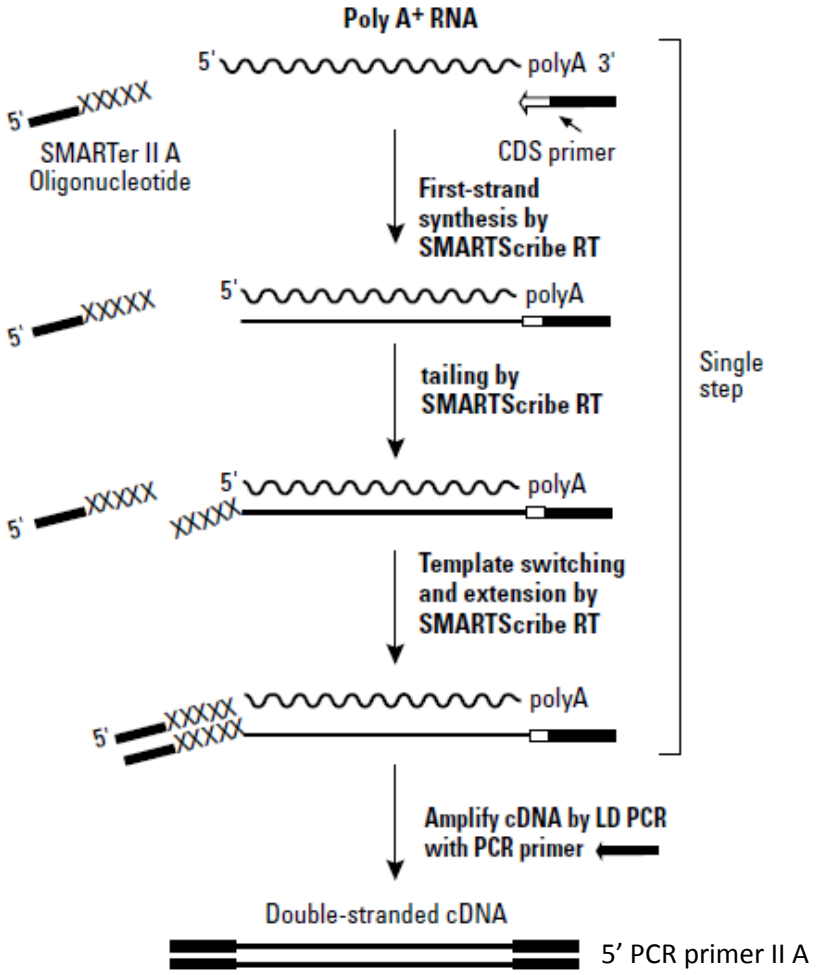


Figure 2.1. Schematic diagram of the Clontech SMARTer™ IIA cDNA synthesis protocol. Poly A (+) enriched RNA is used as template for first strand synthesis by SMARTscribe reverse transcriptase, primed via the 3' CDS primer that binds to the poly(A) tail. The 3' CDS primer contains the SMARTer IIA primer sequence at its 5' end and is therefore incorporated into the 1st strand cDNA. SMARTscribe reverse transcriptase exhibits terminal transferase activity and adds a five nucleotide sequence motif at the 5' end which is recognised by the SMARTer II A oligonucleotide. Template switching by SMARTscribe reverse transcriptase includes the SMARTer IIA sequence in the 1st strand cDNA. The 5' PCR primer IIA recognises the SMARTer IIA sequence at both the 5 and 3' ends of 1st strand cDNA and is used for amplification. Modified image taken from SMARTer™ PCR cDNA synthesis Kit User Manual (PT4097-1 www.clontech.com)

2.18 Pyrosequencing using the 454 platform

Pyrosequencing was performed at the Centre for Genomic Research (CGR), University of Liverpool on the Roche 454 GS FLX Titanium Series platform (454 Life Sciences Corporation, Branford CT, USA). In brief, amplified cDNA was sheared via nebulisation and tagged adaptors ligated to either end of the fragmented cDNA and captured onto beads by limiting dilution. DNA carrying beads were amplified by emulsion PCR, denatured and deposited into wells of half a Picotiter plate for each sample. Sequencing reagents required for pyrophosphate sequencing were then flowed across the plate. Figure 2.2 shows a schematic of the 454 sequencing process.

2.19 Raw read cleaning

Raw reads were cleaned to remove sequences of poor quality (quality score <20), 454 adaptor sequences (A and B), cDNA amplification primer sequences (SMARTer IIA Oligonucleotide, 3' SMARTer CDS primer IIA, 5' PCR primer IIA) and reads less than 30 bp in length using the software Seqclean. All primer and adaptor sequences were supplied using the Seqclean –v option and –A was used to retain poly(A) tails. Reads were then additionally screened for primer and adaptor sequences, as well as mitochondrial derived sequences using Newbler gsMapper. A quality score greater than or equal to 20 indicates there is a base call accuracy of 99 %.

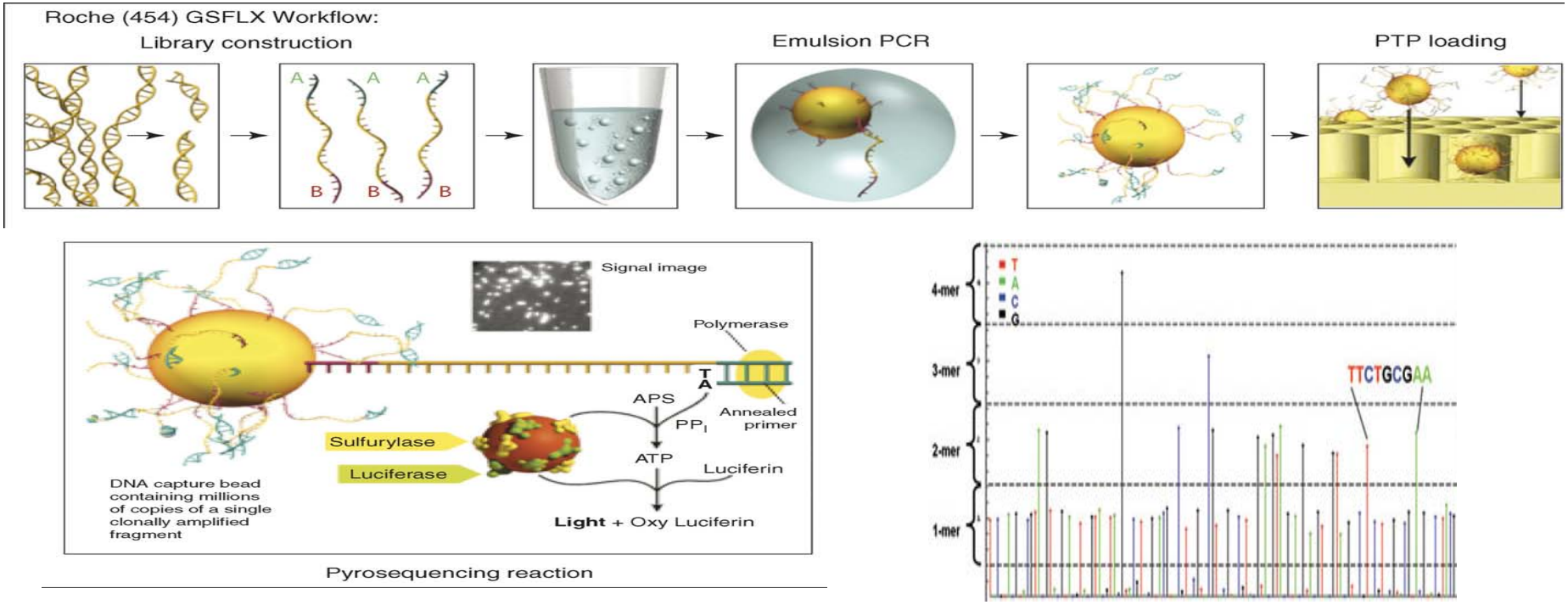


Figure 2.2. Schematic diagram of the 454 sequencing by synthesis protocol. Two adaptors (A and B) are ligated to fragmented template cDNA and bound onto individual beads. Beads are captured in a PCR-reaction-in-oil emulsion and amplified (emulsion PCR), followed by denaturation to ssDNA clones and deposition into wells of a picotiter plate. Reagents necessary for pyrophosphate sequencing are then flowed over the wells followed by sequential addition of nucleotides. Pyrophosphates are generated following nucleotide addition and converted to ATP via sulfurylase. ATP excitation of luciferase generates photons and intensity of the signal indicates the number of bases present. The resulting flowgram (signal intensity of released photons for each nucleotide flow) is then used for sequence analysis. An example of sequence analysis is also shown. Sequential flowing of nucleotides in a defined order ensures each base can be accurately called. The intensity of the peak can then be measured to assess the length of the homopolymer incorporated. Each nucleotide base is represented by a different colour. Figure adapted from (Mardis 2008) and http://454.com/products/technology.asp.

2.20 Mapping of 454 derived reads on to reference sequences

Mapping of reads was performed using the software listed below using default parameters:

Newblerv2.3 gsMapper (Kindly provided by Richard Gregory, CGR)
GMAP/GSNAP (http://research-pub.gene.com/gmap/)

GMAP was used for mapping where a SAM alignment file was required for downstream analysis as Newblerv2.3 does not provide this option.
Databases used for mapping are shown below and were accessed from www.VectorBase.org and abbreviations used in the text are shown in brackets:

agambiae.TRANSCRIPTS-AgamP3.6.fa	**(AgamP3.6)**
agambiae.CHROMOSOMES-PEST.AgamP3.fa (PEST)	**(AgamP3)**

2.21 Integrative Genomics Viewer (IGV) visualisation

SAM alignment files, generated by sequence alignment using GMAP, were loaded into IGV2.0 software (Broad Institute, MA,USA) for visualisation (Robinson *et al.* 2011). The *An. gambiae* GTF (Gene Transfer Format) annotation file was accessed from VectorBase and was set as the default genome, along with the AgamP3 genome FASTA file within IGV. A GTF file contains the genomic location of all annotated transcripts within a transcriptome and an example of a GTF file is shown in Appendix II Table 1. This enabled visualisation of aligned reads to the current *An. gambiae* transcriptome annotation.

2.22 Gene Ontology (GO) classification

BLAST2GO software (http://www.blast2go.com) was used to assign GO terms to sequences (Conesa *et al.* 2005). GO terms are part of a controlled vocabulary to describe the roles of genes within three main independent ontologies: biological processes (biological objective of gene product), cellular compartment (location within the cell where gene product is

active) and molecular function (biochemical activity of gene product) (Ashburner *et al.* 2000).

BLAST2GO relates sequences to functional classes and is capable of handling high-throughput sequence data. In brief, BLAST2GO uses BLAST searches to find similar sequences within either public (e.g. NCBI nr) or custom sequence databases to inputted FASTA sequence files. All GO terms are then retrieved that have previously been assigned to sequences with significant hits (E-value is user defined). Based upon the returned E-values and similarity scores, an evidence code (EC) that indicates the trustworthiness of the GO annotation is given to each sequence to assess the accuracy of the GO term assigned.

In this study, all sequences were used to query the NCBI nr protein database using BLASTX (E-value = $1e^{-3}$) searches. Parameters were set to return the top ten significant hits and default mapping and annotation parameters were used.

2.23 Read Count Analysis

To establish the number of reads (read count) that map to AgamP3.6 transcripts, nucleotide sequences of all 14,974 AgamP3.6 transcripts were accessed from VectorBase (www.VectorBase.org). These were used as a reference to which reads were mapped to using Newbler v2.3 gsMapper default parameters. Read counts were assessed from the 454Refstatus.txt output file and an example of this file is shown in Appendix II Table 4). The file details the number of reads that align to a reference sequence and this represents read count data.

2.24 Contig assembly

Reads were used for assembly via Newblerv2.3 (hereafter Newbler) using default settings and were screened against the *An. gambiae* mitochondrial genome prior to assembly. Unassembled reads were extracted from the Newbler 454Readstatus file and assembled using MIRA default 454 settings (Fig 2.4). Remaining singletons were extracted from the MIRA debris file. Isotigs were also assembled by Newbler and this process is shown in Figure 2.3.

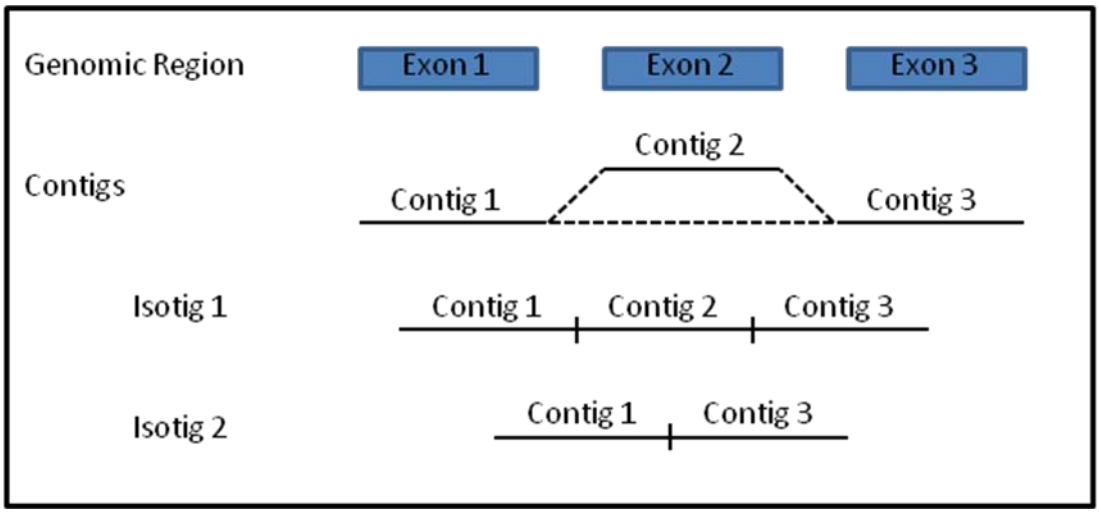


Figure 2.3. Schematic diagram of the isotig assembly process by Newbler. Assembled isotigs represent putative full length transcripts. Contigs represent putative exons, exons that are always co-transcribed with either other exons, untranslated regions or retained introns in primary transcripts. If alternative splice variants are present, reads aligning to a contig may diverge i.e. they will align to a single contig but then 'branch' and be able to align to two further different contigs. Contigs are used to construct isotigs through reads that imply connections between them. Connections between contigs are through 'read spikes' that are composed of greater than ten unique reads. For example, from reads used to assemble contig 1, greater than ten also aligned to the assembled contigs 2 and 3. Therefore Newbler assumes that these reads 'branch' into multiple contigs and these are assumed to be splice variants.

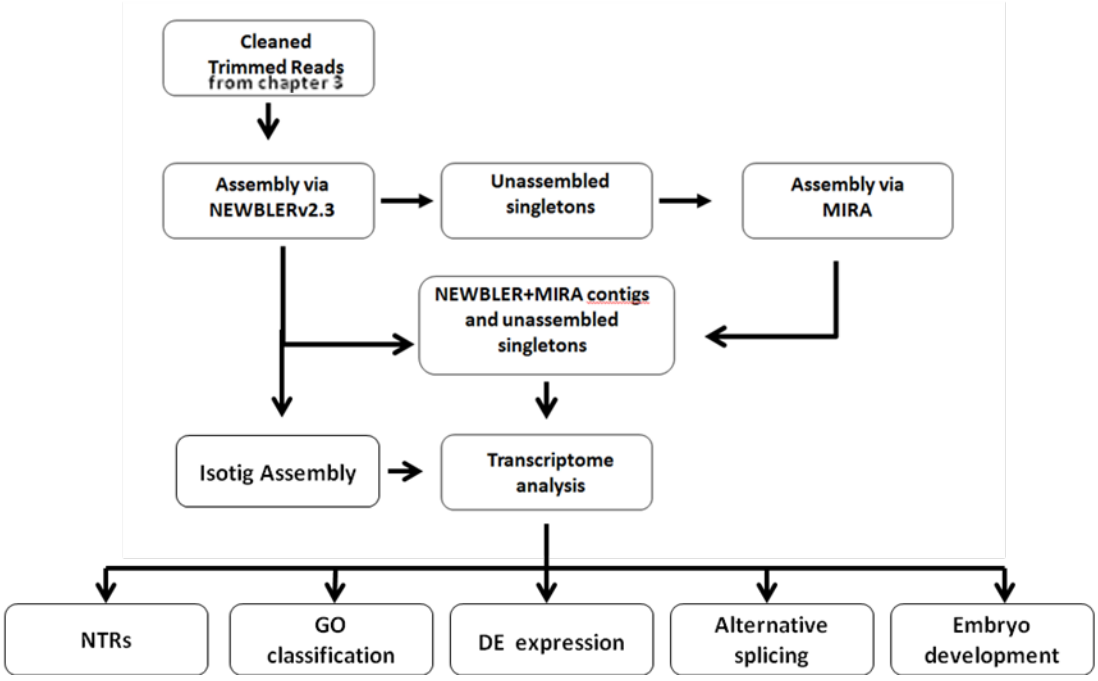


Figure 2.4. Flow chart showing methodology of contig and isotig assembly. The assembly process was done for male and female reads combined (non-sex specific assembly) and for male and female reads individually. Unassembled reads were identified using information from the Newbler Readstatus file and used for MIRA assembly. Remaining singletons were extracted from MIRA output files. Reads not used in assembly are contained within a MIRA debrislist output file and can be subsequently be identified. NTR= new transcribed region. DE = differential expression.

2.25 Standalone BLAST

Standalone BLAST was used from the Wolfson Bio-Linux computer cluster, Life Sciences Building, University of Liverpool. Standalone BLAST databases were created from FASTA files, sourced from VectorBase unless otherwise stated, and are shown below:

agambiae.TRANSCRIPTS-AgamP3.6.fa	(AgamP3.6)
agambiae.CHROMOSOMES-PEST.AgamP3.fa (PEST)	(AgamP3)
agambiae.SCAFFOLDS-M_form.nov09.fa	(MScaffNov09)
agambiae.SCAFFOLDS-S_form.nov09.fa	(SScaffNov09)
agambiae.EST-RAW.dec10.fa	(AgamESTDec10)
aaegypti.PEPTIDES-AaegL1.2.fa	(AaegL1.2)
cquinquefasciatus.PEPTIDES-CpipJ1.2.fa	(CpipJ1.2)
An. darlinig (NCBI PRJNA46227)	(AdarlG)
aaegypti.EST-RAW.dec10.fa	(AaegEST)
cquinquefasciatus.EST-RAW.dec10.fa	(CqEST)
cquinquefasciatus.SUPERCONTIGS-Johannesburg.CpipJ1.fa	(CqSC)
aaegypti.SUPERCONTIGS-Liverpool.AaegL1.fa	(AaSC)

2.26 Calculation of RPKM values

Number of mapped reads was counted via Newbler for all AgamP3.6 transcripts. Length of the transcript was identified via perl script length.pl (appendix II). RPKM values were then calculated using the formula below:

$$\textbf{RPKM} = \frac{\text{Number of mapped reads} \times 1\text{x}10^{3} \times 1\text{x}10^{6}}{\text{Number of total experimental reads}}$$

2.27 Novel splice site detection

To identify novel splice sites, assembled embryo transcripts were compared to a database of AgamP3.6 annotated splice sites. To generate this database, the 14,974 AgamP3.6 gene build transcripts were mapped to the *An. gambiae* genome using the GMAP –f splice site

parameter. This produces a database for the genomic location of each AgamP3.6 splice site. For example:

AGAP004677-RB.exon1/4 chr2L:186860..186859 donor

AGAP004677-RB.exon2/4 chr2L:159367..159366 acceptor

This details the transcript name, the annotated exon, the genomic location of the splice site and whether it is the donor or acceptor site. In total, 104,504 unique splciesites were contained within the database. All embryo transcripts were then mapped to the genome using the GMAP parameters -- splice sites and --novelsplicing to report both known and novel splice sites. Instances of splice sites not represented in the AgamP3.6 database were then extracted.

2.28 Identification of differentially expressed genes: edgeR

edgeR (empirical analysis of digital gene expression in R) is part of the Bioconductor/R statistical package and is used for detection of differentially expressed genes from digital gene expression data. edgeR assumes that the read count data is approximated by a negative-binomial distribution. In brief, the package accepts read count data (number of reads that map to a reference transcript), then normalises the total read library size and finally compares read counts to identify differential expression. As read count data depends upon the expression level of a transcript and also the composition of the RNA population being sampled, library size must be normalised. This normalisation factor is then incorporated into the differential expression analysis.

In addition, edgeR estimates variation between biological and/or technical replicates. Either a common (mean of variation) or tag-wise dispersion value (each gene has an individual variation value) is used to account for differences in read count data. In this study, as no replicates were performed, a dispersion value of zero was assumed. Using edgeR default settings, transcripts with fewer than five mapped reads were removed, as this represents too few reads to reliably detect differential expression.

2.29 Rapid Amplification of cDNA ends (RACE)

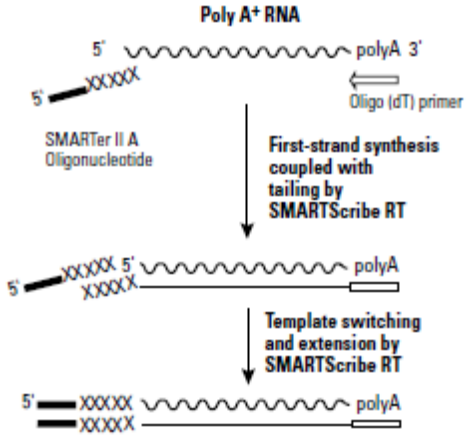
RACE experiments were carried out using Clontech SMARTer cDNA Race Kit (Fig. 2.5). In brief, 1 µg adult male *An. gambiae* RNA was used to synthesise RACE ready cDNA using: 5x First Strand Buffer, DTT(2 mM), dNTPs (1 mM each), template RNA and either 5'-CDS Primer A or 3'-CDS Primer A (1.2 µM). Tubes were incubated for 3 min at 72 °C and then cooled for 2 min at 42 °C. To the 5' cDNA, only SMARTER IIA primer (1.2 µM) was added. To each of the reactions SMARTscribe reverse transcriptase (100 U) and RNase inhibitor (10 U) were added and sterilised water to a 10 µl reaction volume. Samples were incubated for 90 min at 42 °C, followed by 10 min incubation at 70 °C.

1:10 diluted RACE ready cDNA (TE buffer) was used as template using gene specific primers (0.2 µM), Universal Primer Mix (10x), 10X Advantage 2 PCR buffer, dNTPs (0.2 mM), Advantage 2 Polymerase mix (50x) and sterile water to a 50 µl colume. PCR conditions were: 35 cycles of denaturation 94 °C 30 s, annealing 60 °C 30 s, and extension 72 °C 3 min PCR products were visualised via gel electrophoresis. Subsequent nested PCR using nested gene specific primers (GSPs) (0.1 µM) and Nester Universal Primer Mix (10x) was carried out using the same above conditions and 5 µl of 1:50 diluted PCR product as DNA template.

Table 2.2: Primer sequences used 6 for RACE experiments. Sequences and names were taken from the Clontech SMARTer RACE kit manual.

Primer	Sequence
SMARTer II A Oligonucleotide	AAGCAGTGGTATCAACGCAGAGTACXXXXX
3' RACE CDS Primer A	AAGCAGTGGTATCAACGCAGAGTACT(30)N-1N
5' RACE CDS Primer A	$(T)_{25}$ V N
Universal Primer A Mix	CTAATACGACTCACTATAGGGCAAGCAGTGGTATC AACGCAGAGT (long) CTAATACGACTCACTATAGGGC (short)
Nested Universal Primer	AAGCAGTGGTATCAACGCAGAGT

A



B

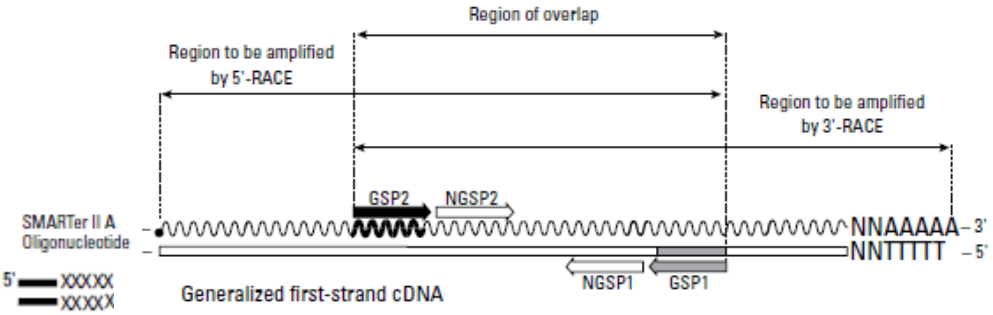


Figure 2.5. Schematic overview of Clontech SMARTer RACE kit protocol.(A) Enrichment for full length 1st strand cDNA sequences. Poly(A)+ RNA contained within total RNA sample is bound by an oligo(dT) primer that binds to the poly(A) tail for first strand cDNA synthesis. The SMARTscribe reverse transcriptase (RT) exhibits terminal transferase activity, adding a three to five nucleotide motif at the end of the 1st strand cDNA. The SMARTer IIA oligonucleotide recognises and binds this motif and template switching by SMARTscribe RT incorporates the SMARTer IIA primer sequence into the 1st strand cDNA. (B) Schematic representation of 5' and 3' RACE experiments. For the 3' reaction, 1st strand cDNA was synthesised using an oligo(dT) primer (3'-RACE CDS primer A) that binds to the poly(A) tail of mRNA transcripts. The 3'-RACE CDS primer A also contains the SMARTER IIA primer sequence at its 5' end, thereby incorporating this sequence in the first strand cDNA. For the 5' reaction, 1st strand cDNA was synthesised using an oligo(dT) primer (5' RACE CDS primer A) that also binds to the poly(A) tail of mRNAs. For 3' RACE, use of a oligo(dT) primer that binds to the 3' poly(A) tail in combination with a gene specific (GSP2) primer amplifies the unknown 3' region. Nested reverse (NGSP1) and forward (NGSP2) can be used for nested PCR amplification of 5' and 3' PCR products respectively. For the 5' RACE reaction, a second primer (SMARTER IIA) binds to the motif added by SMARTscribe RT and template switching by SMARTscribe RT incorporates the SMARTER IIA primer sequence into the end of the 1st strand cDNA. As SMARTscribe RT only adds the motif at the end of the mRNA template, the process enriches for full length transcripts Using gene specific primers in combination with a universal primer mix (UPM), that recognises and binds the SMARTER IIA primer sequence, it is possible to amplify cDNA ends. For 5' RACE, the SMARTer IIA oligonucleotide recognises and binds the sequence added by the

SMARTscribe RT. PCR amplification in combination with a gene specific (GSP1) reverse primer amplifies the unknown 5’ region.

2.30 Open Reading Frame identification

Open reading frames (ORFs) were searched for using the online EMBOSS explorer 'getorf' software using default parameters.

(http://150.185.138.86/cgi-bin/emboss/getorf)

2.31 Cufflinks transcript assembly

Previously published transcriptomic data generated by Solexa (Illumina) sequencing of mixed sex *An. gambiae* adults was downloaded from the NCBI short read archive (SRX014416). Individual reads with sequence similarity to a Y chromosome *An. gambiae* scaffold (NW_045557), identified through BLASTN searches, were mapped to the genome using GMAP default parameters. SAM (sequence alignment/map format) files were then used for a reference based assembly using Cufflinks, following uploading to the Galaxy online cloud (http://main.g2.bx.psu.edu/). This was done to establish if short-read deep sequencing (Solexa) could extend gene sequences identified on the Y chromosome.

2.32 Six-frame translations of nucleotide sequences

Sex-frame translations were performed using the BMC Search Launcher, Baylor College of Medicine using default parameters. http://searchlauncher.bcm.tmc.edu/seq-util/Options/sixframe.html

2.33 Assessment of transcript coding potential

The coding potential of transcript sequences was assessed using the online version of the Coding Potential Calculator (CPS) (http://cpc.cbi.pku.edu.cn/) developed by Kong *et al.* (2007). CPS assesses transcript coding potential through various criteria including matches to protein domain databases and assessment of putative ORFs. Sequences are then classified as coding, weak-coding, weak non-coding or non-coding based upon these criteria.

2.34 Primer Sequences

Table 2.3: Primers and sequences used within chapter 3. Shown are the names of primers and sequences used for amplification of AGAP006798, *doublesex* and *fruitless*.

Primer	Sequence
6798NR	GTGACCGGAATAGCTGTCGT
dsxF	AAAGCACACCAGCGGATCGA
dsxR1	GGACGAGAACATCTCGGTG
dsxR2	CGCAATACCACCCGTCAGAGTGGA
fruFemJxF1	CTACGCAGGCAGGAACAG
fruFemJxR	TAGATGATCGGATGGGGATG
fruFemJxF2	CTACACCTCTGCTGCTGTCG
fruFemF	GTATGCCGACGAGATGTCTG
fruFemR	ATGTGACAGGTGGGCGTCA
fruMalF	ACGCGGACGATATGAACAAT
fruMalR	TTCTGCTGCTGTTTCTCTGC
fruF3	CGCCTGTGACACCAGACCA
fruR3	GGTGAGGAAGCCATGTTGAA
fruR2	GGGCAGCGTTACACCACTAC

Table 2.4: Primer and Adaptor Sequences used within chapter 4. Primer sequences used for embryo separation and the preparation of cDNA for 454 sequencing.

Primer	Sequence
124678F (Sexing primer F)	TTTGAGCATGTGTTTAAAGG
124678R (Sexing primer R)	AGGTTTTGCCGACTACAAT
MagNF	GTACGATGTATTCTGCAACA
MagNR	CAAGGTTTTGTCAAGGTAAT
SMARTer II A Oligonucleotide	AAGCAGTGGTATCAACGCAGAGTACXXXXX
3’ SMART CDS Primer II A	AAGCAGTGGTATCAACGCAGAGTACT(30)N-1N
5’ PCR Primer II A	AAGCAGTGGTATCAACGCAGAGTACGCGGG
454 Adaptor A	CTGAGACAGGGAGGGAACAGATGGGACACGCAGGGATGAGATGG
454 Adaptor B	CTGAGACACGCAACAGGGGATAGGCAAGGCACACAGGGGATAGG

Table 2.5: Primers and sequences used within chapter 5. Primers used for confirmation of NTR contig assembly and alternative splicing candidate AGAP003139.

Contig Name	Forward Primer Sequence	Reverse Prime Sequence
contig06638	AGCGATGTCCACTGCTTCTT	CAGGAAAACCGAACCGAATA
contig08915	CGAACAGAAAGCGAAATGTG	CGTTACGCCTGATGATTGTG
contig09569	GGTCGAAGTTAACGGAAACG	GTGTGCCCTATTTGCTGGAC
contig12387	AGCCGAAGAGTGGCAACAT	CTCACGAGTTGCACGATCAG
contig05828	CATCATAACTACGCGGCAAC	CCCTCCATTTGCTTTGCTTA
singletons_rep_c6734*	CCTTGCTCTTGGTTCTTGCT	GGATCGTTCAATCAGGCATC
singletons_c5769*	CGGAGTCAACGAGTGGAAGT	GATAGCCTTTTGCAGCGACT
contig07915	TTCCCACACCACATTCACAC	TCAACTTGTAGTCGCGTTCG
contig08484	CGGCAGCTCACTAATTCCTC	GGAAGTTGCCGATGTAAAGC
contig13216	ACAGTGAGCGATTGTGTTCG	ATGGAAGTTTGCAACGAACC
AGAP003139F1	GACAGCAAGACGACCGTGTA	N/A
AGAP003139F2	AATCCCGACAATCAGGTGTC	N/A
AGAP003139R1	N/A	CCTCGATCTTGGACTTTTGC

Table 2.6: Primer and sequences used within Chapter 6. All primer sequences below were used for the amplification and characterisation of YOA and YOB genes described in chapter 6.

Name	Transcript (length bp)	RNA size	DNA size	Forward	Reverse
YOA	A (493)	350	3,264	ACCACCGAGTAACCAATCGTGT (F1)	GCAGCATAAACAAGCACAGC
	B (544)	401	3,264	ACCACCGAGTAACCAATCGTGT (F1)	GCAGCATAAACAAGCACAGC
	C (330)	326	3,264	GGGGTTATACCAAGGACCGTAAAG (F2)	GCAGCATAAACAAGCACAGC
YOB	A (681)	464	635	GCCTGATTGTTGTCCGTTCT	TTGTATGTGTCGTACGTTTGTG
	B (593)	385	635	GCCTGATTGTTGTCCGTTCT	TTGTATGTGTCGTACGTTTGTG
	C (843)	547	635	GCCTGATTGTTGTCCGTTCT	TTGTATGTGTCGTACGTTTGTG
	D (755)	635	635	GCCTGATTGTTGTCCGTTCT	TTGTATGTGTCGTACGTTTGTG
08484	N/A	413	575	CGGCAGCTCACTAATTCCTC	GGAAGTTGCCGATGTAAAGC
08484nested	N/A	175	337	GCCTGATTGTTGTCCGTTCT	GCAGCATAAACAAGCACAGC
YOA-A-RACE (Nested)	N/A	N/A	N/A	ACCACCGAGTAACCAATCGTGT (GATATTCGAGTTACGCGGATTCCT)	TACTCGGTTTCTGGGTTCCCTA (CTCGAATATCCGCGTAAGTCAAAT)
YOA-C-RACE (Nested)	N/A	N/A	N/A	GGGGTTATACCAAGGACCGTA (TATAAGCCCGTTTTGACAGCTAGG)	CGATTCTTCTTCCACCTAGCTGTC (ACGGGCTTATAACTTGGCCTGATA)
YOB-A/C (Nested)	N/A	N/A	N/A	N/A	AGAGGGAAAAGGAACTGAGGCAAC (TATGTCTATACAAGGGCGGTGGTT)
YOB-all (Nested)	N/A	N/A	N/A	N/A	GAACTGGAAGTTGCCGATGTAAAG (GAACTGGAAGTTGCCGATGTAAAG)
YOB-D/C (Nested)	N/A	N/A	N/A	N/A	CCGCTGTTGTCGAGAAAGTAATTC (CTGCAGAACGGACAACAATCAG)
YOB-all	N/A	N/A	N/A	CACAAACGTACGACACATACAAAAC	N/A
128125B	N/A	N/A	547	AGCGTGGAGGACATACAA	ATGGCAATTTCGTTTTCA

Table 2.7: Primer sequences used to examine candidate sex-specific AgamP3.6 transcripts. Shown in the tale are primer sequences used to amplify candidate sex-specific transcripts by RT-PCR.

Candidate AgamP3.6 Transcript	Forward Primer	Reverse Primer
Male		
AGAP005551-RA	CTGATCTTCGTGGTGCTGAA	ACCGAGTCAGGGTCAACGTA
AGAP000209-RA	ACCGGGTGTACGACGATTAC	CGCGAGTGTTCGCTGTATAA
AGAP001412-RA	GGCTTCGACTGGGTGTTCTA	TCCATGAACATGAGCACCTC
AGAP005398-RA	CCGTCTACCACGAACAGTTG	CTCACCTTGCGTGCCTATCT
AGAP012308-RA	GCCAAGACACTCACCTAGCC	CATCGCAGCCAAACTTCTTG
AGAP012920-RA	CCGTTTCCGAGTATTTGTGG	AGAAGCGCAAGATCAAGCTG
AGAP001591-RA	CGCGGCTACTTTGACTTTTC	ATGTACCTCCAGCACGGAGT
AGAP007474-RA	AGCTGTGCGTGGAGATGAC	GATCGTAGCTCCTGCGAGAT
AGAP008132-RA	TGTTTCTGCTGCAAATCTCG	GGCTGTCATGGTTGTTGTTG
AGAP005234-RB	GGTGTTCATCGACATCAACG	CGTGTCCGAGTAGGACGTTT
AGAP008684-RA	AGCGGGAGAAGTTCCTCAAC	GTGATCCTTCAGCACCCAAT
AGAP009475-RA	GACGTTTTACACTCGCAAGG	CGAGTTATGCTATCCGCACA
AGAP012308-RA	TACATCAATGACGGCGTGTT	TTAAAGCTGGTCGCTGGATT
AGAP009481-RA	TAAACCGCACGTTTGTGAAA	ACACACATCAGCCTTTCGTG
AGAP003577-RA	TCTCACGAAGCGTTACATCG	CCTCGTAGAACTGGCAGGAG
Female		
AGAP005260-RA	CGGTGGCTCCTCGTCCAG	GTGACCGGAATAGCTGTCGT
AGAP011790-RB	ATCGTTTTGGTTGCGCTAGT	GCATCCGAGGTGTTGGTAGT
AGAP000213-RA	CCACCTGAACTCGGTCATCT	CCAGCTTTTCCAGCAGACTC
AGAP003334-RA	CGTTGTTGCCTGCTTCCT	GCAGAGTGGACGACGGTCT
AGAP003420-RA	ACAAGGCGAAGCAGTACGAT	CTGGATCAGGAAGGGCAGTA
AGAP000065-RA	GGCGATAAGCTGGAAGAGG	GAGAAGAACCTATCGCCCAA
AGAP011948-RA	GCTATCGAACGGTGGACGAA	CGCAAACAATCCTCGATGTA
AGAP001563-RA	GGCAAGGTGAAGGATAGCAG	GTGGTCGGAAGGAATGTTTC
AGAP008018-RA	GACTCACGGCTTGGTGTTCT	GCACGAAGTTTCTCCTGCTT
AGAP007646-RA	CGCAGAAGACCTCACAATCA	TTCCAGCACCGAGGACAT
AGAP009731-RA	TTCCGAGCTGTTTCTGGACT	TCGCGGATCAGATTATAGCC
AGAP006913-RA	GAACTATGGGCTGGGTGAGA	CAAGACCACAGAACCGATCC
AGAP000693-RA	GCTGAAGAAGCTGGGAAAGA	ATGTTAGCAGAGCCGTCGTC
AGAP008288-RA	AAAGCAAGCGATGAAGGAGA	TCCGAGAGTAATTGCGTGTG
AGAP006479-RA	AGCACGAACGAGGAGGACTA	GCAGCTTCTCGTTCTCAACC
AGAP010708-RA	AAACATTGTGCCTGCTGTTG	GGCTCACCGGGATTAAAGTT
AGAP009549-RA	AACGACCGTACGAAGCTGTT	CAGCGAGAACTGTTCGTTGA
AGAP010036-RA	ATGTGGTCGCTAGGCGTTAT	AGCGAACGAATAGTGCATCC
AGAP011604-RA	GTAGTGGCCGTGACAATCCT	CCTGCTTCTCCTTGTCGAAC
AGAP012221-RA	CAGCAGCAGCAGGGAGTG	CATCTCGACTTACCCAAATCC

Table 2.8: Primer sequences used to validate candidate sex-specific new transcribed region transcripts. Shown in the table are primer sequences usedto amplify candidate sex-specific NTRs using RT-PCR.

Name	Forward Primer	Reverse Primer
contig01125	TTGTGGTTCTTCTGCACCAA	GCTTAACAACCGCTGCTTCT
contig03312	GTCCGTGCTGAACAGTTGC	TACACACTACCGCCGTTCAA
contig04026	TCGAGCAGTGAAACCGTG	CTCCGGAACAGATGGTGTCT
contig06155	CGAAGTAGCAACGACGCACT	TCTCACTCACGCTCGAAACA
contig08484	CGGCAGCTCACTAATTCCTC	GGAAGTTGCCGATGTAAAGC
contig09711	TCACTGCTCGATACCTGTGC	CGAATCAGTGAGTGACGCAT

Chapter 3: Sex determination in *An. gambiae* embryos

3.1 Introduction:

Control of sexual dimorphism in insects shows a remarkable degree of similarity and central to all systems is sex-specific splicing of key genes. The molecular control of sex determination has been known since loci involved in *Drosophila melanogaster* were identified, reviewed by Baker *et al.* (1980). At the summit of the *D. melanogaster* sex determination hierarchy, Sex lethal (SXL) is active in females only, and causes female-specific splicing of *transformer* (*tra*) (Nagoshi *et al.* 1988; Erickson and Quintero 2007). In males, non-active SXL results in default male-splicing of *tra* (Sosnowski *et al.* 1989). Female-specific TRA, along with the constitutively expressed non-sex-specific *transformer2* (*tra2*) causes female specific splicing of two downstream transcription factors, *doublesex* (*dsx*) and *fruitless* (*fru*) (Hedley and Maniatis 1991; Hoshijima *et al.* 1991; Heinrichs *et al.* 1998). In males, *dsx* and *fru* follow a default male-specific splicing pattern. *dsx* and *fru* transcription factors then activate downstream targets that control sexual dimorphism (Gailey *et al.* 1991; Keisman *et al.* 2001).

To date, three key homologues have been identified in *An. gambiae*; *sxl*, *dsx* and *fru* (Scali *et al.* 2005; Gailey *et al.* 2006; Traut *et al.* 2006). *Sxl* is not sex-specifically expressed and is therefore presumed non-functional in *An. gambiae* sex determination (Traut *et al.* 2006). *dsx* and *fru* transcripts are sex-specifically spliced and function in the *An. gambiae* sex determination pathway (Scali *et al.* 2005; Gailey *et al.* 2006). No homologue of *tra* or *tra2* has been identified for mosquito species, though identification of TRA/TRA2 binding sites in *fru* and *dsx* suggest that they may be present (Scali *et al.* 2005; Gailey *et al.* 2006). Within our study, we focus specifically upon the embryonic stage of *An. gambiae* sex determination.

Initiation of insect sex determination systems ultimately stems from chromosomal composition of the individual. For example, in *D. melanogaster*, the number of X chromosome linked signal elements (XSEs), in *Ceratitis capitata* the presence of a Y chromosome linked male determining factor and in *Apis mellifera,* a haplodiploid system ultimately controls the sexual fate of each individual (Willhoeft and Franz 1996; Erickson and Quintero 2007; Heimpel and de Boer 2008). Given the varying mechanisms controlling sex determination between species, the diversity of sex determination pathways has been of great interest in evolutionary and developmental biology. Additionally, dissection of sex determination pathways also presents an opportunity to utilise sex specific differences,

such as alternative splicing, for novel transgenic control strategies, as demonstrated using alternate forms of *C. capitata tra* (Fu *et al.* 2007).

A threshold effect of XSEs in *D. melanogaster* ensures sex-specific splicing of sxl at the summit of the sex determination cascade (Sanchez and Nothiger 1982; Erickson and Quintero 2007). Though homologues have been identified in other dipterans (Meise *et al.* 1998; Saccone *et al.* 1998; Lagos *et al.* 2005; Traut *et al.* 2006) and lepidopterans (Niimi *et al.* 2006; Traut *et al.* 2006), *sxl* is not sex-specifically spliced within any other species examined and is presumed only functional in *Drosophila* sex determination to date. In a number of species, a threshold effect of the gene *tra,* not *sxl,* is responsible for determining sexual fate as reviewed by Verhulst *et al* (2010).

tra homologues show limited sequence conservation but high functional conservation and have been identified in *C. capitata* (Pane *et al.* 2002), *Bactrocera oleae* (Lagos *et al.* 2007), *Lucilia cuprina* (Concha and Scott 2009), *Musca domestica* (Pane *et al.* 2002; Hediger *et al.* 2010) and *Anastrepha* species (Ruiz *et al.* 2007). Other than dipterans, *tra* homologues have been identified in the hymenopterans *A. mellifera* and *Nasonia vitripennis* (Hasselmann *et al.* 2008; Verhulst *et al.* 2010). Though sharing limited sequence similarity, successful approaches to identify *tra* homologues have included degenerate primers (Ruiz *et al.* 2007; Concha and Scott 2009; Hediger *et al.* 2010) or bioinformatics approaches using BLAST searches directed at conserved domains (Verhulst *et al.* 2010). Also, *D. melanogaster tra* shares 3'-unstranslated regions with the ring finger protein l(3)73Ah and conserved synteny to *l(3)73Ah* has also been used to isolate *tra* homologues (Irmingerfinger and Nothinger 1995; Pane *et al.* 2002; Lagos *et al.* 2007).

tra2 has been identified in dipterans including *M. domestica* (Burghardt *et al.* 2005), *C. capitata* (Gomulski *et al.* 2008; Salvemini *et al.* 2009), *L. cuprina* (Concha and Scott 2009), *Anastrepha* species (Sarno *et al.* 2009) and *Sciara* species (Martin *et al.* 2011). A *tra2* homologue has also been identified in the lepidopteran *Bombyx mori* (Niu *et al.* 2005). All *tra2* homologues share a conserved RNA recognition motif (RRM) and are responsible, in combination with *tra*, for the sex-specific splicing of the downstream transcription factors *dsx* and *fru*. Expression patterns of *tra2* transcripts are not conserved between different insects. In *D. melanogaster*, three alternatively spliced transcripts function in both the somatic tissues and male germline (Amrein *et al.* 1990; Mattox *et al.* 1990; McGuffin *et al.* 1998). In other dipterans, only a single transcript is present in *C. capitata*, *M. domestica, Sciara ocellaris* and *L. cuprina* (Burghardt *et al.* 2005; Concha and Scott 2009; Salvemini *et*

al. 2009; Martin *et al.* 2011). *tra2* transcripts are present in the embryonic stages of *D. melanogaster*, *C. capitata*, *M. domestica* and *S. ocellaris* (Amrein *et al.* 1988; Burghardt *et al.* 2005; Salvemini *et al.* 2009; Martin *et al.* 2011). Interestingly, a maternal contribution of *tra2* has been identified within *C. capitata* (Gabrieli *et al.* 2010).

In *D. melanogaster* binding of TRA and TRA2 to *fru* transcripts in females activates a 5' splice site causing inclusion of an additional exon containing premature stop codons, producing a non-functional protein (Heinrichs *et al.* 1998; Lee *et al.* 2000). Ectopic expression of *An. gambiae fru* is able to rescue *fru*-deficient *D. melanogaster* mutants and *An. gambiae fru* transcripts are sex-specifically spliced as early as adulthood (Gailey *et al.* 2006). Non-sex-specific isoforms of *fru* are also expressed in *D. melanogaster* and are essential for adult and embryonic neuronal development (Anand *et al.* 2001; Song *et al.* 2002). *fru* homologues have also been identified in *Blattella germanica* (Blattaria: Blattellidae) (Clynen *et al.* 2011), *Schistocerca gregaria* (*Orthoptera: Acrididae*) (Boerjan *et al.* 2011) the dipteran *C. capitata* (Salvemini *et al.* 2009) and the lepidopteran *Nasonia vitripennis* (Bertossa *et al.* 2009).

dsx is a sex-specifically spliced transcription factor responsible for sexual dimorphism and a homologue has been identified in *Aedes aegypti* as well as *An. gambiae* (Waterbury *et al.* 1999; Scali *et al.* 2005; Salvemini *et al.* 2011). Conservation of sex-specific *dsx* splicing extends to other dipterans including *M. domestica* (Hediger *et al.* 2004), *L. cuprina* (Concha and Scott 2009), *Bactrocera* species (Shearman and Frommer 1998; Lagos *et al.* 2005), *C. capitata* (Saccone *et al.* 2008) and *Megaselia scalaris (Kuhn et al. 2000).* Homologues have also been identified within the lepidopteran *B. mori* (Ohbayashi *et al.* 2001) and hymenopterans *A. mellifera* and *N. vitripennis* (Cho *et al.* 2007; Oliveira *et al.* 2009).

Sex-specific splicing of *An. gambiae dsx* is conserved but differs to the *D. melanogaster* mode of splicing. Binding of *tra* and *tra2* to *Drosophila dsx* activates a female specific weak 3' splice site by recruitment of general splicing factors, leading to inclusion of a female specific exon (Hoshijima *et al.* 1991; Inoue *et al.* 1992; Heinrichs and Baker 1997). *An. gambiae* sex-specific splicing occurs due to female specific exon inclusion by activation of a downstream 5' splice site and a male specific exon is transcribed in females as part of the UTR (Scali *et al.* 2005). *dsx* is sex-specifically spliced as early as 1st instar larvae and continues through to adulthood (Magnusson *et al.* 2011). In *D. melanogaster*, sex-specific splicing is known to extend to the embryonic stages, restricted to the somatic gonad precursors (DeFalco *et al.* 2003; Hempel and Oliver 2007). Sex-specific forms of dsx have

also been detected in the embryonic stages of *C. capitata* and *Ae. aegypti* (Gabrieli *et al.* 2010; Salvemini *et al.* 2011).

To study embryonic expression of sex-specific gene forms, male and female insect embryos must be separated using a molecular method as they cannot be sorted using morphological approaches. Previously, repetitive markers present on the male Y chromosome of *C. capitata* have been amplified to separate males and females (Gabrieli *et al.* 2010). Such an approach in *An. gambiae* is only now available following the identification of repetitive fragments on the sex chromosomes (Krzywinski *et al.* 2004; Krzywinski *et al.* 2005).

To examine gene function *in vivo* RNAi studies have been commonplace in laboratories for more than a decade (Fire *et al.* 1998). Previous studies have assessed *tra2* (Burghardt *et al.* 2005; Salvemini *et al.* 2009; Schetelig MF 2011) and *tra* gene function (Pane *et al.* 2002; Lagos *et al.* 2007; Verhulst *et al.* 2010) by using embryo or pupae microinjection of dsRNA. These resulted in sex-reversal phenotype and demonstrated a role for the targeted genes in sex determination. Recently, embryo microinjection of dsRNA has been performed in *An. gambiae* and *Ae. aegypti*, allowing functional characterisation of target genes (Clemons *et al.* 2010; Thailayil *et al.* 2011). Though the genes targeted are not involved in sex determination, the mentioned studies validate the presence of an active *An. gambiae* embryonic RNAi pathway.

3.1.1 Specific Chapter Aims

- Isolate putative *An. gambia*e TRA2 homologues through BLAST approaches and investigate a sex determining role for candidates through RNAi induced reductions in transcript levels.
- Isolate putative *An. gambiae* TRA homologues through comparative BLAST searches and assess evidence supporting classification as putative TRA homologues.
- Investigate the embryonic expression profiles of two genes, *fruitless* and *doublesex*, with known involvement in *An. gambiae* sex-determination.

3.2 Results

3.2.1 Identification of putative transformer 2 homologue

To identify a putative *An. gambiae* homologue of *tra2*, the protein sequence of *D. melanogaster* TRA2 isoform A (NP_476764.1) was used for BLASTP (E-value cutoff $1e^{-10}$) searches against the *An. gambiae* protein database. The top hit (E-value = $6e^{-30}$) corresponded to AGAP006798; alignment of that protein sequence to known TRA2 protein sequences is shown in Figure 3.1. Other best matches included AGAP013051 (E-value = $8e^{-20}$), AGAP008433 (E-value = $5e^{-12}$), AGAP006755 (E-value = $2e^{-11}$), AGAP000977 (E-value = $3e^{-11}$) and AGAP009810 (E-value = $6e^{-11}$). Each additional significant hit was aligned to *D. melanogaster* TRA2 and showed limited sequence similarity (alignments shown in Appendix II Figure 1). All previously identified TRA2 homologues share conserved RNA recognition motif (RRM) and serine/arginine (SR) rich regions shown by yellow and grey boxes respectively in Figure 3.1. Sequence similarity is also present within a linker region, in between the RRM and second RS rich region, indicated by the green box in Figure 3.1. AGAP006798 shows sequence similarity within the RRM and linker regions of known TRA2 proteins. Two ribonucleoprotein identifier sequences, found in all RRM domains, also show sequence similarity between AGAP006798 and TRA2 proteins sequences. Though similarity is reduced within the RS domains, AGAP006798 is abundant in serine and arginine within the same region.

TBLASTN searches also identified an *An. gambiae* EST (BM620287) with similarity to *D. melanogaster* TRA2 isoform A (E-value = $1e^{-28}$). ClustalW 2 alignment of BM620287 to *D. melanogaster* TRA2 is shown in the Appendix II (Fig. 1F). BLASTN searches using BM620287 as query showed that the EST is located in an intergenic region of chromosome3R and contains three putative exons. Amplification of this EST by PCR failed and the candidate was not investigated further. As the current AgamP3.6 transcriptome is unlikely to be complete, TBLASTN (E-value $1e^{-3}$) searches were conducted against the *An. gambiae* genome using *D. melanogaster* TRA2 as query. Sequence similarity was observed within an intergenic region of chromosome3R (E-value $1e^{-25}$) and two *An. gambiae* contig sequences AAAB01064523.1 (E-value = $2e^{-5}$) and AAAB01043769.1 (E-value = $8e^{-5}$).

To investigate the phylogenetic relationship of An. gambiae candidates to insect TRA2 homologues, a phylogenetic tree using the maximum parsimony algorithm was constructed for TRA2 homologue sequences and *An. gambiae* candidates, using Tribolium casteneum TRA2 as an outlier (Fig. 3.2). As expected the dipteran TRA2 homologues (*D. melanogaster*,

L. cuprina, *C. capitata* and *M. domestica*) are clustered together, whereas the lepidopteran *B. mori* is more closely related to the outlier *T. casteneum* (Coleopera). Of the *An. gambiae* candidates identified, AGAP006798 appears to show the highest similarity to the cluster of dipteran TRA2 sequences, with the remaining candidates clustered together, including the identified EST (BM620287). A comparison of the exon/intron structures of insect TRA2 homologues and *An. gambiae* candidates is shown in Table 3.1. All TRA2 homologues contain eight exons, with the exception of *D. melanogaster* that has seven. AGAP006755 contains five exons, the largest number within the *An. gambiae* candidates. It should be noted that the AgamP3.6 transcritpome is unlikely to be complete and therefore RACE experiments would be required to identify the correct number of exons for each candidate. The genomic organisation of four *tra2* homologue gene sequences is shown in Appendix Figure 3. This analysis suggested that the size of the *tra2* homologue gene exon and introns are not conserved, but the positions of exons containing the SR rich regions are. All SR containing exons flank the RRM containing exon, but are separated by an exon not rich in serine/arginine rich residues. For future work, the position of exons containing the SR motifs of new *An. gambiae* candidates should be investigated. In this instance, due to the higher sequence similarity of AGAP006798 to other TRA2 homologues, this candidate was investigated further.

3.2.2 Expression profile of AGAP006798

It was decided to investigate the candidate AGAP006798 further as it had the highest similarity to other TRA2 homologues. Using the primers shown in Figure 3.3, a fragment of AGAP006798 (428 bp) was amplified and cloned following gel electrophoresis and sequencing confirmed the expected identity of the PCR product. To analyze the developmental expression profile of AGAP006798, RT-PCR on total RNA from different life stages was carried out (Fig. 3.3). A single amplification product of expected size was amplified in all developmental stages, including embryonic expression in 4 hr and 20 hr embryos after egg laying (AEL). No obvious difference in expression between males and females was observed with this approach, which suggests that a single transcript is expressed throughout development in males and females. Although this is not qPCR so cannot be sure if there are no differences in quantities.

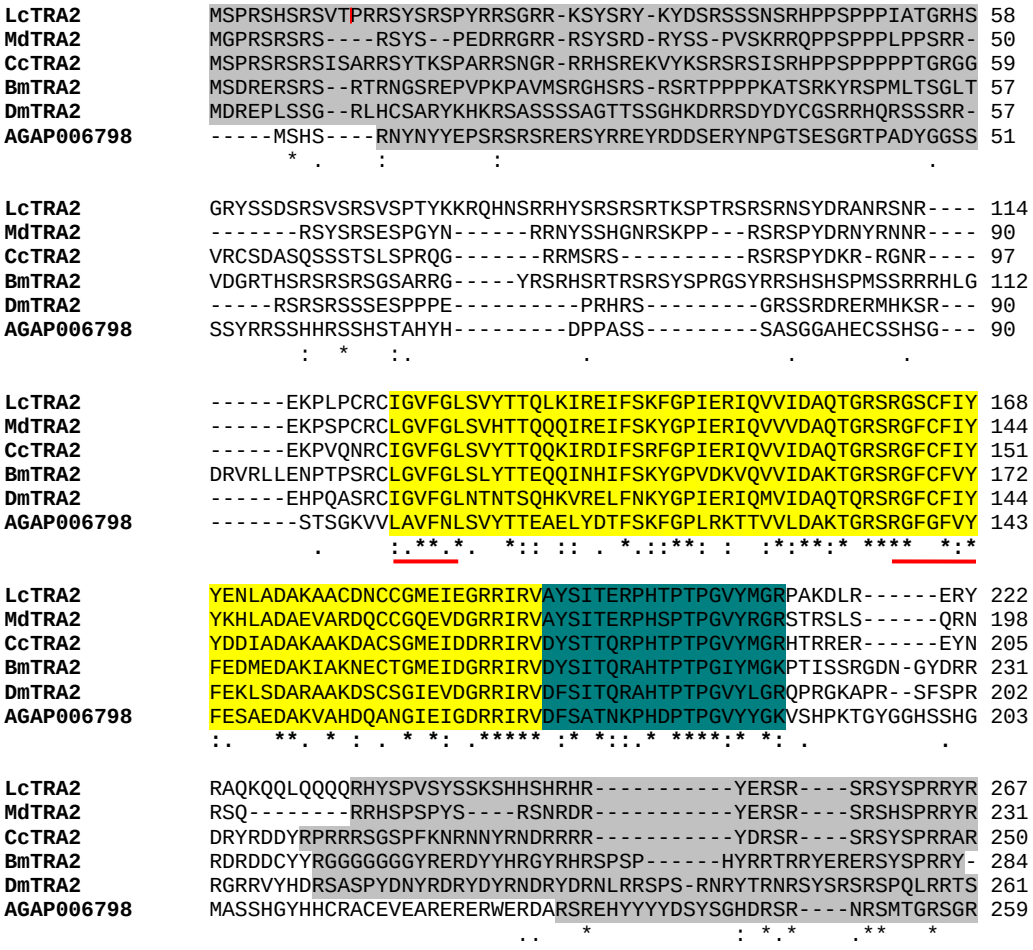


Figure 3.1. Amino acid alignment of AGAP006798 with known TRA2 sequences: The amino acid sequence of AGAP006798 was aligned to known *tra2* sequences via ClustalW2 using the default settings. Asterisk (*) indicates conserved amino acids between all species. Grey boxes represent serine arginine rich regions, yellow boxes represent highly conserved RNA Recognition Motif and green boxes the linker region. LcTRA2, L. *cuprina* (ACS34688), MdTRA2, *M. domestica* (AAW34233), CcTRA2 *C. capitata* (ACH81146), BmTRA2 *B. mori* (NP_001119705), DmTRA2 *D. melanogaster* (AAA62771) and AGAP006798 (EAA04295). Solid red lines indicate ribonucleoprotein identifiers (RNP-1 and RNP-2) common to RRM domains that are essential for RNA binding.

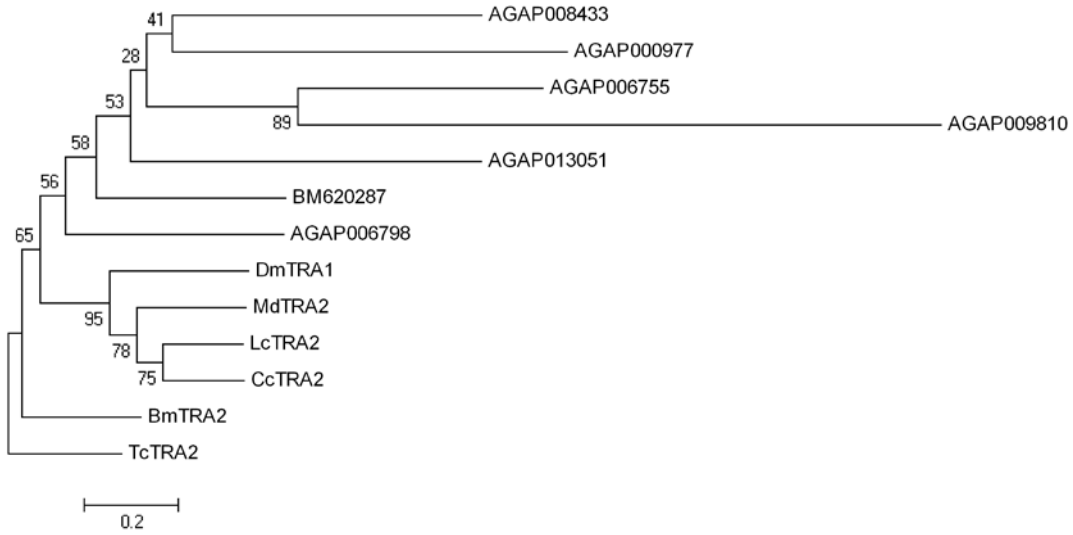


Figure 3.2: Maximum parsimony phylogenetic analysis of Transformer2 (TRA2) based upon amino acid sequences of TRA2 homologues and *An. gambiae* candidates. A rooted maximum parsimony phylogenetic analysis using 1,000 bootstrap replciates, using *T. casteneum* TRA2 as an outlier, of TRA2 homologues and *An. gambiae* candidate sequences. All *An. gambiae* sequences are shown using Genbank accession numbers. LcTRA2, L. *cuprina* (ACS34688), MdTRA2, *M. domestica* (AAW34233), CcTRA2 *C. capitata* (ACH81146), BmTRA2 *B. mori* (NP_001119705), DmTRA2 *D. melanogaster* (AAA62771) and AGAP006798 (EAA04295), TcTRA2 *T. casteneum* (EFA10734).

Table 3.1: Comparison of the exon/intron structure of insect TRA2 homologues and *An. gambiae* candidate sequences. The number of annotated exons and introns for known TRA2 homologues and *An. gambiae* candidates are shown along with accession numbers.

Species	Accession Number	Exon number	Intron number
D. melanogaster	AAA62771	**7**	**6**
C. capitata	ACH81146	**8**	**7**
M. domestica	AAW34233	**8**	**7**
B. mori	NP_001119705	**8**	**7**
An. gambiae	**AGAP006798**	**4**	**3**
An. gambiae	**AGAP008433**	**2**	**1**
An. gambiae	**AGAP000977**	**2**	**1**
An. gambiae	**AGAP006755**	**6**	**5**
An. gambiae	**AGAP009810**	**3**	**2**
An. gambiae	**AGAP013051**	**2**	**1**
An. gambiae	**BM620287**	**3**	**2**

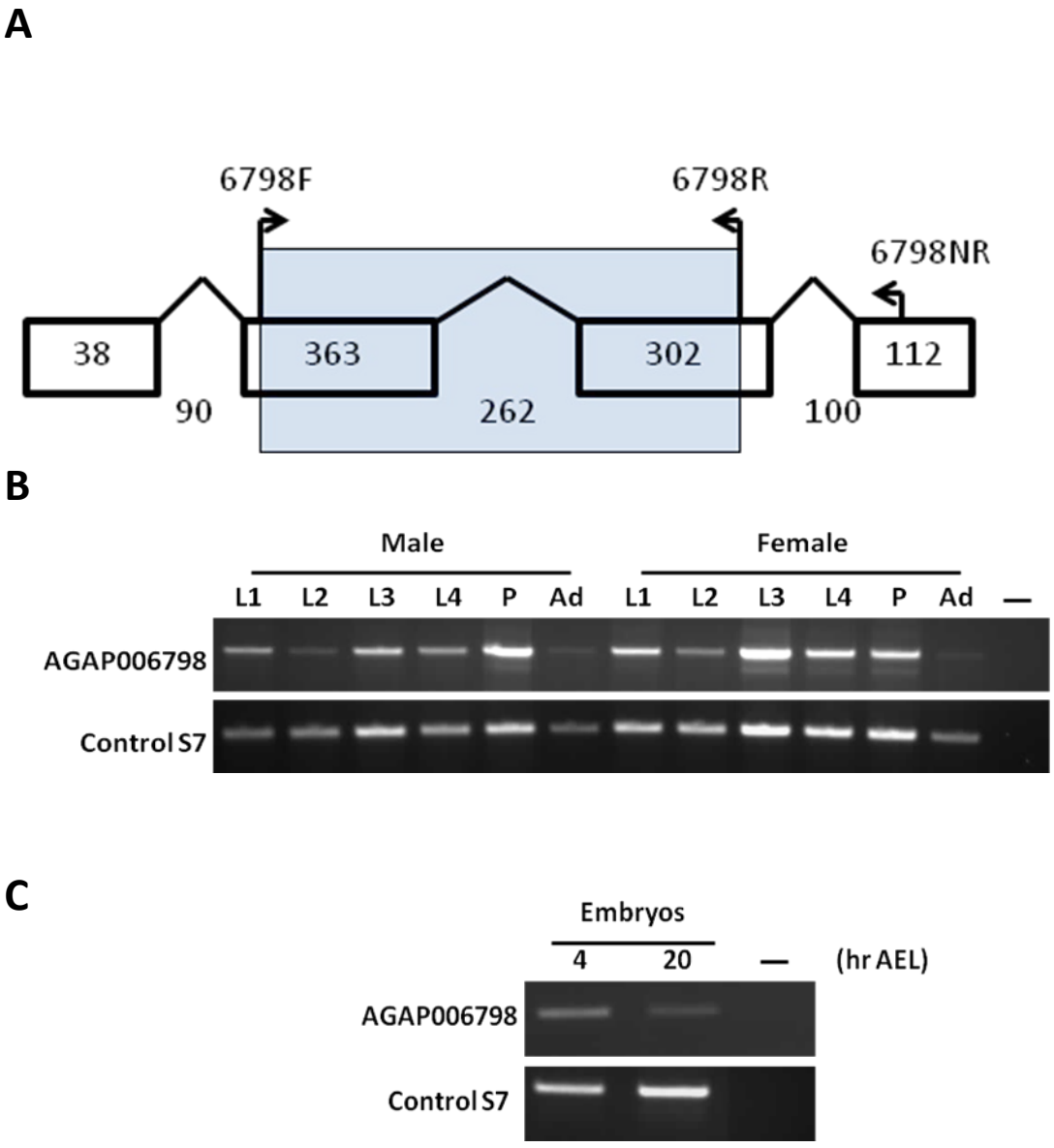


Figure 3.3. Genomic organisation of AGAP006798 and temporal expression profile. (A) T7 tagged primer pair 6798F/6798R was used to amplify the highlighted transcript fragment (pale blue box). Arrows indicate position of primers relative to the transcript. Open boxes represent exons, diagonal lines introns and numbers length in base pairs. (B) PCR amplification of first strand cDNA showed expression of the transcript throughout all developmental stages (C) One-step RT-PCR analysis of mixed sex embryos 4 and 20 hr AEL. Control S7 = amplification of *An. gambiae* ribosomal protein S7 transcripts (L20837) served as a loading control. First (L1), second (L2), third (L3), fourth (L4) instar larvae; P = pupae; Ad = adult. 6798NR primer was used to assess relative transcript abundance in RNAi experiments along with forward primer 6798F. Horizontal line (-), represents PCR or RT-PCR reaction with no cDNA or RNA template respectively.

3.2.3 Functional characterisation of AGAP006798

3.2.3.1 Examination of sex-ratio following AGAP006798 RNAi

RNAi was employed to transiently reduce AGAP006798 transcript levels to determine if the gene has a sex-determining function in *An. gambiae*. A 468 bp dsRNA template, corresponding to the conserved RRM region (highlighted in Fig. 3.3A) was used to synthesise dsRNA (ds6798). Embryo microinjection was performed for both ds6798 and dsGFP controls and hatched larvae were reared to adulthood and sexed by adult morphology. Over two replicates a male bias was seen from injections for ds6798 over dsGFP controls, a male to female ratio of 60:40 and 53:47 was observed for ds6798 and dsGFP respectively. The number of male and females observed for ds6798 and dsGFP is shown in Figure 3.4A. The statistical significance of the male to female ratio for ds6798 injections compared to dsGFP controls over the two replciates was examined using a Chi-squared test and this revealed no statistical difference between the experimental and control groups (P-value >0.05 and <0.1).

3.2.3.2 Examination of embryo microinjection technique

As no statistical difference was observed, the effectiveness of the microinjection technique for dsRNA delivery was assessed, without assessing the sex ratio. Embryos of a transgenic *An. gambiae* line expressing GFP (transgenic line FC (Lycett *et al.* 2011) were injected with either a mixture of dsGFP and ds6798 or dsGFP only as control. FC mosquitoes express GFP within the larval head, ventral nerve cord, chordotonal organs and the gonads of both sexes. FC mosquitoes also express the fluorescent reported dsRED in the larval eye. Observation using a fluorescent microscope of GFP expression knockdown in the hatched larvae was used as a sign of successful dsRNA delivery and gene knockdown in a given individual. Over 3 replicates, the percentage of larvae with no GFP expression varied from 2-11 %. Individuals with knockdown of GFP expression were also observed for dsRED expression using a fluorescent microscope, to confirm they were from the FC line and not contaminating wild type larvae.

The reliable confirmation of knockdown requires complete absence of GFP expression as judgement of partial reduction in intensity of GFP expression is not accurate and was not used. To overcome this, dsRNA was co-injected with a GFP expressing plasmid (FU III) and only those hatched larvae expressing GFP were considered. The FU III plasmid was constructed by B. Tresman (LSTM) and is a fusion of the *An. gambiae* polyubiquitin

(AGAP001971) promoter to the GFP reporter protein coding sequence. Previous embryo microinjection of the plasmid has shown that hatched larvae, from successful injections, express GFP in the posterior segments of the abdomen. Embryo microinjection of FU III into the G3 line was compared to injection of dsGFP into the FC transgenic line for confirmation of successful dsRNA delivery. Microinjection resulted in a higher percentage of hatched larvae expressing GFP (10-50 %), compared to complete silencing of GFP expression (2-11 %) by dsGFP delivery into the FC transgenic line.

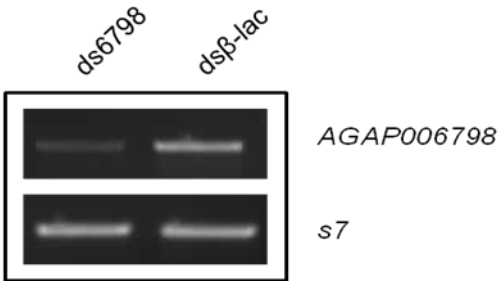
3.2.3.3 Investigation of RNAi targeting AGAP006798 on sexual morphology using GFP expression plasmid FU III as confirmation of delivery

As a higher perecentage of successful injections was observed using FU III, a further round of embryo microinjections were then performed to examine the sexual phenotype of hatched indicviduals using ds6798 and control dsβ-lactamase, co-precipitated with plasmid FU III. β-lactamase was used as control rather than dsGFP, due to the co-injection of a GFP expressing plasmid with the dsRNA. Two replicates of embryo microinjections were performed and firstly, semi-quantitative PCR was conducted using primer pair 6798F/6798NR (Fig. 3.2A) to assess relative transcript abundance following delivery of dsRNA. PCR was conducted using 1st strand cDNA, synthesised from total RNA extracted from GFP positive and negative larvae, 3 days post hatching for ds6798 and dsβ-lactamase (Fig. 3.4B). Primers targeting *An. gambiae* ribosomal S7 protein transcripts (L20837) were used for PCR as loading control. The percentage of hatched larvae expressing GFP ranged from 7-52 % across two replicates. Mean relative expression levels of AGAP006798, compared to S7 loading control, were repressed by 51.7 % ($p = 0.0068$) and 26.3 % ($p = 0.374$) for GFP and non-GFP expressing larvae respectively compared to dsβ-lactamase controls (Fig. 3.4C). Despite a non-significant reduction, a separate RNAi study was conducted where GFP positive individuals were reared to adulthood and sex assigned based upon adult morphology. A total of 7 (8.6 %) and 9 (17.3 %) of hatched larvae expressed GFP for ds6798 and dsβ-lactamase respectively. No phenotypic variation was observed in GFP positive individuals that would indicate an alteration in sexual phenotype by analysis of the developed terminalia. Therefore it cannot be concluded whether injection of ds6798 has an effect on *An. gambiae* sexual development.

A

Replicate Number	Number of embryos injected		Phenotpye of individuals			
	ds6798	dsGFP	ds6798		GFP	
			Male	Female	Male	Female
1	251	203	34	22	25	23
2	229	198	25	17	23	20

B



C

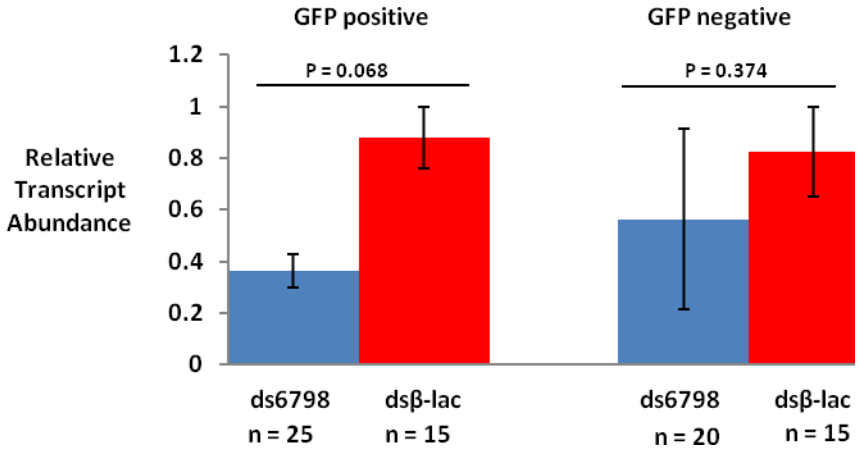


Figure 3.4. Summary of RNAi mediated knockdown of AGAP006798. (A) Summary of dsRNA delivery into FC GFP expressing transgenic line (Lycett *et al.* 2011). Embryos were injected with both ds6798 and dsGFP or dsGFP alone. dsGFP was injected to remove GFP expression from the FC line as an indication of successful dsRNA delivery. The number of larvae with no GFP expression is shown in the table. (B) A representation of semi-quantitative PCR following gel electrophoresis. S7 = *An. gambiae* ribosomal protein S7 transcripts (L20837) served as a loading control. (C) Following embryo microinjection of dsRNA (ds6798 experimental and dsB-lac control) and GFP expressing plasmid FUIII, all hatched larvae (total indicated below each bar) were observed under a fluorescent microscope for GFP expression and total RNA was then extracted from pooled samples across two replicates of GFP positive and negative larvae. PCR was carried out on first strand cDNA for all samples using primer pair 6798F/6798NR and S7 ribosomal control. Band intensities were measured as described in Materials and Methods and were normalised to the S7 loading control. P values calculated using student's t-test. Error bars represent standard error values. n = total number of individuals within the two replicates.

3.2.4 Alignment of TRA homologue protein sequences

Protein sequences of TRA homologues were downloaded from NCBI and aligned using ClustalW2 with default parameters (Fig. 3.5). Alignments revealed two regions, each 16 to 17 amino acids in length, of sequence similarity, though these were not conserved throughout all homologues. All sequences share serine/arginine di-peptide rich regions and a proline rich region at the C-terminus. The first region of similarity (TRA-CAM – shown in green) at the N-terminus is conserved within all species except for *D. melanogaster*. The second region, that is arginine rich (shown in red), is conserved between all species. Within the proline rich region (shown in blue), a further 17 amino acid region has high sequence similarity limited to *C. capitata*, *B. oleae*, *A. obliqua* (Diptera: tephritidae) and *L. cuprina* (Diptera: calliphoridae) Though the sequences of *D. melanogaster*, *G. morsitans* and *M. domestica* do not show sequence similarity within the third region, they are all proline rich. Between the second and proline rich regions, stretches of serine-arginine dipeptides are found for all *tra* homologues.

A bioinformatics approach was undertaken, using PHI BLAST searches against the *An. gambiae* genome using the three conserved domain sequences as query. PHI (Pattern-Hit Initiated) BLAST removes potential false positives as matching hits to the protein sequence must contain the conserved motif. The genomic location of matches for each query to each chromosome was recorded (Fig. 3.6). The sequences of the PHI BLAST query is shown in the Materials and Methods. In total, 9, 93 and 40 matches were identified for the GEG, RS and PP search queries respectively. The results were searched for presence of the matches within a genomic proximity, defined as presence of at least two query sequences within a 100 kb window. Successfully identified candidates were then examined for sequence orientation (positive or negative) and the results are summarised in Table 3.2. All three query sequences were identified on chr2R, but were separated by ~1.5 Mbs. Two candidates had matches to RS and PP within 25 Kb but were in opposite orientations and not considered. The gene boundary size of other insect *tra* genes does not exceed 8 Kb (*M. domestica* largest size), therefore the boundary size of the five candidates displayed in Table 3.2 far exceed that of known *tra* genes. In addition, the BLAST hits of two of the candidates are in the opposite orientation and therefore cannote be considered as a continous transcript sequence. Though not conclusive, these two reasons do suggest that the identified candidates may not be actual TRA homologues.

Figure 3.5. Amino acid alignment of *tra* sequences. Amino acid alignment of TRA sequences using ClustalW2. Conserved regions are indicated by colours (green = TRA-CAM; red = arginine rich; blue = proline rich). Arginine/serine regions are indicated by bold letters. CcTRA *Ceratitis capitata* (AAM88673), BoTRA *Bactrocera. oleae* (CAG29241), AoTRA *Anastrepha obliqua* (ABW04165) LcTRA, *Lucilia cuprina* (ACS34689), DmTRA *Drosophila melanogaster* (AAO38898), GmTRA *Glossina morsitans* (ACY40710) and MdTRA, *Musca domestica* (ACY40709).

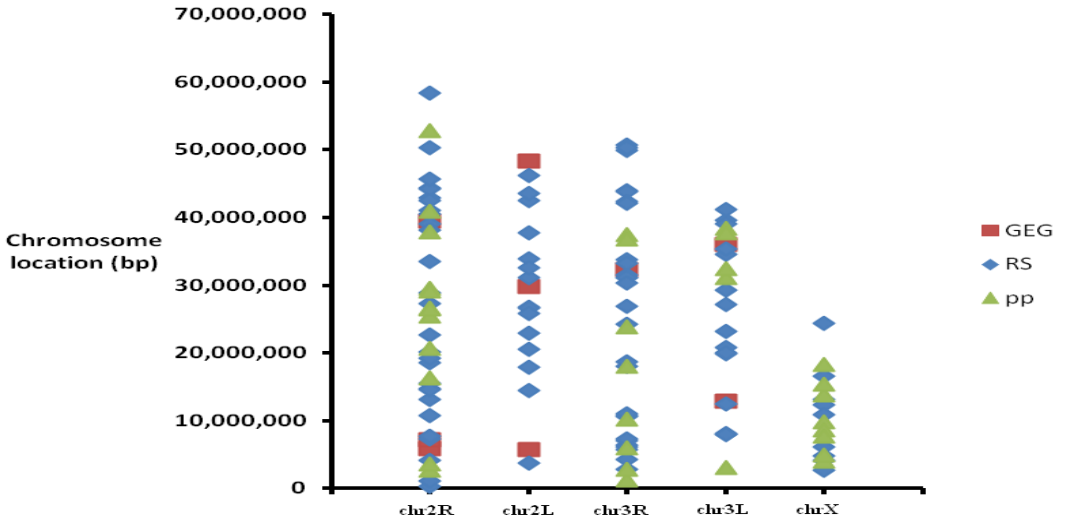


Figure 3.6. Genomic positions of PHI BLAST results for TRA query sequences. Conserved regions of known TRA homologues were identified following Clustal W alignment of protein sequences. Three regions, abbreviated by GEG (red squares), RS (blue diamond) and PP (green triangles) were identified and used to PHI-BLAST against the genome of *An. gambiae* and then grouped according to their chromosome location. Presence of BLAST matches to all three queries in a close genomic location was expected to identify an *An. gambiae* TRA homologue. The full query sequences (GEG, RS and PP) used for PHI-BLAST is shown in Materials and Methods and *Ceratitis capitata* (AAM88673) TRA protein sequence was also used as query to satisfy the PHI-BLAST criteria. The genomic location (base pair position) of each hit was then plotted (Y-axis) against each of the chromosome arms (x-axis) to assess their proximity to each other. Despite the location of PHI-BLAST results appearing to be in close proximity due the scale of the Y-axis, no candidates were identified in close enough proximity to be investigated further.

Table 3.2. Summary of top 5 candidates for homology to TRA query sequences. The chromosome location of each of the query sequences and the orientation and distance between the matches is shown. Genomic locations and orientation were identified using BLASTN nucleotide searches. The distance between the BLAST results is indicative of the gene boundary size.

Chromosome	GEG Position	RS Position	PP Position	Orientation	Predicted gene boundary size (bp)
chr2R	39464667-39464708	40541466-40541621	40944502-40944678	plus	1480011
chr2R	7232766-7232713	7294477-7294331	N/A	minus	61565
chr3R	N/A	2875108-2875248	2853144-2853010	plus/minus	22098
chr3R	N/A	18023760-18023566	18009665-18009778	plus/minus	13982
chr3R	N/A	6079803-6080069	5960917-5960861	minus	118942

3.2.5 Detection of *dsx* sex specific-splice forms

Sequences of the *An. gambiae dsx* transcripts (male isoform AY903307; female isoform AY903308) were downloaded from NCBI and mapped to the genome using BLASTN searches to establish genomic organisation of the transcripts, previously described by Scali *et al.* (2005)and shown in Figure 37A. This was done so that primers to amplify sex-specific forms could be designed (Fig. 3.7A). The forward primer (dsxF) is located in the common fourth exon and the reverse primers are located in the female specific 1.7 Kb exon five (dsxR2) and within the male specific exon (dsxR1).

3.2.6 Temporal expression profile of alternative *dsx* splice forms

A temporal profile of *dsx* expression was then conducted using RT-PCR upon sex-specific embryonic RNA, staged every 4 hr AEL (Fig. 3.7B). The female form is detected immediately after egg laying in both male and female embryos. It is also present in males and females 4 hr AEL, though at reduced levels in male compared to female embryos. The female form is only found in female embryos by 8 hr AEL and this pattern continues through to 32 hr, the latest time point tested. The male form is sex-specifically expressed as early as 8 hr AEL and continues through to 32 hr AEL. The male form is also detected, albeit at lower levels, immediately after egg laying in both sexes (0 hr AEL).

Sex-specific products were successfully amplified by RT-PCR from RNA of 1st instar larvae through to adult stages, as shown in Figure 3.6C. Adult and pupae males and females were separated based upon sexually dimorphic morphology. To separate male and female larvae, males of the transgenic line DA1 (Lycett *et al.* 2011) were mated to female G3 mosquitoes and their progeny inspected for GFP expression under a fluorescent microscope. The DA1 transgenic line has GFP inserted on the X chromosome. Therefore, only female progeny from the cross would express GFP and could be separated from non-GFP expressing males.

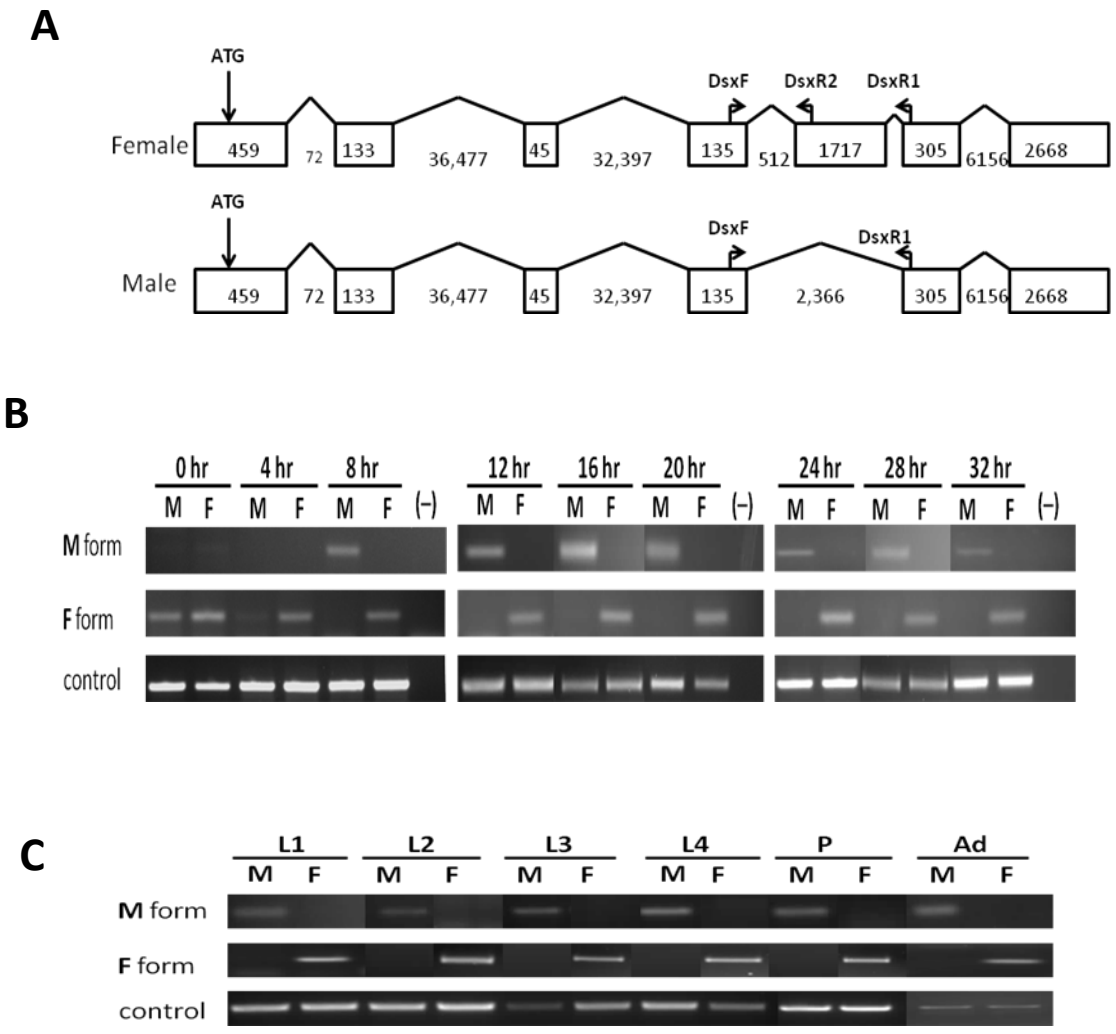


Figure 3.7. Molecular organisation of *An. gambiae dsx* and temporal profile of sex-specific forms. Each of the life stages was examined by RT-PCR using primer pairs dsxF1/dsxR1 (male specific expected product) and dsxF1/dsxR2 (female specific expected product). (A) Genomic organisation of male and female *dsx* transcripts. Arrows indicate position of primers used for RT-PCR. Open boxes represent exons, diagonal lines introns and numbers length in base pairs. ATG indicates the position of the start codon. The dsxR1 primer is present in the female untranslated region and this is not present in the mature mRNA, so an amplification product is not observed. (B) Temporal profile of embryonic expression at 4 hr intervals AEL showing sex-specific forms can be detected from 8 hr AEL. The female form is present in both sexes at 0 hr AEL. (C) Sex-specific expression through post embryonic stages up to adulthood. First (L1), second (L2), third (L3), fourth (L4) instar larvae; P = pupae; Ad = adult. S7 = *An. gambiae* ribosomal protein S7 transcripts (L20837) served as a loading control. M form = male transcript. F form = female transcript. M = male. F = female. Horizontal line (-) represents negative control (no template added).

3.2.7 Detection of *fru* sex specific splice forms

Sequences of the *An. gambiae fru* transcripts (male isoforms AY785361.1 and AY785360.1; female isoform AY725820.1) (Gailey *et al.* 2006) were downloaded from NCBI and BLASTN searches were used to map them to the genome to establish the genomic organisation of each transcript. Primers to amplify sex-specific forms were designed from the genomic organisation, previously described by Gailey *et al* 2006, and shown in Figure 3.8A. Six different combinations of primers were designed that would differentiate between males and females on a basis of size (fruF3/fruR3) or by presence only in males (fruJxF1/fruJxR and fruMalF/fruMalR) or females (fruJx2/fruJxR and fruFemF/fruFemR and fruF3 and fruR2). All primer combinations were used for RT-PCR amplification from sex-specific adult RNA to detect sex-specific forms.

Primer combinations fruJxF1/fruJxR, fruJxF2/fruJxR, fruFemF/fruFemR, fruMalF/fruMalR and fruF3/fruR2 did not detect sex-specific forms, amplifying the fragment of expected size in both sexes (shown in Appendix Fig. 2). The primer combination fruF3/fruR3 was able to discriminate between males and females based upon the size of fragments amplified. The primer pair flanks a 778 bp female specific exon and therefore amplifies a 168 bp fragment in males and a 945 bp fragment in females. Both male and female fragments were amplified in females whereas only the 168 bp male fragment was amplified in males. This still enabled sex discrimination of females from males using the presence of the female specific 945 bp fragment. All fragments were cloned and sequenced to confirm the correct region had been amplified.

3.2.9 Temporal expression profile of alternative *fru* splice forms

A temporal profile of *fru* expression was performed by RT-PCR using primer pair fruF3/fruR3 amplifying RNA from 1st instar larva through to adulthood (Fig. 3.8B). This revealed that sex-specific expression of the male and female isoforms occurred throughout development. The female specific 945 bp fragment was only found in female samples tested, the 168 bp fragment was found in both sexes (3rd and 4th instar larvae and pupal stages), albeit at much lower levels in females. In addition to the expected fragments of *fru* transcripts, multiple products, presumed non-specific products, were amplified in different developmental stages.

Several embryonic stages (4 to 20 hr AEL) were sampled to dissect sex-specific gene expression in these early stages. Amplification via RT-PCR using the same primer combination demonstrated that the sex-specific isoforms were detected in 16 and 20 hr AEL samples (Fig. 3.8B). Other embryonic stages sampled did not reveal the expected sex-specific isoforms because a large number of presumed non-specific fragments were amplified and clear sex-specific amplification was not evident. Despite optimisation of RT-PCR cycling parameters, amplification of non-specific fragments was still evident in other embryonic stages and prevented accurate assessment of *fru* transcript expression.

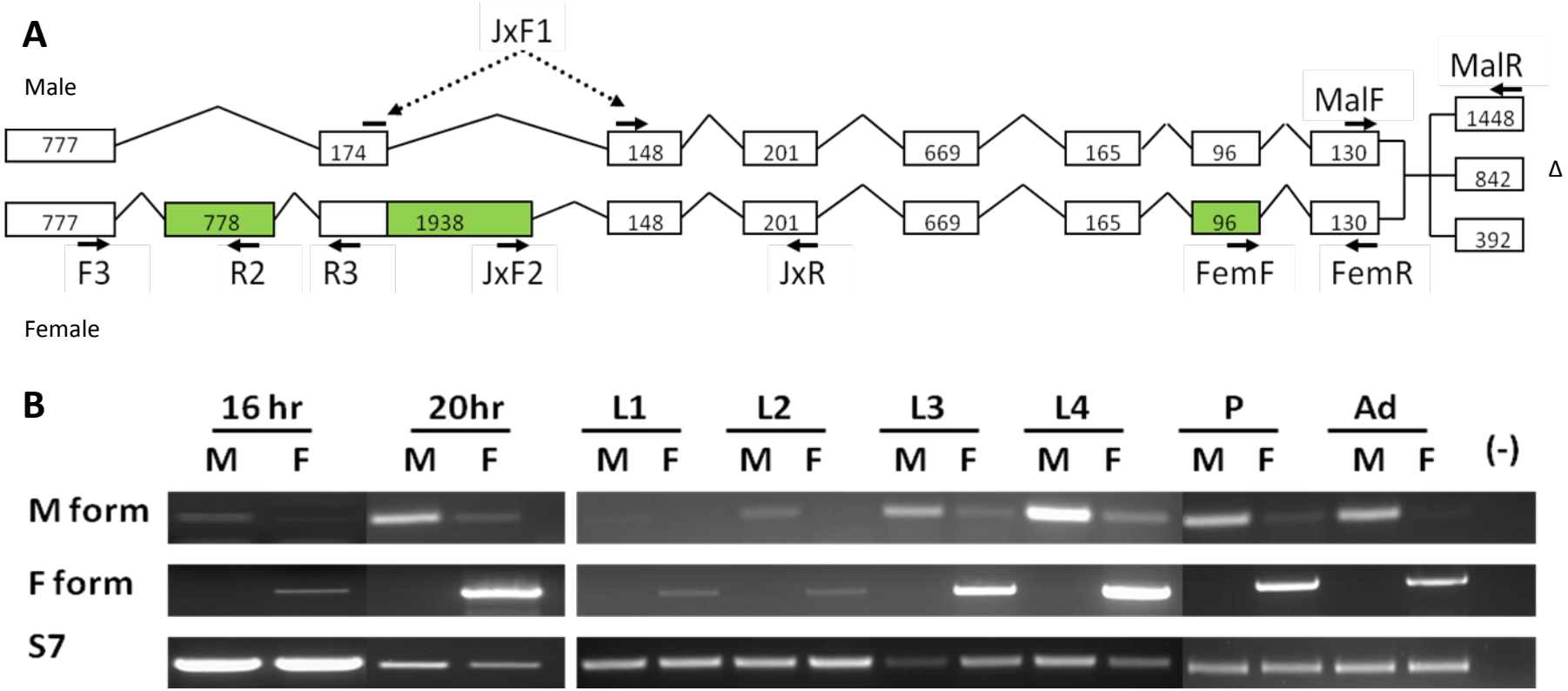


Figure 3.8. Schematic diagram of fruitless genomic organisation and temporal expression profile: Each life stage sampled was examined via RT-PCR using primer pair fruF3/fruR3 (A) Genomic organisation of male (upper) and female (lower) *fru* transcripts according to Gailey *et al* 2006. Arrows indicate position of primers used for RT-PCR. JxF1 primer sequence crossed the exon-exon boundary. Δ indicates alternative 3' ends of transcripts. (B) The female form is only detected in female embryos 16 and 20 hr AEL. The male form is found in both male and female samples of the embryo stages sampled. Sex-specific expression of the female form is found throughout all the post-embryonic developmental stages. The male form is found in female of L3, L4, pupal and adult stages. First (L1), second (L2), third (L3), fourth (L4) instar larvae; P = pupae; Ad = adult. S7 = *An. gambiae* S7 ribosomal protein transcripts (L20837) served as a loading control. M form = male transcript. F form = female transcript. M = male. F = female. (-) = negative control (no template)

3.3 Discussion

Molecular mechanisms underlying sex determination in *An. gambiae* are poorly understood, despite the post-embryonic characterisation of three genes known to be important in other dipteran sex determination pathways. Previously, adult sex-specific splicing and complementation tests in *D. melanogaster* mutants confirmed involvement in the sex determination pathway for *fru* in *An. gambiae* (Gailey *et al.* 2006). Similarly, *dsx* is sex-specifically spliced in post-embryonic stages, suggesting that it is functionally conserved (Scali *et al.* 2005; Magnusson *et al.* 2011). The third gene, *sxl*, is not sex-specifically spliced as in *D. melanogaster* and is therefore presumed to have no function in *An. gambiae* sex determination (Traut *et al.* 2006). Here, we isolated candidate *tra* and *tra2* homologues and analysed the embryonic expression of two *An. gambiae* genes, *dsx* and *fru*.

Currently, no *An. gambiae* homologues of *tra* or *tra2*, the genes responsible for *dsx* sex-specific splicing in *D. melanogaster,* have been identified. The investigation of a putative *tra2* homologue (AGAP006798) has proved inconclusive in assigning a role in *An. gambiae* sex determination. Using protein sequence similarity to *D. melanogaster* TRA2, AGAP006798 was identified within the AgamP3.6 predicted proteome as a putative *tra2* homologue. A single unique fragment of AGAP006798 was amplified in both sexes in all stages from embryonic through to adulthood. Though it should be noted that we did not use primers that cover the entire transcript region. This is contradictory to what is observed in *D. melanogaster*, but similar to other dipterans. Alternatively spliced *tra2* transcripts are found in *D. melanogaster*, that encode for three protein isoforms (Amrein *et al.* 1990; Mattox *et al.* 1990). Two isoforms, $TRA2^{264}$ and $TRA2^{226}$ act with female-specific TRA in the somatic tissues to control female sexual differentiation (Mattox *et al.* 1996). In addition, two isoforms, $TRA2^{226}$ and $TRA2^{179}$ are expressed in the male germline, though it is only the former that is involved in spermatogenesis (Mattox *et al.* 1996). In other dipterans, only a single transcript is present in *C. capitata*, *M. domestica, S. ocellaris and L. cuprina* (Burghardt *et al.* 2005; Concha and Scott 2009; Salvemini *et al.* 2009; Martin *et al.* 2011). Expression of AGAP006798 is therefore similar to other *tra2* genes, with the exception of *D. melanogaster*. Though it should be noted here that we did not examine AGAP006798 expression using different primer combinations. Therefore we cannot conclusively state that no alternative transcripts of AGAP006798 are expressed.

Within our study, initial RNAi experiments did not result in a statistically significant introduced sex-bias that would be expected if a sex-reversal phenotype had occurred. Induction of sex-reversal has been shown in other dipteran species, including *M. domestica* (Burghardt *et al.* 2005) and *C. capitata* (Salvemini *et al.* 2009). It is possible that the 51.7 % reduction in transcript levels observed was insufficient to induce sex-reversal, as this may not correlate to a similar reduction in protein levels (Gygi *et al.* 1999). Previously, *An. gambiae* embryo microinjection of dsRNA targeting eight different male-biased genes resulted in a discernable phenotype for only two candidates, though no information on reduction in transcript abundance was presented (Magnusson *et al.* 2011). A non-significant reduction in transcript levels ($p = 0.068$) may indicate that increasing the amount of dsRNA delivered may increase the level of knockdown to that sufficient to induce sex-reversal. Previous *tra2* RNAi studies do not report the level of transcript knockdown achieved in order to achieve sex-reversal, therefore no frame of reference is available to determine the level required.

Delivery of dsRNA to the embryo should not represent a significant problem as recent knockdown studies relying on dsRNA injection into embryos have produced a downstream phenotype, validating the technique for both *An. gambiae* and *Ae. aegypti* (Clemons *et al.* 2011; Haugen *et al.* 2011; Thailayil *et al.* 2011). Though the genes targeted in the previous studies were not involved in sex determination, targeting a germ cell differentiation gene (*zero population growth*) (Thailayil *et al.* 2011) and two neuronal development gene (*frazzled* and *semaphorin-1a*) (Clemons *et al.* 2011; Haugen *et al.* 2011) demonstrated that an active RNAi pathway is present in mosquito embryos. In addition, knockdown of *zpg* resulted in abnormal gonad development, demonstrating that knockdown can occur through different developing cell populations (Magnusson *et al.* 2011; Thailayil *et al.* 2011). Also, within this study, removal of GFP expression from a small percentage of FC mosquito larvae following dsGFP delivery demonstrates that gene knockdown in widespread tissues is feasible.

D. melanogaster sexually dimorphic terminalia arise from the genital disc at the posterior of the developing embryo (Estrada *et al.* 2003). *dsx* sex-specific splicing in posterior embryonic segments is responsible for formation of male and female structures (Nothiger *et al.* 1987). Therefore, dsRNA must be delivered to specific regions of the embryo to see a phenotypic effect on sexual development. In this study, only adults developing from larvae expressing GFP were analysed for changes in the terminalia as it was assumed dsRNA had

been successfully delivered. What should be noted is that GFP expression from the injected plasmid does not confirm that dsRNA is present in the targeted tissues or throughout the entire embryo. Although, as dsRNA injections were carried out at the posterior end (where the genital disc arises), of pre-blastoderm embryos, it would be expected that dsRNA be present within the targeted tissues, so this may not account for the lack of phenotype. As a suggestion for future RNAi experiments analysing terminalia of successfully injected individuals, use of the FC transgenic line may be favourable. Complete visual knockdown of GFP expression confirms that dsRNA is present in multiple tissues throughout posterior and anterior regions of the embryo and is therefore more informative than introducing GFP expression using the FU III plasmid.

It is also possible that AGAP006798 may not be the true orthologue of *tra2,* despite showing sequence similarity to other TRA2 proteins and containing the highly conserved RRM and serine/arginine rich regions. All TRA2 homologues contain RRM domains that share two ribonucleoprotein identifier sequences, often associated with RRM domain proteins (Bandziulis *et al.* 1989). Partial sequence similarity was also found within the two ribonucleoprotein identifier sequences. It is therefore possible that AGAP006798 may be a homologue of *tra2*, but the gene does not have the same function in *An. gambiae* as it does in other species. Phylogenetic analysis suggested that AGAP006798 is the more closely related of the *An. gambiae* candidates identified to other insect TRA2 sequences. Though it should also be noted that AGAP006798 contains four exons, compared to seven or eight found in all other insect tra2 genes. Though it is possible that the complete gene boundaries of AGAP006798 is incomplete and that RACE experiments may identify additional exons.

A further consideration is that *An. gambiae tra2* remains unannotated in the current AgamP3.6 gene build and that the BLASTP approach taken here would not be expected to lead to its identification. This suggests that *de novo* transcriptome sequencing is required before firm conclusions regarding additional candidates can be made. The future approaches mentioned above, increasing dsRNA concentration or use of FC transgenic line GFP knockdown individuals would have to be explored before any firm conclusions of AGAP006798 function could be made. In addition, had more time been available, a greater number of embryos should have been injected to improve the sample size of GFP knockdown individuals.

We also attempted to isolate an *An. gambiae tra* homologue and, previously, other homologues of *tra* have been identified using degenerate primers (Ruiz *et al.* 2007; Concha and Scott 2009; Hediger *et al.* 2010) and bioinformatics approaches based upon sequence similarity (Verhulst *et al.* 2010). The close linkage of *D. melanogaster tra* with the gene l(3)73Ah has also been exploited to identify homologues. Based upon conserved synteny to *l(3)73Ah, tra* homologues have been identified in *B. oleae* (Lagos *et al.* 2007) and *C. capitata* (Pane *et al.* 2002). It is expected that a homologue of *tra* is present in *An. gambiae* due to the presence of TRA/TRA2 binding sites in *dsx* and *fru* transcripts (Scali *et al.* 2005; Gailey *et al.* 2006). No putative *tra* homologues are found in regions close to *An. gambiae l(3)73Ah* (data not shown) and bioinformatics approaches failed to identify a likely candidate. A likely candidate was not identified as all candidates were separated by very large putative introns and often, we in the opposite orientation. The salient feature of the sex determination pathways is the evolutionary lability of their elements, *tra* sequences have rapidly diverged, even within the *Drosophila* genus despite functional conservation within a number of species (Kulathinal *et al.* 2003). Despite the functional conservation of TRA within dipteran species as well as hymenopterans, sequence similarity is restricted to two amino acid motifs and a proline rich region in the TRA sequences examined within this study. Verhulst *et al.* (2010) suggest that a conserved dipteran domain, a conserved hymenopteran domain and interestingly, a putative TRA autoregulatory domain (TRA-CAM), represent the similarity between TRA proteins in addition to proline rich and serine/arginine di-peptide regions. The rapid evolution of *tra* sequences may prevent identification of its *An. gambiae* homologue via comparative approaches and currently available data. No strong candidates were identified within our study due to large suggested gene boundary sizes and the presence of PHI-BLAST hits in opposite orientations. Therefore, a significantly greater time investment would be required to identify the *An. gambiae tra* homologue.

Targets of TRA and TRA2 regulation, *dsx* and *fru*, have previously been shown to be sex-specifically spliced as early as 1st instar larvae (Magnusson *et al.* 2011) and adults (Gailey *et al.* 2006) respectively. Here, however, we have shown for the first time that *fru* and *dsx* sex-specific splicing also extends to the early embryonic stages. With regards to *dsx*, following maternal deposition, female transcripts are present in both sexes immediately after egg deposition. Degradation of maternally deposited female transcripts in male embryos is coupled with the zygotic expression of the male form, establishing sex-specific

male expression by 8 hr AEL. Though male transcripts are also present immediately after egg laying, we did not detect the male *dsx* form in adult females, suggesting that they are not maternaly deposited. It is possible that the male transcripts are paternally derived, perhaps delivered with the sperm during fertilisation. The male form is absent in males at 4 hr AEL and is presumably degraded, prior to onset of zygotic transcription.

Through either increased stability of maternally deposited female transcripts in females or zygotic transcription of the female form (or a combination of both), the female form is also sex-specific from 8 hr AEL onwards. Stable maternal transcripts have been demonstrated previously in *D. melanogaster* embryos, though the time points examined in our study extend past those previously examined in *D. melanogaster* (Bashirullah *et al.* 1999; De Renzis *et al.* 2007). As we could not discriminate between maternal and zygotic female *dsx* transcripts in this study, we cannot conclude the level of contribution to *dsx* transcript levels from maternal deposition and zygotic transcription To assess the level of zygotic contribution to transcript levels, generation of compound chromosomes to create embryos that lack defined regions of zygotic template have been used for *D. melanogaster* (Merrill *et al.* 1988; De Renzis *et al.* 2007). Such an approach is beyond the scope of this project, but, ablataion of chromosome2R (genomic location of *An. gambiae dsx* gene) and comparision of female *dsx* transcript levels to wild type embryos, would reveal the extent of zygotic transcription contribution. As a less time-consuming alternative, qRT-PCR could be used to assess *dsx* transcript levels. It would be expected that the onset of zygotic transcription would coincide with an increasing level of transcript abundance and this would confirm a zygotic contribution to *dsx* transcription.

The work presented here suggests that *An. gambiae* sexual identity is determined very early in development and temporal *dsx* expression is similar to that of other dipterans. In *C. capitata*, maternally deposited female *dsx* transcripts (4 hr AEL) are degraded prior to sex-specific zygotic expression (10 hr AEL) (Gabrieli *et al.* 2010). In *Ae. aegypti,* of the two female (F1 and F2) and one male (M1) *dsx* transcripts, F1 and M1 are maternally deposited and present throughout development. F2 expression begins 12 hr AEL indicating that sex-specific zygotic expression also takes place in *Ae. aegypti* embryos (Salvemini *et al.* 2011). All these differ to D*. melanogaster*, where no maternal contribution of *dsx* has been reported. *D. melanogaster* male *dsx* splice forms are detected in embryos as early as 10 hr AEL, but are restricted to the developing somatic gonad precursors (DeFalco *et al.* 2003;

Hempel and Oliver 2007). Hempel *et al* (2007) did not detect *dsx* transcripts in embryos 3-10 hr AEL. If more time had been available, it would have been interesting to determine the embryonic spatial distribution of *An. gambiae dsx* transcripts and whether they are localised to the developing somatic gonad as in *D. melanogaster* through *in situ* hybridisation experiments.

Non-sex-specific *fru* transcripts are both maternally deposited and zygotically expressed in *D. melanogaster* embryos (Song *et al.* 2002). Sex-specific transcripts are detected but appear to have no function in the developing embryo; mutations affecting only non-sex specific transcripts produced an abnormal neuronal phenotype and ectopic expression of male transcripts failed to rescue the mutants (Song and Taylor 2003). Similar to *D. melanogaster*, in our study, sex-specific female transcripts have been identified in the *An. gambiae* embryos at least as early as 16 hr AEL. Earlier stages may exhibit sex-specific forms but could not be reliably identified due to amplification of presumed non-specific products preventing accurate assessment following RT-PCR. Though cycling conditions were optimised, additional bands to those expected continued to be present. Low levels of the male *fru* form are observed in female samples in both embryonic and post-embryonic stages and the consequence of this is unclear. The primers used in this study (fruF3/fruR3) flank a female specific 778 bp exon. In males this exon is not present in the mature transcript. Previously, this 778 bp exon has been amplified from male whole bodies, consistent with our findings, but not in male heads (Gailey *et al.* 2006). Male specific *fru* transcripts have been shown to localise to a specific subset of *D. melanogaster* neurons and are required for development of a muscle of Lawrence (MOL) (Gailey *et al.* 1991; Lee *et al.* 2000). Therefore, spatial expression of male specific *fru* transcripts is an additional level of control to that of sex-specific alternative splicing. Expression of male *fru* transcripts in females does not preclude a male specific function as sex-specific post-trancriptional regulation of *fru* transcript has been demonstrated previously (Usui-Aoki *et al.* 2000). Therefore it is likely that an additional level of control, in combination with sex specific splicing, also regulates *An. gambiae fru* function.

This work has confirmed embryonic sex-specific splicing of key *An. gambiae* sex determination genes. Though the identity of *dsx* and *fru* upstream regulator(s) still remains elusive, demonstration that *dsx* is sex-specifically spliced in the embryonic stages provides motivation for the sampling of embryo transcriptiomes by RNA-seq to identify upstream elements of the cascade. In addition, embryonic sex-specific splicing of *dsx* is of great

benefit to a novel genetic control strategy. Fu *et al* (2007) demonstrated that female-specific splicing of *C. capitata tra* can be implemented into a female-specific autocidal control system, driving toxic levels of the tetracycline transactivator protein variant (tTAV) in females only. Here, insertion of the *dsx* female-specific intron within the coding sequence of a lethal gene could be used to drive conditional embryonic lethality. In males, this intron would be retained, interrupting the coding sequence of the lethal gene. Though in principle this would be expected to lead to female-specific mortality, attempts in *D. melanogaster* inserting the female-specific *dsx* intron into the pro-apoptotic gene *hid* resulted in both male and female mortality (Scott *et al.* 2004). Scott *et al.* (Scott *et al.* 2004)hypothesise that a cryptic splice site within the *hid* coding sequence resulted in the female-specific intron being spliced out in both sexes, therefore, following such an approach in *An. gambiae* may not be straight forward.

Additionally, isolation of a female-specific embryonic promoter to ensure sex-specific lethality would be of great benefit. Through comparative sex-specific transcriptome analysis, it may be expected that such a promoter could be identified. Early acting lethality would be of great benefit to a mass release strategy as it would greatly reduce logistical costs. A recent field trial released over 25,000 *Ae. aegypti* transgenic males, sorted mechanically, across a 10 hectare study site (Harris *et al.* 2011). Extrapolating this to a larger area and a longer time frame would greatly increase costs and logistical problems. Sex-specific mortality at the embryonic stage would circumvent this issue and therefore would be of great benefit to a transgenic strategy.

Chapter 4: Establishment of sex-specific *Anopheles gambiae* embryo transcriptomes

4.1 Introduction

Malaria research entered the post-genome age at the turn of the last century. The genomes of human (Venter *et al.* 2001), *Anopheles gambiae* (Holt *et al.* 2002) and *Plasmodium falciparum* (Gardner *et al.* 2002), the triumvirate of a malaria life cycle, were all sequenced using the Sanger sequencing method (Sanger and Coulson 1975; Sanger *et al.* 1977). Genome sequencing has facilitated the dissection of many subjects including but not limited to: host pathogen interactions (Daily *et al.* 2005), parasite drug resistance (Dharia *et al.* 2009; Dharia *et al.* 2010) and vector pathogen interactions (Xu *et al.* 2005). Annotation of genomes has enabled global, whole genome approaches for identification of genes, without the need to first isolate each gene individually by homology before characterisation. For example, sequencing of the *An. gambiae* genome enabled identification of over 200 immunity related genes (Christophides *et al.* 2002) and subsequently, an identified immunity gene, *relish2* (*rel2*) was shown to regulate development of *Plasmodium* parasites (Meister *et al.* 2005). Genome sequencing has therefore provided a wealth of data, allowing detection of gene expression on a global scale, often through use of microarrays.

Previously microarrays have been used for the identification of sex-biased gene expression at various developmental stages and tissues, including midgut tissue (Warr *et al.* 2007), various adult tissues (Baker *et al.* 2011), larval to adult stages (Magnusson *et al.* 2011) and whole adults (Marinotti *et al.* 2006). These studies isolated sex-biased gene expression relating to later stage sexually dimorphic characteristics, such as male-biased spermatogenesis associated genes (Marinotti *et al.* 2006). More recently, *de novo* sequencing approaches using next generation sequencing have isolated sex-specific differences within adult antennae, maxillary palps and whole bodies (Pitts *et al.* 2011). Within our study, it is the aim that sequencing of sex-specific *An. gambiae* embryo transcript libraries will greatly enhance knowledge of sex-specific gene expression, which is currently limited to post-embryonic developmental stages. In addition, the demonstration of *doublesex* and *fruitless* sex-specific embryonic splicing in chapter 2, suggests that further sexually dimorphic gene expression may be identified.

Prior to sequencing of the *An. gambiae* genome, analysis of global changes in transcript abundance using microarrays was restricted to knowledge of expressed sequence tags

(ESTs). In early studies, less than 4,000 putative genes were sampled (Dimopoulos *et al.* 2002). Post-genome sequencing, microarray platforms can sample gene expression of 10,622 protein coding genes of the 13,320 (VectorBase AngamP3.6) currently annotated (Wang *et al.* 2010; Baker *et al.* 2011). Despite an increase in the number of genes, a limitation of microarrays is the requirement for previous annotation of genes for probe design. Therefore analysis may be limited, and could not be comprehensive for entire transcriptome content. High throughput *de novo* sequencing can overcome some of these limitations.

The advent of Next Generation Sequencing (NGS) technology has allowed access to cheaper and faster large scale sequencing, compared to traditional Sanger methods. High throughput sequencing based upon either long read (Roche 454 and Pacific Biosciences® SMRT) or short read (Life Technologies™ SOLiD and Illumina®) technologies vastly increased the amount of sequence data produced. Due to the lower cost compared to short read platforms, and the generation of longer reads, the 454 platform was used in this study. Use of Roche 454 sequencing has enabled many practical applications such as genome sequencing (Wurm *et al.* 2011), single nucleotide polymorphism discovery (Novaes *et al.* 2008; Williams *et al.* 2010), protein-RNA interactions (Sanford *et al.* 2009) and transcriptome sequencing (Vera *et al.* 2008; Alagna *et al.* 2009; Guo *et al.* 2010; Gregory *et al.* 2011) to be routinely carried out.

Massively parallel sequencing using the 454 platform exponentially increases the scale of sequence data achievable compared to Sanger sequencing (Margulies *et al.* 2005). A potential complication of 454 sequencing arises with sequencing of homopolymers. Homopolymers (repetition of a single base) are sequenced by measurement of increasing light intensity and as such, are susceptible to over or under base calling errors (Wicker *et al.* 2006). To compensate for this, measurement of emitted light is also used to assign quality scores to individual bases. This allows the probability of error for each base to be determined and is defined using a Phred quality score (Ewing and Green 1998; Ewing *et al.* 1998). Phred quality scores are logarithmically linked to probability of incorrect base calling. For example, a base with a Phred quality score of 30 has a probability of 1/1000 of being incorrect (Ewing and Green 1998).

Following 454 sequencing, large data files are produced including FASTA files with sequences of each individual read and their quality scores. Removal of low quality reads is important prior to analysis of sequence data as inaccurate base calling can lead to errors in

mapping. Mapping of 454 reads to a reference transcriptome or genome can be performed with a variety of software, including but not restricted to Newbler gsMapper (Margulies *et al.* 2005), BWA-SW (Li and Durbin 2010) and GMAP (Wu and Watanabe 2005). Each software aligns individual reads to a reference, reporting the region of the reference that each read corresponds to. The availability of the *An. gambiae* genome (Holt *et al.* 2002) and access to transcriptome resources (VectorBase AgamP3.6 transcriptome) greatly facilitate this approach. Mapping to a reference generates a sequence alignment/map (SAM) file that details information of each individual read aligned to a reference. An example of a SAM alignment file is shown in Appendix Table 2.

Transcriptome sequencing of coding regions requires isolation of poly(A)+ mRNA from the total RNA population of the sample being studied as mRNA can represent as little as 5 % of the total RNA population (Lamond 2002). The majority of cellular RNA is ribosomal RNA (rRNA) which is not the main focus of the study and must therefore be removed prior to sequencing. Sequencing without rRNA removal would result in low representation of protein-coding genes in the resulting transcript libraries. mRNA can be isolated by poly(A)+ purification (Ma *et al.* 2011; Price *et al.* 2011), use of rRNA depleted random hexamer primers (rRNA priming sequences were removed from a hexamer library) (Armour *et al.* 2009) or depletion of rRNA sequences (Huang *et al.* 2011). In this study, we utilised poly(A) enrichment of mRNA sequences.

It is envisioned that sequencing of male and female poly(A)+ transcriptomes will improve the current understanding of *An. gambiae* sex determination and sexual differentiation through identification of genes transcriptionally active in only one sex. A further benefit of this study is that the embryonic stages of *An. gambiae* have not been sampled using a *de novo* NGS approach previously, and assessment of embryonic transcriptome content will greatly improve knowledge of mosquito genome content. *de novo* sequencing of embryonic stages has been performed in the milkweed bug embryo (Ewen-Campen *et al.* 2011), *C. capitata* (in combination with adult heads) (Gomulski *et al.* 2008), *Aedes aegypti* (Biedler *et al.* 2012) and that of the stable fly (in combination with post-embryonic stages) (Olafson *et al.* 2010). Several of these studies have isolated potential homologues of sex determination genes (Gomulski *et al.* 2008; Olafson *et al.* 2010). For example, putative homologues of the sex determination genes *dsx* and *transformer2* (*tra2*) were identified through 454 sequencing of the stable fly (Olafson *et al.* 2010). It is beneficial to target the embryonic stages for identification of sex determination genes as Graveley *et al.* (2011)

demonstrated that the embryonic transcriptome represents a smaller transcript pool compared to later developmental stages, therefore improving the probability of detecting sex-determining genes.

To enable transcriptome sequencing of sexed embryos, genetic separation is required as embryos are indistinguishable by morphological differences. Molecular approaches to distinguish between embryonic sexes of a species have relied upon markers present on the Y chromosome, as seen for *C. capitata* (Gabrieli *et al.* 2010) and *D. melanogaster* (Lott *et al.* 2011). *An. gambiae* contains differentiated heteromorphic sex chromosomes (X and Y) and repetitive X and Y linked fragments have previously been identified (Krzywinski *et al.* 2004; Krzywinski *et al.* 2005). Thus utilising sex chromosome linked markers, male and female *An. gambiae* embryos may be separated to provide sufficient material for mass sequencing of transcriptomes.

4.1.1 Specific Chapter Aims

- Isolate sufficient mRNA from individual male and female embryos; separated by karyotype to establish their sex.
- Mass sequence sex-specific cDNA using the Roche 454 platform to generate transcript libraries that represent the embryonic transciptome.
- Investigate the success of sampling the embryonic transcriptome through alignment of generated reads to both the *An. gambiae* PEST genome and transcriptome (Agam.P3.6).

4.2 Results

4.2.1 Separation of male and female embryos

A diagnostic PCR was used to assign sexual identity to individual embryos. The PCR amplifies Y linked (AgY477) and X linked (AgX367) repetitive sequences to allow karyotype separation of embryos. A schematic representation of the PCR is shown in Fig. 4.1A. AgX367 shares high sequence similarity to AgY477 aside from a 110 bp indel. Using primer pair 124678F2 and 124678R2, male-specific amplification of a 477 bp fragment of the Y linked AgY477 satellite and non-sex specific ampliciation of a 367 bp fragment of the X linked satellite AgX367 allows discrimination between male (XY) and female (XX) genomic template (Fig. 4.1B). Use of individual embryos as template for PCR produces the same diagnostic banding pattern (Fig. 4.1C), confirming the ability to differentiate between embryos of different sex. By homogenising the remaining embryo material in TRIzol® and storing at -80 °C, assigning sexual identity could be coupled with collection of sex-specific embryonic RNA.

4.2.2 Collection of sex-specific RNA

Large scale identification and collection of sex specific embryonic material was undertaken as shown in Figure 4.2. Two time points of embryonic development were sampled, 4-6 and 18-21 hr AEL. In total, 4,530 embryos were sampled from either 4-6 hr (397) or 18-21 hr (4,133) AEL, of which 53.9 % produced the expected amplicons for either male or female karyotype following PCR amplification. For the remaining samples, the expected banding pattern was either not present, or could not be accurately distinguished, and were therefore discarded. This corresponded to pools of 1,285 and 1,158 embryos for males and females respectively, from which RNA and DNA was extracted. However, PCR using primer pair 124678F2 and 124678R2 upon pooled DNA extracts revealed the presence of a 477 bp product, diagnostic to the Y chromosome, in the female sample, indicating the presence of contaminating male DNA. It was decided that the extracted RNA would not be used for transcriptome sequencing due to this contamination and that more embryos would be processed.

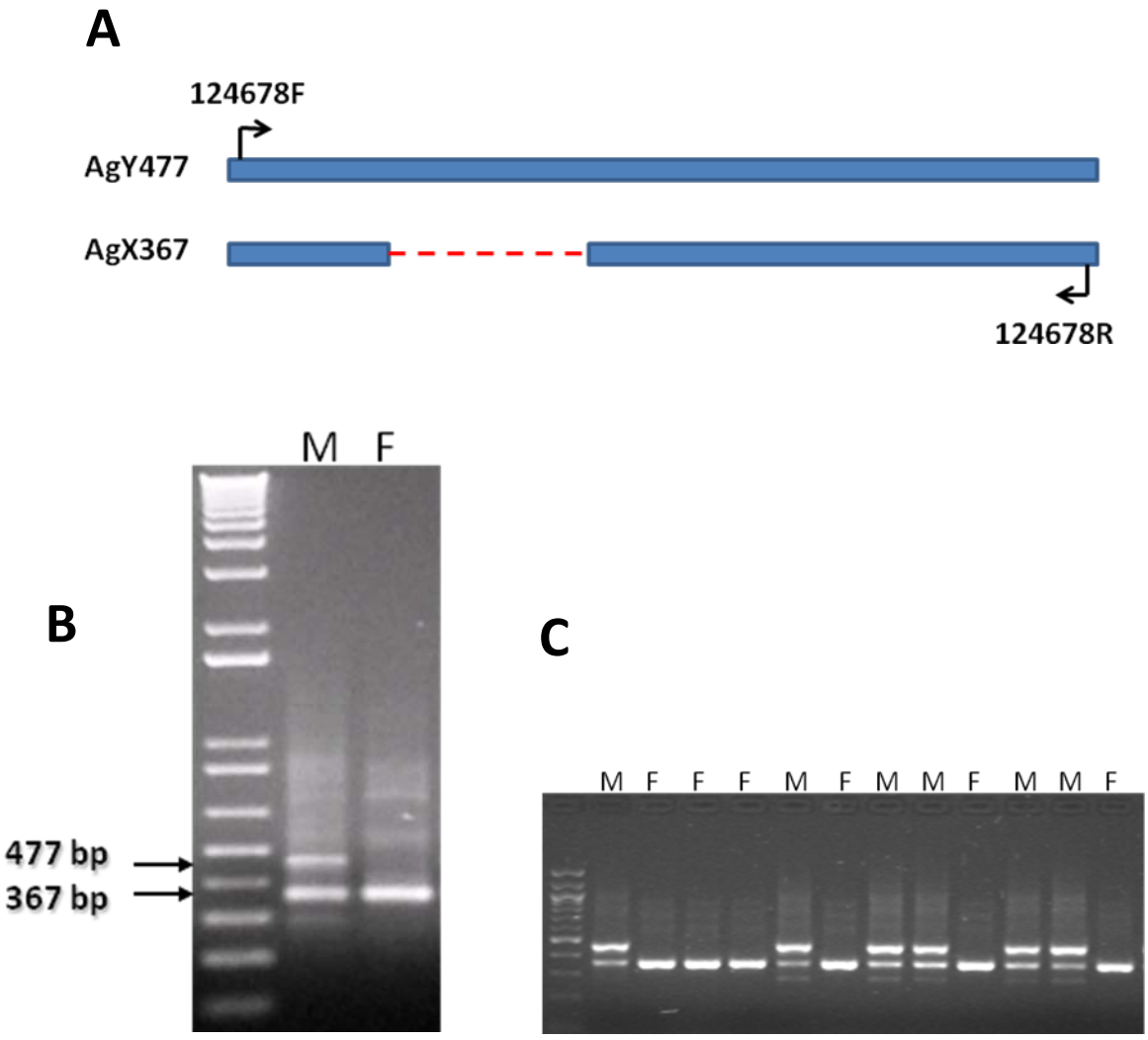


Figure 4.1. Schematic diagram of the sexing PCR reaction and resolving of PCR products via gel electrophoresis. (A) Primer pair 124678F and 124678R amplify highly similar repetitive sequences on the X and Y chromosomes that differ by a 100 bp indel (indicated by the red dashed line) AgY477 = Y-linked repeat sequence; AgX367 = X-linked repeat sequence. (B) Amplification patterns from genomic DNA of individual male and female adults. (C) Amplification patterns from individual embryos. PCR amplification of a 477 and 367bp product indicates a male XY karyotype and a single 367bp band a female XX karyotype. M = male; F = female.

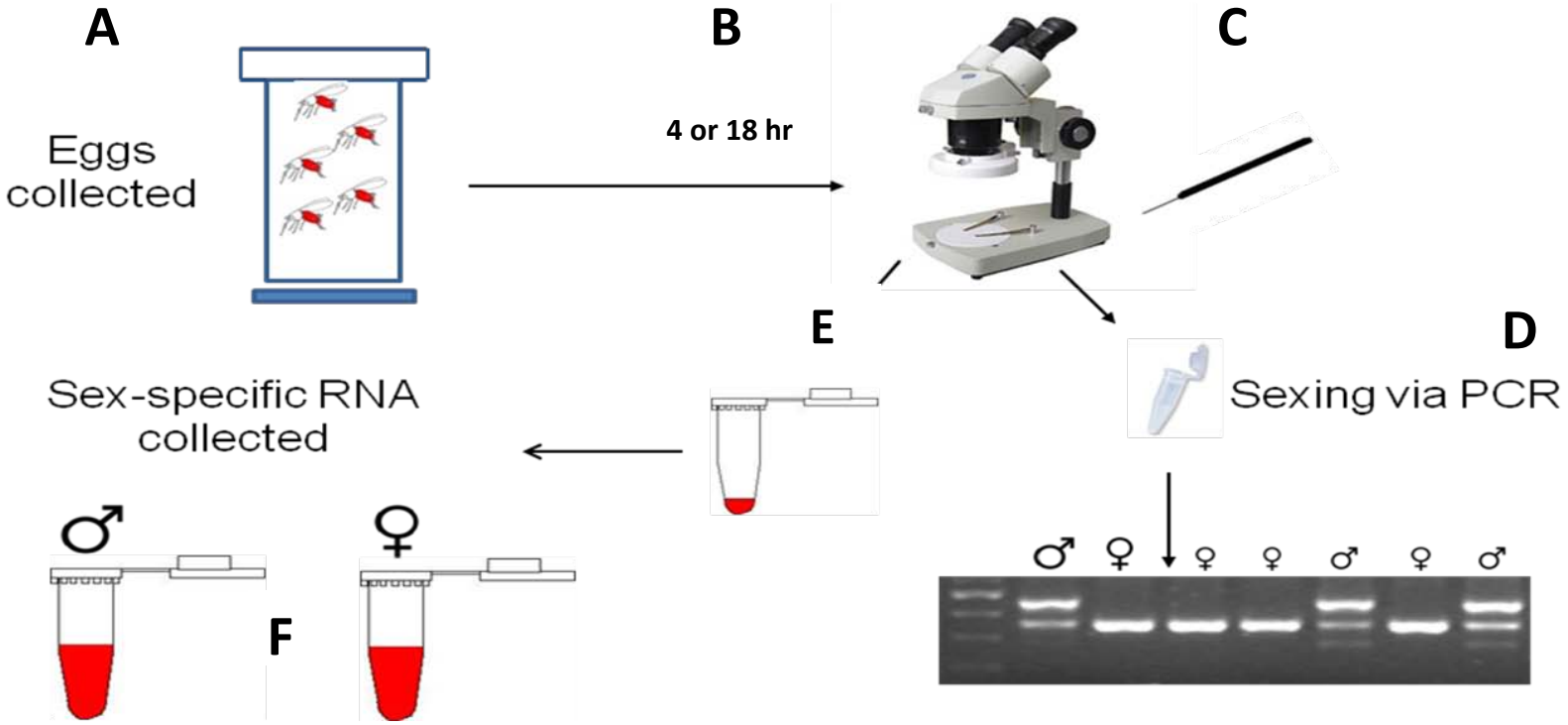


Figure 4.2. Schematic representation of the isolation of sex-specific embryonic RNA. (A) 10-12 mated females, 3 days post blood meal, were forced to lay eggs for 60 min within a laying chamber. (B) Female mosquitoes were removed and collected eggs maintained in insectary conditions for 4 or 18 hr. (C) Aged embryos were individually taken and a needle used to crush them inside a 0.5 ml Eppendorf tube. (D) Residual material on the needle was transferred to a PCR reaction-containing tube and amplified as described in materials and methods. (E) The remaining embryo material within the 0.5 ml Eppendorf tube was then crushed in 2 µl TRIzol reagent and stored at -80 °C. (F) Following analysis of PCR product by gel electrophoresis, single embryos stored in TRIzol were pooled together.

Large scale identification and collection of embryos was repeated and to avoid contamination of entire samples, individual embryos were pooled together in smaller samples and karyotyped prior to final pooling and RNA extraction. A second set of primers (magNF/magNR) that amplify a Y linked *mag*-like retrotransposon integration into a *mdg1*-like retrotransposon, was used along with 124678F2/124678R2 to detect contamination of pooled embryonic DNA. This was done to improve the detection of possible male contamination. In total 2,823 embryos were processed, of which 2,043 (72.4 %) produced the expected male or female amplification pattern following PCR and gel electrophoresis. From this number, 794 extracts from female embryos were not included in the final pool due to detection of Y linked fragments in the corresponding extracted DNA, to leave 821 and 428 male and female samples respectively. The male samples were split 11 and 89% and the females 18 and 82% for 4-6 hr and 18-21 hr AEL embryos respectively.

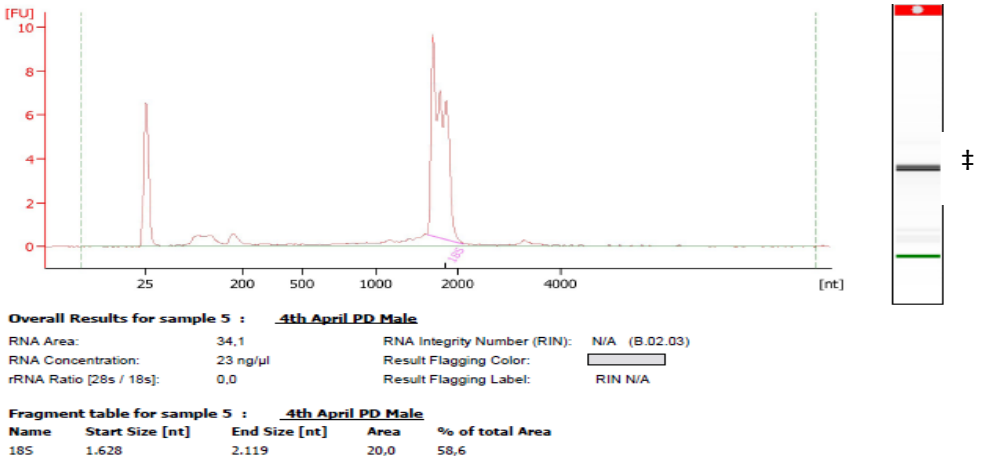
RNA was extracted from the combined sex-specific samples, as summarised in Table. 4.1. Purity of the RNA was assessed by NanoDrop™ spectrophotometer 260/280nm absorbance ratio, with values of 1.87 and 1.82 for males and females respectively. A value of 2.0 is considered to be 'pure' RNA. RNA sample quality was assessed using the Agilent 2100 Bioanalyzer and comparison against the RNA 6000 standard ladder (Fig. 4.3). Analysis of both samples showed a well defined peak of the 18S rRNA and an absence of low molecular weight noise, indicating RNA of high quality. A defined 28S rRNA peak that is observed in other species is not observed in *An. gambiae*. Insect 28S rRNA consists of two separate hydrogen-bonded fragments that separate and co-migrate with the 18S rRNA band during electrophoresis (Winnebeck *et al.* 2010).

A total of 24.84 and 8.02 µg RNA was collected for males and females respectively. Following DNase treatment of samples, purified mRNA (male = 15 ng; female = 3.5 ng) was then used for first strand cDNA synthesis followed by optimisation of cycle number for cDNA amplification, as described in Materials and Methods (Fig. 4.4). Samples were amplified for 23 and 26 cycles for males and females respectively. Due to lower concentration of female mRNA, the cDNA amplification procedure was repeated a second time for female first strand cDNA only. An approximate 1 µg cDNA was purified as described in Materials and Methods and sequenced at the CGR.

4.2.3 Raw Read Analysis

Over 1 million raw reads were generated from the two cDNA libraries, totalling 300 Mbps of sequence information. Total numbers of reads were comparable for both sexes, with 522,495 and 500,012 produced for males and females respectively. Similar raw read length ranges were produced for males and females (Fig. 4.5). The largest raw read found in each of the datasets was 702 (quality score = 12) and 1,084 bp (quality score = 14) for males and females respectively. Following cleaning, 408,851 (78.2 %) and 353,399 (70.7 %) reads remained for males and females respectively. For the remainder of this chapter, cleaned reads will be referred to simply as reads.

A



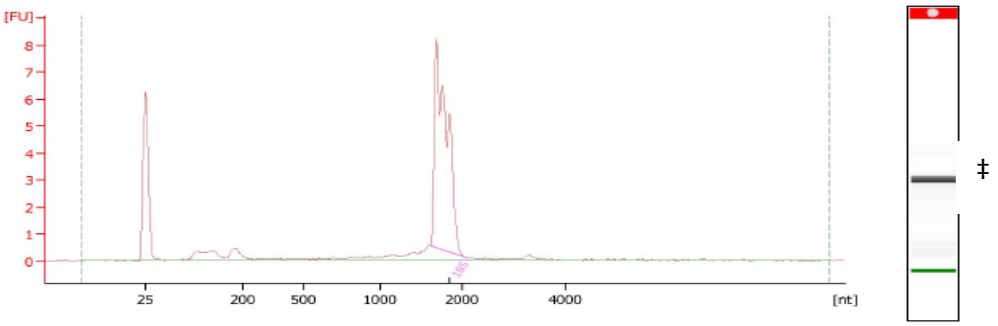
Overall Results for sample 5 : **4th April PD Male**

RNA Area:	34,1	RNA Integrity Number (RIN):	N/A (B.02.03)
RNA Concentration:	23 ng/µl	Result Flagging Color:	
rRNA Ratio [28s / 18s]:	0,0	Result Flagging Label:	RIN N/A

Fragment table for sample 5 : **4th April PD Male**

Name	Start Size [nt]	End Size [nt]	Area	% of total Area
18S	1.628	2.119	20,0	58,6

B



Overall Results for sample 6 : **4th APril PD Female**

RNA Area:	27,5	RNA Integrity Number (RIN):	N/A (B.02.03)
RNA Concentration:	19 ng/µl	Result Flagging Color:	
rRNA Ratio [28s / 18s]:	0,0	Result Flagging Label:	RIN N/A

Fragment table for sample 6 : **4th APril PD Female**

Name	Start Size [nt]	End Size [nt]	Area	% of total Area
18S	1.624	2.116	16,9	61,3

Figure 4.3. Agilent 2100 Bioanalyzer results of RNA extracted from (A) male and (B) female pooled embryo samples. Electropherograms showing defined peaks representing 18S and 28S rRNA and minimal low molecular weight noise are reflective of high quality RNA. Identification of a clear peak for the 18S rRNA for both male (A) and female (B) samples indicates the RNA is of high quality. 28S rRNA peaks are not found in *An. gambiae* samples due to separation into two fragments and co-migration with the 18S band following electrophoresis (Winnebeck *et al.* 2010). Therefore only migration of 18S rRNA can be seen (indicated by ‡). RNA Integrity Numbers (RINs) were not produced for either sample due to the low concentration of samples. The peak at 25 nt represents a standard marker fragment.

Table 4.1. Summary of embryo samples used for synthesis and amplification of cDNA. The number of embryos processed and the final number of each time point used for RNA extraction is detailed, along with RNA quantity and purity.

	Male	Female
Total number of embryos processed	1,172	871
Number of expected amplification patterns	821	428
Number of 18-21 hr embryos	732	349
Number of 4-6 hr embryos	89	79
RNA quantity (µg)	24.84	8.02
RNA 260/280 ratio	1.87	1.82

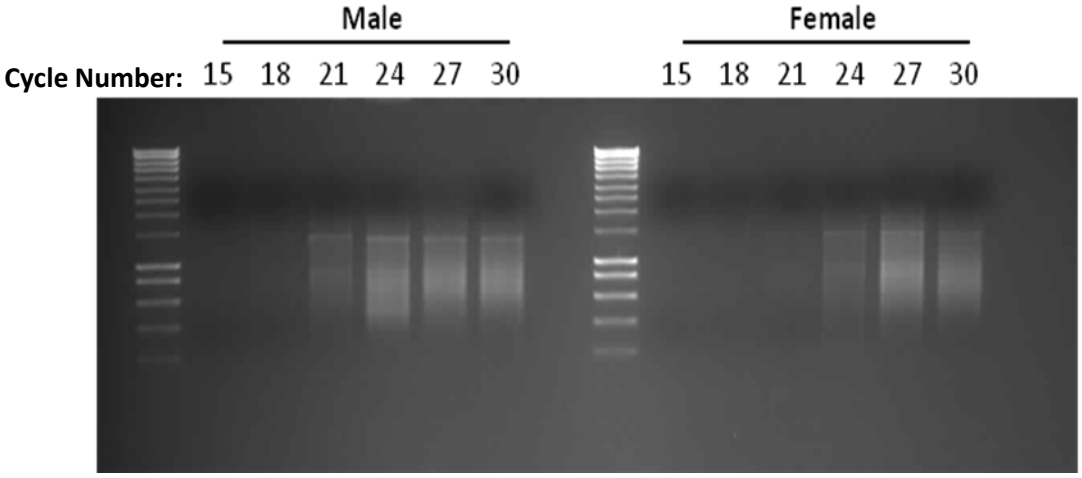


Figure 4.4. Empirical optimisation of PCR cycling conditions for cDNA amplification Empirical calculation of the optimum number of cycles of PCR amplification required for male and female mRNA samples. First strand cDNA was synthesised from mRNA and used for amplification. At each indicated cycle, 5 µl of PCR reaction was analysed via gel electrophoresis. Optimum cycle number was defined as one less than the number resulting in a plateau of cDNA smear intensity. Seen here, a plateau in intensity is found after 24 and 27 cycles for males and females respectively. Therefore optimum cycle number for male and females was 23 and 26 for males and females respectively.

Table 4.2. Summary of raw reads produced from the two RNA-seq datasets. Raw reads produced from the 454 platform are sumamrised and are comparable between male and female datasets.

	Male	Female
Total number of raw reads	522,495	500,012
Minimum Length (bp)	29	29
Max Length (bp)	1,084	702
Total number of nucleotides (bp)	153,802,540	148,610,545

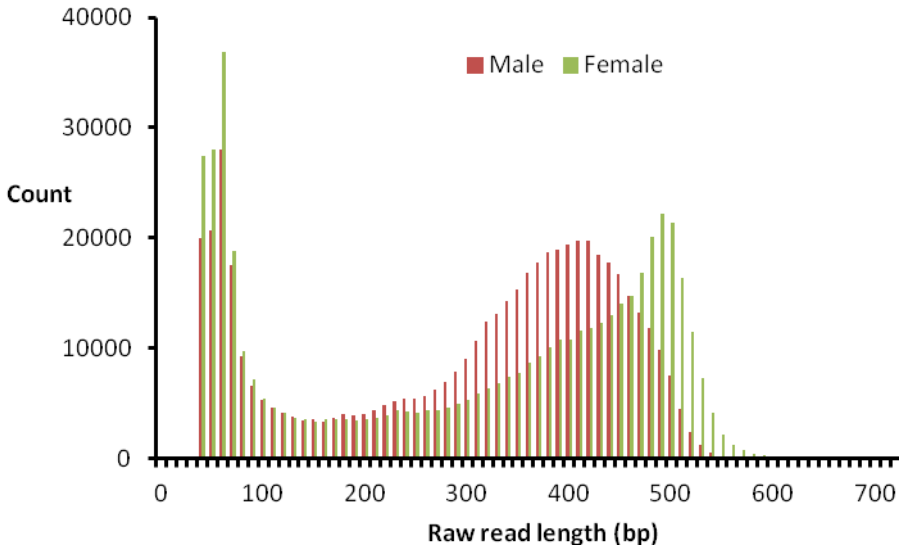


Figure 4.5. Size distribution of 454 raw read lengths for males and female datasets. Histogram of raw red lengths, smoothed by assigning contigs to partitions of 10 bp in length. RNA-seq datasets contain 153 Mbps and 148 Mbps of sequence information for males and females respectively. Females have a greater number of raw reads larger than 500 bp. Both datasets contain a large number (>40,000) of reads less than 50 bp in length. Males and females both show a bimodal distribution in raw read lengths with separated peaks around 70 (both) and 400 (male) or 500 (female) base pairs.

4.2.4 Mapping of male and female combined datasets to the *An. gambiae* PEST genome and transcriptome

Reads originating from male and female embryos were combined and aligned to the *An. gambiae* genome using BLASTN searches to determine their distribution across the genome. In total 715,209 reads (96.8 % of all reads) were aligned to the genome. Allowing for low complexity and repeat sequences by setting the BLAST DUST filter to false, 97.3 % of reads aligned to the genome. The distribution of reads to the genome is shown in Figure 4.6A. A second approach was undertaken using Newbler gsMapper to map the distribution of reads across the AgamP3.6 gene build transcriptome (Fig. 4.6B). Only the 715,209 reads that map to the *An. gambiae* genome were considered. In total, 2 % of reads did not map to the genome, suggesting that gsMapper is more restrictive compared to the BLASTN approach used previously, as all reads used as input for gsMapper were assumed to align to the genome following BLASTN searches.

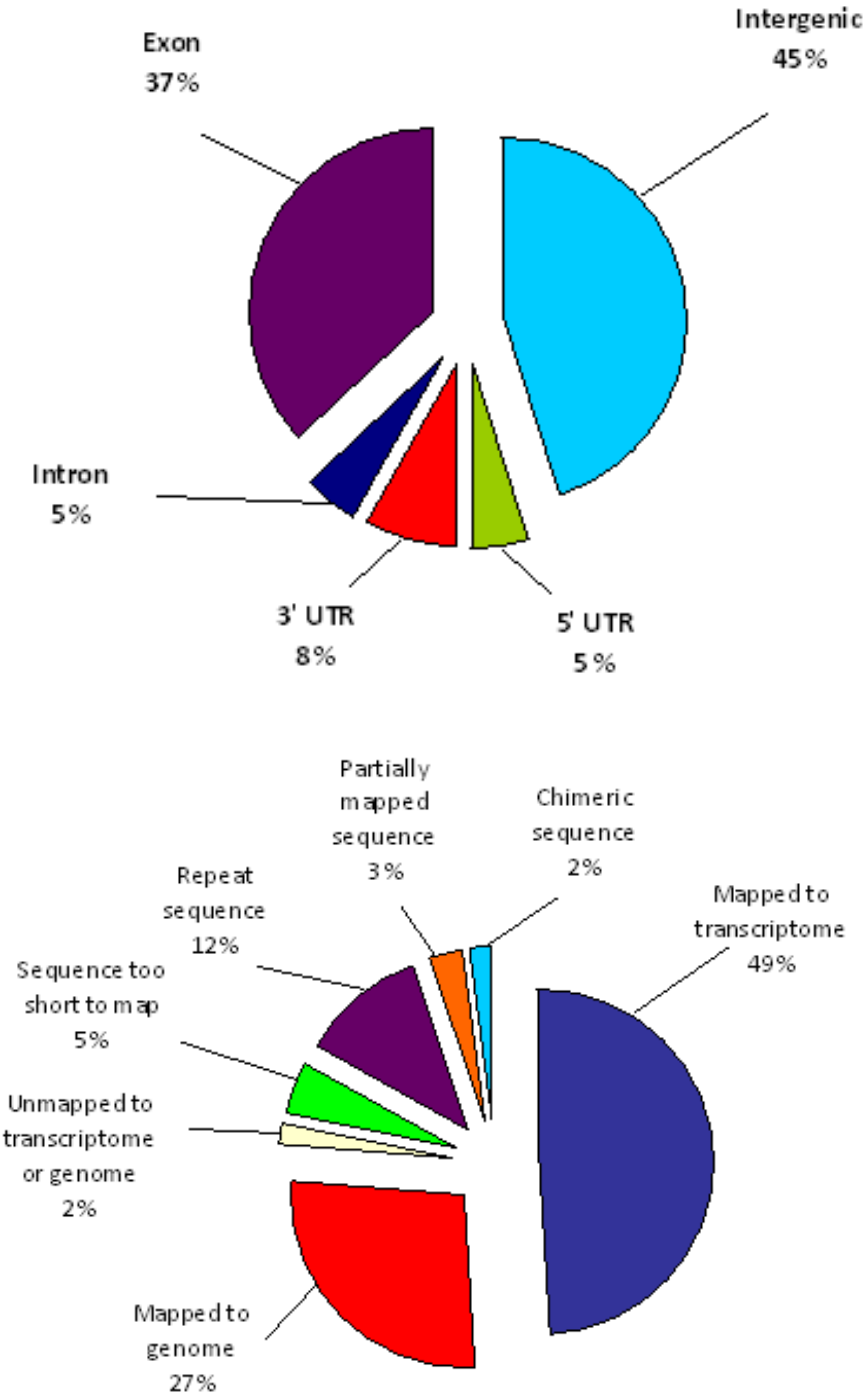


Figure 4.6. Distribution of reads to the genome (A) and transcriptome (B) of *An. gambiae*. Male and female read datasets were mapped to both the transcriptome and genome. (A) Reads were used to query the genome through BLASTN (E-value = $1e^{-10}$) searches against the *An. gambiae* genome to determine read distribution. (B) A second approach was taken using Newblerv2.3 gsMapper default settings. This software assigns a read status to each read to improve the accuracy of mapping. full (>90 % bases aligned to the reference), partial (<90 % of bases aligned to the reference), chimeric (different parts of the read align to different locations), repeat (read aligns to multiple reference locations), unmapped (does not align to the reference) and too short (less than 30 bp in length).

4.2.5 Mapping of individual male and female read datasets to the *An. gambiae* genome

Reads were mapped to *An. gambiae* chromosomes (using GMAP) and their distribution was visualised using the Integrative Genome Viewer (IGV) and wiggle plots, representing distribution of reads along the chromosomes, were created using IGVtools. In this instance, GMAP was used instead of Newbler gsMapper as GMAP allows output of alignments in SAM format that can then be converted into wiggle plots. During the course of this project, Newblerv2.6 was released and this version is able to output BAM files (binary version of SAM file), but due to the timing of its release, was not used in this study. Chromosomal distribution of reads is shown in Fig. 4.7A, across the two autosomes (chromosome2 and chromosome3), the sex chromosomes (chromosomeX and chromosomeY) and unplaced scaffold sequences (chromosomeUNKN). Multiple high degrees of mapping, indicated by an increase in value on the Y-axis, occur throughout the genome and indicate they are areas of high expression, relative to those with lower peaks.

The percentage of reads aligning to each chromosome differs, but, no differences in the percentage of aligned reads between males and females are observed. ChromosomeUNKN sequence scaffolds have not been assigned to any chromosome and over 5 % of reads for both sexes align within them. The percentage of the reads that map to both autosomes, and to the X chromosome, for both sexes is similar. A total of 0.04 % uniquely mapped male reads are located on the Y chromosome (summarised in Fig. 4.7B).

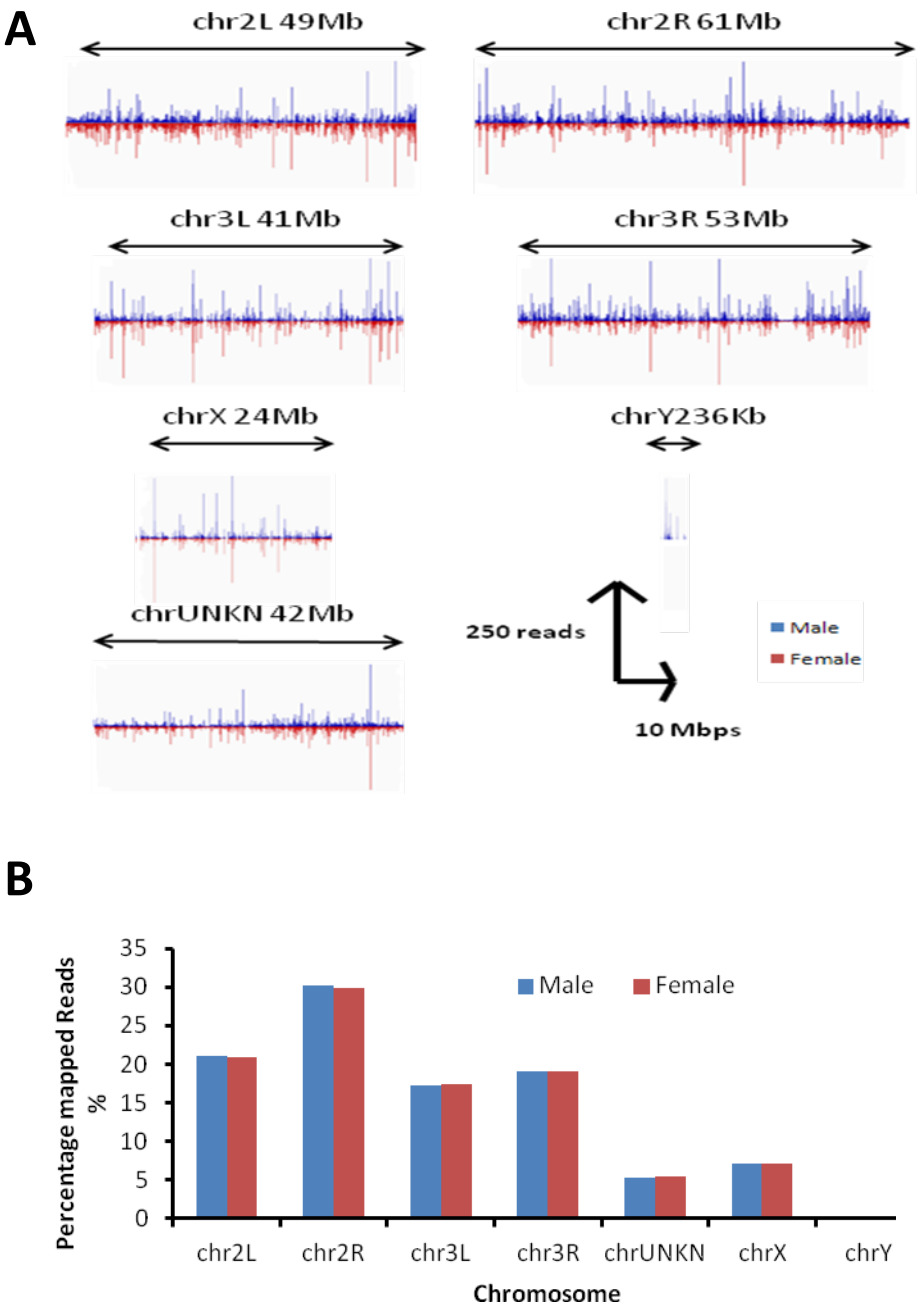


Figure 4.7. Chromosomal distribution of uniquely mapped reads. Alignment of reads to *An. gambiae* chromosomes was performed using GMAP default settings. (A) Wiggle plots of male (blue) and female (red) mapped reads to each of the *An. gambiae* chromosomes. The Y and X axis represent number of mapped reads and relative chromosomal location respectively. A high value on the Y axis represents a higher number of mapped reads.Chromosome length is represented by horizontal arrow and the number of reads by the vertical arrow. Wiggle plots were constructed from GMAP SAM alignment files, converted to WIG files using IGVtools. (B) The percentage distribution of mapped reads across the *An. gambiae* chromosomes. chr = chromosome.

4.2.6 Mapping of individual male and female read datasets to the *An. gambiae* transcriptome

Reads were mapped to the 14,974 transcripts of the AgamP3.6 gene build using Newblerv2.3 gsMapper and this is summarised in Table 4.3. A similar percentage of reads were either fully or partially mapped for the male (63.1 %) and the female (60 %) datasets. Both datasets contained over 85,000 reads that did not align to the transcriptome. A small percentage of reads were too short to be reliably mapped, 3.1 and 4.3 % for males and females respectively. In total the percentage of base pairs mapping to the transcriptome were slightly higher for males compared to females, with 72.9 and 69.6 % respectively. For both sexes, transcripts with a high number of unique mapped reads (>1000) represented predominantly transcripts encoding structural components of ribosomes (Table 4.4). Both sexes had a large number of transcripts with fewer than five unique mapped reads, 3,017 and 2,688 for males and females respectively.

Table 4.3. Summary statistics of individual male and female read datasets mapped to the AgamP3.6 gene build. Reads were mapped to the transcriptome using Newbler for male and female read datasets separately.

	Male	Female
Total number of reads	408,851	353,399
Total number of nucleotides (bp)	131,968,213	122,007,225
Number of mapped reads (% of cleaned reads)	296,938 (72.6)	251,551 (71.2)
Number of mapped nucleotides (bp) (% of cleaned nucleotides)	96,165,157 (72.9)	84,990,569 (69.7)
Fully mapped reads (% of cleaned reads)	169,375 (41.4)	125,070 (35.4)
Partially mapped reads (% of cleaned reads)	88,539 (21.7)	87,011 (24.6)
Unmapped reads (% of reads)	92,261 (22.6)	86,694 (24.5)
Repeat (% of reads)	39,024 (9.5)	32,693 (9.3)
Chimeric (% of reads)	6,839 (1.7)	67,77 (1.9)
Too short (% of reads)	12,813 (3.1)	15,154 (4.3)

Table 4.4. Top hits of AgamP3.6 transcripts with the largest number of unique mapped reads. AgamP3.6 transcripts were prioritised for high numbers of aligned reads and the software BLAST2GO was used to identify the top hit Gene Ontology term. Reads were aligned to the AgamP3.6 transcriptome using Newbler gsMapper default parameters. The number of unique mapping reads and the percentage of mapped reads (from the total read library size) are shown.

AgamP3.6 transcript	Number unique matching reads	Percentage of Mapped Reads	Top Gene Ontology (GO) term
Male			
AGAP007087-RA	5382	0.0181	No associated GO terms
AGAP011802-RA	4972	0.0167	Structural constituent of ribosome
AGAP012334-RA	3312	0.0112	DNA binding activity
AGAP012100-RA	3095	0.0104	Structural constituent of ribosome
AGAP010591-RA	2936	0.0099	Structural constituent of ribosome
Female			
AGAP010461-RA	4616	0.0189	DNA binding
AGAP012100-RA	3933	0.0161	Structural constituent of ribosome
AGAP009634-RA	3634	0.0148	No associated GO terms
AGAP012334-RA	2503	0.0102	DNA binding activity
AGAP006430-RB	1844	0.0075	Sugar binding

4.2.7 Expression levels of AgamP3.6 gene build transcripts by mapped reads

Reads count analysis was performed as described in Materials and Methods. The numbers of unique reads that align to transcripts of the AgamP3.6 gene build represent relative expression levels of the transcripts. In other words, the greater the number of mapped reads, the higher the level of relative expression. Using this criterion of a single fully mapped read being sufficient to deem a transcript expressed, a total of 8,085 and 7,365 transcripts are expressed within the male and female datasets respectively. As shown in Figure 4.8, a small number of, presumably, highly expressed transcripts, are represented by a large fraction of the reads. Over 50 % of the produced reads detect the expression of relatively few (< 5 %) of the total AgamP3.6 transcriptome. Therefore, a small number of abundant AgamP3.6 transcripts represent a large part of the total transcript content of the embryonic stages. As just 10 % of reads align to over 60 % of detected AgamP3.6 transcripts, it also suggests that a large number of low-level expressed transcripts represent a small fraction of the entire transcript content.

4.2.8 Distribution of Mapped Reads along AgamP3.6 transcripts

To investigate the distribution of reads across the length of AgamP3.6 transcripts, from the 5' end to the 3' poly(A) tail; the start, mid and end position of read mapping was extracted from SAM alignment files (example shown in Appendix Table 2). The SAM alignment file provides information regarding the first matching base of a read to an AgamP3.6 reference transcript. This was used as the start site of mapping. Also provided is the number of bases that the read aligns to, therefore the mid and end sites of mapping could be calculated. All reads were mapped to the transcriptome using GMAP as SAM alignment files were required for extraction of mapping information.

A value of zero was set as the 5' end and this indicates that a read maps to the start of a transcript. A value of 1 was set as the 3' end, indicating that reads map to the end of the transcript. Shown in Figure 4.9, an enrichment of reads start alignment at the 5' end of transcripts, 21.9 % of reads mapped to within 10 % of the 5' end of transcripts. An even distribution was found along the transcripts for the midpoint of mapping, increasing from the 5' end to the centre and decreasing towards the 3' end. The end point of mapping was enriched at the 3' end of transcripts, indicated by a higher read count present between the values 0.9 and 1. A mean average coverage of 65 % also indicates that reads map across a majority of the AgamP3.6 transcripts length.

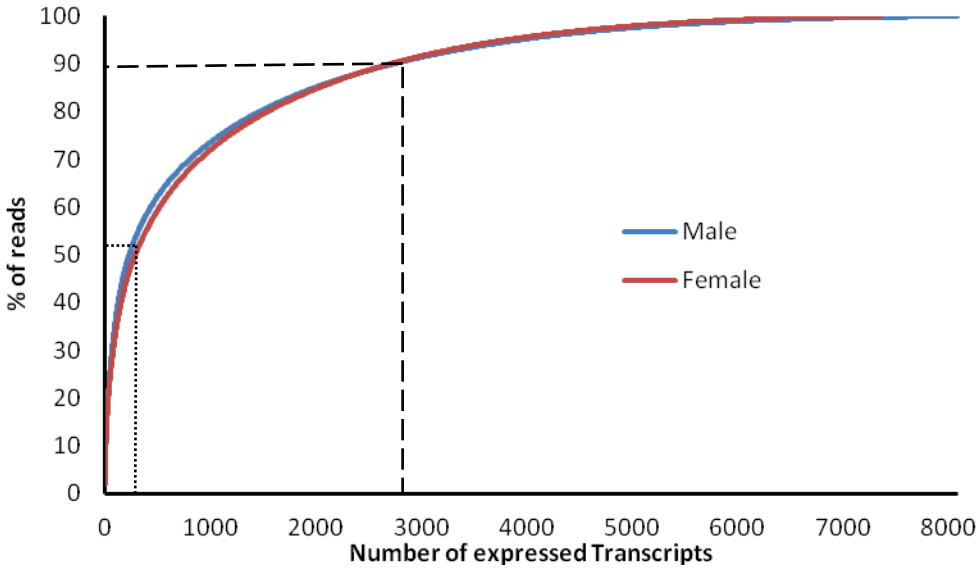


Figure 4.8. Cumulative graph showing the distribution of reads among the AgamP3.6 gene build transcripts. Initially, reads were mapped to transcripts of the AgamP3.6 gene build using Newblerv2.3 gsMapper. Then the cumulative total of reads that align to AgamP3.6 transcripts was calculated as a function of the number of AgamP3.6 transcripts detected. Fifty percent of all reads (dotted line) originate from less than 5 % of AgamP3.6 transcripts. The majority of transcripts are represented by a small number of reads (dashed line).

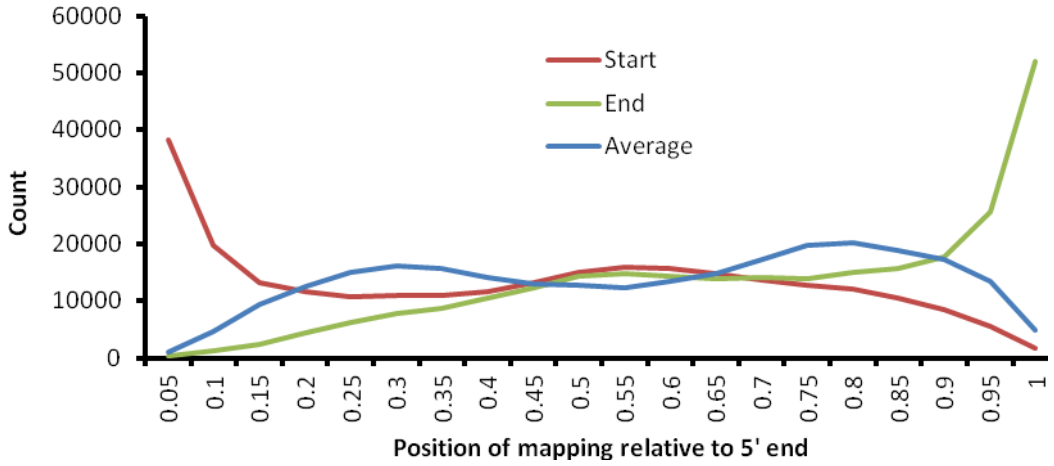


Figure 4.9. Distribution of the start, middle and end position of mapped reads to AgamP3.6 transcripts. Reads were mapped to the *An. gambiae* AgamP3.6 transcriptome using GMAP. Using a value of zero as the 5' start site, reads were classified according to the start position of mapping relative to a 5' zero value, with the 3' end designated 1.

4.2.9 Sequencing effort required to detect AgamP3.6 transcripts

To establish the sequencing effort required to detect AgamP3.6 trasncripts, read count analysis was performed as described in Materials and Methods. Six randomly selected subsets of the total reads (detailed in Fig. 4.10), for males and females, were aligned to the AgamP3.6 transcriptome. This was done to assess the number of reads (sequencing effort) required to detect the expression of AgamP3.6 transcripts. This revealed that 100,000 reads are able to detect the presence of 38.1 % (5,701) and 36.6 % (5,473) of AgamP3.6 transcripts for males and females respectively. Aligning an additional 100,000 reads (200,000 in total) to the transcriptome detects, in total, the presence of 45.6 and 43.8 % of transcripts for males and female respectively. Therefore, 200,000 reads detect a further ~ 7 % of AgamP3.6 transcripts for both males and females, compared to using 100,000 reads. In addition, the lines in Figure 4.10 do not level off, even after sampling >250,000 reads which suggests that the diversity of the embryo transcriptomes have not been exhaustively explored and that a greater number of transcripts are still likely to be detected with further sequencing effort.

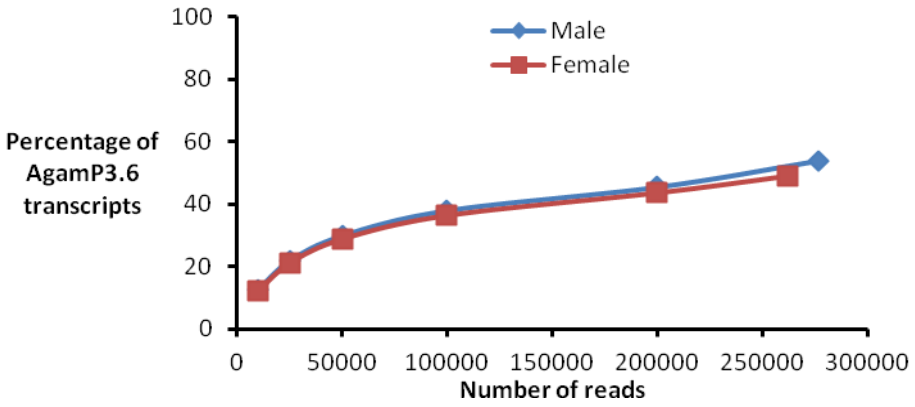


Figure 4.10. Sequencing effort required to sample the AgamP3.6 transcriptome. Read count analysis was performed using Newbler gsMapper to establish the number of AgamP3.6 transcripts that are detected when random subsets of reads are mapped to the AgamP3.6 transcriptome. The total numbers of reads contained within the subsets were 10,000, 25,000, 50,000, 100,000, 200,000 and all reads. As the entire transcript content of the AgamP3.6 transcriptome (14,974 transcripts) would not be expected to be present in the embryonic stages, the two datasets for males and females do not approach 100 %. The final point on the graph indicates the total number of reads generated in this study for males and females that map to the transcriptome. A steep initial rise, followed by a plateau, suggests an increase in sequencing effort (number of reads) is required to detect all expressed transcripts.

4.3Discussion

By utilising a diagnostic PCR, it has been possible to separate *An. gambiae* embryos using an approach similar to that used in *C. capitata* (Gabrieli *et al.* 2010). It has allowed the large scale accumulation of sex-specific pools of RNA and subsequent 454 transcriptome sequencing. Over 300 Mbps of sequence data was produced, with over a million raw reads constituting male and female datasets from a single 454 Picotiter Plate. These datasets represent the first study of sex-specific expression in *An. gambiae* embryos. Previously, sex-specific transcriptional profiling has been restricted to NGS analysis of adult stages (Pitts *et al.* 2011) and post-embryonic stage microarray studies, that are restricted to expression profiling of known gene annotations (Warr *et al.* 2007; Baker *et al.* 2011; Magnusson *et al.* 2011).

cDNA fragmentation carried out during this experiment, as opposed to RNA fragmentation prior to cDNA synthesis, is known to induce a 3' bias upon transcript mapping of oligo-dT primed cDNA (Wang *et al.* 2009). To establish if a bias has been introduced, the position of each base pair from reads can be mapped to the length of reference transcripts. Though we have not mapped the position of each base within our study, the start, middle and end location of mapping has been analysed. This suggested a slight bias of the midpoint of mapping to the 3' end of transcripts, though using this approach, we cannot conclusively state that we have preferentially sampled the 3' of transcripts. The SMARTer cDNA synthesis protocol used enables transcription of full length cDNA and has been shown to produce even coverage across transcripts entire length (Levin *et al.* 2010). Our study also suggests that the technology does enrich for full length transcripts as we detected a high number of reads with mapping initiating at the 5' end. Also, amplified cDNA, fractioned on an agarose gel, produced a smear across a wide range of fragment sizes (100 bp - 5 Kb), suggesting a good representation of full length transcripts. The high transcript coverage is vital for detection of sex-specific alternative splicing events (described further in the thesis). Inadequate coverage would lead to a large number of false positives and, more importantly, may result in missing actual sex-specifically spliced transcripts.

Reads were mapped to the genome to assess their distribution. Uniquely mapped reads (96.8 %) appear to be evenly distributed across the autosomes, X chromosome and unplaced scaffolds for both males and females. Even distribution of autosomal transcripts has previously been shown in sexed embryos of *D. melanogaster* using Illumina sequencing of individual embryos (Lott *et al.* 2011). The percentage of reads mapping to each of the

chromosomes does not markedly differ, except for male only mapping to the Y chromosome. Over 5 % of reads for both males and females map to the chrUNKN scaffolds that have not been placed on any of the chromosomes. Identification of chimeric reads that map to chrUNKN scaffolds as well as an additional location on a separate chromosome may aid in placing these sequences on to their respective chromosomes. Though a number of false positives may be expected using this approach as chimeric reads can be generated following cDNA amplification (Zhang *et al.* 2006).

In total, 8,755 of the currently annotated 14,974 AgamP3.6 transcripts had at least 1 uniquely mapped read in either male (8,085) or female (7,365) datasets. A higher number may be expected in males, due to the larger number of reads within this dataset compared to females. In addition, female cDNA was amplified twice (due to lower concentration of mRNA used for first strand cDNA synthesis), compared to once for males and may account for differences in transcript abundance, preventing firm conclusions being drawn regarding variance in the entire transcript content of male and female transcriptomes. Previously, sequencing of the *D. melanogaster* developmental transcriptome demonstrated embryonic stages expressed far fewer transcripts (7,045) when compared to adult males (12,000) (Graveley *et al.* 2011). This is largely consistent with our data as previous sequencing of alter An. gambiae stages detected a higher number of transcripts than we observed for the embryonic stages. *de novo* Solexa (now Illumina) sequencing (21 million reads) of adult *An. gambiae* transcriptome detected the expression of 95 % of the AgamP3.5 transcriptome, with at least one uniquely mapped read (Gibbons *et al.* 2009). Illumina sequencing (average 30.5 million reads per tissue) of male and female antennae, maxillary palps and whole bodies detected expression of over 12,000 (> 90 %) genes of the AgamP3.6 gene build (Pitts *et al.* 2011).

We have also suggested that the sequencing effort required to detect additional transcripts increases as more reads are produced. This issue has been experienced previously with 454 derived reads (Robert *et al.* 2011). Therefore, though we cannot conclude by how much, it is likely that the entire embryonic transcriptome has not been sampled, but this would be expected when comparing the smaller read library size of 454 sequencing (~1 million) to that of the Illumina platform (~ 20-30 million). We have apparently sampled a large majority of the embryo transcriptome and a considerable further sequencing effort would be required to encompass its full diversity.

The use of short read technologies, producing a greater number of reads compared to the 454 platform, may still be insufficient for full transcriptome assessment. Recently, Illumina sequencing of human B-cells suggested that 500 million reads would be required for accurate detection of all gene expression (Toung *et al.* 2011). Toung *et al.* (2011) used Illumina sequencing of 20 different samples, using an individual lane per sample (there are eight lanes per Illumina flow cell) and the costs involved in using their approach were beyond the scope of our project. Similarly, sampling of 30 *D. melanogaster* development stages failed to identify expression of over 1,000 annotated genes, despite over 4 billion sequencing reads from Illumina, SOLiD and 454 platforms, reflecting the depth of sequencing required to fully characterise a transcriptome (Graveley *et al.* 2011). It is therefore highly likely that a number of embryonic transcripts have not been detected in this study.

What is striking from our datasets it that a relatively high proportion of the reads did not map to previously annotated transcripts (22.6 and 24.5 %) for males and females respectively, but did map to the *An. gambiae* genome, suggesting that a large proportion of the mosquito transcriptome remains to be annotated. The mapping of reads to known intronic regions also suggests that previously unannotated alternative splice forms exist within the current gene build. Previous RNA-seq experiments have suggested that over 90 % of multi-exon genes have more than one splice form and intronic reads may represent additional *An. gambiae* splice forms (Wang *et al.* 2008; Toung *et al.* 2011). In a recent study, an average 42.6 % of reads from various adult tissues did not map to annotated genes but did align to the genome (Pitts *et al.* 2011). The large percentage of transcripts with partial matches to known transcripts suggests that the two datasets may also extend the current annotation of some transcripts.

Partially mapped reads align to a mean 79.5 % of AgamP3.6 transcript length. This suggests that these reads may represent extensions to currently annotated transcript exons or in some cases, novel exons. To validate such extensions, it must be confirmed that these reads map to the genome across their entire length and that partial alignment is not due to poor quality sequences at the extremities of the reads. This information will be important to improve the annotation of the *An. gambiae* genome.

We have shown that a large proportion of the 454 reads represent few, relatively abundant, AgamP3.6 transcripts. Therefore identification of weakly expressed transcripts, including key developmental genes, may prove problematic. Despite this, a previous study

has shown the feasibility of identifying developmental genes involved in sex determination from 454 sequencing (Olafson *et al.* 2010), although normalised cDNA from pooled life stages, including embryonic, were used in that experiment. Normalising cDNA prior to sequencing reduces the representation of highly abundant transcripts and enables improved detection of rare transcripts (Cheung *et al.* 2006). As a large fraction of the reads in our study detect the expression of a small subset of highly abundant transcripts, for future work, cDNA normalisation would be recommended for the detection of rare transcripts, of which sex determination genes may be part of. Within our study, many of the AgamP3.6 transcripts, over 35 %, for both sexes had less than 5 mapped reads. A low number of mapped reads will greatly reduce the confidence at which biological events such as alternative splicing or sex-specificity can be identified from this sequence data. For example, a candidate sex-specific gene with only a single read may represent two scenarios. Firstly, the gene is sex-specific but expressed at very low levels. Or secondly, the transcriptome has not been sampled at a sufficient depth and the transcript in question is actually present in both sexes. A candidate gene with 100 reads provides more confidence that it is sex-specific, and not a consequence of poor sequencing coverage.

The massively parallel sequencing of sex-specific embryonic RNA has produced two large scale RNA-seq datasets and has demonstrated that sufficient RNA could be collected from single embryos for transcriptome sequencing. Reads have been successfully mapped to both the *An. gambiae* genome and transcriptome. The most striking outcome is the large number of reads that map to intergenic regions, supporting the idea that the current gene build of *An. gambiae* is far from complete. These intergenic reads may represent extensions of known transcript exons or correspond to either coding or non-coding new transcribed regions (NTRs). A fraction of these reads may also potentially originate from genomic DNA contaminating the RNA samples, although the possibility is small, because the samples were treated with DNase prior to first strand cDNA synthesis and the first strand synthesis reaction was primed with an oligo-dT primer. Further characterisation of all intergenic reads would be required before conclusions regarding their function can be drawn.

Assembly of all reads into contigs prior to functional characterisation will be required. A reference based assembly could be undertaken for the male and female datasets, following mapping to either the transcriptome or genome. Alignment to the transcriptome would prevent the assembly of intergenic reads into contigs and therefore will not be undertaken.

Similarly, as the genome reference strain (PEST) is different to the strain used for RNA extraction and sequencing (G3), a genome reference assembly will not be used due to potential sequence variations between the two strains. The PEST strain is a chimera of M and S molecular forms and is a hybird of a lab-colonised strain originating from Nigeria crossed with a wild caught Kenyan population (Holt *et al.* 2002). Whereas the G3 strain was colonised from Gambia in 1975 and is also a hybrid of M and S forms (Collins *et al.* 1986). To avoid potential mis-alignment and incorrect assembly, a *de novo* assembly will therefore be undertaken.

Chapter 5: Characterisation of *Anopheles gambiae* embryo transcriptional architecture and identifying sex-specific transcription patterns

5.1 Introduction

The hundreds of thousands of reads produced from the 454 platform in chapter 3 represent a snapshot of the *An. gambiae* embryo RNA landscape. This large sequence data must first be processed to either assemble or align reads to a reference (genome or transcriptome), prior to retrieval of biological information. Alignment identifies regions of similarity to a reference to track gene expression, whereas assembly allows the construction of longer continuous sequences in the form of contigs and isotigs that represent putative co-transcribed exons and transcripts respectively. This study benefits from the opportunity to assemble both sex-specific and non-sex-specific transcript libraries.

Assembly of next generation sequencing reads can either be *de novo* (no reference required), reference based (reads are aligned to a reference genome prior to assembly) or a combination of the two (Martin and Wang 2011). Commonly used software packages for 454 data have been compared previously (Kumar and Blaxter 2010). Assessment of the *de novo* programs CAP3 (Huang and Madan 1999), MIRA (Chevreux *et al.* 2004; Chevreux 2011), Newblerv2.3/v2.5 (Margulies *et al.* 2005), SeqMan (Swindell and Plasterer 1997) and CLC (CLCBIO) suggested that choice of program is largely dependent upon the data set being assembled. It also suggested that any unassembled reads (termed singletons) be assembled using a second assembly software. One strong advantage of the Newbler package is assembly of isotigs that represent putative full length transcripts and this process is summarised in Figure 5.1. Newbler software also attempts to resolve conflicting information within homopolymer (runs of the same nucleotide) regions.

Mapping of reads to a reference is possible using various software for either short read (Illumina and SOLiD) or long read (454) platforms. Long read aligners include Newblerv2.3 gsMAPPER (Roche supported software) (Margulies *et al.* 2005), BWA-SW (Li and Durbin 2010) and GMAP (Wu and Watanabe 2005). Alignment to a reference produces a SAM (Sequence Alignment/Map) alignment file and from this it is possible to extract information about novel transcripts and splice sites (Bonizzoni *et al.* 2011), alternative splicing events (Sorber *et al.* 2011) and differential gene expression (Gregory *et al.* 2011; Swarbreck *et al.* 2011). SAM alignment files contain mapping information within a CIGAR string that details insertions, deletions and skipped regions to the reference. During the alignment process, GTF (Gene Transfer Format) or GFF (General Feature Format) annotation files can be

provided that detail attributes (e.g. genomic co-ordinates) of annotated transcripts, to document expression of known genes. An example of the GTF file format is shown in Appendix II Table 1. Both GMAP and BWA-SW are gapped read aligners and therefore take into account introns during the mapping process, a key feature for long 454 reads. SAM alignment files can be visualised using graphical user interface (GUI) software such as Integrative Genomics Viewer (IGV) (Robinson *et al.* 2011), Tablet (Milne *et al.* 2010) and Bambino (Edmonson *et al.* 2011). Visualisation allows comparison of reads aligned to a reference and the GTF annotation file for detection of differential expression, alternative splicing and single nucleotide polymorphism discovery. In addition, web-based genomics views can be utilised, such as those provided at VectorBase.

For functional annotation of transcriptomes, Gene Ontology (GO) terms can be assigned to sequences using the software BLAST2GO (Conesa *et al.* 2005). GO terms are a controlled vocabulary for standardised description of biological functions of gene products (Ashburner *et al.* 2000). Previous high-throughput sequencing studies have isolated transcription from un-annotated genomic locations and BLAST2GO is able to handle large scale RNA-seq data to assign putative functions to new transcribed regions (Poelchau *et al.* 2011; Price *et al.* 2011). In addition, BLAST2GO software can be used to run InterProScan, searching vast database resources for identification of protein domains to infer putative gene product functions (Gotz *et al.* 2008; Zhang *et al.* 2011).

Previous studies of *An. gambiae* gene expression identified an average 42.6 % of adult tissue derived Illumina reads that did not map to the reference transcriptome (AgamP3.6), suggesting presence of new transcribed regions (NTRs) or perhaps extensions of gene untranslated regions (UTRs) (Pitts *et al.* 2011). A similar situation has been observed from sampling 27 different stages of *D. melanogaster* development, where 1,938 NTRs and 22,965 new splice junctions were identified (Graveley *et al.* 2011).

Currently 14,974 transcripts are annotated in the *An. gambiae* transcriptome (AgamP3.6) and of these, 95.47% are protein coding, with the remaining consisting of snRNA, tRNA and other non-coding sequences. Of the AgamP3.6 transcripts, 37.6 % are computationally predicted without manual annotation, and 11,884 are supported by at least a single EST from the AgamESTDec10 database (containing 217,297 ESTs). Manual confirmation of annotation is often used to detect errors and to identify alternative splicing events as computational gene prediction can often fail to annotate alternative splice forms and 5' and 3' untranslated regions (Rogic *et al.* 2001; Florea *et al.* 2005; Sleator 2010). Recently,

over 90 % of multi-exon genes have been reported to contain more than one alternatively spliced transcript (Toung *et al.* 2011). *De novo* sequencing of the embryonic transcriptome, that has not been sampled previously, may therefore be expected to identify *An. gambiae* NTRs or further extend annotations of untranslated regions or additional exons.

High-throughput sequencing allows assembly of transcripts that may reveal alternative splicing between the male and female transcriptomes. Alternative splicing plays a key role in the sex determination cascade in insects (Salz 2011). Recently, RNA-seq data, including that generated by the Illumina platform (75 bp reads from 30 biological samples) has demonstrated that adult sex-biased splicing of key *D. melanogaster* sex determination genes including *sxl*, *tra* and *dsx* can be identified through transcriptome sequencing (Graveley *et al.* 2011). Detection of alternative splicing between RNA-seq datasets is possible through programs such as Cufflinks (Trapnell *et al.* 2010) and MISO (Katz *et al.* 2010) that require input of a SAM alignment file or HMMSplicer (Dimon *et al.* 2010) that requires input of NGS reads of equal length (therefore generated from either Illumina or SOLiD platforms).

In addition to this, it is possible to identify differential gene expression between two experimental conditions via reads per kilobase of exon model per million mapped reads (RPKM) calculation. This process normalises gene expression levels from read counts (number of mapped reads) to the size of the RNA-seq dataset to allow levels of transcription to be directly comparable between sequence data from different biological conditions (Mortazavi *et al.* 2008). NGS has been shown to be highly comparable to microarrays for dissecting gene expression levels (Marioni *et al.* 2008). Software such as edgeR and DEGseq can also be used to identify differentially expressed genes based upon non-normalised read count data (Robinson *et al.* 2010; Wang *et al.* 2010).

5.1.1 Specific Chapter Aims

- *de novo* assembly of trimmed 454 reads into three datasets: combined assembly (male and female together) and male and female separately.
- Probe the current AgamP3.6 annotation with assembled sequences to identify new transcribed regions and extensions to current gene boundary annotations.
- Functionally characterise the assembled datasets through bioinformatics analysis to include gene ontology annotation and InterProScan searches.

- Compare male and female assembled datasets to isolate sex-specific differences in differential gene expression and alternative splicing events.

5.2 Results

5.2.1 Non-sex specific *de novo* assembly of male and female reads

In total, 762,250 reads were assembled using Newbler, remaining singletons were then assembled using MIRA, as shown in Figure 5.1. The contig N50 value was 1,018 bp (Fig 5.2A) and N50 is defined as the length of the contig such that 50 % of the total nucleotides of the experiment are contained within contigs longer than this value. Additionally, assembly via Newbler resulted in the construction of 12,700 isotigs. To assess whether isotigs actually represent full length transcripts, a ratio of mapped isotig sequence length to AgamP3.6 transcript length was calculated (Fig. 5.2B). Though not likely, it was assumed here that AgamP3.6 transcripts are full length sequences. 56.1 % of isotigs map to over 50 % of the annotated transcript and 44.3 % represent over 75 % of the transcript sequence. A total of 89 isotigs had a value greater than one, indicating that these isotigs carry an insertion or an extension within the annotated transcript.

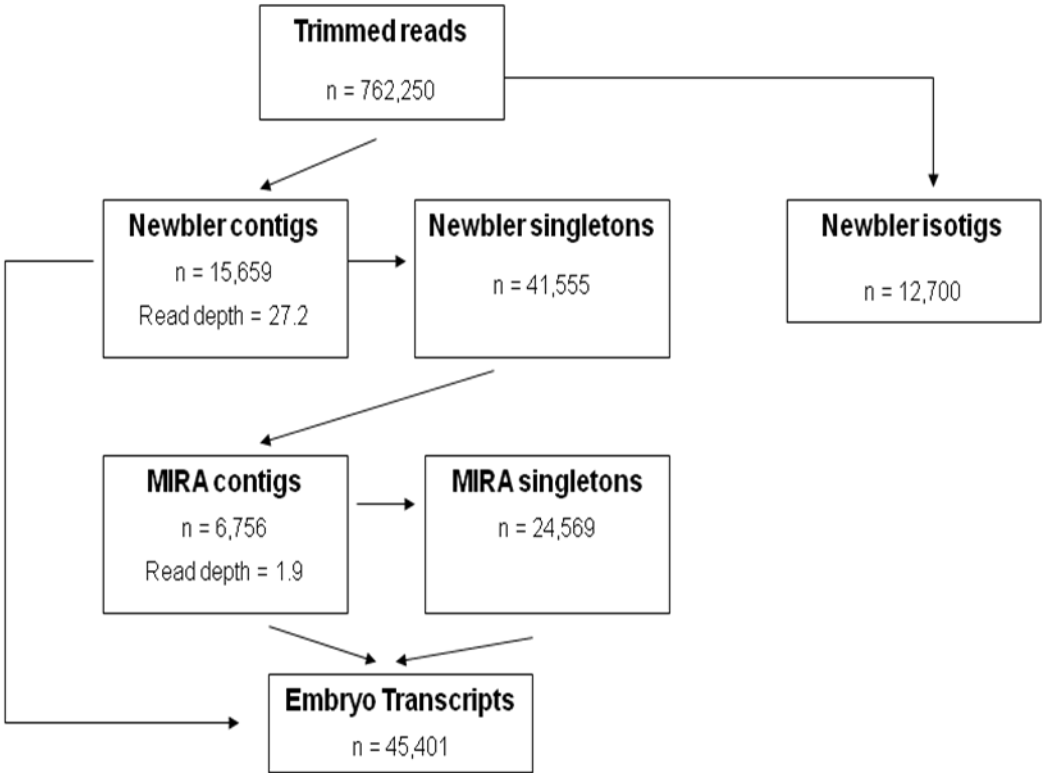


Figure 5.1: Flow chart showing methodology and results of contig and isotig assembly. Trimmed reads from chapter 3 were used for contig assembly using Newbler and then unassembled singletons were extracted from the 454Readstatus file and assembled using MIRA. This resulted in the assembly of a total 45,401 embryo transcripts representing the embryonic transcriptome. In addition Newbler was used for the assembly of isotigs that represent putative full length transcripts.

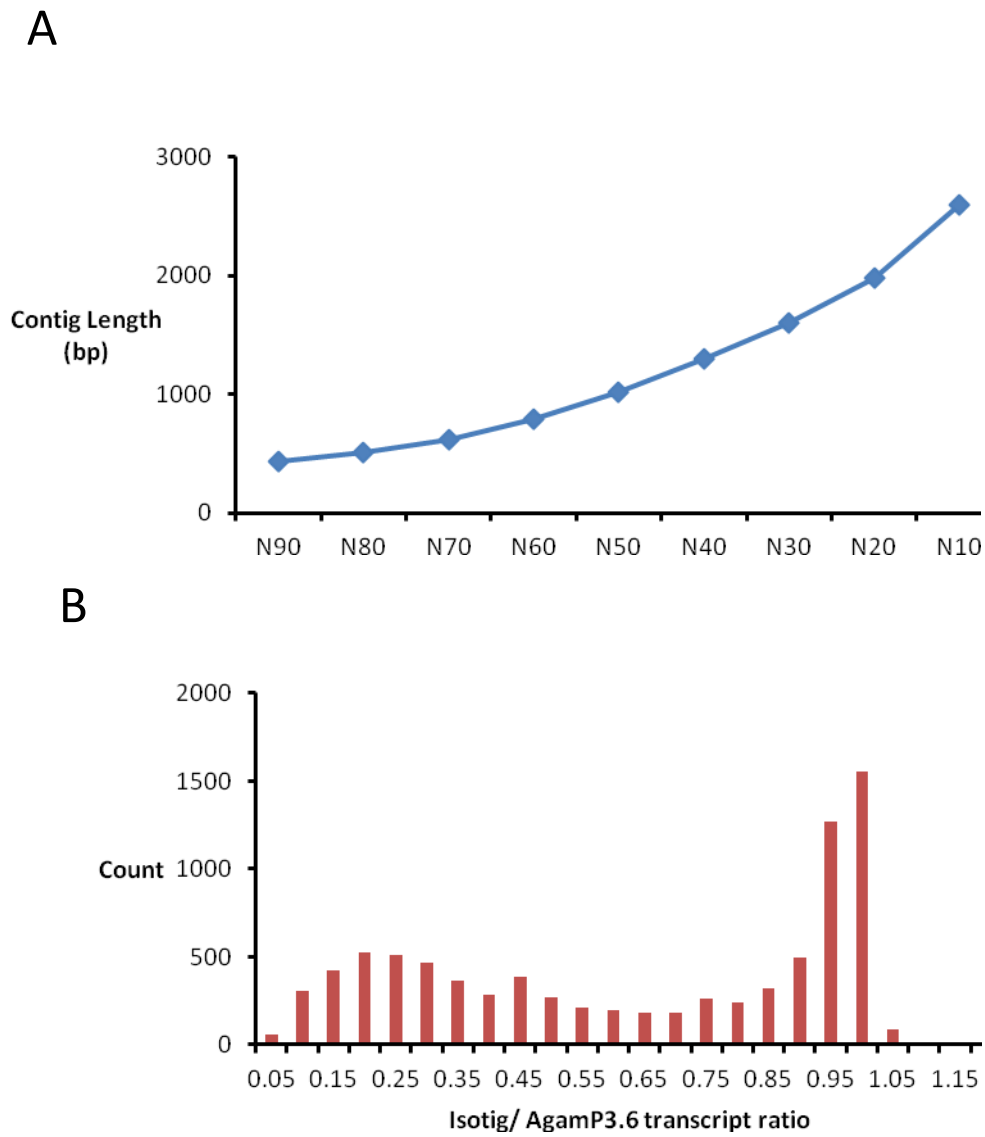


Figure 5.2. Summary of contig N10-90 values and assessment of the representation of full length transcripts by isotigs. (A) N10-N90 values of assembled Newbler and MIRA contigs. N50 is defined as the length of the contig such that 50 % of the total nucleotides of the experiment are contained within contigs larger than this value. Lengths of contigs were extracted from fasta files using perl script length.pl and N values calculated following sorting into ascending order. (B) The ratio of the length of isotig mapping to the length of the annotated AgamP3.6 transcript it maps to. A value of 1 indicates that the isotig covers the entire AgamP3.6 transcript that is maps to. A value greater than one indicates that these isotigs carry an insertion within the annotated transcript.

5.2.2 Detection of non-*An. gambiae* derived sequences

To assess the extent of potential contamination (non-*An. gambiae* derived sequences) found within the transcript library, embryo transcripts were used to query (BLASTN E-value = $1e^{-5}$) the *An. gambiae* PEST genome (AgamP3), both with and without the DUST filter (allowing for matches with low complexity sequences) removed and is summarised in Table 5.1. All embryo transcripts without a hit to any of the *An. gambiae* genomes searched were used for BLASTX comparison to the nr protein database. Only 15 embryo transcripts had a BLASTX match (E-value $-4e^{-6}$ to $6e^{-45}$) to the nr database proteins. Seven of the embryo transcripts matched insect proteins and five matched bacterial proteins. Of particular interest, embryo transcript GLONWJ301C0UR3 matched a protein from *Elizabethkinga anophelis*, and BLASTN searches against the genome of *E. anophelis* (GCA_000240095.2) using GLONWJ301C0UR3 as query revealed a perfect match through most of the transcript's length. *E. anophelis* is a recently described novel species belonging to the Flavobacteriaceae family, identified from *An. gambiae* midgut (Lindh *et al.* 2008; Kampfer *et al.* 2011) and its presence in the embryo may suggest it is transovarially transmitted, or, potentially, may have been sampled from the water in which the eggs were deposited. The four other embryo transcripts with hits to bacterial proteomes were used for BLASTN searches against the *E. anophelis* genome but failed to produce significant hits (E-value $1e^{-5}$). Lack of hits for other transcripts to the *An. gambiae* genomes may result from an incomplete genome sequence coverage or a combination of short length and insufficient quality of the query sequences.

Table 5.1: Summary of non-*An. gambiae* derived sequences. The level of contamination within the embryo transcript dataset was assessed through BLASTN (E-value $1e^{-5}$) searches against available *An. gambiae* genome databases. Transcripts without a significant hit were then compared (BLASTX) to the nr protein database (E-value = $1e^{-3}$)

	Number of Sequences
No BLASTN hit to AgamP3 (DUST off)	2,082
(% of transcripts)	(4.6)
No BLASTN hit AgamP3 (DUST on)	1,508
(% of transcripts)	(3.3)
BLASTN hit *An. gambiae* M form scaffold	111
(% of transcripts with no BLASTN hit)	(0.24)
BLASTN hit *An. gambiae* S form scaffold	94
(% of transcripts with no BLASTN hit)	(0.21)
BLASTN hit *An. gambiae* mitochondrial genome	142
(% of transcripts with no BLASTN hit)	(0.31)
BLASTX hit nr protein database	15
(% of transcripts with no BLASTN hit)	(0.003)

5.2.3 Identification and characterisation of *An. gambiae* new transcribed regions

Here, new transcribed regions (NTRs) are defined as sequences that do not match a transcript of the AgamP3.6 transcriptome but do map to the AgamP3 genome. BLASTN (E-value = $1e^{-10}$) searches identified 26,616 such embryonic transcripts. In anticipation that the current gene build is not exhaustive and that a number of the sequences may represent extensions rather than NTRs, the 26,616 embryo transcripts were used to align to AgamP3.6 transcripts extended to include 10 Kb upstream and downstream genomic regions. In total, 7,284 embryo sequences did not produce a significant hit. Nucleotide searches (BLASTN E-value = $1e^{-10}$) against an *An. gambia*e EST database (AgamESTDec10 – 217,297 sequences) revealed that 848 (11.6 %) NTRs had a previously reported EST. The remaining 6,436 NTRs, with no EST match, were then further characterised.

All NTRs were used to query the proteomes and genomes of other mosquito species. Hits to sequences from other mosquitoes could indicate functional conservation of NTRs and would provide greater evidence they do represent new genes. BLASTX (E-value = 1e-5) searches showed that 254 (3.9 %) unique NTRs had a significant match to a protein database of *Aedes aegypti* (AaegL1.2 – 148 matches), *Culex quinquefasciatus* (CpipJ1.2 – 139 matches) and *An. darlingi* (NCBI accession PRJNA46227 – 181 matches). In addition, 5,559 (86.1 %) and 3,315 (51.5 %) NTRs have predicted ORFs greater than 60 and 120 nucleotides respectively. TBLASTN searches were also conducted as a more sensitive method than BLASTX searches, because potential homologues may not be present in the respective predicted proteomes. All ORFs longer than 120 nt were used for TBLASTN (E-value = $1e^{-6}$) searches against the whole genome shotgun reads of *An. darlingi* (NCBI PRJNA46227), *Cx. quinquefasciatus* (CqSC) and *Ae. aegypti* (AaSC). A total of 581 (17.5 %) unique NTRs had a significant hit to *An. darlingi* (404), *Cx. quinquefasciatus* (368) and *Ae. aegypti* (412).

To determine the number of splice sites within NTRs and hence novel-splice sites, a database of all known AgamP3.6 gene build splice sites (104,504) was created as described in Materials and Methods. NTRs and AgamP3.6 transcripts were then mapped to the genome together using GMAP default settings. GSNAP was then used to identify novel splice sites, not annotated in the AgamP3.6 genebuild using default parameters. From this approach, over 2,000 new splice sites were shown to be present, not previously annotated in the current genebuild.

To confirm that assembled NTRs represent continuous sequences and had not been incorrectly assembled, RT-PCR amplification for selected NTR sequences was carried out with mixed-sex embryonic RNA (18 hr AEL). Primers were designed for 10 NTRs that do not align to an AgamP3.6 transcript. Successful amplification was achieved for seven of the NTRs and of these, six were confirmed following cloning and sequencing of the amplified fragments (Table 5.2).

To infer putative functions of the NTRs, sequences were used to identify similarity to known proteins using the software BLAST2GO. BLASTX (E-value = $1e^{-3}$) searches against the NCBI nr protein database revealed 5.4 % of NTRs had a significant hit. Selected NTRs and their respective top BLASTX hit and GO terms are shown in Table 5.3. InterProScan was also used to search InterPro databases to classify functions of putative gene products. In total, 98 NTRs, of which 91 also had an assigned GO term, were associated with 71 InterPro terms. Signal P software detected the presence of putative signal peptides within 2,179 NTRs. In addition, 1,305 NTRs were assigned transmembrane domains using the TMHMM program (transmembrane helices based on a hidden Markov model). All functional annotations to NTRs are included within the attached CD.

Table 5.2. Summary of RT-PCR amplification for confirmation of NTR contig assembly. Ten randomly selected contigs that did not match an AgamP3.6 transcript were amplified from mixed-sex embryonic RNA (18 hr AEL) via RT-PCR. Successfully amplified products were cloned and sequenced to confirm correct amplification of the selected contig. * indicates MIRA assembled contig. The remaining contigs were assembled via Newbler.

Contig Name	RNA size (bp)	DNA size (bp)	Successful amplification	Sequence confirmed
contig06638	3635	4041	YES	NO
contig08915	2260	2859	YES	YES
contig09569	1745	3725	YES	YES
contig12387	1581	4521	YES	YES
contig05828	3029	3234	YES	YES
singletons_rep_c6734*	711	21924	YES	YES
singletons_c5769*	684	779	YES	YES
contig07915	2062	26257	NO	NO
contig08484	312	433	YES	YES
contig13216	3799	11161	NO	NO

Table 5.3. Summary of selected new transcribed regions with BLASTX hits against the nr protein database. NTRs were used for as query for BLASTX (E-value = $1e^{-3}$) searches against the nr protein database and nucleotide BLASTN (E-value = $1e^{-10}$) searches against the AgamESTDec10 database. Shown is the top BLASTX hit and associated GenBank accession number, top GO term from BLAST2GO software and top hit to *An. gambiae* ESTs.

Contig Name	Length (bp)	Top BLASTX hit (species)	GenBank Accession Number	Top hit Gene Ontology term	e-value of top hit	Top hit to *An. gambiae* EST (E-value)
contig04292	938	Amino acid transporter (*Ae. aegypti)*	XP_001648128	amino acid transmembrane transporter acitivity	$3e^{-176}$	AL694660.1 0
contig01773	999	Gelsolin (*Cx. quinquefasciatus)*	XP_001846164	actin binding	$3e^{-163}$	BM634197.1 0
contig06890	718	Transcription factor castor (*Ae. aegypti)*	EAT43176	Zinc ion binding	$1e^{-119}$	N/A
contig06850	1841	Wd repeat protein (*Ae. aegypti)*	XP_001652768	N/A	$1e^{-106}$	CR527449.1 0
contig08674	438	Sentrin/sumo-specific protease (*Ae. aegypti)*	XP_001653624	SUMO-specific protease activity	1e-68	AJ285126.1 0
singletons_c4499	430	Sodium/chloride dependent transporter (*Ae. aegypti)*	XP_001652213	nuerotransmitter: sodium symporter activity	$3e^{-68}$	N/A
singeltons_c122	525	CG34350 (*D. melanogaster)*	NP_610437	serine-type peptidase activity	$7e^{-45}$	N/A
contig11133	296	Enhancer of yellow 2 (*Cx. quinquefasciatus)*	XP_001847707	transcription coactivator activity	$1e^{-26}$	BM601128.1 (-145)
singletons_c1548	479	Female lethal d, isoform B (*D. melanogaster)*	NP_725327	primary sex determination, soma	$4e^{-24}$	BX016212.1 0
contig04083	827	Werner helicase interacting protein (*Ae. aegypti)*	XP_001647950	helicase activity	$1e^{-21}$	AJ285292.1 0

The NTR sequences were mapped to the AgamP3 genome using GMAP and a selection of those were visualised using the Integrative Genome Viewer (IGV, Broad Institute), along with a track representing the AgamP3.6 GTF annotated transcripts. Due to the high number of NTRs (6,436), only a 25 were visualised using IGV. Figure 5.3 illustrates two NTRs that correspond to intron-containing transcripts. Contig07351 (Fig. 5.3A) maps to chromosome 2L (5,251,107-5,251,844) and is composed of two exons. The second exon contains a gap of 15 bp which may represent an indel compared to the *An. gambiae* PEST reference genome used in this study, because no cannocial splice sites are found within the boundaries of the gap to indicate its status as an intron. BLASTX (E-value = $1e^{-5}$) searches using contig07351 as query against the non-redundant (nr) protein database does not reveal any significant matches. TBLASTX searches (E-value = $1e^{-5}$) conducted against the genomes of *Ae. aegypti, An. darlingi* and *Cx. quinquefasciatus*using the same contig as query revealed no sequence similarity in the respective genomes. Contig11843 (Fig. 5.3B) maps to chromosome3L (7,802,134-7,803,388) and is predicted to contain at least 3 exons, BLASTX searches (E-value = $1e^{-5}$) against the nr protein database revealed no significant hits. In addition, TBLASTX searches revealed no significant similarity when compared to the genomes of *Ae. aegypti, An. darlingi* and *Cx. quinquefasciatus* (E-value = $1e^{-5}$). A single read crosses the first predicted intron and this may represent an alternative splicing event. RT-PCR amplification of the contig would be required to confirm the presence of alternatively spliced transcripts within this contig NTR. Following alignment of contig07351 and contig11843 to the genome, analysis of splicing signal sequences (GT/AG) at the termini of predicted introns confirmed canonical splicing for both contigs.

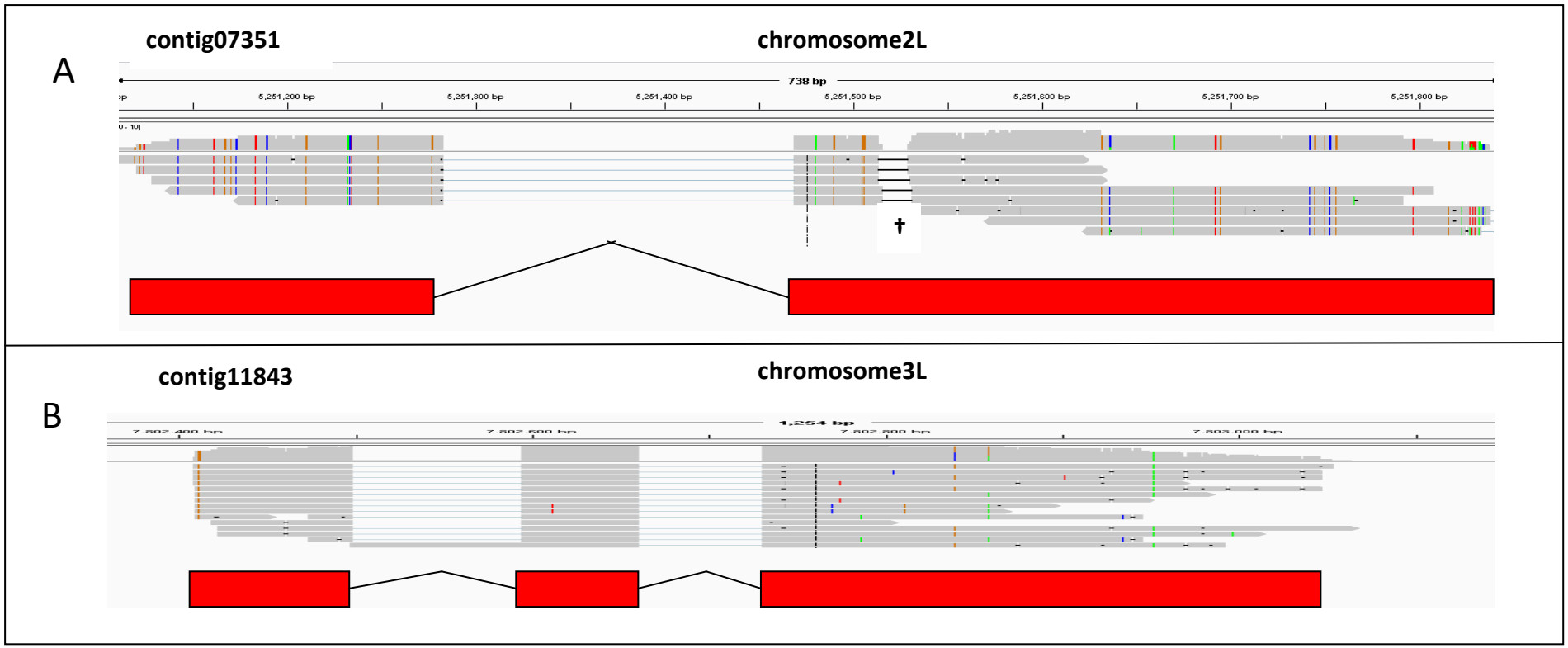


Figure 5.3. Schematic diagram of new transcribed regions (NTRs) mapped to the *An. gambiae* genome. All NTRs were mapped to the AgamP3 genome using GMAP and then the resulting SAM files visualised using IGV. (A) Contig07351 is located at chr2L:5,251,107-5,251,844 and is comprised of at least two exons. † = indicates a 15 bp gap in the reads to the *An. gambiae* PEST reference used. This gap does not contain canonical splice sites indicating that it is not an additional intron. (B) A novel transcript identified by reads contained within contig11843 at the location chr3L:7,802,134-7,803,388. A single read covers the first intron and this may represent an additional splice variant. Grey thin boxes represent individual reads contained within the two contigs. Coloured vertical lines within the individual reads represents single base differences compared to the reference (Green = adenosine; Blue = cytosine; Red = thymine; Brown = guanine; Purple = insertion to the reference. Horizontal dark lines – skipped region to the reference. Both NTRs have canonical splicing signal sequences (GT/AG) and their predicted exon/intron structure is shown by red boxes and diagonal black lines respectively.

5.2.4 NTR isotigs containing insertions to AgamP3.6 anntoated transcripts

Isotigs were used to identify possible insertions to AgamP3.6 transcripts by mapping to the AgamP3.6 transcriptome via GMAP, and the resultant SAM alignment file sorted using custom perl script insertions.pl (appendix). In brief, the script extracts the CIGAR string, detailing mapping information of matches, insertions and deletions to the reference from the SAM alignment file. Insertions to the AgamP3.6 reference transcripts were extracted by searching the CIGAR string for the letter I (denoting insertion) preceded by a value greater than 10 (nucleotides). A total of 641 isotigs that contained an insertion greater than 10 bp in length were returned. The length distribution of the insertions is shown in Figure 5.4. A subset of NTR isotigs with insertions were visualised using IGV and this suggested that insertions may represent either alterations to exon/intron boundaries or novel exons.

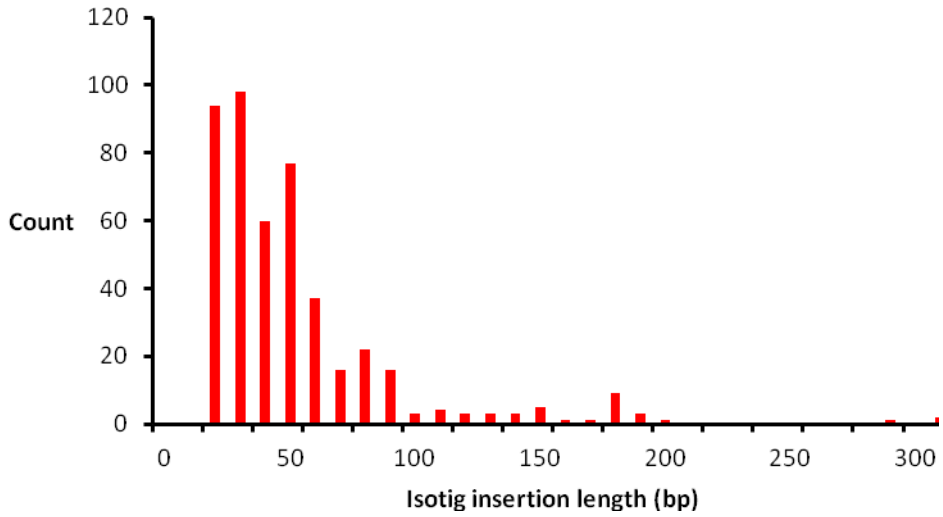


Figure 5.4. Length distribution of isotig insertions to AgamP3.6 transcripts. All isotig sequences were mapped to the transcriptome (AgamP3.6) using GMAP. Length of insertion to the reference transcriptome was identified using perl script insertion.pl (appendix) that extracts insertion information from the CIGAR string of the SAM alignment file. Only insertions greater than 10 bp were considered as values less than this could not be discriminated from sequencing errors e.g. homopolymers. 454 sequencing is prone to homopolymers sequencing errors and over estimation of homopolymer length may account for insertions. Therefore a conservative value of 10 was used.

5.2.4.1 NTR isotig characterisation

To further characterise NTR isotigs, all isotig sequences were used to query (BLASTN E-value = $1e^{-10}$) against AgamP3.6 transcripts that were extended to contain 10 kb upstream and 10 Kb downstream genomic regions. Isotig returning hits that met the E-value cutoff of $1e^{-10}$ (331 in total) were further analysed to asses their sequence length, GC content, number of exons and chromosomal location (Fig. 5.5). NTR isotigs had a shorter mean length 1,050 bp compared to the AgamP3.6 gene build transcripts (1,760 bp) (Fig. 5.5A). The average GC content of NTR isotigs was comparable to that of isotigs that do align to the AgamP3.6 gene build (Fig. 5.5B). NTR isotigs are distributed across all chromosomes, including the sex chromosomes X and Y (Fig. 5.5C). A majority of isotigs consist of either a single or two exons (Fig. 5.5D), with a maximum of six exons found within one isotig sequence. Exons were identified following mapping of NTR isotigs to the genome using the GMAP parameter -- novelsplicing. In total, 36.9 % of NTR isotigs contained an intron. This was fewer compared to isotigs that do align to an AgamP3.6 transcript, 61 % of these contained at least a single intron. When used as queries to search mosquito expressed sequence tag (EST) databases, 48.3 % matched an AgamESTDec10 EST (217,297 ESTs) (BLASTN E-value = $1e^{-10}$) 16.0 % matched an *Ae. aegypti* AaegEST (307,055 ESTs) (TBLASTX E-value = $1e^{-10}$) and 9.1 % matched a *Cx. cquinquefasciatus* CxEST(205,468 ETSs) (TBLASTX E-value = $1e^{-10}$).

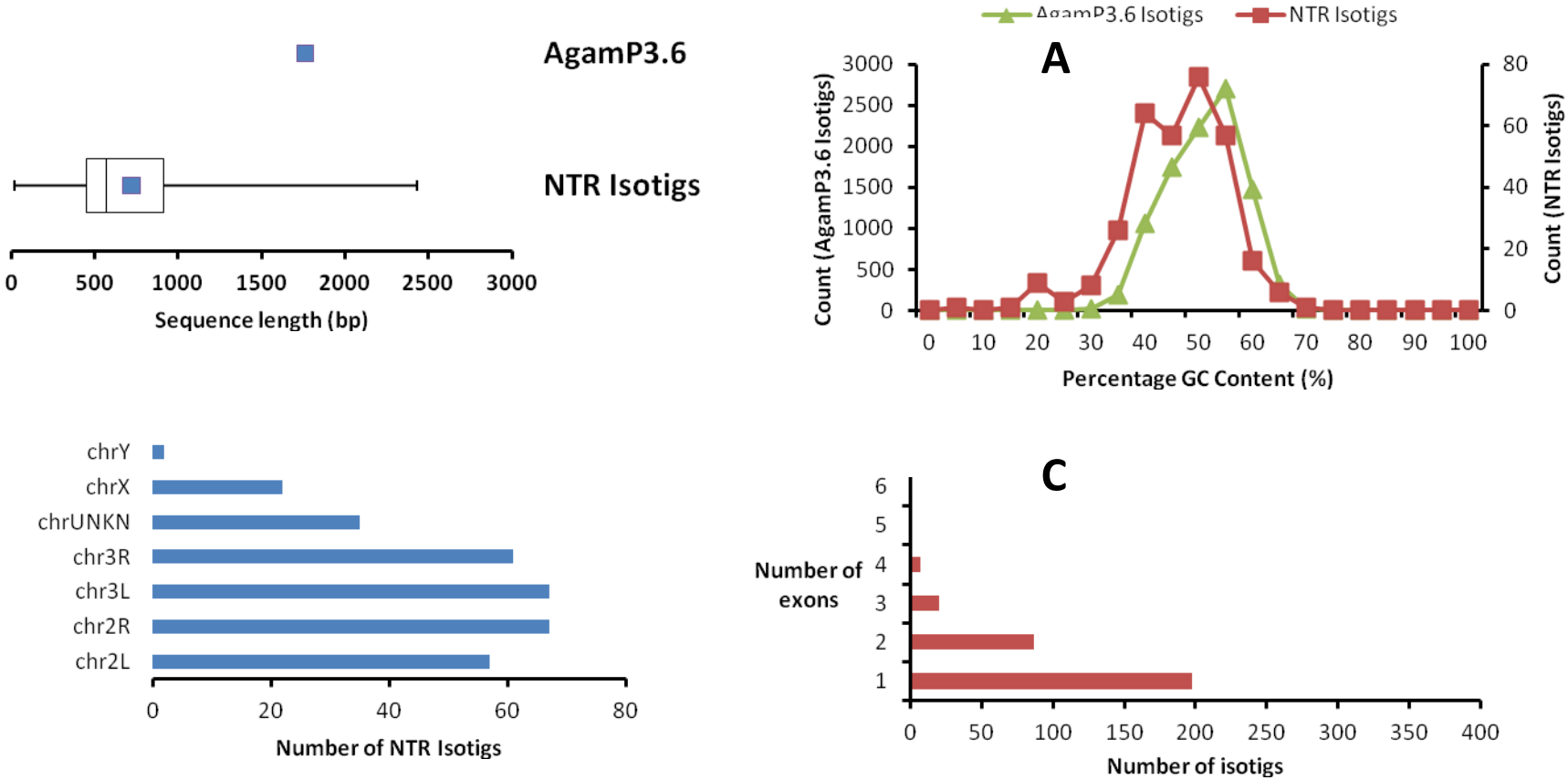


Figure 5.5. Characterisation of NTR isotig sequences. (A) Box plots of the lengths of NTR isotigs compared to annotated AgamP3.6 transcripts. The solid vertical line represents the median, the box shows the upper and lower quartiles and the box plot whiskers the lower and upper limits of sequence length (isotigs and AgamP3.6 transcripts). Mean length is indicated by solid blue boxes. (B) GC content of isotigs matching AgamP3.6 (+10Kb UTRs) and those with no match) (C) Chromosome location identified through mapping to the genome using GMAP. (D) Number of exons identified within the NTR isotigs.

5.2.5 Extension of AgamP3.6 transcript annotations

All embryo transcripts were used to assess whether AgamP3.6 predicted transcripts could be extended. First, 100 randomly selected embryo transcripts were queried against the genome (BLASTN E-value = $1e^{-10}$). This was done to ensure that the mapping of embryo transcript extremities was reliable, so that putative extensions of AgamP3.6 transcripts could be trusted. If embryo transcripts do not align to the genome across their entire length, a putative extension may be the result of poor alignment, rather than a true extension. On average, 6.5 and 11.5 nucleotides from the beginning and end of contigs did not match the genome respectively.

All embryo transcripts were then used to query the AgamP3.6 transcriptome (BLASTN E-value =$1e^{-10}$). Then a perl script transcript.pl (appendix) was used to filter the BLAST output to identify sequences that extend transcripts above a conservative 15 nucleotide threshold. i.e. if the start co-ordinate was >15 or the end co-ordinate was 15 less than the length of the transcript, they were considered as an extension. A value of 15 was selected as this encompasses the 6.5 and 11.5 nucleotides identified as terminal regions of transcripts with poor alignment and is slightly more restrictive to prevent identification of false-positives.

At the 5 prime end, 3,012 transcripts were extended by at least the threshold 15 nucleotides. Increasing this threshold to 50, 100, 500 and 1000 nt, there are 1,672, 1,018, 68 and 9 extended transcripts respectively. Examples of the transcripts extended by greater than 1,000 nt are shown in Figures 5.6 and 5.7. Contig07712 matched the predicted transcript of AGAP008187-RA of the reverse DNA strand (chr3R: 6,468,421-6,470,496 reverse strand) (Fig. 5.6). VectorBase annotation predicts the gene to contain 7 exons and alignment of the contig sequence to the genome confirms presence of all VectorBase exons, and reveals a further two exons 7.7 kb upstream of the predicted 5' end. Analysis of the direction of splicing signal revealed that transcription is occurring from the reverse DNA strand. Contig15579 maps to AGAP009564-RA on the forward DNA strand (chr3R:36,263,269-36,263,427) and extends the known annotation of a single exon to at least 4 exons, two towards the 5' end and a single exon to the 3' end (Fig. 5.7). Splicing signal sequences (GT/AG) confirmed that transcription is occurring from the forward DNA strand.

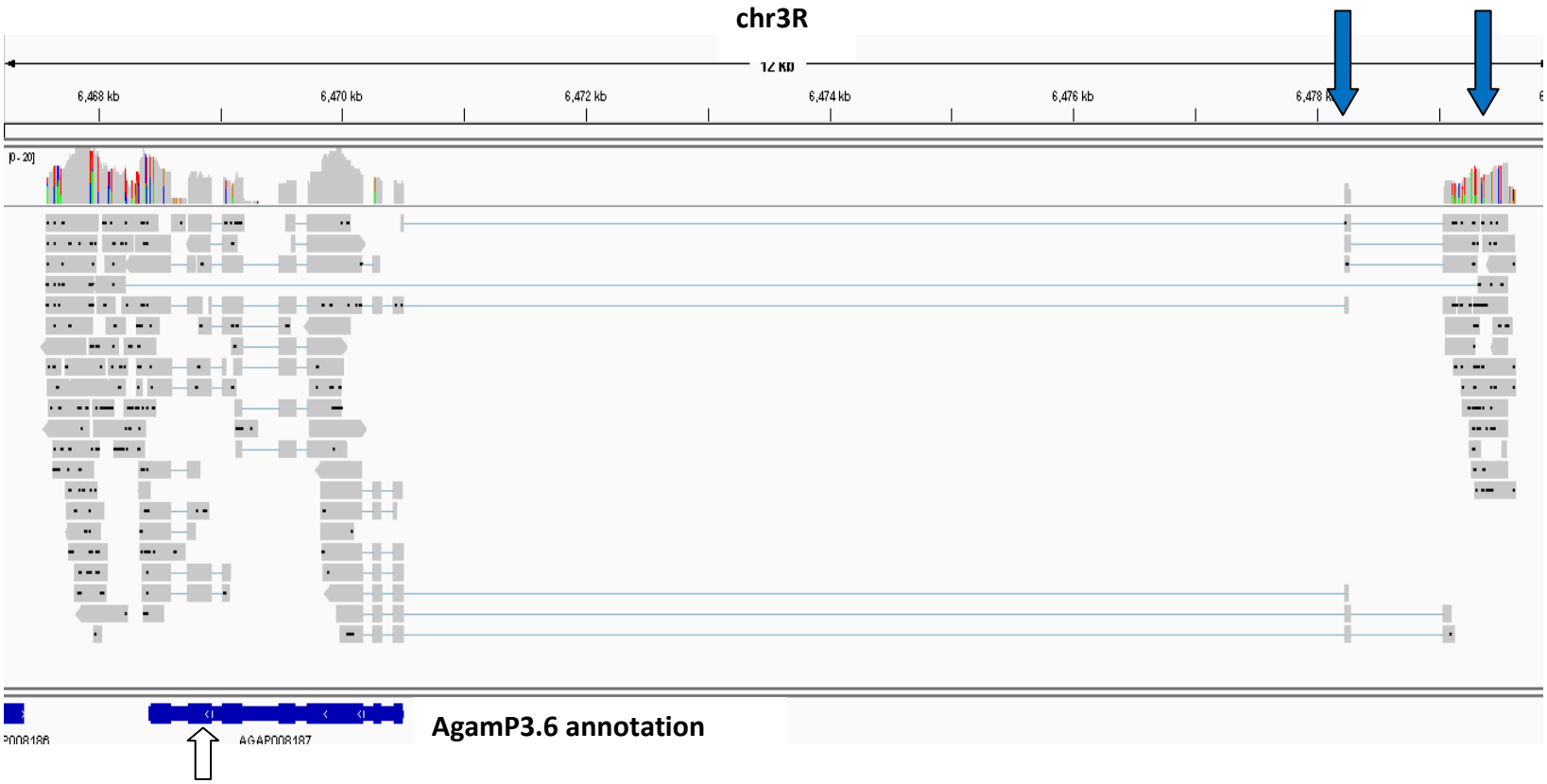


Figure 5.6. Integrative Genome Viewer (IGV) representation of AGAP008187-RA extension by contig07712. SAM alignment files of contig sequences were visualised with IGV along with the AgamP3.6 GTF annotation. Extension of AGAP008187-RA is seen at the 5' end of the predicted gene structure (indicated by the open arrow). Two additional exons (blue arrows) originate from reads that confirm the expression of previously annotated exons. Canonical splice sites are observed in the reverse orientation for both potential additional exons (CT/AC). Grey boxes indicate individual reads that constitute contig07712 and pale blue lines indicate introns. Black squares within reads represent single nucleotide differences to the reference.

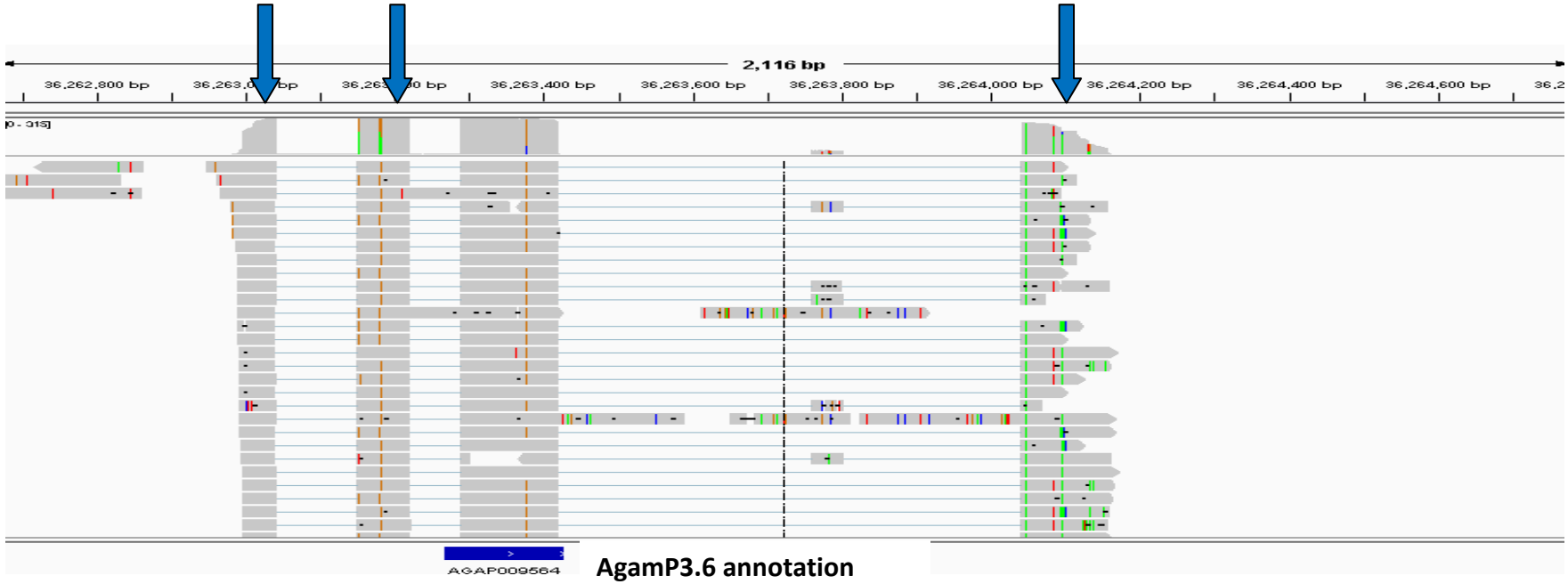


Figure 5.7. Integrative Genome Viewer (IGV) visualisation of AGAP009564-RA extension by contig15579. SAM alignment files of contig sequences were visualised with IGV along with the AgamP3.6 GTF annotation AGAP009564-RA consists of a sole exon (blue bar) and extension of the transcript is seen at both the 5' and 3' ends of the current annotation (indicated by blue arrows). Multiple reads show the presence of two additional exons at the 5' end and an additional exon at the 3' end. Canonical splice sites (GT/AG) are observed for each of the potential additional exon/intron boundaries. Grey boxes indicate individual reads that constitute contig07712 and pale blue lines indicate introns. Multiple reads align to intronic regions and may represent alternative splicing events or retained introns in primary mRNA transcripts. Coloured vertical lines within the individual reads represents single base differences compared to the reference (Green = adenosine; Blue = cytosine; Red = thymine; Brown = guanine; Purple = insertion to the reference.

5.2.6 Functional annotation of RNA-seq data based upon Gene Ontology classification

All embryo transcripts (45,401) were used for functional analysis. Based on BLASTX (E-value = $1e^{-3}$) searches executed through the software BLAST2GO against the NCBI nr protein database, 37.4 % of embryo transcripts had a significant match to a known protein. These sequences had a higher average length (784 bp) compared to those without a significant hit (345 bp). The sequences that did not have a significant BLASTX hit consisted of Newbler contigs (22.7 %), MIRA contigs (15.1 %) and singletons (62.2 %). The top-hits corresponded primarily to *An. gambiae* (87.9 %), followed by *An. darlingi* (6.4 %) and *Ae. aegypti* (2.3 %) proteins. Mapping was then carried out to retrieve GO terms previously assigned to the BLASTX hits and the top hit GO term extracted from this pool. Gene ontology (GO) terms are part of a controlled vocabulary, used to describe the function of a gene product. A total 11,489 (25.3 %) of sequences were associated with 3,418 unique GO terms.

GO terms were assigned within three different ontology categories (2^{nd} level GO terms), cellular compartment, biological process and molecular function and are shown in Figure 5.8, 5.9, and 5.10, respectively. At least one biological process term is associated with 7,971, molecular function with 10,375 and cellular compartment with 6,256 embryo transcripts. There were 4,375 embryonic transcripts with terms from all three GO categories. Within the biological process category, metabolic and cellular processes were the most abundant. Comparatively few sequences represented developmental processes and similarly few sequences were annotated to have involvement in localisation, an important embryonic process. Binding and catalytic activity were highly present in the category of molecular function, accounting for over 75 % of transcripts. In the cellular component category the two most represented were intracellular and intracellular part, with a large number of sequences also attributed to membrane localisation.

Assigned GO terms were searched for descriptions of interest to sex determination and this revealed a number of candidates including: sex (15), male (2), testis (4), female (3), ovary (1), and development (19). Included within the description of sex, *Ae. aegypti* sex determining protein *feminizer-1* (XP_001657331) had significant similarity to singletons_contig1378 (E-value = $1e^{-109}$). Singletons_contig1378 corresponds to the AgamP3.6 gene transcript AGAP00449 and is therefore not a NTR, but has 80 % similarity to

Ae. Aegypti feminizer-1 annotated upon its VectorBase entry. *Ae. aegypti* sex-determining region Y (SRY) (XP_001651645) had significant similarity to contig01241 (E-value = $1e^{-68}$), that aligns to AGAP010919 following BLASTN searches to the transcriptome and is therefore not an NTR. VectorBase confirms the high similarity of AGAP010919 (77 %) to SRY. *Cx. quinqifaceatus* TRA2 (XP_001866711) had significant similarity to a single read GMNRJ6W01D2AB3 (E-value = $6e^{-20}$) and this read aligns to 900 bp at the 5' end of AGAP009648 and may therefore represent an NTR. This sequence was not identified as a TRA2 candidate within chapter 2 and would therefore merit future investigation.

The InterPro database of protein domains was also searched to assign putative functions to the embryo transcripts. A total of 4,697 embryo transcripts were assigned 2,763 InterPro (IPR) terms. Signal P software detected the presence of putative signal peptides within 11,488 embryo transcripts. In addition, 8,113 sequences were assigned transmembrane domains using the TMHMM program. With regards to embryonic sequences without assigned GO terms (35,663 in total), 370 were assigned 261 different IPR term. In addition, 3,942 and 5,755 embryo transcripts with no GO term were annotated as having putative signal peptides and transmembrane domains respectively.

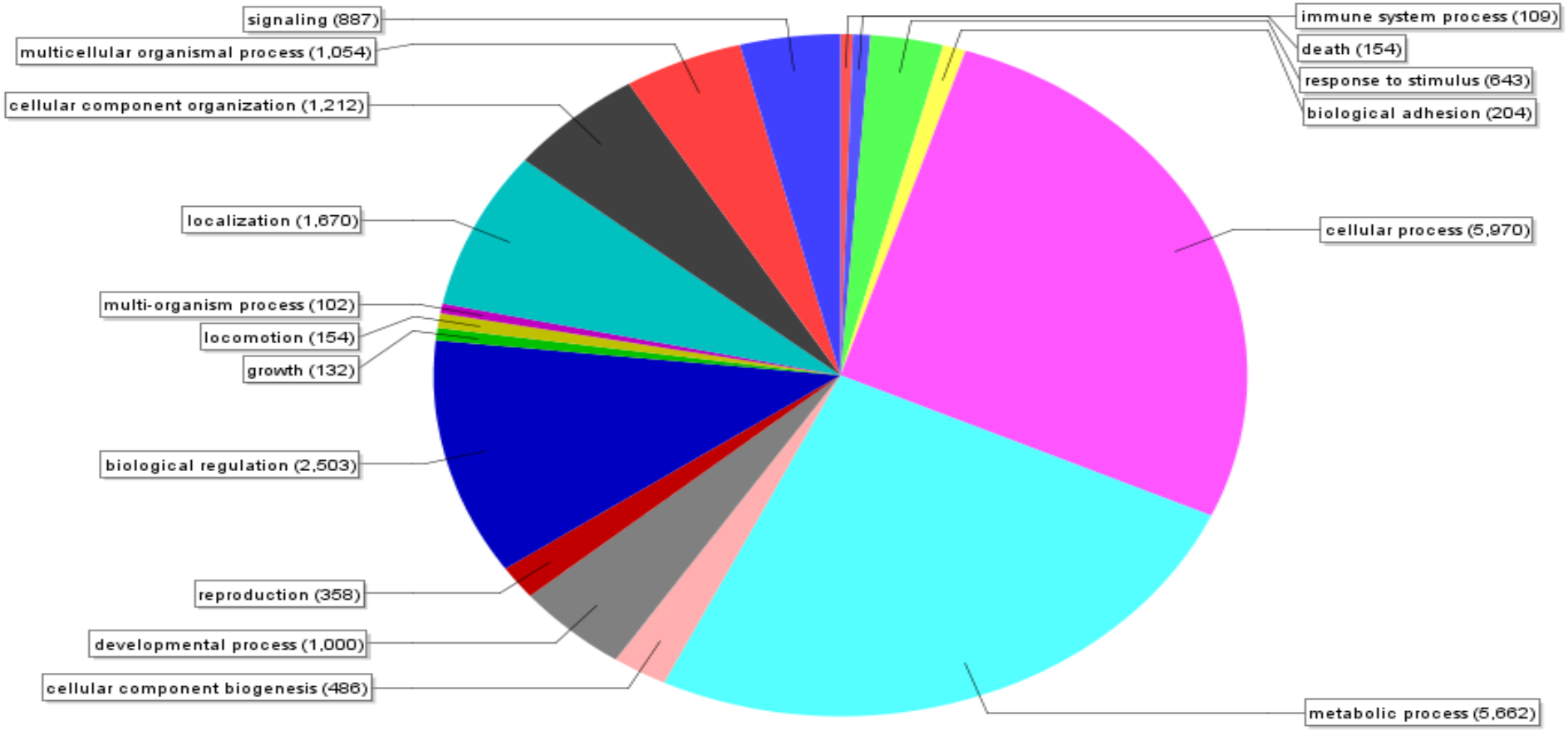


Figure 5.8. Gene ontology graph of Biological Processes (2nd level GO terms) of embryo transcripts. GO classification was based upon BLASTX (E-value = $1e^{-3}$) results generated from BLAST2GO software against the NCBI nr protein database, using non-sex embryo transcripts as query sequences. Mapping and annotation was carried out using default settings.

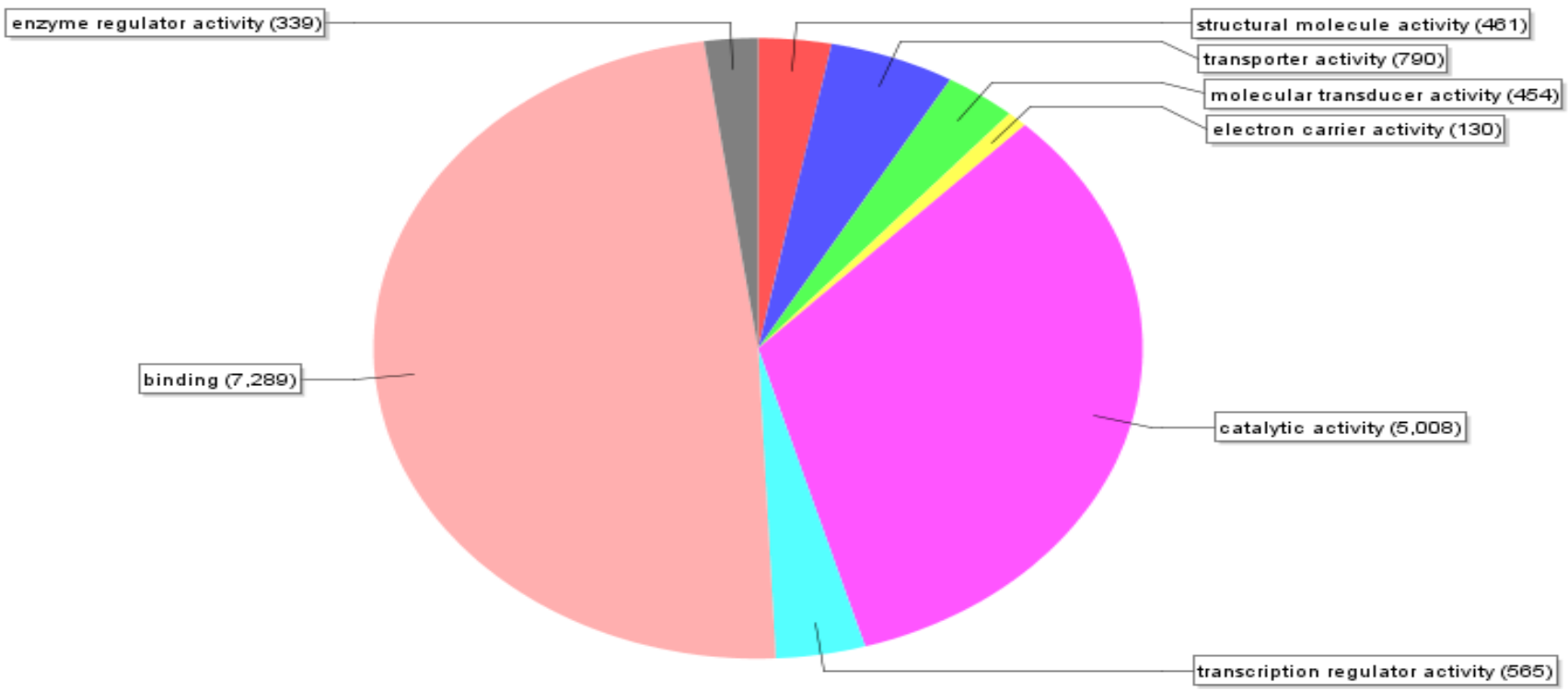


Figure 5.9. Gene ontology graph of Molecular Function (2nd level GO terms) of embryo transcripts. GO classification was based upon BLASTX (E- value = $1e^{-3}$) results generated from BLAST2GO software against the NCBI nr protein database, using non-sex specific embryo transcripts as query sequences. Mapping and annotation was carried out using default settings.

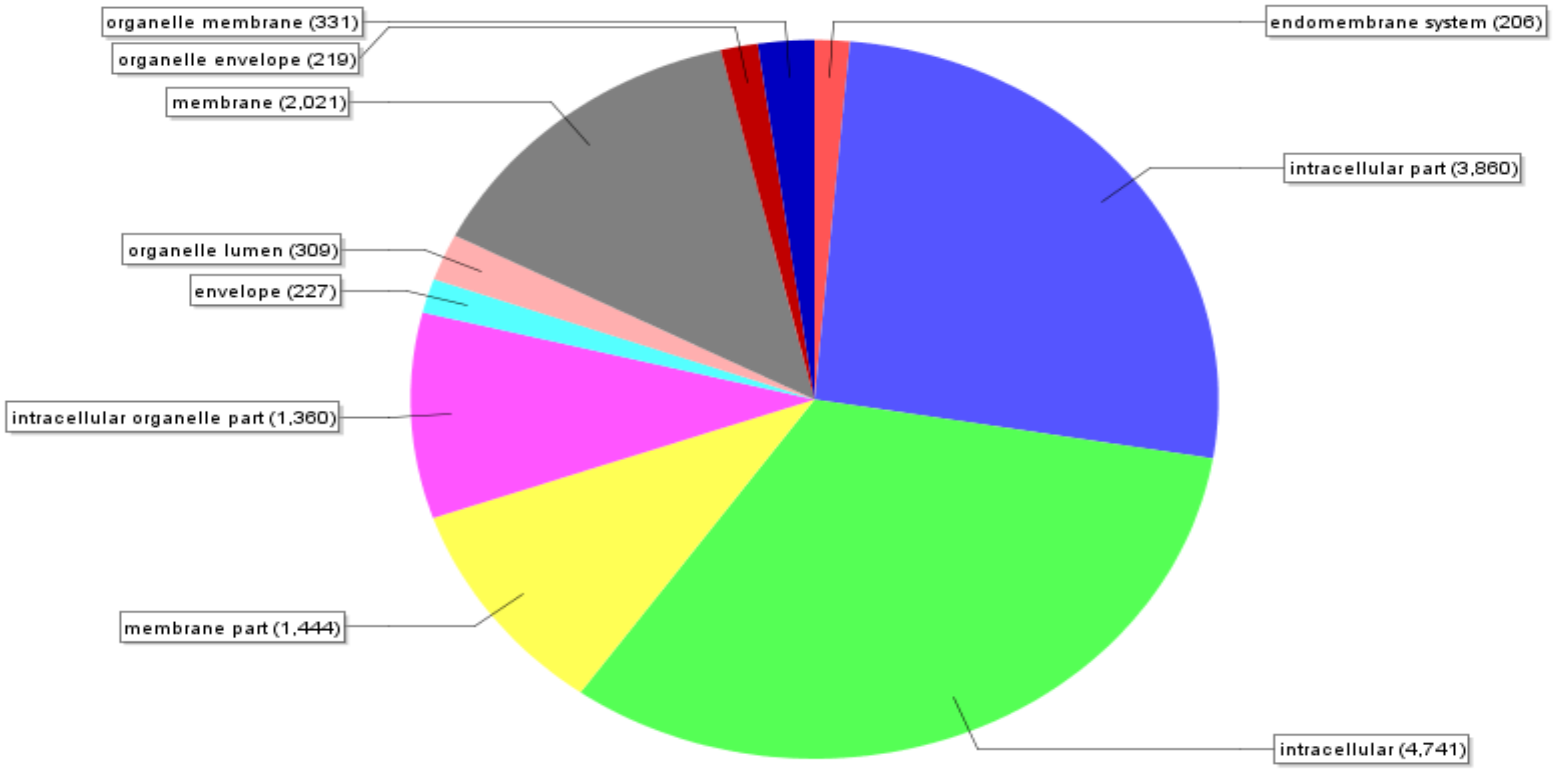


Figure 5.10. Gene ontology graph of Cellular Compartment (2nd level GO terms) of embryo transcripts. GO classification was based upon BLASTX (E-value = $1e^{-3}$) results generated from BLAST2GO software against the NCBI nr protein database, using non-sex specific embryo transcripts as query sequences. Mapping and annotation was carried out using default settings.

5.2.7 Identification of transcripts putatively involved in embryo development

Embryo transcripts were used to identify putative orthologs of genes involved in *D. melanogaster* embryo development processes. Proteins with GO terms sex determination (GO:0007530), pole cell development (GO:0007277), periodic partitioning by pair rule gene (GO:0007366), dorsal-ventral axis specification (GO:0009950) and anterior posterior axis specification (GO:0009948) were downloaded from the FlyBase resource and BLASTX (E-value = $1e^{-5}$) searches against embryo transcripts were used to identify potential *An. gambiae* homologues. Selected results are shown in Table 5.4.

In total, BLASTX analysis identified 350 embryo transcripts with significant similarity to a total of 41 *D. melanogaster* proteins with sex determination GO terms. For example, using the sex determination protein *groucho* as query, significant similarity was identified within contig11518 (E-value = 0). Subsequent BLASTP searches showed high similarity between AGAP010324 (top hit to contig11518) and *D. melanogaster groucho* (E-value = 0). Singleton_contig2858, annotated as *fruitless (fru)* (E-value = e^{-35}), is 458 bp in length and shares 54 % identity to *D. melanogaster fru*. However, this contig does not actually represent *An. gambiae fru.* Instead, it corresponds to *An. gambiae* gene AGAP003537, which belongs to the BTB superfamily and is a homologue of *fushi tarazu* repressor. BLASTN (E-value $1e^{-10}$) searches were then conducted with *An. gambiae fru* transcripts (male isoforms AY785361.1 and AY785360.1; female isoform AY725820.1) as query against our transcript libraries but yielded no significant hits. Thus *fru* transcripts were not identified in our RNA-Seq study.

No embryo transcript has significant similarity to *D. melanogaster doublesex* (*dsx*) following BLASTX searches. To establish if this was a limitation of using *D. melanogaster* DSX as query, nucleotide sequences of *An. gambiae dsx* transcripts were used as query (male isoform AY903307; female isoform AY903308) and also failed to identify a significant match (BLASTN E-value $1e^{-10}$). Therefore, using these approaches, we failed to identify *dsx* transcripts within the dataset.

Contig09350 (E-value = $1e^{-10}$) matched the query transformer2 (tra2) and corresponds to AGAP009810, a different gene to that functionally characterised in chapter 2. AGAP006798, the candidate identified in chapter 2, is present in the transcript libraries; however reads do not align to conserved tra2 domains and would therefore not be expected to be identified using the BLASTX approach taken here. Though AGAP009810 does have

significant sequence similarity to *D. melanogaster* tra2 (E-value = $6e^{-11}$), BLASTP searches showed that tra2 conserved domains were not present (chapter2). Using this approach, we have therefore not identified any additional tra2 *An. gambiae* homologues. The top hit (E-value = $3e^{-67}$) to the *D. melanogaster* sex determination protein Sex-lethal (SXL) was singleton_contig4559 and this contig corresponds to AGAP003899, the *An. gambiae* sx*l* homologue. Though thought not to play a role in mosquito sex determination, the *An. gambiae sxl* homologue is present in our embryonic dataset.

BLASTX analysis identified a total of 500 embryo transcripts with significant similarity to 39 *D. melanogaster* proteins with pole cell GO terms (out of a total 49 used as queries). Contig10039 had a significant hit to *D. melanogaster* nanos (E-value = $2e^{-6}$). Nucleotide alignment of contig10039 to *An. gambiae nanos* (AY873992) revealed 98 % identity across the 599 bp length of contig10039. Contig08113 showed significant similarity to *D. melanogaster runt* and BLASTN searches showed that it corresponds to AGAP001094. Annotation of AGAP001094 present on VectorBase suggests that AGAP001094 shares 40 % target identity to *D. melanogaster runt*, a key gene in the formation of embryo segmentation. Strong candidates for transcripts involved in embryo pattern formation were also identified, notably *cactus* (contig02045, E-value $1e^{-59}$) and the homeobox gene, *bicoid* (contig09056, E-value = $2e^{-10}$). Contig09056 is 1,586 bp in length, and when used to query (BLASTX) the NCBI nr protein database, returns better E-value hits to other homeobox-releated proteins, including *D.melanogaster* Zen2 (E-value = 1e-22), Proboscipedia (E-value 2e-22) and Zen (E-value = 3e-20).

Table 5.4. Selected contigs with sequence similarity to *D. melanogaster* proteins involved in embryo development. The table summarises the top hit contigs from BLASTX searches to proteins involved in *D. melanogaster* embryo development. Shown are the top hit contig to the protein sequence and the top hit AGamP3.6 gene to the contig. Where gapped hits were produced following BLASTX searches, the first E-value was taken and is reported in the table.

Sex determination GO:0007530

D. melanogaster Protein Query	Top hit Contig	Identities %	Length of matching query (bp)	E value of top hit contig	Top hit *An. gambiae* Gene matching to contig (E-value)
sxl-PA;	singletons_c4559	93	115	$3e^{-67}$	AGAP003899 (0)
tra2-PA;	contig09350	35.87	91	$1e^{-10}$	AGAP009810 (0)
ix-PA;	contig07816	50.35	140	$2e^{-34}$	AGAP001237 (0)
fru-PA;	singletons_c2858	54.31	115	$9e^{-35}$	AGAP003537 (0)
gro-PA;	contig11518	88.92	351	0	AGAP010324 (0)

Pole cell GO:0007277

D. melanogaster Protein Query	Top hit Contig	Identities %	Length of matching query (bp)	E value of top hit contig	Top hit *An. gambiae* Genematching to contig (E-value)
aret-PA;	contig09237	73.68	93	$2e^{-37}$	AGAP004950 (0)
ftz-PA;	contig07963	42.18	205	$4e^{-28}$	AGAP013157 (0)
nos-PA;	contig10039	53.33	44	$2e^{-06}$	AGAP006098 (0)
vas-PA;	contig08172	40.62	383	$3e^{-80}$	AGAP004912 (0)
wee-PA;	GMNRJ6W01DCC20	64.62	129	$4e^{-47}$	AGAP007798 (0)

Pair rule GO:0007366

D. melanogaster Protein Query	Top hit Contig	Identities %	Length of matching query (bp)	E value of top hit contig	Top hit *An. gambiae* Gene matching to contig (E-value)
dpn-PA;	GMNRJ6W01AO7CW	61.76	101	$8e^{-33}$	AGAP012346 (0)
en-PA;	singletons_c532	87.95	82	$5e^{-37}$	AGAP008024 ($3e^{-132}$)
eve-PA;	contig11549	72.33	158	$1e^{-59}$	AGAP010279 (0)
odd-PA;	GMNRJ6W01CGL3N	87.27	54	$6e^{-42}$	AGAP007972 (0)
run-PA;	contig08113	77.71	174	$8e^{-84}$	AGAP001094 (0)

Dorsal Ventral Patterning GO:0009950

D. melanogaster Protein Query	Top hit Contig	Identities %	Length of matching query (bp)	E value of top hit contig	Top hit *An. gambiae* Gene matching to contig (E-value)
tkv-PA;	contig02811	79.69	190	$1e^{-108}$	AGAP009329 (0)
cact-PA;	contig02045	49.17	241	$1e^{-59}$	AGAP007938 (0)
dl-PA;	GMNRJ6W01CCYTR	72.52	129	$8e^{-54}$	AGAP009515 (0)

Anterior Posterior Patterning GO:0009948

D. melanogaster Protein Query	Top hit Contig	Identities %	Length of matching query (bp)	E value of top hit contig	Top hit *An. gambiae* Gene matching to contig (E-value)
a-PA;	contig06777	51.46	102	$1e^{-28}$	AGAP008384 (0)
bcd-PD;	contig09056	56.25	47	$2e^{-10}$	AGAP004647 (0)
dpp-PA;	contig11541	32.31	278	$2e^{-33}$	AGAP000824 (0)
cad-PA;	contig07301	53.03	65	$5e^{-15}$	AGAP004660 (0)

5.2.8 *de novo* assembly of male and female transcriptomes

To investigate differences between male and female transcriptomes, cleaned reads were assembled into 10,868 (mean read depth = 18.2) and 11,077 (mean read depth = 16.3) contigs for males and females respectively (Fig. 5.11A) using Newbler. Following assembly, 40,901 and 30,348 singletons remained for males and females respectively. These were then assembled using MIRA software, resulting in assembly of 5,631 and 4,912 contigs for males and females respectively (Fig. 5.11B). Assembly by MIRA resulted in the assembly of ~40 % of the Newbler singleton sequences for both sexes. Read depths were lower for MIRA contigs, with an average of 1.9 and 2.0 for males and females respectively. In addition, a total of 9,708 and 9,694 (isotigs were assembled for males and females respectively (Fig. 5.12C).

In total 43,513 embryo transcripts (16,499 contigs and 27,014 singletons) constituted the male transcriptome. The female transcriptome was represented by fewer embryo transcripts, 33,854 in total (15,989 contigs and 17,865 singletons). This is representative of an initial smaller number of cleaned reads (summarised in Fig. 5.12). Over 90 % of reads were assembled for both male and female datasets. Singletons represent 6.6 and 5.1 % of total reads for males and females respectively. Maximum contig lengths did not differ between males and females, with a length of 6,251 and 6,248 bp for males and females respectively. N50 values for contig assembly were 857 and 947 bp for males and females respectively.

BLASTN searches against AgamP3.6 gene build were carried out using male and female embryo transcripts as query. In total, 23,140 and 17,470 embryonic sequences did not align to a AgamP3.6 transcript for males and females respectively. Of the 14,974 AgamP3.6 transcripts, 8,311 (55.5 %) and 7,529 (50.3 %) were confirmed to be expressed in males and female respectively. Assembly of the sex-specific transcriptomes is summarised in Table 5.5.

Table 5.5. Summary statistics of assembled male and female sequences from Newbler and MIRA. Newbler was first used to assemble cleaned reads and singletons extracted from the 454Readstatus file. Singletons were then used for MIRA assembly and all embryo transcipts were then used to query the AgamP3.6 transcriptome (BLASTN). Shown is a summary for Newbler, MIRA and final assembly.

Contig Assembly Newbler	Male	Female
Total nucleotides (bp)	153,802,540	146,610,545
Cleaned reads (% of raw reads)	408,851 (78.2)	353,399 (70.7)
Total nucleotides cleaned reads (bp) (% total base pairs)	148,075,267 (96.3)	139,585,999 (95.2)
% nucleotides assembled	87.05	88.40
Number of Isotigs	9,708	9,694
Number of contigs	10,868	11,077
Total length of Newbler contigs (bp)	9,600,084	10,370,074
Maximum contig lengths (bp)	6,251	6,249
Number of unassembled reads	40,901	30,348
Assembly of singletons - MIRA		
Number of MIRA contigs	5,631	4,912
% nucleotides of Newbler unassembled reads assembled	37.8	47.4
Total length of MIRA contigs (bp)	2,522,175	2,257,030
Maximum contig length (bp)	937	1,101
Number of unassembled reads	27,014	17,865
Final assembly		
Total number of contigs (Newbler and MIRA combined)	16,499	15,989
% nucleotides assembled	90.7	92.2
Number of AgamP3.6 transcripts to which embryo transcripts align (% of total)	8311 (55.5)	7529 (50.2)
Number of embryo transcripts that do not align to AgamP3.6 transcripts	23,140	17,470
Number of embryo transcripts that extend AgamP3.6 transcripts (> 15 bp	1,714	1,482

B

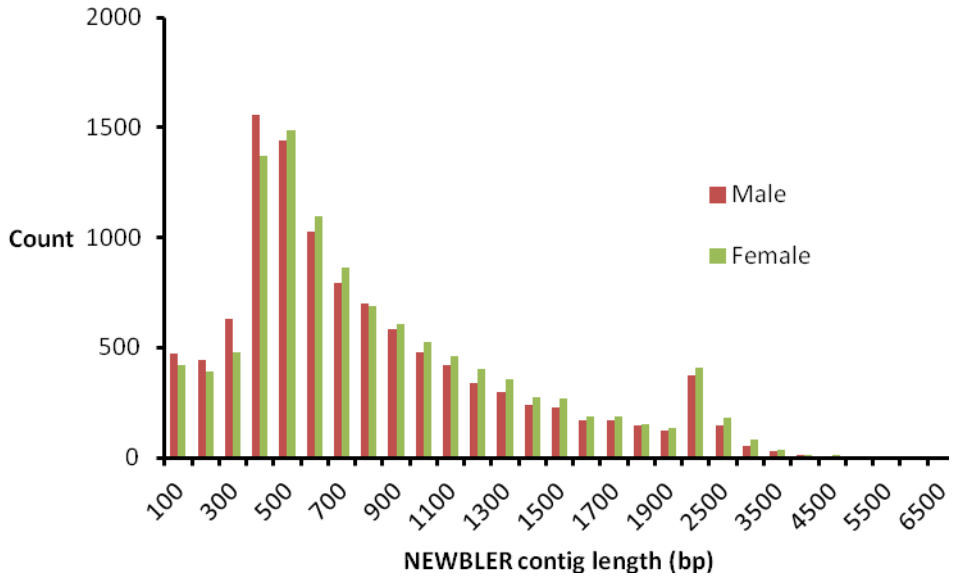
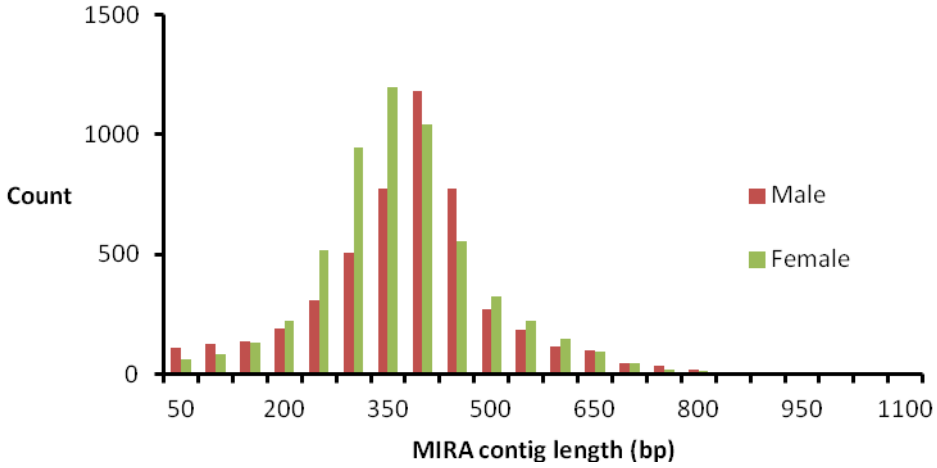


Figure 5.12. Length distribution of contigs assembled from male and female transcript reads using Newbler and MIRA. Cleaned reads were first assembled by Newbler



using default settings. As part of the assembly process, Newbler assigns a 'status' to each read, including unassembled. All sequences with this status were extracted from the Newbler output and assembled using MIRA default 454 settings. Contig lengths were calculated using perl script length.pl (appendix) and used to show distribution of length for A) Newbler and B) MIRA contigs.

C

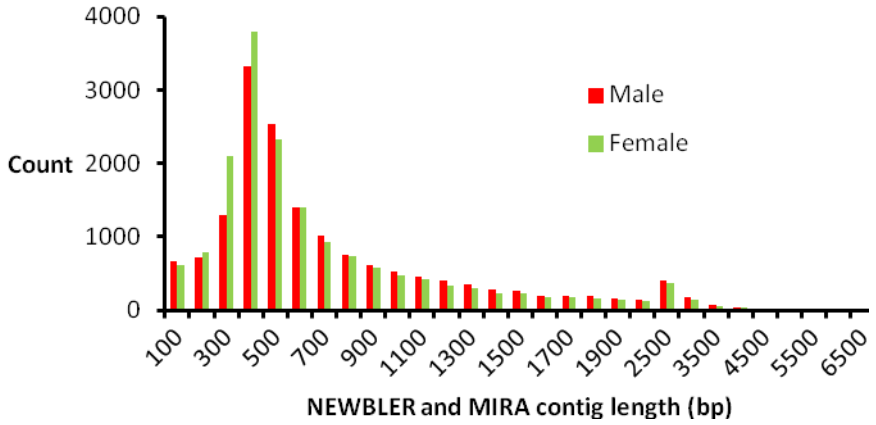
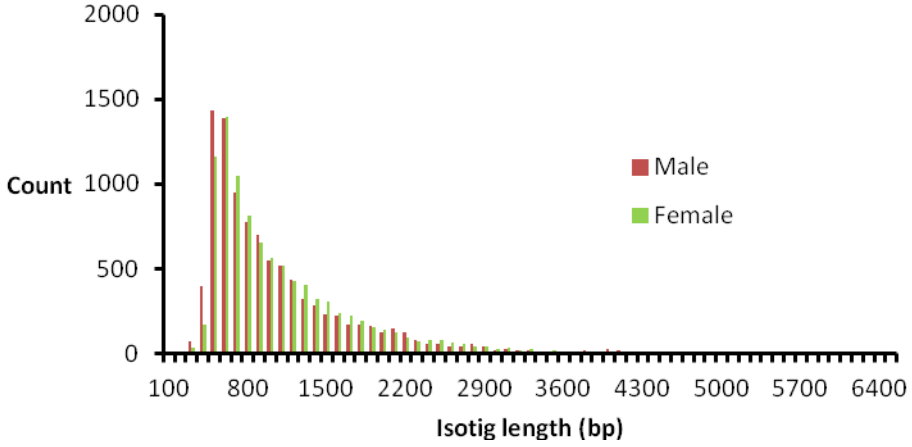
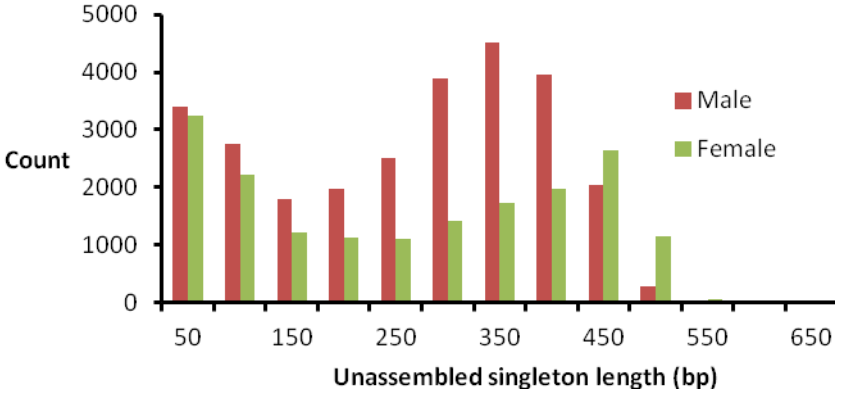


Figure 5.13. Length distribution of contigs, isotigs and unassembled reads from male and female transcript reads. (A) Length distribution of all contigs from Newbler and MIRA. (B) The distribution of isotig lengths assembled using Newbler from male and female



transcript reads. C) Length distribution of residual singletons. Only unassembled reads greater than 50 bp were included.



5.2.9 Sex-specific alternative splicing

5.2.9.1 Cufflinks

Sex-specific SAM alignment files of cleaned reads generated by GMAP were uploaded onto the Galaxy online cloud (http://main.g2.bx.psu.edu/), to utilise the Cufflinks package of software (cufflinks, cuffcomapre and cuffdiff). In brief, the package assembles transcripts from a SAM alignment using a GTF annotation file (AgamP3.6) as reference, to identify different transcripts originating from the same start site. Cuffdiff identified ten genes alternatively spliced between the male and female datasets. Each candidate was examined by visualisation of SAM alignment files using IGV. This was done to determine if candidate sex-specific splicing events had canonical splice sites and that mapping was accurate (e.g. low complexity sequences were searched for). Two of the candidates had canonical splice sites and are summarised below. AGAP003538-RA is predicted to have alternative 5' ends. Male reads (GLONWJ301CPLY0 and GLONWJ301BFO82) map 53 kb and 10 kb past the annotated sequence respectively and female reads (GMNRJ6W01EOR2E and GMNRJ6W01D5287) 59kb and 66kb past the annotated sequence respectively. AGAP003538 is in the forward orientation and each of the sex-specific reads have canonical (GT/AG) splice sites. The length of the region they align to is over 50 bp for all reads. AGAP003538 is annotated at VectorBase as having 90 % similarity to *Ae. Aegypti* 60S ribosomal protein (AAEL003942). We have previously shown that ribosomal protein transcripts have a high number of mapped reads (>2000 shown in Table 4.4) and therefore the single reads suggesting alternative 5' ends did not provide sufficient evidence to investigate this candidate further. A second candidate, AGAP010257 consists of five exons, of which all are detected in both sexes. AGAP010257 also has high similarity (77 %) to an *Ae. aegypti* 60S ribosomal protein (AAEL009825). A single male derived read displays a unique genomic organisation, GLONWJ301AS0DY covers exon one and partially exon two, skips exons three and four and has a partial match to exon five (Fig. 5.13). No female reads display this genomic organisation. AGAP010257-RA is in the reverse orientation but GLONWJ301AS0DY does not have canonical splice sites in either direction at its exon/intron boundary. As a high number of reads would be expected to support a ribosomal transcript alternative splicing event, due to the lack of sequence information supporting these splicing events (i.e. a single read) and the absence of canonical splice sites for GLONWJ301AS0DY they were not investigated further and a different approach, involving custom perl scripts, was undertaken.

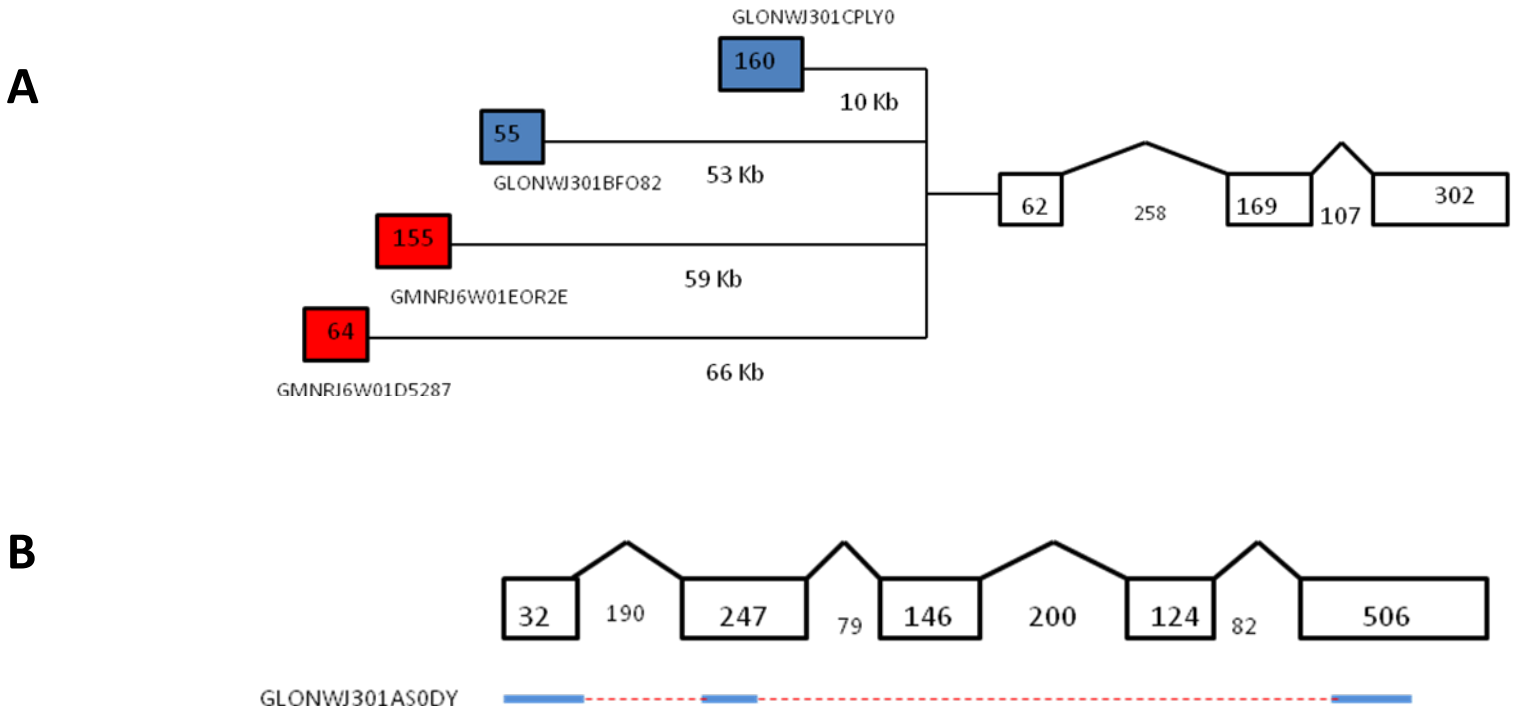


Figure 5.13. Schematic diagram of two Cufflinks identified putative sex-specific splicing events. Each Cufflinks sex-specific splicing candidate was examined through visualisation of SAM alignment files using IGV. (A) AGAP003538-RA has three annotated exons and four reads (two male and two female) suggest potential alternative 5' ends. Two male reads, GLONWJ301CPLY0 and GLONWJ301BFO8, 2 align to 5' upstream regions 10 and 53 Kb past the first annotated exon respectively. Two female reads, GMNRJ6W01EOR2E and GMNRJ6W01D5287 also align to 5' upstream regions, 59 and 66 Kb past the first annotated exon respectively. Each read has canonical splice sites. (B) AGAP010257 has five annotated exons and both male and female reads align to each of these exons, suggesting they are all included in the mature transcript. A single male read, GLONWJ301AS0DY, shows a unique organisation. The read aligns to the first, partially to the second and partial to the fifth exon (blue solid line) with putative introns shown by the dashed red line. These introns do not contain canonical splice sites in either direction. All AgamP3.6 annotated exons are shown by open boxes, introns by diagonal black lines and numbers represent length in base pairs.

5.2.9.2 Read count data across AgamP3.6 transcript exon-exon junctions

In brief, the script (AS.pl appendix) generates a database of 200 bp probes that represent exon-exon boundaries of *An. gambiae* AgamP3.6 transcripts. Additionally, exon boundaries of NTR isotigs were included in the database. The 200 bp probes were then queried against the male and female reads separately to identify the number of reads that cross each boundary (BLASTN E-value $1e^{-10}$). A total of 97,390 and 7,114 probes were used from AgamP3.6 and NTR isotigs respectively. In total, 2,998 and 1,743 putative sex-specific exon-exon junctions were identified for males and females respectively with at least a single read crossing the junction. Selected candidates, with the highest number of reads crossing exon-exon junctions are shown in Table 5.6.

Each of these candidates shown in Table 5.6 were examined using IGV visualisation of SAM alignment files (reads mapped for each sex to the genome) to confirm the number of reads that cross each boundary. In all but one case no clear evidence of sex-specific splicing was evident and the number of reads at each location was over-predicted using BLASTN searches. An example of a visualised candidate with no clear evidence is shown in Figure 5.14. AGAP003139-RA is composed of five exons and exon three is absent in females (Fig. 5.15A). The first exon of AGAP003139-RA is not shown in Figure 5.16 due to the length of the adjacent intron, but is common to both males and females. All exon/intron boundaries have canonical splice sites in the reverse orientation (CT/AC). The sequence of AGAP003139-RA was downloaded from NCBI and mapped to the genome using BLASTN to establish its genomic organisation (Fig. 5.15B). This was done so that primers could be designed to amplify putative sex-specific fragments.

Primer 3139F1 is located in the second, 3139F2 in the fourth and 3139R1 in the fifth exon. Amplification with primer pair 3139F1/R1 would be expected to amplify a fragment 115 bp larger in males compared to females, as they flank the putative male specific third exon. Primer pair 3139F2/R1 would be expected to amplify a fragment of the same size in both males and females. Following RT-PCR amplification using primer pair 3139F1/3139R1 from sex-specific 20 hr AEL embryonic RNA, the fragment of the expected male size (690 bp) was found in both males and females (Fig. 5.15C). The fragment of the female expected size (575 bp) was not amplified from either male or female embryonic RNA. The expected fragment size (394 bp) was amplified from both sexes for primer pair 3139F2/3139R1 also

(Fig. 5.16C). This indicated that exon two is present in AGAP003139-RA transcripts of both males and females and that it is not sex-specifically spliced using the approach taken here.

Table 5.6. Selected putative sex-specific alternatively spliced candidates identified using custom perl script AS.pl. Candidates were prioritised based upon the number of reads that cross the selected exon exon boundary. Where no AgamP3.6 hit was present, probe represents a NTR isotig. This probe identified a novel exon upstream of the first predicted exon of AGAP010588 and had therefore been termed exon 1.

Probe Name Location :Co-ordinates	AgamP3.6 Top Hit (exon boundary)	Number of BLASTN significant hits	
		Male	Female
chr3L:36310736-36310836	AGAP012014-RA (5-6)	125	0
chr3L:13354378-13354478	AGAP010933-RA (1-2)	42	0
chr2R:11738293-11738393	AGAP001872-RB (4-5)	18	0
chr2R:33149295-33149395	AGAP003139-RA (2-3)	16	0
chrUNKN:41201652-41201752	AGAP012938-RA (1-2)	16	0
chr2L:9737338-9737438	AGAP005079-RA (1-2)	0	55
chr3L:35652108-35652208	NTR isotig	0	18
chr3L:667532-667632	AGAP010321-RA (2-3)	0	18
chrX:2993926-2994026	AGAP000179-RA (2-3)	0	18
chr3L:6371792-6371892	AGAP010588-RA (1-2)	0	15

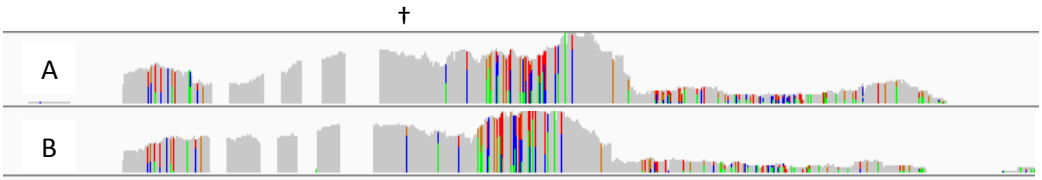


Figure 5.14. Example of an alternatively spliced candidate with no supporting evidence from IGV. All reads were aligned to the genome using GMAP and coverage plots were then produced from SAM alignment files following visualisation using IGV. The height of the coverage plot represents the number of reads aligned to the reference. The coloured lines within the coverage plot represent single base pair differences between aligned reads and the reference transcript. Shown is AGAP012014-RA where exon boundary 5-6 (indicated by †) is expected to be male specific but the coverage plot indicates that this boundary is present in both males (A) and females (B).

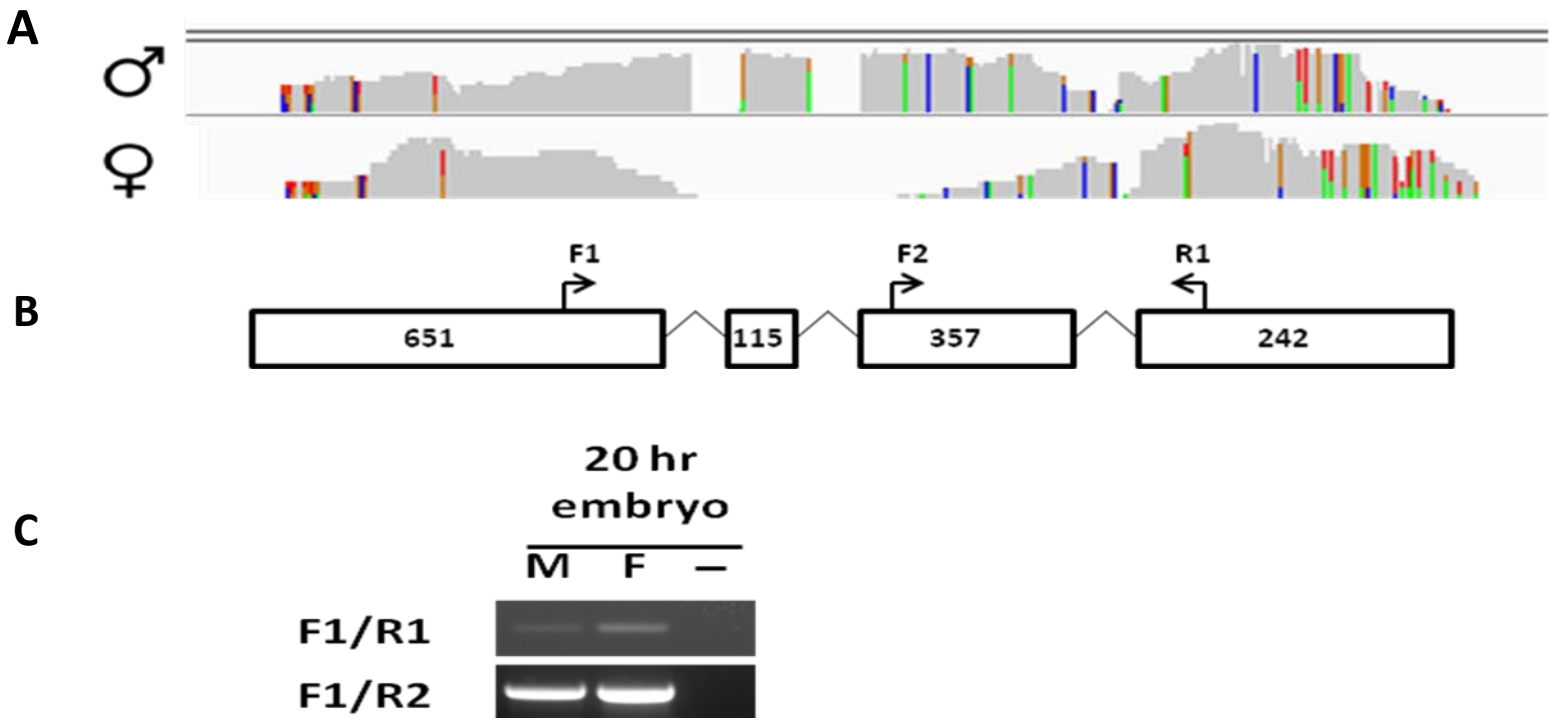


Figure 5.15. IGV visualisation of alternatively spliced candidate AGAP003139-RA and validation by RT-PCR. (A) All reads were aligned to the genome using GMAP and coverage plots were then produced from SAM alignment files following visualisation using IGV. The height of the coverage plot represents the number of reads aligned to the reference. The coloured lines within the coverage plot represent single base pair differences between aligned reads and the reference transcript. (B) Schematic diagram of AGAP003139-RA with positions of primers used for RT-PCR indicated. Open boxes represent exons, diagonal lines introns and numbers length in base pairs. Exon two (115 bp) is a putative male specific exon. (C) RT-PCR amplification was carried out from sex-specific 20 hr AEL embryonic RNA and a fragment of the expected size was amplified from both sexes using primer combinations 3139F1/3139R1 and 3139F1/3139R2. M = male; F = female; - = negative control where no RNA template was added to RT-PCR reaction.

5.2.10 Sex-specific differential transcript expression

To estimate differential transcript expression between male and female datasets, two different approaches were used. The first utilised RPKM values to normalise read count data based upon the length of the reference sequence, as described in Materials and Methods. The second approach used the R/Bioconductor edgeR package to normalise male and female libraries based upon the transcript library sizes.

5.2.10.1 Differential expression based upon RPKM values

Expression levels were first estimated for AgamP3.6 transcripts by establishing read count data. In addition, expression levels of NTR contigs were estimated using the same method. Read counts were first converted to RPKM values for each sex to normalise for total number of reads sampled from each library and then compared to identify differential expression between the sexes. Log fold change (LogFC) values are shown in Figure 5.16 and in total, 987 and 1,267 transcripts had LogFC values greater than one for males and females respectively. Increasing this threshold to greater than two Log_2FC values, 247 and 244 transcripts were differentially expressed between males and females respectively. Using a LogFC > 2 cutoff, male differentially expressed transcripts were composed of 227 and 20 AgamP3.6 and NTR transcripts respectively. Using the same LogFC > 2 cutoff, 203 and 41 female differentially expressed transcripts were AgamP3.6 and NTRs respectively.

Canddiates with the highest fold change are listed in Table 5.7 (male) and Table 5.8 (female). Putative differentially expressed transcripts corresponded to a variety of gene ontology terms, including protein binding, catalytic activity and nucleotide binding. Differences within annotated GO terms between male and female transcripts, with LogFC >1, were tested across all three GO vocabularies using a Fisher's exact test ($p < 0.05$), with no GO term determined as being statistically significant. The transcript displaying the greatest fold change for males AGAP003227-RA (4.4 LogFC) is predicted to be involved in nucleotide binding. The transcript displaying the greatest female bias, AGAP003877-RA (4.6 LogFC) had no associated gene ontology terms.

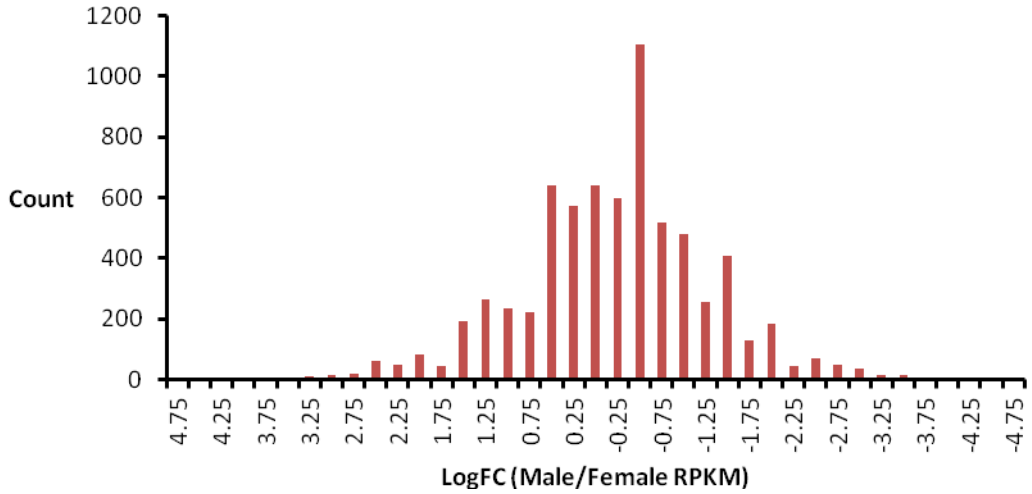


Figure 5.16. The log RPKM expression ratio of AgamP3.6 and NTR contig transcripts present in both male and female transcript libraries. For each AgamP3.6 transcript that had at least one uniquely mapped read from both male and female transcript libraries, RPKM values were calculated from the number of mapped reads, length of transcript and total number of reads in the library. RPKM values were compared to calculate log expression ratios for the RPKM values.

Table 5.7. Top ten male differentially expressed AgamP3.6 transcripts calculated via RPKM values. Number of reads mapping to individual AgamP3.6 transcripts were used to identify transcripts differing in expression between the sexes. These were used to calculate RPKM values for each of the transcripts. N/A = no GO term was assigned to this query.

AgamP3.6 Transcript	GenBank Accession Number	Number of male reads	Number of female reads	Male RPKM value	Female RPKM value	Log_2Fold Enrichment	Top hit Associated GO term
AGAP003227-RA	XM_312935.5	20	1	61.18	3.71	4.04	nucleotide binding
AGAP006235-RA	XM_316302.4	19	1	18.49	1.18	3.97	auxin biosynthetic process
AGAP009431-RA	XM_307843.4	1225	78	10827.93	836.38	3.69	structural component of ribosome
AGAP002489-RA	XM_563683.2	14	1	29.69	2.57	3.53	cell differentiation
AGAP006092-RA	XM_316150.4	13	1	21.06	1.97	3.42	N/A
AGAP008011-RA	XM_001237934.1	13	1	73.70	6.88	3.42	N/A
AGAP004710-RA	XM_001689050.1	574	47	6003.30	596.32	3.33	spermatid development
AGAP011688-RA	XM_001238384.2	12	1	5.60	0.57	3.31	protein binding
AGAP007107-RA	XM_308650.4	12	1	22.18	2.24	3.31	protein folding
AGAP007860-RA	XM_317632.3	12	1	39.74	4.02	3.31	N/A

Table 5.8: Top ten female differentially expressed AgamP3.6 transcripts calculated via RPKM values. Number of mapping reads to AgamP3.6 transcripts were used to identify transcripts showing changes in expression between the sexes. These were used to calculate RPKM values for each of the transcripts. N/A = no GO term was assigned to this query.

AgamP3.6 transcript	GenBank Accession Number	Number of female reads	Number of male reads	Female RPKM	Male RPKM	Log_2 Fold Enrichment	Top hit Associated GO term
AGAP003877-RA	XM_001230765.2	20	1	80.50	3.32	4.60	N/A
AGAP013384-RA	XM_003435969.1	17	1	29.23	1.42	4.37	binding
AGAP007999-RA	XM_317470.5	17	1	15.16	0.74	4.37	structural constituent of ribosome
AGAP009977-RA	XM_001689118.2	15	1	51.20	2.81	4.19	protein transport
AGAP001472-RA	XM_001689294.2	15	1	15.65	0.86	4.19	mesodermal cell migration
AGAP003025-RC	XM_003436310.1	268	18	1452.11	80.40	4.17	structural constituent of ribosome
AGAP000882-RA	XM_316858.4	14	1	43.73	2.57	4.09	oxidation reduction
AGAP000187-RA	XM_310949.5	11	1	15.70	1.18	3.74	microtubule motor activity
AGAP009264-RA	XM_320056.4	20	2	55.47	4.57	3.60	biosynthetic process
AGAP007774-RA	XM_317741.5	10	1	9.21	0.76	3.60	imaginal disc development

5.2.11.2 edgeR differential transcript expression

To further investigate differential gene expression between male and female datasets, the edgeR package from the Bioconductor software development project was used (Robinson *et al.* 2010). Non-normalised AgamP3.6 and NTR contig transcript read count data was used to compare changes in relative expression between the datasets. In total, 268,212 and 223,518 reads mapped to an AgamP3.6 and NTR contig transcripts for males and females respectively. Removing transcripts with less than 5 mapped reads reduced the library size to 264,683 and 221,109 for males and females respectively. Libraries were normalised by a factor of 0.9631 and 1.0383 (this global fold change factor is then incorporated into the fold change analysis) for males and females respectively, indicating that they are very similar in composition. Libraries are normalised based upon the total library size i.e. the normalisation factor compensates for the differences int he number of library reads. As no replicates were performed, a common dispersion value of zero was assumed.

In Figure 5.17, a plot of differential expression between male and female read count data is shown. Using a cutoff value of LogFC >2, a total of 799 differentially expressed genes, 385 and 253 for males and females respectively were identified. Transcripts putatively unique to a single sex (i.e. no read count data in the opposite sex) are shown as a smear in the very left of the graph (orange points), giving them an arbitrary low abundance. Log fold change was concentrated around zero, indicating no differential expression. Few genes are both highly abundant within the datasets and differentially expressed, indicated by limited number of genes to the right of the graph. A selection of top ranking candidates are highlighted in Figure 5.17.

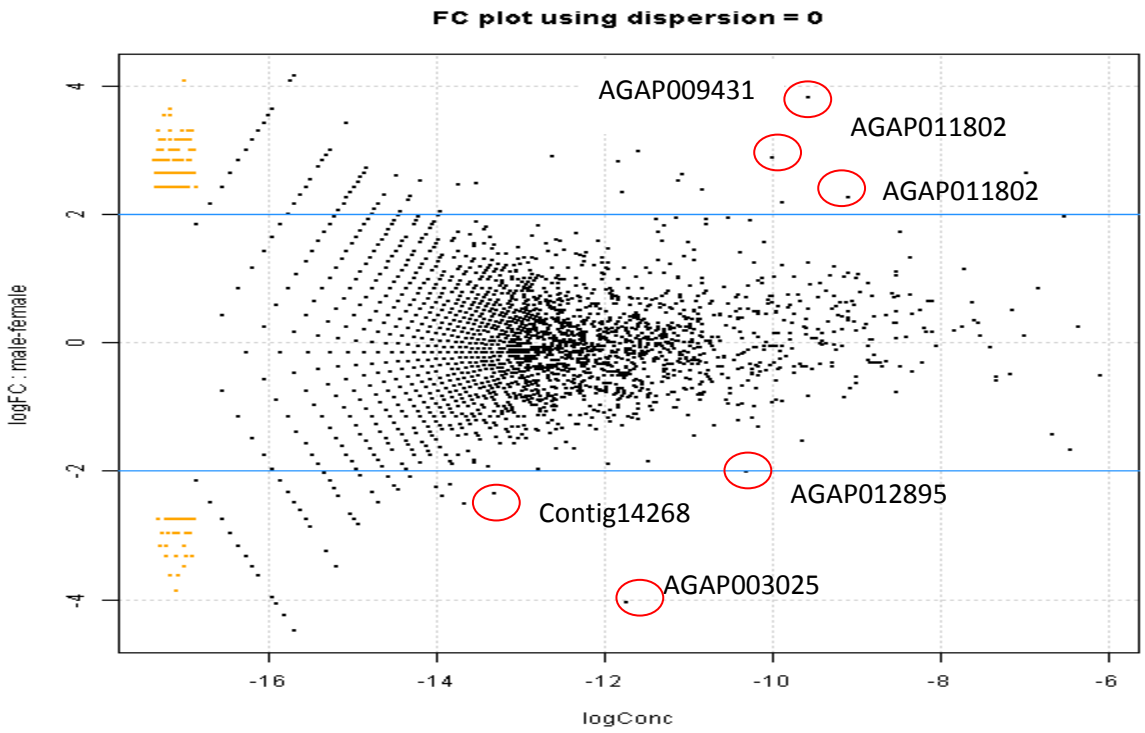


Figure 5.17. Log fold-change (LogFC) expression versus log abundance (logConc) of normalised expression from male and female datasets. Genes with higher expression in males have positive logFC values and genes with higher expression in females have negative values. The x-axis represents the relative abundance of each AgamP3.6 transcript across both male and female libraries. The y-axis is a measure of the differential expression between males and females. The blue horizontal line represents a Log_2FC >2 cutoff. Top candidates for differential expression, with high fold change and abundance within the transcript library are indicated by red circles.

5.3 Discussion

Previously, differences between male and female *An. gambiae* gene expression has been studied in post-embryonic stages. Microarray analysis has shown that gene expression differs between adult tissues (Baker *et al.* 2011), throughout development from 1st instar larvae (Magnusson *et al.* 2011) and within adult midguts (Warr *et al.* 2007). For example, Baker *et al.* (2011) identified a large number of female biased genes associated with hematophagy and Magnusson *et al.* (2011) demonstrated a number of spermatogenesis associated genes that are male biased. Recently, a *de novo* approach using next generation sequencing identified sex-specific differences in transcript expression levels between adult chemosensory appendages (Pitts *et al.* 2011). A global transcriptional study of mixed sex embryos has also been carried out via microarray study (Goltsev *et al.* 2009). Here, we present comprehensive transcript libraries, assembled from 454 derived reads that represent the separate transcriptomes of male and female embryos. In addition we assembled a non-sex specific transcript library that represents the embryonic transcriptome. Therefore this chapter represents the first dissection of both sex and non-sex specific gene expression using a *de novo* approach, to the base pair level in *An. gambiae* embryos.

An assembly pipeline was followed for both male and female reads separately and a non-sex specific approach to generate contig and isotig sequences. Following Newblerv2.3 assembly, a large number of unassembled reads remained for all datasets and these were assembled by MIRA. This strategy has been used previously to assemble Newbler unassembled reads (Kumar and Blaxter 2010; Schwartz *et al.* 2010) and used within our study, resulted in assembly of singletons for each dataset. The recently released Newblerv2.5 has been shown to increase the number of nucleotides included in contigs by 39 % and it has been recommended to repeat assemblies using the later version (Kumar and Blaxter 2010). This finding has been contradicted recently by a relatively modest 1 % increase observed between assemblers (Ewen-Campen *et al.* 2011). Thus we cannot be sure that repeating the assembly process with the latest software would improve the datasets presented here.

Here we employed a *de novo* assembly of the transcriptomes (as opposed to reference-based assembly). The PEST genome, to which alignment would have been undertaken, is a hybrid of the *An. gambiae* M and S molecular forms, that have also recently been

sequenced separately (Holt *et al.* 2002; Lawniczak *et al.* 2010). Whilst the G3 strain is also a hybrid of M and S forms, the strains originate from different geographical locations and have been in laboratory colonies for different lengths of times. As reads are only assembled into contigs if there have over 90 % similarity across a length of 40 bp or greater to another read, incorrect alignment due to genomic differences (e.g. structural variations) between the PEST and G3 strains may therefore prevent accurate contig assembly. In addition, errors in genome assembly can limit the effectiveness of a reference based assembly (Salzberg and Yorke 2005).

A strong advantage of using Newbler is the assembly of isotigs that represent putative full length transcripts. Taking advantage of the availability of the AgamP3.6 predicted transcriptome; an approach was undertaken through comparison of isotigs to AgamP3.6 transcripts. This suggested that over half of isotigs represent greater than 50 % of the transcript it maps to. Using this information, isotigs were demonstrated to contain over 200 insertions to AgamP3.6 transcripts that may represent additional exons, retained introns, partially processed mRNAs or possibly, contaminating genomic sequences. Though RNA samples were treated with DNase, it is possible that intact residual DNA was incorporated into libraries and sequenced. Alternatively, short insertions may represent genomic differences between the G3 strain, from which our samples originiated, and the PEST strain to which isotigs were aligned. This study also allowed identification of a large number of exons, or potentially, novel combinations of exons within transcripts missed from the AgamP3.6 annotation. This was evidenced by isolation of over 2,000 novel canonical splice sites and demonstration that 36.9 % of NTR isotigs contain an intron.

Sampling of the *An. gambiae* embryonic transcriptome using RNA-seq has not been previously performed, so it is therefore not surprising that new transcribed regions have been identified within this study. Of the NTRs, 11.6 % matched an *An. gambiae* EST yet only 5.4 % matched nr database proteins (E-value = $1e^{-3}$) and less than 4 % of NTRs had hits to the protein databases of An. darlingi and and the two culicine species. This somewhat surprising low percentage may be accounted for by several non-mutually exclusive reasons. A proportion of NTRs likely represent either non-coding RNAs (ncRNA) or UTRs of protein coding genes (Wu *et al.* 2008; Wetterbom *et al.* 2010), and thus, their comparisons to protein databases did not yield any hits. Some NTR sequences could have diverged beyond recognition from their orthologs in other organisms. It is also likely that the proteome databases searched are not exhaustive and therefore matching sequences may have been

missed. To further characterise NTRs, full length sequences would be required and could be achieved through RACE or further deep sequencing experiments.

In both mouse and human studies, thousands of putative ncRNA sequences have been identified, of which in mice, over a quarter contain at least a single intron (Okazaki *et al.* 2002; Ota *et al.* 2004). *Drosophila* embryos have been shown to be ncRNA enriched compared to later developmental stages and also, transcripts with very short ORFs that encode short peptides, crucial for embryonic epidermal formation are present (Kondo *et al.* 2010). In fact, a large number of ncRNAs are predicted to play a dual-function role, both as coding sequences and a regulatory ncRNA role (Ulveling *et al.* 2011). Though a majority of NTRs did not match a predicted protein, when NTR contigs and singletons are considered, 10.6 % contained at least a single intron. Presence of introns discriminates these expressed sequences from potential genomic contamination. It also highlights the incomplete annotation of the *An. gambiae* genome and shows that *de novo* NGS can be of value to microarray design, as previously demonstrated using the 454 platform (Bellin *et al.* 2009). Improvement of genome annotation will also have important implications for the wider research community and the transcript resource described here will be made publicly available. All novel sequences and, where available, functional information have been included in a BED file included within the thesis. This file can be viewed using IGV, but also within online viewers such as those at VectorBase. Details for visualisation of the NTRs are included within a README file associated with the BED alignment file.

in silico functional analysis of embryonic transcripts showed a diverse range of gene ontology categories are represented. Gene ontology classification provides gene descriptions that allow putative functions of expressed sequences to be inferred (Conesa *et al.* 2005). Of particular interest to our study was the identification of sex determination-associated genes and we have identified a comparable number of such genes to other transcriptome studies, though *doublesex* and *fruitless* transcripts, known to be involved in *An. gambiae* sex determination, were not sampled. Olafson *et al.* (2010) identified 17 contigs, from a total 12,592 assembled, with sequence similarity to nine *D. melanogaster* sex determination proteins following sequencing of stable fly cDNA. Similary, Gomulski *et al.* (2008) isolated 24 contigs (5,201 total assembled) with sequence similarity to 13 *D. melanogaster* sex determination proteins following sequencing of *Ceratitis capitata* ESTs. Interestingly, the two studies described above used normalised cDNA, a process known to improve detection of low-abundant transcripts (Cheung *et al.* 2006), compared to non-

normalised cDNA use in our study. Despite the demonstration that a majority of our reads originate from a small subset of highly expressed genes, we have detected a comparable number of contigs in our study with sequence similarity to sex determination proteins (350 contigs with similarity to 41 *D. melanogaster* proteins). Therefore, was this study to be repeated we cannot conclude that normalising cDNA prior to sequencing would improve the dection of sex determination genes.

Only 37.4 and 27.4 % of embryo transcripts had a BLASTX match to a predicted protein in the nr database or a GO term respectively. These values are slightly lower than a recent *Ae. aegypti* transcriptome study where 58.5 and 48.5 % of fat body transcripts had a BLASTX match to the nr databse or a GO term respectively (Price *et al.* 2011). That we observed a lower percentage in our study may be attributable to sampling of the embryonic stages, compared to adult stages in the study described by Price *et al.* (2011). As expected, a large number of the embryo transcripts (87.9 %) match An*. gambiae* sequences following BLASTX searches. Some of the AgamP3.6 transcripts are predicted to encode hypothetical proteins; therefore, even if all embryo transcripts have sequence similarity to a AgamP3.6 transcript, not all would be expected to be assigned a GO term.

The male and female transcriptomes were surveyed for genes involved in embryonic development by BLASTP searches using *D. melanogaster* protein sequences as query. This approach demonstrated the valuable information contained within singletons as individual reads had significant sequence similarity to the pair rule genes *deadpan* (E-value = $8e^{-32}$) and *odd skipped* (E-value = $6e^{-42}$). In addition, a putative homologue of *D. melanogaster female lethal d* (*fl(2)d*) has been identified with the associated GO term primary sex determination (GO:0007538). *Fl(2)d* physically interacts with early acting general splicing factors and has been show to interact with *sxl in vitro* (Penn *et al.* 2008). It is also required for the sex-specific splicing of *sxl* and the downstream gene *tra* (Granadino *et al.* 1990; Granadino *et al.* 1996). To further characterise the putative *An. gambiae Fl(2)d* future work should include RACE to determine the full length transcript and then functional RNAi studies initiated to determine if the same function is found in *An. gambiae*. Neither *An. gambiae dsx* or *fru* transcripts were identified in this study, suggesting that these are expressed at too low levels and the depth of sequencing used in this study was insufficient to identify low-level transcript expression. This is supported by the failure to detect *dsx* transcripts from embryos, but the presence of *dsx* transcripts within the adult stages, following sequencing (Illumina) of the *D. melanogaster* embryonic transcriptome by

Graveley *et al.* (2011). As *dsx* transcripts are only expressed in a small subset of embryonic cells (embryonic gonad) (Hempel and Oliver 2007), it may therefore be expected that in our study we would not detect them, due to low-level expression.

Previous studies have identified sex-biased *An. gambiae* gene expression using microarrays. Magnusson *et al.* (2011) showed that a large number of genes are sex-biased in adults (1,752), but relatively few in 1st instar larvae (61). This may be expected as later developmental stages show increasing sexual dimorphism, in the form of morphological and behavioural differences. Within our study, we identified 799 genes differentially expressed between males and females, substantially more than Magnusson *et al.* (2011) identified in early larvae. The reason for this sexually-dimorphic gene expression is unclear, but studies in *D. melanogaster* have shown that embryos express sex-biased genes. Casper *et al.* (2009) identified nine male and one female biased gene in the embryonic germ cells of *D. melanogaster.* In addition, the homeotic gene *abdominal-B* (*abd-B*) and the transcription factor *escargot* (*esg*) are male-biased in *D. melanogaster* embryos (DeFalco *et al.* 2004). Sex-biased genes in the larval genital disc that arises from embryonic germ cells have been shown to be direct targets of female-specific DSX protein (Chatterjee *et al.* 2011). Due to time constraints within our study, no sex-biased genes were characterised. It would have been interesting to carry out functional analysis by RNAi of male and female-biased genes to assess the biological importance of their differential expression. Candidates could be prioritised based upon associated GO terms, such as male-biased AGAP004710 (Log2FC = 3.3) that was assigned a GO term of spermatid development. Further investigation of sex-biased gene expression may improve understanding of how gene expression, from a genome that is largely shared between males and females, can give rise to observed sexual dimorphisms.

It should be noted that sex-biased candidates should be confirmed by quantitative real-time RT-PCR prior to functional work due to several confounding factors. Firstly, female cDNA used for sequencing was derived from two, separate, cDNA amplficiation reactions (to obtain sufficient amounts for sequencing) and may therefore introduce a bias in relative transcript levels. Secondly, read count data for AgamP3.6 transcripts may have been underestimated if the full-length transcript is not currently annotated. Thirdly, though PCR amplification is expected to be linear, female cDNA was amplified for a greater number of cycles and transcript abundances may have been biased. In two recent studies using single NGS experiments, variable validation of 454 read counts to quantitative real time RT-PCR

data was found, one of six (Gregory *et al.* 2011) and six from six (Price *et al.* 2011) differential candidates tested were validated. This suggests the read count data can reflect differential expression but is variable between experiments. Although it should be noted the number of candidates examined in these studies was low.

Alternative splicing is central to dipteran sex determination pathways and the key genes *doublesex* and *fruitless* are known to be sex-specifically spliced in *An. gambiae* embryos (chapter 2). Sex-biased splicing has been documented previously via Cufflinks analysis of NGS data for known sex determination genes including *dsx* and *fru*, as well as a novel gene, *reps* (Graveley *et al.* 2011). In this study, putative sex-specifically spliced candidates have been identified via the software cufflinks and from custom perl scripts. Cufflinks was originally developed for short read (i.e. Illumina) technologies and this may be the reason no reliable candidates were identified using this approach (Trapnell *et al.* 2010). This is demonstrated in the prediction of a ribosomal protein transcript being sex-specifically spliced on the basis of a single read. The software detects transcript levels via reads counts and within this study sufficient read depth has not been produced to use this approach. The use of perl scripts based upon the method of an exon junction expression index (EJEI) previously described revealed a single candidate splicing event (Dong *et al.* 2009). Examination by RT-PCR showed that a putative male specific exon was transcribed in both males and females. Thus, using the two approaches mentioned, no evidence of sex-specific splicing was found within our datasets. Following this work, candidate sex-specific transcription will be searched for through comparison of the male and female transcriptomes.

Chapter 6: Identification of embryonic sex-specific transcripts from RNA-seq data

6.1 Introduction

Regulated gene expression through sex determination pathways is essential for the formation of sexually dimorphic characteristics that define male and female insects. Isolation of genes involved in *An. gambiae* sex determination is therefore of great biological interest and sex-specifically expressed genes are likely to be part of, or targets of this pathway. In is envisioned that comparison of male and female transcriptomes, generated in the previous two chapters, will isolate sex-specific transcriptional events.

Previously, global whole genome expression profiling of *An. gambiae* sex-specific life stages has been conducted using microarray analysis of whole adults (Hahn and Lanzaro 2005; Marinotti *et al.* 2006) and adult tissues including midgut (Warr *et al.* 2007) and multiple somatic and reproductive tissues (Baker *et al.* 2011). Within these studies, sex-biased gene expression was associated with adult sexually dimorphic characteristics. For example spermatogenesis-associated components were up regulated in males (Marinotti *et al.* 2006), blood digestion enzymes such as carboxypepditase and trypsins in female midguts (Warr *et al.* 2007) and genes involved in antigen-mediated defence in females (Hahn and Lanzaro 2005). These studies did not profile early developmental stages, during which gene expression regulating sexual identity originates. Only recently has sex-biased gene expression been analysed as early as first instar larvae, using microarrays (Magnusson *et al.* 2011). Interestingly, Magnusson *et al.* (2011) identified 28 fold fewer sex-biased larval genes compared to adult stages.

Previous whole genome studies isolating sex-biased gene expression utilised microarray platforms that rely upon genes already annotated. RNA-seq offers a *de novo* approach, facilitating sequencing to an unprecedented depth without requirement for *a priori* knowledge of genome annotation. A recent sex-specific gene expression study utilised the *Anopheles* MMC2 microarray platform; containing probes for 7,246 genes of the AgamP3.3 gene build (Magnusson *et al.* 2011). There are currently 13,320 protein coding genes in the AgamP3.6 gene build, a large difference compared to the number assayed in the mentioned study.

Recent comparative RNA-seq studies between male and female transcriptomes have isolated sex-specific gene expression. To do this, reads were assembled, and contigs assembled from reads of only one sex were identified. Using this approach, Hale *et al* 2010

isolated 31 contigs derived from reads of a single sex of the lake sturgeon (Hale *et al.* 2010). Similarly, 20 contigs assembled from male reads only were identified following sex-specific transcriptional profiling of the guppy fish using the 454 platform (Fraser *et al.* 2011). Transcriptome profiling of garter snake male and female tissues identified 195 contigs that were assembled from reads of a single sex (Schwartz *et al.* 2010).

Knowledge of sex-specific gene expression at the embryonic stages is of great biological interest as the sex determination hierarchy establishes sexual identity at this early stage. During the embryonic stages, *dsx* transcripts are sex-specifically spliced in *D. melanogaster* (Hempel and Oliver 2007), *Ceratitis capitata* (Gabrieli *et al.* 2010), *Aedes aegypti* (Salvemini *et al.* 2011) and *An. gambiae* (chapter 2). In *D. melanogaster*, *dsx* expression is required for male-specific expression of *sox100B* in the developing embryonic gonad and female *dsx* null mutants express male specific sox100B (DeFalco *et al.* 2003). In addition to *sox100B*, *male germ line marker 1* (*mgm1*), *unpaired* (*upd*) and *disc proliferation abnormal* (*dpa*) are all expressed exclusively in male developing gonads (Staab *et al.* 1996; Wawersik *et al.* 2005). *D. melanogaster* sexually dimorphic embryonic gonads form at stage 15, 12 hrs after fertilisation (Casper and Van Doren 2006). Therefore comparison of male and female *An. gambiae* embryonic transcriptome, from two developmental time points, may be expected to identify sex-specific gene expression instructive for the formation of the sexually dimorphic gonads.

Isolation of sex-specific transcription is also of great interest to development of novel genetic control strategies. Conventional sterile insect technique (SIT), where irradiated sterilised males are released into a wild population, has been suggested as a vector control option for over 50 years (Knipling 1955). Successful mating of sterilised males with wild females would result in no offspring, leading to population suppression. Issues regarding fitness of irradiated males and undesirable release of biting females have limited the feasibility of this approach (Lofgren *et al.* 1974; Sharma *et al.* 1978). More recently, genetic sterilisation of vector populations though a release of insects carrying a dominant lethal (RIDL) has been proposed (Alphey and Andreasen 2002; Coleman and Alphey 2004). Transgenic RIDL insects carry a conditional, dominant lethal gene that is repressed during laboratory rearing with ac chemical not occurring in nature. Upon release into the wild, offspring from successful matings between RIDL males and wild females will not survive due to absence of the repressor. A proof of principle RIDL study has recently shown population suppression of an island population of *Ae. aegypti,* following release of 25,000

transgenic males within a 10 hectare field study site (Harris *et al.* 2011). One limitation of this approach was the requirement of manual separation of males and females at the pupal stage.

Generation of a transgenic strain that allows genetic separation (genetic-sexing) of males and females would be advantageous over manual separation. *An. gambiae* male-specific promoters have been used to drive expression of a fluorescent reporter gene within pupal testes (β-*tubulin*) allowing sex-separation (Windbichler *et al.* 2008). Previously, an *Ae. aegypti* female-specific flight muscle *actin-4* promoter has been used to drive late lethality in a RIDL control system (Fu *et al.* 2010). Female-specific expression of the tetracycline transactivator protein (tTA) (elevated levels of the tTA subunit VP16 leads to cellular toxicity) in the indirect flight muscles prevented females from flying. Release of males carrying the female specific promoter would result in effectively sterile female offspring as they would be unable to fly and copulate. Using this approach, separation of males for release from females was possible at 4th instar larvae through expression of a fluorescent reporter from the *actin-4* promoter. Late separation of sexes increases rearing costs; costs that would increase markedly in the context of mass release of millions of individuals. Therefore isolation of an early embryonic promoter that is differentially active between males and females would greatly reduce these costs.

Sex chromosome composition of dipteran insects is the primary signal that controls the sex determination pathway. In *An. gambiae*, the male is the heterogametic sex (XY), whereas females are homogametic (XX). The *An. gambiae* Y chromosome is believed to be largely degenerate, containing satellite repeat families, transposable elements and few actively transcribed sequences (Rohr *et al.* 2002; Krzywinski *et al.* 2004; Krzywinski *et al.* 2005). Previous studies of other dipteran heterogametic Y chromosomes have isolated genes integral to male fertility in *D. melanogaster* (Carvalho *et al.* 2000; Sackton *et al.* 2011) and male promoting factors in *C. capitata* (Willhoeft and Franz 1996) and *M. domestica* (Schmidt *et al.* 1997). Through comparison of sex-specific RNA-seq datasets that represent global abundance of poly(A)+ transcripts, expressed sequences may be identified from the heterogametic Y chromosome.

6.1.1 Specific Chapter Aims:

- Comparative bioinformatics approaches to isolate sex-specific transcripts from the current AgamP3.6 gene build and from new transcribed regions identified in previous chapters
- Candidate sex-specific transcripts were the prioritised based upon read count data and absence of hits in putative untranslated regions and then were validated through RT-PCR amplification from sex-specific embryonic RNA.
- Investigate sex-specific candidates validated by RT-PCR to infer a putative function/reason for sex-specific expression in the embryonic stages through RNAi induced reduction in transcript levels.

6.2 Results

6.2.1 Identification of sex-specific candidates: Read count analysis from the AgamP3.6 transcriptome

In this and all subsequent sections, only cleaned reads from chapter 4 were used in analysis and are hereafter referred to simply as reads. Separately, all embryo transcripts for both male and female datasets were used as query for BLASTN (E-value = $1e^{-5}$) searches against the AgamP3.6 transcriptome. One significant matching contig or singleton was considered sufficient to deem a transcript from the current gene build as present within our datasets. AgamP3.6 transcripts present in male and female datasets were compared and a total of 1,527 and 745 were identified as putatively sex-specific for males and females respectively, corresponding to 1,488 and 733 genes. Due to a high number of identified candidates, putative sex-specific transcripts were prioritised by read count data. This was done as the greater the number of reads that produce a significant hit following BLAST searches to each sex-specific candidate, the greater the evidence that they are true sex-specific transcripts. Nucleotide sequences of candidates were downloaded from VectorBase that were extended to contain 1 Kb upstream and downstream regions. This was done in anticipation that full UTRs are not annotated within the current gene build, but may be present in the embryonic transcriptome. Reads that align to a UTR would indicate that a transcript is expressed and therefore they must be included in the analysis.

Sex-specific read datasets (chapter 4) were mapped to the male (1,527) and female (745) candidate sex-specific AgamP3.6 transcripts using Newblerv2.3 gsMapper and the number of unique mapped reads extracted from the 454Refstatus.txt output file (example shown in appendix). Using this approach, nucleotide sequences of selected candidate transcripts were downloaded from NCBI and mapped to the genome using BLASTN searches to establish their genomic organisation. This was done so that primers to amplify a fragment of each candidate could be designed. In total, 15 and 20 primer pairs were designed for male and female AgamP3.6 transcript candidates, flanking an annotated intron where possible. The number of unique mapped reads for each candidate is shown in Table 6.1 and Table 6.2 for males and females respectively. All candidates were examined *in vitro* by RT-PCR amplifying RNA from male and female embryos 20 hr AEL. For all candidates tested, a fragment of the expected size was amplified from both male and female RNA. Therefore,

no candidate AgamP3.6 transcripts were confirmed to be sex-specific. A representative fraction of the candidates tested are shown in Figure 6.1.

Table 6.1. Summary of male specific AgamP3.6 transcript candidates. Shown are the candidate AgamP3.6 transcripts and the number of male reads that map to the transcript. Also shown are the primer sequences and the expected size of the amplification product from RNA and DNA templates. No female reads mapped to the transcript sequence that was extended to contain 1 Kb upstream and downstream regions.

Candidate AgamP3.6 Transcript	Expected RNA amplification product size (bp)	Expected DNA amplification product size (bp)	Number of Aligned reads
AGAP005551-RA	512	661	10
AGAP000209-RA	238	314	6
AGAP001412-RA	224	4916	2
AGAP005398-RA	194	286	1
AGAP012308-RA	513	513	1
AGAP012920-RA	310	444	1
AGAP001591-RA	304	369	5
AGAP007474-RA	831	4223	5
AGAP008132-RA	528	1114	7
AGAP005234-RB	220	324	11
AGAP008684-RA	918	986	4
AGAP009475-RA	278	5206	1
AGAP012308-RA	425	864	9
AGAP009481-RA	555	651	2
AGAP003577-RA	372	595	4

Table 6.2. Summary of female specific AgamP3.6 transcript candidates. Shown are the candidate AgamP3.6 transcripts and the number of female reads that map to the transcript. Also shown are the primer sequences and the expected size of the amplification product from RNA and DNA templates. No male reads mapped to the transcript sequence that was extended to contain 1 Kb upstream and downstream regions.

Candidate AgamP3.6 Transcript	Expected RNA amplification product size (bp)	Expected DNA amplification product size (bp)	Number of Aligned reads
AGAP005260-RA	595	957	1
AGAP011790-RB	757	861	7
AGAP000213-RA	605	746	10
AGAP003334-RA	283	467	5
AGAP003420-RA	321	407	6
AGAP000065-RA	557	722	5
AGAP011948-RA	701	2515	5
AGAP001563-RA	769	2787	4
AGAP008018-RA	404	590	1
AGAP007646-RA	799	1358	2
AGAP009731-RA	647	725	1
AGAP006913-RA	296	376	1
AGAP000693-RA	159	247	1
AGAP008288-RA	410	665	1
AGAP006479-RA	283	2500	4
AGAP010708-RA	308	403	4
AGAP009549-RA	552	895	6
AGAP010036-RA	399	4758	5
AGAP011604-RA	343	343	9
AGAP012221-RA	574	1087	2

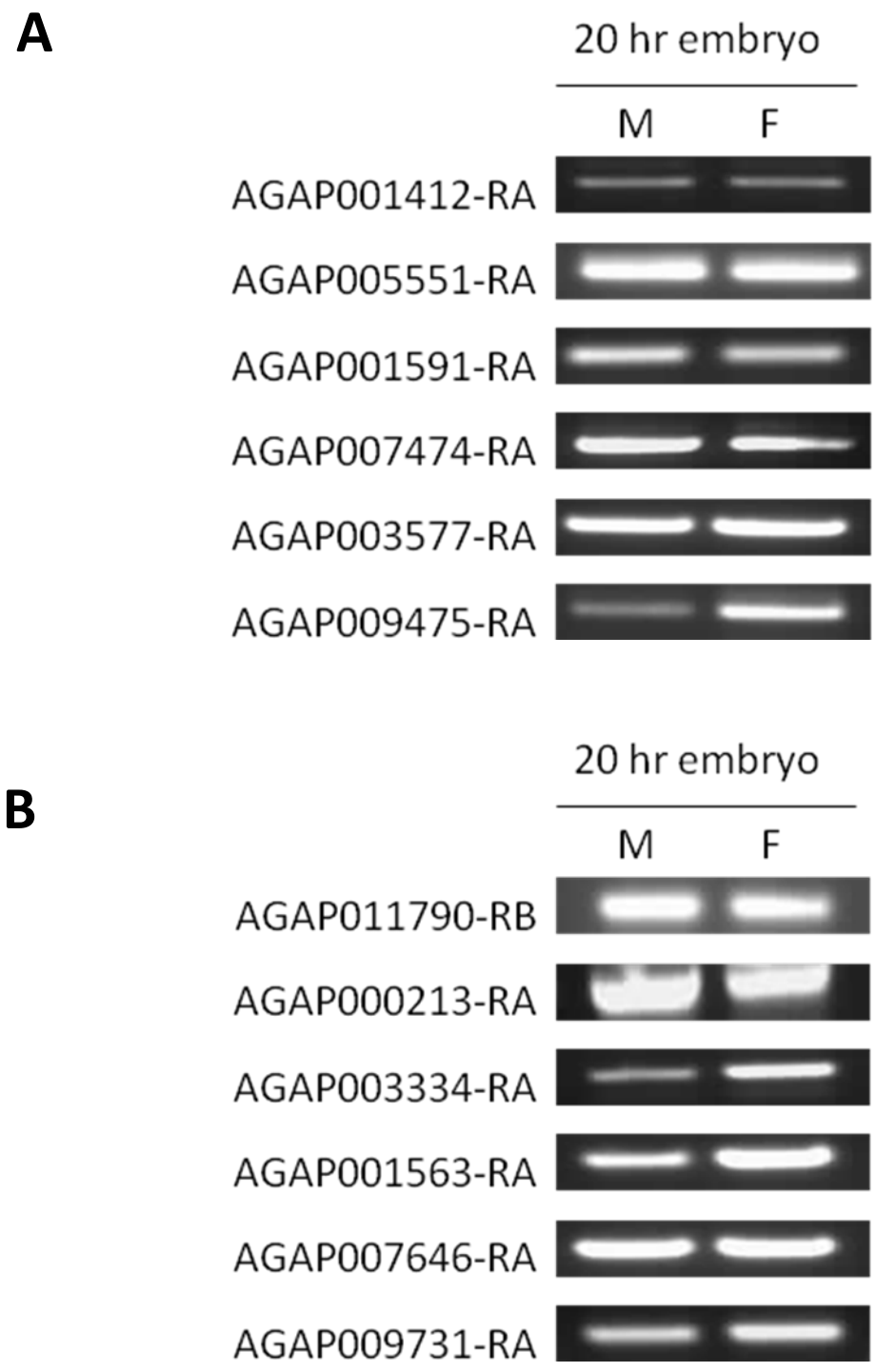


Figure 6.1. Candidate male and female specific AgamP3.6 transcripts examined by RT-PCR. For each sex-specific candidate, primers were designed following BLASTN searches of transcript sequences to the genome to establish their genomic organisation. Primer pairs were used to amplify RNA from male and female embryonic RNA 20 hr AEL by RT-PCR. (A) For all male-specific candidates tested, fragments of the expected size were amplified from both male and female RNA. (B) For all female-specific candidates, fragments of the expected size were amplified from both female and male RNA.

6.2.2 Identification of candidates: Read count analysis from new transcribed regions

Within chapter 4, it was shown that the assembled RNA-seq datasets contained new transcribed regions (NTRs) that may represent previously un-annotated transcripts, or fragments of transcripts. Therefore, the approach taken in 6.3.1 may fail to identify sex-specific transcripts as only currently predicted transcripts from the AgamP3.6 gene build were considered.

To include all sequence data from this study, we attempted to isolate male and female contigs from chapter 4 (including those assembled by Newblerv2.3 and MIRA) that were only present in a single sex. Using standalone BLASTN (E-value = $1e^{-5}$) searches, male reads were used to query female contigs. The same BLAST approach was then repeated but with female reads being used to query male contigs. It was anticipated that absence of reads from the opposite sex aligning to a contig would indicate that it is putatively sex-specific. Contigs with no significant hit were extracted from the BLASTN search results and in total, 1,300 and 1,854 contigs had no sequence similarity to reads of the opposite sex for males and females respectively. Read count analysis was performed for each candidate as described in Materials and Methods. Imposing an arbitrary five read cutoff value, the number of candidates was reduced to 219 and 197 for males and females respectively. A five read cutoff was imposed as it was expected that this or lower number of reads does not provide strong enough evidence for sex-specific transcription.

Using this approach, nucleotide sequences of candidate contigs were mapped to the genome using BLASTN searches to establish their genomic organisation. This was done so that primers to amplify a fragment of the contig could be designed. Primers were designed for six male candidates, flanking introns where possible (Table 6.3). All candidates were examined *in vitro* by RT-PCR amplification of sex-specific 20 hr AEL embryonic RNA. In total, the expected amplification product for five candidate male specific contigs was amplified from both male and female embryonic RNA. For a single contig, contig08484, a fragment of the male expected size was amplified from male RNA only.

BLASTN searches revealed that contig08484 is located on the *An. gambiae* Y chromosome in NCBI scaffold NW_045557 (hereafter Y scaffold) and also revealed that no other similar sequences are present on any other chromosome. BLASTN searches also revealed that all reads originating from the Y chromosome were contained within contig8484 and its

associated isotigs. Contig08484 was assembled into two different isotigs as part of the Newbler assembly process, isotig02394 and isotig02395. The two isotigs sequences were compared to the genome to establish their genomic organisation (Fig 6.2A). Isotig02394 and isotig02395 differ in their exon/intron structure. Both share a common 295 bp first exon but isotig02395 contains an additional 88bp intron within the second exon of isotig02394. Primer pair 08484F/08484R (shown in Fig. 6.2A) was designed flanking the common intron of both isotig02394 and isotig02395 and was expected to amplify a 420 bp male specific fragment. Subsequent RT-PCR amplification revealed the 420 bp fragment within male sample only (Fig. 6.2B). Following amplification, the 420 bp fragment was cloned following gel electrophoresis and sequencing confirmed the identity of the PCR product.

Table 6.3. Summary of male specific NTR contigs. Six candidate male-specific NTR contigs and the number of male reads used to assemble each contig is shown. Shown are the primer sequences and the expected size of the amplification product from RNA and DNA templates. No female reads were used in the assembly of the selected contigs.

Name	Location (chromosome)	Number of Assembled Reads	Expected RNA amplification product size (bp)	Expected DNA amplification product size (bp)
contig01125	(2L) AAAB01008960.1 10284026-10283913 10279860-10279533	6	374	4430
contig03312	(2R) AAAB01008987.1 1509144-1509223 1511950-1512216	8	312	3038
contig04026	(3L) AAAB01008966.1 1260671-1260723 1261724-1261981	7	264	1266
contig06155	(2L) AAAB01008960.1 4742303-4742054 4740066-4739917	6	352	2340
contig08484	(Y) NW_045557 48140-47759 26724-27105 26269-26568 48385-48296	13	420	575
contig09711	(3L) AAAB01008966.1 1636806-1636583 1635271-1635164	12	262	1573

6.2.3 Contig08484 expression profile

Using primer pair 8484F/8484R a temporal profile of contig08484 expression was conducted using RT-PCR upon mixed sex embryonic RNA staged every 30 min AEL (Fig. 6.2C) and sex-specific embryonic RNA staged every 4 hr AEL (Fig. 6.2D). Mixed sex embryonic RNA was used for early stages, as detection of male and female diagnostic banding patterns following PCR (chapter 3) in early stage embryos proved inconclusive. As such, mixed-sex embryonic RNA was used as previously we had shown that contig08484 was located on the Y chromosome. A product of the expected size was amplified from mixed embryonic RNA 3 hr AEL. Male specific products of the expected size were amplified from RNA of male embryos 4 to 36 hr AEL, the latest time point examined. No product of the expected male size was amplified from any embryonic female RNA. A temporal profile of contig08484 expression in post-embryonic stages was then conducted; male specific products of the expected size were amplified using PCR amplification from 1^{st} strand cDNA of 1^{st} instar larvae through to adulthood (Fig. 6.23E). Using RT-PCR amplification from adult male tissue RNA, a fragment of the expected size was amplified in all tissues tested, including somatic and reproductive tissues (Fig. 6.2F). Again, no amplification product of the expected male size was observed from female samples. A 550 bp fragment was also amplified from both male and female embryonic RNA using primer pair 8484F/8484R. Cloning and sequencing showed that the fragment was a result of non-specific amplification from chromosome arm 3R (scaffold CRA_x9P1GAV5DHF: 10,953,496 – 10,953,012).

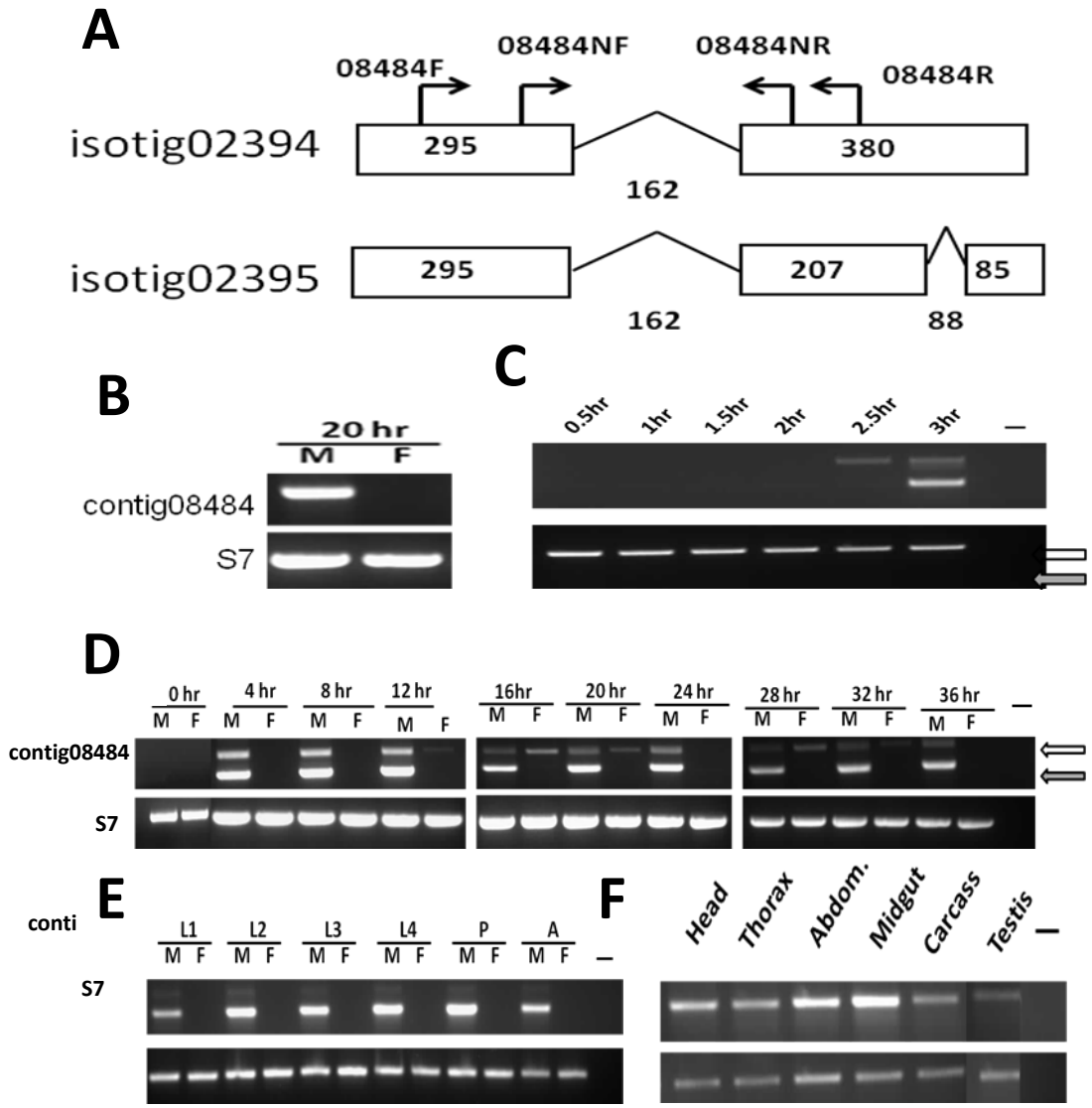


Figure 6.2. Schematic representation of contig08484 and temporal expression profile. (A) Genomic organisation of isotig02394 and isotig02395. Arrows indicate position of primers relative to the transcript. Open boxes represent exons, diagonal lines introns and numbers length in base pairs. Each life stage was examined by RT-PCR or PCR using primer pair 8484F/8484R for expression of contig08484. (B) RT-PCR amplification from sex-specific 20 hr AEL embryonic RNA revealed male specific expression. (C) RT-PCR amplification from mixed sex embryo RNA, 0.5 to 3 hr AEL. A fragment of the expected size was amplified 3 hr AEL. (D) RT-PCR amplification from sex-specific embryonic RNA staged every 4 hr AEL. Male specific expression of contig08484 is first detected at 4 hr AEL and continues through to 36 hr AEL. (E) PCR amplification of male specific contig08484 transcripts from 1st strand cDNA of post-embryonic stages. Amplification of a fragment of the expected size is male specific from 1st instar larvae through to adulthood. (F) RT-PCR amplification from male adult tissues produced a fragment of the expected size in all tissues examined. First (L1), second (L2), third (L3), fourth (L4) instar larvae; P = pupae; Ad = adult; M = male; F = female. S7 = *An. gambiae* ribosomal protein S7 transcripts (Accession L20837) served as a loading control. (-) represents negative control (no RNA or 1st strand cDNA template added for RT-PCR and PCR respectively). The open arrow indiciates the expected 420 bp fragment and the grey arrow the non-specific 550 bp amplification product.

6.2.4 Rapid Amplification of cDNA Ends (RACE) of contig08484

To isolate full length sequences of contig08484, RACE experiments were performed. Firstly, total adult male RNA (1 µg) was used to generate 5' and 3' RACE ready cDNA using SMARTscribe reverse transcriptase (RT). Gene specific primers used are shown in Figure 6.3A. To increase specificity of RACE reactions, nested PCR reactions were performed using a nested universal primer (NUP) and transcript specific nested primers, also shown in Figure 6.3A.

RACE experiments were performed using 3 and 5' ready RACE cDNA, the universal primer mix (UPM) and contig08484 specific primers in the below combinations:

1) 5' cDNA/ Universal primer mix / 08484R
2) 3' cDNA/ Universal primer mix / 08484F

A single visible product was amplified for both reactions (5' = 750bp and 3' = 300bp) following analysis by gel electrophoresis. Nested PCR was then carried out using diluted unpurified PCR product as DNA template using the following combinations:

3) Purified 5' PCR product/NUP/08484NR
4) Purified 3' PCR product/NUP/08484NF

In total, one 5' and five 3' fragments were amplified, cloned and sequenced to establish their identity. Amplified fragment sequences were mapped to the Y scaffold using BLASTN (bl2seq). Analysis of the direction of splicing signal sequences (GT/AG) found at the termini of introns revealed that transcription was occurring from both forward and reverse DNA strands. Therefore, an additional RACE experiment was conducted, reversing the primer combinations used above to ensure that termini of all transcripts involved were isolated:

1) 5' cDNA/UPM/contig8484F <u>followed by:</u>
2) 5' PCR product/NUP/contig8484NF

3) 3' cDNA/UPM/contig8484R <u>followed by:</u>
4) 3' PCR product/NUP/contig8484NR

PCR reactions produced a smear of no distinct amplification products following initial amplification and gel electrophoresis (1 and 3); subsequent nested PCR amplification (2 and 4) produced distinct bands that were purified following gel electrophoresis. In total, three

5’ and three 3’ PCR fragments were cloned following purification and then sequenced to establish their identity. All sequences were mapped to the genome using BLASTN searches to reveal their exon/intron structure. This revealed presence of two transcripts encoded on the forward DNA strand and four transcripts on the reverse strand (Fig. 6.3B). Under this assumption, a further round of RACE experiments was undertaken to ensure that full length sequences of each putative transcript had been identified. Primers were designed to be specific to each transcript where possible and are shown in Figure 6.3C. No additional fragments from transcripts on the reverse DNA strand were amplified, suggesting that the complete structure of transcripts had been identified (Fig. 6.3C). An additional transcript in the forward orientation was identified, but no transcript was extended from previous annotation (Fig. 6.3C).

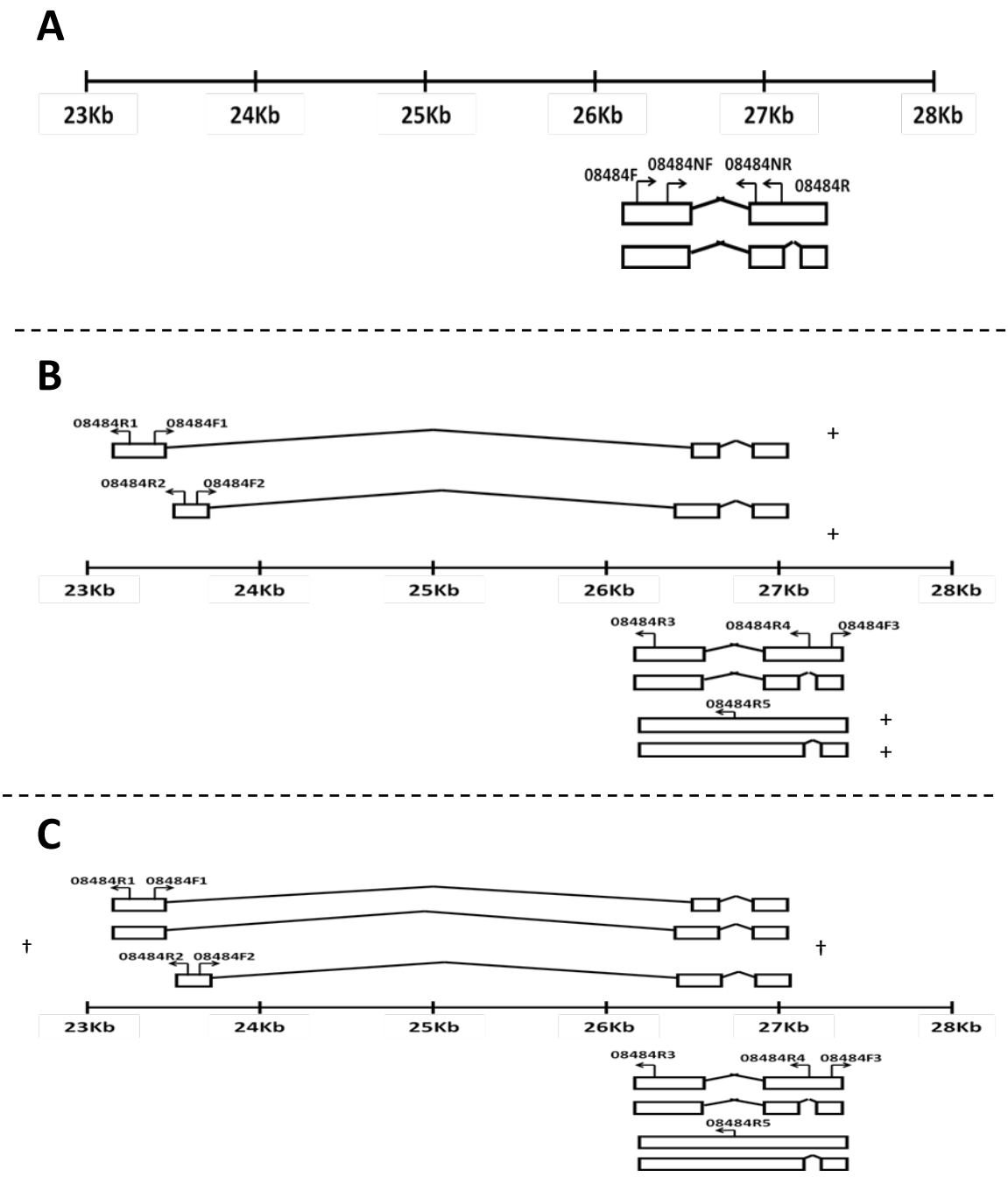


Figure 6.3. Schematic diagram of extension of contig08484 by RACE. Shown in each is a fragment (23 to 28 Kb) of a Y chromosome 48 Kb scaffold (NW_045557). Open boxes represent exons, diagonal lines introns and arrows indicate the position of primers used for amplification. Primers shown in each diagram were used to establish the additional sequences depicted in the subsequent diagram i.e. Sequenced PCR products amplified using primers shown in A established the genomic organisation in B. (A) Genomic organisation of isotig02394 and isotig02395 and location of primers used for initial RACE extension. (B) Mapping of RACE PCR fragments to the genome revealed that two and four transcripts are present in the forward and reverse direction respectively. (C) RACE experiments demonstrated that an additional transcript was present in the forward orientation. + symbol indicates transcripts identified from first RACE experiments. † symbol indicates transcripts identified in the second RACE experiment. Transcripts with no symbol represent those whose genomic organisation was already known.

6.2.5 YOA and YOB genes

The three transcripts in the forward orientation were collective grouped and denoted as transcripts of the gene YOA (Y originating A). Similarly, the four transcripts in the reverse orientation were grouped as YOB (Y originating B) transcripts and the exon/intron structure of the two genes is shown in Figure 6.4.All YOA transcripts consist of three exons spanning 4 Kb of the Y scaffold and YOA-A and YOA-B only differ by a 51 bp extension of exon two. They share the same 258 bp first and 73 bp third exon. YOA-C contains a transcript specific 96 bp first exon and shares exons two and three of YOA-B. The 4 YOB transcripts are found within a 1.5 Kb fragment of the Y scaffold and consist of one (YOB-C), two (YOB-A and YOB-D) or three (YOB-B) exons. The second and third exons of YOA transcripts overlap all YOB transcripts on the opposite strand.

6.2.6 Expression profile of YOA and YOB gene transcripts:

Primer pairs YOA-F1/YOA-R and YOA-F2/YOA-R were used to amplify YOA-A/B and YOA-C respectively (Fig 6.4A). Expected amplification product sizes of YOA-A and YOA-B differ by 51 bp and could therefore be discriminated. All YOB transcripts were amplified with primer pair YOB-F/YOB-R and discriminated between by PCR product size. RT-PCR was carried out upon mixed sex (30 to 180 min AEL) and male-specific (4 to 36 hr AEL) embryonic RNA. In addition, PCR was carried out on male-specific 1st strand cDNA (1st instar larvae through to adulthood) and male-specific tissues 1st strand cDNA. All primers used and expected size of amplification products are shown in the Appendix.

Fragments of the expected size were not amplified for any YOA transcripts from RNA of embryos 30 to 180 min (Fig. 6.4B) or 4 to 36 hr AEL (Fig. 6.4C). Expression of YOA-A and YOA-B transcripts begins during 3rd instar larvae and continues through to adulthood and across all adult tissues tested (Fig. 6.4D-E). Amplification products using for YOA-C were not detected in any of the life stages examined using this approach. Fragments of the expected size for YOB-A and YOB-B were amplified from 180 min AEL embryonic RNA (Fig. 6.4B). Correct size amplification products were also observed from embryonic RNA 4 to 36 hr AEL (Fig. 6.5C) and RNA from 1st instar larvae through to adulthood (Fig. 6.4D). From all male tissues examined, YOB-A and YOB-C amplification products were present. YOB-D transcripts were amplified from head, thorax and testis tissues (Fig. 6.4E). YOB-C transcripts were not detected in any of the life stages examined using this approach.

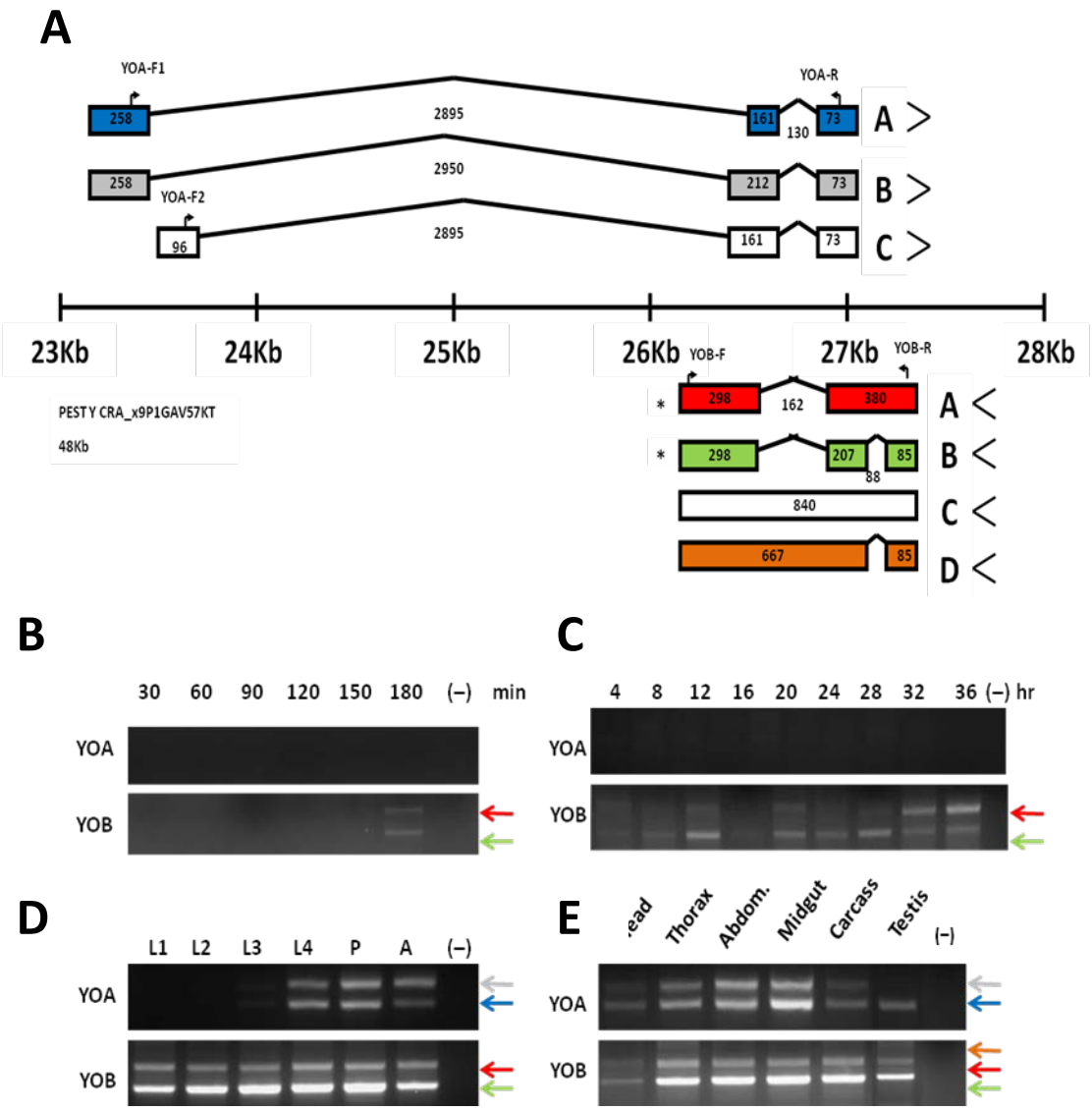


Figure 6.4. Genomic structure of two Y-chromosome genes identified via RACE experiments. Genomic organisation of YOA and YOB identified through RACE experiments and RT-PCR temporal expression profiles of their associated transcripts. (A) Three alternative splice forms of the YOA and the four transcripts of YOB are shown located on scaffold NW_045557. Arrows indicate the position of primers used to amplify transcripts. In B to E: Expression of YOA and YOB transcripts in (B) 30 to 180 min AEL embryos (C) 4 to 36 hr AEL embryos (D) 1st instar larva to adult stages (E) selected adult male tissues. Colour of arrows corresponds to colour of transcripts shown in A. Arrows indicate the position of primers used for amplification by RT-PCR. Open boxes represent exons, diagonal lines introns and numbers length in base pairs. Horizontal line (-) represents negative control, no RNA or cDNA template added for RT-PCR and PCR respectively.

6.2.7 Repetition of YOA and YOB gene transcripts

BLASTN searches of the genome using the complete YOA and YOB transcripts identified a truncated repetition of the two genes on the same Y scaffold (Fig. 6.5). Full length transcript sequences were found as shown in Figure 6.4 but also truncated transcripts, caused by the end of the scaffold sequence, were found in the opposite orientation. The third exon and part of the second exon of all YOA transcripts matched to the edge of the scaffold on the reverse strand, but the first exon is not found for any of the transcripts due to lack of sequence information past the edge of the scaffold. YOB transcripts are all truncated by 215 bp as they map to the very end of the scaffold and no more sequence information is available.

6.2.8 Characterisation of YOA and YOB gene transcripts

To characterise YOA and YOB transcripts, firstly, full length transcripts were used as query for BLASTN (E-value = $1e^{-10}$) searches against 217,297 *An. gambiae* ESTs (database AgESTDec10). Though hits with E-values above the threshold were identified using YOA transcripts as query, none of them correspond to YOA transcripts, as indicated by low sequence identity within the matches. YOB transcripts had no significant hits.

Previously, *An. gambiae* adult transcriptomes have been sequenced (~ 21 million readsin total) using the Solexa (Illumina) platform (Gibbons *et al.* 2009). This dataset was downloaded from the NCBI short read archive (SRA) (Accession: SRX014416) and the reads were searched against the Y scaffold (NW_045557) to identify ESTs of Y chromosome origin using BLASTN. Matching reads were then used for reference based transcript assembly by Cufflinks software following mapping to the genome using GMAP. This produced a further seven putative transcripts, or transcript fragments, aligning to the Y scaffold (Fig. 6.5). Location on either the forward or reverse DNA strand of the Solexa assembled transcripts could not be determined. Four of the transcripts partially overlapped the YOA and YOB transcripts, with the remaining three occurring in introns or upstream of the two genes. Secondly, BLASTN searches were used to query scaffolds of the M and S form genomes to establish whether YOA and YOB transcripts had a significant match to the two genomes. All of the transcripts had a significant match (E-value > $1e^{-30}$) to genomic scaffolds of the M genome (scaffold EQ088672.1) and for S genome (scaffold EQ098449.1).

Thirdly, all YOA and YOB gene transcripts were searched for putative open reading frames (ORF) using EMBOSS:getorf. All transcripts had an ORF greater than 175 nucleotides with a

maximum of 347 nucleotides for YOB-D. Fourthly, TBLASTN searches were also conducted using six-frame translations of all transcripts as query sequence against genomic sequences of *Culex quinquefasciatus* (CpipJ1.2), *Ae. aegypti* (AaegL1.2) and *An. darlingi* (NCBI accession PRJNA46227). This revealed no significant similarity to any nucleotide sequences of the mosquito species examined (E-value = $1e^{-3}$).

Fifthly, six-frame translations of all transcripts were used for BLASTX searches against the NCBI nr protein database (E-value = $1e^{-3}$). No significant matches were found using this approach. SignalP4.0 analysis was used to identify any signal peptide cleavage sites for six frame translations of YOA and YOB transcripts and this analysis did not identify any signal peptide motifs. Finally, the coding potential of all transcripts was assessed using the Coding Potential Calculator as described in Materials and Methods. Using this approach, YOA-A and YOB-B were predicted as non-coding, whereas the remaining transcripts for both YOA and YOB were assigned weak non-coding classification.

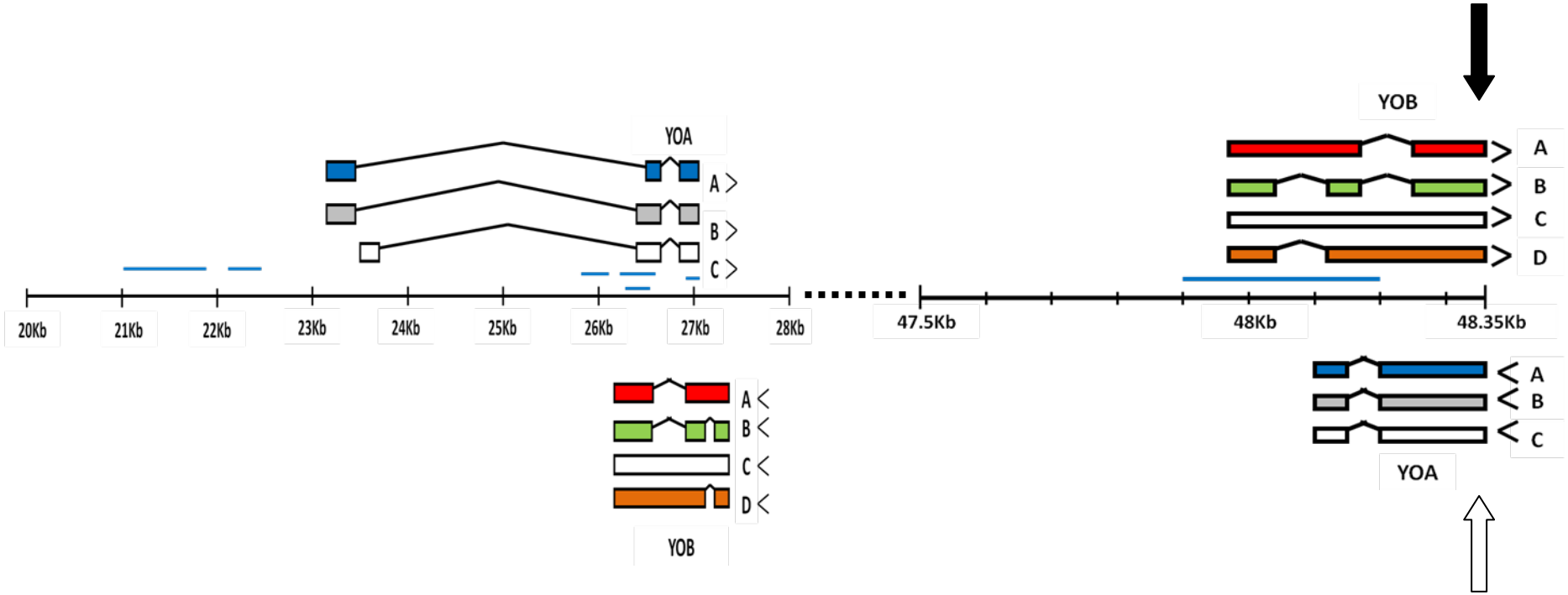


Figure 6.5. Genomic organisation and repetition of YOA and YOB at the extremity of the Y scaffold. YOA and YOB transcripts map to two locations on the 48.35 Kb Y scaffold NW_045557. Full length transcripts are found between 23 and 28 Kb of the scaffold and truncated versions are found at 48 Kb. Black arrow = 215bp truncated region of YOB transcripts. Open arrow = truncated area of YOA transcripts, missing the first exon. Duplicatio of both genes at the extremity of the scaffold are in the reverse orientation to the full length transcripts found between 23 and 28 Kb. Red lines indicate Cufflinks assembled transcripts in the forward orientation and blue lines in the reverse orientation from previously published Solexa ESTs (Gibbons *et al.* 2009).

6.2.9 Functional Characterisation of contig08484

dsRNA corresponding to the region of contig08484 amplified by primer pair 08484F/08484R was synthesised (ds8484). As contig08484 is male specific throughout development and is expressed in the early zygote, it was envisioned that it may function as a male sex promoting factor. Therefore, it was hypothesised that knockdown of contig08484 would result in sex-reversal of successfully injected males. Embryo microinjection was performed for ds8484; dsβ-lactamase gene dsRNA was used as control. dsRNA was co-precipitated with the GFP expression plasmid FU III prior to injection. Plasmid FU III is a fusion of the polyubiquitin (AGAP001971) promoter to the protein coding sequence of the fluorescent reporter GFP. Observation using a fluorescent microscope of GFP protein expression was used as an indication of successful dsRNA delivery. Hatched larvae were screened for GFP expression and GFP positive larvae were reared to adulthood.

The injection process was carried out and is summarised in Fig. 6.6A. In the experiment, 310 and 261 embryos were injected for YOA/YOB and β-lactamase control respectively. Hatch rates for both samples were comparable (20-27%) and the percentage of hatched larvae expressing GFP was 9.2 and 2.8 for YOA/YOB and β-lactamase respectively. In total six individuals (three male and three female) expressed GFP following injection with ds8484 and FU III. For dsβ-lactamase control injections, two (one male and one female) hatched larvae expressed GFP. Emerging adults were phenotyped for their sex and DNA from a leg of GFP positive individuals was used as template for PCR to confirm individual karyotypes using primer pair's 8484NF/8484NR and 128125BF/128125R. Primer pair 128125F/128125R amplifies repetitive fragments from the Y chromosome. It was therefore expected that amplification using the two primer pairs would only amplify a product of the expected size from XY karyotype individuals. Only those individuals displaying a male phenotype contained the male fragment of expected size following gel electrophoresis. All phenotypic females showed no evidence of a XY karyotype, indicated by an absence of the male product of the expected size (Fig. 6.6B).

A

			GFP negative		GFP positive		GFP positive
	Number Injected	Hatch Percentage	Number Male	Number Female	Number Male	Number Female	Hatch percentage
8484	310	20.6	35	24	3	3	9.2
β-lactamase	261	27.2	40	28	1	1	2.8

B

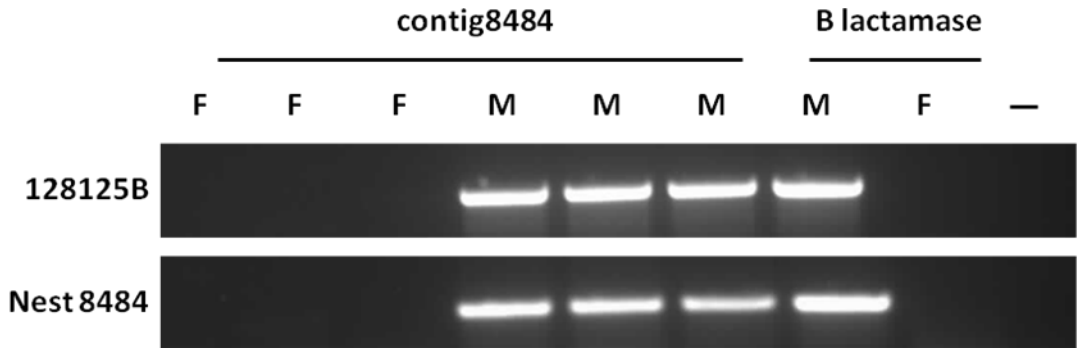


Figure 6.6. Summary of RNAi contig08484 experiments. Two replicates of RNAi injections were performed in pre-blastoderm embryos and GFP positive hatched larva were examined for phenotype changes. Phenotyped adults were used for PCR karotyping using priemr pair 37(B) and 8484NF/8484NR. (A) table summarising the number of injected embryos and the percentage showing GFP expression. (B) Kayrotype analysis of GFP positive phenotype adult individuals with two Y chromosome specific primers for experimental and control injected. (C) Examples of GFP positive individuals. M = male; F = female; (-) horizontal line represents negative control where no genomic DNA template was added.

6.3 Discussion

Two transcript libraries, generated from male and female embryo RNA were compared to identify putative sex-specific transcripts that may improve understanding of the control of *An. gambiae* sexual dimorphism. Additionally, it was hoped that promoters driving differential expression between males and female may be identified that could lend themselves to a genetic sexing strain. This analysis revealed a single contig (08484) that displayed male-specific expression, validated by RT-PCR analysis from sex-specific embryonic RNA. RACE experiments have shown that contig08484 is actually part of two new complete genes on the *An. gambiae* Y chromosome, termed within this study as YOA and YOB. In addition, a large number of potentially sex-specific transcripts have yet to be tested, suggesting that further sex-specific genes may be identified from the transcript libraries.

Identification of sex-specific transcripts within the current AgamP3.6 gene build proved fruitless. Despite candidates being prioritised based upon the number of aligned reads and absence of reads of the opposite sex within extended upstream and downstream regions, we failed to validate any of the 15 male or 20 female candidates examined. As these transcripts represent the top candidates, this would suggest that the read depth within this study was insufficient to identify sex-specific expression of AgamP3.6 transcripts. Though it should be noted that, due to time constraints, a large number of candidates remain to be tested.

In a recent 454 sequencing study of sex-specific guppy fish transcriptomes by Fraser *et al.* (2011), 50 % of candidate sex-specific contigs, each assembled from over 100 reads, were validated. Though RNA-seq studies are not directly comparable due to differences in sample preparation and depth of coverage achieved, the largest number of reads in our study to an AgamP3.6 sex-specific candidate transcript was 11. Though conclusions cannot be drawn through comparison to the study by Fraser *et al.* (2011), it does again suggest that read depth achieved within our study was insufficient. In addition, Hale *et al.* (2010) failed to validate any of the 31 putative sex-specific contigs they identified with greater than 10 aligned reads. Within our study, failure to identify sex-specific AgamP3.6 transcripts may be attributed to the need for amplification of the embryonic cDNA prior to

sequencing. An interesting point is that contig08484, validated as male specific in a second, different approach, was assembled from 13 reads, greater than the number of the top AgamP3.6 candidate.

In the second approach, not based upon the AgamP3.6 gene build, contigs with aligned reads from a single sex were searched for. Candidate contigs were prioritised based upon the number of reads used in their assembly. A total of six male contigs were examined, revealing a single contig with male specific expression. RACE experiments revealed that contig08484 constituted part of two genes, sharing partially overlapping exons and on both the forward and reverse DNA strands. Interestingly, alignment to the genome located the two genes on the Y chromosome. Thus YOA and YOB join a retrotransposon, multiple satellite repeat families and an expressed transcript in calling the Y chromosome home (Rohr *et al.* 2002; Krzywinski *et al.* 2004; Krzywinski *et al.* 2005). Functional characterisation of YOA and YOB has proved problematic. BLASTX searches against the nr protein database failed to identify any significant sequence homology to any known proteins. It is possible that YOA and YOB are not mRNAs but represent a different part of the RNA landscape.

High-throughput RNA-seq studies have demonstrated the complexity of RNA species, demonstrating that the transcriptome is not a simple mRNA to protein relationship. One of these species, non-coding RNA (ncRNA) can either be short (<200) or long (>200) and multi-exon ncRNA have been identified in mice transcriptome studies (Guttman *et al.* 2010). Within short ncRNAs there are a wide range of RNA species, including but not restricted to miRNA, piRNA and snoRNA reviewed by Costa (2010). As YOA and YOB are greater than 200 bp, if they are non-coding, they are likely to be long non-coding RNA (lncRNA) species. Classification of transcripts as ncRNA is problematic, as ncRNAs may contain ORFs purely by chance (Dinger *et al.* 2008). Therefore, though YOA and YOB contain ORFs longer than 50 amino acids, firm conclusions regarding gene function cannot be made without further functional characterisation. Previously, characterisation of intergenic ncRNA has been assayed through changes in downstream gene expression following perturbation knockdown studies, as ncRNA has been shown to regulate expression of surrounding genes (Orom *et al.* 2010; Guttman *et al.* 2011).

Previous transcriptome studies have identified polyadenylated ncRNA, suggesting that in our study that used poly(A) RNA enrichment, ncRNA may be present in the transcript

libraries (Guttman *et al.* 2009). There is previous evidence in the literature for functional roles of male-specific ncRNAs. The well characterised *D. melanogaster RNA on the X 1* and (*roX1*) and *RNA on the X 2* (*roX2*) ncRNAs are part of a male-specific lethal complex, responsible for hyper-transcription of the X chromosome for dosage compensation through chromatin remodelling (Hamada *et al.* 2005; Deng and Meller 2006). It has been suggested that lack of sequence similarity to conserved ORFs in closely related species or conserved protein domains be used to classify ncRNAs (Kong *et al.* 2007; Lin *et al.* 2011). Using the Coding Potential Calculator Kong *et al.* (2007), all YOA and YOB transcripts were given either non-coding or weak non-coding classification within their ORFs. Previously, male-specific ncRNAs have been identified in *D. melanogaster*. Through use of the *D. melanogaster* modENCODE (model organism Encyclopaedia of DNA Elements) whole transcriptome RNA-seq data published by Graveley *et al.* (2011), 1,119 long intergenic non-coding RNAs (lincRNA) were identified, of which 139 were male-specific, with a majority being expressed in the testes (Young *et al.* 2012). Young et al. (2012) postulate that the function of these 139 male-specific lincRNAs is for male or testis-specific development.

Though ORFs within YOA and YOB share no sequence similarity to known proteins, it is possible that they encode short peptides. Short *D. melanogaster* peptide sequences have been identified that are encoded from a previously presumed, polycistronic ncRNA, that show sequence similarity to other insect ORFs (Galindo *et al.* 2007). More recently, small peptides of 11 to 32 amino acids encoding *polished rice* (*pri*) have been shown to regulate expression of the embryonic transcription factor *shavenbaby* (*svb*) for embryonic epidermal differentiation (Kondo *et al.* 2010). Directed mutagenesis or ectopic expression of YOA and YOB ORFs may be avenues of future work worth exploring to provide information regarding their function.

RNAi knockdown with dsRNA corresponding to a region of contig08484 failed to induce an expected sex-reversal phenotype. Absence of phenotype may be attributable to an incorrect expectation of a sex-reversal phenotype. We hypothesised that the early zygotic male-specific expression of contig08484 may indicate its function as a male sex promoting factor. However, previous work showing that a majority of *Drosophila* Y linked genes are necessary for male reproduction and regulation of testis-specific genes involved in sperm formation, suggests that future functional studies should assay for changes in male fertility

(Carvalho *et al.* 2000; Vibranovski *et al.* 2008; Sackton *et al.* 2011). Quantitative RT-PCR analysis should be performed to assess accurate reductions in transcript abundance following gene knockdown. This was not done within our study and insufficient reduction in transcript levels may have resulted in an absence of phenotype.

Though it cannot be concluded what function YOA and YOB have, they have different developmental expression profiles, suggesting they perform distinct functions. YOA transcripts are first expressed in the late larval stages (3^{rd} to 4^{th} instar) and then continue in the pupal and adult stages. YOA expression initiating in the late larval stages, when testis formation occurs, may have indicated that it is a testis specific gene, but RT-PCR analysis demonstrated that it is expressed in all adult tissues examined. YOB expression begins in embryos 180 min AEL, suggesting early zygotic expression and is then continually expressed throughout development. Interestingly, expression of YOB 180 min AEL precedes sex-specific expression of *doublesex* (*dsx*) (8 hr AEL), suggesting that male-specific expression of YOB may not be a direct consequence of *dsx* regulation. Presence of YOA and YOB on the Y chromosome is perhaps surprising as no complete gene sequences have been isolated from the largely heterochromatic male sex chromosome previously. Y-linked *D. melanogaster* genes are thought to have arisen from autosomal gene duplications (Carvalho *et al.* 2000; Carvalho *et al.* 2001). YOA and YOB share no similarity to any sequence within the *An. gambiae* genome and so it cannot be concluded if they arose from duplication followed by an autosome to Y chromosome translocation.

For future characterisation of the YOA and YOB genes, the promoter region of each gene could be analysed through creation of an expression plasmid driving expression of the reporter GFP. Following embryo microinjection of the expression plasmid, transient GFP positive individuals would be expected to all be male. Isolation of GFP positive females would suggest that they are able to activate transcription from the YOB promoter. Absence of GFP positive females would confirm that YOB is male specific and not just expressed exclusively in males due to the location of the gene on the Y chromosome. In addition, as an alternative to RNAi based approaches, further functional characterisation could be performed through transient overexpression of YOA and YOB transcripts from a heat shock or ubiquitously expressed promoter. We have shown that the promoter (*An. gambiae* polyubiquitin AGAP001971) of the expression plasmid FU III can drive GFP expression in

various tissues and a separate expression plasmid with the AGAP001971 prmoter fused to YOA or YOB full length transcripts could be created. Embryo microinjection of this plasmid may be expected to induce male-specific characteristics in females and, if so, would indicate a potential function of the two genes.

Chapter 7: Final Discussion

The major findings of this thesis have been three fold: firstly, the demonstration of sex-specific expression in *An. gambiae* embryos, that starts as early as 4 hr after egg laying (AEL). Secondly, two novel complete Y chromosome genes have been identified and characterised. Thirdly, transcriptome analysis has produced large datasets that will improve genomic annotation of *An. gambiae*. The overall goal of this work was the comparative analysis of male and female embryonic transcriptomes and this has been achieved. It has been possible to isolate sufficient embryonic RNA to perform next generation sequencing analysis of male and female transcript content. This study has demonstrated that male specific gene expression occurs in the embryonic stages and it is hoped that further investigation of these datasets may lead to identification of genes having female specific expression also.

The first part of this study took a targeted gene approach using *doublesex* (*dsx*) and *fruitless* (*fru*), known components of the sex determination pathway, to study sex-specific gene expression. Previously, post-embryonic stages of *An. gambiae* have been shown to express sex-specific *dsx* and *fru* forms (Scali *et al.* 2005; Gailey *et al.* 2006). This study provided detailed information on the temporal patterns of transcription of these two genes in *An. gambiae* embryos.

The An. gambiae FRU protein is able to regulate formation of the male-specific Muscle of Lawrence (MOL) in *fru*-deficient male *D. melanogaster* (Gailey *et al.* 2006). Sex-specific splicing of *An. gambiae fru* has been documented previously in the adult head, but male transcripts are found in female whole bodies (Gailey *et al.* 2006). Similar to the findings of Gailey *et al.* (2006), in our study, male transcripts were detected in female embryos, larvae and adults. This is perhaps not surprising, as post-transcriptional regulation of *fru* transcripts is required in *D. melanogaster*. Binding of TRA and TRA2 prevents translation of female-specific *fru* mRNA and thus transcripts are present, but proteins are not detected (Usui-Aoki *et al.* 2000). Use of immunohistochemistry assays using FRU antibodies to detect alternative splice forms would be of interest, but beyond the scope of this project. Detection of male transcripts, but not male protein, in females would confirm that inhibition of mRNA translation is required to ensure sex-specific functions of the FRU protein.

With regards to *dsx*, we have shown that following deposition of female and male transcripts, signs of sex-specific regulation of the transcript levels are detected 4 h AEL and clearly discernible sex-specific forms are present 8 hr AEL. Sex-specific forms of *dsx* have been previously identified within the embryonic stages of the dipterans *C. capitata* and *Ae. aegypti*, indicating that *An. gambiae* sexual identity is similarly determined at this early stage (Gabrieli *et al.* 2010; Salvemini *et al.* 2011). In *C. capitata* embryos, maternal *dsx* transcripts are degraded and sexual identity effectively reset prior to sex-specific zygotic expression (Gabrieli *et al.* 2010). We have shown in chapter 2 that *An. gambiae* embryonic *dsx* expression is similar to *C. capitata*. A combination of degradation of maternal female transcripts and zygotic expression of male transcripts establishes a sex-specific pattern in males. Stabilising female maternal transcripts, activation of zygotic female form transcription or a combination of the two is responsible for the detection of sex-specific female forms. Zygotic expression of the male form may suggest that zygotic activation of the female form is also involved in establishing female-specific forms.

Zygotic expression of sex-specific *dsx* transcripts suggests that upstream regulators of *dsx* splicing are also expressed in the embryonic stages of *An. gambiae*. *tra* and *tra2* are responsible for *dsx* sex-specific splicing in a number of insect species, reviewed by Sanchez *et al.* (2008). Using the RNAi approach taken in chapter 2, we could not conclude whether a TRA2 candidate was indeed a TRA2 sex determination homologue. What we have shown, (through knockdown of GFP expression in transgenic mosquitoes), is that an active RNAi pathway is present in the embryonic stages and that knockdown in various tissues is possible. Recently, two studies have shown that embryo dsRNA microinjection can induce a visible phenotype and in future RNAi studies, the same genes targeted within these two studies could be used as positive controls (Magnusson *et al.* 2011; Thailayil *et al.* 2011).

Under the assumption that an *An. gambiae* TRA2 homologue were to be identified, the protein-protein interactions of TRA and TRA2 through the serine/arginine (SR) rich domains could be exploited to identify candidate *An. gambiae tra* homologues. Protein pull-down immunoprecipitation assays, using TRA2 as bait, could be performed. Biotin labelled TRA2 protein would be used as bait and incubated with cell lysate, from which TRA would be expected to be present. Biotin labelled TRA2 and bound candidate proteins would then be isolated through binding of biotin to a streptavidin agarose column and then separate bound proteins by gel electrophoresis. Also protein-complex immunoprecipitation (Co-IP)

could be used, where an antibody is raised against a known antigen (TRA2), allowing binding of the protein and any other proteins it is in complex with, from a cell lysate. A further avenue to explore would be oligonucleotide pull-down assays that allow detection of DNA-protein interactions. Biotin-labelled DNA oligonucleotide bait (putative TRA/TRA2 binding sites in *dsx* transcripts described by Scali *et al.* (2005)) could be incubated with a cell lysate and immobilised on streptavidin beads, then proteins identified through western-blot or mass spectrometry. Finally, yeast-two hybrid experiments (Fields and Song 1989) could be performed to isolate candidate TRA proteins from an *An. gambiae* cDNA library. Fusion of the *tra2* coding sequence to the DNA binding domain of GAL4 would allow protein-protein interactions between TRA2 and TRA candidates from the cDNA library (fused to the GAL4 trasncriptional activation domain) through activation of a reporter gene.

TRA protein sequences are rapidly evolving and this has increased the difficulty in identifying an *An. gambiae* homologue. Bioinformatics approaches and use of degenerate primers identified many regions with limited sequence similarity to conserved domains of known TRA sequences. Therefore, with more time, a large scale systematic approach to prioritise candidates would be required, and may identify valuable information contained within the results of using these two approaches.

In the second part of this study, comparison of two RNA-seq datasets identified over 1,500 and 700 putative sex-specific transcripts, including AgamP3.6 transcripts and new transcribed regions (NTRs) for males and females respectively, of which 60 were examined by RT-PCR. A significant number of putative sex-specific candidates have yet to be examined, and future large-scale RT-PCR testing of candidates may be expected to identify genes having sex-specific transcript expression. A single NTR, contig08484, showing male specific expression throughout development was identified. Cloning of the complete cDNA following RACE revealed that contig08484 is part of two novel Y linked genes (YOA and YOB) encoded on opposite strands and having partially overlapping exons. The identification of YOA and YOB has demonstrated that massively parallel sequencing of transcriptomes using the 454 platform can identify sex-specific expression in the embryo.

Y chromosomes are often largely degenerate and contain large regions of heterochromatin and are labelled a graveyard for retrotransposons (Junakovic *et al.* 1998). Males are heterogametic and lack of Y chromosome genetic recombination causes them to be vulnerable to accumulation of deleterious mutations and degeneration (Steinemann and

Steinemann 1992; Charlesworth and Charlesworth 2000). With regard to the *An. gambiae* Y chromosome, 59 Y linked scaffolds, equivalent to 237 Kb, have been identified within the VectorBase AgamP3 genome resource. Prior to our study, a retrotransposon, and a short non-transposable element-like expressed fragment with sequence similarity to putative stop codon containing pseudogenes on other chromosomes, have been isolated on the Y chromosome (Rohr *et al.* 2002; Krzywinski *et al.* 2004). Here, we provide information on two novel full-length Y chromosome genes that show no significant sequence similarity to any known protein and whose developmental expression patterns differ markedly.

As discussed in chapter 5, it is unclear whether YOA and YOB transcripts represent protein coding mRNA, ncRNA, pseudogenes or encode very short peptides (<30 amino acids). Despite attempts to functionally characterise the two genes, by reduction in transcript levels through RNAi, we cannot conclude what function YOA and YOB may perform. It is possible that we may not observe a phenotypic effect, regardless of gene knockdown. Both ncRNAs and protein coding mRNAs are able to perform a variety of functions, including gene regulation (Guttman *et al.* 2011). Therefore future RNAi knockdown studies could be performed and changes in global gene expression assessed, perhaps through microarray experiments, to assess if they have a regulatory role. Recently, the Y chromosome of *D. simulans* has been shown to be involved in genome-wide gene regulation (Sackton *et al.* 2011). In another approach, as previous studies have demonstrated that *D. melanogaster* Y linked genes are necessary for male reproduction and regulation of testis-specific genes involved in sperm formation (Carvalho *et al.* 2000; Vibranovski *et al.* 2008; Sackton *et al.* 2011), future RNAi studies could also investigate a potential effect of YOA and YOB transcript knockdown on male fertility.

Comparision of male and female transcriptomes also revealed a suite of sex-biased candidate genes. We conducted read count analysis and 799 differentially expressed genes (Log_2FC >2) between male and female datasets were identified within this study, using edgeR. With more time, quantitative studies using qPCR would have been performed to confirm sex-biased expression. The RNA-seq libraries generated within the project were from a single 454 sequencing library and cDNA samples for sequencing were prepared differently (different number of amplification cycles and female amplification was repeated twice). Therefore it is essential that candidates are validated through qPCR, prior to any functional analysis, as differences in sample preparation may have affected transcript

abundances. In addition, candidate sex-biased genes have no statistical support (as no sequencing replicates were performed), therefore false positives may be expected within the candidate genes. With more time, but importantly also at a greater cost, replicate RNA-seq experiments would greatly increase the confidence with which differential expression could be identified.

Analysis of the embryo transcripts generated from this study has demonstrated that the current AgamP3.6 gene build is far from complete. Over 7,000 NTRs were detected in our study and, of these, 5 % had sequence similarity to known proteins within the NCBI nr database. As we do not have knowledge of the full length NTR sequences, it is likely further NTRs would have sequence similarity to known proteins if their full transcript sequence was known. It is also possible that the remaining NTR sequences may be extensions of current annotated exons, novel exons of annotated genes, novel protein-codin genes, non-coding RNAs, pseudogenes or other RNA species. All new transcribed regions have been annotated within a BED file, supplied with this thesis. NTRs can be viewed using genomic viewers and upon publication of transcriptome data, all information will be made publicly available from NCBI short read archive and VectorBase to ensure they can be utilised by the wider research community.

Use of selfish-genetic elements, which have an increased rate of transmission to progeny compared to the rest of the genome, have been suggested as a method to drive a pathogen refractory phenotype into a vector population through population replacement, reviewed by Terenius *et al.* (2008). One such element, maternal-effect dominant embryonic arrest (*Medea*), was first identified in wild *Tribolium castaneum* populations that caused maternally-promoted embryonic lethality of progeny, with zygotically activated rescue (Beeman *et al.* 1992). *Medea* elements are capable of rapidly spreading through a population, because among the progeny of *Medea*-bearing females only the individuals that inherit a copy of the element are viable (Beeman *et al.* 1992). Though not known conclusively, it is believed that *Medea* elements consist of two tightly-linked genes, one encoding a maternal toxin and the second a zygotic antidote (Beeman and Friesen 1999).

The molecular identity of this naturally occurring element still remains to be revelase but a synthetic *Medea*-like element has been created in *D. melanogaster*. Use of a maternal-germline specific promoter (*bicoid*) to drive expression of interfering microRNAs (miRNA) that target maternal transcripts of a gene (*myd88)*, essential for correct embryo dorsal-

ventral pattern formation, causes lethality (Chen *et al.* 2007). Zygotic rescue occurs through expression of a miRNA-insensitive *myd88* transcript from an early-zygotic promoter (*bottleneck*). Therefore, only individuals inheriting the zygotic miRNA insensitive myd88 transcript would be viable and Chen *et al.* (2007) demonstrated that all surviving offspring contained the *Medea* construct. Such an approach within a disease vector would aim to use the *Medea* construct to drive a parasite refractory phenotype into a wild population.

Medea elements could also be used for female-specific lethality through use of a YOB promoter in *An. gambiae*, with the aim of suppressing the vector population. A YOB promoter sequence could be utilised to drive early zygotic expression of an antidote, resulting in male-specific zygotic rescue. Recently, it has been demonstrated that the promoter region of an early zygotic kinesin light chain gene (*KLC2.1*) is able to drive expression of a luciferase reporter in the early embryonic stages of *Aedes aegypti* and *An. gambiae* (Biedler *et al.* 2012). Future work to characterise the YOB promoter, coupled with a maternally derived toxin, would ensure that only male offspring would survive. Theoretically, released transgenic homozygous males, carrying a repressible autosomal *Medea* element, would mate with wild type females, resulting in heterozygous progeny carrying the *Medea* element. All heterozygous progeny would be viable, as the toxic gene would not be maternally deposited in the embryo. However, mating of the heterozygous *Medea* female progeny to wild type males would result in delivery of the maternal toxin to all embryos and male-specific zygotic rescue. Such matings would be expected to reduce the population size and therefore vectoral capacity.

This approach would require a promoter to drive germline-specific maternal expression and the *An. gambiae vasa* promoter meets this criterion (Papathanos *et al.* 2009). In addition, knowledge of a maternal gene whose function is required for embryogenesis, but importantly, not oogenesis would also be required. Possible gene candidates include *tramtrack*, a maternally deposited segmentation gene in *D. melanogaster* that is also observed in *An. gambiae* embryos (Goltsev *et al.* 2007) and *oskar*, a maternally deposited gene shown to locate to the posterior and anterior poles of *An. gambiae* embryos (Goltsev *et al.* 2004; Juhn and James 2006). Though beyond the scope of this project, further gene candidates could be identified through transcriptional profiling of eggs shortly after deposition. Alternatively, putative homologues of *D. melanogaster* embryo development

genes could be searched for within the *An. gambiae* genome and potential conservation of function assessed through embryonic RNAi.

References

Ahmad, S. M. and Baker, B. S. (2002). "Sex-specific deployment of FGF signaling in *Drosophila* recruits mesodermal cells into the male genital imaginal disc." Cell **109**(5): 651-661.

Alagna, F., D'Agostino, N., Torchia, L., Servili, M., Rao, R.*, et al.* (2009). "Comparative 454 pyrosequencing of transcripts from two olive genotypes during fruit development." Bmc Genomics **10**: 15.

Alphey, L. and Andreasen, M. (2002). "Dominant lethality and insect population control." Molecular and Biochemical Parasitology **121**(2): 173-178.

Amrein, H., Gorman, M. and Nothiger, R. (1988). "The sex-determining gene *tra-2* of *Drosophila* encodes a putative RNA-binding protein." Cell **55**(6): 1025-1035.

Amrein, H., Maniatis, T. and Nothiger, R. (1990). "Alternatively spliced transcripts of the sex-determining gene *tra-2* of Drosophila encode functional proteins of different size." Embo Journal **9**(11): 3619-3629.

An, W. Q. and Wensink, P. C. (1995). "Integrating sex-specific and tissue-specific regulation within a single *Drosophila* enhancer." Genes & Development **9**(2): 256-266.

Anand, A., Villella, A., Ryner, L. C., Carlo, T., Goodwin, S. F.*, et al.* (2001). "Molecular genetic dissection of the sex-specific and vital functions of the *Drosophila melanogaster* sex determination gene fruitless." Genetics **158**(4): 1569-1595.

Andreasen, M. H. and Curtis, C. F. (2005). "Optimal life stage for radiation sterilization of *Anopheles* males and their fitness for release." Medical and Veterinary Entomology **19**(3): 238-244.

Anosike, J. C., Nwoke, B. E., Ajayi, E. G., Onwulir, C. O., Okoro, O. U.*, et al.* (2005). "Lymphatic filariasis among the Ezza people of Ebonyi State, eastern Nigeria." Annals of Agricultural and Environmental Medicine **12**(2): 181-186.

Armour, C. D., Castle, J. C., Chen, R. H., Babak, T., Loerch, P.*, et al.* (2009). "Digital transcriptome profiling using selective hexamer priming for cDNA synthesis." Nature Methods **6**(9): 647-U35.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H.*, et al.* (2000). "Gene Ontology: tool for the unification of biology." Nature Genetics **25**(1): 25-29.

Auer, P., Srivastava, S. and Doerge, R. (2011). "Differential expression - the next generation and beyond." Briefings in Functional Genomics **doi:10.1093**.

Baker, B. S. and Ridge, K. A. (1980). "Sex and the single cell.1.On the action of major loci affecting sex determination in *Drosophila melanogaster*." Genetics **94**(2): 383-423.

Baker, B. S. and Wolfner, M. F. (1988). "A molecular analysis of *doublesex*, a bifunctional gene that controls both male and female sexual-differentiation in *Drosophila melanogaster*." Genes & Development **2**(4): 477-489.

Baker, D. A., Nolan, T., Fischer, B., Pinder, A., Crisanti, A. and Russell, S. (2011). "A comprehensive gene expression atlas of sex- and tissue-specificity in the malaria vector, *Anopheles gambiae*." Bmc Genomics **12**: 296-308.

Baker, R. H. and Sakai, R. K. (1979). "Triploids and male determination in the mosquito, *Anopheles culicifacies*." Journal of Heredity **70**(5): 345-346.

Bandziulis, R. J., Swanson, M. S. and Dreyfuss, G. (1989). "RNA binding proteins as developmental regulators." Genes & Development **3**(4): 431-437.

Bashirullah, A., Halsell, S. R., Cooperstock, R. L., Kloc, M., Karaiskakis, A.*, et al.* (1999). "Joint action of two RNA degradation pathways controls the timing of maternal transcript

elimination at the midblastula transition in Drosophila melanogaster." Embo Journal **18**(9): 2610-2620.

Beeman, R. W. and Friesen, K. S. (1999). "Properties and natural occurrence of maternal-effect selfish genes ('Medea' factors) in the Red Flour Beetle, *Tribolium castaneum*." Heredity **82**: 529-534.

Beeman, R. W., Friesen, K. S. and Denell, R. E. (1992). "Maternal-effect selfish genes in flour beetles." Science **256**(5053): 89-92.

Bell, L. R., Horabin, J. I., Schedl, P. and Cline, T. W. (1991). "Positive autoregulation of *sex-lethal* by alternative splicing maintains the female determined state in *Drosophila*." Cell **65**(2): 229-239.

Bell, L. R., Maine, E. M., Schedl, P. and Cline, T. W. (1988). "*sex-lethal*, a *Drosophila* sex determination switch gene, exhibits sex-specific RNA splicing and sequence similarity to RNA binding proteins." Cell **55**(6): 1037-1046.

Bellin, D., Ferrarini, A., Chimento, A., Kaiser, O., Levenkova, N., Bouffard, P. and Delledonne, M. (2009). "Combining next-generation pyrosequencing with microarray for large scale expression analysis in non-model species." Bmc Genomics **10**: 9.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J.*, et al.* (2008). "Accurate whole human genome sequencing using reversible terminator chemistry." Nature **456**(7218): 53-59.

Bertossa, R. C., van de Zande, L. and Beukeboom, L. W. (2009). "The *Fruitless* Gene in *Nasonia* Displays Complex Sex-Specific Splicing and Contains New Zinc Finger Domains." Molecular Biology and Evolution **26**(7): 1557-1569.

Biedler, J. K., Hu, W., Tae, H. and Tu, Z. (2012). "Identification of Early Zygotic Genes in the Yellow Fever Mosquito *Aedes aegypti* and Discovery of a Motif Involved in Early Zygotic Genome Activation." Plos One **7**(3): e33933.

Billeter, J. C. and Goodwin, S. F. (2004). "Characterization of *Drosophila* fruitless-gal4 transgenes reveals expression in male-specific fruitless neurons and innervation of male reproductive structures." Journal of Comparative Neurology **475**(2): 270-287.

Boakye, D. A., Wilson, M. D., Appawu, M. A. and Gyapong, J. (2004). "Vector competence, for *Wuchereria bancrofti*, of the *Anopheles* populations in the Bongo district of Ghana." Annals of Tropical Medicine and Parasitology **98**(5): 501-508.

Boerjan, B., Tobback, J., De Loof, A., Schoofs, L. and Huybrechts, R. (2011). "*Fruitless* RNAi knockdown in males interferes with copulation success in *Schistocerca gregaria*." Insect Biochemistry and Molecular Biology **41**(5): 340-347.

Bonizzoni, M., Dunn, W. A., Campbell, C. L., Olson, K. E., Dimon, M. T., Marinotti, O. and James, A. A. (2011). "RNA-seq analyses of blood-induced changes in gene expression in the mosquito vector species, *Aedes aegypti*." Bmc Genomics **12**: 82-95.

Brennecke, J., Aravin, A. A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R. and Hannon, G. J. (2007). "Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*." Cell **128**(6): 1089-1103.

Brennecke, J., Malone, C. D., Aravin, A. A., Sachidanandam, R., Stark, A. and Hannon, G. J. (2008). "An Epigenetic Role for Maternally Inherited piRNAs in Transposon Silencing." Science **322**(5906): 1387-1392.

Burghardt, G., Hediger, M., Siegenthaler, C., Moser, M., Dubendorfer, A. and Bopp, D. (2005). "The transformer2 gene in *Musca domestica* is required for selecting and

maintaining the female pathway of development." Development Genes and Evolution **215**(4): 165-176.

Burtis, K. C. and Baker, B. S. (1989). "*Drosophila* doublesex gene controls somatic sexual-differentiation by producing alternatively spliced messenger RNAs encoding sex-specific polypeptides " Cell **56**(6): 997-1010.

Burtis, K. C., Coschigano, K. T., Baker, B. S. and Wensink, P. C. (1991). "The Doublesex proteins of *Drosophila melanogaste*r bind directly to a sex-specific yolk protein gene enhacner " Embo Journal **10**(9): 2577-2582.

Bushland, R. C., Lindquist, A. W. and Knipling, E. F. (1955). "Eradication of screw-worms through release of sterilized males." Science **122**(3163): 287-288.

Carvalho, A. B., Dobo, B. A., Vibranovski, M. D. and Clark, A. G. (2001). "Identification of five new genes on the Y chromosome of *Drosophila melanogaster*." Proceedings of the National Academy of Sciences of the United States of America **98**(23): 13225-13230.

Carvalho, A. B., Lazzaro, B. P. and Clark, A. G. (2000). "Y chromosomal fertility factors kl-2 and kl-3 of *Drosophila melanogaster* encode dynein heavy chain polypeptides." Proceedings of the National Academy of Sciences of the United States of America **97**(24): 13239-13244.

Casper, A. and Van Doren, M. (2006). "The control of sexual identity in the *Drosophila* germline." Development **133**(15): 2783-2791.

Casper, A. L. and Van Doren, M. (2009). "The establishment of sexual identity in the *Drosophila* germline." Development **136**(22): 3821-3830.

Cator, L. J., Arthur, B. J., Harrington, L. C. and Hoy, R. R. (2009). "Harmonic Convergence in the Love Songs of the Dengue Vector Mosquito." Science **323**(5917): 1077-1079.

Catteruccia, F., Benton, J. P. and Crisanti, A. (2005). "An *Anopheles* transgenic sexing strain for vector control." Nature Biotechnology **23**(11): 1414-1417.

Chanda, E., Masaninga, F., Coleman, M., Sikaala, C., Katebe, C., MacDonald, M., Baboo, K. S., Govere, J. and Manga, L. (2008). "Integrated vector management: The Zambian experience." Malaria Journal **7**: 8.

Charlesworth, B. and Charlesworth, D. (2000). "The degeneration of Y chromosomes." Philosophical Transactions of the Royal Society of London Series B-Biological Sciences **355**(1403): 1563-1572.

Chase, B. A. and Baker, B. S. (1995). "A genetic-analysis of *intersex*, a gene regulating sexual-differentiation in *Drosophila melanogaster* females." Genetics **139**(4): 1649-1661.

Chatterjee, S. S., Uppendahl, L. D., Chowdhury, M. A., Ip, P. L. and Siegal, M. L. (2011). "The female-specific Doublesex isoform regulates pleiotropic transcription factors to pattern genital development in *Drosophila*." Development **138**(6): 1099-1109.

Chen, C. H., Huang, H. X., Ward, C. M., Su, J. T., Schaeffer, L. V., Guo, M. and Hay, B. A. (2007). "A synthetic maternal-effect selfish genetic element drives population replacement in *Drosophila*." Science **316**(5824): 597-600.

Chen, E. H. and Baker, B. S. (1997). "Compartmental organization of the *Drosophila* genital imaginal discs." Development **124**(1): 205-218.

Chen, X. G., Marinotti, O., Whitman, L., Jasinskiene, N. and James, A. A. (2007). "The *Anopheles gambiae* vitellogenin gene (VGT2) promoter directs persistent accumulation of a reporter gene product in transgenic *Anopheles stephensi* following multiple bloodmeals." American Journal of Tropical Medicine and Hygiene **76**(6): 1118-1124.

Cheung, F., Haas, B. J., Goldberg, S. M. D., May, G. D., Xiao, Y. L. and Town, C. D. (2006). "Sequencing *Medicago truncatula* expressed sequenced tags using 454 Life Sciences technology." Bmc Genomics **7**: 10.

Chevreux, B. (2011). "MIRA-Assembler."

Chevreux, B., Pfisterer, T., Drescher, B., Driesel, A. J., Muller, W. E. G., Wetter, T. and Suhai, S. (2004). "Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs." Genome Research **14**(6): 1147-1159.

Cho, S., Huang, Z. Y. and Zhang, J. Z. (2007). "Sex-specific splicing of the honeybee *doublesex* gene reveals 300 million years of evolution at the bottom of the insect sex-determination pathway." Genetics **177**(3): 1733-1741.

Christophides, G. K., Zdobnov, E., Barillas-Mury, C., Birney, E., Blandin, S*., et al.* (2002). "Immunity-related genes and gene families in *Anopheles gambiae*." Science **298**(5591): 159-165.

CLCBIO "CLC Assembly Cell User Manual." http://www.clcbio.com/index.php?id=1393.

Clements, A., N. (1992). "The Biology of Mosquitoes." Chapman and Hall, London.

Clemons, A., Haugen, M., Le, C., Mori, A., Tomchaney, M., Severson, D. W. and Duman-Scheel, M. (2011). "siRNA-Mediated Gene Targeting in *Aedes aegypt*i Embryos Reveals That Frazzled Regulates Vector Mosquito CNS Development." Plos One **6**(1): e16730

Clemons, A., Haugen, M., Severson, D. and Duman-Scheel, M. (2010). "Functional analysis of genes in *Aedes aegypti* embryos." Cold Spring Harbor protocols **2010**(10): pdb.prot5511.

Cline, T. W. (1988). "Evidence that sisterless-a and sisterless-b are two of several discrete ''numerator elements'' of the X/A sex determination signal in *Drosophila* that switch *Sxl* between two alternative stable expression states." Genetics **119**(829-862).

Cline, T. W., Dorsett, M., Sun, S., Harrison, M. M., Dines, J., Sefton, L. and Megna, L. (2010). "Evolution of the *Drosophila* Feminizing Switch Gene Sex-lethal." Genetics **186**(4): 1321-U402.

Clynen, E., Ciudad, L., Belles, X. and Piulachs, M. D. (2011). "Conservation of *fruitless'* role as master regulator of male courtship behaviour from cockroaches to flies." Development Genes and Evolution **221**(1): 43-48.

Cohen, S. M. (1993). Imaginal disc development, Cold Spring Harbor Laboratory Press {a}, 10 Skyline Drive, Plainview, New York 11803, USA.

Coleman, P. G. and Alphey, L. (2004). "Editorial: Genetic control of vector populations: an imminent prospect." Tropical Medicine & International Health **9**(4): 433-437.

Collins, F. H., Sakai, R. K., Vernick, K. D., Paskewitz, S., Seeley, D. C., Miller, L. H., Collins, W. E., Campbell, C. C. and Gwadz, R. W. (1986). "Genetic selection of a plasmodium refractory strain of the malaria vector *Anopheles gambiae*." Science **234**(4776): 607-610.

Concha, C., Li, F. and Scott, M. J. (2010). "Conservation and sex-specific splicing of the *doublesex* gene in the economically important pest species *Lucilia cuprina*." Journal of Genetics **89**(3): 279-285.

Concha, C. and Scott, M. J. (2009). "Sexual development in *Lucilia cuprina* (Diptera, Calliphoridae) is controlled by the *transformer* gene." Genetics **182**(3): 785-98.

Condon, K. C., Condon, G. C., Dafa'alla, T. H., Fu, G. L., Phillips, C. E., Jin, L., Gong, P. and Alphey, L. (2007). "Genetic sexing through the use of Y-linked transgenes." Insect Biochemistry and Molecular Biology **37**(11): 1168-1176.

Conesa, A., Gotz, S., Garcia-Gomez, J. M., Terol, J., Talon, M. and Robles, M. (2005). "Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research." Bioinformatics **21**(18): 3674-3676.

Corbett, J. R. (1974). The biochemical mode of action of pesticides.

Coschigano, K. T. and Wensink, P. C. (1993). "Sex-specific transcriptional regulation by the male and female doublesex proteins of *Drosophila*." Genes & Development **7**(1): 42-54.

Costa, F. F. (2010). "Non-coding RNAs: Meet thy masters." Bioessays **32**(7): 599-608.

Costa, V., Angelini, C., De Feis, I. and Ciccodicola, A. (2011). "Uncovering the Complexity of Transcriptomes with RNA-Seq." Journal of Biomedicine and Biotechnology: 19.

Curtis, C. F., Akiyama, J. and Davidson, G. (1976). "Genetic sexing system in *Anopheles gambiae* species A." Mosquito News **36**(4): 492-498.

Daily, J. P., Le Roch, K. G., Sarr, O., Ndiaye, D., Lukens, A*., et al.* (2005). "*In vivo* transcriptome of *Plasmodium falciparum* reveals overexpression of transcripts that encode surface proteins." Journal of Infectious Diseases **191**(7): 1196-1203.

David, J. P., Strode, C., Vontas, J., Nikou, D., Vaughan, A., Pignatelli, P. M., Louis, C., Hemingway, J. and Ranson, H. (2005). "The *Anopheles gambiae* detoxification chip: A highly specific microarray to study metabolic-based insecticide resistance in malaria vectors." Proceedings of the National Academy of Sciences of the United States of America **102**(11): 4080-4084.

Davies, T. G. E., Field, L. M., Usherwood, P. N. R. and Williamson, M. S. (2007). "DDT, pyrethrins, pyrethroids and insect sodium channels." Iubmb Life **59**(3): 151-162.

De Renzis, S., Elemento, O., Tavazoie, S. and Wieschaus, E. F. (2007). "Unmasking activation of the zygotic genome using chromosomal deletions in the *Drosophila* embryo." Plos Biology **5**(5): 1036-1051.

de Valdez, M. R. W., Nimmo, D., Betz, J., Gong, H. F., James, A. A., Alphey, L. and Black, W. C. (2011). "Genetic elimination of dengue vector mosquitoes." Proceedings of the National Academy of Sciences of the United States of America **108**(12): 4772-4775.

DeFalco, T., Le Bras, S. and Van Doren, M. (2004). "Abdominal-B is essential for proper sexually dimorphic development of the *Drosophila* gonad." Mechanisms of Development **121**(11): 1323-1333.

DeFalco, T. J., Verney, G., Jenkins, A. B., McCaffery, J. M., Russell, S. and Van Doren, M. (2003). "Sex-specific apoptosis regulates sexual dimorphism in the *Drosophila* embryonic gonad." Developmental Cell **5**(2): 205-216.

della Torre, A., Fanello, C., Akogbeto, M., Dossou-yovo, J., Favia, G., Petrarca, V. and Coluzzi, M. (2001). "Molecular evidence of incipient speciation within *Anopheles gambiae* s.s. in West Africa." Insect Molecular Biology **10**(1): 9-18.

Deng, X. X. and Meller, V. H. (2006). "roX RNAs are required for increased expression of X-linked genes in *Drosophila melanogaster* males." Genetics **174**(4): 1859-1866.

Denholm, I., Franco, M. G., Rubini, P. G. and Vecchi, M. (1983). "Identification of a male determinant on the X-chromosome of housefly (*Musca domestica* l) populations in southeast England." Genetical Research **42**(3): 311-&.

Devine, G. J. and Killeen, G. F. (2010). "The potential of a new larviciding method for the control of malaria vectors." Malaria Journal **9**: 4.

Devine, G. J., Perea, E. Z., Killeen, G. F., Stancil, J. D., Clark, S. J. and Morrison, A. C. (2009). "Using adult mosquitoes to transfer insecticides to *Aedes aegypti* larval habitats." Proceedings of the National Academy of Sciences of the United States of America **106**(28): 11530-11534.

Dharia, N. V., Bright, A. T., Westenberger, S. J., Barnes, S. W., Batalov, S*., et al.* (2010). "Whole-genome sequencing and microarray analysis of ex vivo *Plasmodium vivax* reveal selective pressure on putative drug resistance genes." Proceedings of the National Academy of Sciences of the United States of America **107**(46): 20045-20050.

Dharia, N. V., Sidhu, A. B. S., Cassera, M. B., Westenberger, S. J., Bopp, S. E. R*., et al.* (2009). "Use of high-density tiling microarrays to identify mutations globally and elucidate mechanisms of drug resistance in *Plasmodium falciparum*." Genome Biology **10**(2): 13.

Dimon, M. T., Sorber, K. and DeRisi, J. L. (2010). "HMMSplicer: A Tool for Efficient and Sensitive Discovery of Known and Novel Splice Junctions in RNA-Seq Data." Plos One **5**(11): e13875.

Dimopoulos, G., Christophides, G. K., Meister, S., Schultz, J., White, K. P., Barillas-Mury, C. and Kafatos, F. C. (2002). "Genome expression analysis of *Anopheles gambiae*: Responses to injury, bacterial challenge, and malaria infection." Proceedings of the National Academy of Sciences of the United States of America **99**(13): 8814-8819.

Ding, D., Parkhurst, S. M., Halsell, S. R. and Lipshitz, H. D. (1993). "Dynamic HSP83 RNA localization during *Drosophila* oogenesis and embryogenesis." Molecular and Cellular Biology **13**(6): 3773-3781.

Ding, Y. C., Ortelli, F., Rossiter, L. C., Hemingway, J. and Ranson, H. (2003). "The *Anopheles gambiae* glutathione transferase supergene family: annotation, phylogeny and expression profiles." Bmc Genomics **4**: 16.

Dinger, M. E., Pang, K. C., Mercer, T. R. and Mattick, J. S. (2008). "Differentiating Protein-Coding and Noncoding RNA: Challenges and Ambiguities." Plos Computational Biology **4**(11): 5.

Dinglasan, R. R. and Jacobs-Lorena, M. (2008). "Flipping the paradigm on malaria transmission-blocking vaccines." Trends in Parasitology **24**(8): 364-370.

Djouaka, R. F., Bakare, A. A., Coulibaly, O. N., Akogbeto, M. C., Ranson, H., Hemingway, J. and Strode, C. (2008). "Expression of the cytochrome P450s, CYP6P3 and CYP6M2 are significantly elevated in multiple pyrethroid resistant populations of *Anopheles gambiae* s.s. from Southern Benin and Nigeria." Bmc Genomics **9**: 10.

Dong, L. S., Jensen, R. V., De Rienzo, A., Gordon, G. J., Xu, Y. L., Sugarbaker, D. J. and Bueno, R. (2009). "Differentially expressed alternatively spliced genes in Malignant Pleural Mesothelioma identified using massively parallel transcriptome sequencing." Bmc Medical Genetics **10**: 149-158.

Dong, Y., Das, S., Cirimotich, C., Souza-Neto, J. A., McLean, K. J. and Dimopoulos, G. (2011). "Engineered *Anopheles* immunity to Plasmodium infection." PLoS pathogens **7**(12): e1002458.

Dowell, R., Siddiqui, I., Meyer, F. and Spaugy, E. (2000). "Mediterranean Fruit Fly Preventitive Reelase Program in Southern Calafornia." Area-wide control of fruit flies and other insect pests.: 369-375.

Du, W., Awolola, T. S., Howell, P., Koekemoer, L. L., Brooke, B. D., Benedict, M. Q., Coetzee, M. and Zheng, L. (2005). "Independent mutations in the Rdl locus confer dieldrin

resistance to *Anopheles gambiae* and *An. arabiensis*." Insect Molecular Biology **14**(2): 179-183.

Dubendorfer, A., Hediger, M., Burghardt, G. and Bopp, D. (2002). "*Musca domestica*, a window on the evolution of sex-determining mechanisms in insects." International Journal of Developmental Biology **46**(1): 75-79.

Duffy, J. B. and Gergen, J. P. (1991). "The *Drosophila* segmentation gene runt actsas a position-specific numerator element necessary for the uniform expression of the sex-determining gene sex-lethal." Genes Dev. **5**: 2176-2187.

Edgar, B. A. and Schubiger, G. (1986). "Parameters controlling transcriptional activation during early *Drosophila* development." Cell **44**(6): 871-877.

Edmonson, M. N., Zhang, J. H., Yan, C. H., Finney, R. P., Meerzaman, D. M. and Buetow, K. H. (2011). "Bambino: a variant detector and alignment viewer for next-generation sequencing data in the SAM/BAM format." Bioinformatics **27**(6): 865-866.

Edwards, M. J., Lemos, F. J. A., Donnelly-Doman, M. and Jacobs-Lorena, M. (1997). "Rapid induction by a blood meal of a carboxypeptidase gene in the gut of the mosquito *Anopheles gambiae*." Insect Biochemistry and Molecular Biology **27**(12): 1063-1072.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J*., et al.* (2009). "Real-Time DNA Sequencing from Single Polymerase Molecules." Science **323**(5910): 133-138.

Epper, F. and Nothiger, R. (1982). "Genetic and developmental evidence for a repressed genital primordium in *Drosophila melanogaster*." Developmental Biology **94**(1): 163-175.

Erickson, J. W. and Cline, T. W. (1993). "A bZIP protein, sisterless-a, collaborates with bHLH transcription factors early in *Drosophila* development to determine sex." Genes & Development **7**(9): 1688-1702.

Erickson, J. W. and Quintero, J. J. (2007). "Indirect effects of ploidy suggest X chromosome dose, not the X : A ratio, signals sex in *Drosophila*." Plos Biology **5**(12): 2821-2830.

Estrada, B., Casares, F. and Sanchez-Herrero, E. (2003). "Development of the genitalia in *Drosophila melanogaster*." Differentiation **71**(6): 299-310.

Ewen-Campen, B., Shaner, N., Panfilio, K. A., Suzuki, Y., Roth, S. and Extavour, C. G. (2011). "The maternal and early embryonic transcriptome of the milkweed bug *Oncopeltus fasciatus*." Bmc Genomics **12**: 21.

Ewing, B. and Green, P. (1998). "Base-calling of automated sequencer traces using phred. II. Error probabilities." Genome Research **8**(3): 186-194.

Ewing, B., Hillier, L., Wendl, M. C. and Green, P. (1998). "Base-calling of automated sequencer traces using phred. I. Accuracy assessment." Genome Research **8**(3): 175-185.

Fay, R. W., Cole, E. L. and Buckner, A. J. (1947). "Comparative residual effectiveness of organic insecticides against house flies and malaria mosquitoes." Journal of Economic Entomology **40**(5): 635-640.

Fields, S. and Song, O. K. (1989). "A novel genetic system to detect protein protein interactions." Nature **340**(6230): 245-246.

Fillinger, U. and Lindsay, S. W. (2006). "Suppression of exposure to malaria vectors by an order of magnitude using microbial larvicides in rural Kenya." Tropical Medicine & International Health **11**(11): 1629-1642.

Fillinger, U., Ndenga, B., Githeko, A. and Lindsay, S. W. (2009). "Integrated malaria vector control with microbial larvicides and insecticide-treated nets in western Kenya: a controlled trial." Bulletin of the World Health Organization **87**(9): 655-665.

Fire, A., Xu, S. Q., Montgomery, M. K., Kostas, S. A., Driver, S. E. and Mello, C. C. (1998). "Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*." Nature **391**(6669): 806-811.

Flickinger, T. W. and Salz, H. K. (1994). "The *Drosophila* sex determination gene *snf* encodes a nuclear-protein with sequence and functional similarity to the mammalian u1a snrnp protein." Genes & Development **8**(8): 914-925.

Florea, L., Di Francesco, V., Miller, J., Turner, R., Yao, A*., et al.* (2005). "Gene and alternative splicing annotation with AIR." Genome Research **15**(1): 54-66.

Foe, V. E. and Alberts, B. M. (1983). "Studies of nuclear and cytoplasmic behavior during the 5 mitotic-cycles that precede gastrulation in *Drosophila* embryogenesis." Journal of Cell Science **61**(MAY): 31-70.

Fraser, B. A., Weadick, C. J., Janowitz, I., Rodd, F. H. and Hughes, K. A. (2011). "Sequencing and characterization of the guppy (*Poecilia reticulata*) transcriptome." Bmc Genomics **12**: 202-216.

Fu, G. L., Condon, K. C., Epton, M. J., Gong, P., Jin, L., Condon, G. C., Morrison, N. I., Dafa'alla, T. H. and Alphey, L. (2007). "Female-specific insect lethality engineered using alternative splicing." Nature Biotechnology **25**(3): 353-357.

Fu, G. L., Lees, R. S., Nimmo, D., Aw, D., Jin, L*., et al.* (2010). "Female-specific flightless phenotype for mosquito control." Proceedings of the National Academy of Sciences of the United States of America **107**(10): 4550-4554.

Gabrieli, P., Falaguerra, A., Siciliano, P., Gomulski, L. M., Scolari, F., Zacharopoulou, A., Franz, G., Malacrida, A. R. and Gasperi, G. (2010). "Sex and the single embryo: early deveopment in the Mediterranean fruit fly, *Ceratitis capitata*." Bmc Developmental Biology **10**: 12-26.

Gailey, D. A., Billeter, J. C., Liu, J. H., Bauzon, F., Allendorfer, J. B. and Goodwin, S. F. (2006). "Functional conservation of the fruitless male sex-determination gene across 250 Myr of insect evolution." Molecular Biology and Evolution **23**(3): 633-643.

Gailey, D. A., Taylor, B. J. and Hall, J. C. (1991). "Elements of the fruitless locus regulate development of the muscle of lawrence, a male-specific structure in the abdomen of *Drosophila melanogaster* adults." Development **113**(3): 879-890.

Galindo, M. I., Pueyo, J. I., Fouix, S., Bishop, S. A. and Couso, J. P. (2007). "Peptides encoded by short ORFs control development and define a new eukaryotic gene family." Plos Biology **5**(5): 1052-1062.

Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M*., et al.* (2002). "Genome sequence of the human malaria parasite *Plasmodium falciparum*." Nature **419**(6906): 498-511.

Garrett-Engele, C. M., Siegal, M. L., Manoli, D. S., Williams, B. C., Li, H. and Baker, B. S. (2002). "*intersex*, a gene required for female sexual development in *Drosophila*, is expressed in both sexes and functions together with *doublesex* to regulate terminal differentiation." Development **129**(20): 4661-4675.

Gempe, T., Hasselmann, M., Schiott, M., Hause, G., Otte, M. and Beye, M. (2009). "Sex Determination in Honeybees: Two Separate Mechanisms Induce and Maintain the Female Pathway." Plos Biology **7**(10): 11.

Gibbons, J. G., Janson, E. M., Hittinger, C. T., Johnston, M., Abbot, P. and Rokas, A. (2009). "Benchmarking Next-Generation Transcriptome Sequencing for Functional and Evolutionary Genomics." Molecular Biology and Evolution **26**(12): 2731-2744.

Goltsev, Y., Fuse, N., Frasch, M., Zinzen, R. P., Lanzaro, G. and Levine, M. (2007). "Evolution of the dorsal-ventral patterning network in the mosquito, *Anopheles gambiae*." Development **134**(13): 2415-2424.

Goltsev, Y., Hsiong, W., Lanzaro, G. and Levine, M. (2004). "Different combinations of gap repressors for common stripes in *Anopheles* and *Drosophila* embryos." Developmental Biology **275**(2): 435-446.

Goltsev, Y., Rezende, G. L., Vranizan, K., Lanzaro, G., Valle, D. and Levine, M. (2009). "Developmental and evolutionary basis for drought tolerance of the *Anopheles gambiae* embryo." Developmental Biology **330**(2): 462-470.

Gomulski, L. M., Dimopoulos, G., Xi, Z. Y., Soares, M. B., Bonaldo, M. F., Malacrida, A. R. and Gasperi, G. (2008). "Gene discovery in an invasive tephritid model pest species, the Mediterranean fruit fly, *Ceratitis capitata*." Bmc Genomics **9**: 15.

Gong, P., Epton, M. J., Fu, G. L., Scaife, S., Hiscox, A*., et al.* (2005). "A dominant lethal genetic system for autocidal control of the Mediterranean fruitfly." Nature Biotechnology **23**(4): 453-456.

Gonzalez, A. N., Lu, H. and Erickson, J. W. (2008). "A shared enhancer controls a temporal switch between promoters during *Drosophila* primary sex determination." Proceedings of the National Academy of Sciences of the United States of America **105**(47): 18436-18441.

Gossen, M. and Bujard, H. (1992). "Tight control of gene expression in mammalian cells by tetracycline-responsive promoters." Proceedings of the National Academy of Sciences of the United States of America **89**(12): 5547-5551.

Gossen, M., Freundlieb, S., Bender, G., Muller, G., Hillen, W. and Bujard, H. (1995). "Transcriptional activation by tetracyclines in mammalian cells." Science **268**(5218): 1766-1769.

Gotz, S., Garcia-Gomez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H*., et al.* (2008). "High-throughput functional annotation and data mining with the Blast2GO suite." Nucleic Acids Research **36**(10): 3420-3435.

Granadino, B., Campuzano, S. and Sanchez, L. (1990). "The *Drosophila melanogaster fl(2)d* gene is needed for the female-specific splicing of sex-lethal RNA." Embo Journal **9**(8): 2597-2602.

Granadino, B., Penalva, L. O. F. and Sanchez, L. (1996). "The gene *fl(2)d* is needed for the sex-specific splicing of *transformer* pre-mRNA but not for *doublesex* pre-mRNA in *Drosophila melanogaster*." Molecular & General Genetics **253**(1-2): 26-31.

Graveley, B. R. (2000). "Sorting out the complexity of SR protein functions." Rna-a Publication of the Rna Society **6**(9): 1197-1211.

Graveley, B. R., Brooks, A. N., Carlson, J., Duff, M. O., Landolin, J. M*., et al.* (2011). "The developmental transcriptome of *Drosophila melanogaster*." Nature **471**(7339): 473-479.

Gregory, R., Darby, A. C., Irving, H., Coulibaly, M. B., Hughes, M*., et al.* (2011). "A De Novo Expression Profiling of *Anopheles funestus*, Malaria Vector in Africa, Using 454 Pyrosequencing." Plos One **6**(2): e17418.

Grossman, G. L., Rafferty, C. S., Clayton, J. R., Stevens, T. K., Mukabayire, O. and Benedict, M. Q. (2001). "Germline transformation of the malaria vector, *Anopheles gambiae*, with the piggyBac transposable element." Insect Molecular Biology **10**(6): 597-604.

Guerrero, F. D., Down, S. E., Djikeng, A., Wiley, G., Macmil, S., Saldivar, L., Najar, F. and Roe, B. A. (2009). "A Database of Expressed Genes From *Cochliomyia hominivorax* (Diptera: Calliphoridae)." Journal of Medical Entomology **46**(5): 1109-1116.

Guo, S. G., Zheng, Y., Joung, J. G., Liu, S. Q., Zhang, Z. H*., et al.* (2010). "Transcriptome sequencing and comparative analysis of cucumber flowers with different sex types." Bmc Genomics **11**: 384-397.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F*., et al.* (2009). "Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals." Nature **458**(7235): 223-227.

Guttman, M., Donaghey, J., Carey, B. W., Garber, M., Grenier, J. K*., et al.* (2011). "lincRNAs act in the circuitry controlling pluripotency and differentiation." Nature **477**(7364): 295-U60.

Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J*., et al.* (2010). "*Ab initio* reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs." Nature Biotechnology **28**(5): 503-U166.

Gygi, S. P., Rochon, Y., Franza, B. R. and Aebersold, R. (1999). "Correlation between protein and mRNA abundance in yeast." Molecular and Cellular Biology **19**(3): 1720-1730.

Hahn, M. W. and Lanzaro, G. C. (2005). "Female-biased gene expression in the malaria mosquito *Anopheles gambiae*." Current Biology **15**(6): R192-R193.

Hale, M. C., Jackson, J. R. and DeWoody, J. A. (2010). "Discovery and evaluation of candidate sex-determining genes and xenobiotics in the gonads of lake sturgeon (*Acipenser fulvescens*)." Genetica **138**(7): 745-756.

Hamada, F. N., Park, P. J., Gordadze, P. R. and Kuroda, M. I. (2005). "Global regulation of X chromosomal genes by the MSL complex in *Drosophila melanogaster*." Genes & Development **19**(19): 2289-2294.

Harris, A. F., Nimmo, D., McKemey, A. R., Kelly, N., Scaife, S., Donnelly, C. A., Beech, C., Petrie, W. D. and Alphey, L. (2011). "Field performance of engineered male mosquitoes." Nature Biotechnology **29**(11): 1034-U109.

Hasselmann, M., Gempe, T., Schiott, M., Nunes-Silva, C. G., Otte, M. and Beye, M. (2008). "Evidence for the evolutionary nascence of a novel sex determination pathway in honeybees." Nature **454**(7203): 519-U7.

Haugen, M., Flannery, E., Tomchaney, M., Mori, A., Behura, S. K., Severson, D. W. and Duman-Scheel, M. (2011). "Semaphorin-1a Is Required for *Aedes aegypt*i Embryonic Nerve Cord Development." Plos One **6**(6): e21694

Hediger, M., Burghardt, G., Siegenthaler, C., Buser, N., Hilfiker-Kleiner, D., Dubendorfer, A. and Bopp, D. (2003). "Sex determination in *Drosophila melanogaster* and *Musca domestica* converges at the level of the terminal regulator *doublesex*." Development Genes and Evolution **214**(1): 29-42.

Hediger, M., Burghardt, G., Siegenthaler, C., Buser, N., Hilfiker-Kleiner, D., Dubendorfer, A. and Bopp, D. (2004). "Sex determination in *Drosophila melanogaster* and *Musca domestica* converges at the level of the terminal regulator doublesex." Development Genes and Evolution **214**(1): 29-42.

Hediger, M., Henggeler, C., Meier, N., Perez, R., Saccone, G. and Bopp, D. (2010). "Molecular Characterization of the Key Switch F Provides a Basis for Understanding the Rapid Divergence of the Sex-Determining Pathway in the Housefly." Genetics **184**(1): 155-170.

Hedley, M. L. and Maniatis, T. (1991). "Sex-specific splicing and polyadenylation of *dsx* pre-messenger-RNA requires a sequence that binds specifically to tra-2 protein *in vitro*." Cell **65**(4): 579-586.

Heimpel, G. E. and de Boer, J. G. (2008). Sex determination in the Hymenoptera. Annual Review of Entomology. Palo Alto, Annual Reviews. **53:** 209-230.

Heinrich, J. C. and Scott, M. J. (2000). "A repressible female-specific lethal genetic system for making transgenic insect strains suitable for a sterile-release program." Proceedings of the National Academy of Sciences of the United States of America **97**(15): 8229-8232.

Heinrichs, V. and Baker, B. S. (1997). "*In vivo* analysis of the functional domains of the *Drosophila* splicing regulator RBP1." Proceedings of the National Academy of Sciences of the United States of America **94**(1): 115-120.

Heinrichs, V., Ryner, L. C. and Baker, B. S. (1998). "Regulation of sex-specific selection of fruitless 5 ' splice sites by transformer and transformer-2." Molecular and Cellular Biology **18**(1): 450-458.

Helinski, M. E. H., Hassan, M. M., El-Motasim, W. M., Malcolm, C. A., Knols, B. G. J. and El-Sayed, B. (2008). "Towards a sterile insect technique field release of *Anopheles arabiensis* mosquitoes in Sudan: Irradiation, transportation, and field cage experimentation." Malaria Journal **7**: 10.

Hempel, L. U. and Oliver, B. (2007). "Sex-specific Doublesex(M) expression in subsets of *Drosophila* somatic gonad cells." Bmc Developmental Biology **7**: 15.

Hiroyoshi, T. (1964). "Sex-limited inheritance + abnormal sex ratio in strains of housefly." Genetics **50**(3): 373-&.

Holt, R. A., Subramanian, G. M., Halpern, A., Sutton, G. G., Charlab, R*., et al.* (2002). "The genome sequence of the malaria mosquito *Anopheles gambiae*." Science **298**(5591): 129-149.

Horabin, J. I. and Schedl, P. (1993). "*sex-lethal* autoregulation requires multiple cis-acting elements upstream and downstream of the male exon and appears to depend largely on controlling the use of the male exon 5' splice-site." Molecular and Cellular Biology **13**(12): 7734-7746.

Horn, C. and Wimmer, E. A. (2003). "A transgene-based, embryo-specific lethality system for insect pest management." Nature Biotechnology **21**(1): 64-70.

Hoshijima, K., Inoue, K., Higuchi, I., Sakamoto, H. and Shimura, Y. (1991). "Control of *doublesex* alternative splicing by transformer and transformer-2 in *Drosophila*." Science **252**(5007): 833-836.

Huang, R., Jaritz, M., Guenzl, P., Vlatkovic, I., Sommer, A*., et al.* (2011). "An RNA-Seq Strategy to Detect the Complete Coding and Non-Coding Transcriptome Including Full-Length Imprinted Macro ncRNAs." Plos One **6**(11): 13.

Huang, X. Q. and Madan, A. (1999). "CAP3: A DNA sequence assembly program." Genome Research **9**(9): 868-877.

Inoue, H. and Hiroyoshi, T. (1982). "A male-determining factor on autosome-1 and occurrence of male-recombination in the housefly, *Musca domestica*." Japanese Journal of Genetics **57**(3): 221-229.

Inoue, K., Hoshijima, K., Higuchi, I., Sakamoto, H. and Shimura, Y. (1992). "Binding of the *Drosophila* transformer and transformer-2 proteins to the regulatory elements of *doublesex* primary trnascripts for sex-specific RNA processing " Proceedings of the National Academy of Sciences of the United States of America **89**(17): 8092-8096.

Inoue, K., Hoshijima, K., Sakamoto, H. and Shimura, Y. (1990). "Binding of the Drosophila *sex-lethal* gene product to the alternative splice site of *transformer* primary transcript." Nature **344**(6265): 461-463.

Irmingerfinger, I. and Nothinger, R. (1995). "The *Drosophila melanogaster* gene *lethal(3)73ah* encodes a ring finger protein homologous to the oncoproteins mel-18 and bmi-1." Gene **163**(2): 203-208.

Ito, H., Fujitani, K., Usui, K., ShimizuNishikawa, K., Tanaka, S. and Yamamoto, D. (1996). "Sexual orientation in *Drosophila* is altered by the satori mutation in the sex-determination gene *fruitless* that encodes a zinc finger protein with a BTB domain." Proceedings of the National Academy of Sciences of the United States of America **93**(18): 9687-9692.

James, A. A., Benedict, M. Q., Christophides, G. K., Jacobs-Lorena, M. and Olson, K. E. (2006). Evaluation of drive mechanisms (including transgenes and drivers) in different environmental conditions and genetic backgrounds. Dordrecht, Springer.

Jinks TM, Polydorides AD, Calhoun G and P., S. (2000). "The JAK/STAT signaling pathway is required for the initial choice of sexual identity in *Drosophila melanogaster*." Mol. Cell. **5**(3): 581-587.

Juhn, J. and James, A. A. (2006). "*oskar* gene expression in the vector mosquitoes, *Anopheles gambiae* and *Aedes aegypti*." Insect Molecular Biology **15**(3): 363-372.

Junakovic, N., Terrinoni, A., Di Franco, C., Vieira, C. and Loevenbruck, C. (1998). "Accumulation of transposable elements in the heterochromatin and on the Y chromosome of *Drosophila simulans* and *Drosophila melanogaster*." Journal of Molecular Evolution **46**(6): 661-668.

Kampfer, P., Matthews, H., Glaeser, S. P., Martin, K., Lodders, N. and Faye, I. (2011). "Elizabethkingia anophelis sp. nov., isolated from the midgut of the mosquito *Anopheles gambiae.*" **11**(61): 2670-2675.

Katz, Y., Wang, E. T., Airoldi, E. M. and Burge, C. B. (2010). "Analysis and design of RNA sequencing experiments for identifying isoform regulation." Nature Methods **7**: 1009-1015.

Keisman, E. L. and Baker, B. S. (2001). "The *Drosophila* sex determination hierarchy modulates wingless and decapentaplegic signaling to deploy dachshund sex-specifically in the genital imaginal disc." Development **128**(9): 1643-1656.

Keisman, E. L., Christiansen, A. E. and Baker, B. S. (2001). "The sex determination gene *doublesex* regulates the A/P organizer to direct sex-specific patterns of growth in the *Drosophila* genital imaginal disc." Developmental Cell **1**(2): 215-225.

Keyes, L. N., Cline, T. W. and Schedl, P. (1992). "The primary sex determination signal of *Drosophila* acts at the level of transcription." Cell **68**(5): 933-943.

Knipling, E. F. (1955). "Possibilities of insect control or eradication through the use of sexually sterile males." Journal of Economic Entomology **48**(4): 459-462.

Kondo, T., Plaza, S., Zanet, J., Benrabah, E., Valenti, P., Hashimoto, Y., Kobayashi, S., Payre, F. and Kageyama, Y. (2010). "Small Peptides Switch the Transcriptional Activity of *Shavenbaby* During *Drosophila* Embryogenesis." Science **329**(5989): 336-339.

Kong, L., Zhang, Y., Ye, Z. Q., Liu, X. Q., Zhao, S. Q., Wei, L. and Gao, G. (2007). "CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine." Nucleic Acids Research **35**: W345-W349.

Koudou, B. G., Koffi, A. A., Malone, D. and Hemingway, J. (2011). "Efficacy of PermaNet (R) 2.0 and PermaNet (R) 3.0 against insecticide-resistant *Anopheles gambiae* in experimental huts in Cote d'Ivoire." Malaria Journal **10**: 10.

Krafsur, E. S., Whitten, C. J. and Novy, J. E. (1987). "Screwworm eradication in North and Central America." Parasitology Today **3**(5): 131-137.

Kramer, S. G., Jinks, T. M., Schedl, P. and Gergen, J. P. (1999). "Direct activation of *Sex-lethal* transcription by the *Drosophila* Runt protein." Development **126**(1): 191-200.

Krzywinski, J., Nusskern, D. R., Kern, M. K. and Besansky, N. J. (2004). "Isolation and characterization of Y chromosome sequences from the African malaria mosquito *Anopheles gambiae*." Genetics **166**(3): 1291-1302.

Krzywinski, J., Sangare, D. and Besansky, N. J. (2005). "Satellite DNA from the Y chromosome of the malaria vector *Anopheles gambiae*." Genetics **169**(1): 185-196.

Kuhn, S., Sievert, V. and Traut, W. (2000). "The sex-determining gene *doublesex* in the fly *Megaselia scalaris*: Conserved structure and sex-specific splicing." Genome **43**(6): 1011-1020.

Kulathinal, R. J., Skwarek, L., Morton, R. A. and Singh, R. S. (2003). "Rapid evolution of the sex-determining gene, *transformer*: Structural diversity and rate heterogeneity among sibling species of *Drosophila*." Molecular Biology and Evolution **20**(3): 441-452.

Kumar, S. and Blaxter, M. L. (2010). "Comparing *de novo* assemblers for 454 transcriptome data." Bmc Genomics **11**: 571-582.

Lagos, D., Koukidou, M., Savakis, C. and Komitopoulou, K. (2007). "The transformer gene in *Bactrocera oleae*: the genetic switch that determines its sex fate." Insect Molecular Biology **16**(2): 221-230.

Lagos, D., Ruiz, M. F., Sanchez, L. and Komitopoulou, K. (2005). "Isolation and characterization of the *Bactrocera oleae* genes orthologous to the sex detennining Sex-lethal and doublesex genes of *Drosophila melanogaster*." Gene **348**: 111-121.

Lam, B. J., Bakshi, A., Ekinci, F. Y., Webb, J., Graveley, B. R. and Hertel, K. J. (2003). "Enhancer-dependent 5 '-splice site control of *fruitless* pre-mRNA splicing." Journal of Biological Chemistry **278**(25): 22740-22747.

Lamond, A. I. (2002). "Molecular biology of the cell, 4th edition." **417**(6887): 383.

Lanciotti, R. S., Ludwig, M. L., Rwaguma, E. B., Lutwama, J. J., Kram, T. M., Karabatsos, N., Cropp, B. C. and Miller, B. R. (1998). "Emergence of epidemic O'nyong-nyong fever in Uganda after a 35-year absence: Genetic characterization of the virus." Virology **252**(1): 258-268.

Lapuk, A., Marr, H., Jakkula, L., Pedro, H., Bhattacharya, S.*, et al.* (2010). "Exon-Level Microarray Analyses Identify Alternative Splicing Programs in Breast Cancer." Molecular Cancer Research **8**(7): 961-974.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A.*, et al.* (2007). "Clustal W and clustal X version 2.0." Bioinformatics **23**(21): 2947-2948.

Lawniczak, M. K. N., Emrich, S. J., Holloway, A. K., Regier, A. P., Olson, M.*, et al.* (2010). "Widespread Divergence Between Incipient *Anopheles gambiae* Species Revealed by Whole Genome Sequences." Science **330**(6003): 512-514.

Lee, G., Foss, M., Goodwin, S. F., Carlo, T., Taylor, B. J. and Hall, J. C. (2000). "Spatial, temporal, and sexually dimorphic expression patterns of the *fruitless* gene in the *Drosophila* central nervous system." Journal of Neurobiology **43**(4): 404-426.

Lee, G., Hall, J. C. and Park, J. H. (2002). "*Doublesex* gene expression in the central nervous system of *Drosophila melanogaster*." Journal of Neurogenetics **16**(4): 229-248.

Lehmann, T. and Diabate, A. (2008). "The molecular forms of *Anopheles gambiae*: A phenotypic perspective." Infection Genetics and Evolution **8**(5): 737-746.

Lenhart, A., Eigege, A., Kal, A., Pam, D., Miri, E. S.*, et al.* (2007). "Contributions of different mosquito species to the transmission of lymphatic filariasis in central Nigeria: implications for monitoring infection by PCR in mosquito pools." Filaria journal **6**: 14.

Levin, J. Z., Yassour, M., Adiconis, X. A., Nusbaum, C., Thompson, D. A., Friedman, N., Gnirke, A. and Regev, A. (2010). "Comprehensive comparative analysis of strand-specific RNA sequencing methods." Nature Methods **7**(9): 709-715.

Li, H. and Baker, B. S. (1998). "*her*, a gene required for sexual differentiation in *Drosophila*, encodes a zinc finger protein with characteristics of ZFY-like proteins and is expressed independently of the sex determination hierarchy." Development **125**(2): 225-235.

Li, H. and Baker, B. S. (1998). "*hermaphrodite* and *doublesex* function both dependently and independently to control various aspects of sexual differentiation in *Drosophila*." Development **125**(14): 2641-2651.

Li, H. and Durbin, R. (2010). "Fast and accurate long-read alignment with Burrows-Wheeler transform." Bioinformatics **26**(5): 589-595.

Liang, H. L., Nien, C. Y., Liu, H. Y., Metzstein, M. M., Kirov, N. and Rushlow, C. (2008). "The zinc-finger protein Zelda is a key activator of the early zygotic genome in *Drosophila*." Nature **456**(7220): 400-U67.

Lin, M. F., Jungreis, I. and Kellis, M. (2011). "PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions." Bioinformatics **27**(13): I275-I282.

Lindh, J. M., Borg-Karlson, A. K. and Faye, I. (2008). "Transstadial and horizontal transfer of bacteria within a colony of *Anopheles gambiae* (Diptera: Culicidae) and oviposition response to bacteria-containing water." Acta Tropica **107**(3): 242-250.

Lloyd, V. K., Dyment, D., Sinclair, D. A. R. and Grigliatti, T. A. (2003). "Different patterns of gene silencing in position-effect variegation." Genome **46**(6): 1104-1117.

Lobo, N. F., Clayton, J. R., Fraser, M. J., Kafatos, F. C. and Collins, F. H. (2006). "High efficiency germ-line transformation of mosquitoes." Nature Protocols **1**(3): 1312-1317.

Lofgren, C. S., Dame, D. A., Breeland, S. G., Weidhaas, D. E., Jeffery, G., Kaiser, R., Ford, H. R., Boston, M. D. and Baldwin, K. F. (1974). "Release of chemosterilized males for control of *Anopheles albimanus* in El Salvador .3. field methods and population-control." American Journal of Tropical Medicine and Hygiene **23**(2): 288-297.

Lombardo, F., Lycett, G. J., Lanfrancotti, A., Coluzzi, M. and Arca, B. (2009). "Analysis of apyrase 5' upstream region validates improved *Anopheles gambiae* transformation technique." BMC research notes **2**: 24.

Long, J. C. and Caceres, J. F. (2009). "The SR protein family of splicing factors: master regulators of gene expression." Biochemical Journal **417**: 15-27.

Lott, S. E., Villalta, J. E., Schroth, G. P., Luo, S. J., Tonkin, L. A. and Eisen, M. B. (2011). "Noncanonical Compensation of Zygotic X Transcription in Early *Drosophila melanogaste*r Development Revealed through Single-Embryo RNA-Seq." Plos Biology **9**(2): e1000590.

Luo, S. Z. D., Shi, G. W. and Baker, B. S. (2011). "Direct targets of the *D. melanogaster* DSX(F) protein and the evolution of sexual development." Development **138**(13): 2761-2771.

Lycett, G. J., Amenya, D. and Lynd, A. (2011). "The *Anopheles gambiae* alpha-tubulin-1b promoter directs neuronal, testes and developing imaginal tissue specific expression and is a sensitive enhancer detector." Insect Molecular Biology **21**(1): 79-88.

Lycett, G. J., Kafatos, F. C. and Loukeris, T. G. (2004). "Conditional expression in the malaria mosquito *Anophels stephensi* with tet-on and tet-off systems." Genetics **167**(4): 1781-1790.

Lynch, K. W. and Maniatis, T. (1996). "Assembly of specific SR protein complexes on distinct regulatory elements of the *Drosophila doublesex* splicing enhancer." Genes & Development **10**(16): 2089-2101.

Ma, X., Zhu, Y. J., Li, C. F., Shang, Y. L., Meng, F. J., Chen, S. L. and Miao, L. (2011). "Comparative transcriptome sequencing of germline and somatic tissues of the *Ascaris suum* gonad." Bmc Genomics **12**: 11.

Magnusson, K., Mendes, A. M., Windbichler, N., Papathanos, P. A., Nolan, T., Dottorini, T., Rizzi, E., Christophides, G. K. and Crisanti, A. (2011). "Transcription Regulation of Sex-Biased Genes during Ontogeny in the Malaria Vector *Anopheles gambiae*." Plos One **6**(6): e21572.

Mardis, E. R. (2008). "The impact of next-generation sequencing technology on genetics." Trends in Genetics **24**(3): 133-141.

Mardis, E. R. (2008). Next-generation DNA sequencing methods. Annual Review of Genomics and Human Genetics. Palo Alto, Annual Reviews. **9:** 387-402.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S*., et al.* (2005). "Genome sequencing in microfabricated high-density picolitre reactors." Nature **437**(7057): 376-380.

Marinotti, O., Calvo, E., Nguyen, Q. K., Dissanayake, S., Ribeiro, J. M. C. and James, A. A. (2006). "Genome-wide analysis of gene expression in adult *Anopheles gambiae*." Insect Molecular Biology **15**(1): 1-12.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. and Gilad, Y. (2008). "RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays." Genome Research **18**(9): 1509-1517.

Martin, I., Ruiz, M. F. and Sanchez, L. (2011). "The gene transformer-2 of *Sciara* (*Diptera*, *Nematocera*) and its effect on *Drosophila* sexual development." Bmc Developmental Biology **11**: 19-32.

Martin, J. A. and Wang, Z. (2011). "Next-generation transcriptome assembly." Nature Reviews Genetics **12**(10): 671-682.

Martinez-Torres, D., Chandre, F., Williamson, M. S., Darriet, F., Berge, J. B., Devonshire, A. L., Guillet, P., Pasteur, N. and Pauron, D. (1998). "Molecular characterization of pyrethroid knockdown resistance (kdr) in the major malaria vector *Anopheles gambiae* S.S." Insect Molecular Biology **7**(2): 179-184.

Mattox, W., McGuffin, M. E. and Baker, B. S. (1996). "A negative feedback mechanism revealed by functional analysis of the alternative isoforms of the *Drosophila* splicing regulator *transformer-2*." Genetics **143**(1): 303-314.

Mattox, W., Palmer, M. J. and Baker, B. S. (1990). "Alternative splicing of the sex determination gene *transformer-2* is sex-specific in the germ line but not in the soma." Genes & Development **4**(5): 789-805.

McDonald, P. T., Hausermann, W. and Lorimer, N. (1977). "Sterility introduced by release of genetically altered males to a domestic population of *Aedes aegypti* at Kenya coast." American Journal of Tropical Medicine and Hygiene **26**(3): 553-561.

McGuffin, M. E., Chandler, D., Somaiya, D., Dauwalder, B. and Mattox, W. (1998). "Autoregulation of *transformer-2* alternative splicing is necessary for normal male fertility in *Drosophila*." Genetics **149**(3): 1477-1486.

McRobert, S. P. and Tompkins, L. (1985). "The effect of *transformer*, *doublesex* and *intersex* mutations on the sexual behavior of *Drosophila melanogaster*." Genetics **111**(1): 89-96.

Meise, M., Hilfiker-Kleiner, D., Dubendorfer, A., Brunner, C., Nothiger, R. and Bopp, D. (1998). "*Sex-lethal*, the master sex-determining gene in *Drosophila*, is not sex-specifically regulated in *Musca domestica*." Development **125**(8): 1487-1494.

Meister, S., Kanzok, S. M., Zheng, X. L., Luna, C., Li, T. R*., et al.* (2005). "Immune signaling pathways regulating bacterial and malaria parasite infection of the mosquito *Anopheles gambiae*." Proceedings of the National Academy of Sciences of the United States of America **102**(32): 11420-11425.

Meldrum, C., Doyle, M. A. and Tothill, R. W. (2011). "Next-generation sequencing for cancer diagnostics: a practical perspective." The Clinical biochemist. Reviews / Australian Association of Clinical Biochemists **32**(4): 177-195.

Meredith, J. M., Basu, S., Nimmo, D. D., Larget-Thiery, I., Warr, E. L*., et al.* (2011). "Site-Specific Integration and Expression of an Anti-Malarial Gene in Transgenic *Anopheles gambiae* Significantly Reduces *Plasmodium* Infections." Plos One **6**(1): 9.

Merrill, P. T., Sweeton, D. and Wieschaus, E. (1988). "Requirements for autosomal gene activity during precellular stages of *Drosophila melanogaster*." Development **104**(3): 495-509.

Merrill, P. T., Sweeton, D. and Wieschaus, E. (1988). "Requirements for autosomal gene activity during precellular stages of *Drosophila melanogaster*
" Development **104**(3): 495-509.

Metzker, M. L. (2010). "Applications of Next Generation Sequencing technologies - the next generation." Nature Reviews Genetics **11**(1): 31-46.

Miller, L. H. and Hoffman, S. L. (1998). "Research toward vaccines against malaria." Nature Medicine **4**(5): 520-524.

Milligan, J. F., Groebe, D. R., Witherell, G. W. and Uhlenbeck, O. C. (1987). "Oligoribonucleotide synthesis using t7 RNA-polymerase and synthetic DNA templates." Nucleic Acids Research **15**(21): 8783-8798.

Milne, I., Bayer, M., Cardle, L., Shaw, P., Stephen, G., Wright, F. and Marshall, D. (2010). "Tablet-next generation sequence assembly visualization." Bioinformatics **26**(3): 401-402.

Mitsuhashi, M. (1996). "Technical report .2. Basic requirements for designing optimal PCR primers." Journal of Clinical Laboratory Analysis **10**(5): 285-293.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. and Wold, B. (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq." Nature Methods **5**(7): 621-628.

Mufunda, J., Nyarango, P., Usman, A., Gebremeskel, T., Mebrahtu, G*., et al.* (2007). "Roll back malaria - an African success story in Eritrea." Samj South African Medical Journal **97**(1): 46-50.

Muller, P., Warr, E., Stevenson, B. J., Pignatelli, P. M., Morgan, J. C*., et al.* (2008). "Field-Caught Permethrin-Resistant *Anopheles gambiae* Overexpress CYP6P3, a P450 That Metabolises Pyrethroids." Plos Genetics **4**(11): 10.

Munhenga, G., Brooke, B. D., Chirwa, T. F., Hunt, R. H., Coetzee, M., Govender, D. and Koekemoer, L. L. (2011). "Evaluating the potential of the sterile insect technique for malaria control: relative fitness and mating compatibility between laboratory colonized and a wild population of *Anopheles arabiensis* from the Kruger National Park, South Africa." Parasites & Vectors **4**: 11.

Murray, C., Rosenfeld, L., Lim, S., Andrews, K., Foreman, K*., et al.* (2012). "Global malaria mortality between 1980 and 2010: a systematic analysis." The Lancet **379**: 413-431.

Nagengast, A. A., Stitzinger, S. M., Tseng, C. H., Mount, S. M. and Salz, H. K. (2003). "Sex-lethal splicing autoregulation in vivo: interactions between SEX-LETHAL, the U1 snRNP and U2AF underlie male exon skipping." Development **130**(3): 463-471.

Nagoshi, R. N., McKeown, M., Burtis, K. C., Belote, J. M. and Baker, B. S. (1988). "The control of alternative splicing at genes regulating sexual differentiation in *Drosophila melanogaster*." Cell **53**(2): 229-236.

Ndiath, M. O., Cohuet, A., Gaye, A., Konate, L., Mazenot, C., Faye, O., Boudin, C., Sokhna, C. and Trape, J. F. (2011). "Comparative susceptibility to *Plasmodium falciparum* of the molecular forms M and S of *Anopheles gambiae* and *Anopheles arabiensis*." Malaria Journal **10**: 269-276.

Nielsen, R., Paul, J. S., Albrechtsen, A. and Song, Y. S. (2011). "Genotype and SNP calling from next-generation sequencing data." Nature Reviews Genetics **12**(6): 443-451.

Nien, C. Y., Liang, H. L., Butcher, S., Sun, Y. J., Fu, S. B., Gocha, T., Kirov, N., Manak, J. R. and Rushlow, C. (2011). "Temporal Coordination of Gene Networks by Zelda in the Early *Drosophila* Embryo." Plos Genetics **7**(10): 16.

Niimi, T., Sahara, K., Oshima, H., Yasukochi, Y., Ikeo, K. and Traut, W. (2006). "Molecular cloning and chromosomal localization of the *Bombyx Sex-lethal* gene." Genome **49**(3): 263-268.

Nirmala, X., Marinotti, O. and James, A. A. (2005). "The accumulation of specific mRNAs following multiple blood meals in *Anopheles gambiae*." Insect Molecular Biology **14**(1): 95-103.

Niu, B. L., Meng, Z. Q., Tao, Y. Z., Lu, S. L., Weng, H. B., He, L. H. and Shen, W. F. (2005). "Cloning and alternative splicing analysis of *Bombyx mori* transformer-2 gene using silkworm EST database." Acta Biochimica Et Biophysica Sinica **37**(11): 728-736.

Nolan, T., Petris, E., Muller, H. M., Cronin, A., Catteruccia, F. and Crisanti, A. (2011). "Analysis of Two Novel Midgut-Specific Promoters Driving Transgene Expression in *Anopheles stephensi* Mosquitoes." Plos One **6**(2): 8.

Nothiger, R., Leuthold, M., Andersen, N., Gerschwiler, P., Gruter, A., Keller, W., Leist, C., Roost, M. and Schmid, H. (1987). "Genetic and developmental analysis of the sex-determining gene *double sex* (*dsx*) of *Drosophila melanogaster*." Genetical Research **50**(2): 113-123.

Novaes, E., Drost, D. R., Farmerie, W. G., Pappas, G. J., Grattapaglia, D., Sederoff, R. R. and Kirst, M. (2008). "High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome." Bmc Genomics **9**: 14.

Nyren, P., Pettersson, B. and Uhlen, M. (1993). "Solid-phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay." Analytical Biochemistry **208**(1): 171-175.

O' Neil, M. T. and Belote, J. M. (1992). "Interspecific comparison of the *transformer* gene of *Drosophila* reveals an unusually high degree of evolutionary divergence." Genetics **131**(1): 113-128.

Ohbayashi, F., Suzuki, M. G., Mita, K., Okano, K. and Shimada, T. (2001). "A homologue of the *Drosophila doublesex* gene is transcribed into sex-specific mRNA isoforms in the silkworm, *Bombyx mori*." Comparative Biochemistry and Physiology B-Biochemistry & Molecular Biology **128**(1): 145-158.

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H*., et al.* (2002). "Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs." Nature **420**(6915): 563-573.

Olafson, P. U., Lohmeyer, K. H. and Dowd, S. E. (2010). "Analysis of expressed sequence tags from a significant livestock pest, the stable fly (*stomoxys calcitrans*), identifies transcripts with a putative role in chemosensation and sex determination." Archives of Insect Biochemistry and Physiology **74**(3): 179-204.

Oliveira, D., Werren, J. H., Verhulst, E. C., Giebel, J. D., Kamping, A., Beukeboom, L. W. and van de Zande, L. (2009). "Identification and characterization of the *doublesex* gene of *Nasonia*." Insect Molecular Biology **18**(3): 315-324.

Orom, U. A., Derrien, T., Beringer, M., Gumireddy, K., Gardini, A*., et al.* (2010). "Long Noncoding RNAs with Enhancer-like Function in Human Cells." Cell **143**(1): 46-58.

Ota, T., Suzuki, Y., Nishikawa, T., Otsuki, T., Sugiyama, T*., et al.* (2004). "Complete sequencing and characterization of 21,243 full-length human cDNAs." Nature Genetics **36**(1): 40-45.

Pane, A., De Simone, A., Saccone, G. and Polito, C. (2005). "Evolutionary conservation of *Ceratitis capitata transformer* gene function." Genetics **171**(2): 615-624.

Pane, A., Salvemini, M., Bovi, P. D., Polito, C. and Saccone, G. (2002). "The *transformer* gene in *Ceratitis capitata* provides a genetic basis for selecting and remembering the sexual fate." Development **129**(15): 3715-3725.

Pannuti, A., Kocacitak, T., Lucchesi, J. C. and Kenghong, T. (2000). "*Drosophila* as a Model for the Study of Sex Determination in Anopheline and Aedine Mosquitoes." Area-wide control of fruit flies and other insect pests.: 263-269.

Papathanos, P. A., Windbichler, N., Menichelli, M., Burt, A. and Crisanti, A. (2009). "The *vasa* regulatory region mediates germline expression and maternal transmission of proteins in the malaria mosquito *Anopheles gambiae:* a versatile tool for genetic control strategies." Bmc Molecular Biology **10**: 65-78.

Pareek, C. S., Smoczynski, R. and Tretyn, A. (2011). "Sequencing technologies and genome sequencing." Journal of Applied Genetics **52**(4): 413-435.

Patterson, R., Weidhaas, D. E., Ford, H. R. and Lofgren, C. S. (1970). "Suppression and elimination of an island population of *Culex pipiens quinquefasciatus* with sterile males." Science **168**(3937): 1368-&.

Penn, J. K. M., Graham, P., Deshpande, G., Calhoun, G., Chaouki, A. S., Salz, H. K. and Schedl, P. (2008). "Functioning of the *Drosophila* Wilms'-tumor-1-associated protein homolog, Fl(2)d, in sex-lethal-dependent alternative splicing." Genetics **178**(2): 737-748.

Phillips-Howard, P. A., Nahlen, B. L., Kolczak, M. S., Hightower, A. W., ter Kuile, F. O*., et al.* (2003). "Efficacy of permethrin-treated bed nets in the prevention of mortality in young children in an area of high perennial malaria transmission in western Kenya." American Journal of Tropical Medicine and Hygiene **68**(4): 23-29.

Phuc, H. K., Andreasen, M. H., Burton, R. S., Vass, C., Epton, M. J*., et al.* (2007). "Late-acting dominant lethal genetic systems and mosquito control." Bmc Biology **5**: 11.

Pitts, R. J., Rinker, D. C., Jones, P. L., Rokas, A. and Zwiebel, L. J. (2011). "Transcriptome profiling of chemosensory appendages in the malaria vector *Anopheles gambiae* reveals tissue- and sex-specific signatures of odor coding." Bmc Genomics **12**: 17.

Poelchau, M., Reynolds, J., Denlinger, D., Elsik, C. and Armbruster, P. (2011). "A de novo transcriptome of the Asian tiger mosquito, *Aedes albopictus*, to identify candidate transcripts for diapause preparation." Bmc Genomics **12**(1): 619.

Powers, A. M., Brault, A. C., Tesh, R. B. and Weaver, S. C. (2000). "Re-emergence of chikungunya and o'nyong-nyong viruses: evidence for distinct geographical lineages and distant evolutionary relationships." Journal of General Virology **81**: 471-479.

Price, D. P., Nagarajan, V., Churbanov, A., Houde, P., Milligan, B., Drake, L. L., Gustafson, J. E. and Hansen, I. A. (2011). "The fat body transcriptomes of the yellow fever mosquito *Aedes aegypti*, pre- and post- blood meal." Plos One **6**(7): e22573.

Pultz, M. A., Carson, G. S. and Baker, B. S. (1994). "A genetic-analysis of *hermaphrodite*, a pleiotropic sex determination gene in *Drosophila melanogaster*." Genetics **136**(1): 195-207.

Raghavendra, K., Barik, T. K., Reddy, B. P. N., Sharma, P. and Dash, A. P. (2011). "Malaria vector control: from past to future." Parasitology Research **108**(4): 757-779.

Ranson, H., Jensen, B., Vulule, J. M., Wang, X., Hemingway, J. and Collins, F. H. (2000). "Identification of a point mutation in the voltage-gated sodium channel gene of Kenyan *Anopheles gambiae* associated with resistance to DDT and pyrethroids." Insect Molecular Biology **9**(5): 491-497.

RBM (2008). "The global malaria action plan for a malaria free world. http://www.rollbackmalaria.org/gmap."

RBM (2011). "Refined/Updated GMAP Objectives, Targets, Milestones and Priorities Beyond 2011: Roll Back Malaria, 2011. http://www.rbm.who.int/gmap/gmap2011update.pdf."

Rendon, P., McInnis, D., Lance, D. and Stewart, J. (2004). "Medfly (Diptera : Tephritidae) genetic sexing: Large-scale field comparison of males-only and bisexual sterile fly releases in Guatemala." Journal of Economic Entomology **97**(5): 1547-1553.

Rideout, E. J., Billeter, J. C. and Goodwin, S. F. (2007). "The sex-determination genes *fruitless* and *doublesex* specify a neural substrate required for courtship song." Current Biology **17**(17): 1473-1478.

Riehle, M. M., Guelbeogo, W. M., Gneme, A., Eiglmeier, K., Holm, I*., et al.* (2011). "A Cryptic Subgroup of *Anopheles gambiae* Is Highly Susceptible to Human Malaria Parasites." Science **331**(6017): 596-598.

Robert, C., Nieminen, J., Dufort, I., Gagne, D., Grant, J., Cagnone, G. and Sirard, M. (2011). "Combining resources to obtain a comprehensive survey of the bovine embryo transcriptome through deep sequencing and microarrays." Mol Reprod Dev.: 651-664.

Robinett, C. C., Vaughan, A. G., Knapp, J. M. and Baker, B. S. (2010). "Sex and the Single Cell. II. There Is a Time and Place for Sex." Plos Biology **8**(5): 16.

Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G. and Mesirov, J. P. (2011). "Integrative genomics viewer." Nature Biotechnology **29**(1): 24-26.

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

Rogic, S., Mackworth, A. K. and Ouellette, F. B. F. (2001). "Evaluation of gene-finding programs on mammalian sequences." Genome Research **11**(5): 817-832.

Rohr, C. J. B., Ranson, H., Wang, X. L. and Besansky, N. J. (2002). "Structure and evolution of mtanga, a retrotransposon actively expressed on the Y chromosome of the African malaria vector *Anopheles gambiae*." Molecular Biology and Evolution **19**(2): 149-162.

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M. and Nyren, P. (1996). "Real-time DNA sequencing using detection of pyrophosphate release." Analytical Biochemistry **242**(1): 84-89.

Ronaghi, M., Uhlen, M. and Nyren, P. (1998). "A sequencing method based on real-time pyrophosphate." Science **281**(5375): 363-+.

Rosinski-Chupin, I., Briolay, J., Brouilly, P., Perrot, S., Gomez, S. M*., et al.* (2007). "SAGE analysis of mosquito salivary gland transcriptomes during *Plasmodium* invasion." Cellular Microbiology **9**(3): 708-724.

Ross, R. (1897). "On some Peculiar Pigmented Cells Found in Two Mosquitos Fed on Malarial Blood." British medical journal **2**(1929): 1786-8.

Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W*., et al.* (2011). "An integrated semiconductor device enabling non-optical genome sequencing." Nature **475**(7356): 348-352.

Rouget, C., Papin, C., Boureux, A., Meunier, A. C., Franco, B., Robine, N., Lai, E. C., Pelisson, A. and Simonelig, M. (2010). "Maternal mRNA deadenylation and decay by the piRNA pathway in the early Drosophila embryo." Nature **467**(7319): 1128-U144.

Ruiz, M. F., Milano, A., Salvemini, M., Eirin-Lopez, J. M., Perondini, A. L. P., Selivon, D., Polito, C., Saccone, G. and Sanchez, L. (2007). "The gene transformer of *Anastrepha* fruit flies (Diptera, *tephritidae*) and its evolution in insects." Plos One **2**(11): e1239.

Rwaguma, E. B., Lutwama, J. J., Sempala, S. D. K., Kiwanuka, N., Kamugisha, J*., et al.* (1997). "Emergence of epidemic O'nyong-nyong fever in southwestern Uganda, after an absence of 35 years." Emerging Infectious Diseases **3**(1): 77-77.

Ryner, L. C., Goodwin, S. F., Castrillon, D. H., Anand, A., Villella, A., Baker, B. S., Hall, J. C., Taylor, B. J. and Wasserman, S. A. (1996). "Control of male sexual behavior and sexual orientation in *Drosophila* by the fruitless gene." Cell **87**(6): 1079-1089.

Saccone, G., Peluso, I., Artiaco, D., Giordano, E., Bopp, D. and Polito, L. C. (1998). "The *Ceratitis capitata* homologue of the *Drosophila* sex-determining gene Sex-lethal is structurally conserved, but not sex-specifically regulated." Development **125**(8): 1495-1500.

Saccone, G., Salvemini, M., Pane, A. and Polito, L. C. (2008). "Masculinization of XX *Drosophila* transgenic flies expressing the *Ceratitis capitata* Doublesex(M) isoform." International Journal of Developmental Biology **52**(8): 1051-1057.

Saccone, G., Salvemini, M. and Polito, L. C. (2011). "The transformer gene of *Ceratitis capitata*: a paradigm for a conserved epigenetic master regulator of sex determination in insects." Genetica **139**(1): 99-111.

Sachs, J. and Malaney, P. (2002). "The economic and social burden of malaria." Nature **415**(6872): 680-685.

Sackton, T. B., Montenegro, H., Hartl, D. L. and Lemos, B. (2011). "Interspecific Y chromosome introgressions disrupt testis-specific gene expression and male reproductive phenotypes in *Drosophila*." Proceedings of the National Academy of Sciences of the United States of America **108**(41): 17046-17051.

Sadowski, I., Ma, J., Triezenberg, S. and Ptashne, M. (1988). "GAL4-VP16 is an unusually potent transcriptional activator." Nature **335**(6190): 563-564.

Salvemini, M., Mauro, U., Lombardo, F., Milano, A., Zazzaro, V., Arca, B., Polito, L. C. and Saccone, G. (2011). "Genomic organization and splicing evolution of the *doublesex* gene, a *Drosophila* regulator of sexual differentiation, in the dengue and yellow fever mosquito *Aedes aegypti*." Bmc Evolutionary Biology **41-52**: 19.

Salvemini, M., Robertson, M., Aronson, B., Atkinson, P., Polito, L. C. and Saccone, G. (2009). "*Ceratitis capitata* transformer-2 gene is required to establish and maintain the autoregulation of *Cctra*, the master gene for female sex determination." International Journal of Developmental Biology **53**(1): 109-120.

Salz, H. K. (2011). "Sex determination in insects: a binary decision based on alternative splicing." Current Opinion in Genetics & Development **21**(4): 395-400.

Salz, H. K. and Erickson, J. W. (2010). "Sex determination in *Drosophila* The view from the top." Fly **4**(1): 60-70.

Salzberg, S. L. and Yorke, J. A. (2005). "Beware of mis-assembled genomes." Bioinformatics **21**(24): 4320-4321.

Sanchez, L. (2008). "Sex-determining mechanisms in insects." International Journal of Developmental Biology **52**(7): 837-856.

Sanchez, L., Gorfinkiel, N. and Guerrero, I. (2001). "Sex determination genes control the development of the *Drosophila* genital disc, modulating the response to Hedgehog, Wingless and Decapentaplegic signals." Development **128**(7): 1033-1043.

Sanchez, L. and Nothiger, R. (1982). "Clonal analysis of *sex-lethal*, a gene needed for female sexual development in *Drosophila melanogaster*." Wilhelm Rouxs Archives of Developmental Biology **191**(3): 211-214.

Sanders, H. R., Evans, A. M., Ross, L. S. and Gill, S. S. (2003). "Blood meal induces global changes in midgut gene expression in the disease vector, *Aedes aegypti*." Insect Biochemistry and Molecular Biology **33**(11): 1105-1122.

Sanford, J. R., Wang, X., Mort, M., VanDuyn, N., Cooper, D. N., Mooney, S. D., Edenberg, H. J. and Liu, Y. L. (2009). "Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts." Genome Research **19**(3): 381-394.

Sanger, F. and Coulson, A. R. (1975). "Rapid method for determining sequences in DNA by primed synthesis with DNA-polymerase." Journal of Molecular Biology **94**(3): 441-448.

Sanger, F., Nicklen, S. and Coulson, A. R. (1977). "DNA sequencing with chain-terminating inhibitors." Proceedings of the National Academy of Sciences of the United States of America **74**(12): 5463-5467.

Santolamazza, F., Calzetta, M., Etang, J., Barrese, E., Dia, I.*, et al.* (2008). "Distribution of knock-down resistance mutations in *Anopheles gambiae* molecular forms in west and west-central Africa." Malaria Journal **7**: 8.

Sarno, F., Ruiz, M. F., Eirin-Lopez, J. M., Perondini, A. L. P., Selivon, D. and Sanchez, L. (2009). "The gene transformer-2 of *Anastrepha* fruit flies (Diptera, *Tephritidae*) and its evolution in insects." Bmc Evolutionary Biology **10**: 140-154.

Scali, C., Catteruccia, F., Li, Q. X. and Crisanti, A. (2005). "Identification of sex-specific transcripts of the *Anopheles gambiae* doublesex gene." Journal of Experimental Biology **208**(19): 3701-3709.

Schetelig MF, M. A., Saccone G,Handler AM. (2011). "Male only progeny in *Anastrepha* suspensa by RNAi-induced sex reversion of chromosomal females." Insect Biochemistry and Molecular Biology.

Schmidt, R., Hediger, M., Roth, S., Nothiger, R. and Dubendorfer, A. (1997). "The Y-chromosomal and autosomal male-determining M factors of *Musca domestica* are equivalent." Genetics **147**(1): 271-280.

Schutt, C. and Nothiger, R. (2000). "Structure, function and evolution of sex-determining systems in Dipteran insects." Development **127**(4): 667-677.

Schwartz, T. S., Tae, H., Yang, Y., Mockaitis, K., Van Hemert, J. L., Proulx, S. R., Choi, J. H. and Bronikowski, A. M. (2010). "A garter snake transcriptome: pyrosequencing, *de novo* assembly, and sex-specific differences." Bmc Genomics **11**: 694-710.

Scott, M. J., Heinrich, J. C. and Li, X. L. (2004). "Progress towards the development of a transgenic strain of the Australian sheep blowfly (Lucilia cuprina) suitable for a male-only sterile release program." Insect Biochemistry and Molecular Biology **34**(2): 185-192.

Seawright, J. A., Kaiser, P. E., Dame, D. A. and Lofgren, C. S. (1978). "Genetic method for preferential elimination of females of *Anopheles albimanus*." Science **200**(4347): 1303-1304.

Semotok, J. L., Cooperstock, R. L., Pinder, B. D., Vari, H. K., Lipshitz, H. D. and Smibert, C. A. (2005). "Smaug recruits the CCR4/POP2/NOT deadenylase complex to trigger maternal transcript localization in the early drosophila embryo." Current Biology **15**(4): 284-294.

Sharma, V., Razdan, R. and Ansari, M. (1978). "*Anopheles stephensi*: effect of gamma-radiation and chemosterilants on the fertility and fitness of males for sterile male releases." Journal of Economic Entomology **71**(3): 449-450.

Shearman, D. C. A. (2002). "The evolution of sex determination systems in Dipteran insects other than *Drosophila*." Genetica **116**(1): 25-43.

Shearman, D. C. A. and Frommer, M. (1998). "The *Bactrocera tryoni* homologue of the *Drosophila melanogaste*r sex-determination gene *doublesex*." Insect Molecular Biology **7**(4): 355-366.

Shirangi, T. R., Dufour, H. D., Williams, T. M. and Carroll, S. B. (2009). "Rapid Evolution of Sex Pheromone-Producing Enzyme Expression in *Drosophila*." Plos Biology **7**(8): e1000168.

Sievert, V., Kuhn, S., Paululat, A. and Traut, W. (2000). "Sequence conservation and expression of the Sex-lethal homologue in the fly *Megaselia scalaris*." Genome **43**(2): 382-390.

Sievert, V., Kuhn, S. and Traut, W. (1997). "Expression of the sex determining cascade genes *Sex-lethal* and *doublesex* in the phorid fly *Megaselia scalaris*." Genome **40**(2): 211-214.

Sinkins, S. P. and Gould, F. (2006). "Gene drive systems for insect disease vectors." Nature Reviews Genetics **7**(6): 427-435.

Sleator, R. D. (2010). "An overview of the current status of eukaryote gene prediction strategies." Gene **461**(1-2): 1-4.

Smibert, C. A., Lie, Y. S., Shillinglaw, W., Henzel, W. J. and Macdonald, P. M. (1999). "Smaug, a novel and conserved protein, contributes to repression of *nanos* mRNA translation in vitro." Rna-a Publication of the Rna Society **5**(12): 1535-1547.

Song, H. J., Billeter, J. C., Reynaud, E., Carlo, T., Spana, E. P., Perrimon, N., Goodwin, S. F., Baker, B. S. and Taylor, B. J. (2002). "The fruitless gene is required for the proper formation of axonal tracts in the embryonic central nervous system of *Drosophila*." Genetics **162**(4): 1703-1724.

Song, H. J. and Taylor, B. J. (2003). "Fruitless gene is required to maintain neuronal identity in evenskipped-expressing neurons in the embryonic CNS of *Drosophila*." Journal of Neurobiology **55**(2): 115-133.

Sorber, K., Dimon, M. T. and DeRisi, J. L. (2011). "RNA-Seq analysis of splicing in *Plasmodium falciparum* uncovers new splice junctions, alternative splicing and splicing of antisense transcripts." Nucleic Acids Research **39**(9): 3820-3835.

Sosnowski, B. A., Belote, J. M. and McKeown, M. (1989). "Sex-specific alternative splicing of RNA from the *transformer* gene results from sequence-dependent splice site blockage." Cell **58**(3): 449-459.

Staab, S., Heller, A. and SteinmannZwicky, M. (1996). "Somatic sex-determining signals act on XX germ cells in *Drosophila* embryos." Development **122**(12): 4065-4071.

Stein, L. D. (2010). "The case for cloud computing in genome informatics." Genome Biology **11**(5): 7.

Steinemann, M. and Steinemann, S. (1992). "Degenerating Y-chromosome of *Drosophila miranda* - A trap for retrotransposons." Proceedings of the National Academy of Sciences of the United States of America **89**(16): 7591-7595.

Stevenson, B. J., Bibby, J., Pignatelli, P., Muangnoicharoen, S., O'Neill, P. M.*, et al.* (2011). "Cytochrome P450 6M2 from the malaria vector *Anopheles gambiae* metabolizes pyrethroids: Sequential metabolism of deltamethrin revealed." Insect Biochemistry and Molecular Biology **41**(7): 492-502.

Strauss, J. H. and Strauss, E. G. (1994). "The alphaviruses - gene-expression, replication, and evolution." Microbiological Reviews **58**(3): 491-562.

Sturtevant, A. H. (1945). "A gene in *Drosophila Melanogaster* that transforms females into males " Genetics **30**(3): 297-299.

Su, Z. Q., Li, Z. G., Chen, T., Li, Q. Z., Fang, H.*, et al.* (2011). "Comparing Next-Generation Sequencing and Microarray Technologies in a Toxicological Study of the Effects of Aristolochic Acid on Rat Kidneys." Chemical Research in Toxicology **24**(9): 1486-1493.

Sugarbaker, D. J., Richards, W. G., Gordon, G. J., Dong, L., De Rienzo, A.*, et al.* (2008). "Transcriptome sequencing of malignant pleural mesothelioma tumors." Proceedings of the National Academy of Sciences of the United States of America **105**(9): 3521-3526.

Sutherland, D., Samakovlis, C. and Krasnow, M. A. (1996). "*Branchless* encodes a *Drosophila* FGF homolog that controls tracheal cell migration and the pattern of branching." Cell **87**(6): 1091-1101.

Swarbreck, S. M., Lindquist, E. A., Ackerly, D. D. and Andersen, G. L. (2011). "Analysis of Leaf and Root Transcriptomes of Soil-Grown *Avena barbata* Plants." Plant and Cell Physiology **52**(2): 317-332.

Swindell, S. R. and Plasterer, T. N. (1997). SEQMAN: Contig assembly. Methods in Molecular Biology; Sequence data analysis guidebook. S. R. Swindell, Humana Press Inc. {a}, Suite 808, 999 Riverview Drive, Totowa, New Jersey 07512, USA. **70:** 75-89.

Tadros, W., Goldman, A. L., Babak, T., Menzies, F., Vardy, L.*, et al.* (2007). "SMAUG is a major regulator of maternal mRNA destabilization in *Drosophila* and its translation is activated by the PAN GU kinase." Developmental Cell **12**(1): 143-155.

Takahashi, M., Altschmied, L. and Hillen, W. (1986). "Kinetic and equilibrium characterization of the Tet repressor tetracycline complex by fluorescence measurements - evidence for divalent metal-ion requirement and energy-transfer." Journal of Molecular Biology **187**(3): 341-348.

Tawfik, D. S. and Griffiths, A. D. (1998). "Man-made cell-like compartments for molecular evolution." Nature Biotechnology **16**(7): 652-656.

ten Bosch, J. R., Benavides, J. A. and Cline, T. W. (2006). "The TAGteam DNA motif controls the timing of *Drosophila* pre-blastoderm transcription." Development **133**(10): 1967-1977.

ter Kuile, F. O., Terlouw, D. J., Phillips-Howard, P. A., Hawley, W. A., Friedman, J. F*., et al.* (2003). "Impact of permethrin-treated bed nets on malaria and all-cause morbidity in young children in an area of intense perennial malaria transmission in western Kenya: Cross-sectional survey." American Journal of Tropical Medicine and Hygiene **68**(4): 100-107.

Terenius, O., Marinotti, O., Sieglaff, D. and James, A. A. (2008). "Molecular Genetic Manipulation of Vector Mosquitoes." Cell Host & Microbe **4**(5): 417-423.

Thailayil, J., Magnusson, K., Godfray, H. C. J., Crisanti, A. and Catteruccia, F. (2011). "Spermless males elicit large-scale female responses to mating in the malaria mosquito *Anopheles gambiae*." Proceedings of the National Academy of Sciences of the United States of America **108**(33): 13677-13681.

Thomas, D. D., Donnelly, C. A., Wood, R. J. and Alphey, L. S. (2000). "Insect population control using a dominant, repressible, lethal genetic system." Science **287**(5462): 2474-2476.

Tomancak, P., Beaton, A., Weiszmann, R., Kwan, E., Shu, S*., et al.* (2002). "Systematic determination of patterns of gene expression during *Drosophila* embryogenesis." Genome Biology **3**(12): RESEARCH0088.

Torres, T. T., Metta, M., Ottenwalder, B. and Schlotterer, C. (2008). "Gene expression profiling by massively parallel sequencing." Genome Research **18**(1): 172-177.

Toung, J. M., Morley, M., Li, M. Y. and Cheung, V. G. (2011). "RNA-sequence analysis of human B-cells." Genome Research **21**(6): 991-998.

Townson, H., Nathan, M. B., Zaim, M., Guillet, P., Manga, L., Bos, R. and Kindhauser, M. (2005). "Exploiting the potential of vector control for disease prevention." Bulletin of the World Health Organization **83**(12): 942-947.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J. and Pachter, L. (2010). "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation." Nature Biotechnology **28**(5): 511-515.

Traut, W., Mimi, T., Ikeo, K. and Sahara, K. (2006). "Phylogeny of the sex-determining gene sex-lethal in insects." Genome **49**(3): 254-262.

Ulveling, D., Francastel, C. and Hube, F. (2011). "When one is better than two: RNA with dual functions." Biochimie **93**(4): 633-644.

Usui-Aoki, K., Ito, H., Ui-Tei, K., Takahashi, K., Lukacsovich, T*., et al.* (2000). "Formation of the male-specific muscle in female *Drosophila* by ectopic fruitless expression." Nature Cell Biology **2**(8): 500-506.

Valcarcel, J., R., S., Zamore, P., D. and Green, M., R. (1993). "The protein Sex-lethal antagonizes the splicing factor U2AF to regulate alternative splicing of *transformer* pre-mRNA." Nature **362**(6416): 171-175.

Vargas-Teran, M., Hursery, B. and Cunningham, E. (1994). "Eradication of the screwworm from Libya using the sterile insect technique." Parasitology Today **10**(3): 119-122.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J*., et al.* (2001). "The sequence of the human genome." Science **291**(5507): 1304-1351.

Vera, J. C., Wheat, C. W., Fescemyer, H. W., Frilander, M. J., Crawford, D. L., Hanski, I. and Marden, J. H. (2008). "Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing." Molecular Ecology **17**(7): 1636-1647.

Verhulst, E. C., Beukeboom, L. W. and van de Zande, L. (2010). "Maternal Control of Haplodiploid Sex Determination in the Wasp *Nasonia*." Science **328**(5978): 620-623.

Verhulst, E. C., van de Zande, L. and Beukeboom, L. W. (2010). "Insect sex determination: it all evolves around *transformer*." Curr Opin Genet Dev **20**(4): 376-83.

Vibranovski, M. D., Koerich, L. B. and Carvalho, A. B. (2008). "Two new Y-linked genes in *Drosophila melanogaster*." Genetics **179**(4): 2325-2327.

Vontas, J., Blass, C., Koutsos, A. C., David, J. P., Kafatos, F. C., Louis, C., Hemingway, J., Christophides, G. K. and Ranson, H. (2005). "Gene expression in insecticide resistant and susceptible *Anopheles gambiae* strains constitutively or after insecticide exposure." Insect Molecular Biology **14**(5): 509-521.

Vreysen, M., Saleh, K., Ali, M., Abdulla, A., Zhu, Z*., et al.* (2000). "*Glossina austeni* (Diptera: Glossinidae) eradicated on the island of Unguja, Zanzibar, using the sterile insect technique." Journal of Economic Entomology **93**(1): 123-135.

Vrontou, E., Nilsen, S. P., Demir, E., Kravitz, E. A. and Dickson, B. J. (2006). "*fruitless* regulates aggression and dominance in *Drosophila*." Nature Neuroscience **9**(12): 1469-1471.

Wang, C. and Lehmann, R. (1991). "*nanos* is the localized posterior determinant in Drosophila." Cell **66**(4): 637-647.

Wang, E. T., Sandberg, R., Luo, S. J., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P. and Burge, C. B. (2008). "Alternative isoform regulation in human tissue transcriptomes." Nature **456**(7221): 470-476.

Wang, L. K., Feng, Z. X., Wang, X., Wang, X. W. and Zhang, X. G. (2010). "DEGseq: an R package for identifying differentially expressed genes from RNA-seq data." Bioinformatics **26**(1): 136-138.

Wang, M.-H., Marinotti, O., Vardo-Zalik, A., Boparai, R. and Yan, G. (2010). "Genome-wide transcriptional analysis of genes associated with acute desiccation stress in *Anopheles gambiae*." Plos One **6**(10): e26011.

Wang, Z., Gerstein, M. and Snyder, M. (2009). "RNA-Seq: a revolutionary tool for transcriptomics." Nature Reviews Genetics **10**(1): 57-63.

Warr, E., Aguilar, R., Dong, Y. M., Mahairaki, V. and Dimopoulos, G. (2007). "Spatial and sex-specific dissection of the *Anopheles gambiae* midgut transcriptome." Bmc Genomics **8**: 37-48.

Waterbury, J. A., Jackson, L. L. and Schedl, P. (1999). "Analysis of the doublesex female protein in *Drosophila melanogaster*: Role in sexual differentiation and behavior and dependence on intersex." Genetics **152**(4): 1653-1667.

Wawersik, M., Milutinovich, A., Casper, A. L., Matunis, E., Williams, B. and Van Doren, M. (2005). "Somatic control of germline sexual development is mediated by the JAK/STAT pathway." Nature **436**(7050): 563-567.

Weill, M., Malcolm, C., Chandre, F., Mogensen, K., Berthomieu, A., Marquine, M. and Raymond, M. (2004). "The unique mutation in ace-1 giving high insecticide resistance is easily detectable in mosquito vectors." Insect Molecular Biology **13**(1): 1-7.

Wetterbom, A., Ameur, A., Feuk, L., Gyllensten, U. and Cavelier, L. (2010). "Identification of novel exons and transcribed regions by chimpanzee transcriptome sequencing." Genome Biology **11**(7): 16.

WHO (1997). "Lymphatic Filariasis: Reasons for Hope." Edited by Dzenowagis J.

WHO (2003). "Space spray application of insecticides for vector and public health pest control: A practitioner's guide."

WHO (2006). "Pesticides and their application for the control of vectors and pests of public health importance " WHO Pesticide evaluation scheme (WHOPES)World Health Organisation, Department of Control of Neglected Tropical Diseases

WHO (2011). "World Malaria Report."

Wicker, T., Schlagenhauf, E., Graner, A., Close, T. J., Keller, B. and Stein, N. (2006). "454 sequencing put to the test using the complex genome of barley." Bmc Genomics **7**: 11.

Wieschaus, E. and Sweeton, D. (1988). "Requirements for X-linked zygotic gene activity during cellularization of early Drosophila embryos." Development **104**(3): 483-493.

Wilkins, A. S. (1995). "Moving up the hierarchy - a hypothesis on the evolution of a genetic sex determination pathway." Bioessays **17**(1): 71-77.

Willhoeft, U. and Franz, G. (1996). "Identification of the sex-determining region of the *Ceratitis capitata* Y chromosome by deletion mapping." Genetics **144**(2): 737-745.

Williams, L. M., Ma, X., Boyko, A. R., Bustamante, C. D. and Oleksiak, M. F. (2010). "SNP identification, verification, and utility for population genetics in a non-model genus." BMC genetics **11**: 32-46.

Williams, M., Woodhall, J., Corbet, P. and Gillett, J. (1965). "O'nyong-nyong fever: an epidemic virus disease in East Africa. VIII. Virus isolations from *Anopheles* mosquitoes." Transactions Royal Society Tropical Medicine and Hygiene **59**: 300-306.

Williams, M. C., Woodall, J. P. and Porterfield, J. S. (1962). "O'nyong-nyong fever; an epidemic virus disease in East Africa. V Human antibody studies by plaque inhibition and other serological tests." Transactions of the Royal Society of Tropical Medicine and Hygiene **56**: 166-72.

Williams, T. M., Selegue, J. E., Werner, T., Gompel, N., Kopp, A. and Carroll, S. B. (2008). "The regulation and evolution of a genetic switch controlling sexually dimorphic traits in *Drosophila*." Cell **134**(4): 610-623.

Windbichler, N., Papathanos, P. A. and Crisanti, A. (2008). "Targeting the X Chromosome during Spermatogenesis Induces Y Chromosome Transmission Ratio Distortion and Early Dominant Embryo Lethality in *Anopheles gambiae*." Plos Genetics **4**(12): 9.

Winnebeck, E. C., Millar, C. D. and Warman, G. R. (2010). "Why does insect RNA look degraded?" Journal of Insect Science **10**: 159-166.

Wu, J. Q., Du, J., Rozowsky, J., Zhang, Z., Urban, A. E., Euskirchen, G., Weissman, S., Gerstein, M. and Snyder, M. (2008). "Systematic analysis of transcribed loci in ENCODE regions using RACE sequencing reveals extensive transcription in the human genome." Genome Biology **9**(1): R3.

Wu, T. D. and Watanabe, C. K. (2005). "GMAP: a genomic mapping and alignment program for mRNA and EST sequences." Bioinformatics **21**(9): 1859-1875.

Wurm, Y., Wang, J., Riba-Grognuz, O., Corona, M., Nygaard, S*., et al.* (2011). "The genome of the fire ant *Solenopsis invicta*." Proceedings of the National Academy of Sciences of the United States of America **108**(14): 5679-5684.

Xu, X. J., Dong, Y. M., Abraham, E. G., Kocan, A., Srinivasan, P*., et al.* (2005). "Transcriptome analysis of *Anopheles stephensi-Plasinodium berghei* interactions." Molecular and Biochemical Parasitology **142**(1): 76-87.

Young, R., Marques, A., Tibbit, C., Haerty, W., Bassett, A., Liu, J. and Ponting, C. (2012). "Identification and properties of 1,119 candidate lincrna loci in the *Drosophila melanogaster* genome." Genome Biology and Evolution.

Zaim, M., Aitio, A. and Nakashima, N. (2000). "Safety of pyrethroid-treated mosquito nets." Medical and Veterinary Entomology **14**(1): 1-5.

Zhang, K., Martiny, A. C., Reppas, N. B., Barry, K. W., Malek, J., Chisholm, S. W. and Church, G. M. (2006). "Sequencing genomes from single cells by polymerase cloning." Nature Biotechnology **24**(6): 680-686.

Zhang, Z. P., Wang, Y. L., Wang, S. H., Liu, J. T., Warren, W., Mitreva, M. and Walter, R. B. (2011). "Transcriptome Analysis of Female and Male *Xiphophorus maculatus* Jp 163 A." Plos One **6**(4): 6.

Appendix I

A

```
AGAP009810      MTSDKRTVYVGGLSEEVTEKLITDAFIPFGDLVDIQMPIDYESQKHRGFAFIEFENAEDA 60
DmTra2          --MDREPLSSG----------------------RLHCSARYKHKRSASSS----SAGTTS 32
                   *:..:  *                      :: .  *: ::  . :    . .  :

AGAP009810      AAAVDNMNDSELCGRTIRVNTAKPQRIKEGSNRPVWADDNWLQKHAGATLKNDGENGAEQ 120
DmTra2          SGHKDRRSDYDYCGSRRHQRSSSRRRSRSRSSSESPPPEPRHRSGRSSRDRERMHKSREH 92
                :.  *. .* : **   : .::. :* :. *.    . :   :.  .:  ::  .:. *:

AGAP009810      PMDTGTDEVQPKEPKAAEEKKRNPQVFFDIRIGNSDVGRIIMVLRADVVPKTAENFRALC 180
DmTra2          PQASRCIGVFGLNTNTSQHKVR------ELFNKYGPIERIQMVIDAQTQRSRGFCFIYFE 146
                *  :    *   :.:::::.* *      ::    . : ** **: *:.  . .  *  :

AGAP009810      TGEQGFGYKGSTFHRIIPEFMCQGGDFTANNGSGGKSIYGKKFADENFILKHTGFGVLSM 240
DmTra2          KLSDARAAKDSCSG-IEVDGRRIRVDFSITQR-AHTPTPGVYLGRQPRGKAPRSFSPRRG 204
                . .:. . *.*    *  :      **: .:  . ..  *  :. :       .*.

AGAP009810      ANSGPNTNGSQFFICTEKTDWLDGKHVVFGNVISGADVVRKMERCGSKGGRVQQKVTIVA 300
DmTra2          RRVYHDRSASPYDNYRDRYDYRNDR---YDRNLRRSPSRNRYTRNRSYSRSRSPQLRRTS 261
                 .   : ..* :    :: *: :.:   :.. :  :   .:  *  * .   . ::  .:

AGAP009810      CGEVKG 306
DmTra2          SRY--- 264
                    .
```

B

```
DmTra2          MDREPLSSGRLHCSARYKHKRSASSSSAGTTSSGHKDRRSDYDYCGSRRHQRSSSRRRSR 60
AGAP00097       --MNPMTN------MKNVTKLSEQDLKLGGKSSWHDQYKS-------------------- 32
                   :*::.       :   * * .. . * .** *.: :*

DmTra2          SRSSSESPPPEPRHRSGRSSRDRERMHKSREHPQASRCIGVFGLNTNTSQHKVRELFNKY 120
AGAP00097       ----------------------------------SAWIFVGGLPYDLTEGDVLCVFSQY 57
                                                 *  * * **  : :: .*  :*.:*

DmTra2          GPIERIQMVIDAQTQRSRGFCFIYFEKLSDARAAKDSCSGIEVDGRRIRVDFSITQRAHT 180
AGAP00097       GEIVNVNLVRDKATGKSKGFAFICYEDQRSTVLAVDNLNGIKLVGKTLRVDHVQDYRPPK 117
                * * .::: *  * :*:**.** :*.  .:  * *. .**:: *: :***.    *. .

DmTra2          PTPGVYLGRQPRGKAPRSFSPRRGRRVYHDRSASPYDNYRDRYDYRNDRYDRNLRRSPSR 240
AGAP00097       ET---------------DKTDDETRQLYMEGCAPKAEHPRP-----------EKQKDPTV 151
                 *               . :  . *::* : .*.  :: *           : ::.*:

DmTra2          NRYTRNRSYSRSRSPQLRRTSSRY 264
AGAP00097       KSEKRHKSIKKE------------ 163
                :  .*::* .:.
```

C

```
DmTra2          MDREPLSSGRLHCSAR--YKHKRSASSSSAGTTSSGHKDRRSDYDYCGSRRHQRSSSRRR 58
AGAP006755      MTQYLPSNLLALFAPRDPIPYLPPPDKLPHEKKSQGYLGVGAFLDCFEDPKDTPPPTRVE 60
                * :   *.     :.*    :  .... .  ..*.*: .  :  *   . :.  ..:* .

DmTra2          SRSRSSSESPPPEPRHRSGRSSRDRERMHKS---REHPQASRCIGVFGLNTNTSQHKVRE 115
AGAP006755      TREERVERRRRERAEQVAYKLEREIATWDPNQLLEATEDPFKTLFVARINYDTSESKLRR 120
                :*..  ..    ...: : : .*:    . .   .   :. : : *  :* :**: *:*.

DmTra2          LFNKYGPIERIQMVIDAQTQRSRGFCFIYFEKLSDARAAKDSCSGIEVDGRRIRVDFSIT 175
AGAP006755      EFEIYGNIKKIIMINDKDTGKPRGYAFIEYEHERDMHAAYKHADGKKIDGKRVLVDVERA 180
                 *: ** *::* *: * :* :.**:.** :*:  * :** . ..* ::**:*: **.. :

DmTra2          QRAHTPTP---GVYLGRQPRG------KAPRSFSPRRGRRVYHDRSASPYDNYRDRYDYR 226
AGAP006755      RTVKGWLPRRLGGGLGGTRRGGPDVNIKHSGREDNERERERYRLERERERDRDRRERDRPE 240
                : .:   *   *  **   **      * .   . .* *. *: .      : *:*   .

DmTra2          NDRYDRNLRRSPSRNRYTRNRSYSRSRSPQLR--------------------------- 258
AGAP006755      NNRFERRRSRSRDRDRRHRSRSKDRSRRRRSREREVDDRGDRWEKERDRERDRSRDRDKD 300
                *:*::*.  ** .*:*  *.** .***  : *

DmTra2          ----RTSSRY---------------------------------------------------- 264
AGAP006755      RKRKRSRSRDRDRAEKSRHNKRGDRDRGERDRGERGGGGGGGGGASGNGGAERKERKPEF 360
                    *: **

DmTra2          ------------------------------------------------------------
AGAP006755      REGEIKIKEEPIDDYPDYSAQYSQYQSQVKYEDGEEEKYRPNEPPPHLRADNGRGYVTGN 420


DmTra2          --------------
AGAP006755      GDDDDGDYDDGQGY 434
```

D

```
DmTra2          ---------------------------------MDREPLSSGRLHCS----------AR 16
AGAP008433      SNSHRSRSRERERERERDRDRERDRERERERDRERDRDGEKASKKHRSGSPGRNRDKERR 60
                                                 **:  .:.: * *           *

DmTra2          YKHKRSASSSSAGTTSSGHK------------------DRRSDYDYCGSRRHQRS-SSRR 57
AGAP008433      RSKDRSKSRSPARRDRSKDRSKEKDRGKSDHHRREVVVEKRRSRDRVDHRRRSRERDYRR 120
                 .:.** * *.*    * .:                  ::* . *  . **:.*. . **

DmTra2          RSRSRSSSE-------SPPPEPRHRSGRSS----RDRERMHKSREHPQASR--------- 97
AGAP008433      RSRSRDGGRGMGRGRRSMSPKPYRGRGRGGSGYYRDRSPLEEMSQEDRDARTVFCMQLSQ 180
                *****....       * .*:* :  **..    ***. :.:  :. : :*

DmTra2          ----------------------------------------------CIGVFG------- 103
AGAP008433      RIHARDLEEFFSSVGKVRDVRLITCNKTKRFKGIAYIEFKDPESVALALGLSGQKLLGIP 240
                                                           .:*: *

DmTra2          ---------------------------------LNTNTSQHKVRELFNKYGPIERIQM 128
AGAP008433      ISVQHTQAEKNRMASQPPVAPPKNPSGPMRLYVGSLHFNITEDMLNGIFEPFGKIDNIQL 300
                                                 *: * ::. :. :*: :* *:.**:

DmTra2          VIDAQTQRSRGFCFIYFEKLSDARAAKDSCSGIEVDGRRIRVD-----FSITQRAHTPT- 182
AGAP008433      IMDADTGRSKGYGFITFHNADDAKKALEQLNGFELAGRPMKVGNVTERLDVTTHASLDTD 360
                ::**:* **:*: ** *.: .**: * :. .*:*: ** ::*.     :.:* :*   *

DmTra2          ----PGVYLGRQ-----------------PRGKAPRSFSPRRGRRVYHDRSASP------ 215
AGAP008433      EMDRSGIELGATGRLQLMFKLAEGAGLAVPRAAADALLATAPQPVPQQPIMQSPPIATQC 420
                    .*: **                   **. *   ::.      :    **

DmTra2          --YDNYRDRYDYRNDRYDRNLR-------------------RSPSRNRYTRNRSYSRSR 253
AGAP008433      FLLSNMFDPATETNPNWDLEIQDDVIEECNKHGGVQHVYVDKQSPSGNVYVKCPSIATAV 480
                   .*  *     * .:* :::                   :*** * *.:  * : :
```

```
DmTra2          SPQLRRTSSRY---------------------------------- 264
AGAP008433      LAVNALHGRWFAGRVIGAAYVPLINYYNLFPEAAQSVTMLQPKRSG 526
                 .     .  :
```

E

```
DmTra2          ------------------------------------------------------------
AGAP013051      MGDILWSPAMNEPSTSQPQLPRFTENDQLFDELLENIVYVGNLPKQIMLTDLIELFKYAG 60

DmTra2          ------------------------------------------------------------
AGAP013051      RIERVAWYGEHRTDINTKVAFIRFRHSRHAKEAAKWDRIRYQDSILIVMQIFKDQWFDMS 120

DmTra2          -----------------------------------------------------MDREPLS 7
AGAP013051      VSIMVRNIRDDTTDWQLYEAFRRFGKIYGILIPTHGTAYVGFYYEAETQCALEMNNNMFN 180
                                                                    *:.: :.

DmTra2          SGRLHCSARYKHKRSASSSSAGTTSSGHKDRR----------------------SDYDY 44
AGAP013051      GNRMRVEMLRRNLPLQQIDIFDITKNEVRLLRDMLLEVDEKLEQFYSAYDEAFNYSSKDY 240
                ..*:: .   ::    . .  . *..  :  *                       *. **

DmTra2          CGSRRHQRSSSRRRSRSRSSSESPPPEPRHRSGRSSRDRERMHKSREHPQASRCIGVFGL 104
AGAP013051      RRWKRKRRRVTPLLSDSSSSSESSSSSVSEISNRSATEVRRILCG--SNEENRCLGIFGM 298
                   :*::*  :   * * *****....  . *.**: : .*:  .    : .**:*:**:

DmTra2          NTNTSQHKVRELFNKYGPIERIQMVIDAQTQRSRGFCFIYFEKLSDARAAKDSCSGIEVD 164
AGAP013051      NPDTTEKTLMKLFSRYGHVKDIKLIYDGKTNVSRGYSFIYFKHASDARRAQRKLNGTMLE 358
                *.:*::::.: :**.:** :: *::: *.:*: ***:.****:: **** *: . .*  ::

DmTra2          GRRIRVDFSITQRAHTPTPGVYLGRQPRGKAPRSFSPRRGRRVYHDRSASPYDNYRDRYD 224
AGAP013051      GRKVRVDFSRSK-PHEPR------ADRRKRVSPTVASTSADRCCHRRHGATHH------- 404
                **::***** :: .* *        : * :.. :.:.  . *  * * .:.:.

DmTra2          YRNDRYDRNLRRSPSRNRYTRNRSYSRSRSPQLRRTSSRY 264
AGAP013051      ---------------KKRKKRRVCYSSSDSE--------- 420
                              ::* .*. .** * *
```

F

```
BM620287        PTRPPHRTFSTCCELSCSKRILRKMTSSSYRSSDRRTSRRYDEDGDRHHRHRSRSVDRCH 60
DmTRA2          ----MDREPLSSGRLHCSARYKHKRSASSSSAGTTSSG----------HKDRRSDYDYCG 46
                     .*   :. .* ** *  :* ::**  :.   :.          *:.*  . * *

BM620287        RSRRSRSDSRDRYRRNRSASRSPRGRSRHRKSSTAPNASRSRIRVDSPEPSRCLGVFGLS 120
DmTRA2          SRRHQRSSSR-RRSRSRSSSESPPPEPRHRSGRSSRDRERMHKSREHPQASRCIGVFGLN 105
                  *:.**.** *  *.**:*.**  ..***.. :: : .* :   : *:.***:*****.

BM620287        VYTTEPYLNDIFCHFGTVEKSVVIYDAKTRLSRGFGFVYFKSQAEASIARANCNGLQIHG 180
DmTRA2          TNTSQHKVRELFNKYGPIERIQMVIDAQTQRSRGFCFIYFEKLSDARAAKDSCSGIEVDG 165
                . *::  :.::* ::*.:*:  :: **:*: **** *:**:. ::*  *: .*.*:::.*

BM620287        RRIRVDYSITDQPHPPTPG--------------------RVHGPKAVQPLSVPAEPITL 219
DmTRA2          RRIRVDFSITQRAHTPTPGVYLGRQPRGKAPRSFSPRRGRRVYHDRSASPYDNYRDRYDY 225
                ******:***::.*.****                    **:  ::..* .   :

BM620287        R------------------------------------ 220
DmTRA2          RNDRYDRNLRRSPSRNRYTRNRSYSRSRSPQLRRTSSRY 264
                *
```

Figure 1: Amino acid alignment of significant hits to *D. melanogaster* TRA. The amino acid sequences of top hit *An. gambiae* proteins and an EST were aligned to known *tra2* sequences via ClustalW2 using the default settings. Asterisk (*) indicates conserved amino acids between all species. DmTRA2 *D. melanogaster* (AAA62771). A six-frame translation of the An.

gambiae EST BM620287 was performed and each frame aligned to DmTRA2, the alignment with the greatest similarity is shown.

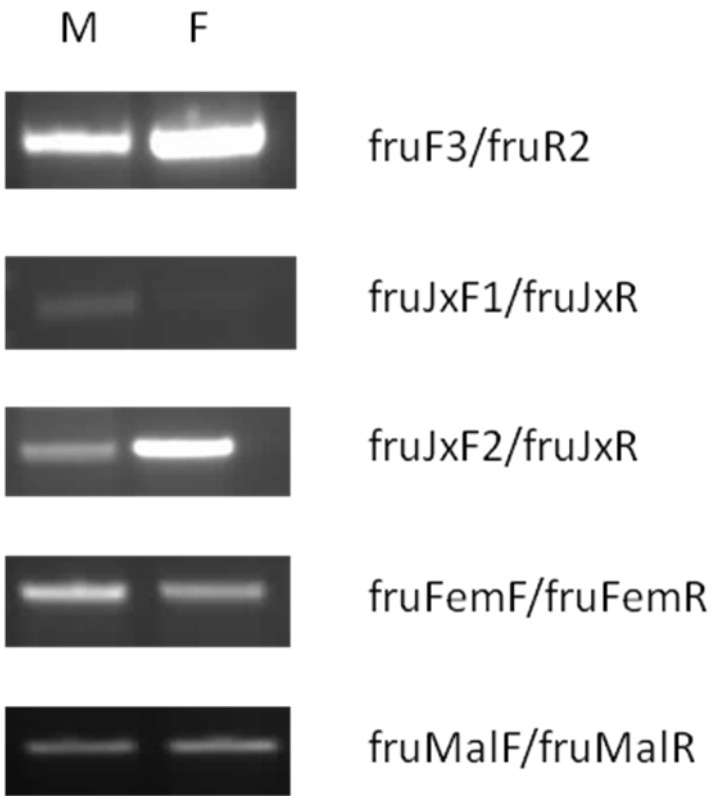


Figure 2: RT-PCR amplification of putative sex-specific fruitless transcripts by RT-PCR. Each primer pair was used for RT-PCR analysis amplifying from embryonic male and female RNA (20 hr AEL).

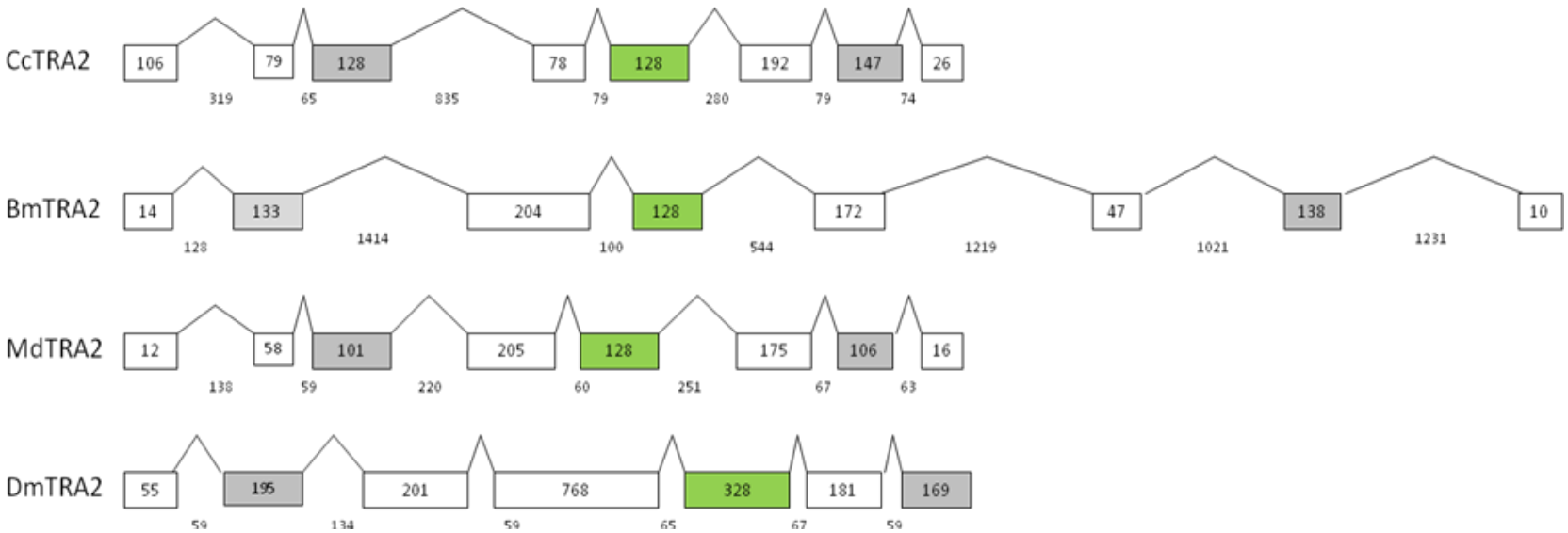


Figure 3: The genomic organisation of four *tra2* homologues. The exon/intron structure of four tra2 homologues was deduced following BLASTN searches against genomic databases. Boxes represent exons, diagonal lines introns and numbers the size of the feature in base pairs. Grey boxes indicate the position of SR rich regions and green the presence of a RNA recognition motif (RRM). SR rich regions flank the RRM, with a single interrupting exon at both the 5 and 3 prime sides. Conservation of this structure may aid identification of further *An. gambiae* TRA2 candidate homologues. MdTRA2, *M. domestica* (NM_001126233), CcTRA2 *C. capitata* (EU437408), BmTRA2 *B. mori* (NM_001126234), DmTRA2 *D. melanogaster* (NM_136602).

Table 1: Example of *An. gambiae* AgamP3.6 GTF file. Shown in the first column is the name of the reference sequence, in this instance chromosome2L. The GTF file was constructed by VectorBase and thus the source in column two is designated as such. The annotated sequences are described based upon either gene, mRNA, exon, coding sequence (CDS) and three and five prime untranslated regions (UTRs). The fourth and fifth columns describe the location of the described sequence relative to the genome with start and end positions respectively. In the final column, the name of the annotated transcript is shown.

Reference Sequence	Source	Description	Start position	End position	Transcript ID
chr2L	VectorBase	gene	157348	186936	ID=AGAP004677;
chr2L	VectorBase	mRNA	157348	186936	ID=AGAP004677
chr2L	VectorBase	exon	186860	186936	ID=E013802B;Parent=AGAP004677-RB;
chr2L	VectorBase	exon	159192	159366	ID=E013803B;Parent=AGAP004677-RB;
chr2L	VectorBase	exon	157679	158297	ID=E013804B;Parent=AGAP004677-RB;
chr2L	VectorBase	exon	157348	157623	ID=E013805B;Parent=AGAP004677-RB;
chr2L	VectorBase	five_prime_utr	186860	186936	Parent=AGAP004677-RB;
chr2L	VectorBase	five_prime_utr	159357	159366	Parent=AGAP004677-RB
chr2L	VectorBase	CDS	159192	159356	Parent=AGAP004677-RB;
chr2L	VectorBase	CDS	157679	158297	Parent=AGAP004677-RB;
chr2L	VectorBase	CDS	157496	157623	Parent=AGAP004677-RB;
chr2L	VectorBase	three_prime_utr	157348	157495	Parent=AGAP004677-RB

Table 2: Example of a SAM alignment file. Shown is an example of cleaned reads mapped to the AgamP3.6 transcriptome. In the first column a unique identifier of the sequence being aligned to the reference is shown. The bitwise flag is a bitwise flag giving information regarding the alignment of the transcript. In our study, the values of 16 (predicted forward orientation) and 0 (predicted reverse orientation) were considered. The third column indicates to which reference sequence the read has aligned to and the fifth the first position to which the read aligns. A value of 255 in the mapping quality column indicates that no quality information was available. The CIGAR string gives detailed information about the alignment. Values indicate the number of bases that meet the letter parameter following the number. M = alignment match; I = insertion to the reference; D = deletion to the reference; N = skipped region to the reference. The following two columns, RNext and PNext refer to paired-end reads and these were not studies in this thesis. A value is absent from the TLen column as this is only used if all reads map to a single reference sequence. A truncated version of the aligned sequence is shown in column 10, the full length sequence is normally present. In the final column, quality scores are not presented as FASTA files and not FASTQ were used for alignment.

Sequence Name	Bitwise Flag	Reference Name	Position	Mapping Quality	CIGAR String	RNext	PNext	TLen	Sequence	Quality Scores
GLONWJ301DCXW6	16	AGAP000002-RA	57	255	161M1I169M21S	*	0	0	TGGCCTCATTTACGCAGGCG	*
GLONWJ301EV7VQ	0	AGAP000002-RA	122	255	6S263M	*	0	0	CATGGGATATAATAAAGTG	*
GLONWJ301CIQ5V	0	AGAP000002-RA	1268	255	257M1D175M	*	0	0	GACTGCAACTGCGCCAGGA	*
GLONWJ301D50HO	0	AGAP000002-RA	1269	255	26M1I275M1I44M1I15M1I12M	*	0	0	ACTGCAACTGCGCCAGGAG	*
GLONWJ301ENCQG	0	AGAP000002-RA	2085	255	163M1I31M	*	0	0	AGCCTCCGTTCATGCTGCCA	*
GLONWJ301D62F8	0	AGAP000002-RA	2505	255	4S407M	*	0	0	CTATGGGGCAGATTGGGTTG	*
GMNRJ6W01BM4T7	0	AGAP000002-RA	2861	255	6S63M1I27M1I45M	*	0	0	ATGGGGATATCGTCAGCCG	*
GMNRJ6W01EF3JH	0	AGAP000002-RA	3081	255	2S329M1D48M1D78M	*	0	0	ATGGGTGATGAAGAAGAG	*

Table 3: Example of a BED alignment file. Shown is an example of new transcribed regions (NTR) mapped to the AgamP3 genome sequence. The first columns give information of where the NTR sequences aligns to. Columns two and three indicate the genomic location of the NTR within that reference. Column five gives the unique identifier of the aligned sequences and a predicted orientation is shown in the last column.

Chromosome	Start position	End position	Sequence name	Predicted orientation
chr2L	26906	27360	singletons_c5614	-
chr2L	98956	99396	GMNRJ6W01E0SDN	-
chr2L	99954	100380	GLONWJ301CF4O7	+
chr2L	319328	319671	GLONWJ301AMN1H	+
chr2L	410590	410782	GLONWJ301EWAA2	+
chr2L	410590	410854	GLONWJ301EG468	+
chr2L	410590	410949	GLONWJ301EBRHR	+
chr2L	543270	543762	singletons_c2230	+
chr2L	575674	576010	GLONWJ301DKO2O	+

Table 4: Example of 454RefStatus.txt file. The 454RefStatus.txt file is part of the Newblerv2.3 output, detailing read count data following alignment to a reference. Shown is alignment of reads to the AgamP3.6 transcriptome. In the first column the AgamP3.6 reference transcript is shown and the number of aligned reads in the second column. The table also provides information regarding the number of aligned reads as a percentage of all aligned reads (column 3) and of total reads (column 4). In the last column, the percentage coverage of the reference transcript is shown.

Reference Accession	Num Unique Matching Reads	Pct of All Unique Matches	Pct of All Reads	Pct Coverage of Reference
AGAP007087-RA	5377	2.10%	1.30%	97.33%
AGAP011802-RA	4967	1.90%	1.20%	100.00%
AGAP012334-RA	3309	1.30%	0.80%	100.00%
AGAP012100-RA	3090	1.20%	0.70%	100.00%
AGAP010591-RA	2927	1.10%	0.70%	79.63%
AGAP007406-RA	2120	0.80%	0.50%	100.00%
AGAP010257-RA	1731	0.70%	0.40%	99.81%
AGAP010461-RA	1611	0.60%	0.40%	100.00%
AGAP000278-RA	1565	0.60%	0.40%	99.35%

Perl Scripts

#insertion.pl

```
#Read in a sam alignment file and return isotigs that contain an insertion to the
#reference sequence, identified via the CIGAR string.
use strict;
my$i;
my$file_in= "mid2mid3isotigs.sam";
open(INPUT,$file_in)|| die "Cannot open $file_in\n";;
while(my $line = <INPUT>){
            chomp($line);
      my @temp = split(/\t/,$line);
                  my$query=$temp[0];
                  my$strand=$temp[1];
                  my$chr=$temp[2];
                  my$co_ord=$temp[3];
                  my$map_quality=$temp[4];
                  my$cigar=$temp[5];
        my @temp = $cigar;
                        my @values = (50..250);
                        my$value;
                        my$counter=1;
                        foreach $value (@values){
                                    if (grepis   $cigar){
                                                print "score $value $query $cigar\n";

                                    }
                                    }
                                    }
```

#transcript.pl

```
#Perl script to identify contigs from a blast m -8 format that show extension to the
#currently annotated AgamP3.6 transcriptome.
use strict;
open(INPUT,"../mid2_mid3_combined/454AllContigs_sing_contigs_sing_raw.fna_100_blas
tn_transcripts");
my$threshold=<STDIN>;
while(my $line = <INPUT>){
	chomp($line);
	my @temp = split(' ',$line);
	my$query=$temp[0];
	my$sbjct=$temp[1];
	my$q_start=$temp[6];
	my$q_end=$temp[7];
	my$s_start=$temp[8];
	my$s_end=$temp[9];

		if ($q_start>$threshold & $s_start==1) {
			print "Contiig with a extension longer than $threshold = $query\n";
			}

				}
```

#Length.pl

```
#!/usr/bin/perl
#Script courtesy of Anton Tsyganov-Bodounov
#Perl script to return lengths of fasta sequence files
use warnings;
open FASTAFILE, "mid3_out.unpadded.fasta" or die "Could not open file: $!";
%dna =();
$seqid ='';

while (<FASTAFILE>){
	chomp;
	if (/^>(\S+)\s*.*$/) { #found seqid in line
		$seqid = $1;
	}else {
		$dna{$seqid} .= $_;
	}
}
close FASTAFILE;
foreach $seqid (keys %dna) {
	my$length = length $dna{$seqid};

  print "$seqid $length\n";


}

#@ids = keys %dna;
#@seq = values %dna;
#print "@ids\n@seq";
```

AS.pl

```
use strict;
open(INPUT,"nathan_splice_sites.iit");
my$threshold=15;
my$add=100;
while(my $line = <INPUT>){
	chomp($line);
	 my @temp = split(' ',$line);
	my$exon=$temp[0];
		my$chr_start_end=$temp[1];
		my$class=$temp[2];
		my$length=$temp[3];
	my @array = split(':',$chr_start_end);
		my$chr=$array[0];
		my$start_end=$array[1];
	my @beans = split('\..',$start_end);
		my$start=$beans[0];
		my$end=$beans[1];
my$ed;
my@array;
	if ($class eq "acceptor"){
			$ed=$end-$add;
		push @array, $chr, $ed, $end, $class;
	}
		elsif ($class eq "donor"){
			$ed=$end+$add;
		push @array, $chr, $end, $ed, $class;
			}
my@array2 = join(":", @array);

open(OUTPUT,">>for_samtools.txt");
foreach my$array(@array2){
print OUTPUT "@array2\n";
}
close (OUTPUT);
}

open(INPUT2, "for_samtools.txt");
while(my $line2 = <INPUT2>){
	chomp($line2);
	 my @temp2 = split(':',$line2);
	my$chr=$temp2[0];
		my$start=$temp2[1];
		my$end=$temp2[2];
		my$info=$temp2[3];
my$genome  = "agambiae.CHROMOSOMES-PEST.AgamP3.fa";
my$command="../samtools-0.1.16/samtools faidx $genome $chr:$start-$end";

if ($info eq "donor"){
```

```
        system($command);
}

}

# Script is ran for both donor and acceptor splice sites
# Command line is then used as below to merge the two sequences
# paste -d '' acceptor_Seq.fasta donor_seq.fasta | perl -pe 's/(?<=\w)>.*//' >
acceptor_donor_seq
# Combined fasta file is then used for BLAST searches against cleaned reads of males and
females separately
# The command cut -f1 acceptor_donor_seq_blastn_mid2 | sort | uniq -c >hits_mid2 is
then used to total the number of hits for each sex
```