

AN INVESTIGATION OF INSULATOR PROTEINS IN MOSQUITO GENOMES

A Thesis

by

MICHAEL JOSEPH JOHANSON

Submitted to the Office of Graduate Studies of  
Texas A&M University  
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,  
Committee Members,

Craig Coates  
David Stelly  
Keith Maggert  
Michel Slotman

Interdisciplinary Faculty Chair,

Craig Coates

August 2013

Major Subject: Genetics

Copyright 2013 Michael Joseph Johanson

## ABSTRACT

Transgenic mosquitoes are beneficial for the design and implementation of various pathogen control programs. However, low and variable expression of transgenes caused by position effects is a hindrance to the characterization and effective use of transgenes in mosquito species. The use of insulator sequences to flank transgenes may have the ability to overcome position effects caused by the genomic environment surrounding the insertion site. CTCF is a multifunctional protein, conserved from humans to *Drosophila*. Its role as an enhancer blocker in the *Drosophila* bithorax complex and its proximal binding to other insulator proteins on *Drosophila* chromosomes makes it a good candidate for identifying insulator sequences throughout the mosquito genome that may be used to improve mosquito transgenesis. Its multifunctionality as a transcription factor and genome organizer also makes CTCF worthy of investigation for an improved understanding of the regulation of the mosquito genome. This study uses chromatin immunoprecipitation with an *An. gambiae* CTCF antibody followed by Illumina deep sequencing (ChIP-Seq) to identify regions of CTCF binding throughout the *An. gambiae* genome. A subset of the CTCF binding site peaks was validated using ChIP-PCR. Another subset of this data set, including the ChIP-PCR validated peaks, was input into the motif finding tool, AlignACE, in order to identify a CTCF binding site consensus. Four motifs were identified, none of which were found in more than 11.9% of the ChIP-Seq data set. These results lead us to conclude that *An. gambiae* CTCF binds to a wider variety of sequences compared to *Drosophila* CTCF. This work also includes a comparison of the expression profiles of the dipteran insulator proteins, Su(Hw) and CP190, with that of CTCF across multiple life stages in *Ae. aegypti*. The results of this study suggest the possibility of genomic colocalization, as has been recently discovered in *Drosophila*. The identification of CTCF binding site peaks throughout the *An. gambiae* genome provides a large data set of potential insulator sequences that may be used to improve mosquito transgenesis, and provide a new model for the study of CTCF function in a species with medical significance.

## TABLE OF CONTENTS

	Page
ABSTRACT .....	ii
TABLE OF CONTENTS .....	iii
LIST OF FIGURES .....	v
LIST OF TABLES .....	vi
CHAPTER I INTRODUCTION .....	1
1.1 The Need for a Solution.....	1
1.2 Vector Control .....	3
1.3 Genetic Means of Control .....	4
1.4 The Natures of the Mosquito Genomes .....	10
1.5 Chromatin Maintenance and Regulation .....	14
1.6 CTCF as an Insulator Protein .....	22
CHAPTER II IDENTIFICATION OF REGIONS OF CTCF BINDING IN THE	
<i>Anopheles gambiae</i> GENOME.....	27
2.1 Introduction .....	27
2.2 Materials and Methods .....	29
2.2.1 Identification of potential CTCF binding sites in silico .....	29
2.2.2 Anti-CTCF ChIP using Sua 4 cultured cells.....	30
2.2.3 ChIP-Seq library preparation.....	32
2.2.4 Real-time PCR for validation of the ChIP-Seq library .....	32
2.2.5 Sequencing and analysis of the CTCF immunoprecipitated ChIP-Seq library .....	32
2.3 Results .....	33
2.3.1 In silico identification and validation of a CTCF binding site.....	33
2.3.2 CTCF binding site peaks identified via ChIP-Seq are over-represented near genes .....	36
2.3.3 CTCF binding site chromosome map .....	39
2.3.4 Some immune response genes may be regulated by CTCF .....	43
2.3.5 CTCF binds near some heme-peroxidase genes .....	44
2.3.6 CTCF and sex differentiation genes .....	46
2.3.7 CTCF binding site peaks at the <i>Anopheles gambiae</i> bithorax complex .....	47
2.3.8 Four sequence motifs identified among a subset of the CTCF binding site peaks .....	52
2.4 Discussion.....	55

	Page
2.4.1 The distribution of CTCF binding sites likely reflects its multiple putative functions .....	55
2.4.2 What effects does CTCF have on neighboring genes?.....	57
2.4.3 CTCF may regulate genes important for immunity.....	62
2.4.4 Some heme-peroxidase genes may be regulated by CTCF .....	64
2.4.5 CTCF binding site peaks are located near important sex differentiation genes .....	65
2.4.6 The CTCF binding site profile of the <i>Anopheles gambiae</i> bithorax complex .....	67
2.4.7 <i>An. gambiae</i> CTCF is associated with a variety of DNA sequence motifs .....	68
2.5 Conclusions .....	69
<b>CHAPTER III EXPRESSION PROFILES OF THE INSULATOR PROTEINS</b>	
CP190 AND SU(HW) IN <i>Aedes aegypti</i> .....	71
3.1 Introduction .....	71
3.2 Materials and Methods .....	74
3.3 Results .....	75
3.4 Discussion.....	77
3.5 Conclusions .....	78
<b>CHAPTER IV CONCLUSION AND FUTURE DIRECTIONS .....</b>	
<b>REFERENCES .....</b>	
<b>APPENDIX .....</b>	

## LIST OF FIGURES

FIGURE		Page
1	Dominant malaria vector species in Africa.....	2
2	Three models for enhancer-promoter interaction for gene activation.....	17
3	Luciferase activity from insulated and uninsulated transgene .....	22
4	Chromosome 2R from an <i>An. gambiae</i> ovarian nurse cell stained with the <i>An. gambiae</i> CTCF antibody.....	34
5	ChIP-PCR result for potential CTCF binding region found on Chromosome 2R at position 59,0160524 bp.....	36
6	Distribution of CTCF ChIP-Seq peaks in the <i>An. gambiae</i> genome.....	39
7	Distribution of CTCF binding site peaks along the five <i>An. gambiae</i> chromosome arms .....	41
8	Distribution of CTCF binding site peaks at chromosome bands compared to chromosomes immunostained with the CTCF antibody.....	42
9	Maps of genomic regions with CTCF binding site peaks in relation to selected genes of interest.....	49
10	Logos representing the motifs identified from among the 212 CTCF binding site peaks.....	54
11	Expression profiles of <i>An. gambiae cp190</i> and <i>ctcf</i> .....	74
12	Expression profiles of <i>Ae. aegypti cp190</i> and <i>su(Hw)</i> .....	77
13	Diagram of an insulated <i>attP</i> site.....	82

## LIST OF TABLES

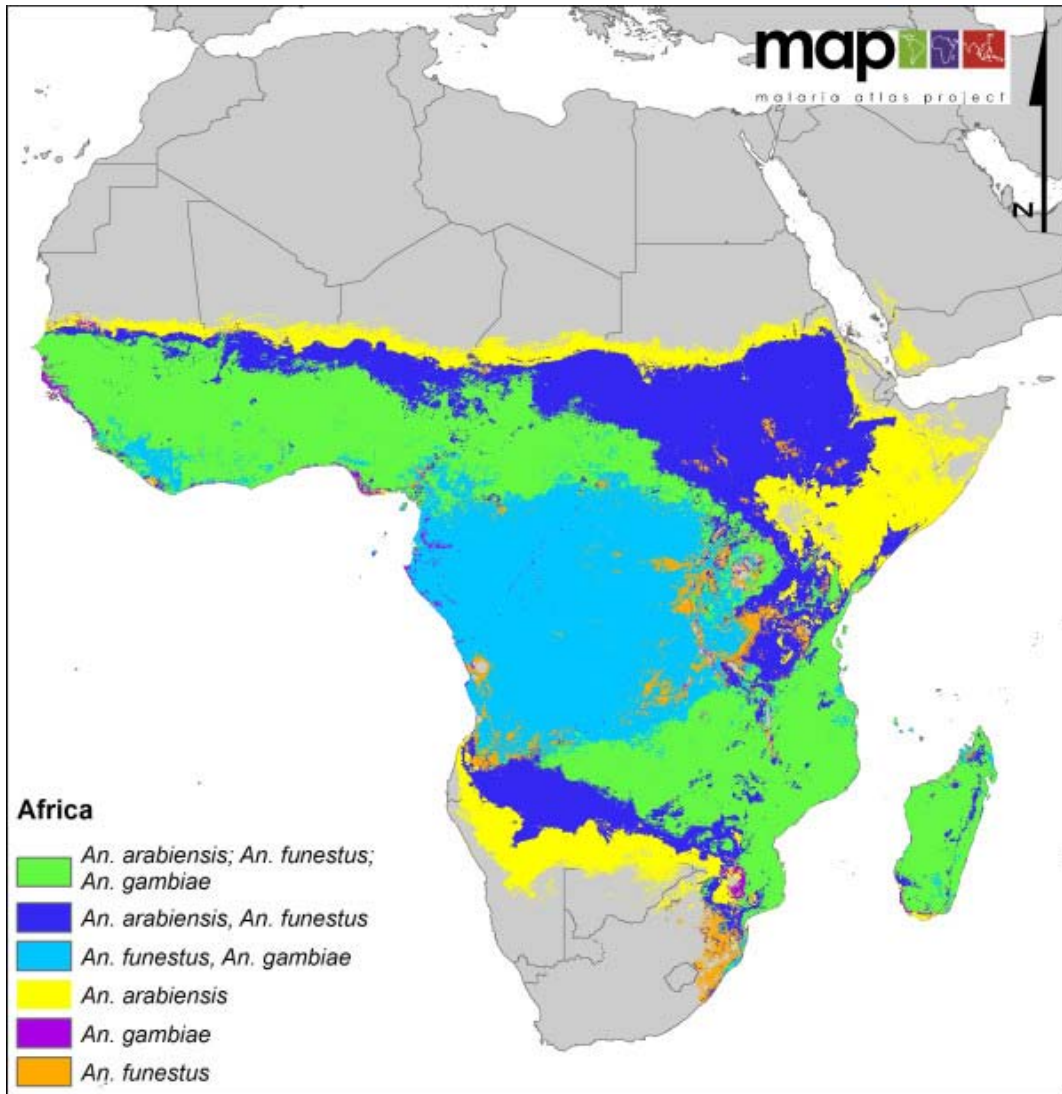
TABLE		Page
1	Effects of the WARI insulator on yellow and white expression .....	21
2	ChIP-Seq identified CTCF binding sites validated by ChIP-PCR .....	43

## CHAPTER I

### INTRODUCTION

#### 1.1 The Need for a Solution

Mosquitoes are responsible for more human death and illness than any other animal on the planet. The malaria vector, *Anopheles gambiae*, is one of the principle species responsible for these deaths. As the vector of *Plasmodium falciparum*, one of the parasites that cause human malaria, *An. gambiae* is one of the principle mosquito vectors responsible for malaria transmission on the continent of Africa [1]. This disease is a threat to 3.3 billion people, nearly half of the world's population [2]. Malaria is the cause of 20% of all childhood deaths in Africa. An African child is estimated to have between 1.6 and 5.4 episodes of malaria fever per year [2]. Pregnant women are also at risk. Not only is death due to complications of the disease a risk; spontaneous abortion, premature delivery, and stillbirth are also risks associated with the disease [2]. Malaria causes approximately 250 million illnesses per year and nearly one million deaths per year. More than 90% of these occur in sub-Saharan Africa [2]. Figure 1 shows the range of the three most dominant vectors of malaria in Africa.



**Figure 1: Dominant malaria vector species in Africa.** This map shows the distribution of the three most dominant malaria vector species in Africa. Sinka *et al.* *Parasites & Vectors* 2012 [1]

*Anopheles gambiae* is an efficient transmitter of malaria. It blood feeds almost exclusively on humans. The larvae develop in pools of water created by human activities, and the adults rest in human dwellings [3]. The degree of adaptation of *Anopheles gambiae* to humans has enabled the *Plasmodium* parasite to take advantage of the mosquito-human relationship to enhance its own parasitic relationship with humans. Anti- malaria drugs such as



quinine, chloroquine, Fansidar, mefloquin, and halofantrine have been used to treat malaria patients. However, over time, the plasmodium parasite has developed resistance to anti-malarial drugs. Currently, artemisinin drugs are the most effective treatments. Artemisinins are combined with longer acting malaria drugs for drug therapies known as artemisinin combination therapies (ACTs). Although effective, studies have not ruled out adverse reactions being linked to ACTs, and further studies are necessary to ensure that these drug therapies are safe [4, 5]. Such circumstances emphasize the need to focus the malaria control effort on the mosquito vector.

*Aedes aegypti* is the vector of dengue fever and yellow fever. Both of the diseases are caused by viruses and no effective treatments are available. Although relatively few deaths occur from these diseases, dengue can develop into dengue hemorrhagic fever, a complication that can often result in death. There is an effective vaccine for yellow fever, and it can be kept under control through vaccine campaigns; however, political instability in some countries results in the disruption of the vaccine's distribution. This was observed to be the case recently in Cote d'Ivoire [6]. On the other hand, no effective vaccine has been developed for the dengue virus, and thus this disease can only be controlled through control of the vector mosquito population [7]. Both *Anopheles gambiae* and *Aedes aegypti*, along with their respective pathogens, are currently limited to tropical and subtropical regions. However, it has been estimated that as global temperatures increase, the number of people at risk for these diseases will increase by 3-5% (several hundred million people) [8].

## **1.2 Vector Control**

The current means of mosquito vector control is insecticide treated bed nets and indoor residual spraying. Insecticide treated bed nets successfully repel and control mosquitoes. The insecticide treated bed nets are either given to the public or sold at a low cost in order for all people to have access to this protection and reduce the rate of disease transmission in endemic

areas. Indoor residual spraying is performed for all human dwellings in an endemic area. The goal of indoor residual spraying is to end the life of the mosquito before the end of the parasite's extrinsic incubation period (10 to 14 days for *Plasmodium*) [9], so that the *Plasmodium* parasite does not have a chance to develop in the mosquito and infect another human being. Once a female mosquito blood feeds on a human host, it becomes lethargic and rests on the walls of the human dwelling. This results in exposure to the insecticide, which results in the death of the mosquito before it has a chance to transmit mature *Plasmodium* to another human host. Both of these strategies have been effective in controlling vector-borne disease transmission; however, both rely on the use of insecticides. Insecticide resistance among mosquito species is rendering these strategies ineffective [3]. The distribution of these control methods is also vulnerable to civil unrest. Alternative genetic strategies are necessary for effective mosquito vector disease control.

### **1.3 Genetic Means of Control**

There are two ways to control the transmission of vector-borne diseases. The first is vector population suppression, and the second is vector population replacement [10]. Suppression is performed using insecticides, as mentioned above, in addition to alternative methods such as the sterile insect technique (SIT) and release of insects with a dominant lethal (RIDL). SIT is a method in which males are sterilized, usually via irradiation, and released into the wild to mate with wild type females. SIT is well suited for *Anopheles* mosquitoes because females tend to only mate once. Therefore, mating with a sterile male will result in the mating female potentially not producing any offspring for its entire lifetime. This approach results in a lower vector population and thus lower disease transmission [11]. Disadvantages to this include lower fitness of irradiated males, rendering them less effective at mating with the wild females, and the difficulty of separating the sterile males from the females.

Recently, Oxitech, an Oxford-based research firm has developed an alternative SIT strategy. Male mosquitoes are genetically modified in a way such that their offspring are sterile. This has the same effect as irradiation; however, genetically modified males and their sterile offspring can better compete for mating. Genetically modified males were released beginning in November and December of 2009 and followed up with a larger release between May and October of 2010. The study showed that the endogenous mosquito population had been suppressed by 80% by August 2010 [12].

RIDL, the release of insects carrying a dominant lethal, uses a different approach by using males carrying a female-specific dominant lethal gene that produces only male offspring, resulting in the suppression of the vector population [10, 13, 14]. RIDL uses a transcription factor gene under the control of a female-specific promoter or enhancer, which is necessary for the transcription of a toxic transgene. Alternatively, a transcription factor can also be used to drive the expression of a gene that is only lethal to females. The primary advantage of this method is that the effect can carry over into the next generation, as males are fertile. This method also eliminates the need to raise and eliminate females from a sterile insect strain and has the advantage of using transgenic males, which have a fitness advantage over irradiated males [15].

For both methods, it is important that only males be released, as females could contribute to an unwanted increase in mosquito populations, reduce the efficacy of the trial by mating with sterile males, and transmit disease. Although RIDL eliminates females before release, most systems do not eliminate them until the adult stage resulting in the extra cost of raising unwanted females up to this life stage. In the earlier part of the last decade, transgenic tagging systems were developed such that fluorescent transgenes that were only expressed in the testes could be used to sort males from females using a flow cytometer. This system was limited by the fact that the fluorescence could only be detected at the late larval stage. Recently this

system has been enhanced such that early sex-specific transgenic markers can be detected at early stages of development of *Anopheles gambiae* mosquitoes. Larval populations can be sorted by sex, transgenic/non-transgenic, heterozygous/homozygous, transgenic females/non-transgenic males. The system also has no effect on the mating ability of the adult males, thus improving the productivity of population suppression systems [11].

Population replacement is based on a strategy using transgenic mosquitoes that are resistant to the vector-borne disease to replace the current mosquito population [10]. This avenue of vector control shows some promise with several advances having been made in recent years [16]. The use of transposable elements for inserting transgenes into mosquito genomes has been successfully pursued. Six mosquito species, *Aedes aegypti*, *Aedes fluviatilis*, *Anopheles albimanus*, *Anopheles gambiae*, *Anopheles stephensi*, and *Culex quinquefasciatus*, have all been genetically transformed with transposons carrying transgenes [16]. Some of these transgenes, such as [SM1]4, PLA2, and Cecropin A have conferred some level of *Plasmodium* resistance to the transformed mosquito species[16]. Transgenesis using dsRNA constructs may be used to silence genes necessary for *Plasmodium* transmission. An *Ae. aegypti* strain expressing dsRNA targeting *RELI*, an innate immune response gene, experienced *RELI* inhibition [16], demonstrating the potential of this approach.

Paratransgenesis, the use of genetically modified symbionts to reduce vector competence, also shows possibilities for disease transmission control. *Asaia* sp. bacteria have been successfully transformed with enhanced green fluorescent protein (EGFP) expressing plasmids and introduced to adult mosquitoes through a sugar or blood meal resulting in infection. Larvae were also able to be infected with the bacteria from their aquatic environment [16]. *Asaia* are found in the mosquito midgut and salivary glands, which are the sites of pathogen development and transmission. The bacteria are also transmitted from male to female during

mating and then transmitted vertically to the offspring[16]. To demonstrate the proof of principle, *Escherichia coli* expressing the anti-malaria molecules SM1 and PLA2 were able to inhibit *P. berghei* development in *An. Stephensi* [16].

Even with all of the recent success in mosquito transgenesis, many challenges remain. Blockage of *Plasmodium* infection through mosquito transgenesis has not yet been achieved for human parasites [16]. To date, mosquito transgenesis has only been shown to block *P. berghei*, the rodent parasite and one of these cases was in a non-natural mosquito-parasite pair, *An. stephensi*-*P. berghei*. The one successful demonstration of a transgenic insect impairing the development of a human pathogen is a transgenic strain of *Ae. Aegypti*, which inhibits the dengue virus development [16]. The relative fitness of transgenic mosquito populations has also been an issue. As mentioned above in regard to SIT, a less fit transgenic population would be unable to drive the transgene through the natural population. Although it was shown that *An. stephensi* mosquitoes hemizygous for the SM1 transgene exhibited higher fitness than wild type mosquitoes when fed on *P. berghei* infected mice, mosquitoes that were homozygous for the SM1 transgene, exhibited lower fitness [16].

On a more optimistic note, transgenic *An. stephensi* mosquitoes generated with the  $\Phi$ C31 integrase system, expressing a fluorescent marker gene, showed no significant difference in fitness when compared with wild type mosquitoes [17]. This provides an integration system that may be useful for effective mosquito transgenesis. This is particularly helpful in controlling position effects as the *attP* docking site provided by the  $\Phi$ C31 integrase system allows for site specific integration into the genome [18]. This results in the transgene integrating at the same position in every transgenic line, thus avoiding variable expression levels due to position effects caused by random integration of transposable elements.

Other problems exist in relation to using transposons to insert transgenes into insect genomes. Non-canonical transposition reactions can result in integration of donor plasmid fragments throughout the insect genome. Transgene size can influence transposon activity. Transposons may also remobilize into somatic tissues and cause damage in some regions of the genome [16]. A serious ecological problem may also result in the event of horizontal transfer of the transgene to a sibling or non-related species [16]. Another problem with the use of transposons for transgene integration is the variable expression of transgenes due to random integration into the genome that results from the cut and paste transposase system. Random integration results in both position effects (PE) and position effect variegation (PEV). Position effects are the result of a transgene being affected by regulating elements near the insertion site. For example, insertion of the transgene near an enhancer may result in over expression; insertion near a silencer may result in reduced or no expression. Position effects result in various expression levels between transgenic lines due to varying chromosomal environments at each insertion site. Position effect variegation is the result of repression of transgene expression due to heterochromatin spreading at the insertion site, leading to silencing of the transgene. The amount of heterochromatin spreading at each genomic location is variable between different cells and tissues within the organism; therefore, PEV causes variable expression of the transgene within a transgenic line. These two issues are of concern for many applications of mosquito transgenesis.

Examples of position effects have occurred in transgenic lines of *Ae. aegypti* involving two different transposable elements. In two separate studies, the *Hermes* and *Mariner* elements were used to insert the *D. melanogaster* wildtype *cinnabar* gene ( $cn^+$ ) into the *Ae. aegypti* genome. All transformed mosquitoes were mutants for the sex-linked white gene  $kh^w$ ; therefore colored eyes in subsequent generations indicated insertion of the gene [19]. In the *Hermes* study, four founder families and one pool produced  $G_1$  progeny with colored eyes with varying eye

color from light to dark red. In the *Mariner* study, five founder families produced G<sub>1</sub> progeny with colored eyes. Eye color varied between families ranging from a light orange to purple/black. These results indicate that the different positional insertions across these families resulted in different levels of transgene expression due to different chromosomal environments. One family from each study had progeny with varying eye color among them. Such results suggest position effect variegation due to the variation in gene expression among individuals with a common insertion site [20, 21].

Benedict [22] identified a PEV phenotype in *An. gambiae* in a cross of *pink eye* (*p*) females with irradiated males. This variegated phenotype consists of patches of wild-type ommatidia over a pale pink background. This phenotype was named *Mosaic* (*Mos*). Genetic studies confirmed that *Mos* was sex linked, and suggested that recombination occurs between *Mos* and *pink eye* (*p*). The estimated distance between the two was 14.4 cM. Crosses were also conducted to determine if *Mos* would be expressed in a *white* mutant background. The crosses revealed that *w* is epistatic over *Mos*. Cytogenetic analysis of ovarian nurse-cell polytene chromosomes of *Mos/Mos*<sup>+</sup> and *Mos/Mos* females revealed an insertion of euchromatin into the heterochromatic region of division 6 on chromosome X. Based on the genetic analysis, it was suspected that a wild type *pink eye* (*p*<sup>+</sup>) duplication might be involved in the insertion. The region of the chromosome at which *pink eye* is located, 2B, was compared to the euchromatic insertion at division 6 and the cytological appearance was similar. To confirm that the insertion was indeed a *p*<sup>+</sup> duplication, two mapped cDNAs of the 2B region, c51 and c81, were hybridized independently to ovarian polytene chromosomes of *Mos* homozygotes. Indeed, c51 consistently hybridized to the insertion and 2B, confirming that the insertion was a duplication of *pink eye* [22]. Benedict[22] believes that this insertion was the result of a transposition event in which *p*<sup>+</sup> was inserted into the heterochromatic region of division 6 on chromosome X, thus deleting the

$p^+$  allele such that it would not complement the pink eye mutation on its homologue. The new position of the  $p^+$  allele resulted in a PEV phenotype.

#### 1.4 The Natures of the Mosquito Genomes

The genomes of *An. gambiae* and *Ae. aegypti* are quite different from that of *Drosophila melanogaster*, in which the majority of insect genome studies have been performed. The genome sizes among the three vary substantially in size with *An. gambiae* having two-fold the genome size at 272.8 Mb compared to 118 Mb for *Drosophila* [23]. The *Ae. aegypti* genome is 5 times larger than that of *An. gambiae* at 1,376Mb [23]. *Aedes aegypti* has nine times the average length of intergenic region compared to *Drosophila* (six times compared to *An. gambiae*) and four times the average intron length in *Drosophila* (three times compared to *An. gambiae*) [23]. Some of these differences are due to loss of non-coding DNA from the *D. melanogaster* genome. This is supported by the fact that all *Anopheles* species have genome sizes between 240Mb and 290Mb and all other culicids have genomes of 500 Mb or greater, and all except two *Drosophilid* species have genome sizes of 230Mb or greater [3]. The number of coding genes, exons, and coding lengths vary by less than 20% between *Drosophila* and *Anopheles* [2]. The variation in genome size is likely due to the loss of non-coding DNA sequence from *D. melanogaster* and the insertion of transposable elements throughout the two mosquito genomes over evolutionary time [3, 23, 24].

The *An. gambiae* genome has approximately 40 different identified types of transposons. Most of these are Class I repeats; particularly long terminal repeat retrotransposons (LTRs), small interspersed nuclear elements (SINEs), and miniature inverted repeat transposable elements (MITEs). All of the major Class II transposon families are also represented [3]. Transposon densities differ according to the chromosomal arm. The X chromosome has the highest transposon density with 59 transposons per Mb. Chromosomal arms 2R, 2L, 3R, and 3L



have 37, 46, 47, and 48 transposons per Mb respectively. The large number of paracentric inversions on chromosomal arm 2R may be related to its lower density of transposable elements as recombination is more frequent in regions where transposon density is lower [3]. Transposons make up approximately 16% of the genome's euchromatin and 60% of its heterochromatin, compared to 2% and 8%, respectively of the *D. melanogaster* genome. Transposons present in heterochromatin are highly fragmented; therefore, 60% is likely an underestimate. It has been noted that there must be a mechanism within the heterochromatin that promotes transposon loss from these regions in order to balance the insertion of new copies [3].

The *An. gambiae* and *D. melanogaster* genomes have 12,981 one-to-one orthologs and 1,779 many-to many-orthologs [24]. This supports the notion that most of the differences in genome sequence are due to intergenic non-coding DNA as mentioned above. In addition to intergenic regions, introns are also important in accounting for this difference, as well as some genes unique to mosquito biology.

A comparison of protein coding genes between *Ae. aegypti*, *An. gambiae*, and *D. melanogaster* show that the mosquito species have a significant number of unique genes shared exclusively between them. This demonstrates the unique biology shared exclusively among mosquitoes. Comparison of orthologs among these three species reveals that 67% of the *Ae. aegypti* proteins have an ortholog in the *An. gambiae* genome, with only 58% having an ortholog in the *D. melanogaster* genome [23]. Comparison of three way single copy orthologs showed 74% average amino acid identity between the mosquito species compared to 58% identity between mosquito and fruit fly. Approximately 2,000 orthologs are shared only between the two mosquito species, possibly representing functions unique to mosquito biology [23]. It seems likely that the mosquito's hematophagy would contribute to the difference in coding genes compared to *Drosophila*. Interestingly, only one gene family, the peroxidases, demonstrates

major differences in gene copy numbers between *An. gambiae* and *D. melanogaster*. Preliminary analysis indicates that peroxidases are important during the invasion of the midgut by malaria parasites [24]. Peroxidases have also been linked to blood feeding [25].

Genome organization is varied greatly between *Anopheles* and *Drosophila* such that only small gene neighborhoods have been retained. This is known as microsynteny [24]. Numerous local inversions, translocations and gene duplications have resulted in two very different genomes. Such events may have led to the loss of non-coding DNA from the *Drosophila* genome as well as to the relatively rapid evolution of this non-coding DNA, thus leading to the divergence in genome structure. Insertion of transposable elements is also likely to have led to this divergence. Overall, 4,099 *Anopheles* genes and 4,244 *Drosophila* genes are assigned to 948 confirmed microsynteny blocks. The fraction of orthologs that remain within microsynteny blocks determined to exist between these two species is 34% in *Anopheles*. This figure represents a significant amount of local neighborhood conservation between *Anopheles* and *Drosophila*; however, it is considerably lower than the corresponding fraction (40%-50%) between puffer fish and mammals [24]. This highlights the much higher rate of insect evolution compared to vertebrates.

The most conserved pair of chromosomal arms between *An. gambiae* and *D. melanogaster* is *Dm2L* and *Ag3R*, with 95% of microsynteny blocks in *Dm2L* mapping to *Ag3R*. The remaining 5% represent exchanges with other arms which fail to have a significant signal above random expectation [24]. The chromosomal arm *Ag3R* microsynteny also maps significantly to *Dm2L* at 83%. Dual correspondence is detected in other arms, with one arm of a species corresponding with two arms of the other. For example, the *Anopheles* 2L arm contains approximately 42% and 54% of the contents of the *Drosophila* 2R and 3L arms, respectively [24]. This further illustrates the genomic rearrangement that has occurred between these two

species over evolutionary time. The loss of non-coding DNA in *D. melanogaster* and the lower number of transposon insertions are likely to have led it to have a much more condensed genome with most of its heterochromatin localized to the centromeres and telomeres in comparison to *An. gambiae*, resulting in a smaller number of blocks of euchromatin. As mentioned previously, the difference in genome size between *Drosophila* and *Anopheles* is likely due to differences in intergenic DNA sequence due to a higher number of transposable element insertions into the *An. gambiae* genome. The presence of interstitial blocks of heterochromatin along euchromatic chromosome arms leads to the possibility of stretches of intergenic heterochromatin flanking active euchromatic genes, or legitimate gene repression is carried out by blocks of intercalary heterochromatin. Alternatively, some of the *Anopheles* genes may have adapted to being expressed in heterochromatin, similar to the *Drosophila* gene, *light*.

The amount of repetitive sequence and high amount of genetic variation within both of the above mentioned mosquito genomes have caused much difficulty in assembling them. Although most of the *An. gambiae* gene sequence has been mapped to chromosomes, a number of unassembled chromosome fragments are classified as unknown. Also, the repetitiveness of the *An. gambiae* genome makes analysis of intergenic regions difficult. The *Ae. Aegypti* genome sequence is yet to be mapped to chromosomes due to the large number of repeats that have hindered its assembly. The *Ae. aegypti* Liverpool strain genome is organized into contigs, sequences that have been mapped together to form long stretches of assembled sequence. The 1.3 Gb genome is organized into 4,758 supercontigs. The average length of these supercontigs is 1500 kb. Smaller assembled sequences are known as contigs and have a length of 82 kb [26]. Although supercontigs are helpful for knowing where many sequences are in relation to many other sequences on a single contig, the gaps between contigs make effective study of the whole genome problematic. Such features have hampered genomic analysis, eliminating such tools as

microarrays with sufficient coverage of the genome to analyze gene regulatory regions, which have been quite useful in *Drosophila*.

### **1.5 Chromatin Maintenance and Regulation**

Eukaryotic genomes are organized into domains of active and silenced chromatin. These chromatin domains can be defined differently; either as actively transcribed versus inactive chromatin domains, as DNase I-sensitive versus DNase I-resistant chromatin domains, or by the distribution of specific histone variants. Active domains tend to have a higher concentration of acetylated core histones as well as histone H3 methylated at lysine 4 (H3K4) [27]. Silent domains tend to have a higher concentration of histone H3 methylated at lysine 9 (H3K9) and histone H3 methylated at lysine 27 (H3K27). Active and silenced domains can be referred to as euchromatin and heterochromatin respectively. These are defined cytologically with the tightly condensed dark bands of chromosomes referred to as heterochromatin and the clear interbands referred to as euchromatin.

Two types of heterochromatin exist in eukaryotic organisms. Constitutive heterochromatin is the chromatin that remains silent in almost all cell types and primarily resides at the centromeres and telomeres of chromosomes. Constitutive Intercalary heterochromatin may also be present throughout the chromosome arms. Facultative heterochromatin refers to regions of euchromatin that are silenced during cellular development [27]. The genomic regions of facultative heterochromatin vary from cell type to cell type, and play an important role in cell differentiation. An example of the function of facultative heterochromatin is X chromosome inactivation in female mammals.

Heterochromatin is tightly condensed, resistant to crossing over, late replicating and is unable to be transcribed in most cases, although exceptions do exist [27]. It is believed that heterochromatin is important for gene regulation [28]. Models for heterochromatin formation

and spreading are based on experiments in fission yeast, *Drosophila*, and mammals. The generation of double-stranded RNA is required for initiation of pericentric heterochromatin nucleation. The double-stranded RNA is processed by the RNAi mechanism into small interfering RNAs (siRNA) [27, 29]. The siRNAs are necessary to target proteins necessary for heterochromatin formation to the centromere. It has been shown that deletion of RNAi machinery such as *ago1*, *dcrl*, and *rdp1* is correlated with loss of pericentric heterochromatin and transcription of pericentric reporter genes [29, 30]. Transcription of repetitive sequences at the centromeres leads to recruitment of RNAi machinery and accumulation of siRNAs. This accumulation of siRNAs results in recruitment of histone modifying proteins such as Rpd3, Hda1, Su(var3-9) and HP1. These proteins form heterochromatin by deacetylating, methylating, and binding to Histone H3 at lysine 9 (H3K9) [29, 30]. HP1 then assembles nucleosomes into heterochromatin, and helps it spread by recruiting RNAi machinery and HDACs to continue the process bidirectionally. [29, 31].

Gene silencing in intercalary heterochromatin is independent of H3K9 methylation. In some regions, methylated H3K27 and polycomb-group proteins govern the silencing process. Intercalary heterochromatin is composed of unique sequences, including transposons in some cases. Intercalary heterochromatin is generally the sum of multiple silenced genes in close proximity to one another, with synchrony in replication, resulting in a visible band of heterochromatin. Such groups of genes in *Drosophila* include the homeotic genes, which are regulated throughout development. These sequences include a Polycomb response element (PRE), which recruits Polycomb group proteins [32, 33]. This recruitment is mediated by HMT (EZ), which catalyzes H3K27 methylation. A Polycomb protein complex binds to H3K27me<sub>2/3</sub> via its chromodomain, similar to the manner in which HP1 binds to H3K9me<sub>2/3</sub>. The polycomb proteins interact with transcription initiation proteins to maintain repression of transcription [34].

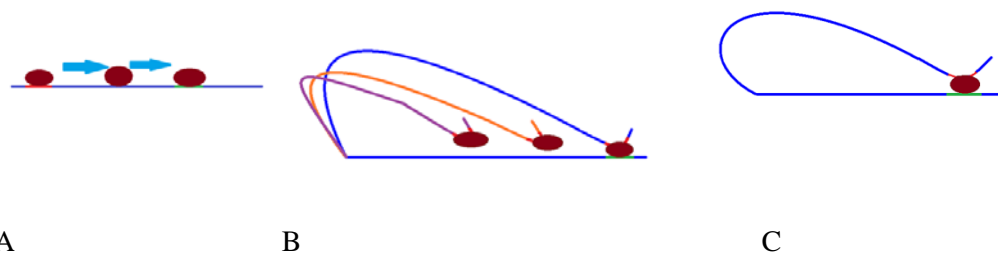
Pericentric and intercalary heterochromatin are necessary for regulation of the genome.

Pericentric heterochromatin is necessary for establishment of the centromere in order to maintain chromosomal segregation. Intercalary heterochromatin is necessary to restrict the expression of specific genes to specific tissues throughout the development and life of the organism.

The maintenance of heterochromatin/euchromatin boundaries is also important for proper gene regulation. Studies of cHS4, a complex vertebrate insulator located at the extreme 5' end of the chicken  $\beta$ -globin locus, support a model for a molecular mechanism that blocks the spread of heterochromatin. An insulator is a DNA sequence that regulates gene expression. The barrier function at cHS4 is established by the sequence-specific DNA-binding proteins upstream transcription factor 1 (USF1) and USF2. These two proteins recruit histone acetyl transferases and histone methyl transferases, which lead to cHS4 mediated acetylation and H3K4 methylation of nucleosomes. Mutations of cHS4 that disrupt the binding of USF1 and USF2 eliminate recruitment of HATs and HMTs and abolish barrier activity at cHS4. This led to the proposal that acetylation and H3K4 methylation of nucleosomes renders them resistant to H3K9 methylation and HP1 binding, thus stopping the spread of heterochromatin. Similar mechanisms may be based on other heterochromatin histone modifications. [35]

The formation and maintenance of active chromatin is necessary for gene transcription and is an active process. Specific DNA elements are required to recruit chromatin remodeling and modifying enzymes to open the domain so that enhancers can communicate with promoters. Factors necessary for activating chromatin vary from one locus to another [27]. The activation of chromatin is necessary for gene transcription; however, chromatin activation does not guarantee gene transcription. The presence of specific transcription factors may be necessary for transcription of a particular gene to occur. Once a chromatin domain is activated, an enhancer sequence is able to communicate with the promoter(s) of the gene(s) it regulates. Considering

that enhancers may be on the order of 100 kb upstream or downstream of the gene(s) they regulate, the manner in which communication occurs is of much interest. Some models suggest looping of the intermediate chromatin, while others suggest the tracking of recruited activator proteins along the chromatin fiber to the promoter region where it interacts with the transcription initiation complex. Another, known as the hopping model, involves the random sampling of an enhancer bound activator along the intervening chromatin until it encounters the target promoter [27]. Figure 2 illustrates these three models of gene activation. The formation of transcription factories is another version of chromatin looping in which an enhancer sequence, such as a locus control region (LCR), recruits multiple promoters and the necessary transcription machinery to facilitate the transcription of multiple genes [36].



**Figure 2: Three models for enhancer-promoter interaction for gene activation.** A) The tracking model suggests that once the transcription machinery (maroon oval) is recruited to the enhancer sequence (red) it travels the length of the sequence until it reaches the promoter sequence (green) at which point transcription can be initiated. B) The hopping model suggests that once the transcription machinery is recruited to the Enhancer, it samples the intervening sequence (purple and orange lines represent unsuccessful sampling attempts) until it makes contact with the appropriate promoter sequence. C) The looping model suggests that the enhancer sequence directly recruits the transcription machinery to the promoter sequence.

Active domains allow for the communication of enhancers with promoters. Given that enhancers can interact with multiple promoters, the need arises for regulation of this interaction

to control appropriate gene expression throughout the genome. One of the ways this is facilitated is through the partitioning of the genome into active and inactive domains. This partitioning of the genome maintains the expression of genes in one active domain to be controlled only by the enhancers of that domain. Gene expression within the genome may also need to be regulated within an active chromatin domain. This calls for the need of an enhancer blocker. This concept is illustrated at the human *Igf2*/*H19* locus at which an insulator sequence regulates the expression of these two genes. *H19* is paternally imprinted and *Igf2* is maternally imprinted. In this case, the two genes share enhancers that are located downstream of the two genes. The imprinting of one of the two genes is determined by differential methylation patterns of the parent of origin. At the paternal locus, a region downstream of *Igf2* and upstream of *H19* is differentially methylated, thus blocking the binding of CTCF, an insulator binding protein. This allows for the enhancer to communicate with *Igf2* and bypass the methylated *H19* promoter, thus *Igf2* is expressed and *H19* is imprinted. At the maternal locus, this differentially methylated region is not methylated, allowing for the binding of CTCF which inhibits the communication of the enhancers with the *Igf2* promoter and allows interaction with the *H19* promoter. Thus the binding of CTCF to the differentially methylated domain acts as an enhancer blocker at the maternal locus[37]. Enhancer blocking is an important regulatory function throughout the genome, and is one of the two functions of insulators.

As mentioned above, an insulator is a DNA sequence that regulates the genome by neutrally regulating gene expression. It neither specifically silences nor activates genes. Its function is to regulate the influence of silencing and activating elements through enhancer blocking and/or maintaining a barrier between euchromatin and heterochromatin. The example above illustrates the enhancer blocking function of an insulator. Another example of an insulator acting as an enhancer blocker is the case of an insulator discovered on the 3' end of the



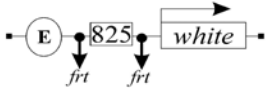

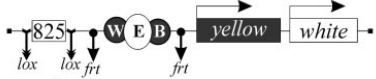
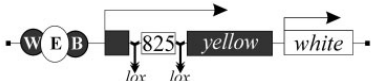
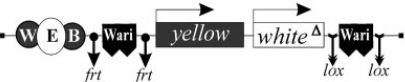

*Drosophila white* gene, known as *white* abutting resident insulator (Wari) [38]. Multiple transgenic experiments were performed to characterize the enhancer blocking function of this insulator sequence. The results of these experiments are summarized in Table 1. Chetverina *et al.* [38] began by excising the insulator sequence and inserting it between the eye enhancer and the *mini-white* gene. This reduced expression of *mini-white*. Excision of the insulator sequence from this position restored *mini-white* expression as expected. They also performed a similar experiment with a different construct in which *yellow* preceded *white* and the insulator was placed downstream of the wing, eye, and body enhancers and upstream of *white* and *yellow*. The expression of *yellow* and *white* were reduced, showing that the insulator was not specific to the *mini-white* gene. When the insulator was placed upstream of the enhancers such that it was no longer between the enhancers and the two reporter genes, there was no effect on the expression of either gene. Insertion of the insulator in an intron of the *yellow* gene resulted in *white* experiencing weakened expression, while *yellow* was unaffected. These experiments show that the function of this insulator is position dependent [38].

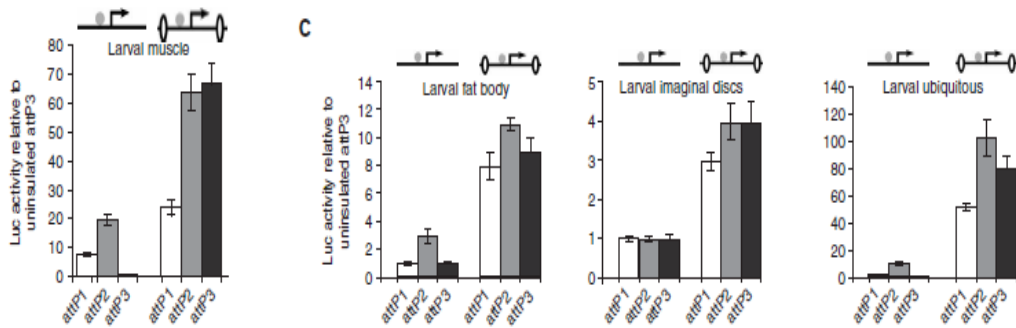
Assays also showed that a pair of Wari sequences, flanking the *yellow* and *white* gene sequences, decreased gene expression for both reporter genes even more so than a single Wari sequence positioned between the enhancers and gene promoters. Other assays showed that if the second insulator sequence was placed between the *yellow* and *white* genes, only the gene sequence flanked by the two insulators will have reduced expression. These results are summarized in Table 1. If *yellow* was flanked by the two insulators, only *yellow* had reduced expression, leaving *white* unaffected or with increased expression. Chetverina *et al.* [38] deduced that the interaction between the two insulators, resulting in the looping out of the *yellow* gene, would thus result in its insulation from the enhancer and its decrease in expression. This looping due to insulator interaction results in the enhancer bypassing the insulators and the activation of

the *mini-white* gene. The looping would also result in decreasing the distance between the eye enhancer and the mini-white gene, thus increasing the expression of mini-white [38].

Due to their ability to block interactions of enhancers and silencers with promoters, as well as block the spread of heterochromatin, it has been proposed that flanking a transgene with insulators will ameliorate the problem of position effects [39]. Site specific integration has also been proposed to solve this problem; however, predetermined integration sites are not always optimal for sufficiently high levels of transgene expression [39]. The  $\Phi$ C31 integrase system integrates transgene containing plasmids with *attB* sites at *attP* landing sites within the genome. However, the *attP* landing sites are randomly integrated into the genome. Therefore, position effects can still affect the expression of the transgene [39]. It has been shown that position effects vary greatly from one tissue to another at any *attP* site in *Drosophila* [39]. However, Markstein *et al.* (2008) [39] have shown that flanking the transgene with the gypsy insulator results in significantly increased gene expression at three different *attP* sites compared to the un-insulated loci (Figure 3) [39]. This provided the opportunity to create transgenes at a single locus that can be highly expressed in most, if not all, tissues in *Drosophila* [39]. The use of insulators in conjunction with  $\Phi$ C31 sites may provide the same opportunity for *An. gambiae* and other mosquito species.

**Table 1: Effects of the WARI insulator on *yellow* and *white* expression.** Construct maps from Chetverina *et al.* show orientation of the Wari insulator in relation to wing, body, and eye enhancers (W, E, B) and *yellow* and *white*. The box containing 825 in the first four diagrams refers to an 825 bp sequence containing the Wari insulator. Reduced or Increased refers to expression levels with reference to wild type expression. Images from Chetverina *et al.*(2008) [38].

Construct Map	<i>yellow</i> expression	<i>white</i> expression
	N/A	Reduced
	Reduced	Reduced
	Wild Type	Wild Type
	Wild Type	Reduced
	Reduced	Reduced
	Reduced	Wild Type or Increased



**Figure 3: Luciferase activity from insulated and uninsulated transgenes.** Luciferase activity from the un-insulated (left) and insulated (right) (ovals represent gypsy insulator sequences) induced UAS::Luciferase transgene induced in larval muscle, larval fat body, larval imaginal discs, and larval ubiquitous. (Markstein *et al.* (2008) [33])

### 1.6 CTCF as an Insulator Protein

CTCF is the only known insulator protein in vertebrates. It is a zinc finger protein, which is also known to act as an activator, a repressor, and a chromatin organizer. It is important for cell functions such as growth, proliferation, differentiation, apoptosis, nucleosome positioning, X-chromosome inactivation, cell cycle regulation and imprinting [40, 41]. CTCF has been described as a genome organizer which binds to multiple sequences throughout the genome, which, along with chromatin context, dictate the function of CTCF at each binding site [42]. CTCF is also found in insects and has been shown to play a role in insulator function [43]. Therefore, we see CTCF as an important factor for regulation of the *An. gambiae* genome, which deserves further study for the purposes of understanding *An. gambiae* gene regulation and identifying binding sequences that could be used to improve *An. gambiae* transgenesis.

CTCF has been shown to bind at important insulators in the bithorax complex of *D. melanogaster*. However, other insulator proteins such as Su(Hw), BEAF-32, Zw5, GAGA factor, MOD(MDG4) and CP190 also exist in insects. The presence of multiple insulator proteins in

insects, specifically dipterans, implies a need for a variety of insulator proteins in insect genomes. How the presence of additional insulator proteins affects the role of CTCF in insects as compared to its role in vertebrates is yet to be elucidated; however, *Drosophila* ChIP-Seq data show that CTCF, Su(Hw), BEAF-32, MOD(MDG4), and CP190 bind in tandem at many genomic locations suggesting a synergistic relationship among these proteins for insulating activity [44].

CTCF has been shown to be necessary for the function of the *Frontal abdominal 8 (Fab-8)* insulator in the bithorax complex (BX-C) of *Drosophila* between *iab 7* and *iab 8* [43]. CTCF binding to *Fab 8* was confirmed by a methylation interference assay. Mobility shift assays with negative control or mutant CTCF binding site sequences and wild type sequences revealed that the wild type sequences were retarded by protein binding and the mutated sequences and negative control sequences were not, showing that the *Fab 8* sequence is necessary for the binding of CTCF at the *Fab 8* insulator [43, 45]. Chromatin immunoprecipitation assays in conjunction with PCR using primers flanking the *Fab 8* sequence and sequences not shown to bind to CTCF revealed that CTCF binding was unique to *Fab 8* [43, 45]. Enhancer blocking assays have also confirmed the role of *Fab 8* as an insulator when binding CTCF [43]. Experiments with mutated CTCF binding sites in the *Fab 8* sequence incorporated into an EGFP reporter gene construct transfected into stable S2 cell lines, as well as a CTCF knockdown using RNAi in stable S2 cell lines with a similar EGFP reporter gene construct containing wild type *Fab 8* insulators showed that CTCF is necessary for the function of the *Fab 8* insulator [46].

Smith *et al.* found that CTCF tends to bind near the promoters of genes and between gene promoters that are transcribed in opposite directions and those that are spatially or temporally divergent [47]. ChIP-on-chip data showed that CTCF binds between the *Drosophila* ortholog of the human gene implicated in Alzheimer's disease,  *$\beta$  amyloid protein precursor-like*

(*Appl*) and an uncharacterized transcript *CG4293*. Affymatrix expression data shows that *Appl* is expressed in the embryo 6 hours into development and that *CG4293* is likely maternally loaded and is transcribed early in embryogenesis. The ChIP-on-chip data shows a strong CTCF peak between the two divergently transcribed promoters. An example of spatially divergent genes separated by CTCF is the case of the divergently transcribed *bicoid* (*bcd*) and *Amalgam* (*Ama*) genes. *bcd* is restricted to the anterior of the early embryo, whereas *Ama* is expressed in embryogenesis and is localized in the dorsal region and neural ectoderm of the embryo. Such CTCF binding patterns suggest that CTCF may be necessary for the differential regulation of closely positioned genes [47].

The experiments described above apply specifically to the enhancer blocking insulator function of dCTCF in *Drosophila*. Experiments in vertebrates show that CTCF also plays an important role as a chromatin barrier. The human tumor suppressor gene, p16<sup>INK4</sup>(p16), is flanked by heterochromatin borders from approximately 2kb upstream of the transcription start site to approximately 1kb to 4kb downstream of the transcription start site. In breast cancer cells with aberrantly silenced p16 genes, these heterochromatin borders are absent [48]. Examination of the sequence 3' of the heterochromatin border revealed the presence of a CTCF binding site. CTCF was observed to be associated with this region in p16 expressing cell lines, yet not in p16 non-expressing cell lines [48]. Chromatin immunoprecipitation analysis revealed that in p16 expressing cells, CTCF clearly binds downstream of heterochromatin in the p16 promoter region. No CTCF binding occurred near -7 Kb or +4 Kb in relation to the p16 transcription start site. Interestingly, although CTCF is not present at this region in p16 non-expressing cells, it is present at other genes such as *c-myc* [48].

Knock-down of CTCF with shRNA was performed in p16 expressing cells to observe its effects on p16. The mRNA levels of p16 as well as H19 were reduced. However, *c-myc*

remained impervious to the loss of CTCF, suggesting a different function for CTCF at this locus. H2A.Z was absent from the p16 promoter and a 3' shift of H4K20 also occurred. Transcription of p16 could be restored with the drug, AZA; however, AZA does not restore CTCF binding [48].

Orthologs for CTCF have been identified in *Anopheles gambiae* and *Aedes aegypti* [49]. These orthologs have 38% identity and 56% similarity with *D. melanogaster* and *H. sapiens* across all eleven zinc finger domains, respectively, and 68% of the critical binding residues are conserved. The expression of the two mosquito CTCF proteins span across all life stages with increased expression in the embryo and ovary of the blood fed female, which is consistent with its potential role as an insulator protein at important developmental stages [50]. They are also believed to be bound to nuclear structures and expressed in ovarian nurse cells [49, 50]. CTCF Immunostained chromosome spreads have revealed a low resolution distribution of CTCF binding along the chromosomes of *Anopheles gambiae* ovarian nurse cells. This work identifies regions of CTCF binding in the *Anopheles gambiae* genome at a higher resolution through the ChIP-Seq technique. The expression profiles of *Aedes aegypti cp190* and *su(Hw)* were also determined and examined in the attempt to initiate further work that may lead to the identification of yet more effective insulator sequences that could be used for the improvement of mosquito transgenesis by overcoming the challenges of Position Effects and Position Effect Variegation.

The prospect of this work is that the identification of CTCF binding regions will provide another resource for genetic engineers to establish more reliable lines of transgenic mosquitoes which could be used for disease vector suppression or replacement. Insulator sequences could be incorporated into a transgene construct, flanking the transgene in order to insulate it from elements causing PE and PEV. If known CTCF binding sites prove to be effective insulators,

mosquito strains could be developed with  $\phi$ C31 *attP* docking sites flanked with CTCF binding sites by randomly inserting *attP*  $\phi$ C31 docking sites with flanking CTCF binding sites at multiple genomic locations via an effective transposable element. This would create strains of mosquitoes that could be transformed using site specific insertions that are insulated at every site.



## CHAPTER II

### IDENTIFICATION OF REGIONS OF CTCF BINDING IN THE *Anopheles gambiae* GENOME

#### 2.1 Introduction

The zinc finger protein, CTCF, has been shown to be associated with repressor, activator, and an insulator functions in human, chicken, mouse, and *D. melanogaster* [42, 51]. The insulator property can be divided into two functions: enhancer blocking and heterochromatin barrier. As explained in Chapter I, an insulator would be advantageous for the production of transgenic mosquito strains by incorporating CTCF binding site sequences into a transgene construct, such that a transgene would be insulated from genomic elements that may cause position effects or position effect variegation. An understanding of the potential activator and repressor functions of mosquito CTCF will enhance the understanding of individual gene and genome wide transcriptional regulation, in disease vector species such as *An. gambiae*. CTCF is known to bind to a variety of DNA sequences [42, 51]. It has been proposed that this is the result of the use of different subsets of zinc fingers to bind to each DNA sequence. Ohlsson *et al.* (2010) [42] has proposed that the binding of a particular set of the zinc fingers results in a particular function; with the function of CTCF being determined by the DNA sequence at the binding site[42].

The first step in understanding the multiple roles CTCF plays in regulating *An. gambiae* gene expression is the identification of CTCF binding sites throughout the genome. The discovery of a functional insulator sequence has the potential to provide insulating sequences for transgene constructs that would maintain stable expression of refractory transgenes that block malaria transmission. In addition to this primary purpose, identification of CTCF binding sequences proximal to genes and gene clusters will facilitate further opportunities for the study

of mosquito gene regulation. The identification of a CTCF binding site map will guide future studies of gene regulation and CTCF function.

Identifying transcription factor binding sites in the *An. gambiae* genome is an ambitious task. To date, the primary tool for identifying global transcription factor binding sites has been Chromatin Immunoprecipitation (ChIP) of transcription factor bound DNA, followed by hybridization to a microarray of known DNA sequences covering much of the subject genome. This technique, also referred to as ChIP-on-chip, has been used to identify CTCF binding sites throughout the *Drosophila melanogaster* genome [45, 52]. Genomic DNA microarrays for *An. gambiae* are available; however, regulatory regions are not well represented. Impetus for the development of a microarray that includes the intergenic sequences necessary for the global analysis of transcriptional regulation has been lacking due to the challenges of obtaining sufficient DNA oligonucleotides from the incompletely sequenced genome, as well as the relatively limited number of researchers that would find use for such an array.

In recent years, the advent of next generation sequencing has provided a means for wider coverage of the genome. Chromatin Immunoprecipitation followed by parallel sequencing (ChIP-Seq) allows for a significant number of the immunoprecipitated DNA fragments to be sequenced in parallel. This approach also eliminates the bias associated with hybridization to a microarray with a limited number of sequences. ChIP-Seq provides an increased base pair resolution, suffers from less noise, provides greater coverage, and has the ability to capture heterochromatin and microsatellites, which are abundant in this genome and may include regions of CTCF binding. These advantages have led to the increasing use of ChIP-Seq for genome wide transcription factor and histone modification mapping studies in vertebrates [53].

Mosquito genes important for development, immunity, blood feeding, and sex differentiation are of particular interest for the purpose of implementing novel strategies for the

control of pathogen transmission. The bithorax complex in *Drosophila melanogaster* has been well studied, revealing seven frontal abdominal insulator sequences. Among these sequences, six have been shown to be bound by or predicted to be bound by CTCF, with *Fab 7* being the only exception [45]. Three more CTCF binding sites have also been identified in the region [45]. *Frontal abdominal 8 (Fab 8)* has been shown to have enhancer blocking activity [43] and has been implicated in the transcriptional regulation of genes within the bithorax complex. An understanding of CTCF's role in the transcriptional regulation of the *An. gambiae* bithorax complex may lead to an improved understanding of the gene networks governing mosquito development. An understanding of the regulation of immunity genes may lead to the use of the mosquito's own immune system to combat pathogen infection and thus reduce transmission. Genes which are important for blood feeding, such as some heme- peroxidase genes which produce anticoagulants and vasodilators, are important for increasing host susceptibility to pathogen infection and could similarly be manipulated for disease transmission control [25, 54]. Sex differentiation genes could be manipulated to improve efficiency of the sterile insect technique [15]. A genome-wide CTCF binding map would be useful for the study of these important systems in order to potentially take advantage of them for pathogen or vector control purposes.

## **2.2 Materials and Methods**

### *2.2.1 Identification of potential CTCF binding sites in silico*

Dr. Igor Sharakhov (Virginia Tech University) provided images from polytene chromosome spreads mapped with BAC clones and bound with a polyclonal antibody generated against the C-terminal region of *An. gambiae* CTCF [50], which were used to identify regions of CTCF binding. Dr. Sharakhov's lab also provided nucleotide coordinates for the euchromatic and heterochromatic regions of each chromosome. Fluorescent *in situ* hybridization (FISH) was

used to hybridize and localize the PCR products of gene fragments believed to be located near euchromatin/heterochromatin boundaries, thus identifying the most accurate boundary coordinates possible [55]. This data was used to identify heterochromatin/euchromatin transition zones.

DNA sequences identified from the above data and a position specific scoring matrix for the *Drosophila* CTCF consensus published by Holohan *et al.* (2007) [45] were input into Patser [56] to identify potential CTCF binding sites in the *Anopheles gambiae* genome. DNA sequences identified as similar to the consensus with Patser scores greater than 11 were manually analyzed for similarity with the *Drosophila* CTCF consensus published by Bartkuhn *et al.* (2009) [52]. The DNA sequences with the highest similarity were selected for validation using ChIP-PCR by designing primers flanking putative binding sites and allowing the amplification of 200 to 400 base pair regions. Chromatin Immunoprecipitation was performed using the Upstate ChIP Assay Kit (Temecula, CA). The targeted amplicons were amplified using the following PCR amplification conditions; step 1: 3 minutes at 94°C, step 2: 15 seconds at 94°C, step 3: 30 seconds at the determined annealing temperature, step 4: 1 minute at 72°C, step 5: steps 2-4 repeated 39 times, step 6: 5 minutes at 72°C, step 7: held at 4°C.

#### 2.2.2 Anti-CTCF ChIP using Sua 4 cultured cells

Approximately  $1 \times 10^6$  Sua 4 (neonate larval) cells were cultured in a 25cm<sup>2</sup> cell culture flask with 10 ml of Schneider's media with 20% fetal bovine serum containing antibiotics and fungizone. Proteins were crosslinked to genomic DNA by adding 1% formaldehyde to each flask of cells and incubating the cells at 37°C for 10 minutes. The cells were then washed twice with ice cold PBS containing protease inhibitors and scraped into a conical tube. The cells were then centrifuged and the supernatant was removed. The cells were resuspended in lysis buffer and incubated on ice for ten minutes. The chromatin was sheared by sonication into fragments of 100

to 400 base pairs. After sonication, 1% of the sample (20µl) was collected to be used as a control and checked for shearing efficiency. The control sample was incubated for 4 hours with 1µl of 5M NaCl to reverse the crosslinks. DNA was purified and extracted using phenol/chloroform extraction and ethanol precipitation. The DNA was loaded onto a 1.5% gel and run at 96 volts for 30 minutes. Visualization of the DNA following ethidium bromide staining of the gel indicated that the DNA fragments were between 100 and 200 base pairs.

The remainder of the sonicated sample was pre-cleared with 75 µl of Protein A Agarose/Salmon sperm DNA for 30 minutes at 4°C with agitation. The Protein A Agarose/Salmon sperm DNA beads were pelleted by centrifugation and the supernatant was collected and immunoprecipitated with 6µl of *An. gambiae* CTCF antiserum (1:300 dilution) overnight at 4°C with rotation. The immunoprecipitated sample was incubated with 60 µl of Protein A agarose/Salmon sperm DNA with rotation for 1 hour. The agarose beads were pelleted by centrifugation. The supernatant was removed and the protein A agarose/antibody/histone complex was washed for 5 minutes with each of the following reagents, in the following order: Low Salt Immune Complex Wash Buffer, High Salt Immune Complex Wash Buffer, LiCl Wash Buffer, and TE Buffer (two washes)(Millipore, Temecula, CA). The DNA/protein complex was eluted by washing the agarose bead pellet twice in 250 µl of elution buffer (1% SDS, 0.1M NaCO<sub>3</sub>) for 15 minutes. The crosslinks of the combined eluates were then reversed by incubating the sample with 20 µl 5M NaCl for 4 hours at 65°C. Proteinase K digestion was performed to eliminate any remaining antibody and chromatin proteins. DNA was extracted and purified by phenol/chloroform extraction and ethanol precipitation.

### 2.2.3 ChIP-Seq library preparation

The immunoprecipitated DNA was used to prepare a ChIP-Seq library using the Illumina ChIP-Seq DNA Sample Prep kit (SanDiego, CA). Briefly, end repair of the DNA fragments was performed with T4 DNA ligase and dNTPs, followed by the addition of adenine bases to the repaired fragment ends. Adapters were then ligated to the DNA fragments. PCR amplification was performed with adapter specific primers and the library was run on a 2% gel at 120V for 1 hour. A 200 to 300 base pair sized fragment was excised from the gel and the DNA was gel extracted using the Qiagen gel extraction kit (Qiagen Sciences, Maryland).

### 2.2.4 Real-time PCR for validation of the ChIP-Seq library

A real-time PCR experiment was performed using two samples each with the same mass of DNA template. One sample consisted of DNA from the ChIP-Seq library prepared from the CTCF immunoprecipitated DNA and the other consisted of input DNA prepared with adapters. Both samples were amplified with Sybr Green (Thermo Scientific) and primers flanking the 2R EuHet fragment, which was validated for CTCF binding using a gel based PCR assay. Each sample was assayed in triplicate on a real-time PCR plate using the Applied Biosystems 7300. The following PCR protocol was used: 3 minutes at 95°C; 40 cycles of 30 seconds at 95°C followed by 30 seconds at 60°C; 30 seconds for each of the following temperatures: 70°C, 75°C, 80°C, 85°C, 90°C, 95°C; held at 4°C [57].

### 2.2.5 Sequencing and analysis of the CTCF immunoprecipitated ChIP-Seq library

Sequencing of the prepared library was performed at the IGSP Genome Sequencing and Analysis Core Resource at Duke University in Durham, NC using Illumina GAIIX sequencing technology. The sequence data was modified using various tools in the Galaxy suite [11, 58, 59]. FASTQ Groomer[60] was used to format the data and the data was aligned to the ensembl AgamPEST3 [3, 61] version of the *Anopheles gambiae* genome using Bowtie[62].

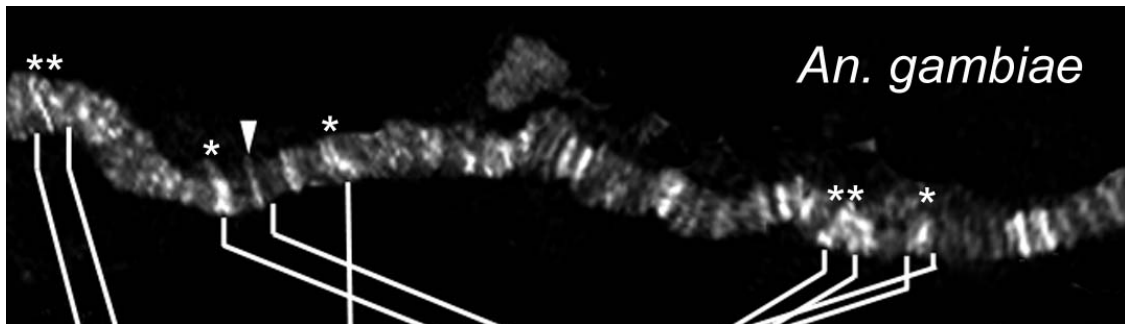
The control library was prepared from input DNA from the same Sua 4 cell line using the same extraction and purification protocol as described above, without the immunoprecipitation step. The input DNA was sent to the IGSP Genome Sequencing and Analysis Core Resource at Duke University in Durham, NC for Illumina ChIP-Seq library preparation and Illumina GAIIx sequencing. After alignment to the AgamPEST3 [3, 61] version of the genome using Bowtie [62], Model-based Analysis for ChIP-Seq (MACS)[63] was used to identify peaks of sequence enrichment from the CTCF antibody immunoprecipitated library throughout the genome, using the input library as a background control, as described in [63]. Throughout this work these peaks will be referred to as CTCF binding site peaks. It is important to note that these CTCF binding site peaks may contain one or multiple CTCF binding sites within the given coordinates identified by MACS [63]. For parameters used to run MACS, see supplementary material. Annotation of the MACS data was performed with annotationPeaks from the Homer suite [44] in conjunction with the ensemble AgamPEST3 [3, 61] version of the *An. gambiae* genome.

## **2.3 Results**

### *2.3.1 In silico identification and validation of a CTCF binding site*

Dr. Igor Sharakhov's lab (Virginia Tech University) provided images of polytene chromosome spreads obtained from *An. gambiae* ovarian nurse cells bound with a polyclonal antibody generated against the C-terminal region of *An. gambiae* CTCF [50]. Figure 4 is an image of part of the 2R chromosomal arm treated with this antibody. This procedure identified CTCF binding regions along at least two of the chromosome arms. Chromosome arms 2L and 2R were hybridized with BAC clones as reference points to identify the chromosomal positions of the CTCF binding regions distributed along the chromosome arms. Regions of fluorescence along the chromosomes indicated regions of CTCF antibody accumulation, identifying regions

of CTCF binding with a resolution of approximately one megabase. Coordinates for the euchromatic and heterochromatic regions of *An. gambiae* ovarian nurse cell chromosomes, provided by the Sharakhov lab were used to identify euchromatin/heterochromatin transition zones. The chromosome spreads treated with the CTCF antibody indicate that euchromatin/heterochromatin transition zones may be regions of enriched CTCF binding.



**Figure 4: Chromosome 2R from an *An. gambiae* ovarian nurse cell stained with the *An. gambiae* CTCF antibody.** CTCF binding regions are indicated by white bands. Asterisks indicate regions that are syntenic with CTCF binding regions in *An. stephensi*. The arrow head indicates a CTCF binding region unique to *An. gambiae*.

To identify potential CTCF binding sites, we used a position specific scoring matrix (PSSM) for the *Drosophila* consensus, as published by Holohan *et al.* [45] based on their *Drosophila* CTCF microarray data. The PSSM was input into the bioinformatics perl program, Patser [56], with sequences corresponding to regions of the chromosomes identified as CTCF binding regions via the CTCF antibody-stained chromosome spreads, or euchromatin/heterochromatin border regions identified by the Sharakhov lab[44]. The Patser output showed multiple DNA sequences with similarity to the *Drosophila* consensus, with scores corresponding to their level of similarity.



To identify the output sequences that were most likely to be actual sites of CTCF binding, a lower threshold Patser score was established by entering sequences from the *Drosophila* genome that contained CTCF binding sites that were experimentally identified by microarray analysis [45] into Patser with the *Drosophila* CTCF consensus PSSM [45], and identifying the sequences in the Patser output that corresponded to the experimentally identified sequences. These sequences had a Patser score of 11 or higher. Sequences with scores lower than 11 did not correspond to experimentally identified CTCF binding sites. Therefore, a score of 11 was used as a minimum threshold for output sequences to be considered DNA sequences with likely similarity to the *Drosophila* CTCF binding consensus. Each one of the output sequences with a score of at least 11 was compared to the *Drosophila* CTCF binding consensus published by Bartkuhn *et al.* (2009) [52]. This consensus was used based on its increased accuracy due to it being generated from a data set of 300 experimentally identified CTCF binding sites, compared to only 33 for the Holohan consensus. Those output sequences with the highest similarity to the consensus were considered potential CTCF binding sites and were tested for confirmation by ChIP-PCR.

One potential binding site was identified on chromosome 2R within a euchromatic region, near a euchromatin/heterochromatin border at base pair position 59,016,524. This potential binding site was validated with ChIP-PCR using primers that flanked the potential binding site, producing a 200 base pair amplicon (Figure 5). Input DNA was amplified to be used as a positive control, DNA immunoprecipitated with normal rabbit serum was used as a negative control reaction and a no template control was also utilized. The resulting gel, shown in figure 5, clearly showed PCR products from the input and CTCF immunoprecipitation reactions. No PCR products were generated from the normal rabbit serum immunoprecipitation and the no

template control reactions. This validated potential binding site was used as a positive control for ChIP-Seq library preparation (see Materials and Methods).



**Figure 5: ChIP-PCR result for potential CTCF binding region found on chromosome 2R at position 59,016,524 bp.** Input is amplification using non-immunoprecipitated DNA as template. CTCF IP: Immunoprecipitated with CTCF antiserum. NRS IP: Immunoprecipitated with normal rabbit serum. NTC: no template control

### 2.3.2 CTCF binding site peaks identified via ChIP-Seq are over-represented near genes

The annotationPeaks tool from Homer [44] used the MACS [63] data and the ensembl AgamPEST3 [3, 61] reference genome to identify whether CTCF binding site peaks were found within an intron, promoter, exon, transcription termination site (TTS), 5'UTR, 3'UTR, or intergenic region. Of the entire data set, 51% are within intergenic regions, and 28% of the intergenic peaks are within 10kb upstream of a promoter, and 14% of them are within 10kb downstream of a promoter. The large amount of intergenic sequence in the genome could explain this result by random distribution. However, given that 42% of CTCF binding site peaks are located in close proximity to genes and gene clusters, as well as the gene insulating property of CTCF identified in other organisms, it is likely that CTCF is acting as an insulator at many of these intergenic positions. Further experimental analysis is necessary to determine the function of the CTCF binding site peaks at intergenic positions.

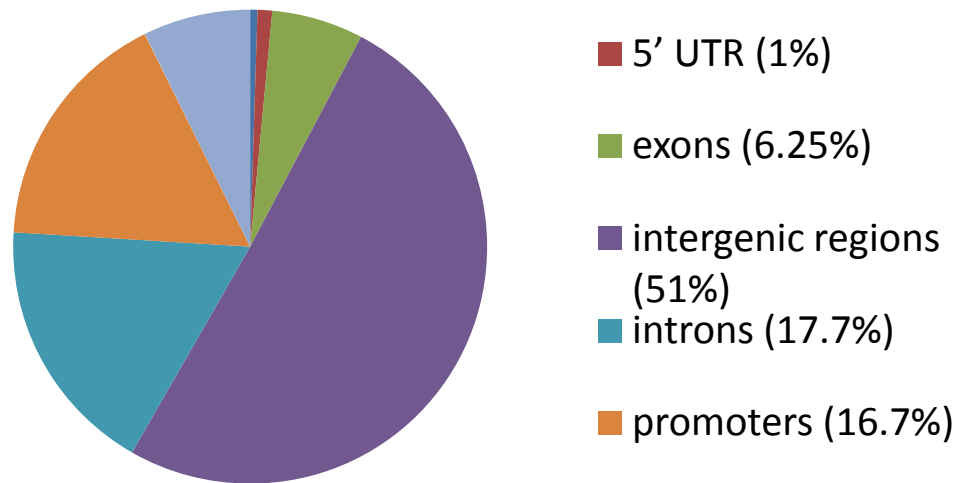
Another interesting result is the presence of 16.7% of the identified CTCF binding site peaks within the promoter region of a gene, defined by the Homer software, annotationPeaks, as the range of 1000 base pairs upstream of the transcription start site to 1000 base pairs downstream of the transcription start site [64]. This is consistent with earlier evidence of the transcriptional activating and repressing roles of CTCF. The identification of the functional roles of these identified CTCF binding site peaks will require further study.

The Homer annotation also revealed that 17.7% of the CTCF binding site peaks in the data set were found within introns, 6.25% of the peaks were found within exons, and 7.3% were found at transcription termination sites (TTSs). Shukla *et al.* [65] provide an explanation as to how intragenic regulation by CTCF may occur, the details of which are presented in the discussion section. RNAi depletion of CTCF confirmed that CTCF binding to an exon was responsible for alternative splicing of the human gene, *CD45*, resulting in the inclusion of the exon [65]. Other models have been put forth, suggesting that insulators bound by CTCF within the introns of genes are responsible for repression of these genes. It has been shown that methylation of a CTCF binding site within an intron inhibits CTCF binding and leads to increased transcription of the gene. This is the case for the human oncogene *BCL6*, which is upregulated due to aberrant methylation in the first intron, leading to a lack of CTCF binding at this locus in lymphoma cells [66]. This supports the notion that a CTCF binding site at an intron may act as a repressor.

The results of this ChIP-Seq study show that the majority of CTCF binding sites in the *An. gambiae* genome are situated near or within genes. Based on the amount of intergenic sequence within the genome, a random distribution of CTCF binding sites should result in close to 90% of the peaks being intergenic. This value was estimated using data from the AgamPest3 genome on vectorbase.org [3, 67, 68]. The average length of an *An. gambiae* gene was estimated

by comparing the longest and shortest genes in the genome. This average length was estimated on the high end in order to err on the side of less intergenic sequence. The average gene length was multiplied by the number of genes in the AgamPest3 genome. This number of base pairs represents the amount of the genome that is gene sequence (10% of the genome). This was subtracted from the total number of base pairs in the genome to estimate the amount of intergenic sequence in the genome. The number of intergenic peaks is well below this random expectation, at 51%. Also, 17.7% of the CTCF binding site peaks are found within introns and 16.7% are found within promoters. Thus, nearly half of the potential CTCF binding sites are within genes, similar to the binding pattern of other transcription factors, such as CREB1 [69]. Figure 6 illustrates the frequency of CTCF at the six genic contexts defined by the Homer annotationPeaks software [64]. This suggests that CTCF may be necessary for maintaining appropriate transcript levels of many genes throughout the *An. gambiae* genome, as has been shown in other organisms [42]. It is important to note that percentages of CTCF binding sites mapped to their relative genic positions in the *An. gambiae* genome are similar to those in the human genome, which has 46% mapped to intergenic regions, 20% mapped to promoters, 12% mapped to exons, and 22% mapped to introns [42]. According to microarray data, *D. melanogaster* has a similar distribution of CTCF binding sites [52]. Thus, CTCF appears to be conserved at least in regard to its binding site positions relative to genes throughout the genome from humans to dipterans. Further studies will be needed to determine whether this conservation extends to the multiple functional roles of CTCF throughout the genome.

## Distribution of CTCF ChIP-Seq peaks in the *Anopheles gambiae* genome



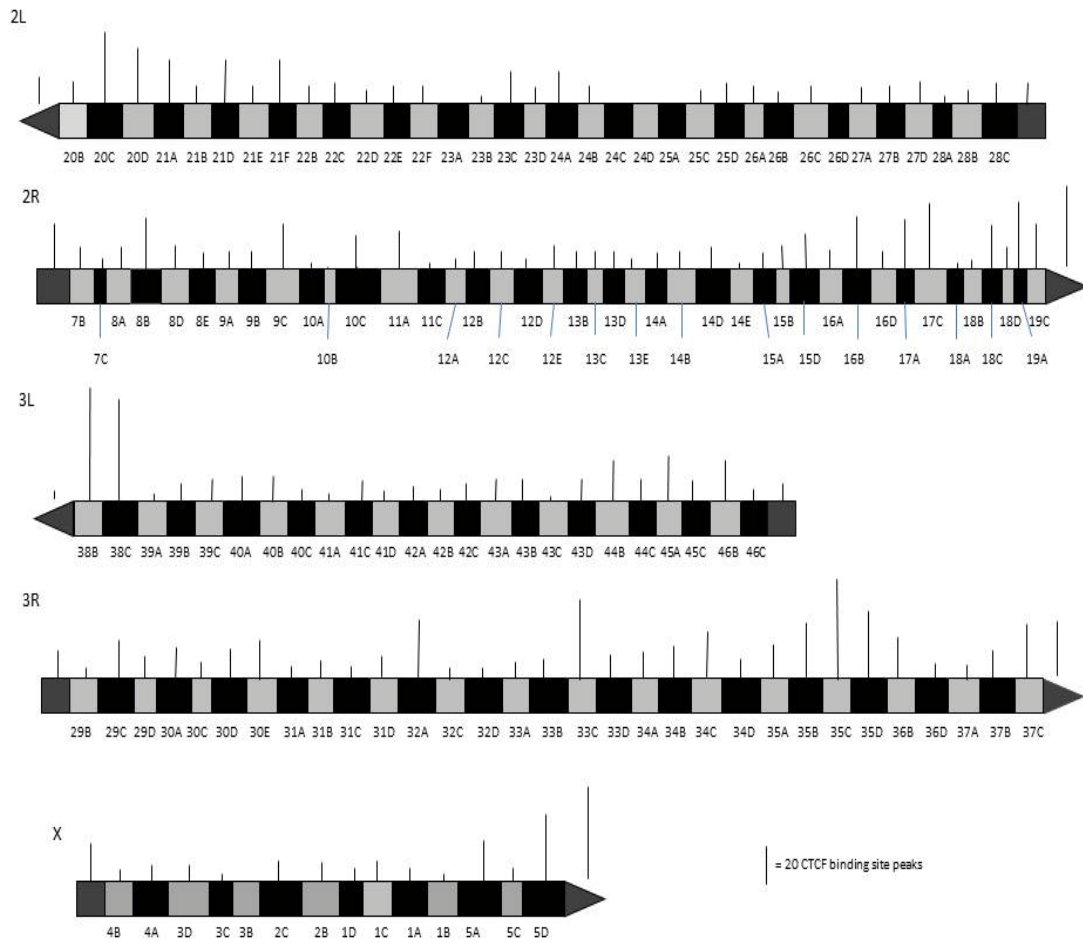
**Figure 6: Distribution of CTCF ChIP-Seq peaks in the *An. gambiae* genome.** 5'UTR and 3' UTR represent peaks located within the 5' and 3' untranslated regions of annotated genes, exons represent peaks located within coding regions of annotated genes, excluding promoter sequences. Intergenic regions represent peaks located outside annotated genes, also excluding promoter sequences. Introns represent peaks located within non-coding DNA between exons. Promoters represent peaks found within 1000 bp upstream of an annotated transcription start site and 100 bp downstream of a transcription start site. TTSs represent peaks located at a transcription termination site

### 2.3.3 CTCF binding site chromosome map

The CTCF binding site ChIP-Seq data was mapped to individual chromosome arms according to the chromosome band coordinates for the AgamP3 genome available on vectorbase.org [3, 67, 68]. Some of the chromosome bands are not represented on vectorbase.org. According to vectorbase.org personnel (personal communication), all of the assembled sequence is present; however, sequences that would be designated as part of the

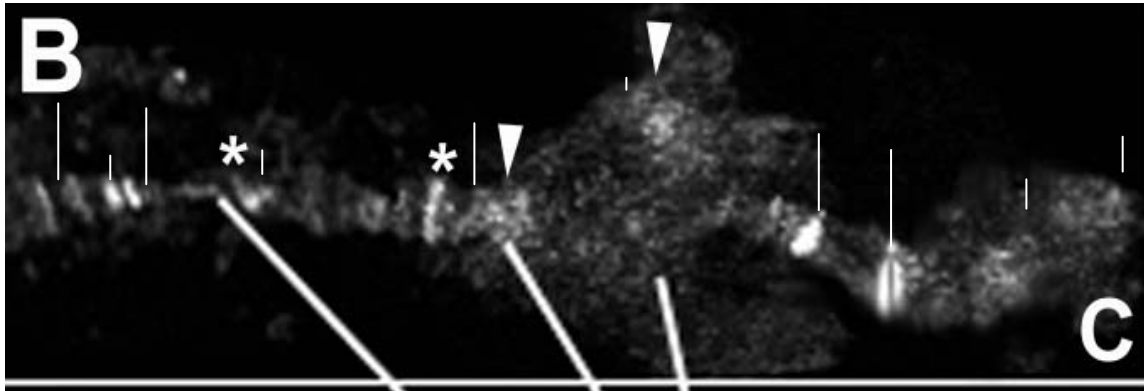
missing chromosome band have been designated as part of a neighboring band. Therefore, some of the peaks assigned to specific chromosome bands in this work may actually be present in other chromosome bands not identified on vectorbase.org. Figure 7 shows all five chromosome arms with their respective bands as identified on vectorbase.org with lines of varying lengths indicating the relative number of CTCF binding site peaks identified within the corresponding chromosome band.

The CTCF immunostained chromosome spreads were compared to the chromosome map. This comparison showed some correlation between regions of CTCF binding identified on the chromosome spreads with the ChIP-Seq data identified along the chromosomes. Figure 8 shows the chromosome spreads with lines of varying length along them indicating the number of CTCF binding site peaks for the vectorbase.org chromosome band corresponding to those particular chromosomal regions. The intensity of the immunostaining signal does not always correlate with the number of ChIP-Seq peaks at a particular chromosome band. This may be caused by differences in CTCF affinity among the different sites, resulting in lower signal intensity on the chromosome spreads. The different cell types used in these experiments may also explain these differences [70]. The chromosome spreading procedure can also affect the resolution of the immunostaining signals along the chromosomes. However, some chromosome bands showed signals of CTCF binding on the chromosome spreads correlating with the number of CTCF binding site peaks. These included 8A, 8B, 8E (at BAC clone 12\_G10), 9A, 9C, 11A, 11C, 12B, 12C, and 12E on chromosome 2R, and 20A (designated as the centromere on vectorbase.org), 20C, 21B (at BAC clone 02A19), and 21D on chromosome 2L.

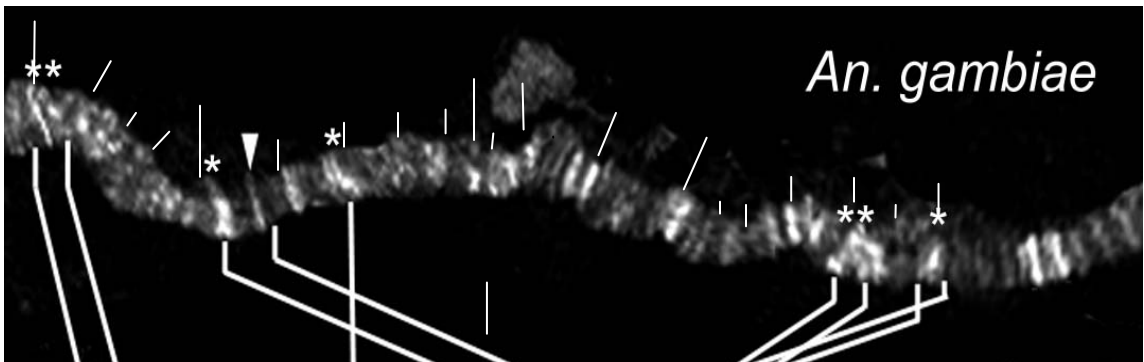


**Figure 7: Distribution of CTCF binding site peaks along the five *An. gambiae* chromosome arms.** Varying lengths of lines for the chromosome bands as identified on vectorbase.org indicate the relative abundance of identified CTCF binding site peaks for each chromosome band.

A)



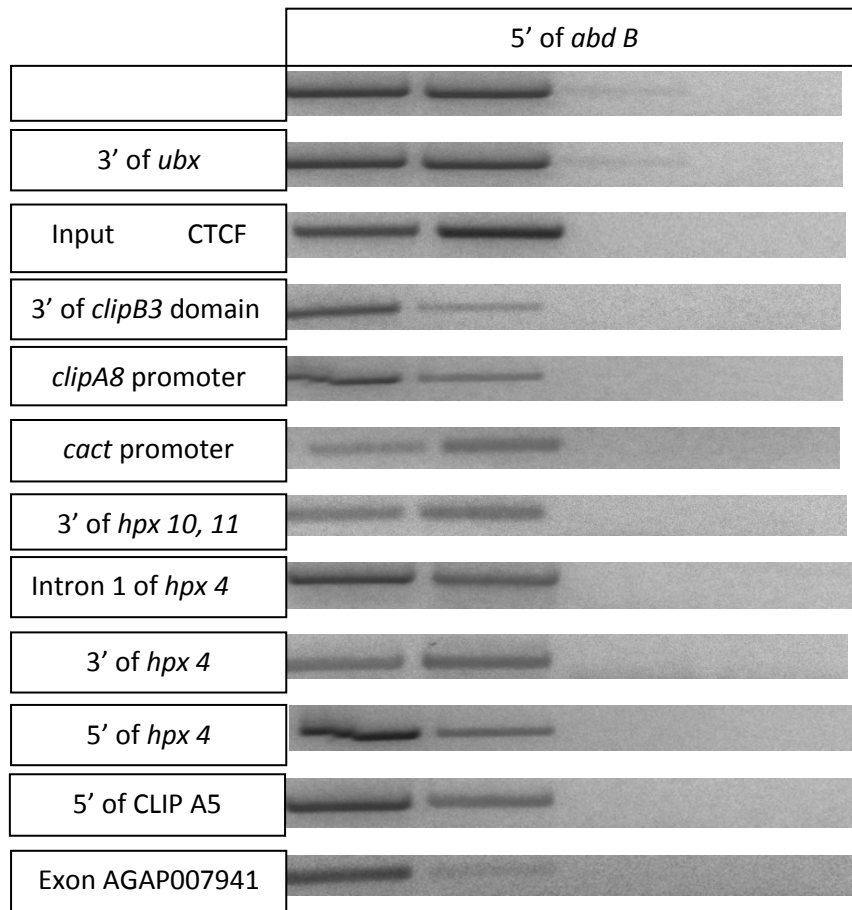
B)



**Figure 8: Distribution of CTCF binding site peaks at chromosome bands compared to chromosomes immunostained with the CTCF antibody.** A) Chromosome 2L chromosome spread. B) Chromosome 2R chromosome spread. White immunostained bands indicate regions of CTCF binding. White vertical lines of varying lengths indicate the relative abundance of CTCF binding site peaks at each chromosome band at that location.



**Table 2: ChIP-Seq identified CTCF binding sites validated by ChIP-PCR.** The left column identifies the genomic location of the CTCF binding site peaks relative to nearby genes of significance. Top: The template used for the individual PCR reactions. Input refers to non-immunoprecipitated DNA. CTCF IP refers to DNA immunoprecipitated using the antibody raised against *An. gambiae* CTCF. NRS IP refers to DNA that was Immunoprecipitated using normal rabbit serum. NTC refers to no template control



#### 2.3.4 Some immune response genes may be regulated by CTCF

CTCF binding sites proximal to some immune response genes were selected for validation (see Table 2) because of the possibility that CTCF may play a role in the regulation of these genes. An improved understanding of the regulation of immune response genes will aid in the development of a malaria transmission control program. The CLIP genes are a family of

immune response genes that regulate melanization. One cluster of eight CLIP genes is between two CTCF binding site peaks 60 kb apart. CLIPB3 and CLIPB4 have been shown to be necessary for ookinete melanization. Two other important CLIP genes, CLIP A2 and CLIP A5, are found within a gene cluster with ten other CLIP genes between two CTCF binding site peaks, 89.5 kb apart. CLIPA2 and CLIPA5 are known to block ookinete melanization. In both cases, the flanking positions of the CTCF binding site peaks, relative to the CLIP gene clusters, suggest an insulating role for CTCF at these genomic locations based upon results in other organisms that indicate that genes and clusters of genes are insulated by flanking insulator sequences, such as the genes of the bithorax complex in *D. melanogaster* [71].

CTCF binding site peaks are also located at the promoter of two CLIP genes. The CTCF binding site peak associated with the CLIPB17 promoter is isolated with no other CTCF binding site peaks within 500 kb in the upstream direction and nearly 400 kb in the downstream direction. In contrast, the CTCF binding site peak at the CLIPA8 promoter is accompanied by another CTCF binding site peak located just 4.4 kb upstream of the promoter.

Another gene important for immunity is *cact* (*cactus*), which has a CTCF binding site peak at its promoter. Just downstream of *cact*, another CTCF binding site peak is present within exon 2 of the novel gene, AGAP007941. Although the function of this gene is unknown, it presents an opportunity to study the possible role of CTCF as an intragenic regulator.

### 2.3.5 CTCF binds near some heme-peroxidase genes

Heme-peroxidase genes have been shown to be associated with immunity and blood feeding in mosquito species. Therefore, further understanding of the transcriptional regulation of these genes may be useful for malaria transmission control. With this in mind, a subset of CTCF binding sites near some Heme-peroxidase genes were validated by ChIP-PCR, as shown in Table 2. An Immunomodulatory peroxidase (IMPer), identified as HPX15 in the *An. gambiae* genome,

has been shown to combine with dual oxidase to protect the midgut lumen from immune responses which allows for the survival of commensal bacteria, consequently allowing *Plasmodium* ookinetes to survive in the midgut [72]. HPX 15 is located within a 337.8 kb segment of the genome between two CTCF binding site peaks with HPX 14 and 10 novel genes. Peroxidase activity has also been observed in female *Anopheles albimanus* mosquitoes before blood feeding and has been shown to be associated with the inactivation of vasoconstricting substances necessary to form hemostats. [25]. Some Heme-peroxidase genes in the *An. gambiae* genome have CTCF binding site peaks in close proximity. Further study of the function of CTCF, as well as the identification of other proteins in complexes at these CTCF binding regions may provide more insight regarding the regulation of genes such as HPX15 and others that may be important for blood feeding.

Heme-peroxidase 6 (HPX 6) is positioned within a 53.67 kb region between two CTCF binding site peaks that also contains 6 novel genes. The positioning of the peaks suggests that they may insulate the genes within the intervening sequence. According to CTCF looping models previously described in Chapter I [42], the sequence between the two CTCF binding site peaks may form a loop that is isolated from inappropriate enhancer or silencer interactions. This would maintain HPX 6 and the other genes within the 53.67 kb region at their appropriate expression levels. HPX 16 and HPX 2 are also located within similar regions between flanking CTCF binding site peaks. HPX 10 and HPX 11 are located within another region between CTCF binding site peaks, which they share with several cuticular proteins. In these cases the genomic region is no larger than 136 kb and no smaller than 10 kb. The looping model says that in order for a loop to be formed, a genomic region needs to be at least 10 kb in length [42]. However, looping is not the only possible mechanism for insulating genes from enhancers. Single insulator

sequences have also been shown to effectively block the activity of genomic elements that may alter gene transcription [38].

Another heme-peroxidase gene, HPX 4, has a CTCF binding site positioned within the intron of an alternative splice variant and 22.2kb upstream of the other splice variant. A second CTCF binding site peak is 1.7kb downstream of both splice variants. The CTCF binding site positioned at the first intron of the longer variant may be responsible for distinguishing between the two variants, as is the case mentioned previously in which CTCF has been shown to be necessary for distinguishing between splice variants depending on whether or not the binding site is bound by CTCF [65]. The short variant does not begin to be transcribed until the sixth exon and has one extra intron. Further study will be necessary to determine if and how CTCF distinguishes between these two splice variants. This presents an opportunity for the study of the possible intragenic regulatory function of *An. gambiae* CTCF.

### 2.3.6 CTCF and sex differentiation genes

CTCF binding sites were validated and or analyzed near sex differentiation genes due to the possibility that they may play a role in the transcriptional regulation of such genes. Understanding the transcription of sex differentiation genes would be useful for improving the efficiency of the sterile insect technique. Doublesex is responsible for the normal expression of secondary sexual characteristics in *Drosophila* [73]. The CTCF binding site peak data was compared with a consensus sequence for the *Aedes aegypti* doublesex binding site, provided by Dr. Helen Benes (University of Arkansas, Little Rock), using Patser [56]. Doublesex has well conserved DNA binding sites across dipteran species [73]. Patser identified 97 sequences among the 2,416 CTCF binding site peak sequences (~4% of total peaks), that are similar to the position specific scoring matrix (PSSM) of the *Aedes aegypti* doublesex binding motif.

Male sex lethal 2 (MSL-2) is the protein necessary for dosage compensation in male *Drosophila*. Dosage compensation is accomplished in male *Drosophila* by doubling the amount of mRNA produced from the X chromosome. In females, ectopic MSL-2 expression results in the doubling of mRNA on both X chromosomes and thus results in lethality. MSL-2 expression in females is regulated by the Sex-lethal protein, which binds to the 5' and 3' UTR of *msl-2* mRNA to inhibit its translation.

The *An. gambiae msl-2* ortholog is found within a 34.8 kb region flanked by two CTCF binding site peaks, suggesting a possible insulator function for these two binding sites. The *An. gambiae* ortholog for *sex-lethal* has a CTCF binding site peak at the promoter of its shorter splice variant, which is within the first intron of the longer splice variant. The gene is also positioned near the middle of a 58.9 kb sequence flanked by two CTCF binding site peaks. The positioning of these peaks suggests an insulator function for the two flanking peaks. The peak found at the promoter/intron of the two splice variants suggests a possible activator/repressor function or intragenic regulation role, perhaps in directing alternative splicing events. This is another candidate region for the further study of *An. gambiae* CTCF function.

### 2.3.7 CTCF binding site peaks at the *Anopheles gambiae* bithorax complex

The bithorax complex in insects is important for determining the insect body plan during development. Expression levels of specific genes must be maintained at specific embryonic locations and developmental time points to guarantee correct development of each body segment. At the *Drosophila* bithorax complex, CTCF has been shown to play an important role in insulating specific bithorax complex transcriptional regulatory regions from one another. CTCF binding sites have been identified at nine different locations within the *Drosophila* bithorax complex [45]. Two are between *ultrabithorax (ubx)* and *abdominal A (abd-A)*, the binding site in the 5' direction coincides with the predicted insulator known as *frontal abdominal*

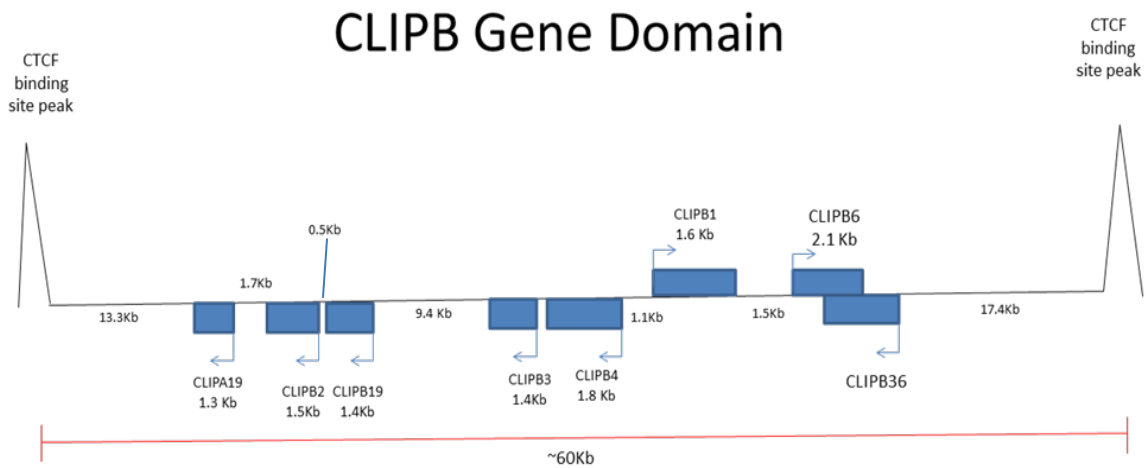
2 (Fab-2). The binding site that is downstream of Fab-2 has the designation of A. Five more are between *abd-A* and *abdominal B (abd-B)* and correspond to the predicted insulators Fab-3 and Fab-4, as well as to the genetically identified insulators *Mcp (Miscadastral pigmentation)*, *Fab-6*, and *Fab-8* insulators [45]. Two more binding sites are upstream of *abd-B* and have been designated as B and C. *Fab-8* has been shown to be an important boundary between the *Abd-B* transcriptional regulatory regions of *iab-7* and *iab-8*. The *iab-7* and *iab-8* regulatory sequences initiate and maintain the specific expression patterns of *Abd-B* for parasegments 12 and 13 respectively. Removal of the *Fab-8* sequence results in the fusion of these transcriptional regulatory regions, resulting in the loss of the parasegment 12 expression pattern, which is replaced by Parasegment 11 and parasegment 13 expression patterns in the *Drosophila* embryo [74]. Not only does CTCF bind to the *Fab-8* sequence, it has been shown to be necessary for maintenance of the enhancer blocking activity of the *Fab-8* barrier between the *iab-7* and *iab-8* transcriptional regulatory regions [43, 46].

With this in mind, we looked at the CTCF binding site peaks in the *An. gambiae* bithorax complex. Figure 9E illustrates the *An. gambiae* bithorax complex with black peaks representing CTCF binding site peaks, compared with CTCF binding sites previously identified in *Drosophila*, represented with blue peaks. The *An. gambiae* ChIP-Seq data reveal two CTCF binding site peaks flanking a 584 kb region containing the bithorax complex, composed of the orthologs of *ubx*, *abd-A*, and *abd-B*. Three more CTCF binding site peaks were identified between *abd-A* and *abd-B*. Based on their location with respect to the orthologous genes, these peaks appear to correspond to the Fab-3, Fab-4, and *Fab-8* insulators in *D. melanogaster* [45, 52]. The peak upstream of *abd-B* was validated by ChIP-PCR, as shown in Table 2. The peak downstream of *ubx* was unable to be validated due to the repetitive sequences within the peak region. However, primers were able to be designed for a 211 base pair region, 47 base pairs

downstream of the called peak region. This region was shown to bind CTCF using ChIP-PCR, as shown in Table 2. The three peaks identified between the two flanking peaks were unable to be validated by ChIP-PCR due to highly repetitive sequences in the region. These data suggest that the *An. gambiae* bithorax complex has a similar organization to that of *D. melanogaster*.

Unfortunately, the repetitive sequence in the vicinity of these three peaks prevented effective primer design for complete ChIP-PCR validation.

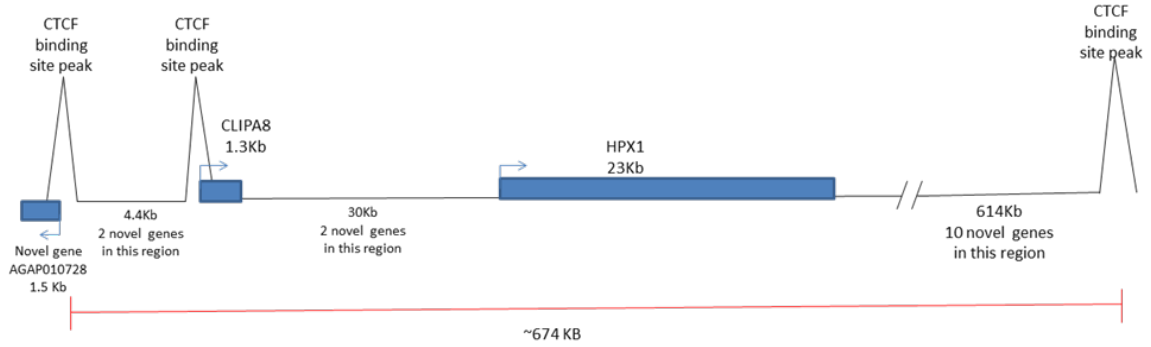
A



**Figure 9: Maps of genomic regions with CTCF binding site peaks in relation to selected genes of interest.** Pointed peaks represent CTCF binding site peaks identified by ChIP-Seq (width of peak is not to scale). The black line represents the DNA fragment under examination. The blue rectangles represent genes along the indicated DNA strand. The length and annotation of each gene are noted above or below the respective rectangles. Other genomic distances refer to distances between genes. Genes on top of the black line are on the forward strand and genes on the bottom of the black line are on the reverse strand. Promoters are indicated by arrows pointing in the direction of transcription. The red line at the bottom indicates the approximate base pair length for the indicated length. Figure 9D shows more detail of the individual transcripts with blue rectangles indicating exons and blue lines indicating introns. The white box at the 5' end of the long transcript indicates the 5' untranslated region. Figure 9E labels *An. gambiae* CTCF binding site peaks and *Drosophila* CTCF binding sites with the homologous *D. melanogaster* insulator identifier. Light blue peaks indicate *Drosophila* CTCF binding sites. Base pair distances refer to distances between genes and CTCF binding site peaks identified in *An. gambiae*. Insulators in quotes refer to predicted insulators. Italicized insulators refer to insulators that have been genetically identified.

B

## CLIPA8 and HPX Gene Domain



C

## Sex-lethal genomic region

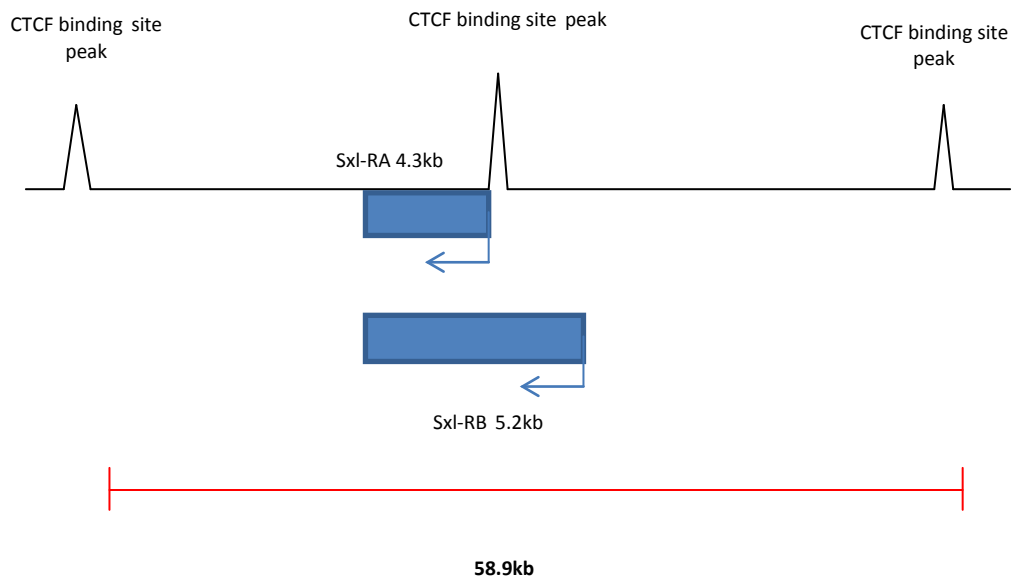
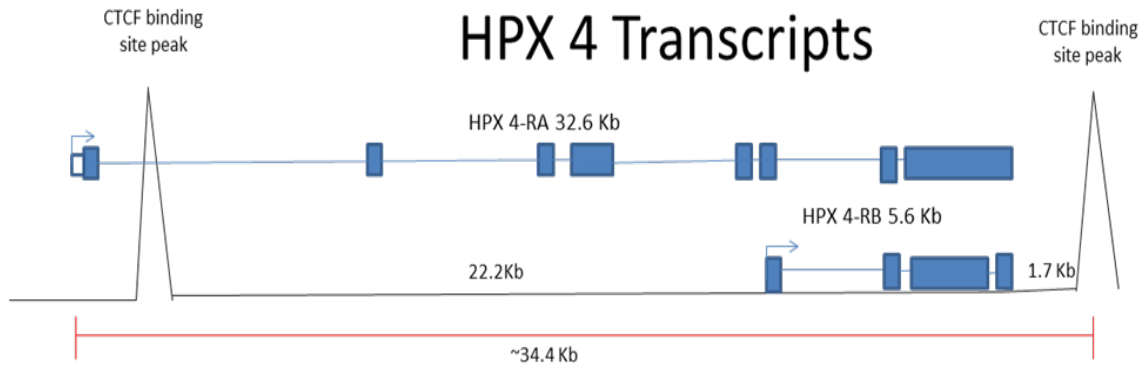


Figure 9 continued



D



E

*Anopheles gambiae* Bithorax complex

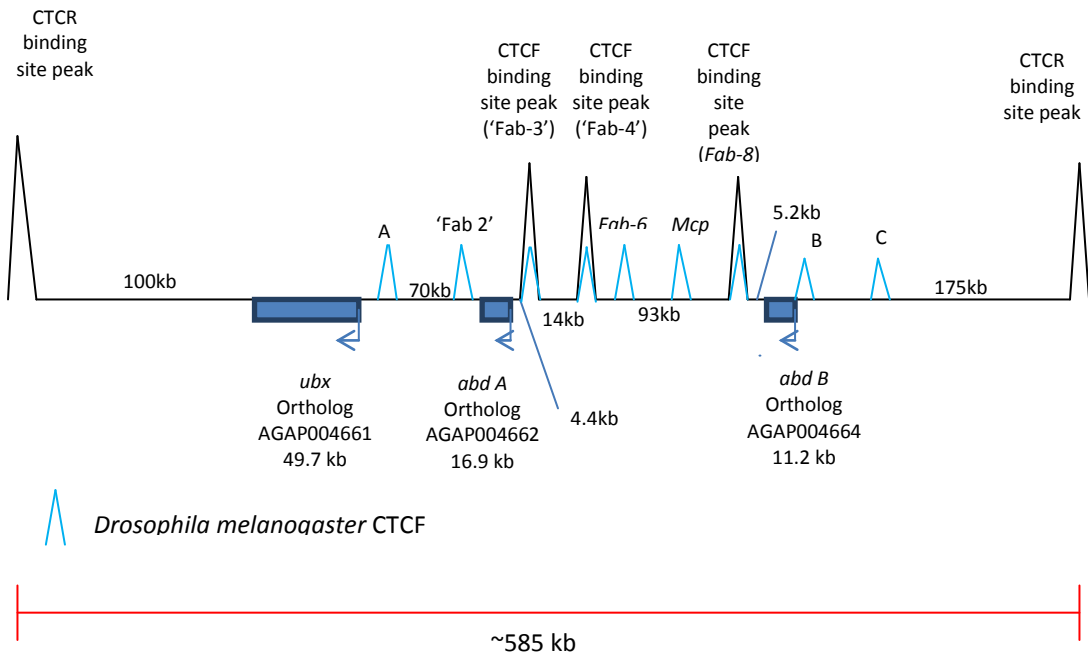


Figure 9 continued

### *2.3.8 Four sequence motifs identified among a subset of the CTCF binding site peaks*

The following criteria were used to select a subset of CTCF binding site peak sequences for analysis to identify a CTCF binding site consensus: (1) the CTCF binding site peak sequences corresponded to a chromosome band that showed enrichment for CTCF on the CTCF immunostained chromosome spreads, and/or (2) were validated for CTCF binding by ChIP-PCR. There was only one biological replicate of CTCF ChIP-Seq data; therefore, the CTCF immunostained chromosome spread data and the ChIP-PCR validated sequences provided an additional biological replicate of a set of sequences representing CTCF binding sites with high confidence. Two-hundred-twelve peaks met these criteria, these were analyzed using the motif finding tool, AlignACE [75] from the Tmod software suite (Toolbox of motif discovery) [76], which uses a Gibbs sampling algorithm to identify motifs from multiple sequences through alignment of similar sequences. Fifty-four motifs were identified from the data set. Four of these motifs, motif 17, motif 24, motif 27, and motif 29 were relatively conserved across at least 3 base pair positions with relatively low repetitiveness and represented 24%, 19.3%, 25%, and 19.8% of the 212 sequences input into AlignACE, respectively. After identifying and subtracting duplicate sequences among the four motifs, it was shown that these four motifs are represented in approximately 54% of the sequences input into AlignACE.

Figure 10 shows the Logos representation of the four discovered motifs, which illustrate the conservation of the individual nucleotides at the specified positions by the height of the letters representing each nucleotide. Motif 27 is represented in 25% of the sequences input into AlignAce and is 82% unique from the motif that accounts for the next highest percentage of input sequences, motif 17. These two motifs, together, are represented in 45% of the input sequences after subtracting duplicate sequences. The other two motifs, 24 and 29 combined, are represented in 33% of the input sequences. Twenty-three of the input sequences are unique to

motif 17, 16 are unique to motif 27, 11 are unique to motif 24, and 13 are unique to motif 29. Therefore, 63 of the 115 CTCF binding site peak sequences with one of these motifs (55%) contain a single motif. The remaining 45% of the CTCF binding site peak sequences contain multiple motifs. The CTCF binding site peak labeled as 2R\_171, which is found at the promoter sequence of AGAP002418, which codes for a cytochrome P450 protein, contains all four motifs within its 216 bp sequence. Other CTCF binding site peak sequences contain more than one of the identified motifs; eighteen contain three of the motifs and 27 contain two of them.

The Perl Program, Patser [56], was used to search for the four motifs among the 2,416 identified CTCF binding site peaks. To do this, alignment matrices were constructed using the aligned sequences from the AlignACE output which were input into the enoLogos online software [77]. A position specific scoring matrix for each of the four motifs was constructed and input into Patser with the 2,416 CTCF binding site peak sequences. Motifs 17, 24, 27, and 29 were identified among only 11.9%, 7.8%, 8.3%, and 8.4% of the entire ChIP-Seq dataset, respectively.

Motif 17



Motif 24



Motif 27



Motif 29



**Figure 10: Logos representing the motifs identified from among the 212 CTCF binding site peaks.** These motifs had relatively low repetitiveness and longer sequences of continuously conserved bases compared to the remaining motifs identified by AlignACE.

## 2.4 Discussion

### 2.4.1 The distribution of CTCF binding sites likely reflects its multiple putative functions

This ChIP-Seq experiment has identified 2,416 regions of CTCF binding throughout the *An. gambiae* genome ranging in size from 48 bp to 1,970 base pairs, with 62% of the data set between 100 and 200 base pairs. The identification of regions of CTCF binding, defined in this work as CTCF binding site peaks, throughout the *An. gambiae* genome will aid in the identification of insulator sequences that could be used to insulate a transgene that is inserted randomly into the genome. It will also aid in further research of the transcriptional regulation of genes that are important for developing an effective genetic strategy for controlling mosquito borne disease transmission.

CTCF has been shown to be an important factor across multiple species from humans to *Drosophila* [43]. Orthologs have been identified in human, chicken, mouse, *Xenopus*, zebra fish, cattle, tammar wallaby, platypus, central bearded dragon, *D. melanogaster*, *Ae. aegypti*, and *An. gambiae* [42, 78]. CTCF is highly conserved across these species within its eleven zinc finger domains, which is believed to provide its multifunctional properties. CTCF has been suggested to be a master weaver of multicellular genomes, with possible functions in nucleosome positioning, enhancer blocking, maintaining heterochromatin boundaries, mediating *cis* and *trans* long range chromatin interactions, imprinting, and X chromosome inactivation [41]. A model put forth to explain CTCF's multiple functions is the CTCF code, which identifies its ability to bind to multiple sequences by using different combinations of Zinc fingers as the mechanism responsible for its multiple functions [42]. Therefore, according to this model, the sequence of the binding site determines the function of CTCF by exposing different protein binding domains. Ohlsson *et al.* (2010) [42] also note that chromatin context influences CTCF

protein interaction and function. It has been shown that CTCF binding sites display insulator function in a plasmid-based insulator assay; however, these sites might have alternative properties in other assays, due to different chromatin contexts[42]. Transcription factor binding sites such as the thyroid hormone response elements can modulate CTCF function [42, 59]. Methylation is also known to antagonize CTCF binding. The correlative data of CTCF's presence at H3K27me3/H2AK5ac borders implies a need for cooperation between CTCF and chromatin modifiers [42, 62, 63]. Thus, in addition to the underlying sequence of the binding site, multiple factors in the chromatin context also appear to contribute to determining the function of CTCF [42].

CTCF has multiple protein binding partners including other CTCF molecules, which are determined by the available protein binding domains. The binding sequence and the chromatin context appear to determine the availability of a particular protein binding domain at a particular genomic position. This ability to expose different protein binding domains, determined by the DNA sequence it binds and the chromatin context of the binding site, also leads to different post translational conformations such as poly(ADP-ribosyl)ation, SUMOylation, and phosphorylation. These modifications are associated with different CTCF functions such as insulator function, transcriptional repression, and growth inhibition [42].

The primary function of CTCF may be to spatially organize the genome with multiple DNA sequences and chromatin contexts throughout the genome, appropriately guiding the necessary interactions to do so. Enhancer blocking, chromatin boundary function, transcriptional activation and repression, as well as other functions mentioned above may simply be appendages of this primary function. It appears that the genome has a code to dictate the appropriate function of CTCF at the appropriate genomic location via specific sequence and chromatin context [42]. This makes the function of CTCF at individual binding sites difficult to predict, especially in

insect species that have multiple insulator proteins that interact with CTCF at a subset of the binding sites, which also appear to affect its function [44, 58].

The distribution of CTCF binding site peaks in relation to genes and within genes may be an indication of the role CTCF plays at each of its varying locations. As mentioned previously, CTCF is known to function as a repressor, an activator, an insulator, and more recently has been identified as playing a role in determining splice variants [65]. A majority of the identified CTCF binding site peaks are located in intergenic regions of the genome. We expected this to be the case, considering that one of CTCF's functions is to act as an insulator between genes and gene transcriptional regulatory regions. However, considering the large amount of intergenic sequence in the *An. gambiae* genome (90%), it is interesting to note that only 51% of the identified CTCF binding site peaks are found to be intergenic. This suggests that at nearly half of the CTCF binding site peaks, CTCF is likely to be acting within genes, perhaps performing an intragenic regulatory function rather than or as well as insulating nearby genes from the genomic environment.

#### 2.4.2 What effects does CTCF have on neighboring genes?

This data also showed that many of the intergenic CTCF binding site peaks were located proximal to genes, with 28% of them within 10 kilobases upstream of a promoter, and 14% within 10kb downstream of a promoter. The proximity to genes may indicate selective forces maintaining CTCF binding sites at these locations. It is likely that CTCF is important for regulating transcription of those genes near CTCF binding site peaks, as well as those genes containing CTCF binding site regions within their coding region. It is important to keep in mind that CTCF also has been shown to be involved in long distance interactions, which cannot be inferred from this data set. Ohlsson *et al.* (2010) [42] explain that CTCF function is likely determined by sequence and chromatin environment, and it has not been implied that genomic

position in relation to genes can definitively predict the effect of CTCF on neighboring genes [42]. Such questions regarding the effect of CTCF on nearby genes will have to be answered by testing hypotheses at each individual locus through genetic engineering techniques. Keeping in mind that the CTCF binding site peaks identified in this study may contain one or multiple binding sites, the relationship between CTCF and neighboring genes can be fairly complicated. However, previously identified CTCF binding sites identified near genes in other organisms can provide some insights.

For CTCF binding sites located at the promoter region of genes, CTCF may likely be functioning as an activator or a repressor, such as has been identified in vertebrates at the APB $\beta$  promoter and *c-Myc* promoter respectively [79, 80]. For binding sites within intergenic regions, it is likely that CTCF may act as an insulator for nearby genes, isolating inappropriate enhancers from gene promoters. This is the case at the bithorax complex, whereby boundary elements maintain appropriate expression levels of *ultrabithorax*, *abd A*, and *abd B* in the appropriate parasegments, based on the location of insulator regions between the parasegment specific enhancers [73]. Other examples of intergenic CTCF binding sites have also been identified as insulators in vertebrates, such as the *Igf2/H19* [37] locus and the chicken HS4  $\beta$ -globin locus [81]. It should be noted that although CTCF has been shown to only function as an enhancer blocker at the cHS4 insulator and is independent of silencer blocking, the insulator sequence is responsible for silencer blocking [82]. Therefore, CTCF binding may be a useful genomic landmark for identifying insulator sequences even if a portion of the insulator function is independent of CTCF. To determine the function of CTCF at its DNA binding sites throughout the *An. gambiae* genome, it will be necessary to perform assays for individual binding sites similar to those used to identify the function of CTCF in the above mentioned organisms.



An EMSA assay, to determine the sequence bound by CTCF followed by a methylation interference assay, to determine the specific nucleotides required for CTCF binding, are two of the assays necessary to determine the affinity of CTCF at a given binding site. Once the binding sequence is determined through a competitive mobility shift assay, a mutated sequence could be synthesized to compare functionality with the wild type sequence in the appropriate transgenic assays necessary to determine the hypothesized function.

To test repression or activation, as was done for the *APBβ* promoter and *c-myc* promoter [79, 80] a reporter gene assay could be performed by independently inserting the wild type and mutated sequences adjacent to the promoter of a reporter gene such as CAT (chloramphenicol resistant gene) or EGFP (enhanced green fluorescent protein), followed by transfection of the recombinant plasmid construct into cultured cells and integration into a chromosome. A similar assay could be performed to determine insulator function by inserting the sequences between the promoter of a reporter gene and a functional enhancer as performed by Li *et al.* (2008) [83]. These experiments can provide an indication of the role of CTCF at a genomic location based solely upon the binding site sequence; however, the sequences are not incorporated into their natural chromatin environment, and thus the CTCF functional role cannot be exactly determined at each natural binding site.

Site directed homologous recombination would be useful in determining the function of a CTCF binding site *in vivo*; however, for dipterans this is a difficult procedure and is not routinely performed even in *D. melanogaster*. The biological system necessary for efficient homologous recombination in mouse is not present in dipterans making such an endeavor extremely challenging, not to mention the repetitive nature of mosquito genomes at many of the CTCF binding sites adding to the difficulty of the task. The best option for determining the effects of chromatin environment on CTCF function would be to independently insert a wild

type CTCF binding site sequence and a mutated CTCF binding site sequence into the same location of the genome via a site specific integration system, such as  $\Phi$ C31, to determine how CTCF functions in a particular chromatin environment. Multiple  $\Phi$ C31 docking sites throughout the mosquito genome would allow for a variety of chromatin contexts within which different CTCF binding site sequences could be assayed for CTCF function. This experiment is based on results that showed that insulator sequences integrated into the genome can determine the nuclear localization of DNA, as determined by Gerasimova *et al.*(2000) [84] in the case of the *gypsy* insulator [84].

Binding site regions found within introns and exons have at least two possible functions based on previously identified binding sites at similar locations in other genomes. CTCF has been shown to function as a repressor when bound to introns by blocking RNA polymerase II and stalling transcription, such that the full RNA transcript is not completed and thus gene expression is repressed. An example of this occurs at the *BCL6* locus of the human genome [66]. Two CTCF binding sites are located at the 5' end, one at the 3' end, and multiple putative CTCF binding sites are located within intron 1. The intron 1 putative sites have shown robust enrichment for CTCF in H929 cells. These same putative sites show enrichment in Raji cells when methylation is removed by treatment with 5-Aza-C. A CTCF knockdown with a *ctcf* short hairpin RNA resulted in an increase of *BCL6* expression. By comparing cell types with varying levels of methylation at a particular CTCF binding site and analyzing its ability to bind CTCF and its effects on *BCL6* transcription, this study was able to demonstrate that CTCF can act as a repressor when bound to intronic sequences[66]. Perhaps further study of CTCF binding in multiple tissues of *An. gambiae*, resulting in the identification of variable CTCF binding sites across cell types will provide similar opportunities to study CTCF function at introns in this species.

Other evidence has shown that CTCF bound to the exon of a gene causes it to be included in the transcript by pausing RNA polymerase II, then resuming transcription such that the bound exon is included in the transcript [65]. This study compared cell types with varying expression levels of the splice variants of human *CD45*. A comparison of published CTCF ChIP-Seq results revealed variation in CTCF binding at exon 5 of *CD45* across the cell types. ChIP data revealed that CTCF binding at exon 5 and the inclusion of exon 5 in the transcript were shown to be linked. ChIP data also revealed that RNA polymerase II pausing just upstream of the CTCF binding site was shown to correspond to CTCF binding and exon 5 inclusion, as well as an increase in exon 4/5 and 5/6 junctions in the transcript. RNA Pol II has shown it can resume transcription when CTCF is bound at exon 5. These data show that CTCF binding is important to the inclusion of exon 5. When the CTCF binding site is methylated, CTCF does not bind and RNA polymerase II binding does not occur. This leads to the exclusion of exon 5, evidenced by the loss of exon 4/5 and 5/6 junctions and an increase of 4/6 junctions. Thus CTCF has been shown to play a role in distinguishing between splice variants [65]. The potential intronic and exonic CTCF binding identified in our data set may lead to an improved understanding of gene regulatory mechanisms in *An. gambiae*. More CTCF ChIP-Seq data from other *An. gambiae* cell types will be necessary to conduct such studies.

At the HPX 6 locus, a binding site is located upstream of the promoter and appears to be a candidate insulator for HPX 6, insulating it from cross talk outside of its transcriptional regulatory region. However, this binding site is also located at the promoter of a novel gene and appears that it could act as an activator or repressor of this gene. Such an example could be studied to determine how CTCF functions when multiple scenarios are possible. If Ohlsson *et al.* (2010) [42] is correct and the underlying sequence determines the function of CTCF, then perhaps identification of the sequence at such positions will provide improved predictions of

function compared to the position relative to genes. Furthermore, recent work suggests that other proteins bound to adjacent DNA sequences forming complexes with CTCF may play a role in determining the functional role of CTCF at a given binding site [78].

As mentioned above, CTCF function cannot be predicted from relative proximal genomic position to genes alone. However, this work provides the necessary data to begin the study of the effects of CTCF on important genes and gene clusters. Further study of the function of CTCF at these varying types of binding site regions will require individual analysis of each CTCF binding site peak. Therefore, our analysis is focused on genes that may be helpful for establishing transgenic strategies for controlling the transmission of mosquito borne diseases. As such, genes important for immunity, blood feeding, sex differentiation, and development have been included in this analysis.

#### *2.4.3 CTCF may regulate genes important for immunity*

The CLIP genes are serine protease inhibitors that play a role in regulating the melanization response, which is a process that encapsulates *Plasmodium* at the ookinetes stage in the midgut of the mosquito. Knockdowns of CLIP genes have shown that some enhance melanization while others have been shown to reduce melanization [85]. Four CLIP gene regions are proximal to identified CTCF binding site peaks. Two different CLIP gene regions, one containing primarily CLIPB genes on Chromosome 2R and the other containing primarily CLIPA genes on Chromosome 3L, are each flanked by CTCF binding site peaks. Interestingly, two genes in the CLIPB, *clipB3* and *clipB4* region are necessary for activation of the melanization process and two genes in the CLIPA region, *clipA2* and *clipA5* block melanization [85]. In both cases, the flanking positions of the CTCF binding site peaks, relative to the CLIP gene clusters, suggest an insulating role for CTCF at these genomic locations based upon results in other organisms that indicate that genes and clusters of genes are insulated by flanking

insulator sequences, such as the genes of the bithorax complex in *D. melanogaster*. CTCF may play an important regulating role in maintaining an appropriate balance between the expressions of these two groups of genes in order to regulate the melanization process. Further understanding of this process may be helpful in the development of mosquito strains refractory to disease transmission.

Two other CTCF binding site peak positions near CLIP genes suggest another possible function for CTCF. CLIPA8 and CLIPB17 have CTCF binding site peaks within their promoter regions, indicating that CTCF may act as a repressor or an activator for these two genes. Also, the CTCF binding site peak 4.4kb upstream of CLIP A8 poses an interesting question regarding CTCF function when it is bound near two genes and another CTCF binding site. This nearby peak may have an insulating influence upon the expression of CLIPA8. However, it is important to note that this same CTCF binding site peak is also located within the promoter region of novel gene AGAP010728 and may simply function as an activator or repressor of that gene. More research needs to be performed to determine how neighboring CTCF binding sites regulate gene expression in the same genomic vicinity.

The other immunity gene of interest is *cactus* (*cact*). Identified as a negative regulator of the immune response Toll pathway in *D. melanogaster*, RNA interference-mediated silencing of *cact* results in Toll pathway activation. This has been shown to significantly decrease the *P. berghei* burden, and the removal of the negative regulator can induce an immune response without a pathogen challenge [86]. A CTCF binding site peak is located at the *cact* promoter in *An. gambiae*. It appears that CTCF may play an important role in the regulation of *cact* expression. Downstream of *cact*, the novel gene, AGAP007941, also has a CTCF binding site peak at its 2<sup>nd</sup> exon. Although its function is unknown, further study could improve the understanding of intragenic gene regulation by CTCF. The location of these peaks suggests that

CTCF may act as an activator, a repressor, or as an intragenic regulator distinguishing between splice variants in *An. gambiae*. Understanding this relationship may be useful in the development of a *Plasmodium* resistant strain of mosquitoes, as well as an informative model of CTCF function.

#### 2.4.4 Some heme-peroxidase genes may be regulated by CTCF

CTCF binding site peaks have been identified near several heme-peroxidase genes. As stated in Chapter I, the peroxidase gene family is the only gene family that demonstrates a significant difference in copy number when comparing *An. gambiae* and *D. melanogaster* [24]. Most of the CTCF binding sites are found flanking heme- peroxidase genes in a relatively small genomic region. One is found within the intron of the long splice variant of HPX 4 and upstream of the short splice variant. In the malaria mosquito, *An. albimanus*, peroxidase activity has been localized to the posterior lobe of the salivary gland of female mosquitoes just before blood feeding [25]. It was also detected in nitrocellulose membranes probed by hungry mosquitoes. Peroxidase activities were lower in salivary glands of mosquitoes after probing and blood feeding. Thus, it was suggested that in salivary glands, heme-peroxidase functions as an antagonist to vasoconstricting substances [25]. The HPX 4 gene is of particular interest given the position of the CTCF binding site peak that suggests CTCF may play a role in regulating the expression of the two splice variants. The function of these genes in *An. gambiae* is still speculative; however, understanding the role of CTCF in regulating the expression profiles of these genes that may be important for blood feeding and pathogen transmission may lead to improved transgenic strategies for pathogen transmission control.

One heme peroxidase gene identified in *An. gambiae*, HPX 15, known as an Immunomodulatory peroxidase (IMPer) has been discovered to assist in the formation of a peritrophic matrix with dual oxide (Duox) upon blood feeding in female *An. gambiae*. The

peritrophic matrix forms around the blood bolus in the midgut to prevent contact of blood cells and dietary bacteria with the mid gut epithelium, as a first line of defense against pathogens, and as a means of allowing dietary bacteria into the midgut without eliciting an immune response. IMPer and Duox are secreted from the midgut epithelium and catalyze protein crosslinking in the mucin layer to form the peritrophic matrix. The peritrophic matrix has been shown to reduce the permeability of immune elicitors against bacteria and plasmodium parasites. When IMPer is silenced via dsRNA, the median number of *P. berghei* oocysts present 7 days post infection is reduced by 9.2 fold. Ookinetes invade the midgut in IMPer silenced individuals; however, they are killed and appear fragmented. Silencing of IMPer in *An. gambiae* and *An. Stephensi* also reduced *P. falciparum* infection. It was shown that when IMPer is silenced, NOS, an enzyme that generates nitrous oxide, which is a potent antiplasmodium effector molecule, is induced. Similar results were obtained when Duox was silenced, thus Duox and IMPer are necessary to form the peritrophic matrix in the midgut which protects Plasmodium ookinetes. Thus, reducing the expression of IMPer (HPX 15) allows for the *An. gambiae* immune system to protect itself from *Plasmodium* infection via induction of the antiplasmodium molecule, NOS, due to lack of formation of the peritrophic matrix [72]. Increased understanding of CTCF's role in transcriptional regulation of HPX15 may provide more insight that would be useful for investigating the mosquito immune system as a means to reduce plasmodium transmission.

#### 2.4.5 CTCF binding site peaks are located near important sex differentiation genes

Doublesex is a well conserved protein with a well conserved binding site among dipterans, and is responsible for the normal expression of secondary sexual characteristics. A sex specific variant binds near genes controlling secondary sexual characteristics to regulate them in a sex-specific manner. The percentage (~4%) of doublesex binding motifs identified among the

CTCF binding site peak sequences was the same as that identified throughout the entire genome. Therefore, there appears to be no correlation between CTCF binding and doublesex binding.

Two other genes important for sex differentiation are *sex lethal (sxl)* and *msl-2*, which work in concert to control dosage compensation in *D. melanogaster*. Little is known about dosage compensation in *An. gambiae*; however, orthologs for these genes exist in *An. gambiae* and an improved understanding of their transcriptional regulation could lead to a better understanding of the mechanisms governing sex differentiation and dosage compensation in this species. The *An. gambiae msl-2* ortholog is located between two CTCF binding site peaks only 34.8kb apart. Both peaks are found within intergenic sequences suggesting that they may insulate the *msl-2* gene.

The ortholog for *sxl* is flanked by two CTCF binding site peaks, 58.9 kb apart, with an additional CTCF binding site peak within intron 1 of the long splice variant and at the promoter of the short splice variant. This peak may be involved in regulating the expression levels of the two splice variants, under the control of differential methylation of the binding site, as has been described in mammalian systems [65]. As for the flanking binding sites, one is within the TTS of the novel gene AGAP003897 and the other is within an intron of the novel gene AGAP003901. The flanking binding site regions found within the novel genes may be regulating the expression of the two respective genes, depending on which sequences they bind, rather than acting as insulators for *sxl*, or they may be functioning as insulators and intragenic expression regulators simultaneously. Further study may provide more insight as to how CTCF distinguishes between its multiple functions. Given the interconnected function of MSL-2 and SXL, study of the transcriptional regulation of these two genes for the purpose of producing all male mosquito populations is an attractive field of research for the implementation of SIT.



#### 2.4.6 The CTCF binding site profile of the *Anopheles gambiae* bithorax complex

The bithorax complex (BX-C) of *D. melanogaster* has been well studied. Multiple insulator sequences binding different insulator proteins are necessary for regulating the expression of *ultrabithorax* (*ubx*), *abdominal A* (*abd-A*), and *abdominal B* (*abd-B*) within nine different body segments in order to maintain the *Drosophila* body plan [87]. Interaction of the nine different transcriptional regulatory regions is responsible for their normal expression among the nine different body segments. CTCF binds to six of the insulator sequences in the BX-C, as well as three other locations. Cloned binding sites have been shown to interact with one another; however, deletion of the binding sites only partially reduced the ability of insulator elements to interact. The regulation of the BX-C does not appear to be one of simple insulation and relief of insulation. Its complexity lies in the involvement of other factors, including polycomb group proteins. Many other DNA sequences, such as polycomb response elements, promoter targeting sequences, and promoter targeting elements, as well as enhancers, initiators and promoters are involved. Therefore, a complete understanding of the function of CTCF at the BX-C is yet to be elucidated [61].

In the *An. gambiae* ChIP-Seq data, three peaks are located within the bithorax complex that appear to be homologous to three of the binding sites identified among the *Drosophila* insulators, *Fab-3*, *Fab-4*, and *Fab-8*. This suggests conservation of the bithorax complex across these two species. The other three CTCF bound insulators may also be present in the *An. gambiae* bithorax complex; however, the sensitivity of the ChIP-Seq experiment may not have been sufficient to detect them. These binding sites may or may not have the same function as their likely homologs in *Drosophila*. Further assays, as outlined above, will be necessary to determine their function

#### 2.4.7 *An. gambiae* CTCF is associated with a variety of DNA sequence motifs

CTCF binding site peak sequences that correlate with CTCF antibody enrichment on chromosome bands, and CTCF binding site peak sequences that have been validated using ChIP-PCR were used to identify four potential CTCF binding site motifs. These four motifs are found in 54% of the 212 sequences input into AlignACE. The percentages of motifs 17, 24, 27, and 29 among the entire CTCF binding site data set were 11.9%, 7.8%, 8.3%, and 8.4%, respectively. These low percentages were unexpected given that the primary *Drosophila* CTCF consensus is found among 50% of *Drosophila* CTCF binding sites identified by ChIP-Seq, and second and third motifs were identified among another 40% and <10% of the putative *Drosophila* CTCF binding sites. The total number of CTCF binding site peaks identified in *An. gambiae* (2,416) is comparable to the total number of CTCF binding sites identified in *Drosophila* (2,871)[44], both using a ChIP-Seq technique. As both species have comparable numbers of genes in their respective genomes (13,460 and 17,864) [3, 67, 68, 88], these results suggest a low rate of false positives from the ChIP-Seq identification method. The use of CTCF binding site peak sequences that correlate with the CTCF immunostained chromosomes or have been validated with ChIP-PCR should have eliminated most false positives from the dataset input into AlignAce, thus providing the most accurate analysis possible.

Van Bortle *et al.* [44] identified three CTCF binding site motifs based on their ChIP-Seq data. The secondary and tertiary motifs were representative of binding sites with lower CTCF occupancy and the presence of binding sites for insulators nearby. The variable consensus were believed to be affected by these other insulator binding sites, CP190 being considered the most likely due to the necessity of CP190 for CTCF to bind at a subset of its binding sites [44]. This may also be the case for *An. gambiae*; however, the motifs identified with the *An. gambiae* data are not as conserved as those identified for *Drosophila*. Each of these was only present in

11.9% or less of the 2,416 identified CTCF binding site peaks. Therefore, nearby binding sites of other insulator proteins does not likely solely explain the high variability of sequences that is seen in the *An. gambiae* ChIP-Seq dataset. CTCF may have been detected at some of the identified sequences due to indirect interactions with other bound proteins. However, this would also be expected to occur in *Drosophila*, resulting in a less conserved consensus. Therefore, assuming neither of these possibilities are the primary cause for variability among the *An. gambiae* CTCF binding site peak sequences, these results suggest that *An. gambiae* CTCF binds to a wider variety of DNA sequence motifs than *Drosophila* CTCF.

Further experiments identifying the binding sites of other insulator proteins, such as CP190, and Su(Hw), will provide more insight as to whether or not these proteins are also associated with any of the four identified motifs. Chromatin Immunoprecipitation (ChIP) Chromosome Conformation Capture (3C) could be used to identify interactions between proteins bound by these motifs. ChIP-Seq with three biological replicates with the CTCF antibody can be performed, as well as using an alternative *An. gambiae* CTCF antibody would increase the number of high confidence CTCF binding site peak sequences. Electrophoretic Mobility Shift Assays will be necessary to identify actual nucleotide sequences bound by CTCF. These experiments will aid in identifying candidate insulator sequences to insulate transgenes.

## **2.5 Conclusions**

The findings of this study accomplished the goal of identifying a large sample of CTCF binding sites throughout the *An. gambiae* genome, of which a subset can be assayed for insulator function and eventually incorporated into a transgene construct to evaluate their effectiveness in overcoming position effects in transgenic mosquitoes. In addition to our primary goal, the identification of CTCF binding sites throughout the genome reveals some of the genes that may be regulated by CTCF. Variable positions of CTCF binding sites in relation to neighboring genes

provides insight as to the role CTCF may play in regulating the transcription of these genes. Further studies will be necessary to determine the functional role of CTCF at individual genomic loci. The model for the diverse functional roles of CTCF put forth by Ohlsson *et al.* (2010) [42] states that the CTCF binding site sequence determines the function of CTCF at a particular position[42]. Some binding sites, based on their position, appear as though they could have more than one function. Such situations will require functional assays to determine the role of CTCF at those loci. The variety of sequence motifs among the CTCF binding site peaks is unexpected. This suggests that *An. gambiae* CTCF binds to a more variable set of DNA sequences than observed in *Drosophila*. Electrophoretic Mobility Shift Assays will need to be performed to identify actual nucleotide sequences bound by *An. gambiae* CTCF.

## CHAPTER III

### EXPRESSION PROFILES OF THE INSULATOR PROTEINS, CP190 AND SU(HW) in *Aedes aegypti*

#### 3.1 Introduction

CTCF is the only known insulating protein in vertebrate genomes. Dipteran genomes are much more compact and have at least five insulator proteins in addition to CTCF: BEAF-32, Zw5, Su(Hw), GAGA factor and CP190 [70]. BEAF-32 and Zw5 bind to the scs' and scs insulating elements respectively. Su(Hw) binds to a specific sequence within the gypsy transposable element, as well as other sequences throughout the *Drosophila* genome [73]. GAGA factor binds to insulator sequences throughout the *Drosophila* genome, including *Frontalabdominal 7 (Fab 7)* [89]. CP190 has been shown to have overlapping binding sites with other insulator proteins such as CTCF and Su(Hw) throughout the fly genome [52, 90]. CP190 occupancy has been shown to be responsible for H3 depletion, and CP190/dCTCF double occupancy sites have been detected at the borders of H3K27me3 islands, suggesting that it plays a role in chromatin remodeling with CTCF [52]. CP190 has been shown to be necessary for CTCF binding at some binding sites [90].

CTCF and CP190 are important for the regulation of body patterning in development. *Drosophila* studies show that *dctcf* mutants show a homeotic phenotype and pharate lethality [90]. These studies showed that most dCTCF binding sites are also occupied by CP190, including insulators within the bithorax complex. The enhancer blocking ability of *Fab 8* was tested in an enhancer blocking assay using the white enhancer and the mini-white reporter gene. All CP190 mutants resulted in increased eye pigmentation, thus revealing the need for CP190 at the *Fab 8* insulator for proper insulator function [90].

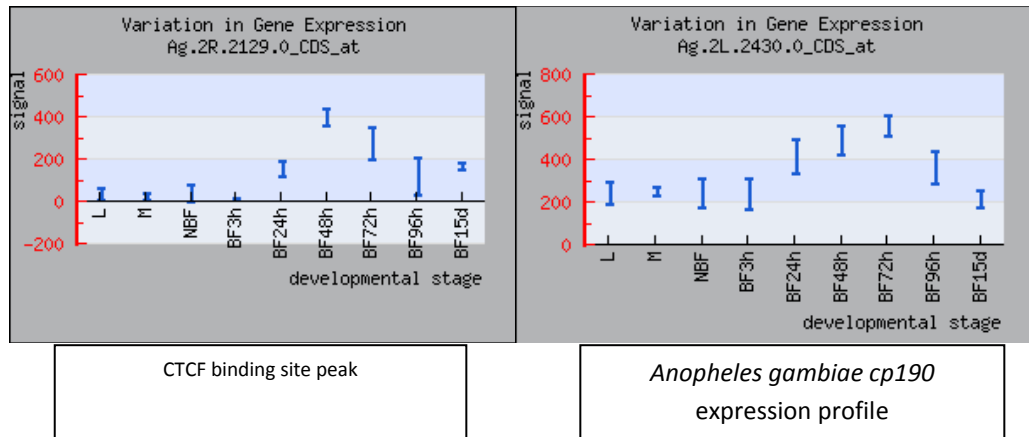
Su(Hw) has been shown to bind throughout the *Drosophila* genome at sequences within the *gypsy* retrotransposon, as well as at non-*gypsy* sequences [73]. These two different sequences may reflect different functions for Su(Hw) [11]. Previous immunohistochemistry studies showed that CP190 colocalized with Su(Hw); however Su(Hw) did not appear to colocalize at sequences bound with CTCF [90]. However, more recent ChIP-Seq data revealed that Su(Hw) and CTCF bind at adjacent sequences, within 200 to 300 base pairs, and may be responsible for important chromatin architecture and insulator activity at the borders of H3K27 rich regions of the genome. The presence of thousands of independent Su(Hw) sites likely biased the earlier analysis to lead the authors to believe that CTCF and Su(Hw) did not colocalize [58].

This *Drosophila* ChIP-Seq analysis of insulator proteins also identified three motifs for CTCF binding. The most common, or primary motif was found to generally bind only CTCF. However, at the secondary and tertiary motifs, CTCF colocalized with Su(Hw), BEAF-32, CP190, MOD(MDG4), and other cofactors at the borders of H3K27 enriched regions. As shown at the *Fab 8* insulator, CP190 and other insulator proteins may be necessary for insulator function at some CTCF binding sites [43].

Putative orthologs for *cp190* and *su(Hw)* exist in *Anopheles gambiae* and *Aedes aegypti*. According to the *An. gambiae* expression profile database at UC Irvine (Figure 11) [54, 61, 71, 78], the expression profile of the *cp190* ortholog is similar to that of the *Anopheles gambiae ctf* ortholog profile across life stages from larvae to 15 days post blood feeding, with time points at 24 hours post blood feeding, 48 hours post blood feeding, 72 hours post blood feeding, and 96 hours post blood feeding. Adult male and non-blood fed adult female expression profiles were also compared. Although overall expression levels of *ctcf* are higher than *cp190* expression levels across all life stages in *Anopheles gambiae*, the data show that the expression levels of both profiles increase with blood feeding and gradually decrease over time. Expression of *cp190*

reaches its peak at 48 hours and *ctcf* expression reaches its peak at 72 hours. Both patterns indicate an important role in embryo development, and are consistent with an interdependent function of one another, as is suggested by the previously mentioned *Drosophila* data [44].

A similar expression profile for the *Ae. aegypti* *ctcf* ortholog was determined using gel based RT-PCR analysis across the life stages from embryo  $\leq 1$  hour post oviposition to ovaries of blood fed females. The life stages and time points examined were  $\leq 1$  hour post oviposition, 24 hours post oviposition, larvae, female pupae, male pupae, female adult (non-blood fed), male adult, ovaries non-blood fed, and ovaries blood fed [50]. The expression profile for *ctcf* across these life stages is similar to that of *ctcf* and *cp190* in *Anopheles gambiae*. Both embryo stages show increased expression levels, with that of *ctcf* expression at  $\leq 1$  hour post oviposition showing the most elevated expression level. Ovaries post blood feeding show a significant amount of expression compared to non-blood fed ovaries. These data are consistent with *ctcf* having an important role in development. Considering the *Drosophila* and *An. gambiae* data regarding *ctcf* and *cp190* expression, as well as the dependence of CTCF binding on CP190 at some CTCF binding sites throughout the *Drosophila* genome, it is likely that the *Ae. aegypti* ortholog to *cp190* would have a similar expression profile to the *Ae. aegypti* ortholog to *ctcf*. In light of the recent ChIP-Seq data suggesting interaction of Su(Hw) with CTCF at secondary and tertiary CTCF binding motifs, it would also be likely that *Ae. aegypti* *su(Hw)* would have an expression profile similar to those of *ctcf* and *cp190*. This chapter summarizes expression profiles for the *Ae. aegypti* *cp190* and *su(Hw)* orthologs using RT-PCR across eight life stages and an additional three ovarian stages for *Aedes aegypti* *cp190*.



**Figure 11: Expression profiles of *An. gambiae cp190* and *ctcf*.** Data from the UC Irvine *Anopheles gambiae* expression profile database [54, 61, 71, 78] shows expression levels for *cp190* and *ctcf* across larvae, male, non-blood fed female, 24, 48, 72, 96 hrs., and 15 days post blood feeding life stages.

### 3.2 Materials and Methods

Total RNA was extracted from embryos 2 to 3 hours post oviposition, embryos 24 hours post oviposition, late larval stage, male pupae, female pupae, male adults, non-blood fed female adults, female adults post blood feeding, ovaries from non-blood fed females, ovaries from females 48 hours post blood feeding, and ovaries from females 72 hours post blood feeding. All RNA was extracted using a standard Trizol method. Each sample of RNA was used to synthesize cDNA for each life stage from which RNA was extracted using the High Capacity Reverse Transcription Kit (Foster City, CA). The cDNA concentration for each sample was measured on a Nanodrop spectrophotometer (Nanodrop Technologies, Wilmington, DE).

Primers were designed at the 5' end of the transcript of the *Ae. aegypti* ortholog for *cp190* (AAEL002771-RA), flanking a 416 base pair region from nucleotide 1317 to nucleotide 1732. Primer sequences are as follows Forward: 5'- CCCTTGGCTGTGTCTACGTT-3', Reverse: 5'- ATTCATCGTCCGAGAAATCG-3'. Primers were designed in the middle of the



transcript of the *su(Hw)* *Ae. aegypti* ortholog (AAEL002145-RA), flanking a 308 base pair sequence from nucleotide 946 to nucleotide 1235. Primer sequences are as follows: Forward: 5'-ACTGGTGAACGACCTATCG-3', Reverse: 5'-CTTCCGGATGAACGACTTTG-3'. Each PCR mixture consisted of 100 nanograms of template cDNA, 0.8 $\mu$ l of each primer for a final concentration of 0.8 $\mu$ M, and 10 $\mu$ l of 2x Go TAQ master mix in a total volume of 20 $\mu$ l. The PCR amplification conditions utilized were as follows: 1: 95°C for 5 minutes, 2: 95°C for 15 seconds, 3: 52°C (*cp190*) and 51°C (*su(Hw)*) for 15 seconds, 4: 72°C for 30 seconds, 5: repeat steps two through four 29 times for a total of 30 cycles, 6: 72°C for 2 minutes. PCR products were separated on a 1.5% electrophoresis gel at 100 volts at 45 mAmps for 30 minutes. Each experiment was performed twice giving the same results.

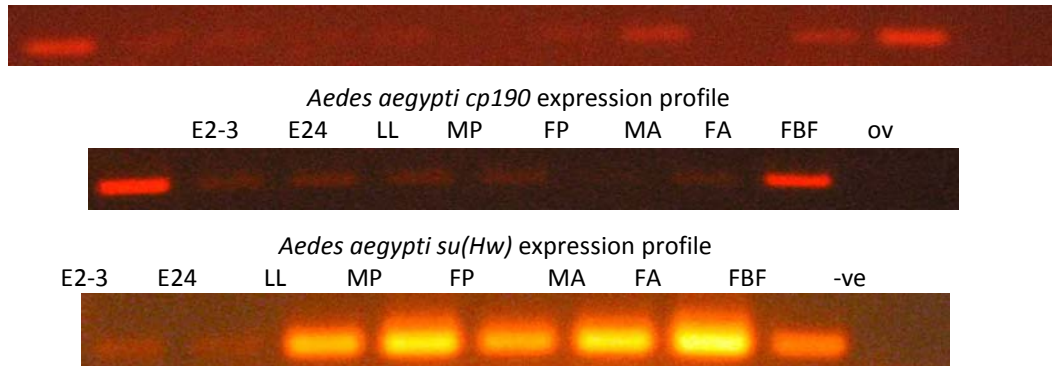
### 3.3 Results

Expression of the *Ae. aegypti cp190* ortholog (AAEL011409) (Figure 12) was observed across seven of the eight life stages examined. Amplification was not observed for the male adult template. Of the ovarian tissue examined, amplification was not observed for the non-blood fed samples. The lack of amplification of for these two life stages is likely due to low levels of gene expression. Both blood fed samples at 48 hours post blood feeding and 72 hours post blood feeding resulted in amplification products, revealing expression of *cp190* in these tissues.  *$\beta$ -actin* expression was used as a control across eight of the life stages, from embryo 2 to 3 hours post oviposition to female adult blood fed. Expression of  *$\beta$ -actin* was consistent across 5 of the life stages from late larva to non-blood fed female adult. Lower expression of  *$\beta$ -actin* was observed in the embryonic stages. This may be explained as they are developmental stages in which  *$\beta$ -actin* would be expected to have low expression levels due to the lack of muscle development at these stages. A slight decrease in  *$\beta$ -actin* expression in the blood fed female sample may be explained by lethargy in the females after a blood meal. Amplification products of the *cp190*

ortholog observed at 2 to 3 hours post oviposition, female adult post blood feeding, ovaries 48 hours post blood feeding, and 72 hours post blood feeding, were at higher levels compared to those of the remaining life stages, mirroring the *Aedes aegypti ctcf* ortholog expression profile. Ovaries 72 hours post blood feeding and the embryos 2 to 3 hours post oviposition were the stages at which *cp190* appeared to be most highly expressed. These results are consistent with *cp190* playing an important role in embryo development.

Expression of *Ae. aegypti su(Hw)* (Figure 12) was observed across all eight life stages. The embryo 2 to 3 hours post oviposition and female blood fed templates showed a significantly higher expression level compared to all other life stages. All remaining life stages showed a minimal amount of expression. These data mirror the *Ae. aegypti ctcf* and *cp190* expression profiles, as well as the *An. gambiae* expression profiles for the same orthologs. These data are consistent with the notion that Su(Hw), CTCF, and CP190 may interact at some binding sequences, as is suggested by the recent *Drosophila* ChIP-Seq data [44].

### *Anopheles gambiae* *ctcf* expression profile



**Figure 12: Expression profiles of *Ae. aegypti* *cp190* and *su(Hw)*.** RT-PCR amplification of cDNA fragments of *Ae. aegypti* orthologs for *cp190*, *su(Hw)*, and  $\beta$ -*actin* (control) across the following life stages: embryo 2-3 hrs. post oviposition (E2-3), embryo 24 hrs. post oviposition (E24), late larvae (LL), male pupae (MP), female pupae (FP), male adult (MA), female adult (FA), female blood fed (FBF), ovaries non-blood fed (ov NBF), ovaries 48 hrs. post blood feeding (ov 48), ovaries 72 hrs. post blood feeding (ov 72) and negative no template control (-ve).

### 3.4 Discussion

The expression profiles of the *Ae. aegypti* orthologs of *cp190* and *su(Hw)* were visualized across eight life stages, as well as blood fed and non-blood fed ovary tissue for *cp190*. This work has revealed that *Aedes aegypti* *cp190* is highly expressed in the early embryo, blood fed ovaries, and blood fed whole animal adult female compared to the other life stages assayed. This profile mirrors the expression profile of the *An. gambiae* ortholog for *cp190*, suggesting that its expression throughout the life cycle is conserved across these mosquito species. This profile also mirrors the expression profile of the *Ae. aegypti* ortholog for *ctcf*. The correlation of expression profiles suggests that CTCF and CP190 may have similar interdependent functions as has been shown in *Drosophila* [44].

The *Ae. aegypti* ortholog of *su(Hw)* was also shown to have a similar expression profile to *cp190* and *ctcf* in both mosquito species. Given the increased expression levels of these genes in early embryo and blood fed ovaries above all other life stages observed, it appears that the products of these genes may play important roles of gene regulation at developmental life stages. These data are also consistent with the *Drosophila* ChIP-Seq insulator protein data which show that in *Drosophila*, these proteins colocalize at a subset of insulator sequences binding to a secondary or tertiary CTCF binding motif. Many of these motifs are located at the borders of H3K27 enriched regions and CTCF has been shown to be necessary for maintaining these regions [44]. CP190 has also been shown to be necessary for the enhancer blocking function of CTCF at *Fab 8* [90]. Therefore, the interaction of multiple insulator proteins may be necessary for proper chromatin organization and insulator activity at some CTCF binding sites.

### 3.5 Conclusions

The expression profiles of the *Ae. aegypti* orthologs of *cp190* and *su(Hw)* mirror the *Ae. aegypti* expression profile for the *ctcf* ortholog and the expression profiles for the *An. gambiae* orthologs for *ctcf* and *cp190*. These data indicate that *cp190* is likely to be important for development based on its increased expression levels in embryos, blood fed females, and blood fed ovaries compared to all other life stages assayed. The *Aedes aegypti* ortholog of *su(Hw)* had a similar expression profile suggesting that it is also likely to be important for development. The parallel expression profiles across life stages with *ctcf* suggest that CP190 may have a similar role in facilitating binding of CTCF to its binding site sequences, based on data from experiments with *D. melanogaster* in which it was shown that CP190 was necessary for CTCF binding to a subset of its binding sites [90]. These data are also consistent with recent ChIP-Seq data for the *Drosophila* insulator proteins CP190, CTCF, and Su(Hw), which show that all three bind within 200-300 base pairs of one another at secondary and tertiary CTCF binding motifs,

along with the *Drosophila* specific insulator protein BEAF-32 [44]. It has been suggested that these proteins act synergistically to maintain chromatin architecture and insulator activity [44]. This may also be the case in mosquitoes. Further experiments will be necessary to identify spatial and functional relationships between CP190, Su(Hw), and CTCF in mosquitoes. ChIP-Seq experiments with antibodies raised against the mosquito orthologs to CP190 and Su(Hw) need to be performed to identify any colocalization of the three proteins within the genomes. Also, RNAi knock downs of each of these three insulator proteins would be useful in identifying any lack of dependence of these three proteins, for DNA binding at any of the identified binding sites. For the primary purpose of identifying optimal sequences for insulating transgenes in order to minimize position effects, sequences with multiple bound insulator proteins could be compared with sequences with single bound insulator proteins in an enhancer blocking assay. The sequences identified as the most effective enhancer blockers could be empirically tested in transgenic experiments with whole mosquitoes.

## CHAPTER IV

### CONCLUSION AND FUTURE DIRECTIONS

CTCF has been well studied in vertebrates and *Drosophila* over the past two decades [41]. These studies have provided much insight regarding the potential functional roles of CTCF in mosquito genomes since the discovery of its ortholog in *An. gambiae* and *Ae. aegypti* [49]. CTCF's potential role as an insulator protein could enable the identification of insulator sequences that could be used for improving mosquito transgenesis techniques. Furthermore, its likely role as a genome organizer has also provided the impetus for this work to identify regions of CTCF binding in the *An. gambiae* genome. Not only does the multitude of high resolution CTCF binding regions identified in this study provide potential insulator sequences that can be tested for potential use in the improvement of mosquito transgenesis, it also provides a new model organism for the study of CTCF function. Armed with the knowledge of the genomic locations of CTCF binding in *An. gambiae*, scientists can study CTCF in a species that is relatively closely related to *Drosophila* that has a different chromatin structure. Furthermore, the identification of CTCF binding sites in related Anopheline species that have undergone recent speciation events will provide insight into the role of chromatin organization in this process. This will lead to further insight regarding the evolution of chromatin organization.

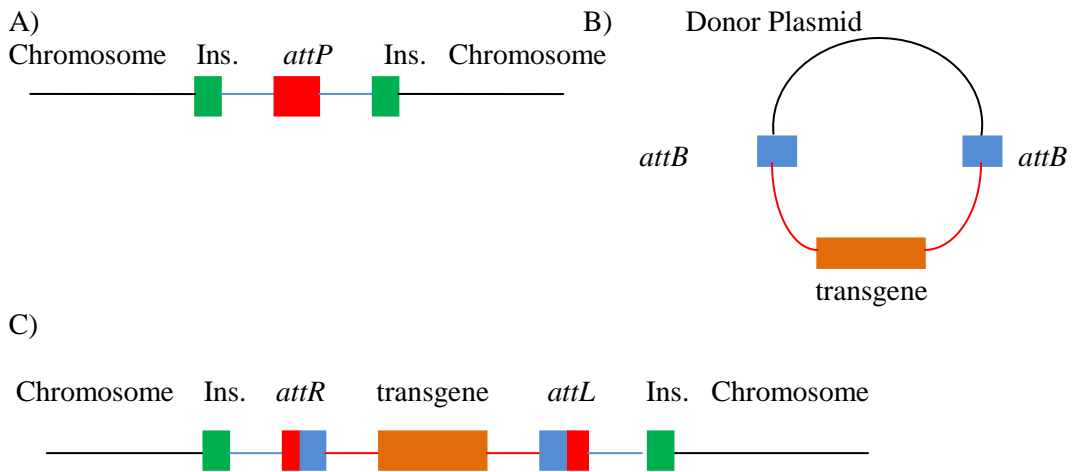
Analysis of the CTCF ChIP-Seq data revealed that some of the chromosome bands on the CTCF immunostained chromosome spreads correlated with the identified CTCF binding site peaks. Analysis of these CTCF binding site peaks and those that were validated by ChIP-PCR identified four motifs. These four motifs represent a small percentage of the total number of identified CTCF binding site peaks. This was unexpected considering that the consensus discovered for *Drosophila* CTCF represented nearly 100% of the identified sequences. This leads to the conclusion that *An. gambiae* CTCF binds to a wider variety of sequence motifs

throughout its genome. Further experiments will be necessary, including EMSA, to identify which sequences bind to CTCF.

This study only scratches the surface of potential future advances in mosquito transgenesis and improving our understanding of chromatin organization, as well as the role insulator proteins play in gene regulation. The multiple genomic positions of CTCF binding site peaks in relation to genes and one another provide opportunities to study gene regulation and chromatin organization in this species. One of the deficiencies of the approach taken to date is the lack of insight regarding CTCF long range interactions. Chromatin Immunoprecipitation (ChIP) Chromosome Conformation Capture (3C) analysis would provide an extra dimension to the knowledge this study has provided. In addition to the identification of CTCF binding regions, Su(Hw) and CP190 were shown to have similar expression profiles with CTCF across multiple life stages for *Ae. aegypti*, similar to the comparison of the expression profile data between CTCF and CP190 from the UC Irvine *An. gambiae* gene expression database [54, 61, 71, 78]. These data are consistent with the notion that insulator proteins bind adjacent to one another at some genomic locations, as has been shown in *Drosophila* [44], implying a cooperative role in regulating gene expression.

With regard to using CTCF binding site peak sequences to improve mosquito transgenesis, an enhancer blocking assay performed in cultured cells, similar to that performed by Li *et al.* [83] in S2 cells, would narrow down the candidate pool of potential sequences that may be used to flank a transgene to improve its expression. The candidate sequences with the most consistent and effective rates of enhancer blocking would be incorporated into a transgene construct flanking the transgene and any associated regulatory sequences. To test the effectiveness of the insulator sequences, insulated and uninsulated transgenes would be integrated into the embryonic germ line at the same chromosomal locations using the  $\Phi$ C31

integrase system [18] to ensure that transgene expression is evaluated within the same chromatin environment [39]. Effective insulator sequences could be used to create a strain of mosquitoes with insulated *attP* docking sites throughout the genome by randomly integrating *attP* sites flanked with insulator sequences. This would provide researchers with a strain of mosquitoes into which any effector or reporter transgene would be flanked by insulators when integrating transgenes with the  $\Phi$ C31 integrase system. Figure 13 illustrates how an insulated *attP* site would be used to insulate a site specifically integrated transgene flanked with *attB* sites using the  $\Phi$ C31 integrase system.



**Figure 13: Diagram of an insulated *attP* site.** A) An *attP* docking site randomly integrated into the genome with flanking insulator sequences (Ins.). B) A donor plasmid containing a transgene flanked by *attB* sites. C) The integrated transgene flanked by *attR* and *attL* sequences, formed by the recombination of the *attP* and *attB* sites, and the insulator sequences.



The distribution of CTCF binding sites identified in this work suggests multiple roles for CTCF throughout the genome. Many of the intergenic binding locations suggest a role as an insulator, while the CTCF binding site peaks found within promoter regions suggest that it may be acting as an activator or a repressor at these genomic locations. Also, a number of CTCF binding site peaks are within introns and exons suggesting an intragenic regulatory role. Individual functional assays will be necessary to identify the likely function of CTCF at specific genomic locations. Activation and repression activities can be assayed by inserting CTCF binding sequences proximal to reporter genes that will be transfected and incorporated into the chromosomes of either cultured cells or whole animals [79, 80]. Insulator assays similar to the one designed by Li *et al.* [83] can be used to determine enhancer-blocking function. Intragenic regulation could be tested by comparing gene expression in cells with differential CTCF binding patterns at the binding site peak in question. Importantly, CTCF binding site peak sequences can be inserted into multiple types of chromatin environments using the  $\Phi$ C31 integrase system to determine how the different chromatin contexts affect CTCF binding and ultimately CTCF function [84].

In *Drosophila*, CTCF, CP190, and Su(Hw) have been shown to have cell-type specific distributions throughout the genome [70]. These differences in insulator binding may be responsible for tissue specific gene expression. As suggested by the data in the current study and experiments in other organisms [65, 66], the expression of splice variants may be regulated by differential patterns of CTCF binding. To investigate this possibility, it would be necessary to perform ChIP-Seq for CTCF in multiple tissues. The neonate larval cells used in the current study provide a snapshot sample of CTCF throughout the genome, and there is some data that indicates that CTCF binding is conserved between the ChIP-Seq data from the neonate larval cell line and the immunostained ovarian nurse cell chromosomes. However, in order to identify cell-

type specific CTCF binding sites, more tissues will need to be assayed. Ovarian nurse cells are a likely candidate due to their importance in development and their increased expression level of CTCF in blood fed ovaries above other life stages [50]. The salivary glands and midgut would be a good choice due to their role in pathogen transmission. An understanding of how CTCF functions in these tissues will lead to a better understanding of the regulation of genes that may be useful for preventing pathogen transmission.

Comparisons of expression profiles of the insulator proteins CTCF and CP190 [54, 61, 71, 78] in *An. gambiae* showed similar patterns across nine life stages from larva through six life stages post blood feeding. In *Ae. aegypti*, gel based RT-PCR expression profiles from early embryo through blood fed females for Su(Hw), and through two stages of post blood fed ovarian tissue for CTCF [50] and CP190, showed increased levels of expression for early embryo, blood fed female and blood fed ovaries above all other life stages assayed. This data is consistent with the notion that CTCF colocalizes with CP190 and Su(Hw) at a subset of binding sites as is the case in *Drosophila* [44]. This being the case, further work can be performed to identify optimal insulator sequences that may contain binding sites for multiple insulator proteins. ChIP-Seq can be performed for CTCF in *Ae. aegypti* to identify insulator sequences in this species, and ChIP using antibodies for *An. gambiae* CP190 and Su(Hw) could be performed in both species, followed either by PCR with primers flanking a subset of the identified CTCF binding site peaks, or by performing Illumina parallel sequencing with the chromatin immunoprecipitated DNA to identify colocalization among the three insulator proteins. It would be interesting to identify sequences bound by one, two, and all three of the insulator proteins, and test them for insulator function using the assays outlined above to evaluate whether one or two of the three proteins, or a combination of all three bind to a sequence producing a more effective insulator than other potential insulator sequences. Ultimately, this work will be useful in identifying an ideal

insulator sequence that could flank an *attP* docking site and be used to create a mosquito strain ideal for mosquito transgenesis. Additionally, it will further understanding of the roles and interactions CTCF has in managing genome wide chromatin architecture and regulating gene expression.

## REFERENCES

1. Sinka, M.E., Michael J Bangs, Sylvie Manguin, Yasmin Rubio-Palis, Theeraphap Chareonviriyaphap, Maureen Coetzee, Charles M Mbogo, Janet Hemingway, Anand P Patil, William H Temperley, Peter W Gething, Caroline W Kabaria, Thomas R Burkot, Ralph E Harbach and Simon I Hay, *A global map of dominant malaria vectors*. Parasites and Vectors, 2012. **5**(69).
2. *10 facts on malaria*. 2009 March 2009 [accessed 2010 September 9, 2010]; Available from: <http://www.who.int/features/factfiles/malaria/en/>.
3. Holt, R.A., G.M. Subramanian, A. Halpern, G.G. Sutton, R. Charlab, D.R. Nusskern, P. Wincker, A.G. Clark, J.M. Ribeiro, R. Wides, *et al.*, *The genome sequence of the malaria mosquito Anopheles gambiae*. Science, 2002. **298**(5591): p. 129-149.
4. Staedke, S.G., P. Jagannathan, A. Yeka, H. Bukirwa, K. Banek, C. Maiteki-Sebuguzi, T.D. Clark, B. Nzarubara, D. Njama-Meya, A. Mpimbaza, P.J. Rosenthal, M.R. Kanya, F. Wabwire-Mangen, G. Dorsey, A.O. Talisuna, *Monitoring antimalarial safety and tolerability in clinical trials: A case study from Uganda*. Malaria Journal, 2008. **7**(107).
5. van Agtmael, M.A., T.A. Egelte, C.J. van Boxtel, *Artemisinin drugs in the treatment of malaria: from medicinal herb to registered medication*. Trends in Pharmacological Sciences, 1999. **20**(5): p. 199-205.
6. *Political unrest hampering Cote d'Ivoire's yellow fever vaccine campaign*. 2011 [accessed 2011 January 7, 2011]; Available from: [www.medicalnewstoday.com/articles/212945.php](http://www.medicalnewstoday.com/articles/212945.php).
7. *Dengue and dengue haemorrhagic fever*. 2009 [accessed 2010 9 September 2010]; Available from: [www.who.int/mediacentre/factsheets/fs117/en/](http://www.who.int/mediacentre/factsheets/fs117/en/).
8. *Climate change and human health: Risks and response*. 2003, World Health Organization. p. 17.
9. Paaijmansa, K.P., Andrew F. Read, and Matthew B. Thomas, *Understanding the link between malaria risk and climate*. PNAS, 2009. **106**(33): p. 13844–13849.
10. Christophides, G.K., *Transgenic mosquitoes and malaria transmission*. Cellular Microbiology, 2005. **7**(3): p. 325-333.
11. Marois, E., Christina Scali, Julien Soichot, Christine Kappler, Elena A. Levashina, Flaminia Catteruccia, *High-throughput sorting of mosquito larvae for laboratory studies and for future vector control interventions*. Malaria Journal, 2012. **11**(302).
12. Enserink, M., *GM mosquito trial alarms opponents, strains ties in Gates-funded project*. Science, 2010. **330**(6007): p. 1030-1031.

13. Wilke, A.B.B., *Control of vector populations using genetically modified mosquitoes*. Rev Saúde Pública, 2009. **43**(5): p. 869-874.
14. Catteruccia, F., *Malaria vector control in the third millennium: progress and perspectives of molecular approaches*. Pest Management Science, 2007. **63**(7): p. 634-640.
15. Thomas, D.D., Christl A. Donnelly, Roger J. Wood, Luke S. Alphey, *Insect population control using a dominant, repressible, lethal genetic system*. Science, 2000. **287**(5462): p. 2474-2476.
16. Coutinho-Abreu, I.V., Kun Yan Zhub, Marcelo Ramalho-Ortigao, *Transgenesis and paratransgenesis to control insect-borne diseases: Current status and future challenges*. Parasitology International, 2009. **59**(1): p. 1-8.
17. Amenya, D.A., M. Bonizzoni, A.T. Isaacs, N. Jasinskiene, H. Chen, O. Marinotti, G. Yan, A. A. James, *Comparative fitness assessment of Anopheles stephensi transgenic lines receptive to site-specific integration*. Insect Molecular Biology, 2010. **19**(2): p. 263-269.
18. Bischof, J., Robert K. Maeda, Monika Hediger, François Karch, and Konrad Basler, *An optimized transgenesis system for Drosophila using germ-line-specific  $\phi$ C31 integrases* PNAS, 2007. **104**(9): p. 3312–3317.
19. Bhalla, S.C., *White eye, a new sex-linked mutant of Aedes aegypti*. Mosquito News, 1968. **28**(3): p. 380-385.
20. Coates, C.J., Nijole Jasinskiene, Linda Miyashiro, and Anthony A. James, *Mariner transposition and transformation of the yellow fever mosquito, Aedes aegypti*. PNAS, 1998. **95**(3748–3751).
21. Jasinskiene, N., Craig J. Coates, Mark Q. Benedict, Anthony J. Cornel, Cristina Salazar Raffery, Anthony A. James, and Frank H. Collins, *Stable transformation of the yellow fever mosquito, Aedes aegypti, with the Hermes element from the housefly*. PNAS, 1998. **95**(7): p. 3743–3747.
22. Benedict, M.Q., *Mosaic: A position-effect variegation eye-color mutant in the mosquito Anopheles gambiae*. The Journal of Heredity, 2000. **91**(2): p. 128-133.
23. Nene, V., Jennifer R. Wortman, Daniel Lawson, Brian Haas, *et al.*, *Genome sequence of Aedes aegypti a major arbovirus vector*. Science, 2007. **316**(5832): p. 1718-1723.
24. Zdobnov, E.M., Christian von Mering, Ivica Letunic, David Torrents *et al.*, *Comparative genome and proteome analysis of Anopheles gambiae and Drosophila melanogaster*. Science, 2002. **298**(5591): p. 149-159.

25. Ribeiro, J.M.C., Roberto H. Nussenzveig, *The salivary catechol oxidase/oxidase activities of the mosquito Anopheles albimanus*. The Journal of Experimental Biology, 1993. **179**(1): p. 273-287.
26. *Aedes aegypti (Aedes aegypti): Assembly*. 2010 December 14, 2010 [accessed 2011 January 7, 2011]; Available from: [www.vectorbase.org/Aedes\\_aegypti/Info/Index](http://www.vectorbase.org/Aedes_aegypti/Info/Index).
27. Valenzuela, L., Rohinton T. Kamakaka, *Chromatin insulators*. Annual Review of Genetics, 2006. **40**(1): p. 107-138.
28. Eissenberg, J.C., Arthur J. Hilliker, *Versatility of conviction: heterochromatin as both a repressor and an activator of transcription*. Genetica, 2000. **109**(1-2): p. 19-24.
29. Perrod, S., S.M. Gasser, *Long-range silencing and position effects at telomeres and centromeres: parallels and differences*. Cellular and Molecular Life Sciences, 2003. **60**(11): p. 2303-2318.
30. Huisinga, K.L., Sarah C.R. Elgin, *Small RNA-directed heterochromatin formation in the context of development: What flies might learn from fission yeast*. Biochimica et Biophysica Acta, 2008. **1789**(1): p. 3-16.
31. Danzer, J.R., Lori Wallrath, *Mechanisms of HP1-mediated gene silencing in Drosophila*. Development 2004. **131**(15): p. 3571-3580.
32. Belyaeva, E.S., E.N. Andreyeva, S.N. Belyakin, E.I. Volkova, I.F. Zhimulev *et al.*, *Intercalary heterochromatin in polytene chromosomes of Drosophila melanogaster*. Chromsoma, 2008. **117**(5): p. 411-418.
33. Pirrotta, V., *Chromatin-silencing mechanisms in Drosophila maintain patterns of gene expression*. Trends In Genetics, 1997. **13**(8): p. 314-318.
34. Zhang, Y., R. Cao, L. Wang, and R.S. Jones, *Mechanism of polycomb group gene silencing*. Cold Spring Harbor Symposia on Quantitative Biology, 2004. **69**: p. 309-318.
35. Gaszner, M. and Gary Felsenfeld, *Insulators: exploiting transcriptional and epigenetic mechanisms*. Nature Reviews: Genetics, 2006. **7**(9): p. 703-713.
36. Marenduzzo, D., Ines Garo-Tindade, Peter R. Cook, *What are the molecular ties that maintain genomic loops?* Genetics, 2009. **23**(3): p. 126-133.
37. Kurukuti, S., Vijay Kumar Tiwari, Gholamreza Tavoosidana, Elena Pugacheva, Adele Murrell, Zhihu Zhao, Victor Lobanenkov, Wolf Reik, Rolf Ohlsson, *CTCF binding at the H19 imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to Igf2*. PNAS, 2006. **103**(28): p. 10684-10689.

38. Chetverina, D., Ekaterina Savitskaya, Oksana Maksimenko, Larisa Melnikova, Olga Zaytseva, Alexander Parshikov, Alexander V. Galkin, Pavel Georgiev, *Red flag on the white reporter: a versatile insulator abuts the white gene in Drosophila and is omnipresent in mini-white constructs*. Nucleic Acids Research, 2008. **36**(3): p. 929-937.
39. Markstein, M., Chrysoula Pitsouli, Christians Villalta, Susan E Celniker & Norbert Perrimon, *Exploiting position effects and the gypsy retrovirus insulator to engineer precisely expressed transgenes*. Nature Genetics, 2008. **40**(4): p. 476-483.
40. Lee, Bum-Kyu and V.R. Iyer, *Genome-wide Studies of CCCTC-binding Factor (CTCF) and cohesin provide insight into chromatin structure and regulation*. The Journal of Biological Chemistry, 2012. **287**(37): p. 30906–30913.
41. Ohlsson, R., Marek Bartkuhn and Rainer Renkawitz, *CTCF shapes chromatin by multiple mechanisms: the impact of 20 years of CTCF research on understanding the workings of chromatin*. Chromosoma, 2010. **119**(4): p. 351-360
42. Ohlsson, R., Victor Lobanenko, and Elena Klenova, *Does CTCF mediate between nuclear organization and gene expression?* BioEssays, 2010. **32**: p. 37-50.
43. Moon H, F., D. Loukinov, E. Pugacheva, Q. Chen, S.T. Smith, A. Munhall, B. Grewe, M. Bartkuhn, R. Arnold, L.J. Burke, R. Renkawitz-Pohl, R. Ohlsson, J. Zhou, R. Renkawitz, V. Lobanenko, *CTCF is conserved from Drosophila to humans and confers enhancer blocking of the Fab-8 insulator*. EMBO, 2005. **6**(2): p. 165-70.
44. Van Bortle, K., Edward Ramos, Naomi Takenaka, Jingping Yang, Jessica E. Wahi, and Victor G. Corces, *Drosophila CTCF tandemly aligns with other insulator proteins at the borders of H3K27me3 domains*. Genome Research, 2012. **22**: p. 2176-2187.
45. Holohan, E., Camilla Kwong, Boris Adryan, Marek Bartkuhn, Martin Herold, Rainer Renkawitz, Steven Russell, Robert White, *CTCF genomic binding sites in Drosophila and the organization of the bithorax complex*. PLoS Genetics, 2007. **3**(7): p. 1211-1222.
46. Ciavatta, D., Steve Rogers and Terry Magnuson, *Drosophila CTCF is required for Fab-8 enhancer blocking activity in S2 cells*. Journal of Molecular Biology, 2007. **373**(2): p. 233–239.
47. Smith, S.T., Priyankara Wickramasinghea, Andrew Olsonb, Dmitri Loukinovc, Lan Lina, Joy Denga, Yanping Xionga, John Ruxa, Ravi Sachidanandamd, Hao Suna, Victor Lobanenko, Jumin Zhou, *Genome wide ChIP-chip analyses reveal important roles for CTCF in Drosophila genome organization*. Developmental Biology, 2009. **328**(2): p. 518-528.
48. Witcher, M. and Beverly M. Emerson, *Epigenetic silencing of the p16<sup>INK4a</sup> tumor suppressor is associated with loss of CTCF binding and chromatin boundary*. Molecular Cell, 2009.**34**(3): p. 271-284.

49. Gray, C.E., Craig J Coates, *Cloning and chracterization of cDNAs encoding putative CTCFs in the mosquitoes, Aedes aegypti and Anopheles gambiae*. BMC Molecular Biology, 2005. **6**(16).
50. Gray, C.E., *Promoters, Enhancers and Insulators for Improved Mosquito Transgenesis*, in *Genetics*. 2005, Texas A&M University: College Station, TX. p. 176.
51. Heger, Peter, Birger Marin, Marek Bartkuhn, Einhard Schierenberg, and Thomas Wiehe, *The chromatin insulator CTCF and the emergence metazoan diversity*. PNAS, 2012. **109**(43): p. 17507-17512.
52. Bartkuhn, M.T.S., Martin Herold, Mareike Herrmann, Christina Rathke, Harald Saumweber, Gregor D Gilfillan, Peter B Becker and Rainer Renkawitz, *Active promoters and insulators are marked by the centrosomal protein 190*. The EMBO Journal, 2009. **28**(7): p. 877-888.
53. Park, P.J., *ChIP-seq: advantages and challenges of a maturing technology*. Nature Reviews: Genetics, 2009. **10**(10): p. 669-680.
54. Marinotti, O., E. Calvo, Q.K. Nguyen, S. Dissanayake, J. M. C. Ribeiro, & A. A. James, *Genome-wide analysis of gene expression in adult Anopheles gambiae*. . Insect Mol Biol, 2006. **15**(1): p. 1-12.
55. Sharakhova, M.V., Phillip George, Irina V. Brusentsova, Scotland C. Leman, Jeffrey A. Bailey, Christopher D. Smith, and Igor V. Sharakhov *Genome mapping and characterization of the Anopheles gambiae heterochromatin*. BMC Genomics, 2010. **11**(459).
56. Hertz, J., Jaques van Helden. *RSA-tools-patser*. Available from: [http://rsat.ccb.sickkids.ca/patser\\_form.cgi](http://rsat.ccb.sickkids.ca/patser_form.cgi).
57. O'Geen, H., Seth Fietze, and Peggy J. Farnham, *Using ChIP-Seq technology to identify targets of zinc finger transcription factors in Engineered Zinc Finger Proteins*. Methods in Molecular Biology, 2010. **649**: p. 437-455.
58. Ohlsson, R., Rainer Renkawitz and Victor Lobanekov, *CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease*. Trends In Genetics, 2001. **17**(9): p. 520-527.
59. Lutz, M., Les J. Burke, Pascal LeFevre, Fiona A. Myers, Alan W. Thorne, Colyn Crane-Robinson, Constanze Bonifer, Galina N. Filippova, Victor Lobanekov, and Rainer Renkawitz, *Thyroid hormone-regulated enhancer blocking: Cooperation of CTCF and thyroid hormone receptor*. The EMBO Journal, 2003. **22**(7): p. 1579–1587.



60. Blankenberg, D, A. Gordon, G. Von Kuster, N. Coraor, J. Taylor, A. Nekrutenko, Galaxy Team. *Manipulation of FASTQ data with Galaxy*. Bioinformatics, 2010. **26**(14):1783-5.
61. Dissanayake, S., O. Marinotti, J.M.C. Ribeiro, & A.A. James, *Anopheles gambiae gene expression database with integrated comparative algorithms for identifying conserved DNA motifs in promoter sequences*. BMC Genomics 2006. **7**(116).
62. Barski, A., Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E. Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao<sup>1</sup>, *High-resolution profiling of histone methylations in the human genome*. Cell 2007. **129**: p. 823-837.
63. Cuddapah, S., Raja Jothi, Dustin E. Schones, Tae-Young Roh, Kairong Cui and Keji Zhao, *Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains*. Genome Research, 2009. **19**(1): p. 24-32.
64. Heinz S, B.C., N. Spann, E. Bertolino, *et al.*, *Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities*. Molecular Cell, 2010. **38**(4): p. 576-589.
65. Shukla, S., Ersen Kavak, Melissa Gregory, Masahiko Imashimizu, Bojan Shutinoski, Mikhael Kashlev, Philipp Oberdoerffer, Rickard Sandberg, Shalini Oberdoerffer, *CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing*. Nature, 2011. **479**: p. 74-79.
66. Lai, A.Y., Mehrnaz Fatemi, Achana Dhasarathy, Christine Malone, Steve E. Sobol, Cissy Geigerman, David L. Jaye, Deepak Mav, Ruchir Shah, Liping Li, Paul A. Wade, *DNA methylation prevents CTCF-mediated silencing of the oncogene BCL6 in B cell lymphomas*. The Journal of Experimental Medicine, 2010. **207**(9): p. 1939-1950.
67. Mongin, E., Christos Louis, Robert A. Holt, Ewan Birney, Frank H. Collins, *The Anopheles gambiae genome: an update*. Trends in Parasitology, 2004. **20**(2): p. 49-52.
68. Sharakhova, M.V., Martin P Hammond, Neil F Lobo, Jaroslaw Krzywinski, Maria F. Unger, Maureen E. Hillenmeyer, Robert V. Bruggner, Ewan Birney and Frank H. Collins, *Update of the Anopheles gambiae PEST genome assembly*. Genome Biology, 2007. **8**(1): R5.
69. Euskirchen, G., Thomas E. Royce, Paul Bertone, Rebecca Martone, John L. Rinn, F. Kenneth Nelson, Fred Sayward, Nicholas M. Luscombe, Perry Miller, Mark Gerstein, Sherman Weissman, and Michael Snyder, *CREB Binds to Multiple Loci on Human Chromosome 22*. Molecular and Cellular Biology, 2004. **24**(9): p. 3804-3814.
70. Bushey, A.M., Edward Ramos and Victor Corces, *Three subclasses of a Drosophila insulator show distinct and cell type-specific genomic distributions*. Genes & Development, 2009. **23**(11): p. 1338–1350.

71. Sieglaff, D.H., W.A. Dunn, X.S. Xie, K. Megy, O. Marinotti, and A. A. James, *Comparative genomics allows the discovery of cis-regulatory elements in mosquitoes*. Proc Natl Acad Sci, 2009. **106**(9): p. 3053-3058.
72. Kumar, S., Alvaro Molina-Cruz, Lalita Gupta, Janneth Rodrigues, Carolina Barillas-Mury, *A peroxidase/dual oxidase system modulates midgut epithelial immunity in Anopheles gambiae*. Science, 2010. **327**(5973): p. 1644-1648.
73. Kuhn-Parnell, E.J., Cecilia Helou, David J. Marion, Brian L. Gilmore, Timothy J. Parnell, Marc S. Wold and Pamela K. Geyer, *Investigation of the properties of non-gypsy Suppressor of Hairy-wing binding sites*. Genetics, 2008. **179**(3): p. 1263–1273.
74. Barges, S., Jozsef Mihaly, Mireille Galloni, Kirsten Hagstrom, Martin Müller, Greg Shanower, Paul Schedl, Henrik Gyurkovics and François Karch, *The Fab-8 boundary defines the distal limit of the bithorax complex iab-7 domain and insulates iab-7 from initiation elements and a PRE in the adjacent iab-8 domain*. Development, 2000. **127**(4): p. 779-790.
75. Roth, F., et al., *Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation*. Nat. Biotechnol., 1998. **16**: p. 939-945.
76. Hanchang Sun, Y.Y., Yibo Wu, Hui Liu, Jun S. Liu and Hongwei Xie, *Tmod: toolbox of motif discovery*. Bioinformatics, 2010. **26**(3): p. 405-407.
77. Workman, C.T., Y. Yin, D.L. Corcoran, T. Ideker, G.D. Stormo, P.V. Benos, *enoLOGOS: a versatile web tool for energy normalized sequence logos*. Nucleic Acids Res., 2005. **Jul 1**(33(Web Server Issue)): p. W389-92.
78. Marinotti, O., Q.K. Nguyen, E. Calvo, A.A. James, J.M.C. Ribeiro, *Microarray analysis of genes showing variable expression following a blood meal in Anopheles gambiae*. Insect Mol Biol, 2005. **14**(4): p. 365-373.
79. Vostrov, A.A. and Wolfgang W. Quitschke, *The zinc finger protein CTCF binds to the APB $\beta$  domain of the Amyloid  $\beta$ -protein precursor Promoter. Evidence for a role in transcriptional activation*. The Journal of Biological Chemistry, 1997. **272**(52): p. 33353–33359.
80. Filippova, G.N., Sara Fagerlie, Elena M. Klenova, Cena Myers, Yvonne Dehner, Graham Goodwin, Paul E. Neiman, Steve J. Collins, and Victor V. Lobanenko, *An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of Avian and Mammalian c-myc oncogenes*. Molecular and Cellular Biology, 1996. **16**(6): p. 2802–2813.
81. Bell, A.C., Adam G. West, and Gary Felsenfeld, *The Protein CTCF is required for the enhancer blocking activity of vertebrate insulators*. Cell, 1999. **98**(3): p. 387-396.

82. Yao, S., Cameron S. Osborne, Rikki R. Bharadwaj, Peter Pasceri, Tanya Sukonnik, Dylan Pannell, Felix Recillas-Targa, Adam G. West and James Ellis, *Retrovirus silencer blocking by the cHS4 insulator is CTCF independent*. Nucleic Acids Research, 2003. **31**(18): p. 5317-5323.
83. Li, M., Vladimir E. Belozherov and Haini N. Cai, *Analysis of chromatin boundary activity in Drosophila cells*. BMC Molecular Biology, 2008. **9**(109)
84. Gerasimova, T.I., Keith Byrd, Victor G. Corces, *A chromatin insulator determines the nuclear localization of DNA*. Molecular Cell, 2000. **6**(5): p. 1025-1035.
85. Volz, J., Hans-Michael Muller, Agniexzka Zdanowicz, Fotis C. Kafatos, Mike A. Osta, *A genetic module regulates the melanization of Anopheles to Plasmodium*. Cellular Microbiology, 2006. **8**(9): p. 1392-1405.
86. Cirimotich, C.M., Yuemei Dong, Lindsey S. Garver, Shuzhen Sim, George Dimopoulos, *Mosquito immune defenses against Plasmodium infection*. Developmental and Comparative Immunology, 2010. **34**(4): p. 387-395.
87. Maeda, R.K. and F. Karch, *Chapter 1 The bithorax complex of Drosophila: an exceptional Hox cluster*, in *Current topics in developmental biology*, P. Olivier, Editor. 2009, Academic Press. p. 1-33.
88. S.J. Marygold, P.C. Leyland, R.L. Seal, J.L. Goodman, J.R. Thurmond, V.B. Strelets, R.J. Wilson and the FlyBase Consortium, *FlyBase: improvements to the bibliography*. Nucleic Acids Res., 2013. **41**(D1): p. D751-D757.
89. Maeda, R.K. and François Karch, *Making connections: boundaries and insulators in Drosophila*. Current Opinion in Genetics & Development, 2007. **17**(5): p. 394-399
90. Mohan M, B., M. Herold, A. Philippen, N. Heinl, I. Bardenhagen, J. Leers, R.A. White, R. Renkawitz-Pohl, H. Saumweber, R. Renkawitz, *The Drosophila insulator proteins CTCF and CPI90 link enhancer blocking to body patterning*. EMBO, 2007. **26**(19): p. 4203-14.

## APPENDIX

### SUPPLEMENTARY DATA

A-1

MACS parameters

#### ARGUMENTS

#### LIST:

band width = 200

ChIP-Seq file = /galaxy/main\_pool/pool5/files/003/628/dataset\_3628209.dat

control file = /galaxy/main\_pool/pool5/files/003/627/dataset\_3627165.dat

d = 46

effective genome size = 2.60e+08

format = BAM

model fold = 13

name = MACS\_in\_Galaxy

pvalue cutoff = 1.00e-05

Ranges for calculating regional lambda are : peak\_region,1000,5000,10000

tag size = 36

This file is generated by MACS

total tags in control: 29669847

total tags in treatment: 2245360

unique tags in control: 23661990

unique tags in treatment: 1159423

A-2

CTCF binding site peaks near genes of interest

\*=validated with ChIP-PCR

Peak ID	Start	End	# tags	Fold Enrichment	P-value $\times 10^6$	Annotation	Detailed Annotation	Distance to TSS	Nearest Promoter ID	Description of genes of interest
3L_301	33239277	33239404	11	19.15	-7.544	promoter-TSS (AGAP011779-RA)	NA	67	AGAP011779-RA	CLIPA5 cluster
*3L_302	33328699	33328803	6	12.39	-5.314	promoter-TSS (AGAP011798-RA)	NA	-447	AGAP011798-RA	CLIPA5 cluster
*2R_254	34439881	34440012	9	13.7	-6.484	Intergenic	NA	-10520	AGAP003244-RA	ClipB3(8 Clip genes)
2R_255	34499771	34499895	6	16.53	-5.198	TTS (AGAP003259-RB)	NA	985	AGAP003258-RA	ClipB3(8 Clip genes)
*3L_156	9015497	9015605	6	19.5	-5.623	promoter-TSS (AGAP010731-RA)	NA	32	AGAP010731-RA	CLIPA8
2R_75	7278745	7278901	34	17.93	-18.599	promoter-TSS (AGAP001648-RA)	NA	51	AGAP001648-RA	CLIPB17
*3R_30	3214284	3214400	6	11.04	-5.086	promoter-TSS (AGAP007937-RA)	NA	-106	AGAP007938-RA	CACT
*3R_31	3228650	3228758	7	16.93	-6.115	exon (AGAP007941-RA)	exon 2 of 8	349	AGAP007941-RA	CACT & AGAP007941
2R_380	48912076	48912207	7	17.87	-6.179	intron (AGAP004052-RA)	intron 1 of 5	35728	AGAP004052-RA	dblsx AGAP004050
2R_379	48627045	48627123	9	16.9	-6.484	promoter-TSS (AGAP004047-RA)	NA	34	AGAP004047-RA	dblsx AGAP004050
X_61	9532791	9532909	6	24.37	-5.83	Intergenic	NA	5938	AGAP013283-RA	msl-2

A-2 continued

\*=validated with ChIP-PCR

Peak ID	Start	End	# tags	Fold Enrichment	P-value $1 \times 10^{\wedge}$	Annotation	Detailed Annotation	Distance to TSS	Nearest Promoter ID	Description of genes of interest
X_62	9567412	9567541	6	21.54	-5.232	Intergenic	NA	12321	AGAP000534-RB	msl-2
2R_336	45747249	45747384	6	16.05	-5.089	intron (AGAP003901-RA)	intron 2 of 10)	5809	AGAP003901-RA	sxl
*2R_335	45716399	45716519	8	13.28	-5.145	promoter-TSS (AGAP003899-RA)	NA	671	AGAP003899-RB	sxl
2R_334	45688191	45688311	16	78	-22.94	TTS (AGAP003897-RA)	NA	2140	AGAP003897-RA	sxl
*2R_521	60012716	60012822	6	14.62	-6.002	Intergenic	NA	91094	AGAP004660-RB	BC-X
2R_522	60297839	60297962	7	24.37	-6.553	TTS (AGAP004663-RA)	NA	344	AGAP004663-RA	BC-X
2R_523	60312345	60312475	9	8.04	-5.221	Intergenic	NA	14854	AGAP004663-RA	BC-X
2R_524	60405670	60405838	6	18.02	-5.314	Intergenic	NA	16512	AGAP004664-RA	BC-X
*2R_525	60597396	60597506	6	19.5	-5.579	Intergenic	NA	89531	AGAP004665-RA	BC-X
3R_347	42680969	42681106	7	18.63	-8.097	Intergenic	NA	-4227	AGAP009769-RA	GPR-CAL1
3R_348	42700001	42700083	15	12.23	-8.212	Intergenic	NA	-6579	AGAP009770-RA	GPR-CAL1
*2L_330	44476004	44476157	7	24.37	-6.175	intron (AGAP007237-RA)	intron 1 of 8	4728	AGAP007237-RA	HPX 4
*3L_182	12724084	12724201	8	22.18	67.45	Intergenic	NA	-8799	AGAP010895-RA	HPX 10,11
3L_161	10914598	10914701	65	8.97	-18.841	Intergenic	NA	5738	AGAP010811-RA	HPX 15,14
3R_168	24362783	24362914	6	14.52	-5.785	TTS (AGAP009033-RA)	NA	2167	AGAP009033-RA	HPX 2
*2L_331	44505634	44505763	8	18.84	-5.778	exon (AGAP007238-RA)	exon 2 of 2	1033	AGAP007238-RA	HPX 4
*3L_183	12860633	12860749	6	18.43	-5.314	Intergenic	NA	-1749	AGAP010909-RA	HPX 10,11
3L_160	10576666	10576751	14	19.83	-8.51	intron (AGAP010800-RA)	intron 4 of 4	6779	AGAP010800-RA	HPX 15,14