The organization and evolution of heterochromatin in the *Anopheles gambiae* complex

Phillip George

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Entomology

Igor V. Sharakhov, Chair
Daniela Cimini
Jinsong Zhu
Carlyle C. Brewster
Sally L. Paulson
Loke T. Kok, Department Head

February 24th, 2014
Blacksburg, Virginia

Keywords: *Anopheles gambiae*, heterochromatin, piRNA, polytene chromosomes

Copyright 2014, Phillip George

The organization and evolution of heterochromatin in the *Anopheles gambiae* complex

Phillip George

Abstract

The *Anopheles gambiae* complex is comprised of the most important vectors of malaria in Sub-Saharan Africa. Most current control methods involve the use of chemicals that help to limit potential contact with these mosquitoes. However, these control methods still have risks that include insect resistance, environmental toxicity, human health, as well as animal health. In order to develop new strategies that either produce novel targeted insecticides or transgenic mosquitoes that can replace current mosquito populations, it is important to acquire as much biological information about the vector as possible. The reduction in cost and speed of high-throughput sequencing has brought forth many new sequenced genomes that can provide a wealth of information about individual populations as well as their respective evolutionary histories. However, in order to fully understand a genome, these sequences must be assembled properly.

One of the largest challenges toward fully assembling a genome is the abundance of repetitive sequences. These sequences, typically part of gene poor regions known as heterochromatin, are generally left as unassembled scaffolds that are neglected in many genomic studies. Heterochromatin is a biologically important chromatin state that has roles in gene regulation and genome stability. Exclusion of these chromatin domains from experimental assays can provide an incomplete picture in regards to organismal biology. A lack of information regarding heterochromatin, even in *An. gambiae*, necessitates further understanding and characterization of this chromatin type that can provide valuable information about the mosquito's biology.

Heterochromatin is organized differently amongst different species. Some species with compact genomes, like *Drosophila melanogaster*, exhibit rigid organization of heterochromatin, with repetitive elements being confined to peri-centromeric and sub-telomeric regions of the chromosome. Larger genomes such as *Aedes aegypti*, have a much less structured heterochromatin pattern, with repetitive elements being dispersed across the genome. However, *An. gambiae*'s genome is more intermediate in size as well as transposable element content. These factors may have an impact in controlling how heterochromatin is organized within the *An. gambiae* genome. Does *An. gambiae* compensate for the increased genome size by expanding past the peri-centromeric heterochromatin into new intercalary compartments?

In *An. gambiae*, heterochromatin had yet to be identified separately from euchromatin. Morphologically, some regions of *An. gambiae* chromosomes exhibited characteristics similar to transcriptionally active puffs or peri-centromeric heterochromatin. We characterize these regions, as well as the rest of the genomic landscape, by using morphological and genetic features to identify various chromatin types. Peri-centromeric heterochromatin and new regions of intercalary heterochromatin were identified. Genomic coordinates representing the transition from euchromatin to heterochromatin were also identified. By finding these heterochromatin-euchromatin boundaries, various genetic features could be assigned to either heterochromatin or euchromatin. Critical genes associated with heterochromatin formation and basic genomic functions were identified. These data help to better understand features that are associated with the different environments created by chromatin compaction.

This study also looks at the Piwi-interacting RNA (piRNA) pathway and its role in *An. gambiae*. The piRNA pathway is associated with transposable element (TE) suppression in many species, where clusters of vestigial TEs provide some of the RNA necessary for the pathway to function. These clusters are primarily associated with heterochromatin in *Drosophila melanogaster*. We identify piRNA clusters in *An. gambiae* and see a similar shift from primarily peri-centromeric compartmentalization toward the presence of intercalary regions located within the euchromatin. Transposable elements are maintained in secondary heterochromatin regions that exhibit similar morphology and features to peri-centromeric

heterochromatin. The piRNA pathway also has implications in gene regulation, germline development, and anti-viral immunity. Three candidate genes associated with spermatogenesis and embryogenesis have been identified. These genes showed piRNA enrichment, and upon further analysis show up-regulation after a blood meal is taken. These genes could potentially prove useful in vector control as targets of transgenic experiments.

Heterochromatin is an important, yet neglected aspect of the genome. These studies attempt to provide data to stimulate the study of heterochromatin through characterization of heterochromatin-related genomic features.

Acknowledgements

I would like to thank my advisor, Dr. Sharakhov, for providing the opportunity to begin my career in research in his lab. You've been a steadying presence for the last six years that has helped to keep me focused. I cannot say that, when I began my research in your lab, I would have wanted to pursue a career in cytogenetics. However, your passion coupled with the atmosphere you projected in the lab has really opened my eyes to the potential impact of this kind of science. For that, I am forever grateful. I hope that other graduate students will be able to see the qualities that make you such a fantastic mentor and decide to continue their studies in your lab.

To my committee members- Dr. Carlyle Brewster, Dr. Daniela Cimini, Dr. Sally Paulson, and Dr. Jinsong Zhu- I sincerely want to thank you for all of the support and advice you have given me over the past five and a half years. There have been times where I have felt a bit lost or confused about the direction of my work, but all of you have had encouraging words of wisdom that have put me back on track to finish my projects. You have made a lasting impression on the importance of understanding the bigger picture and importance of my work.

I'd like to thank the members of the Sharakhov lab, both past and present, for all of the help and support they have given me. It has been a very demanding five and a half years, and all of you certainly have made those years much more manageable and fulfilling.

To Dr. Ai Xia, thank you for getting me acclimated to graduate school. I was thoroughly unprepared for what to expect, but you helped me in many ways to prepare for the struggles associated with doctoral research.

To Dr. Maryam Kamali, we had a good run together! It was a pleasure to have spent a majority of my time in the lab with such a sweet, motivated person. You have always been there when I needed someone to talk to, and I cannot say how much this means to me. Your support has always meant a lot to me.

To Ashley Peery, you have been a true source of inspiration for me over the years. When you came to the lab, I was beginning to feel burned out in regards to my research. Your eagerness and hard-working attitude helped me persevere through those down times. You've been a great friend and I really appreciate all of the support and help you've given me.

To Atashi Sharma, thank you for being a great friend. All of your support has alleviated a lot of the academic tension that I've had. I'll miss all of the randomness that helped to brighten my days. I'm quite lucky to be able to call you a friend.

To Nick Kinney, I can't thank you enough for all of the help you've provided. I'm not sure how some of this work would have turned out without your help and support.

To Vladimir Timoshevskiy, Anastasia Noumenko, and Maria Sharakhova, I cannot thank you enough for all of the help you have provided me. I have learned so much from all of you both science- and life- related, and only wish I could return the favor.

To all of the undergraduates that I have worked with over the years, thank you all for helping to complete my projects. It was an honor to pass on some of my knowledge to all of you.

To Dr. Leman, I appreciate the time you have spent helping with the statistical analyses of my datasets. It was a pleasure to work with you and hear some of your incredible life stories.

To all of the friends that I have made during my time here, both science related and not, I truly appreciate having you all in my life. Calvin Waldron, Lawrence Stawicki, Kristoefer Clegg, Shawn Savage, Naeem Thompson, Dr. Rebecca Ortiz, Dr. Titi Denloye, Dr. Brenna Traver, Gade Kimsawatade, Dr. Nikki Lewis, Dhaval Khumar, Lindsay Key, and Mikolaj Krol- you all have been incredible friends, and I wish you all the best in the future. Jessica Bowman, Christian Laourdakis, Sanjay Basu, Lisa Moore, John Machen, Frank Criscione, Twyla Smith, and Scott O'Neill- thank you for making my days in lab that much better. Being around all of you has been truly wonderful, and I am lucky to have been around such a fantastic group of people.

Finally, I would like to thank my parents for everything they have done. I can't imagine I'd be in the position that I am in now without all of your support and understanding. I will be forever grateful and love you both. Thank you for pushing me to succeed and strive to make an impact in the world.

Table of Contents

Chapter 1. Introduction **1**

1.1. Advances and limitations of malaria control ***1***
1.1.1. Overview of malaria 1
1.1.2. Classical malaria prevention/control methods 3
1.1.3. Genetic approaches to vector control 6

1.2. Genomics and cytogenetics of Anopheles gambiae ***9***
1.2.1. The *Anopheles gambiae* complex 9
1.2.2. *Anopheles gambiae* chromosomes 11
1.2.3. *Anopheles gambiae* genome assembly 13

1.3. Epigenomics ***14***
1.3.1. Chromatin types 15
1.3.2. Transposable elements (TEs) and their role in the genome 18
1.3.3. Diverse functions of the Piwi-interacting RNA (piRNA) pathway 23

1.4. Methods for the analysis of individual chromosomes ***29***
1.4.1. Laser capture microdissection (LCM) 29
1.4.2. Whole genome amplification (WGA) 32
1.4.3. PCR-based amplification 33
1.4.4. Multiple displacement amplification (MDA) 34
1.4.5. Chromosome painting 36
1.4.6. Sequencing of microdissected chromosomes 38

Chapter 2. Identification and classification of chromatin types in *Anopheles gambiae s.s.* **40**

2.1. Introduction ***40***

2.2. High-throughput Physical Mapping of Chromosomes using Automated in situ Hybridization ***41***
2.2.1. Abstract 41
2.2.2. Protocol 43
2.2.3. Representative Results 47
2.2.4. Discussion 53

2.3. High-Resolution Cytogenetic Map for the African Malaria Vector Anopheles gambiae ***56***
2.3.1. Abstract 56
2.3.2. Introduction 57
2.3.3. Results 59
2.3.4. Discussion 66
2.3.5. Experimental Procedures 68

2.4. Genome mapping and characterization of the Anopheles gambiae heterochromatin ***71***
2.4.1. Abstract 71
2.4.2. Background 72
2.4.3. Results and Discussion 76
2.4.4. Methods 103

2.5. A molecular feature based model for the prediction of chromatin types in An. gambiae *106*
2.5.1. Abstract 106
2.5.2. Introduction 107
2.5.3. Results and Discussion 111
2.5.4. Methods 120

Chapter 3. Organization of piRNAs and evolution of piRNA clusters in the African malaria vector *Anopheles gambiae* **123**

3.1. Summary *123*

3.2. Introduction *125*

3.3. Results *129*
3.3.1. Small RNA libraries of *An. gambiae* 129
3.3.2. piRNA clusters 132
3.3.3. TE composition of clusters 136
3.3.4. piRNAs mapping to genes 137
3.3.5. Piwi immunostaining 141

3.4. Discussion *143*
3.4.1. Evolution of piRNA spatial organization 143
3.4.2. Spatial organization within closely related forms 145
3.4.3. Likely biological functions of genes enriched in piRNAs 145

3.5 Methods *146*
3.5.1. Purification of small RNAs and library construction 146
3.5.2. Library processing 147
3.5.3. Mapping sequences to reference genomes 147
3.5.4. TE, gene ontology (GO), and expression data analysis 148
3.5.5. PIWI antibody generation and immunostaining 148

Chapter 4. Summary **149**

4.1. General review *149*

4.2. Review of chapter 2 ..... *150*

4.3. Review of chapter 3 ..... *152*

Bibliography **154**

Appendix **177**

List of Tables

Table 2.1. BLAST results of the BAC-end sequences against the genome assemblies of An. gambiae.....49

Table 2.2. Major landmarks of the banding pattern for the An. gambiae chromosomal arm recognition 61
Table 2.3. Chromosomal landmarks for recognizing common polymorphic inversions in An. gambiae.....64

Table 2.4. Localization of HP1 and lamin on An. gambiae chromosomes 82
Table 2.5. Boundaries between heterochromatin and euchromatin in the An. gambiae genome 85
Table 2.6. Coverage (%) of molecular elements in chromatin types of An. gambiae 91

Table 2.7. Molecular features used in initial chromatin prediction model.....112
Table 2.8. Molecular features used in final chromatin prediction model.....116
Table 2.9. Novel intercalary heterochromatin regions found on the autosomes 119

Table 3.1. List of An. gambiae annotated genes with more than 150 piRNAs mapping to them 138
Table 3.2. Table 3.2 Identification of PIWI and HP1 localization sites in An. gambiae via immunolocalization.....143

Table S1. Genome coordinates of DNA markers mapped to polytene chromosomes of Anopheles gambiae ...179

List of Figures

Figure 1.1. – piRNA biogenesis through primary amplification 27
Figure 1.2. - Ping-pong amplification of piRNAs 27

Figure 2.1. - Schematic representation of high-pressure chromosome preparation 50
Figure 2.2. A scheme representing automated FISH, slide scanning, and chromosome mapping of genomic scaffolds 51
Figure 2.3. A spread of polytene chromosome 3 of An. gambiae obtained using the high-pressure technique 51
Figure 2.4. Comparison of An. gambiae polytene chromosomes prepared using the traditional (top) and high-pressure (bottom) technique 52
Figure 2.5. FISH of BAC clones to polytene chromosomes of An. gambiae 52

Figure 2.6. Cytogenetic photomap of An. gambiae ovarian nurse cell chromosomes 60
Figure 2.7. Comparison of An. gambiae polytene chromosome spreads prepared using the high-pressure (top) and traditional (bottom) technique 60
Figure 2.8. Fluorescent in situ hybridization of fragments of heterochromatic genes to the polytene chromosomes 2L and 3L of An. gambiae 66

Figure 2.9. The pericentric and intercalary heterochromatin of polytene chromosomes shown on a standard cytogenetic map of An. gambiae 78
Figure 2.10. Localization of HP1 and lamin Dm0 Drosophila antibodies on An. gambiae chromosomes 79
Figure 2.11. Schematic representation of the heterochromatin amount in the An. gambiae genome 86
Figure 2.12. Overrepresented GO terms in genes within the cytologically confirmed heterochromatin (a) and within "unknown chromosome" (b) of An. gambiae 89
Figure 2.13. Median values of gene density and repetitive element coverage in chromatin types of An. gambiae 93
Figure 2.14. Genome landscapes of the An. gambiae heterochromatin and euchromatin 95
Figure 2.15. Median values of gene density and repetitive element coverage in "unknown chromosome" of An. gambiae 101

Figure 2.16. Transposable element and gene content across the An. gambiae landscape 113
Figure 2.17. Recent TE content across the An. gambiae landscape 114

Figure 3.1. piRNA and TE distribution across An. gambiae genome assembly 130
Figure 3.2. Distribution of transposable elements in the M form of An. gambiae 131
Figure 3.3. Distribution of piRNA clusters across An. gambiae M and S form chromosomes 133
Figure 3.4. The composition of the top 15 piRNA producing clusters 137
Figure 3.5. *Gene ontology (GO) term analyses of gene subset with more than 150 piRNA enrichment* 139
Figure 3.6. Expression analysis of reproduction related GO term annotation cluster genes 140
Figure 3.7 Immunolocalization of Piwi1 and HP1 antibodies to chromosomal squashes 142

Chapter 1. Introduction

1.1. Advances and limitations of malaria control

Most current efforts to control malaria involve the use of insecticides that are sprayed indoors, either on the walls, or on bed nets that protect users while they sleep. While effective, these methods will ultimately not result in the elimination of malaria. These methods, as well as other techniques that target larval stage mosquitoes, are expensive and require local adoption in order to be effective. Genetic modification presents an alternative option that is not only species-specific, but is also biologically-friendly toward the environment. Until an inexpensive vaccine is generated that can be provided across third-world countries, new methods must be developed in order to reduce the global burden of malaria.

1.1.1. Overview of malaria

Malaria, or "bad air," has plagued humans for the past 4000 years. The first published cases of malaria date back to ancient China, where approximately in 2700 BC, Chinese writings including *Nei Ching, The Canon of Medicine* depicted an illness with the symptoms of malaria that were caused by biting insects. Malaria is one of the world's most common life threatening diseases. According to the World Health Organization [1], in 2010 a predicted 216 million cases of malaria were reported- 81% of those cases occurring in Africa. Of those 216 million cases, an estimated 655,000 ultimately ended in death. 91% of these deaths occurred in Africa [1]. Malaria, however, is not constrained only to the African continent, it is also currently endemic in South-East Asia and South America [2]. A predicted 95+ countries ranging from Africa, Central and South America, and South-East Asia all have shown cases of human malaria [3-5]. These data support that approximately 3 billion+ people are at risk to suffer from malaria, with a higher

incidence for those living in tropical countries. The most at risk populations in these countries include young children, pregnant women, and those with HIV/AIDs.

Malaria is caused by protozoan parasites belonging to the genus *Plasmodium.* The human disease is caused by parasites belonging to one of five species of *Plasmodium*: *P. falciparum, P. vivax, P. ovale, P. malariae* [6], and most recently *P. knowlesi* [7,8]. There are over 120 other species of *Plasmodium*; however, few are considered a serious threat to humans [9]. The two primary species of *Plasmodium* responsible for malaria worldwide are *P. vivax* and *P.* falciparum. Geographically, *P. vivax* is the most widespread of the parasites [3-5], spanning a majority, if not the entirety, of the countries at risk for malaria transmission, while *P. falciparum,* although detected in 85 countries [10], is primarily of concern in sub-Saharan Africa. *P. ovale* is predominantly found in West Africa, but can span across Asia and Africa [11]. Distribution of *P. malariae* is fairly similar to that of *P. falciparum* [12], and *P. knowlesi* is localized in South-East Asia [8].

Approximately 3,500 species of mosquitoes exist worldwide. These mosquitoes are classified into three subfamilies: *Anophelinae, Culicinae,* and *Toxorhynchitinae*. The mosquitoes serving as vectors of the malaria parasite are all anophelines. Of over 450 species of *Anopheles,* 70 are known to be competent vectors of malaria [13]. Primary vectors based on region include: *A. darlingi* - South America, *A. stephensi, A. sinensis* - South East Asia, and *A. albimanus* in Central America [14]. Members of the *Anopheles gambiae* complex are primarily responsible for the spread of malaria across the African continent, with *A. arabiensis,* and *A. gambiae s.s.* being the most important vectors within the complex. *A. funestus,* outside of the *A. gambiae* complex, also is an important vector in Africa. This complex, in conjunction with the

other *Anopheline* mosquitoes present in Asia, provides a major challenge in controlling the spread of malaria.

1.1.2. Classical malaria prevention/control methods

Multiple combinations of different vector control methods including insecticidal regulation, source reduction control of larval populations, and biological control of both larval and adult populations; as well as human protection via repellents, chemoprophylaxis, and the potential for malaria vaccinations are currently implemented to avert further spread of the disease. Researchers estimate that since the early 1900's, the affected area of global malaria has been reduced by approximately 50%, from 53% to 27% of the land on Earth [15] through these techniques. However, these methods of mosquito control are still limited in their impact on malaria. Current control methods may help to reduce the burden of malaria, but it is clear that other intervention techniques are necessary in order to relieve areas of the disease. Many different approaches are being taken to reduce the malaria load to a level that is low enough to not result in regional epidemics. Many of these approaches involve the use of insecticides and other chemicals that attempt to kill malaria vectors. However, these strategies may not be able to fully combat malaria as mosquitoes continue to evolve.

The Global Malaria Eradication Program, assembled in 1955 to eradicate malaria in endemic countries excluding sub-Saharan Africa and Madagascar [16], successfully eliminated malaria from North America, Europe, parts of South-Central America, parts of Asia, and the Caribbean [17] through the combinatory use of DDT, residual spraying and surveillance. A total of 37 endemic countries successfully eliminated malaria, and many countries saw a marked decrease in malaria cases [16]. DDT, the main insecticide used during the elimination program,

was cited as a harmful toxin to wildlife, the environment, and humans [18] and subsequently banned from use in 1972.

Indoor residual spraying (IRS) and insecticide treated nets (ITNs) are becoming more commonly used within homes in Africa. IRS is an effective method for killing endophilic (indoor resting) and/or endophagic (indoor biting) mosquitoes. A majority of mosquitoes that come in contact with the insecticide treated walls and ceilings will be killed. ITNs are also most effective against endophilic and/or endophagic vectors. ITNs and IRS are both very effective; however, these methods are not enough to completely remove the burden of malaria. Both IRS and ITN require local adopters that can afford these vector control methods. Vectors of malaria are adapting to current vector control methods, becoming more resistant to insecticides that are currently implemented [19]. With the mosquitoes' burgeoning resistance to these insecticides, it is necessary to rotate through multiple chemicals while developing new ones in order to maintain a high level of vector control. These treatments also neglect exophilic mosquitoes that don't enter homes. Outdoor spraying, mentioned above, targets the exophilic mosquitoes that are not killed by indoor treatments.

Where insecticidal control of mosquitoes target adult mosquitoes, particularly those that are endophilic, the objective of source reduction is to diminish larval population levels. Source reduction is the process of eliminating likely egg laying and breeding sites, effectively removing areas where mosquitoes can develop from larvae into adults. Destruction of these sites is both cheap and cost effective. However, in order for source reduction to be effective, it is necessary to have knowledge of the vector and what its preference toward breeding sites is. Other drawbacks for source reduction include having to remove larger pools of water, as well as the inability to locate and identify smaller sites. Chemical larvicides have also been put into practice

to lower larval populations. Placing larvicides into known breeding grounds can be much easier than removing the water source; however, they can still prove a threat to both the environment and organisms located nearby. Bio-pesticides can also be used as a more natural form of source reduction. Bio-pesticides, the use of naturally occurring substances and organisms like the bacteria *Bacillus thuringiensis* [20,21], fungus *Coelomomyces* [22], nematodes [23,24], microsporidian *Nosema* [25], and natural predators like larvivorous fish [26] offer another method for controlling mosquito populations, while typically being less toxic to the environment. Three species of larvivorous fish, *Poecilia reticulata* (Guppy), *Puntius bimaculatus* (Ipilli Kadaya) and *Rasbora caveri* (Dandiya) showed complete or near complete (80%) reduction of larvae when subjecting one to three fish in a population of *Aedes aegypti* [26]. Bio-pesticides can be cheaper and more cost effective than traditional pesticide management; but, typically work slow, can be ineffective, and in some cases, present the opportunity for vector resistance. The introduction of foreign organisms into habitat can negatively impact the existing ecosystem. Bio-pesticides must be applied frequently in order to maximize effectiveness, and lack control of pupal and adult stages of mosquito development.

Classical vector control approaches represent an effective way to regulate vector populations in regions that can afford to maintain them. ITNs and IRS work very well in protecting people inside their houses. However, many of the regions that need these preventative strategies the most are some of the poorest worldwide. In addition to expenses, mosquitoes are developing resistance to insecticides, or adapting to avoid control measures altogether. Larval control through habitat destruction, in theory, could completely wipe out populations of mosquitoes. However, it is nearly impossible to identify all of the minute breeding sites that can produce large populations of mosquitoes very rapidly. Bio-pesticides could also severely cripple

mosquito populations. However, costs combined with potential ecosystem impact limit the effectiveness of this control method as well. Deficiencies of these classical approaches to vector control call for searching novel methods such as those that rely on mosquito genetics.

1.1.3. Genetics approaches to vector control

Genetic approaches to vector control likely will become the best way to control populations of vectors. With the increasing library of information in regards to organism genetics, genes associated with traits like vector competence and sex differentiation are becoming identified. These genes serve as excellent targets for genetic control studies that can lead to vector control.

Manipulating the genetics of current Anopheline populations could be one of the major candidate approaches for a more permanent form of vector control. Two major classes of this control exist: population suppression and population replacement. Both methods involve the use of transgenic mosquitoes in order to reduce population numbers (population suppression) or reduce vector competence (population replacement) of a natural population.

Sterile insect technique (SIT) provides a method of biologically controlling vector populations. The goal of SIT is to release irradiated or chemically sterilized males into natural populations, with the idea that mating with wild females will reduce total future progeny. The process is considered to be safe, environmentally friend, and effective [27]. SIT has been used since the 1950's with varying results. In the early 1960's, sterilized *A. quadrimaculatus* [28] and *Aedes aegypti* [29] were released in Florida with little overall effect. Reasons for ineffectiveness included changed mating behaviors and reduced competitiveness from the radiation treatment.

However, releases of *Culex quinquefasciatus* [30] in India and Florida and *Culex tarsalis* [31] in California proved to be more effective in terms of competitiveness. Although SIT is an eco-friendly method of biological control, many factors can lead to ineffective results, including optimization of sterilization protocols, mass rearing of the target species, and competitiveness of the males.

A form of sterile insect technique, 'Release of Insects carrying a Dominant Lethal' or RIDL, involves the insertion of a promoter controlled dominant-lethal gene into females, which kills female progeny [32]. The inserted gene can be later removed by the treatment of tetracycline. RIDL is much like sterile insect technique in that it targets to lower populations by reducing progeny numbers via affected parents. This method is highly species specific, environmentally safe, relatively cheap, and highly efficient [33]. Compared to the technology needed for population replacement, techniques like SIT and RIDL are much simpler. The mechanism behind RIDL also could address one of the limitations of standard irradiation/chemosterilization SIT: the lack of a need for sexing during mass-rearing. Sterilized females are seen as a detriment to SIT [34], and thus excluding the need for sexing out females would create a more efficient and effective program. RIDL can potentially remove the need for sexing, as female-specific lethality pre-eclosion will naturally remove females from the population. Without the need for irradiation, the overall fitness of the sterile males should be much more in-line with wild population mosquitoes. The efficiency of RIDL is proposed to be increased as more dominant-lethal gene loci are inserted into the target as reviewed by Black *et al.* [35].

Genetic factors have already been linked to host feeding habits and transmission competence of *Plasmodium* in mosquitoes [36]. Some members of the *A. gambiae* complex are highly efficient at transmitting malaria, while others are not. By determining the genetic basis for

the difference between those highly competent and less competent species, transgenic mosquitoes with genes helping to lower overall transmission capability can be made.

Manipulating the vector to block amplification and transmission of the *Plasmodium*, involving antibodies targeting the parasite, and taking advantage of genes in natural populations that confer refractoriness are all potential modes of creating transgenic mosquitoes that cannot transmit malaria to humans [37]. By enhancing the immune response to the *Plasmodium* parasite in the mosquito, researchers can potentially create mosquitoes that cannot efficiently transmit malaria. Proteins encoded by genes like TEP1 [38], APL1 [39], LRRD7 [40], and FBN9 [41] have all been linked to the immune response against *Plasmodium* infection. Creating transgenic mosquitoes with a phenotype refractory to malaria transmission suffers from the same drawbacks as transgenics created using RIDL. Mosquitoes must be able to out-compete wild-type mosquitoes, and be produced efficiently.

Genetic manipulation is a biologically-friendly method for the control of vector populations. These strategies are species specific, do not introduce exotic species or agents into nature, and would not have harmful side effects on human populations. However, there are still drawbacks. Releasing sterile insects requires a high level of fitness to outcompete natural male populations. It is also typically necessary to constantly introduce more insects to keep population levels low. Genetically modified mosquitoes can suffer from the same drawbacks. Modified mosquitoes must be able to out-compete wild-type mosquitoes; otherwise the modification will not be fixed within the population. In the case of both SIT and genetic modification, public opinion can limit the effectiveness of these techniques as well. Genetic modification still is an issue with many people, so introduction of modified populations of insects can create political issues. The success in the development of novel genetics-based

approaches to malaria control will depend on the availability of genomic information and on our understanding of genetic processes.

1.2. Genomics and cytogenetics of *Anopheles gambiae*

In order to develop genetic control studies that may help reduce the burden of malaria, genomic and cytogenetic studies must be performed to understand the biology behind the vectors of malaria. The *Anopheles gambiae* complex provides an excellent model for cytogenetic and genomic studies, while also being an economically important group of organisms. Members of this complex represent the largest threat of malaria in Sub-Saharan Africa. Gene targets that allow some of these species to be such efficient vectors can be used to control the spread of malaria.

1.2.1. The *Anopheles gambiae* complex

The *Anopheles gambiae* complex serves as an excellent model for studying the genetics behind vectorial capacity of malaria in closely related species. A better understanding of the genetic background of these mosquitoes can provide essential information to developing a control strategy that can further relieve or eliminate the burden of malaria across the world. Evolutionary dynamics between members of this complex can potentially bring to light factors associated with vector competence that can lead to transgenic strategies or insecticides that can further reduce the burden of malaria.

The *An. gambiae* complex is comprised of seven cryptic species that are morphologically indistinguishable. These species exhibit different adaptations to their ecological background- residing in different geographical regions and exhibiting different behaviors- that lead to differences in vectorial capacity. Within the complex, *An. gambiae sensu stricto* (*An. gambiae s.s.*) and *An. arabiensis* represent the two most important vectors of malaria. Both display anthropophily and both are freshwater species; however, the geographical distribution of these two vectors can be very different, expanding the potential range for malaria transmission. Although both species exist in sympatry throughout Sub-Saharan Africa, *An. gambiae s.s.* is distributed through the rainy, subtropical regions of the contintent, while *An. arabiensis* is spread across the arid regions [42-44]. Of the lesser vectors, *An. merus* and *An. melas* are both salt-water species found on the coasts of Africa. *An. bwambae* is restricted to geothermal springs in the Semliki Forest of Uganda [45], while *An. quadriannulatus* A and B are found in southern Africa and Ethiopia, respectively [46]. *An. quadriannulatus* A and B are both zoophilic and thus are not considered vectors of malaria [47-49]. More recently, two incipient species of *An. gambiae* undergoing partial reproductive isolation have been identified [50]. These two species, termed the "M" and "S" forms, differ genetically in three regions dubbed "speciation islands" [51] that had elevated divergence between the two forms that was not present elsewhere across the genome. These two incipient species make an excellent model for evolutionary study of ecological diversification in species currently believed to be undergoing speciation.

The M and S forms may provide the most significant evolutionary information yet regarding *Anopheles gambiae* mosquitoes. This complex exhibits very different adaptations that are reflected through their genetic makeup. Bioinformatic and cytogenetic analyses are

continuing to elucidate the genetic factors responsible for these adaptations, which may eventually lead to genetic vector control strategies.

1.2.2. *Anopheles gambiae* chromosomes

For many species, cytogenetic analyses are restricted to small mitotic chromosomes that may hide important chromosomal rearrangements that are responsible for different adaptations. Study of mitotic chromosomes typically relies on specialized staining techniques to fully resolve different regions of chromatin like euchromatin and heterochromatin. These chromatin domains provide essential environments for cellular processes to occur. The *An. gambiae* complex does not just provide an excellent model for studying species with closely related lineages. These species also make an excellent organism for cytogenetic study due to the presence of polytenized chromosomes.

These giant, well-banded interphase chromosomes, form as a result of multiple rounds of replication without cell division. The lack of cell division causes the sister chromatids to remain together, resulting in single, large chromosomes. These chromosomes are found in many Dipteran species in various tissue types including the salivary glands, ovarian nurse cells, and Malphigian tubules [52]. These characteristic banding of these chromosomes results from different levels of chromatin compaction, and can be used to readily identify many species.

The *Anopheles gambiae* complex share a mitotic karyotype of 2N=6. Two autosomes divided into four total arms: 2R, 2L, 3R, and 3L complement the X and Y sex chromosomes. The polytene nature of chromosomes in the *An. gambiae* complex provides a unique ability to

study organism chromosomal structure at a higher resolution than is possible with standard mitotic chromosomes.

Cytogenetic maps detailing the chromosome complement of *Anopheles gambiae s.s.* were developed originally from larval salivary glands [53,54], and later in ovarian nurse cells [55,56]. A total of 46 numbered divisions and 151 lettered subdivisions were assigned to classify each of the chromosome arms [57].

The pairing of tools like *in situ* hybridization and restriction mapping with cytogenetic maps has provided researchers with an important method of linking genes and other genetic sequence features to the physical chromosomes of an organism. Microdissected DNA [58], microsatellite markers [59], randomly amplified polymorphic DNA markers [60,61], cosmids [61], and cDNA markers [61] have all been used to generate physical maps for *An. gambiae* chromosomes. More recently, the physical map of the *An. gambiae* genome has been updated to link 67 scaffolds to chromosomal regions [62], with an additional 16 scaffolds being mapped within the peri-centromeric heterochromatin [63].

Polytene chromosomes are a powerful tool that can be used to combine bioinformatic information with physical cytogenetics analyses. Polytene chromosomes have morphologically distinct chromatin types that can be further studied or isolated using techniques like microdissection to improve the overall genome assembly.

1.2.3. *Anopheles gambiae* genome assembly

DNA sequencing projects result in limited length scaffolds and contigs that must be assembled together to create a genomic assembly. These assemblies require properly oriented and placed contigs/scaffolds in order to truly represent an organism's genome. Misplaced, misoriented, and unplaced contigs all contribute to inaccuracies and gaps within the genomic assembly that can provide misinformation when the assembly is used for genomic analyses.

The *Anopheles gambiae* PEST strain, a mutant line containing a fixed, standard chromosomal arrangement and a sex-linked pink eye mutation that allowed for identification of colony contamination, was sequenced using plasmid libraries derived from 330 males or 430 females in 2002 [62]. The line was found to be polymorphic for molecular markers representative of both M and S forms, thus suggesting the PEST strain is a hybrid between the two forms [62]. The total size of the initial genome assembly was 278 megabases with 10.2 fold sequence coverage [62]. Approximately 2000 bacteria artificial chromosome (BAC) clones were physically mapped using *in situ* hybridization to orient and assemble the scaffolds. 187,844,042 bp over 30 scaffolds were originally mapped and oriented, with an additional 45,266,526 bp over 112 scaffolds being mapped, but not oriented [62]. The resulting assembled genome covered approximately 84% of the sequenced genome. Fifteen additional scaffolds were mapped within the peri-centromeric heterochromatin, increasing the total size of peri-centromeric heterochromatin from 3.3 Mb to 8.64 Mb and bringing the total mapped genome size to approximately 264 Mb [63]. As the costs for sequencing continue to dwindle, more assemblies have been produced for *An. gambiae*, including independent M and S form assemblies [64], as well as an effort to sequence and assemble many more Anopheline species [65]. These data will be extremely valuable in analyzing the evolution of vectorial competence amongst related species.

The *An. gambiae* genome, as with any other currently available genome, can still be improved. The "unknown chromosome," a collection of unassembled scaffolds represents 42 megabases of the genome that have not been properly assembled within the genome. These sequences are likely heterochromatic, and could provide a wealth of information regarding chromatin dynamics and heterochromatin-related genes.

Studies of model organisms such as *D. melanogaster* revealed that the functional organization of the genome could not be understood without studying higher levels of genetic regulation. For this reason, epigenomics has recently received much attention from researchers. Epigenomics studies epigenetic modifications in an entire genome at the DNA, RNA and chromatin levels. Different chromatin types have been identified in *Drosophila*, but little information is available for mosquitoes.

1.3. Epigenomics

Epigenomics is a relatively new field that explores how epigenetic modifications, heritable non-genetic changes that result in genomic dynamics, affect the genome. Chromosome dynamics resulting from epigenetic modifications create environments that regulate gene expression within different cell types. Epigenomics can shed light on how disease initiates epigenetic changes within the organism, establishing causal roles for the parasite. These studies can improve targeting strategies within transgenic experiments.

1.3.1. Chromatin types

The definitions of chromatin types vary from species to species. Many definitions offer similar characterizations, however no unanimous description of chromatin types has truly been made. These regions are extremely important, as they provide various environments that affect cellular dynamics.

Chromatin is the genetic material, both DNA and interacting proteins, found in nuclei that condenses into structures known as chromosomes. Chromatin is organized into multiple higher order structures that allow for varying conditions of chromosomal interaction, including transcription, recombination, and DNA repair [66]. Originally seen by Heitz in 1928 [67], chromatin showed differing levels of compaction. Less condensed, typically transcriptionally active chromatin with an abundance of unique coding genes was termed euchromatin. Heterochromatin was defined by highly condensed, gene poor regions that were typically transcriptionally repressed.

Heterochromatin is classically defined by its late-replicating nature due to the condensation of the chromatin in these regions. It is surmised that the replication machinery has a hard time accessing the underlying DNA and requires a longer period of time to complete. Previously, heterochromatin was labeled as junk DNA [68] that took up space in the genome. The highly repetitive, gene poor regions provided a buffer within the genome that was believed to serve no other purpose. This perception has changed, as studies have shown the biological significance of heterochromatin in regulating genes, providing chromosomal stability, affecting nuclear organization [69], or even providing an environment necessary for specific gene expression. Heterochromatin can be further classified into two types: constitutive

heterochromatin and facultative heterochromatin. Constitutive heterochromatin maintains its heterochromatic nature through all stages of development and thus is considered permanent. This form of heterochromatin is primarily found at the centromeres and telomeres of chromosomes, and is composed of highly repetitive DNA including satellite sequences [70] and vestigial transposable elements (TEs) [71]. Facultative heterochromatin, dynamic chromatin capable of changing chromosomal states based on gene activity and other signals from the cell, does not exhibit a high density of repetitive DNA. Facultative heterochromatin can be present in some cell types, while not present in others. Potential reasons for decondensation and increased transcription include: change in developmental state, nuclear relocalization, and heritable factors [72].

Chromatin is comprised of four core histones, H2A, H2B, H3, and H4 that are arranged in an octamer with the DNA wrapped around it. These histones are able to undergo different types of post-translational modifications that are proposed to organize genomes into specific domains [73]. These modifications, which can often be associated with different chromatin types or states, alter the chromatin structure by signaling for changes in chromatin condensation. These structural changes help facilitate gene transcription and silencing. Modifications including methylation of Histone 3 Lysine 4 (H3K4me) and histone acetylation has long been linked to transcription [74,75], and in many cases has also been associated with transcriptionally active regions of the chromosome [76,77]. Hypoacetylation, tri-methylation of Histone 4 at lysine 20, and tri-methylation of Histone 3 at lysine 9 (H3K9me3) are epigenetic marks associated with peri-centromeric heterochromatin, and thus believed to be essential for heterochromatin formation. These epigenetic marks are important in recruiting effector molecules to their specific chromatin locations [78]. The most notable of these effector molecules is HP1. Heterochromatin protein 1

(HP1) is a family of conserved proteins found in many eukaryotic species that is important in heterochromatin formation. HP1 stabilizes heterochromatin domains [79,80] and aids in chromatin remodeling [81]. The HP1 group is made up of chromosomally bound proteins that were identified as a suppressor of a phenomenon known as position effect variegation in *D. melanogaster* [82]. This phenomenon describes changes in expression pattern of genes located near or within heterochromatin [83] that can be altered as a result of HP1 or other suppressor of variegation protein knockdown [84,85]. HP1 is stably localized to the peri-centromeric and telomeric regions of chromosomes, but can also be found located at many discrete euchromatic loci [86,87]. Many of the binding sites found in *D. melanogaster* appear to be conserved across *Drosophila* species, and thus may suggest a role in gene regulation or suppression amongst its other roles [88].

Cytologically, heterochromatin is easily discernable from euchromatin due to the chromatin's differences in condensation. Peri-centromeric heterochromatin of *An. gambiae* is characterized by diffuse, non-banded chromatin that forms into a puff-like structure. Morphologically, the intercalary heterochromatin (IH) is different in *An. gambiae* compared to *D. melanogaster*. IH is characterized by extremely condensed blocks or diffuse, non-banded structures that resemble puffs in *An. gambiae*, a contrast to the dark bands exhibited in *D. melanogaster*. Intercalary heterochromatin in *D. melanogaster* is characterized by under-replication in polytene chromosomes of larval salivary glands and can be found dispersed throughout euchromatin [89]. Due to this under-replication, IH regions appear as "weak points" characterized by chromosomal breaks caused by incomplete polytenization [90]. These regions, much like peri-centromeric heterochromatin are defined by late-replication. This heterochromatin is also essential in forming attachments that anchor the chromosomes to the nuclear periphery [91-93]. The visible morphological structures of intercalary heterochromatin in

polytene chromosomes of *An. gambiae* provide a unique ability to cytologically study this chromatin type at a higher resolution than is possible in other species.

Heterochromatin represents a significant portion of the genome, yet it is one of the most neglected aspects of genomic studies. The importance of heterochromatin in cellular functions including gene regulation and chromosome stability illustrate the necessity to study this chromatin type in genomic studies. By isolating or incorporating specialized studies that elucidate how heterochromatin works, a better understanding of nuclear organization and gene expression can be achieved.

One of the prominent features of heterochromatin is the abundance of transposable elements. Therefore, studying transposable element regulation and function will help with understanding how the genome is organized and regulated.

1.3.2. Transposable elements (TEs) and their role in the genome

Transposable elements (TEs) are genetic elements that are capable of moving to new genetic loci through a cut and paste mechanism (DNA transposons) or through RNA intermediates (retrotransposons). The discovery of jumping genes in the 1940's by Barbara McClintock led to the identification of TEs. Shortly after finding jumping genes, the first documented transposable elements, *Ac* and *Ds*, were characterized in "The Origin and Behavior of Mutable Loci in Maize" [94].

These elements can represent a significant portion of eukaryotic genomes; TE composition has been shown to vary from 1% in *Apis mellifera* [95], to 45% in *Bombyx mori* [96,97],

up to over 85% in some species of grass, including maize [98]. In the case of important mosquito vectors, *Anopheles gambiae* contains approximately 16% TEs [62], while TEs comprise almost half (47%) of the *Aedes aegypti* genome [99]. TEs can even be responsible for the size difference of genomes of closely related species. Of 12 *Drosophila* genomes, TE content ranges from 2.7% to 25% [100].

TEs are traditionally classified into class I elements, known as retrotransposons, and class II elements, known as DNA transposons. Retrotransposons typically consist of a gag-pol genome, containing the necessary domains for protease, reverse transcriptase, RNaseH, and integrase; in some cases, an env-like open reading frame exists [101]. Retrotransposons transcribe DNA to RNA, creating an RNA intermediate that is then reverse transcribed back into DNA that can transpose elsewhere in the genome. These elements are typically more abundant than their DNA transposon counterparts. Class II elements contain a transposase gene that is flanked by terminal inverted repeats (TIRs). Rather than requiring a RNA intermediate like retrotransposons, the transposase excises the TE, allowing them to be inserted and ligated into a new position in the genome. A majority of these lacks the capability to transpose as they settle into their final resting place in the genome, a state known as epigenetically silent, yet there are still elements that are able to actively move. In the case of humans, the only type of TE that is still known to be active is non-LTR (long terminal repeat) retrotransposons, which comprise almost one-third of the entire genome [102]. In *Anopheles gambiae*, results suggest that there are still multiple active elements in at least three subfamilies of TEs: Pao-Bel, Copia, and Gypsy [103]. All of these are subfamilies of LTR-retrotransposons. Technological advancements in high-throughput sequencing technology have led to resurgence in TE studies. As transposable elements typically comprise a large portion of many genomes, the large increase in genome

sequencing projects, due to availability and significant decrease in costs, has provided an abundance of data regarding these mobile elements.

TEs can affect the genome in multiple ways. The most evident of these effects is the ability to alter the genome's overall size. TEs are often linked to the stark differences in the genome size of many species. TE proliferation and elimination play a vital role in altering the size of a genome. The presence or absence of certain TE families, as well as variation in copy number of specific TEs has led to differentiation amongst 12 different *Drosophila* species [104,105].

The importance and function of these mobile elements has always been debated, ever since their initial discovery. For decades, misnomers like "junk DNA" [106,107] and "selfish elements" [108] have been assigned to TEs. These labels result from the capablility for autonomous replication associated roles in genomic disruption [106,107]. Classical experiments have shown that TEs are capable of genomic destabilization and genomic rearrangements. Ectopic recombination between repeat copies of transposable elements can lead to genomic rearrangements within different species [109-111]. TEs have been linked to a significant quantity of differences between humans and chimpanzees, a result of deletions caused by ectopic recombination of *Alu* elements, which belong to the short interspersed element (SINE) family [112]. Transposable elements have been linked to rearrangement breakpoints in *D. melanogaster* [113] as well as in *A. gambiae s.l.* [114]. Due to their insertion near these breakpoints, TEs can cause inversions, leading to genomic recombination.

Another phenomenon, called hybrid dysgenesis, was originally described in *Drosophila melanogaster*. This phenomenon comprised a range of aberrant traits including chromosomal aberrations, mutation, and sterility [115]. Eventually, these traits were linked to the mobilization of

P-elements, later classified as DNA transposons, and I-elements, classified as non-LTR retrotransposons [115-117].

Some argue that the presence of TEs throughout the genome distributes promoters and enhancers that further enhance or repress gene expression levels of host genes [118,119]. TEs can render protein coding genes nonfunctional [120] as well as affect gene expression levels [121]. In fact, studies suggest many promoters originate from transposable elements [122,123].

However, not all TEs have negative effects on the host genome. A relationship has been established between some bacterial transposons that confer antibiotic resistance. In *Drosophila*, TEs also represent a significant portion of the chromosomes' telomeres, extending the length of the telomere to prevent chromosome shortening [124]. There is also evidence that well conserved TEs in the centromeres provide stability as well as the opportunity for evolution [125].

As mentioned before, TEs were widely considered parasitic and conferring no potential benefits to the host genome [106,107]. This hypothesis also suggested that the only reason for a TE's survival is its ability to successfully transpose elsewhere into the genome. Many studies have purported this outlook- increased TE density [126] and TE mobilization [127,128] have both led to decreased host fitness. The genome, then, is suggested to undergo "genome-level" selection [106,107], a non-phenotypic selection resulting from differential reproductive success of TEs. Differences in element density from same family TEs, *ie*. *Copia* and *gypsy* LTR-retrotransposons [129], suggest that TE transposition or removal rate, presumably though selective pressure, differs from one element to the next. TE removal rate may also be linked to the chromosomal location at which the TE is inserted. TEs typically insert themselves outside of gene-rich areas [130], reducing fitness costs associated with genic insertions and likelihood of disrupted recombination.

However, once a TE inserts itself into heterochromatin, it typically cannot escape. Two theories explain this phenomenon. The first, more classical, theory says that removal of the TEs is not efficient in low recombination regions [131]. Thus, the TEs remain in transcriptionally repressed areas. The second theory suggests that TEs target heterochromatic regions [132]. Positive selection for elements not recognized by the host could lead to amplification and transposition of the element until its eventual silencing [133,134]. Three models have been proposed that consider TE population density based on the element's effect toward the genome: the ectopic exchange [134], deleterious insertion [135], and host fitness cost [136]. However, all of these models assume that these elements have no phenotypic effect on the host.

Two databases provide the most comprehensive database of mosquito TEs. Repbase contains collections of TEs from most eukaryotic genomes that are currently sequenced and available [137-139]. This database is a tremendous resource for masking and annotating repetitive sequences in general, as consensus sequences for elements across several genomes are included. Another TE database called Tefam contains sequences for elements from three mosquito genomes: *Anopheles gambiae*, *Aedes aegypti*, and *Culex quinquefasciatus.*

TEs are an important, yet overlooked portion of the genome. They can be used transgenic drive mechanisms, and are implicated in genomic evolution. TEs can also create new genes that can profoundly affect the genome. However, the annotation of TEs in *An. gambiae* is still limited compared to model organisms like *D. melanogaster*.

TEs must be regulated in order to prevent the entire genome from being overrun by these elements. Recent studies discovered a major role of the Piwi-interacting RNA (piRNA) pathway in regulating the TE activity in eukaryotes.

1.3.3. Diverse functions of the Piwi-interacting RNA (piRNA) pathway

The Piwi-interacting RNA (piRNA) pathway has been identified in multiple Dipteran species as a mechanism for transposable element regulation. As the pathway has been recently identified, many of the core processes behind the pathway still remain unclear.

The piRNA pathway is one of three small non-coding RNA mechanisms that have been found in many metazoan organisms. These pathways provide post-transcriptional silencing through translational repression and mRNA degradation. The piRNA pathway is quite different from both the microRNA (miRNA) and small interfering RNA (siRNA) pathways. Where the siRNA and miRNA pathways both require the RNase III enzyme Dicer to process the double stranded RNA precursors into their final small RNA products, the piRNA pathway eschews the necessity for this enzyme, relying on slicer-mediated cleavage to produce functional piRNAs. As with the miRNA and siRNA pathways, the piRNA pathway associates with proteins from the Argonaute family. Both the miRNA and siRNA pathways interact with Argonaute proteins that are found ubiquitously in all tissue types. The piRNA pathway associates with proteins from the PIWI clade that are primarily found within the germline. These proteins, made up of PIWI, Aubergine (Aub), and Argonaute 3 (Ago3), were found to be part of the larger Argonaute family [140,141]. While Piwi [142] can be found within the nuclei of both somatic and germline cells, Aubergine [143] and Argonaute 3 [144] are both cytoplasmic proteins found in the nuage, or cytoplasmic cement, of germline cells only.

The piRNA pathway's primary function is posited to help protect the germline against the mobilization of transposable elements. In the germline, it is essential to protect the genome

against deleterious TE mobilization that can result in disrupted gametogenesis that could potentially lead to sterility [145,146].

Mutations in any of the three Piwi proteins results in transposable element de-repression in the germ line. Transposon levels are elevated as a result of Piwi protein reduction [147-149]. This finding indicates that all three proteins are essential to functional TE silencing, and that these proteins are working in a non-redundant manner [149-152]. In mice, the piRNA pathway confers a slightly different method of protection than is seen in *D. melanogaster*. All three of the piwi-related proteins in mice, MIWI2, MILI, and MIWI, appear to have significant roles in spermatogenesis [153-156]. MIWI2 and MILI are essential to control long interspersed nuclear elements (LINEs) and long terminal repeat (LTR) retrotransposons [155,156]. However, rather than protect the female germline as is found in the fruit fly, these proteins are essential to maintaining the male germline, as well as assisting gametogenesis. MIWI2 and MILI both are implicated in transposable element silencing; however, MIWI appears to only function in spermatogenesis. In MIWI2 and MILI mutants, it has been observed that CpG methylation, an important factor in TE silencing, is decreased.

piRNAs do not solely associate with transposable elements. These small RNAs can target protein-coding genes as well as other non-TE sequences. In *D. melanogaster*, genes including *Stellate* [157], *Fasciclin 3* [158,159], and *vasa* [160], all interact with part of the piRNA population. These findings suggest that the piRNA pathway may not only target transposable elements for silencing, but can also repress or silence both repetitive (*Stellate*) and non-repetitive (*Fascicilin 3* and *vasa*) genes. In *Aedes aegypti*, a shift occurs where piRNAs localize to transposable protein-coding genes and other non-TE sequences. Of the 30 top genes with

piRNA enrichment, two are homologous to *Stellate* and *Traffic jam*, both of which are found in *D. melanogaster* [161].

Studies have also shown that altering the piRNA pathway also results in germline stem cell loss or sterility, which suggest a strong role in germline development and function [142,153,155,162,163]. Piwi proteins specifically interact with piRNAs during gametogenesis [164-166]. Aubergine, Argonaute 3, and Piwi also all have established roles in gonad development. Piwi is essential to fertility in *Caenorhabditis elegans* [167,168], *Danio rerio* [169], *and Mus musculus* [153].

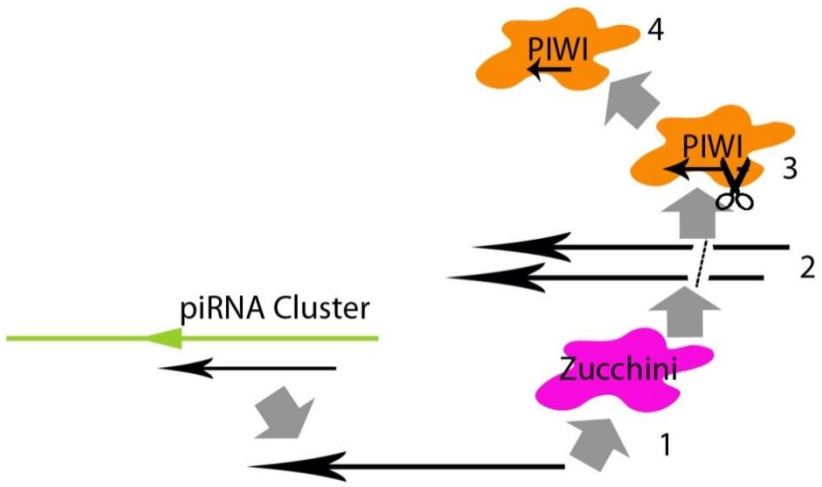
Mounting evidence has begun to suggest that the piRNA pathway may also have a role in anti-viral immunity in some species of *Aedes* mosquitoes. After injection of *Aedes albopictus and Aedes aegypti* with Sindbis virus, elevated quantities of small RNAs ranging from 23-30 nucleotides were resolved [170-174]. Viral piRNAs (vpiRNAs) were found to be produced both through primary biogenesis as well as through the ping-pong amplification cycle [174]. Cells with normal siRNA responses and cells deficient with an integral siRNA protein, Dicer-2, both exhibited a similar piRNA response, suggesting that the piRNA pathway's role in anti-viral immunity is more than a redundant backup immune response [174].

piRNAs are 21-30 nucleotide single-stranded RNAs that can be linked to one of the three PIWI-clade Argonaute proteins. These piRNAs exhibit very different characteristics as compared to miRNAs. miRNAs are well conserved, even amongst species, and are not too genetically diverse. piRNAs lack sequence homology and are not conserved, leading to an extremely complex population of small RNAs that are all unique. piRNAs are expressed in germline of *Drosophila melanogaster* [144] and *Danio rerio* [169], as well as in the testes of *Mus musculus* [175,176] and *D. rerio* [169].

These small RNAs are primarily produced from specific loci termed piRNA clusters. piRNA clusters are primarily located within the peri-centromeric and sub-telomeric heterochromatin of the *D. melanogaster* genome. These clusters are made up of vestigial TEs, mostly partial or defective elements that have localized in loci ranging from 2-250 kilobases in size [144]. piRNA clusters generate long single stranded transcripts that become the precursors for the final piRNA pool [144,177]. Germline clusters exhibit mostly bi-directional transcription- the generated piRNAs come from both genomic strands. In somatic cells, however, there is a noted increase in uni-directional linear transcription of piRNAs [144,178].

In *Drosophila melanogaster*, piRNAs are generated through two independent mechanisms, yet many aspects of these mechanisms have not been entirely elucidated. The piRNA pathway is the most recently discovered of the three small RNA silencing mechanisms, and as such is still the least understood. Much of the current understanding in regards to biogenesis and function is derived from experiments in *D. melanogaster*.

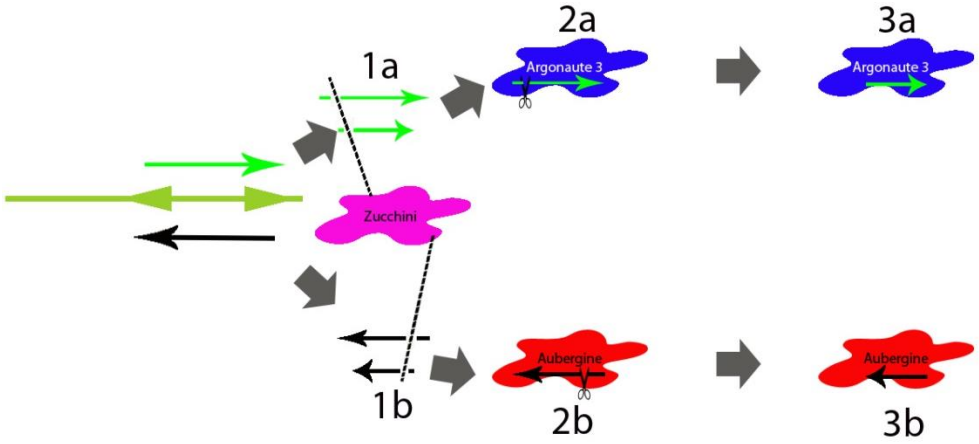
To summarize the mechanism, piRNA clusters produce long single stranded RNA transcripts, ranging in size from kilobases to hundreds of kilobases that are processed into primary piRNAs through primary biogenesis. The biogenesis of these long transcripts, termed precursor piRNAs, has yet to be elucidated, but a study in mice suggests RNA Polymerase II is needed to transcribe these clusters [179]. Promoter mutations in piRNA clusters of *D. melanogaster* have been linked to the lengthy transcripts [144]. It has also been shown that a homologue of HP1, Rhino, is required at the piRNA cluster site to initiate piRNA production [180]. Rhino forms a complex with the protein Cutoff that is also essential to piRNA production [181].



piRNA Biogenesis in somatic cells:

1. Anti-sense transcripts from piRNA clusters are cleaved by Zucchini to create medium length transcripts- intermediate piRNAs (2).
3. The intermediate piRNAs are loaded onto PIWI proteins and the piRNA fragment is trimmed to created a mature PIWI piRNA complex (4).

Figure 1.1. piRNA biogenesis through primary amplification. This pathway is primarily found in somatic cells.



piRNA Biogenesis in germline cells:

1a/b. Transcripts from multi-oriented loci produce transcripts that are cleaved by Zucchini into piRNA fragments in both sense (a) and anti-sense (b) orientation.
2a. Sense fragments are loaded onto Argonaute 3.
2b. Anti-sense fragments are loaded onto Aubergine.
3a/b. Respective complexes trim down piRNA, forming mature piRNA complex for both strand orientations.

Figure 1.2. Ping-pong amplification of piRNAs. This pathway is found in germ-line cells.

Depicted in **Figure 1.1 and Figure 1.2** are the generalized mechanisms through which piRNAs are generated in *D. melanogaster*.

The mechanism for which piRNAs are generated within somatic cells is usually linked to the initial production of piRNAs. In *D. melanogaster*, many of the somatic piRNAs originate from uni-directional clusters that provide anti-sense transcripts that can bind to the Piwi protein. As mentioned previously, of the three proteins found within the Piwi-clade, Piwi is the only one found in somatic cells. Primary biogenesis is thought to occur through cleavage of these long precursor piRNAs that can be then loaded onto the Piwi protein and trimmed to the typical 24-30 nucleotide small RNA. This processing requires Zucchini, a cytoplasmic endonuclease that reduces the piRNA size prior to Piwi loading [182,183].

Secondary piRNAs are produced through an amplification cycle referred to as the ping-pong cycle that provides a large portion of the representative piRNAs in the germline [144]. Rather than interacting solely with Piwi, the ping-pong cycle utilizes Aubergine and Argonaute 3, binding anti-sense and sense piRNAs, respectively. Aubergine-piRNA complexes can target transposon RNAs, cleaving them small enough to load onto Argonaute 3. Argonaute 3-piRNA complexes can, in turn, target transcripts produced from piRNA clusters.

The resulting cycle can produce large quantities of piRNAs that exhibit two motifs that can help identify at least a portion of the total piRNA pool. First, the piRNAs associated with Aubergine tend to have a Uridine at the 5' end of the sequence, while the piRNAs associated with Argonaute 3 typically contain an Adenine at position 10 of the 5' end. Second, there is a 10 basepair overlap that occurs between complementary opposite sense piRNAs [144,184].

The piRNA is an essential pathway in regulating transposable elements of many metazoan species. The pathway has been implicated in secondary roles including gene regulation and anti-viral immunity, though these roles need be better understood. However, comparative analyses may shed light on how this pathway is evolving within species.

1.4. Methods for the analysis of individual chromosomes

As sequencing projects get more efficient and cheaper, more researchers are conducting studies on individuals within a population to identify polymorphisms and mutations leading to phenotypic change. Microdissection offers the opportunity to not only isolate cells from a single individual, but to isolate individual chromosomes that can be used in sequencing and fluorescent *in situ* hybridization (FISH) experiments. Both of these techniques can be used to analyze chromosomal rearrangements. Microdissection can produce chromosomal paints that can be used to study nuclear organization of cells. Together, microdissection and whole genome amplification can provide a powerful tool in cytogenetic analyses.

1.4.1. Laser capture microdissection (LCM)

Laser capture microdissection (LCM) provides a versatile method for cytogenetic isolation of DNA. For most FISH experiments, researchers must rely on PCR generated or bacterially raised DNA fragments in order to label DNA of interest. The isolation of DNA using LCM provides a unique opportunity to produce DNA that not only can be used in FISH

experiments without the need for previously known DNA sequences, but also can provide DNA from an individual for use in sequencing.

Microdissection was originally developed to isolate DNA from bands of *Drosophila melanogaster* polytene chromosomes [185]. The technique involved the use of glass microdissection needles set in a micromanipulator that would scratch a glass slide containing these chromosomes. The operator could effectively cut out a segment of chromosome that could be then used in further genetic studies. This process has been applied to metaphase chromosomes as well, where chromosomes from virtually all species can be isolated through microdissection [186,187]. This technique required an experienced operator skilled in identifying and dissecting target chromosomal regions. As technology advanced, innovations to the microdissection platform have allowed for simpler procedures that were under computer control. The first initial change to the microdissection platform saw the use of a laser microbeam that replaced the need for micromanipulation. A microcomputer was used to control an argon-ion laser beam that could cut and isolate target chromosomes [188]. At this time, the equipment needed to perform successful microdissection was cumbersome and still required substantial maintenance. Further advances in microdissection equipment have resulted in two technologies that are still currently being used today: laser capture microdissection via infrared capture and ultraviolet systems.

As the technology in this field increases, so too does the practicality of using microdissection as a technique to isolate chromosomes and cells of interest. In the procedure's nascent stage, which required the use of glass dissection needles and a manually operated microcontroller, the potential risk of contamination was much higher than contamination stemming from laser capture microdissection. The removal of moving parts that came into

contact with the sample, as well as a user friendly platform, resulted in an efficient process that has made microdissection a practical technique for isolating cells and chromosomes. The range of uses for this platform has expanded as a result of technological advances. As mentioned above, the technique was originally used primarily to isolate chromosomes from an entire chromosomal complement to be used as DNA markers. Current microdissection capabilities allow users to extract single cells [189-191], which result in the capture of DNA, RNA, and/or proteins of interest [192]. Advancements in the technology allow for the differentiation and isolation of cancer cells nested in clusters of normal cells [193-195], as well as the segregation of individual chromosomes for use in fluorescent chromosome painting [196-203].

Microdissection can be an invaluable tool when using the technique to isolate DNA for downstream applications. The technique when used in cytogenetic analysis does not require the knowledge of genetic sequences of interest, nor does it require the use of vectors, ie. bacteria/ yeast artificial chromosomes (BACs and YACs), to amplify the DNA. Regions of interest can be identified based on morphological characteristics, as well as genomic assembly, which can then be used in conjunction with chromosomal paints, as mentioned above, smaller probes that help in constructing linkage [204] and physical maps [205,206], and development of expressed sequence tags (ESTs) [207-209].

DNA extracted in LCM has many potential down-stream applications. Particularly, it can provide unique information regarding a single individual through sequencing, or serve as the probe for DNA labeling experiments. However, in order to fully take advantage of this technology, an efficient method of DNA amplification after LCM is needed to amplify the DNA to levels suitable for downstream applications.

1.4.2. Whole genome amplification (WGA)

In most genomic assays, one of the limiting reagents for the technique is the initial quantity of DNA available. Technological advancements in some genomics applications like high-throughput sequencing have resulted in a notable decrease in initial DNA requirements. As of 2010, some of the most common high-throughput sequencing techniques required at least 500 ng (454 sequencing) up to 1 mg of DNA (capillary sequencing) in order to guarantee a successful sequencing run (http://www.sanger.ac.ukresourcestechnologiesdna_requirements_for_sequencing_june2010.pdf). Fluorescent *in situ* hybridization (FISH) experiments generally require between 500 ng and 1µg of total DNA in order to create sufficient probes for an experiment. Similarly, the method for procuring the DNA can be extremely difficult which results in precious samples that have a limited number of potential applications. Due to the fact that DNA is an exhaustible resource that can be difficult to initially produce, molecular diagnostic laboratories must overcome these limitations through advancing amplification techniques. Multiple techniques currently exist that allow users to greatly increase DNA quantity from a small initial stock. These reactions are anticipated to produce micrograms (µg) of DNA from initial quantities of picograms-nanograms. Currently, there are multiple variants of two primary methods of whole genome amplification: polymerase chain reaction (PCR) based strategies and multiple displacement amplification (MDA).

1.4.3. PCR-based amplification

Three variants to standard polymerase chain reaction (PCR) experiments have been used to successfully carryout a WGA on small quantities of DNA: degenerate oligonucleotide PCR (DOP-PCR), primer extension PCR (PEP), and adaptor-ligation PCR. All three variants are similar to a standard PCR reaction in that they require the use of Taq polymerase and primer sets to amplify DNA. Developed by Telenius *et al.*, DOP-PCR, which has become less commonly used for WGA experiments, incorporates semi-degenerate primer sets in conjunction with an increasing annealing temperature [210]. These primers are capable of binding to many sites at a low initial temperature, but increase in specificity as the temperature is increased. PEP PCR was originally designed by Zhang *et al.* in order to amplify DNA from single sperm cells that could be used in haplotyping assays [211]. The technique requires a pre-amplification step that creates binding sites on the DNA fragments for 15-base oligonucleotide primers to anneal to at low temperature. Adaptor-ligation PCR, originally established in 1989 by Ludecke *et al.* [212] uses restriction enzymes to digest DNA, leaving the necessary sites for adaptors to ligate to the ends of the DNA fragments. In all three techniques, there is an inherent limitation in fragment size – DNA can be amplified to lengths of 3 kb, but most fragment sizes average approximately 400-500 base pairs in length- and error rate – the use of standard Taq polymerase leaves the potential for the introduction of errors into the amplified sequence. This error rate is especially significant in repeat rich regions like the heterochromatin [213,214]. Of the three techniques mentioned above, PEP suffers the greatest from amplification bias – primer binding is not uniform during annealing and extension, resulting in some loci being discriminated against. All of these PCR variants have been successfully applied to whole genome amplification of single cells. PEP PCR has been used to amplify DNA to analyze disease genetic loci from single cells [215-217], while

DOP-PCR and some variants have been used to amplify DNA for chromosomal paints [218] from single chromosomes or provide sufficient DNA for pre-implantation genetic diagnoses [219]. DOP-PCR has been modified to accommodate issues with amplicon fragment size [220] to create longer amplified fragments by Pwo DNA polymerase and modified thermocycling conditions. PEP had similar modifications resulting in longer amplicons as well [221]. Commercial kits have been produced that incorporate PCR amplification of cells from small initial DNA quantities. The GenomePlex Single Cell Whole Genome Amplification Kit, developed by Sigma-Aldrich, has been successfully used to amplify DNA from single cells of humans [222,223] and mice [224]. More recently, the SeqPlex DNA amplification kit has been developed to further improve the PCR based whole genome amplification platform. SeqPlex offers the ability to remove the adapters required to carry out PCR.

These PCR-based amplification methods are excellent when working with unique, non-repetitive DNA. In the case of the GenomePlex and SeqPlex kits, it is even possible to directly label amplified DNA with a modified nucleotide mix. However, most eukaryotic genomes have an abundance of repetitive sequences dispersed across the genome. Especially in heterochromatic regions, these amplification strategies are inferior to other amplification methods.

1.4.4. Multiple displacement amplification (MDA)

Multiple displacement amplification (MDA) takes advantage of Phi 29 polymerase, a DNA polymerase that has 3' -> 5' exonuclease activity, in conjunction with random hexamer primers that results in increased fidelity when compared to standard Taq DNA polymerases [225].

This polymerase, derived from the *Bacillus subtilis* bacteriophage ϕ29, leads to many potential advantages over standard PCR based amplification methods. MDA has been reported to yield large quantities of DNA from very small quantities of template DNA [226]. Amplified fragment sizes upon amplification can reach over 10 kb in length because of the high rate of processing and lack of strand dissociation of the phi29 polymerase (Sigma). The error rate of MDA amplification is also lower than PCR based amplification due to phi29s native exonuclease activity- standard Taq polymerase generally creates three errors per 10,000 nucleotides added (350) compared to the one nucleotide per 1,000,000-10,000,000 error rate of the Phi29 polymerase. Phi29 is also capable of handling different secondary structures that would cause native Taq polymerase to slip or dissociate from the template. This slippage can result in low coverage of amplified loci, shorter DNA fragments as mentioned above, and artifact amplification. The ability to process longer DNA fragments also results in less amplification bias compared to PCR-based whole genome amplification techniques [226,227]. The polymerase allows the genome to be covered, resulting in uniform amplification across the entire DNA complement. There are currently three primary WGA commercial kits that use phi29 based amplification: REPLI-g (Qiagen), GenomiPhi (GE Healthcare), and TempliPhi (GE Healthcare). The MDA protocol still has limitations that hinder its capabilities to effectively amplify large template DNA fragments. The MDA technique typically produces abundant product called template independent product (TIPs). There is also notable dropout of certain loci, especially in heterochromatic regions that are rife with repetitive elements [227]. This dropout, known as allele drop-out (ADO), is defined as the random non-amplification of an allele found in a heterozygous sample. The average ADO rate can range from 0-60%, as described in [228,229]. Originally, MDA procedures used a large number of cells to produce enough starting material to carry out the

reaction, but recent experiments have shown that the technique can be applied to single cells with relative success [230-233].

The primary benefit of MDA amplification over PCR-based strategies is the ability to amplify heterochromatic regions at a higher rate. The generation of longer fragments of DNA that have low error rate make amplifying repetitive sequences much more efficient. However, a higher quantity of background is typically associated with Phi29, which can potentially interfere with sequencing efforts. Both PCR-based and Phi29-based amplification strategies have benefits and limitations that make them better suited for different applications. Thus, it is important to determine which aspects of amplification are most needed, and select a method of amplification based off this information.

1.4.5. Chromosome painting

Chromosome painting refers to the labeling of an entire chromosome arm in an effort to distinguish arms for genome analyses. This technique can be used to identify chromosomal rearrangements in a single individual, as well as in comparative genomic analyses.

This technique for painting was originally described by multiple research labs in 1988 [234-236] as a method for labeling human chromosomes for cytogenetic analyses. These studies used cloned DNA libraries to look for chromosome abnormalities and aberrations that could potentially lead to disease. These techniques revolutionized cytogenetic assays looking for chromosomal rearrangements, as previous studies required studying chromosome banding patterns. Probes for these painting experiments started with libraries generated from phages [237,238] and plasmids [237,239], and evolved into DNA isolated using flow sorting with DOP-PCR

[240,241] and microdissection [196-203]. These techniques have been used for cytogenetic analyses to the point where commercial libraries have been developed for human assays [242]. Probes produced through any of these amplification methods are fluorescently labeled and used in fluorescent *in situ* hybridization (FISH) experiments. Labeled chromosomes would stand out in comparison to non-target chromosomes, so rearrangements like inversions and translocations would be obvious to the researcher. More recently, chromosome paints were used to study nuclear organization of cells. Chromosomes maintain specific territories within the cell, creating nuclear dynamics that affect gene regulation and expression [243-245]. Each individual chromosome is labeled with an individual color or mix of colors to produce a territory map within the cell. A technique called spectral karyotyping (SKY) or multicolor FISH (m-FISH) mixes various ratios of fluorescently labeled DNA to create a spectral map initially for the entire human genome [246,247]. This same technique was later applied to mice [248,249]. These techniques considerably enhanced the ability to find chromosomal defects leading to disease phenotypes in humans. Although disease phenotypes are not widely studied in mosquitoes, multiple potential applications make chromosome painting a valuable tool within mosquito cytogenetic studies. As mentioned previously, cytogenetic analyses of nuclear organization can complement more bioinformatics approaches looking at chromosome-chromosome and chromosome-nuclear envelope interactions through bioinformatics approaches. Secondly, chromosome paints can be used to identify chromosome inversions that lead to selective advantages within the mosquito. The 2La inversion, one of ten fixed inversions in the *Anopheles gambiae* complex, has been shown to provide resistance to aridity and dessication [250,251]. These inversions are identified through unique banding patterns that are a hallmark of polytene chromosomes. Chromosome paints can be generated for species without cytogenetic maps delineating chromosomal

arrangements, as well as mitotic chromosome experiments that do not benefit from higher-resolution banding patterns.

1.4.6. Sequencing of micro dissected chromosomes

Microdissected material can be used for generating probes in fluorescent *in situ* hybridization (FISH) experiments, but the isolated material also is very useful in sequencing projects. Current whole genome amplification (WGA) techniques, as mentioned previously, have become very efficient in producing high quality amplicons. When combined with microdissection, an abundance of DNA can be produced from a single chromosome. This DNA can then be used in high-throughput sequencing to isolate a target region from the rest of the genome. Recently, multiple studies have used microdissected material in order to sequence potentially interesting regions. Weise *et al.* [252] took advantage of microdissection to sequence tumor rearrangements, whose sequences are not normally available. Highly repetitive regions including sub-telomeric, intercalary, and peri-centromeric heterochromatin, which are generally hard to map in genome assemblies, can be specifically microdissected to capture potential unique sequences and, in the case of humans, look at disease and aging factors related to these heterochromatic regions [253,254]. Smaller, undetectable rearrangements may also be identified as a result of these microdissection experiments. Also, genome assemblies can be improved through microdissection sequencing [255]. Previously unmapped scaffolds, which typically represent heterochromatic regions, can be identified by isolating and sequencing individual regions and aligning to those unmapped scaffolds. As the importance of heterochromatin within a genome becomes more understood, it is important for genome assemblies to provide access to these DNA

sequences that are currently left unassembled. The *Anopheles gambiae* genome currently has a 42.4 megabase "unknown" chromosome [256], that is comprised of unmapped scaffolds that are speculated to be mostly heterochromatic in nature [257]. Higher quality assembly of this genome can provide a wealth of information that can benefit genomic studies that may be crucial in the fight against malaria.

Chapter 2. Identification and classification of chromatin types in *Anopheles gambiae s.s.*

2.1. Introduction

The *Anopheles gambiae* complex represents a unique opportunity to study the evolutionary dynamics of closely related species, particularly in the case of the M and S molecular forms of *An. gambiae s.s.* In order to make the most accurate conclusions about these genomes, an accurate representation of each genome must be generated. The advancement of genomics-related technology has greatly increased the wealth of information available, especially in the number of genome sequences available from high-throughput sequencing. However, this data can be misleading in the sense that although it is available, without proper quality control and assembly, it may not provide the best representation of that genome as possible. Heterochromatin, an essential chromatin state plays a role in regulating genes, providing chromosomal stability, and affecting nuclear organization [69]. However, due to its highly repetitive nature, much of the heterochromatin in many genomes [257,258] remains unassembled as a separate "unknown chromosome." It is important to include these data in whole-genome analyses, as they can provide hidden links to species evolution, mutations, and genomic rearrangement.

The following subchapters represent attempts to improve the understanding of the *Anopheles gambiae* genome through cytogenetic and bioinformatics approaches. In sub-chapter 2, we detail a protocol for efficient, automated method for physically mapping and detecting fluorescently labeled DNA probes that will aid in producing high-resolution physical maps. Sub-chapter 3 presents a method to increase detail in polytene chromosome maps that can lead to more accurate mapping of DNA probes. Here, we also make a link from genetic coordinates to

their physical location on polytene chromosomes. In sub-chapter 4, heterochromatin-euchromatin boundaries and molecular features across the genomic landscape of *An. gambiae* are identified. For this paper, I take responsibility for the fluorescent *in situ* hybridization (FISH) results that identify the heterochromatin-euchromatin boundaries. Finally, in sub-chapter 5, a model is used to predict a chromatin type based on molecular features. This model successfully predicted a majority of previously classified heterochromatic regions, and identified at least seven novel regions.

By advancing current techniques and protocols, we can make a reasonable attempt to improve currently available information, rather than introduce an abundance of new information that must be parsed. The methods outlined below represent attempts to improve the *Anopheles gambiae* genome using currently available data, in order to better understand the genetic basis behind the malaria mosquito.

2.2. High-throughput Physical Mapping of Chromosomes using Automated *in situ* Hybridization

Phillip George, Maria V. Sharakhova, Igor V. Sharakhov

as published in the Journal of Visualized Experiments (JoVE), June 2013

Video available at http://www.jove.com/video/4007/high-throughput-physical-mapping-chromosomes-using-automated-situ

2.2.1 Abstract

Projects to obtain whole-genome sequences for 10,000 vertebrate species [259] and for 5,000 insect and related arthropod species [260] are expected to take place over the next five years. For example, the sequencing of the genomes for 15 malaria mosquito species is currently being done using an Illumina platform [261]. This *Anopheles* species cluster includes both vectors and

non-vectors of malaria. When the genome assemblies become available, researchers will have the unique opportunity to perform comparative analysis for inferring evolutionary changes relevant to vector ability. However, it has proven difficult to use next-generation sequencing reads to generate high-quality *de novo* genome assemblies [262]. Moreover, the existing genome assemblies for *Anopheles gambiae*, although obtained using the Sanger method, are gapped or fragmented [63,261].

Success of comparative genomic analyses will be limited if researchers deal with numerous sequencing contigs, rather than with chromosome-based genome assemblies. Fragmented, unmapped sequences create problems for genomic analyses because: (i) unidentified gaps cause incorrect or incomplete annotation of genomic sequences; (ii) unmapped sequences lead to confusion between paralogous genes and genes from different haplotypes; and (iii) the lack of chromosome assignment and orientation of the sequencing contigs does not allow for reconstructing rearrangement phylogeny and studying chromosome evolution. Developing high-resolution physical maps for species with newly sequenced genomes is a timely and cost-effective investment that will facilitate genome annotation, evolutionary analysis, and re-sequencing of individual genomes from natural populations [263,264].

Here, we present innovative approaches to chromosome preparation and fluorescent *in situ* hybridization (FISH) that facilitate rapid development of physical maps. Using *An. gambiae* as an example, we demonstrate that the development of physical chromosome maps can potentially improve genome assemblies and, thus, the quality of genomic analyses. First, we use a high-pressure method to prepare polytene chromosome spreads. This method allows the user to visualize more details on chromosomes than the regular squashing technique [265,266]. Second, a fully automated, front-end system for FISH is used for high-throughput physical genome

mapping. The automated slide staining system runs multiple assays simultaneously and dramatically reduces hands-on time [267]. Third, an automatic fluorescent imaging system, which includes a motorized slide stage, automatically scans and photographs labeled chromosomes after FISH [268]. This system is especially useful for identifying and visualizing multiple chromosomal plates on the same slide. In addition, the scanning process captures a more uniform FISH result. Overall, the automated high-throughput physical mapping protocol is more efficient than a standard manual protocol.

2.2.2. Protocol

1. High-pressure Polytene Chromosome Preparation from Ovarian Nurse Cells

Dissect half-gravid *Anopheles* females at 25 hours post-blood feeding under a dissection microscope and fix ovaries in fresh modified Carnoy's solution with methanol (100% methanol: glacial acetic acid, 3:1). At this stage the ovaries are at Christophers' III stage of development when follicles have an oval shape and a transparent area with nurse cells within follicles has a round shape13. Put ovaries from approximately five females into 500 µl of modified Carnoy's solution in a 1.5 ml Eppendorf tube and keep them at room temperature for 24 hours. Transfer ovaries to -20 °C for a long-term storage.

Prepare modified Carnoy's solution with ethanol (100% ethanol: glacial acetic acid, 3:1) and 50% propionic acid just prior to making slides. Place one pair of ovaries on a dust-free slide in one drop of modified Carnoy's solution. Split ovaries into approximately 6 sections with dissecting needles and place them into drops of 50% propionic acid on clean slides under a dissection MZ6 Leica stereomicroscope. Use a separate slide for each section.

Separate follicles dissecting using needles and wipe out remaining tissue with filter paper or paper towel under a dissection microscope. Add a new drop of 50% propionic acid to the follicles and allow them to sit for 3-5 minutes at room temperature. Place a coverslip on top of the 50% propionic acid droplet. Let the slide stand for approximately 1 minute.

Wrap the slide with filter paper and plastic. Using a Dremel 200 rotary tool with a Flex-Shaft attachment and soft plastic tip set between 3000-5000 RPM, express the follicles by swirling the tip in circles and lightly pressing the coverslip to evenly spread the nuclei. This step should take approximately 1 minute. Check the spread quality with an Olympus CX41 Phase Microscope using a 20x objective.

Prepare a sandwich by placing an additional coverslip next to the coverslip covering the chromosomes, and cover them with a second microscope slide. This reduces the chance of crushing the slide in the vise. Wrap the slides with a plastic sheet that come in glass slide containers and filter paper to hold the sandwich and to protect the slide from scratching due to the vise.

Apply pressure to the slides via mechanical vise. A pressure of 85-120 inch-lbs is sufficient and is achieved by using a torque wrench. This step is necessary for flattening the chromosomes as much as possible.

Remove the second microscope slide and the additional coverslip. Heat the slide with the coverslip covering the chromosomes to 55 °C on a slide denaturation/hybridization system for 10-15 minutes to further flatten the chromosomes. Dip the slide in liquid nitrogen for at least 15 seconds, and, when bubbling has stopped, proceed to quickly remove the coverslip with a razor blade. Immediately place slides in cold 50% ethanol for 5 minutes. Dehydrate slides in 70%, 90%, 100% ethanol for 5 minutes each. Air dry slides.

2. FISH Using an Automated Slide Staining System

Program is set up with the Xmatrx automated slide staining system to run the following steps, except the probe labeling. The detailed nick-translation labeling protocol is described in the documentation provided with DNA polymerase from Fermentas.

Before FISH, prepare fluorescent probes by labeling genomic BAC DNA with a fluorochrome using a nick-translation protocol. Mix 1 μg of DNA, 0.05 mM each of unlabeled dATP, dCTP, and dGTP and 0.015 mM dTTP, 1 μl Cy3- or Cy5-dUTP, 0.05 mg/ml BSA, 5 μl of 10x nick-translation buffer, 20 units of DNA polymerase I, 0.0012 units of DNase I, and nuclease-free water to 25 μl. The DNA polymerase I/DNase I ratio is selected empirically to obtain probes with a size range from 300 to 500 bp. Incubate the mix at 15 °C for 1 hour.

Put slides and reagents into the Xmatrx automated slide staining system and start the program to run the following steps. Apply 800 μl of 1x PBS for 20 min. Blow slides with air. Perform formalin fixation by applying 450 μl of 4% formalin in 1x PBS for 1 min followed by washes with 100% ethanol for 1 sec twice and for 2 min once. Blow slides with air.

Heat slides at 45 °C for 2 min to avoid bubbling when applying the probes. Apply 20 μl of the DNA probes, add drops of mineral oil to avoid evaporation of a hybridization solution, and place a coverslip on top. Denature chromosomes and DNA probes by heating slides at 90 °C for 10 min.

For hybridization, incubate slides at 42 °C for 14 hours with coverslips on. The hybridization solution consists of 2x SSC, 100 mM sodium phosphate, 1x Denhardt's solution, 100 μg/ml of sodium azide, and 10% dextran sulfate in formamide.

For stringency washes, heat slides at 42 °C for 2 min, remove coverslips, wash slides in 2x SSC for 1 sec 4 times. Blow slides with air. Apply 800 µl of 0.4x SSC at 42 °C for 10 min 2 times. Wash slides in 2x SSC at 25 °C for 10 min.

Perform chromosome staining by applying 50 µl of 1 µM YOYO-1 in 1x PBS. Apply drops of mineral oil to avoid evaporation of the staining solution, put on coverslips, and incubate at 25 °C for 10 min. Remove coverslips, wash slides in 2x SSC for 1 sec 4 times. Blow slides with air. Apply 15 µl of ProLong Gold antifade reagent. Put on coverslips.

3. Slide Reading with an Automatic Fluorescent Imaging System

This section details the use of the Duet software, which comes standard with the ACCORD PLUS automated scanning system. Instructions begin after turning on in this order: Olympus U-RFL-T power supply for a halogen bulb, computer Dell precision T3500, microscope Olympus BX61 with a connected camera Olympus U-CMAD3.

For setting up 10x Pre-Scan, open the Duet software in the ACCORD PLUS automated scanning system. Click "Online" button. Enter new Case ID and assign a Slide ID. Click the dot labeled "BF." This is the brightfield option. Set a scan choice to "10x Prescan." Use "2500x circle," "10000x circle" or "rectangle." Click "Set&run" for 10x Pre-Scan.

Click the "OK" button to run 10x Pre-Scan. Follow the prompts to adjust the scanning properly. Click "Finish" to start the scan. Press the "Main" tab to go back to the main screen. Click "Offline" button. Find the Case ID and Slide ID that was assigned and click "Offline Scan." In the black box at the top left area click on an arrow and select "10x Prescan." Using the arrows (< || Δ > a.k.a. "back," "pause," "play," "forward" buttons), go through the scanned images.

After finding an image of interest, double click on the screen the middle of the target region and press "Snap." This will target an image for capture later on. After selecting all targets, click "Classify." Select "10x Prescan." Right click an image and get the chance to classify the images. Select "Polytene."

Set up 40x Pre-Scan in the Duet software. Click "Main" to go back to the main menu. Select "Online" again. Select a slide. Change "BF" to "FL." Change the task name to "Revisit-X40-RG." Change the section right below the last setting to "Revisit-ALL." Click "Set&Run." Press "OK." Follow the prompts again to set up the automation. Click the "Start matching views" button to match 10x and 40x images. Click "Finish" to start the scan. Once done, click "Classify" and look at the images.

2.2.3 Representative Results

Figure 2.1 graphically depicts the scheme of the high-pressure chromosome preparation. This step involves the process of squashing and flattening chromosomes using a Dremel tool and mechanical vise, as well as chromosome visualization using a phase-contrast microscope. **Figure 2.2** illustrates the hybridization of a fluorescently labeled probe to target DNA on the chromosome slide preparations using an automated slide staining system, the use of an automated scanning microscope for visualizing and mapping the probes after the FISH experiment, and placing and orienting genomic scaffolds on the chromosomes.

Preparations of ovarian nurse cell polytene chromosomes from females of *An. gambiae* were made using the high-pressure technique. This method does not damage or change most of the chromosome structure **(Figure 2.3)**. It flattens bended chromosomes and, thus, reveals hidden fine bands that are not seen on regular preparations **(Figure 2.4)**.

The probes used in this protocol are genomic BAC DNA clones obtained from a ND-TAM BAC library generated from *An. gambiae* PEST strain DNA. The genomic DNA for this library was extracted from newly hatched first instar larvae of both sexes. **Figure 2.5** shows the results of FISH using BAC clones hybridized to polytene chromosomes of *An. gambiae*. This procedure was performed using the Xmatrx automated slide staining system. Using a cytogenetic photomap for An. gambiae >George, 2010 #4780<, two BAC clones, 102B24 (GenBank: BH372701, BH372694) and BAC 142O19 (GenBank: BH368703, BH368698), were localized in subdivisions 16C and 16D of the 2R arm, respectively. However, the BLAST search against the An. gambiae PEST strain AgamP3 assembly [269] identified the sequences homologous to 102B24 and 142O19 in subdivisions 16B and 17A of the 2R arm, respectively **(Table 2.1)**. Therefore, the correspondence between the genomic coordinates and the cytogenetic subdivisions can now be adjusted according to our mapping data. The BLAST search against the An. gambiae M form and S form genome assemblies found that the two BAC clones are located in different contigs, but in the same scaffolds within each form **(Table 2.1)**. The identified contigs can now be associated with specific chromosomal locations. Moreover, the identified scaffolds can now be properly oriented within the cytological subdivisions 16CD. Interestingly, the distances between the 102B24 and 142O19 sequences in the scaffolds are 1,892,981 bp, 1,658,391 bp, and 1,688,426 bp in the PEST strain, M-form, and S-form genome assemblies, respectively. Because, the PEST strain is a hybrid between the M and S forms, this difference is likely due to incorrect assembly of the PEST genome.

BAC (accession)	PEST-strain AgamP3	M-form contigs	M-form scaffolds	S-form contigs	S-form scaffolds
	Coordinates	Coordinates	Coordinates	Coordinates	Coordinates
102B24	2R:AAAB01008844_26	ABKP020247	EQ090167.1	ABKQ010123	EQ099711.1
(BH372694)	(16B)	53.1		32.1	
	43,681,342-43,681,874	32,513-33,045	1,858,377-1,858,909	4,299-4,832	215,891-216,424
102B24	2R:AAAB01008844_28	ABKP020247	EQ090167.1,	ABKQ010123	EQ099711.1,
(BH372701)	(16B)	51.1		35.1	
	43,788,213-43,788,969	24,816-25,575	1,772,683-1,773,442	27,782-28,540	305,514-306,272
142O19	2R:AAAB0100	ABKP020247	EQ090167.1	ABKQ010123	EQ099711.1
(BH368698)	8805_5 (17A)	1.1		76.1	
	45,428,580-45,429,264	2,499-3,185	315,806-316,492	49,242-49,942	1,791,625-1,792,325
142O19	2R:AAAB0100	ABKP020220	EQ090167.1	ABKQ010123	EQ099711.1
(BH368703)	8805_8 (17A)	17.1		79.1	
	45,560,099-45,574,323	30,971-32,469	200,518-202,016	27,760-28,508	1,903,569-1,904,317

Table 2.1. BLAST results of the BAC-end sequences against the genome assemblies of An. gambiae.

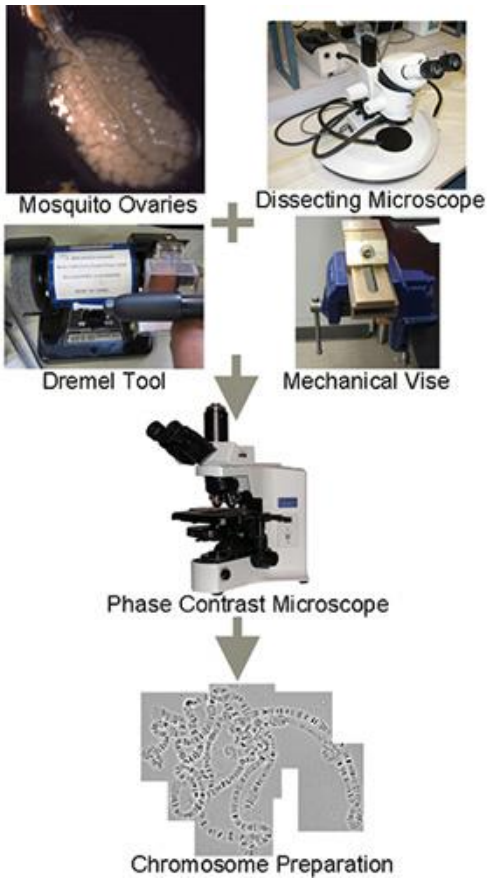


Figure 2.1. Schematic representation of high-pressure chromosome preparation. **Mosquito ovaries are shown at the correct stage of development.**

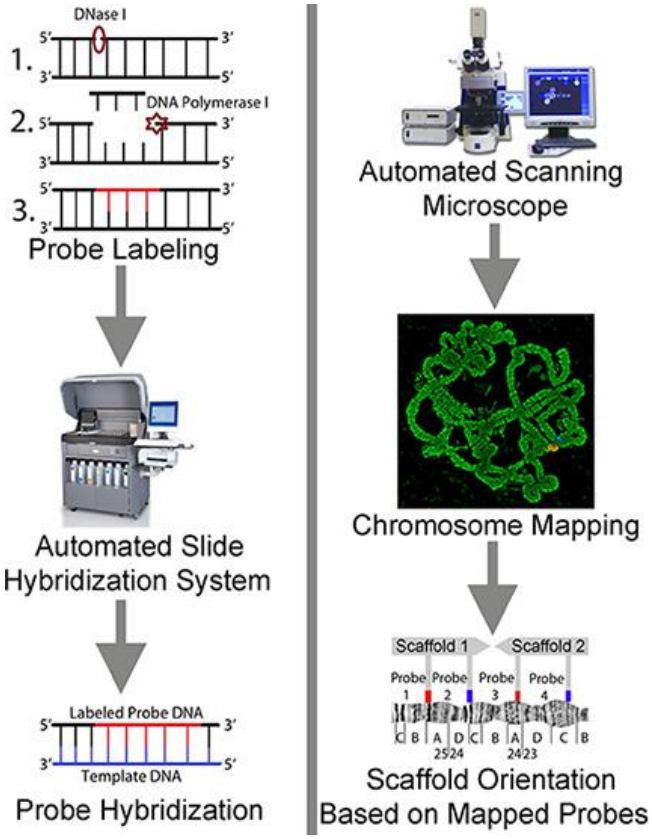


Figure 2.2. A scheme representing automated FISH, slide scanning, and chromosome mapping of genomic scaffolds.

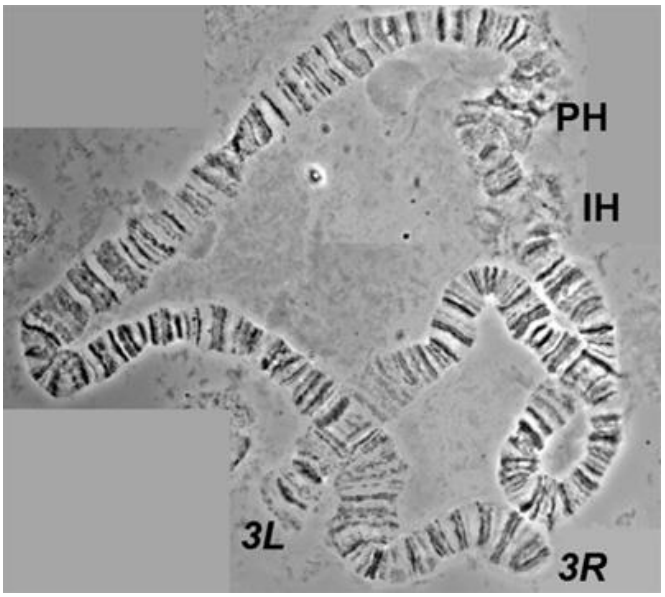


Figure 2.3. A spread of polytene chromosome 3 of An. gambiae obtained using the high-pressure technique.
3L and 3R mark left and right arms at their telomeres. PH and IH indicate pericentric and intercalary heterochromatin, respectively.

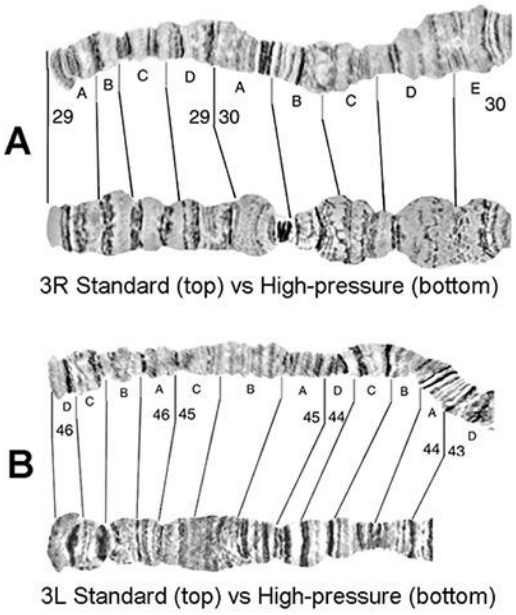


Figure 2.4. Comparison of An. gambiae polytene chromosomes prepared using the traditional (top) and high-pressure (bottom) technique.
The images cover subdivisions 29A-30E of arm 3R (A) and 43D-46D of arm 3L (B).

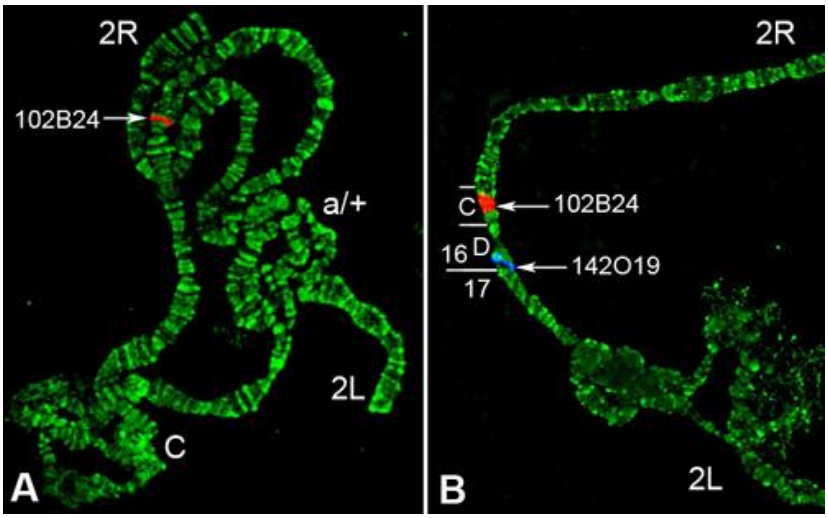


Figure 2.5. FISH of BAC clones to polytene chromosomes of An. gambiae.
A) Hybridization of 102B24 (red signal) with the 2R arm. B) Dual-color FISH of 102B24 (red signal) and 142O19 (blue signal) to subdivisions 16C and 16D of the stretched 2R arm, respectively. Arrows indicate signals of hybridization of BAC clones labeled with Cy3 (red) and Cy5 (blue). C-the centromeric region. a/+ shows the heterozygote 2La inversion. Chromosomes were counterstained with the fluorophore YOYO-1.

2.2.4. Discussion

The most critical step of the high-pressure procedure is the proper squashing of ovarian nurse cells isolated at Christopher's' III stage of ovarian development [270]. Improper squashing can lead to insufficiently spread chromosomes, which can cause problems when trying to determine probe locations after FISH. If the chromosomes are over-squashed, they can be become broken or elongated to the point where banding patterns are lost. Production of multiple slides should lead to a consistency when attempting to squash the slides using the Dremel tool, which will increase overall slide production efficiency. The high-pressure technique was first developed for freshly isolated salivary glands of D. melanogaster [264]. However, ovaries of mosquitoes are routinely preserved in modified Carnoy's fixative solution (3 methanol: 1 glacial acetic acid by volume) before they are used for chromosomal preparations. Therefore, we modified the existing high-pressure protocol to make it suitable for the fixed ovarian nurse cell polytene chromosomes of the malaria mosquito An. gambiae. Because the high pressure is applied using a precision vise possessing a highly parallel work surface of the entire slide, it takes significantly less time to prepare the chromosome squash than using a traditional tapping technique with a pencil's eraser.

Other important steps include sufficiently soaking follicles in 50% propionic acid and heating the slide. Both of these steps are essential in helping to flatten the chromosomes. If they are neglected, chromosomes can appear shiny after dehydration, which potentially leads to an overabundance of background that can be mistaken for a signal in FISH. The high-pressure squashing technique also works well using the same solutions when making mitotic chromosome preparations. The additional spreading and flattening helps to better express the chromosomes

from their nuclei. Adding equal pressure should flatten them accordingly to allow for measurements and potentially better resolution of bands visible from staining. The details about isolating mitotic chromosomes from mosquitoes are given elsewhere [271]. Although regular squash preparations are sufficient for many purposes including FISH [272] and immunostaining [273], the high-pressure method not only lowers potential variance from one slide to the next, but also increases overall chromosome quality, leading to higher detail when mapping chromosomes. This procedure can also be used for preparing polytene chromosome membrane slides for laser-capture microdissection.

Limitations to the high-pressure method include slide breakage and overstretching of the chromosomes. Slide breakage caused by placing too much stress on the slide via the vise is possible, but is limited by using the pressures denoted in the article. For overstretching of the chromosomes, if the Dremel tool is applied to the slide for an extended period of time, there is the possibility that the chromosomes can become too stretched out and lose resolution. This can be remedied by applying the tool for a brief period of time, checking under the microscope, and applying more time with the Dremel tool if the chromosomes are insufficiently spread.

A traditional FISH protocol includes a number of washing and incubation steps, which are usually 5-20 min long and require almost full attention of a researcher for the whole duration of the experiment. Moreover, the number of slides that can be handled manually is usually limited to a few slides in a given experiment. In contrast, an automated slide staining system performs all steps (including washing, incubation, probe application, denaturation, hybridization, coverslip application and removal) automatically. This frees up to 6-8 working hours for a researcher in a FISH experiment. Moreover, an automated FISH system can significantly increase the throughput. For example, the Xmatrx system is capable of processing up to 40

assays on single preparation slides and up to 80 assays on dual preparation slides simultaneously. Among limitations of this system is that it is not efficient for a small number of FISH experiments as it requires preparing large volumes of solutions. In addition, the FISH protocol programmed in the system may require modifications and adjustments for new applications.

A slide scanning system is an automated version of a fluorescent microscope. Automated stage moving and a simple microscope control panel free a researcher's time and make operating the microscope extremely simple. For instance, the software in the ACCORD PLUS scanning system allows for easy capturing of multiple channels of fluorescence, and easy to manage image acquisition. An example of this is the inclusion of z-stack capturing, rather than taking a single image; the software captures a configurable z-stack of images to ensure that at least one image remains in focus. Although this system is made more for multiple image acquisition (a lot of cells on a single slide), it still makes finding chromosomes on the slide much easier than navigating through the slide manually.

Both the automated slide staining system and the scanning system are routinely used for FISH diagnostics of chromosomal abnormalities in human cells by cytogenetics clinics. In this study, we developed a protocol that utilizes these automated systems for a basic research application. The automated high-throughput physical mapping protocol can facilitate rapid development of physical chromosome maps for any species of interest. Although, in this report, we used polytene chromosomes for physical mapping, mitotic chromosomes can also be utilized with these systems. It would be useful for the protocol to be adjusted to mitotic chromosomes because the majority of organisms do not develop readable polytene chromosomes. However, the polytene chromosomes, when available, can provide useful information about the correspondence of functional genome domains with the chromosome structure at the highest

resolution. The organizational principles of polytene chromosomes, such as patterns of bands and interbands, have recently been likened to that of regular nonpolytene (interphase) chromosomes [274]. Therefore, the detailed physical mapping performed on high-pressure chromosome preparations has the potential to link DNA sequences to specific chromosomal structures such as bands, interbands, puffs, centromeres, telomeres, and heterochromatin; thus, creating chromosome-based genome assemblies.

2.3 High-Resolution Cytogenetic Map for the African Malaria Vector *Anopheles gambiae*

Phillip George, Maria V. Sharakhova, and Igor V. Sharakhov*

as published in the Insect Molecular Biology, Oct. 2010

2.3.1 Abstract

Cytogenetic and physical maps are indispensible for precise assembly of genome sequences, functional characterization of chromosomal regions, and population genetic and taxonomic studies. We have created a new cytogenetic map for *Anopheles gambiae* by using a high-pressure squash technique that increases overall band clarity. To link chromosomal regions to the genome sequence, we attached genome coordinates, based on 302 markers of BAC, cDNA clones, and PCR-amplified gene fragments, to the chromosomal bands and interbands at approximately a 0.5-1 Mb interval. In addition, we placed the breakpoints of seven common polymorphic inversions on the map and described the chromosomal landmarks for the arm and inversion identification. The map's improved resolution can be used to further enhance physical mapping, improve genome assembly, and stimulate epigenomic studies of malaria vectors.

2.3.2 Introduction

Anopheles gambiae is a prominent vector in the transmission of the malaria parasite Plasmodium falciparum in Africa, responsible for more than one million deaths annually [275]. Understanding the genetics of mosquito may explicate why *An. gambiae* is such a proficient vector of the disease. Cytogenetic maps, which depict chromosome structures and banding patterns, provide a basis for many types of genome analysis including: physical mapping, synteny investigation, whole genome sequence assembling, and positional cloning [263]. Anopheline mosquitoes, as other *Diptera*, have giant polytene chromosomes that undergo endoreplication and can be found in various tissues [52]. These chromosomes are characterized by various levels of compaction that are evident from light and dark bands, as well as by diffuse puffs and specific structures of heterochromatic regions. Banding patterns are mostly consistent within a species and, in some cases, somewhat consistent between closely related species. Drawn and photo chromosomal maps have been developed for about 50 species from the genus *Anopheles* [276]. During the last decade, cytogenetic photomaps based on digital imagining were created for the polytene chromosomes *of An. albimanus* [277], *An. funestus* [92], and *An. stephensi* [278].

The major malaria vector *An. gambiae* has karyotype 2n=6. Because of homologous chromosome paring, the polytene chromosome complement includes 5 arms: X, 2R, 2L, 3R, and 3L. The first preliminary cytogenetic map for *An. gambiae* polytene chromosomes from larval salivary glands was developed in 1956 [279]. Later, more detailed drawn chromosomal maps were created for *An. gambiae* and for the X chromosome of *An. arabiensis*, known at that time as species B [53]. The fixed inversions identified based on the banding pattern became a key for the species diagnostics. Finding polytene chromosomes in ovarian nurse cells significantly improved

the quality of the chromosomal images and simplified the procedure of slide preparation [280]. In the 1970's and 1980's, chromosomal maps became the major tool for investigating the mosquito's genetics and evolution. Cytogenetic analysis of chromosomal inversions helped to identify species within the *An. gambiae* complex [54,281,282] and led to the discovery of chromosomal forms within *An. gambiae* s.s. [283,284].

In 2002, a newly drawn high-quality map of polytene chromosomes from ovarian nurse cells was published [56]. This map, together with unassembled digital photo images of chromosomes (VectorBase.org), was utilized for the physical mapping and assembly of the *An. gambiae* genome [63,256]. The positions of ~2000 BAC clones were assigned to the chromosomal locations. However, the current genome assembly suffers from a number of physical gaps, incorrect scaffold orientation, and the presence of 42 Mb of unmapped sequences. These deficiencies hinder further accurate annotation and functional characterization of the *An. gambiae* genome. The majority of the gaps are located in gene-poor repeat-rich heterochromatic regions. More detailed physical mapping is needed to improve the genome assembly and to link heterochromatin to the genome sequence. A high-resolution cytogenetic photomap should be useful for precise physical mapping and for studying epigenetic chromatin modifications and their relationships with the genome sequence.

A novel high-pressure method of polytene chromosome preparation was developed and tested on *Drosophila melanogaster* [265]. In our study, a modified high-pressure method was applied to the chromosomes of the malaria mosquito *An. gambiae*. By increasing the overall resolution of the cytogenetic map and chromosome squashes in general, we were able to better determine specific gene locations along the chromosome. The new map will make it possible to study how different levels of chromatin compaction affect functionality of chromosomal regions.

Therefore, the improved resolution can be used to further enhance physical mapping and epigenomic studies of malaria vectors.

2.3.3 Results

Development of the high-resolution map

Here, we present a high-resolution cytogenetic photomap of ovarian nurse cell polytene chromosomes for the major malaria vector *An. gambiae* **(Figure 2.6)**. The map was developed using a modified version of a novel high-pressure technique [265] that applies ~150 kg/cm2 of pressure to a slide placed in a vice. Approximately 90% of slides created with this method were still intact and functional after adding mechanical pressure. The 10% failure rate was due to slide breakage from the high pressure. This pressure provided a better resolution image so that the banding pattern became clearer and many new small bands became visible **(Figure 2.7A)**. The high-pressure method does not damage or change most of the chromosome structure. It flattens bended chromosomes, and thus, reveals hidden bands. The diffuse structure of the heterochromatin was more evident through the addition of high-pressure flattening. However, often the heterochromatin and other chromosomal regions ended up broken and unusable for mapping. Therefore, several regions on the map were taken from images using squash procedures with less pressure. Chromosomes from ten slides were used to create this map. The chromosome map includes images for five chromosomal arms and mostly retains previously described divisions and subdivisions [56]. For the convenience of physical mapping, region 6 on chromosome X was additionally subdivided into three lettered subdivisions -- A, B, and C, -- which are absent from the 2002 map. However, this region included four lettered subdivisions on the original drawn map (published in 1967) for the salivary gland chromosomes [54]. To make

viewing easier, chromosomal images were straightened and shaped similarly to those on the drawn map [56] as shown in **(Figure 2.7B)**.

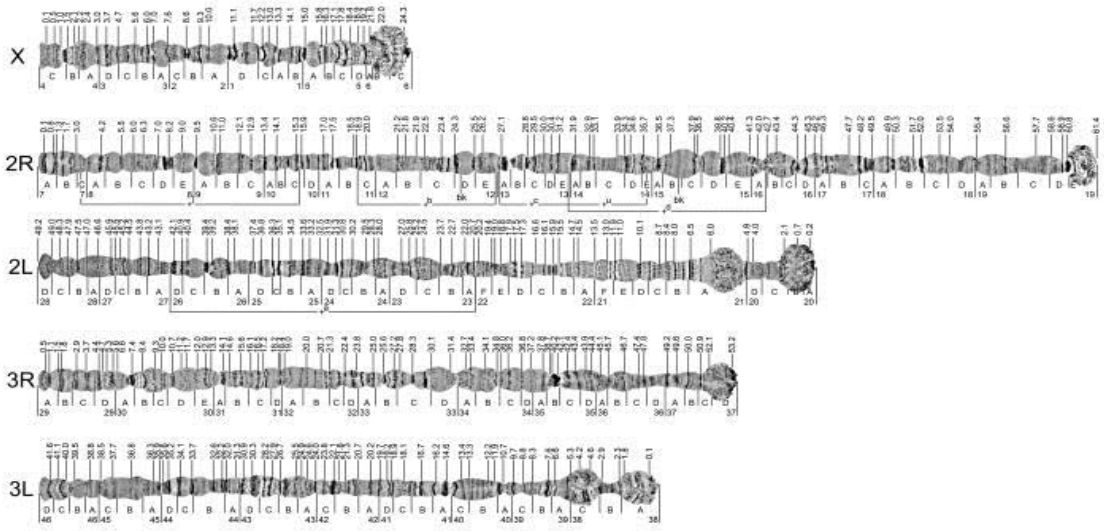


Figure 2.6. Cytogenetic photomap of *An. gambiae* ovarian nurse cell chromosomes. Genome coordinates are shown above X, 2R, 2L, 3R, and 3L chromosomal arms. Divisions and subdivisions are indicated by numbers and letters below the chromosomes. The location of common polymorphic inversions +j, +b, +c, +bk +u, +d, and +a is shown in their standard orientation by brackets below the chromosomes.

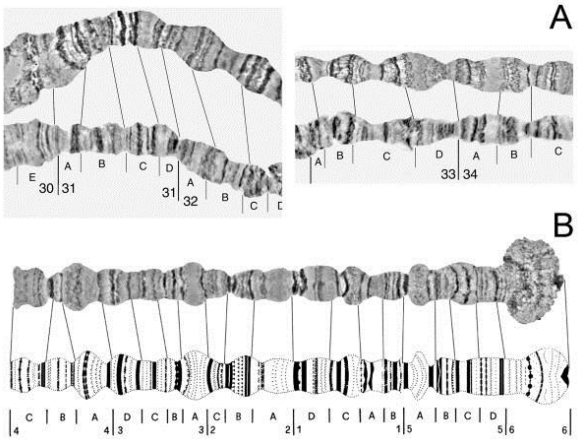


Figure 2.7. (A) Comparison of *An. gambiae* polytene chromosome spreads prepared using the high-pressure (top) and traditional (bottom) technique. The images cover divisions 30–34 of the 3R arm. (B) Comparison of the high-resolution photomap (top) and drawn map (Coluzzi et al. 2002) (bottom) of the X chromosome of *An. gambiae*. The vertical lines connect the same regions.

Landmarks for recognition of chromosomal arm

Table 2.1 provides the description of the landmarks routinely used in our laboratory for chromosomal arm identification. Chromosome X is usually located separately from other arms on the chromosomal preparations. Chromosomes 2 and 3 form a diffuse chromocenter, which can be easily destroyed by squashing the slide [92]. Nevertheless, left and right arms of both chromosomes typically remain attached to each other. The most robust landmarks for all chromosomal arms are the length of the arm and morphology of the telomeric and centromeric regions.

Chromosome	Telomere	Centromere	Additional Landmarks
X	Flared light zone	Diffuse granulated area ending with dark round band in NOR, sometimes asynaptic	Large puffs in 4A, 3A, 5A regions, two large dark bands in region 1D
2R	Two couples of dark thick bands in region 7A	Dark granulated area in region 19E with dark large band in region 19D	Two large bands followed by three thick bands in region 17C
2L	Very light flared area	Dark granulated area	Diffuse light granulated area in region 21A
3R	Dark slightly flared dark band in the distal area	Dark granulated area	Dark block in region 35B
3L	Light slightly flared zone with dark band in the proximal area	Dark granulated area	Large granulated area in region 38C

NOR – Nuclear Organizing Region

Table 2.2. Major landmarks of the banding pattern for the *An. gambiae* chromosomal arm recognition

Chromosome X, the shortest in the chromosomal complement, can be easily recognized by the flared, lightly stained telomeric end in region 1A and the diffuse granulated heterochromatic area on the centromeric end in region 6. Heterochromatin of region 6 ends with a dark round band on the very proximal end, which corresponds to the nucleolus organizing region [63]. The centromeric region can be asynaptic in some squashes. In addition, several specific structures can also be considered as landmarks for the X chromosome: large puffs in regions 4A, 3A, and 5A and two large dark bands in region 1D.

Chromosome 2 is the longest in the polytene complement. The 2R arm can be easily recognized by two couples of dark thick bands in region 7A, a dark granulated centromeric area in region 19E, and a dark large band in region 19D. An additional robust landmark for this arm is a very stable structure with two large bands followed by three thick bands in region 17C. The 2L chromosomal arm has always had a light flared telomeric end (28D region) and a large dark granulated centromeric end (20A-B). An additional specific landmark for 2L arm is a diffuse light granulated area in region 21A, which usually forms the attachment of the chromosome to the nuclear envelope [92].

Chromosome 3 is relatively shorter than chromosome 2. Both arms of chromosome 3 have slightly flared telomeric ends; 3R has a dark band in the distal region of the telomere, and 3L has a very light telomere and a dark band located proximally in region 46D. Centromeric areas for both arms have a very similar, dark granulated morphology. Additional landmarks for chromosome 3 include a very dark block in region 35B of the 3R arm and a large granulated area in region 38C of the 3L arm.

Landmarks for recognition of polymorphic inversions

The location of the six common polymorphic inversions on the map is indicated by brackets below the chromosomes: +j, +b, +c, +u, and +d on chromosome arm 2R and +a on chromosome arm 2L **(Figure 2.6)**.

Breakpoints for the seventh common polymorphic inversion 2Rbk are indicated by lines because this inversion is based on the preexisting 2Rb inversion [56]. Inversion j is easily recognizable by a series of four dark bands, located in the proximal half of region 8D, and by a large, light puff in region 9A surrounded by a narrow dark band (9A distal) and two wide bands (9A proximal and 9B distal) **(Table 2.3)**. Two distinct bands in region 8B followed by four lightly granulated double bands in region 8C can be utilized as additional landmarks for inversion j. A landmark for inversion b is a light-bulb-looking puff in region 12D surrounded by four dark bands on the left in subdivisions 12C-12D and one dark band on the right in region 12E. An additional landmark for the b inversion is a light area with four diffuse bands in region 12A which has no dark bands. Landmarks for inversion c are the very dark wide band followed by a narrow band in region 13A and two distinct thick bands in regions 13D-13E. Distal parts of regions 13B and 13C form two narrow zones, which make the whole structure look like a tightened asymmetrical ribbon. Inversion u also has a narrow zone in the middle of region 14C, one wide band in region 14A, two thick bands located close to each other in region 14B, and three distinct bands located equidistantly in regions 14D-14E. Inversion d is overlapped with inversion c in region 14. In addition to the c inversion landmarks, two large puffs surrounded by dark bands on both sides in regions 15 BC and 15D can be utilized as landmarks for the d inversion. Inversion bk overlaps with the b, c, u, and d inversions; therefore, the same landmarks can be used for its recognition. Inversion a on the 2L arm is the biggest common inversion in the

An. gambiae genome. This inversion can be identified by two light puffs surrounded by very dark bands located distally in region 26. On the opposite proximal part of the inversion, two large puffs in regions 24A and 23C have a dark granulated structure.

Inversion	**Major landmarks**	**Additional landmarks**
2R*j*	Four dark bands located in proximal half of the region 8D; large light puff in region 9A surrounded by the narrow dark band in distal 9A; two wide bands located in regions 9A proximal - 9B distal	Two distinct bands in region 8B followed by four light granulated double bands in region 8C
2R*b*	"Bulb" - a puff in region 12D surrounded by four dark bands in regions 12C-12D and one dark band in region 12E	Light area with four diffuse bands in region 12A, no dark bands around
2R*c*	"Ribbon" - very dark wide band followed by narrow band in region 13A; distal parts of regions 13B and 13C form two narrow zones - the middle part of the "ribbon"	Two distinct thick bands in regions 13D-13E
2R*u*	One wide band in region 14A and two thick bands located close to each other in region 14B and three distinct bands located at the same distance in regions 14D-14E	Narrow zone in the middle of region 14C
2R*d*	Narrow zone in the middle of region 14C with couple of double bands in regions 14A-14B located distally and three distinct bands located proximally in regions 14D-14E	Two large puffs surrounded with dark bands on both sides in regions 15B-C and 15D
2L*a*	Two light puffs surrounded by very dark bands located distally in the inversion in region 26	Two large granulated puffs in regions 24A and 23C with no dark bands around

Table 2.3. Chromosomal landmarks for recognizing common polymorphic inversions in *An. gambiae*

Genome coordinates on the cytogenetic map

To make the cytogenetic map suitable for genomic analyses, bands and interbands were given approximate genome coordinates **(Figure 2.6, Table S1)**. Most of these coordinates represent the genome positions of BAC and cDNA clones, which were physically mapped to chromosomes [63,256]. Although the *An. gambiae* genome was mapped by *in situ* hybridization of ~2,000 BAC clones with polytene chromosomes [256], a large portion of BAC clones hybridized to overlapping or multiple chromosomal locations. Moreover, signals from nonfluorescent images often covered several bands and interbands. We chose markers that hybridized specifically to only one band or interband. The resolution of this physical map varies in the different locations of the chromosomes depending on the precision of the *in situ* hybridization. We attempted to provide at least one coordinate per 0.5 Mb of chromosome length. Because heterochromatic regions were not sufficiently covered with markers, we performed additional physical mapping of PCR-amplified gene fragments **(Figure 2.8, Table S1)**. The size of the mapped portion of the *An. gambiae* genome is approximately 230.5 Mb (VectorBase.org), giving the map an overall marker density of 0.76 Mb (230.5 Mb/302 markers). Individual chromosomal arm resolutions come out to be 0.76 Mb for X (24.4/32), 0.78 Mb for 2R (61.5/79), 0.71 Mb for 2L (49.4/70), 0.78 Mb for 3R (53.2/68), and 0.79 Mb for 3L (42.0/53).

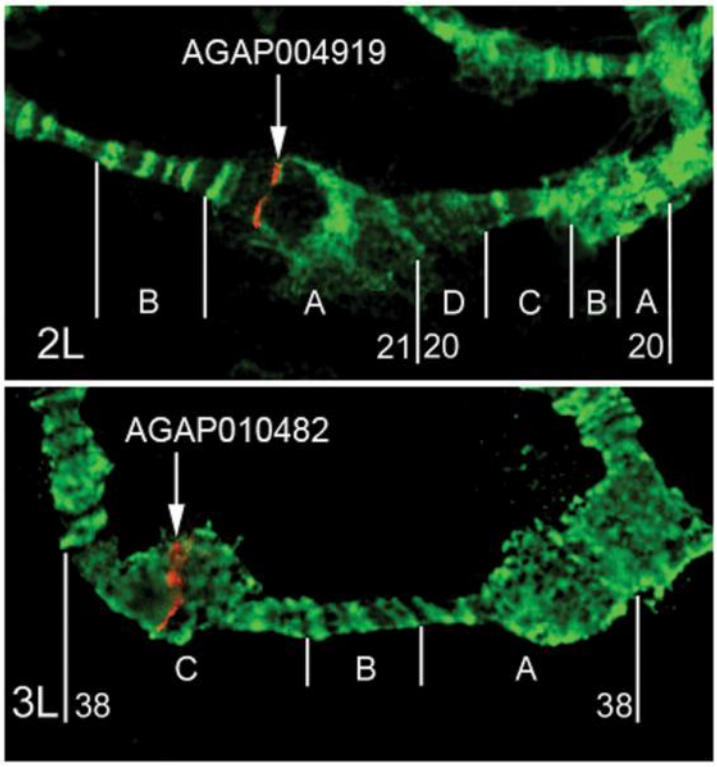


Figure 2.8. Fluorescent in situ hybridization of fragments of heterochromatic genes to the polytene chromosomes 2L and 3L of *An. gambiae*. Arrows indicate signals of hybridization in diffuse intercalary heterochromatin.

2.3.4 Discussion

We present in this study a new cytogenetic photomap for the major malaria vector *An. gambiae*. This map is based on a modified high-pressure technique that provides greater details of chromosomes and more accurately represents the observable banding patterns **(Figure 2.7)**. This technique was first developed for freshly isolated salivary glands of *D. melanogaster* [265]. However, ovaries of mosquitoes are routinely preserved in the Carnoy's fixative solution (3 methanol: 1 glacial acetic acid by volume) before they are used for chromosomal preparations. Therefore, we modified the existing high-pressure protocol to make it suitable for the fixed ovarian nurse cell polytene chromosomes of the malaria mosquito *An. gambiae*. Several

technical problems were encountered and successfully eliminated. Slide breakage was overcome by placing a second coverslip on the opposite side of the slide as a counterbalance. Excessive bubbling was removed by using a new solution (40% propionic acid, 40% water, 20% acetic acid) instead of a 50% propionic acid solution in our original protocol. Our results indicate that the high-pressure method produces extremely flat chromosomes and significantly improves structural resolution of the banding pattern. To make a spread, we used the eraser end of a pencil to apply mechanical force to the coverslip, expressing the chromosomes from the nuclei. The high-pressure squash was achieved by placing a slide into a vice and applying ~150 kg/cm^2 of pressure. This pressure provided a higher resolution image so that the banding pattern became clearer and many new small bands became visible **(Figure 27A)**. Because high pressure is applied using precision vises possessing highly parallel work surfaces, several chromosomal preparations can be made at the same time to increase throughput.

To make this map suitable for population genetics studies, we indicated chromosomal positions for seven common polymorphic inversions and described landmarks for their recognition. Inversions were shown in their standard orientation to make this map consistent with previous maps. The standard orientation of inversions in our map may not correspond to their ancestral arrangement. For example, the derived state of the standard 2L+a arrangement has been demonstrated in several independent studies [278,285,286].

To facilitate further physical mapping and genome assembly of the *An. gambiae* genome, we described landmarks for chromosomal arm recognition and put approximate genome coordinates on bands, interbands, and heterochromatic regions. This photomap can now be used for various applications in cytogenetics and comparative genomics. For example, the molecular content of intercalary heterochromatin, euchromatin-heterochromatin borders, puffs, and

inversions can now be determined at a higher resolution. The packaging of DNA into euchromatin and heterochromatin of *An. gambiae* could have major implications for understanding how genome sequences function and respond to *Plasmodium* infection. Further use of high-pressure chromosomal preparations for microdissection and sequencing of specific regions will directly link the chromatin structure to the genome sequence. Similarly, using high-pressure chromosomal preparations for immunostaining will improve the precision of mapping chromosomal proteins and histone modifications involved in DNA replication, gene expression, gene silencing, and inheritance.

The high-pressure technique can also be successfully applied to other species from the genus *Anopheles* and, probably, could improve the quality of chromosomal preparations from *Anopheles* salivary glands. We would also recommend using this method for species from the genus *Culex* because the *Culex* polytene chromosomes are hardly spreadable with traditional techniques [287,288]. However, a very low yield of chromosomal preparations from salivary glands or Malpighian tubules of *Aedes aegypti* [289] will make it more difficult to use the high-pressure method for this species.

2.3.5 Experimental procedures

Mosquito strain

A laboratory SUA strain of *An. gambiae* was used in this study. These mosquitoes are the M form of *An. gambiae*. Mosquitoes were reared at 28°C at 80% humidity. Mosquitoes were grown at a low density (500-750 mosquitoes per 4 liter pan) to obtain better quality chromosomes. Larvae were fed ad libitum. Adults were given sugar water through dampened cotton balls that were removed at least 2 hours pre-blood feeding to ensure that most mosquitoes

had taken a blood meal. To obtain the chromosomal preparations, females were bloodfed twice with a Guinea pig.

Chromosome preparation

The experiment involved the use of fixed ovaries. Both the Carnoy's solution, 3:1 methanol: acetic acid, and 50% propionic acid were made fresh prior to slide preparation. Ovaries of *An. gambiae* SUA mosquitoes were dissected approximately 25 hours post-blood feeding at Christopher's Stage II of development. Only ovaries displaying a slightly oval shape were collected; elongated and circular ovaries were discarded. Dissected ovaries were placed in fresh Carnoy's solution for 24 hours at room temperature, followed by a decrease in temperature to −20°C until time of preparation.

Fixed ovaries were divided in Carnoy's solution under a dissection microscope. Any tissue other than follicles was removed from the slide via tissue paper. After tissue removal, the divided sections of follicles were placed onto slides containing 40% propionic acid, 40% water, 20% acetic acid (usually about four to six slides per set of ovaries, or half to one third of an ovary per slide). After approximately 5 to 10 minutes, the acid mix was replaced by a drop of 50% propionic acid to further remove any remaining tissue. Follicles were then separated from each other, and the drop of acid was replaced once again. A coverslip was lightly placed on top of the ovaries in solution, and light mechanical-force tapping created by using the eraser end of a pencil was applied to the entire surface to express the chromosomes from the nuclei. Once sufficiently spread, a mechanical vise was used to evenly apply pressure to further flatten chromosomes on the preparation. Approximately 150 kg/cm^2 of pressure was applied through the vise for a 30-second interval. Banding patterns of the chromosomes were viewed under an Olympus BX-41 phase-contrast microscope on wet slides.

Chromosome imaging and assembling

Chromosomes that showed a suitable level of polytenization were imaged by an Olympus BX-41 (Olympus America, Inc., Melville, NY, USA) with an attached Olympus Q-Color 5 camera and Q-Imaging software. Images were combined, straightened, shaped, and cropped using Adobe Photoshop. To link the chromosomal bands to genome sequences, previous images of BAC and cDNA clone *in situ* hybridizations were utilized [63,290].

Fluorescent *in situ* hybridization

PCR probes were chosen from poorly assembled regions of the *An. gambiae* genome. Many of these probes were based on genes located near expected heterochromatin - euchromatin boundaries on each chromosome arm. Primers were designed using the Primer3 program (http://frodo.wi.mit.edu/primer3/). PCR products ranged from 400-600 bp in size. The *in situ* hybridization procedure was done as previously described [291]. PCR products were gel purified using the Geneclean kit (Qbiogene, Inc., Irvine, CA). The DNA was labeled with Cy3-AP3-dUTP (GE Healthcare UK Ltd., Buckinghamshire, England) using Random Primers DNA Labeling System (Invitrogen Corporation, Carlsbad, CA, USA). DNA probes were hybridized to the chromosomes at 39°C overnight in hybridization solution (Invitrogen Corporation, Carlsbad, CA, USA). Then the chromosomes were washed in 0.2XSSC (Saline-Sodium Citrate: 0.03M Sodium Chloride, 0.003M Sodium Citrate) counterstained with YOYO-1, and mounted in DABCO. Fluorescent signals were detected and recorded using a Zeiss LSM 510 Laser Scanning Microscope (Carl Zeiss MicroImaging, Inc., Thornwood, NY, USA).

2.4 Genome mapping and characterization of the *Anopheles gambiae* heterochromatin

Maria V Sharakhova, Phillip George, Irina V Brusentsova, Scotland C Leman, Jeffrey A Bailey, Christopher D Smith and Igor V Sharakhov

as published in the BMC Genomics , Aug 2010.

2.4.1 Abstract

Background

Heterochromatin plays an important role in chromosome function and gene regulation. Despite the availability of polytene chromosomes and genome sequence, the heterochromatin of the major malaria vector *Anopheles gambiae* has not been mapped and characterized.

Results

To determine the extent of heterochromatin within the *An. gambiae* genome, genes were physically mapped to the euchromatin-heterochromatin transition zone of polytene chromosomes. The study found that a minimum of 232 genes reside in 16.6 Mb of mapped heterochromatin. Gene ontology analysis revealed that heterochromatin is enriched in genes with DNA-binding and regulatory activities. Immunostaining of the *An. gambiae* chromosomes with antibodies against *Drosophila melanogaster* heterochromatin protein 1 (HP1) and the nuclear envelope protein lamin Dm0 identified the major invariable sites of the proteins' localization in all regions of pericentric heterochromatin, diffuse intercalary heterochromatin, and euchromatic region 9C of the 2R arm, but not in the compact intercalary heterochromatin. To better understand the molecular differences among chromatin types, novel Bayesian statistical models were developed to analyze genome features. The study found that heterochromatin and

euchromatin differ in gene density and the coverage of retroelements and segmental duplications. The pericentric heterochromatin had the highest coverage of retroelements and tandem repeats, while intercalary heterochromatin was enriched with segmental duplications. We also provide evidence that the diffuse intercalary heterochromatin has a higher coverage of DNA transposable elements, minisatellites, and satellites than does the compact intercalary heterochromatin. The investigation of 42-Mb assembly of unmapped genomic scaffolds showed that it has molecular characteristics similar to cytologically mapped heterochromatin.

Conclusions

Our results demonstrate that *Anopheles* polytene chromosomes and whole-genome shotgun assembly render the mapping and characterization of a significant part of heterochromatic scaffolds a possibility. These results reveal the strong association between characteristics of the genome features and morphological types of chromatin. Initial analysis of the *An. gambiae* heterochromatin provides a framework for its functional characterization and comparative genomic analyses with other organisms.

2.4.2 Background

Located in pericentric, telomeric, and some internal chromosomal regions, heterochromatin plays an important role in cell division [292], meiotic pairing [293], regulation of DNA replication, and gene expression [294]. Among insect species, the most detailed analysis of heterochromatin has been performed in *Drosophila* [295-298]. Molecular analysis has determined that pericentric heterochromatic regions are enriched with highly and moderately repetitive DNA sequences, and are extremely depleted of genes [258,299,300]. Mapping of heterochromatic scaffolds is difficult because the heterochromatin is underreplicated and poorly banded in polytene

chromosomes of salivary glands. Special efforts had to be directed towards the assembly and annotation of heterochromatin in *Drosophila* [258,301-304]. Bioinformatic analysis of the heterochromatic portion of the *Drosophila* genome revealed the presence of more than 200 genes. Interestingly, heterochromatic genes are enriched specific functional domains, including putative membrane cation transporters domains and domains involved in DNA or protein binding [302]. This finding suggests that pericentric heterochromatin may encode genes involved in the establishment or maintenance of alternative chromatin states. In addition to the pericentric heterochromatin, *Drosophila* has intercalary heterochromatin, which is interspersed throughout the euchromatin and characterized, in part, by underreplication in polytene chromosomes of larval salivary glands [305,306]. A study of a genome-wide profile of underreplication in polytene chromosomes identified 52 underreplication zones, which were colocalized with regions of intercalary heterochromatin. These underreplication zones varied from 100 to 600 kb in length, and each contained from 6 to 41 unique genes [307].

One of the important problems of chromosome biology is to understand the relationships between the morphology of the chromatin and the DNA and protein composition. Two morphological types of the heterochromatin have been described in the pericentromeric regions of *Drosophila* polytene chromosomes: proximal condensed, α-, and distal diffuse, β-heterochromatin [308]. The compact central part of the chromocenter (α-type) is enriched with satellite DNA, while the distal diffuse area (β-type) contains mostly transposable elements (TEs) [309,310]. Biochemical studies have discovered that heterochromatic regions have a specific histone code, characterized by hypoacetylation and methylation of the histone H3 at lysine 9 [311]. This modification of the histone H3 is a docking site for the heterochromatin protein 1 (HP1) [312,313], a major component of heterochromatin first described in *Drosophila* [314]. Comparative studies of

Drosophila polytene chromosomes have discovered differences in the chromatin state suggesting the switching of chromatin states during evolution. For instance, when staining patterns of HP1 on polytene chromosomes were compared, it was found that the heterochromatic fourth chromosomes of *D. melanogaster* and *D. pseudoobscura* bind to HP1, while the euchromatic fourth chromosome of *D. virilis* does not. Interestingly, the level of CA/GT repeats on chromosome 4 of *D. virilis* is 20 fold higher than the level on chromosome 4 of *D. melanogaster*. Moreover, the density of TEs in this chromosome is significantly higher for *D. melanogaster* than for *D. virilis* [315-317].

A number of studies have demonstrated direct associations between heterochromatin and the nuclear envelope (NE) [318-323]. In *Drosophila* salivary gland nuclei, pericentromeric heterochromatin attaches permanently to the NE, while intercalary heterochromatin forms high-frequency contacts to NE [324]. Chromatin fibers of diffuse heterochromatin form visible attachments to the NE in *Drosophila* [320] and *Anopheles* [325,326]. The chromosomal regions that attach to the NE may depend on the presence of specific DNA. For example, repetitive matrix attachment regions (MARs) specifically bind to lamin, the major protein of the nuclear periphery [318,327-330]. It has been shown that MAR DNA is several fold richer in heterochromatin than in euchromatin [331-333].

Although the *Drosophila* studies provided important insights into the structural and functional organization of heterochromatin, the organization of heterochromatin in other insects remains poorly understood. Malaria mosquitoes are an excellent system for studying heterochromatin because they possess well-developed polytene chromosomes with clear morphology. Sequencing of the genome of the major African malaria vector *An. gambiae* [256] provides an opportunity to analyze the molecular structure of the heterochromatin and to study

genomic determinants of heterochromatin formation, maintenance, and function. In malaria mosquitoes, the heterochromatin size and morphology vary significantly among species and within species [334-336], affecting mating behavior and fertility [337,338]. In the *An. gambiae* complex, one of the species, *An. gambiae sensu stricto*, is subdivided into two subtaxa: the M and S molecular forms [339]. These two partially isolated subtaxa predominantly breed within their own form and differ in behavior and environmental adaptations [340]. A DNA microarray analysis revealed that two pericentric regions on X and 2L were the major islands of fixed genomic differentiation between the M and S molecular forms [341]. A more recent microarray study based on the improved AgamP3 assembly and AgamP3.4 gene build provided better estimates for the number and size of diverged pericentric islands between the M and S forms [342]. The study found three islands of genomic divergence: a ~4-Mb region on the X chromosome, a ~2.5-Mb region on the 2L arm, and a 1.7-Mb region on the 3L arm. However, it is not clear if the pericentric islands of genomic divergence are located within heterochromatin or mostly overlap with euchromatin of *An. gambiae*.

According to the CoT analysis, about 86 Mb (33% of 260-Mb genome) of the *An. gambiae* genome corresponds to repetitive elements, which are mostly located in heterochromatic areas of the chromosomes [343]. However, only 3.3 Mb were identified as heterochromatin in the first publication of *An. gambiae* genome [256]. Using cDNA clones for the physical mapping of the heterochromatic scaffolds, an additional 5.3 Mb were mapped to the pericentromeric regions in the chromosomes [344]. Nevertheless, the more precise chromosomal and genomic mapping, as well as detailed analysis of the molecular organization of the *Anopheles* heterochromatin, has yet to be conducted.

In this study, the boundaries of the heterochromatin-euchromatin junctions of all morphologically defined pericentric and intercalary heterochromatin regions were determined for each of the five chromosomal arms of *An. gambiae*. The large regions of intercalary heterochromatin were morphologically different: 0.7-Mb and 0.8-Mb regions of 2L and 3L were diffuse, while a 2.9-Mb region of 3R was a compact heterochromatin. Because the *An. gambiae* genome assembly successfully captured not only the euchromatin, but a significant portion of the heterochromatin, comparative analysis of chromatin types was possible. We provided evidence that heterochromatin and euchromatin differ in gene density and the coverage of retroelements and segmental duplications (SDs). Gene ontology (GO) analysis revealed that heterochromatin is enriched in genes with DNA-binding and regulatory activities. The pericentric heterochromatin had the highest coverage of retroelements and tandem repeats, while intercalary heterochromatin was enriched with SDs. We also demonstrated that the diffuse intercalary heterochromatin binds to HP1 and lamin and has a higher coverage of DNA TEs, minisatellites, and satellites than does the compact intercalary heterochromatin. The investigation of 42-Mb assembly of unmapped genomic scaffolds ("unknown chromosome") demonstrated that it has molecular characteristics similar to cytologically mapped heterochromatin. Finally, the locations and sizes of pericentric heterochromatin regions closely matched the locations and sizes of pericentric islands of genomic divergence between M and S incipient species of *An. gambiae*.

2.4.3 Results and Discussion

Morphological types of the *An. gambiae* heterochromatin

The diploid number of the chromosomes in malaria mosquitoes is six, which includes two pairs of autosomes as well as the X and Y sex chromosomes. The polytene chromosome

complement of a female mosquito has five chromosomal arms: four autosomal arms 2R, 2L, 3R, 3L, and one arm of the X chromosome. In this study, morphological identification of the heterochromatin for the African malaria mosquito *An. gambiae* was performed for the first time. The following criteria were used to distinguish heterochromatic and euchromatic regions in the polytene chromosomes from ovarian nurse cells **(Figure 2.9)**. We considered a region as heterochromatic if it (i) consisted of a compact condensed block or (ii) had a diffuse granulated structure with no banding pattern. These two types of heterochromatin can be distinguished from euchromatic regions, which have a clear banding pattern or puffy non-granulated areas. Pericentric regions of all chromosomes matched these morphological criteria of heterochromatin. The pericentric heterochromatin of the X chromosome has a large diffuse granulated area in region 6, which is similar to the β-heterochromatin of *Drosophila* **(Figure 2.10a)**. The diffuse granulated heterochromatin **(Figure 2.10a)** is morphologically distinct from the euchromatic non-granulated puff in subdivision 9C of the 2R arm **(Figure 2.10b)**. In addition, region 6 of the X chromosome has a dark compact band in the tip of the chromosome **(Figure 2.9)**, which was previously described as a nucleolar organizer region because ribosomal genes were mapped to this area by in situ hybridization [344]. The polytene chromosome 2 has a dark compact proximal heterochromatin surrounded by abundant diffuse heterochromatin in regions 19E-20A **(Figure 2.10c)**. A dark heterochromatic band is also present in region 19D of the 2R arm. The pericentric heterochromatin of chromosome 3 spans subdivisions 37D-38A. Chromosomes 2 and 3 form a diffuse chromocenter via their pericentric heterochromatin [326].

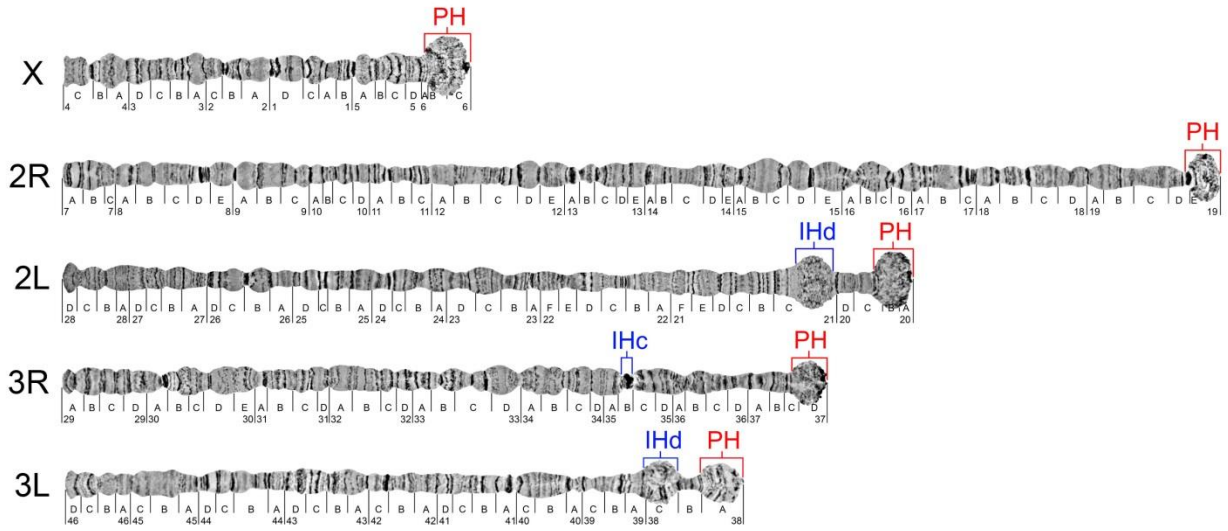


Figure 2.9. The pericentric and intercalary heterochromatin of polytene chromosomes shown on a standard cytogenetic map of *An. gambiae* [345]. PH--pericentric heterochromatin, IHc--compact intercalary heterochromatin, IHd--diffuse intercalary heterochromatin.

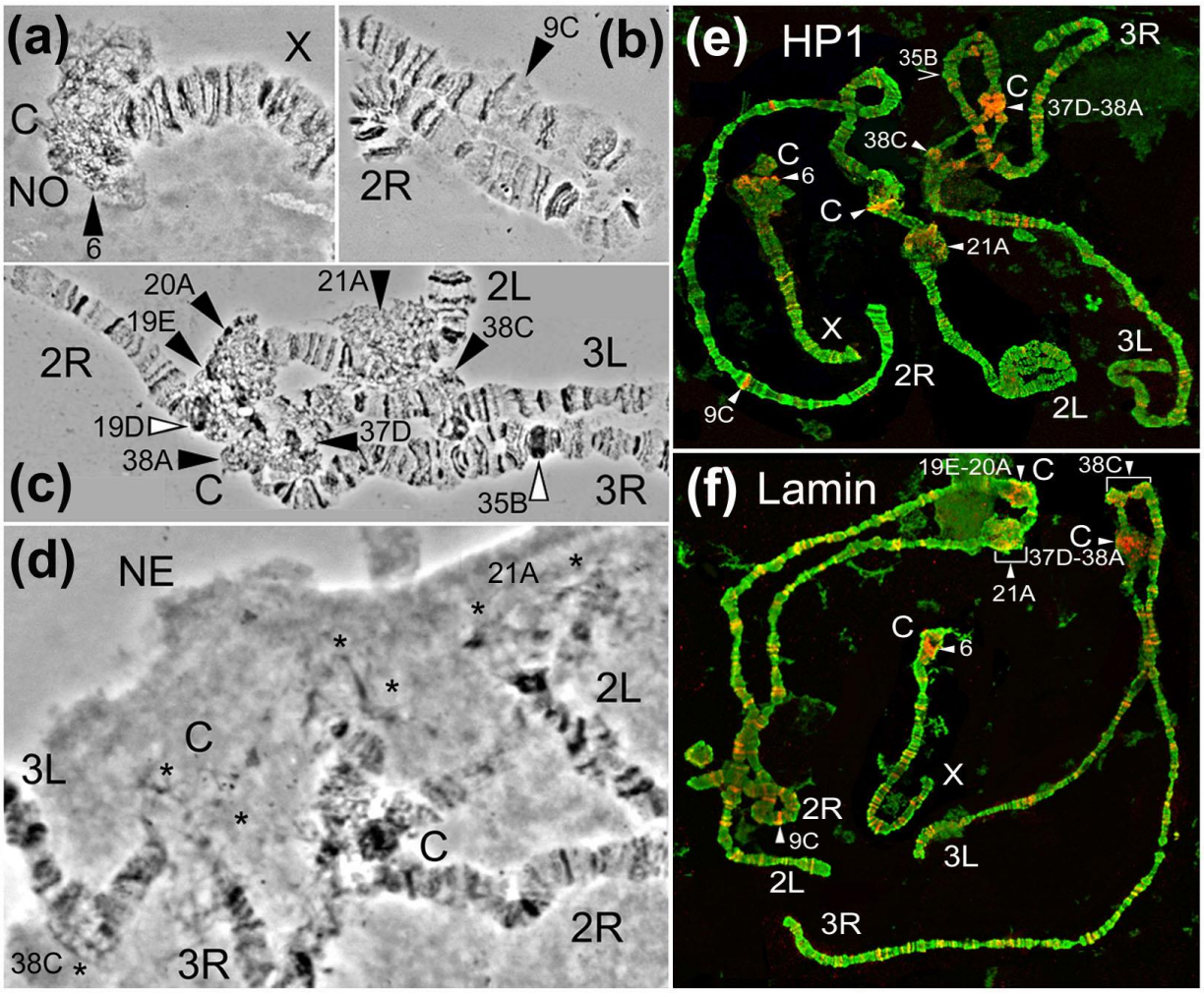


Figure 2.10. Localization of HP1 and lamin Dm0 *Drosophila* antibodies on *An. gambiae* chromosomes. Small numbers and letters indicate subdivisions of the chromosome map. The diffuse type of heterochromatin is shown by black arrowheads (a, b, c). The white arrowheads show compact heterochromatin (c) and sites of HP1 and lamin localization (e, f). Asterisks (d) show attachments of diffuse heterochromatin to the NE. X, 2R, 2L, 3R, 3L - chromosomal arms, C - centromeric areas.

Three regions of intercalary heterochromatin are visible on arms 2L, 3R, and 3L **(Figure 2.10c)**. The subdivision 21A of 2L chromosomal arm forms a large, lightly granulated puff-like structure with no banding pattern. The middle area of subdivision 38C of 3L arm has a similar morphology, but it is slightly smaller and darker. Both regions of intercalary diffuse heterochromatin are located in close proximity to the pericentric regions. The third region of

intercalary heterochromatin is in subdivision 35B of the 3R arm and is located 10 subdivisions away from the centromere. Unlike intercalary heterochromatin of 2L and 3L, this region has a compact dense structure, which is similar to α-heterochromatin of *Drosophila*. In malaria mosquitoes, diffuse and compact types of heterochromatin were previously described in the *Anopheles maculipennis* subgroup [325,346]. Interestingly, the large blocks of compact heterochromatin or the diffuse intercalary heterochromatin regions have not been seen in most species of *Drosophila*. The intercalary heterochromatin in salivary gland nuclei of *D. melanogaster* is strongly underreplicated and has the morphology of "weak" points, which are able to form ectopic contacts [347]. These properties are less prominent in ovarian nurse cell nuclei of the *D. melanogaster otu11* strain where the bands of intercalary heterochromatin are morphologically similar to euchromatic bands [348]. Large blocks of intercalary heterochromatin have been described in polytene chromosomes of *D. immertensis* and species from genera *Chironomus* and *Anopheles* [295,346]. Although the morphology of pericentric heterochromatin is similar in *An. gambiae* and *D. melanogaster*, the presence of two distinct types of intercalary heterchromatin in *An. gambiae* makes this species a unique model system for studying genomic determinants of chromatin morphology.

Chromosomal localization of HP1 and lamin in *An. gambiae*

HP1 is an evolutionarily conserved protein and a good marker of heterochromatic regions [315]. One HP1a ortholog is present in *An. gambiae* (VectorBase gene ID: AGAP009444). The *An. gambiae* protein AGAP009444-PA is 70.4% similar to the *D. melanogaster* HP1a protein in the 206 overlapping amino acids. The antibodies for HP1 were localized in the chromocenter, chromosome 4, telomeric, and some euchromatic regions in *D. melanogaster* [314,349]. In order to

examine the association of HP1 with heterochromatin in *An. gambiae*, we hybridized the primary antibody C1A9 against *D. melanogaster* HP1 to *An. gambiae* polytene chromosomes. This antibody correctly recognized HP1 even in more distantly related species such as the mealybug *Planococcus citri* [350]. Several positively stained loci were invariable, i.e., they were found on every examined chromosome and on every slide. Similar to *Drosophila*, the major invariable sites of HP1 localization were the pericentric regions in *An. gambiae* **(Figure 2.10)**. In addition, diffuse intercalary heterochromatin of regions 21A and 38C were always stained positively for HP1. Only one major invariable HP1-binding site was identified in a large interband of the euchromatic region 9C of 2R arm **(Figure 2.10b)**. All other positive euchromatic sites were variable, and a total of 122 HP1 binding sites were detected on *An. gambiae* chromosomes **(Table 2.4)**. Based on the previous *An. gambiae* genome mapping coordinates, we analyzed the molecular content of the euchromatic site of HP1/lamin binding in region 9C (genome coordinates 12874430-13778780). The analysis found no enrichment of any class of TE. The only heterochromatic molecular feature of this region was a 4.5-kb block of satellite DNA, which consisted of 228-bp units repeated 40 times. Similarly, one major invariable site of HP1 binding was found in euchromatic region 31 of the 2L arm in *D. melanogaster* [86]. However, the molecular analysis of this region found no enrichment in any repetitive DNA. About 200-300 actively expressing loci related to developmentally important and heat-shock genes were positively stained for HP1 in *Drosophila* chromosomes, suggesting a positive role for HP1 in euchromatic gene expression [351-353]. However, only 20 HP1-positive euchromatic sites were invariable among strains, natural populations, and individuals of *D. melanogaster* [86]. Unlike in *Drosophila*, telomeric localization of HP1 was found only on chromosome X in *An. gambiae*, but even this site was variable. Surprisingly, no HP1 binding was detected in the compact

intercalary heterochromatin of subdivision 35B of the 3R chromosome, suggesting that this region has a distinct molecular composition or is strongly underreplicated, and thus, HP1 presence is below the level of detection. Subdivision 35B was morphologically described as heterochromatic based on very dense dark structure **(Figure 2.9 and 2.10c)**. The genomic analysis confirmed its repeat-rich gene-poor heterochromatic nature (see "Difference in molecular content among chromatin types of *An. gambiae*").

Chromosome	Invariable sites	Variable sites	Average number of variable sites per Mb	Chromosome analyzed
X: HP1	1	1-17	0.33	10
X: Lamin	1	1-19	0.41	5
2R: HP1	2	13-53	0.55	3
2R: Lamin	2	7-49	0.46	5
2L: HP1	2	2-20	0.22	4
2L: Lamin	2	17-31	0.49	5
3R: HP1	1	1-18	0.18	3
3R: Lamin	1	29	0.55	2
3L: HP1	2	2-14	0.19	2
3L: Lamin	2	30	0.71	1

Table 2.4. Localization of HP1 and lamin on *An. gambiae* chromosomes

Association of heterochromatin with the NE has been demonstrated in a number of studies [318-323]. Attachment of pericentric regions to the NE in ovarian nurse cell nuclei of *An. gambiae* has also been demonstrated [326]. In our study, the attachments to the nuclear periphery were detected in all pericentric regions, and diffuse intercalary heterochromatin in regions 21A (2L) and 38C (3L) **(Figure 2.10d)**. To test whether heterochromatin binds to the NE, mosquito chromosomes were stained with antibody ADL67.10 against NE protein lamin Dm0 of *D. melanogaster*. We found only one lamin Dm0 ortholog in the *An. gambiae* genome (VectorBase

gene ID: AGAP011938). The *An. gambiae* protein AGAP011938-PA is 78.2% similar to the *D. melanogaster* lamin Dm0 protein in the 628 overlapping amino acids. The antibody against lamin Dm0 successfully hybridized to the *An. gambiae* chromosomes and colocalized with the HP1 antibody in all major invariable sites and in most of the variable sites (Figure 2f). However, the total number of sites was higher for lamin Dm0 (158 sites) than for HP1 (122 sites) **(Table 2.4)**. The major sites for lamin Dm0 were found in the pericentromeric areas, diffuse intercalary heterochromatin regions, and euchromatic interband in region 9C. No lamin Dm0 antibody was detected in region 35B of the 3R chromosome of *An. gambiae*.

Thus, the immunostaining of the antibodies for HP1 and lamin Dm0 has demonstrated that both proteins are primarily associated with the diffuse pericentric and intercalary heterochromatin, but not with the compact intercalary heterochromatin of *An. gambiae*. Two binding motifs, chromo and chromoshadow domains, provide HP1 with the ability to be broadly involved in chromatin and protein binding [354-356]. In vitro studies revealed a direct interaction between HP1 and the lamin B receptor in mammalian cells [323,357,358]. However, in Drosophila, similar direct associations of HP1 with lamin have not been shown, and these proteins have been found associated with different genomic regions [359]. Therefore, despite the colocalization of HP1 and lamin in heterochromatin of *An. gambiae*, the actual protein binding sites in the genome may differ as suggested by the additional regions of lamin binding.

Heterochromatin-euchromatin boundaries in the *An. gambiae* genome

The cytological identification of heterochromatin allowed us to determine the location of heterochromatin-euchromatin boundaries in the *An. gambiae* genome. The approximate coordinates were found based on the genome positions of BAC and cDNA clones, which were

physically mapped to chromosomes near heterochromatin-euchromatin boundaries [256,344]. Because heterochromatic regions were not sufficiently covered with markers, additional PCR-amplified gene fragments were designed and utilized as DNA probes for physical mapping. Fluorescent in situ hybridization (FISH) was used to hybridize multiple PCR products thought to be located near the heterochromatin-euchromatin boundary of each major heterochromatic region of the five chromosome arms **(Table 2.5)**. This allowed for more exacting definition of the boundaries, based on the outermost heterochromatin and euchromatin markers, defining a transition zone with an average size of 78 kb (range: ~15 to 226 kb). Based on these boundaries, a total of ~16.6 Mb was defined as a heterochromatin in the currently mapped genome assembly of *An. gambiae* **(Figure 2.11a)**. The mapped portion of the heterochromatin within defined chromosomes now comprises ~6.4% of the ~260-Mb genome [256,344] and contains 232 (~1.8%) of the ~13,000 total predicted genes. For comparison, no less than 230 genes were annotated in 24 Mb of *D. melanogaster* heterochromatin (release 5.1) [302]. In addition, the sizes of intercalary heterochromatin were also determined. The diffuse heterochromatic regions were 0.7 Mb and 0.8 Mb in 2L and 3L, respectively, and the compact heterochromatin on 3R was 2.9 Mb long. The relatively short sizes of regions of intercalary diffuse heterochromatin as compared to regions of condensed heterochromatin suggest incomplete genome assembly of the diffuse type. However, these sizes exceed the sizes of intercalary heterochromatin known in Drosophila, which range from 100 to 600 kb [307]. The higher repeat content of the mosquito genome may be responsible for the larger sizes of intercalary heterochromatin in *An. gambiae*.

Chromosome Chromatin type	Genome coordinates Start	 End	Mapped markers Start	 End	Size (bp)	Gene number
X						
EU	1	19,928,574	Telomere	AGAP001035	19,928,573	1035
CH	20,009,764	24,393,108	AGAP001039	Centromere	4,383,344	56
2R						
EU	1	58,969,802	Telomere	AGAP004644	58,969,801	3550
CH	58,984,778	61,545,105	26D02	Centromere	2,560,327	32
2L						
CH	1	2,431,617	Centromere	AGAP004707	2,431,616	31
PEU	2,487,770	5,042,389	AGAP004711	AGAP004892	2,554,619	183
IH	5,078,962	5,788,875	AAAB01008948_1	AGAP004905	709,913	12
EU	6,015,228	49,364,325	AGAP004919	Telomere	43,349,097	2812
3R						
EU	1	38,815,826	Telomere	AGAP009690	38,815,825	1960
IH	38,988,757	41,860,198	AGAP009696	AGAP009730	2,871,441	35
EU	41,888,356	52,131,026	BAC 30P16	BAC 25H11	10,242,670	554
CH	52,161,877	53,200,684	AGAP010287	Centromere	1,038,807	23
3L						
CH	1	1,815,119	Centromere	AGAP010342	1,815,118	33
PEU	1,896,830	4,235,209	AGAP010344	AGAP010481	2,338,379	138
IH	4,264,713	5,031,692	AGAP010482	AGAP010491	766,979	10
EU	5,133,257	41,963,435	AGAP010505	Telomere	36,830,178	1923

Table 2.5. Boundaries between heterochromatin and euchromatin in the *An. gambiae* genome.

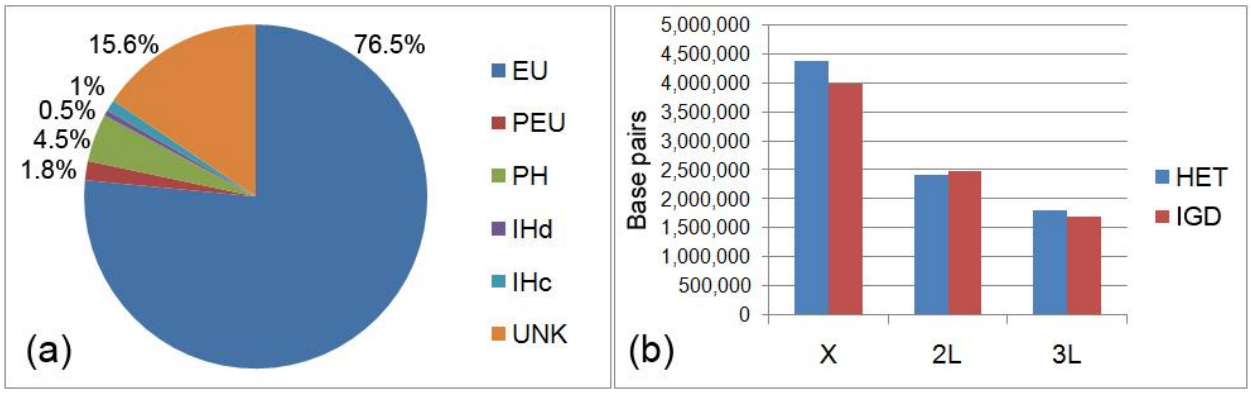


Figure 2.11. Schematic representation of the heterochromatin amount in the *An. gambiae* genome. (a) Relative proportions of mapped chromatin types and unmapped sequences in the assembly. PH--pericentric heterochromatin, IHc--compact intercalary heterochromatin, IHd--diffuse intercalary heterochromatin, PEU--proximal euchromatin, EU--euchromatin, UNK--"unknown chromosome." (b) Comparison of sizes and positions of islands of genomic divergence (IGD) and regions of pericentric heterochromatin (HET) in the X chromosome, the 2L arm, and 3L arm. Position of a putative centromere corresponds to 0 bp.

Heterochromatin and pericentric regions of genomic divergence in incipient species

Three pericentric islands of genomic divergence were found in chromosomes X, 2L, and 3L in two partially isolated subtaxa - the M and S molecular forms of *An. gambiae* s.s. [342]. Our analysis showed that the positions of islands of genomic divergence mostly correspond to the positions of physically mapped regions of pericentric heterochromatin **(Figure 2.11b)**. The sizes of the pericentric heterochromatin were the following: 4.4 Mb of the X chromosome, 2.4 Mb of the 2L arm, and 1.8 Mb of the 3L arm. Thus, the overlaps with islands of genomic divergence are 91% in the X chromosome, 97% in the 2L arm, and 94% in the 3L arm. This observation suggests that heterochromatic sequences diverge rapidly during speciation of malaria mosquitoes. Earlier cytological studies showed the presence of significant intra- and interspecific differences in amount and location of heterochromatin in the *An. gambiae* complex [334,337]. A genome-wide microsatellite study of members of the *An. gambiae* complex has determined a high level of genetic introgression among species [360]. However, the *An. gambiae* microsatellites

at six loci of X, 3L, and 3R could not be amplified in all sibling species, indicating significant sequence divergence from the major malaria vector. These loci were identified as heterochromatic in our study. Fast changes in heterochromatic DNA can be accompanied by the rapid evolution of heterochromatic proteins. Although HP1 is an evolutionarily conserved protein, other heterochromatin- and centromere-associated proteins demonstrate rapid adaptive evolution [361,362]. For example, an LHR protein encoded by lhr (Lethal hybrid rescue) colocalizes with HP1 in heterochromatic regions and has diverged extensively in sequence between *D. melanogaster* and *D. simulans* species in a manner consistent with positive selection. Interestingly, F1 hybrids between these species demonstrate altered chromatin structure, probably attributable to the effects of species-specific differences in TEs and other repetitive DNAs [363], suggesting a role for heterochromatin in speciation.

Overrepresentation of gene ontology terms in the *An. gambiae* heterochromatin

To characterize gene content of the *An. gambiae* heterochromatin, we utilized GO terms [364]. The frequencies of GO terms assigned to genes in heterochromatin were compared to frequencies for all GO-annotated genes in the peptide dataset of *An. gambiae* **(Figure 2.12a)**. After Bonferroni correction for multiple tests, this analysis revealed significant enrichment for molecular functions in heterochromatin, including DNA binding (12 genes) and sequence-specific DNA binding (12 genes). Protein products of 29 heterochromatic genes constitute membrane, representing a significant enrichment of the GO cellular location. Finally, heterochromatin had overrepresentation of several gene types, including those encoding for proteins involved in biological regulation (24 genes) and regulation of metabolic processes (17 genes) (biological processes). The GO analysis of the "unknown chromosome" (sequence

assembly lacking chromosomal assignment) identified enrichment in a number of interesting genes **(Figure 2.12b)**. We found that genes residing in the "unknown chromosome" had significant overrepresentation of GO terms in biological processes, including chromosome organization (15 genes), DNA packaging (15 genes), and nucleosome assembly (15 genes). Transcription initiation factor activity (four genes) was among several molecular functions overrepresented in the genes within the "unknown chromosome." Analysis of the heterochromatic portion of the Drosophila genome revealed the overrepresentation of similar GO terms [302]. These studies suggest that heterochromatin of insects may accumulate genes important for its own establishing, maintaining, or modifying chromatin structure.

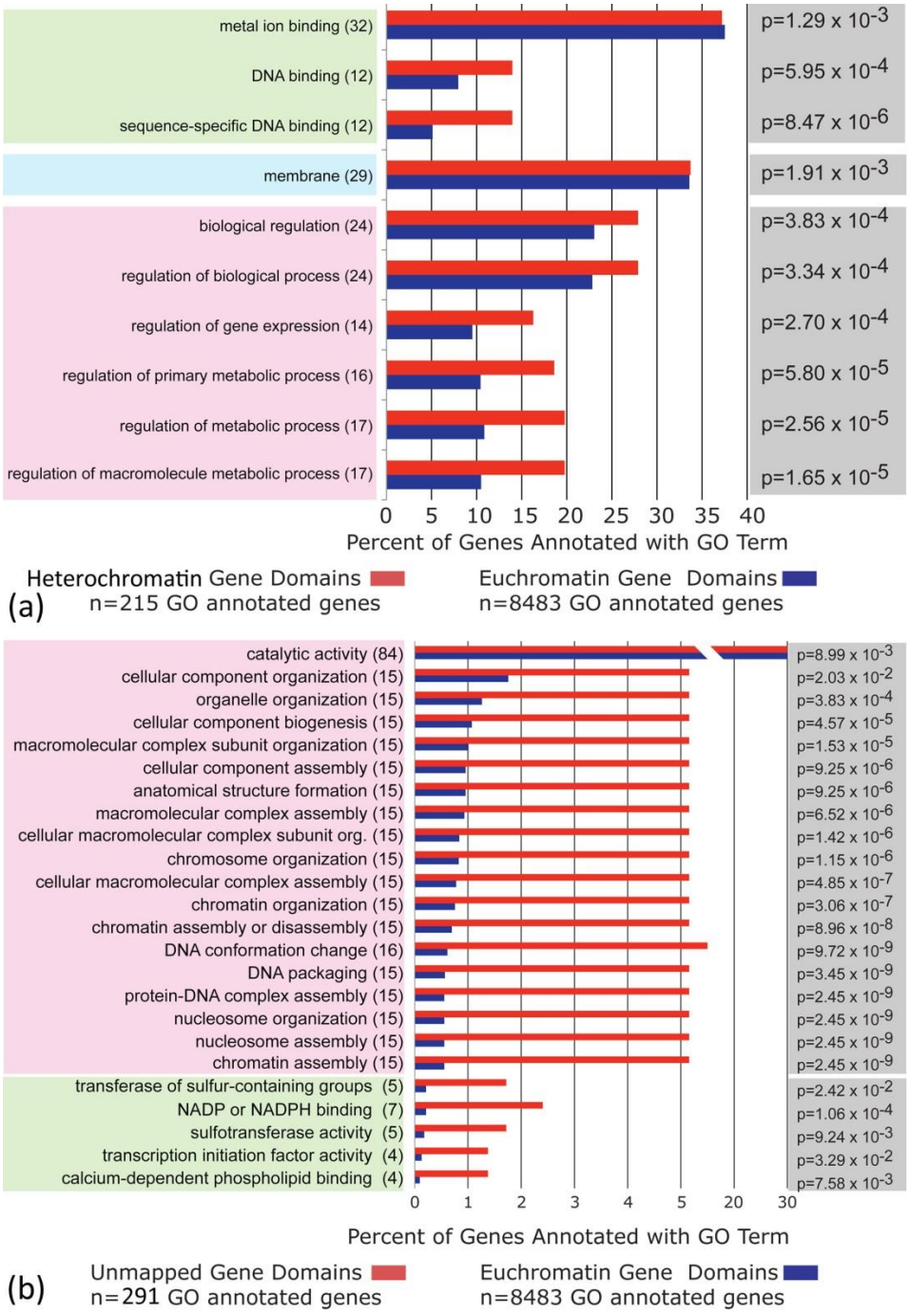


Figure 2.12. Overrepresented GO terms in genes within the cytologically confirmed heterochromatin (a) and within "unknown chromosome" (b) of *An. gambiae*. The percentages of heterochromatic (red) and euchromatic (blue) genes containing the listed GO biological process (pink shading), cellular location (blue shading), and molecular function (green shading) terms are indicated. Numbers in parentheses refer to the actual number of heterochromatin or unmapped genes annotated with the listed GO domain. GO-Term-Finder, Bonferroni corrected p-value scores are shown to the right (grey shading).

Difference in molecular content among chromatin types of *An. gambiae*

Using Bayesian statistical model and procedure for discerning differences between chromatin types, eight molecular features were analyzed: genes, DNA-mediated TEs (DNA TEs), RNA-mediated TEs (RNA TEs), SDs, micro- and minisatellites, satellites, and MARs. These molecular features were compared among five distinct chromatin types: 1) pericentric heterochromatin of all chromosomes; 2) diffuse intercalary heterochromatin in regions 21A of 2L and 38C of 3L; 3) compact intercalary heterochromatin, region 35B of 3R; 4) proximal euchromatin, located between pericentric and diffuse intercalary heterochromatin, includes subdivisions 20CD of 2L and 38B of 3L; and 5) euchromatin in all remaining regions in the chromosomes. For this analysis, the data that distinguishes both the counts and the overall base-pair coverage were incorporated for each molecular feature into the genomic windows of each of the five chromatin types. Dominant model selection procedures gave us the ability to compare all possible competing models and to select between parsimonious models by maximizing the posterior distribution.

Heterochromatin had a uniformly low concentration of genes. On average, the gene density was 4.7 times lower in the heterochromatin than in the euchromatin **(Table 2.6)**. Our analysis showed that heterochromatin significantly exceeds euchromatin in the coverage of RNA TEs and SDs. RNA TEs were the most abundant features in the mosquito genome **(Figure 2.13)**. The pericentric heterochromatin had the highest coverage of RNA TEs, microsatellites, minisatellites, and satellites. The intercalary heterochromatin had a higher coverage of SDs than all other chromatin types. The diffuse intercalary heterochromatin had a higher coverage of TEs, minisatellites, and satellites than did the compact intercalary heterochromatin. The enrichment of TEs in the pericentric heterochromatin and diffuse intercalary heterochromatin as compared to

the compact intercalary heterochromatin can explain the pattern of HP1 localization in polytene chromosomes of *An. gambiae*. Pericentric and diffuse intercalary heterochromatin, but not the compact type, was HP1 positive. Similarly, the fourth chromosomes of *D. melanogaster* and *D. pseudoobscura* bound to HP1, while the fourth chromosome of *D. virilis* did not. The density of TEs in this chromosome was significantly higher for *D. melanogaster* than for *D. virilis* [315-317]. The proximal euchromatin had a higher coverage of DNA TEs, MARs, and SDs but a lower coverage of satellites than the rest of the euchromatin. These differences can probably be explained by the close distance of the proximal euchromatin to the centromere.

DNA transposons					
	Euchromatin	Pericentric heterochromatin	Compact intercalary heterochromatin	Proximal euchromatin	Diffuse intercalary heterochromatin
Mean coverage	1.558	9.374	3.235	4.297	9.720
Median coverage	1.340	9.325	3.100	3.100	10.840
Mean ranks	1.269	4.277	2.162	2.728	4.564
RNA transposons					
	Euchromatin	Pericentric heterochromatin	Compact intercalary heterochromatin	Proximal euchromatin	Diffuse intercalary heterochromatin
Mean coverage	3.269	23.774	11.975	6.488	17.628
Median coverage	2.400	22.185	13.430	4.895	17.325
Mean ranks	1.194	4.766	3.102	2.103	3.835
Segmental duplications					
	Euchromatin	Pericentric heterochromatin	Compact intercalary heterochromatin	Proximal euchromatin	Diffuse intercalary heterochromatin
Mean coverage	6.095	9.837	12.332	7.795	14.930
Median coverage	4.870	9.214	11.575	7.935	14.775
Mean ranks	2.149	3.196	3.274	2.611	3.770

Microsatellites					
	Euchromatin	Pericentric heterochromatin	Compact intercalary heterochromatin	Proximal euchromatin	Diffuse intercalary heterochromatin
Mean coverage	0.441	0.693	0.452	0.433	0.370
Median coverage	0.370	0.685	0.410	0.465	0.330
Mean ranks	2.453	4.560	3.149	2.727	2.111
Minisatellites					
	Euchromatin	Pericentric heterochromatin	Compact intercalary heterochromatin	Proximal euchromatin	Diffuse intercalary heterochromatin
Mean coverage	1.022	1.687	0.912	0.918	1.400
Median coverage	0.720	1.715	0.870	0.830	1.370
Mean ranks	2.397	4.259	2.400	2.212	3.732
Satellites					
	Euchromatin	Pericentric heterochromatin	Compact intercalary heterochromatin	Proximal euchromatin	Diffuse intercalary heterochromatin
Mean coverage	0.296	0.573	0.088	0.147	0.405
Median coverage	0.230	0.580	0.090	0.075	0.200
Mean ranks	3.175	4.548	1.956	2.185	3.136
MARs					
	Euchromatin	Pericentric heterochromatin	Compact intercalary heterochromatin	Proximal euchromatin	Diffuse intercalary heterochromatin
Mean coverage	3.393	6.373	5.850	7.487	6.320
Median coverage	3.400	5.895	5.900	7.075	6.540
Mean ranks	1.508	3.119	2.788	4.165	3.420
Genes					
	Euchromatin	Pericentric heterochromatin	Compact intercalary heterochromatin	Proximal euchromatin	Diffuse intercalary heterochromatin
Mean coverage	5.684	1.440	1.287	6.400	1.455
Median coverage	5.780	1.400	0.820	6.230	1.580
Mean ranks	4.306	2.230	1.734	4.689	2.041

Table 2.6. Coverage (%) of molecular elements in chromatin types of *An. gambiae*.

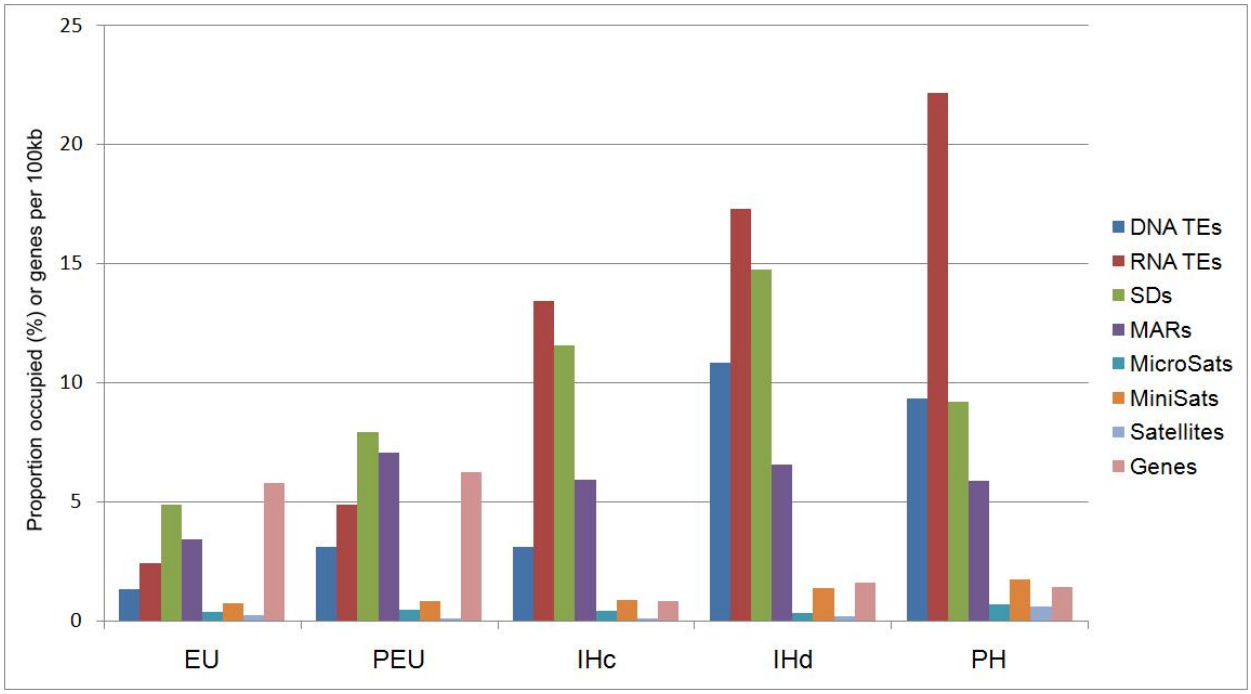


Figure 2.13. Median values of gene density and repetitive element coverage in chromatin types of *An. gambiae*. Percentage of region length occupied per 1 Mb are indicated for all repetitive elements. PH--pericentric heterochromatin, IHc--compact intercalary heterochromatin, IHd--diffuse intercalary heterochromatin, PEU--proximal euchromatin, EU--euchromatin.

Chromatin types and genome landscape in *An. gambiae*

In addition to the overall differences among chromatin types, the distribution of molecular features within chromosomal arms was analyzed. A high density of genes was seen outside of the heterochromatin boundaries believed to be euchromatin, followed by a transition zone and a heterochromatic region with a low gene density. The distribution of TEs densities had the opposite pattern. The highest coverage of SDs was detected in intercalary heterochromatin

with peaks in some euchromatic regions of the 2R, 3R, and 3L arms **(Figure 2.14)**. MARs were found concentrated in the pericentric heterochromatic and proximal euchromatic regions of all arms, but they were also abundant in distal euchromatic regions of the 2L, 3R, and 3L arms. We observed the high coverage of predicted MARs in heterochromatic regions, which are associated with the NE [331-333]. Moreover, the increase in MAR coverage seen in euchromatic regions of the 2L, 3R, and 3L arms correlated positively with the higher density of lamin-positive sites in these arms detected by immunostaining **(Table 2.4)**. The highest coverage of MARs was found in proximal euchromatin, which was not stained by the lamin antibody. Also, the two types of heterochromatin were not significantly different in MAR coverage. However, the lamin-positive pericentric and diffuse intercalary heterochromatic regions were significantly enriched with TEs. The coverage of DNA TEs was about two times higher in pericentric and diffuse intercalary heterochromatin than in other chromatin types **(Table 2.6)**.

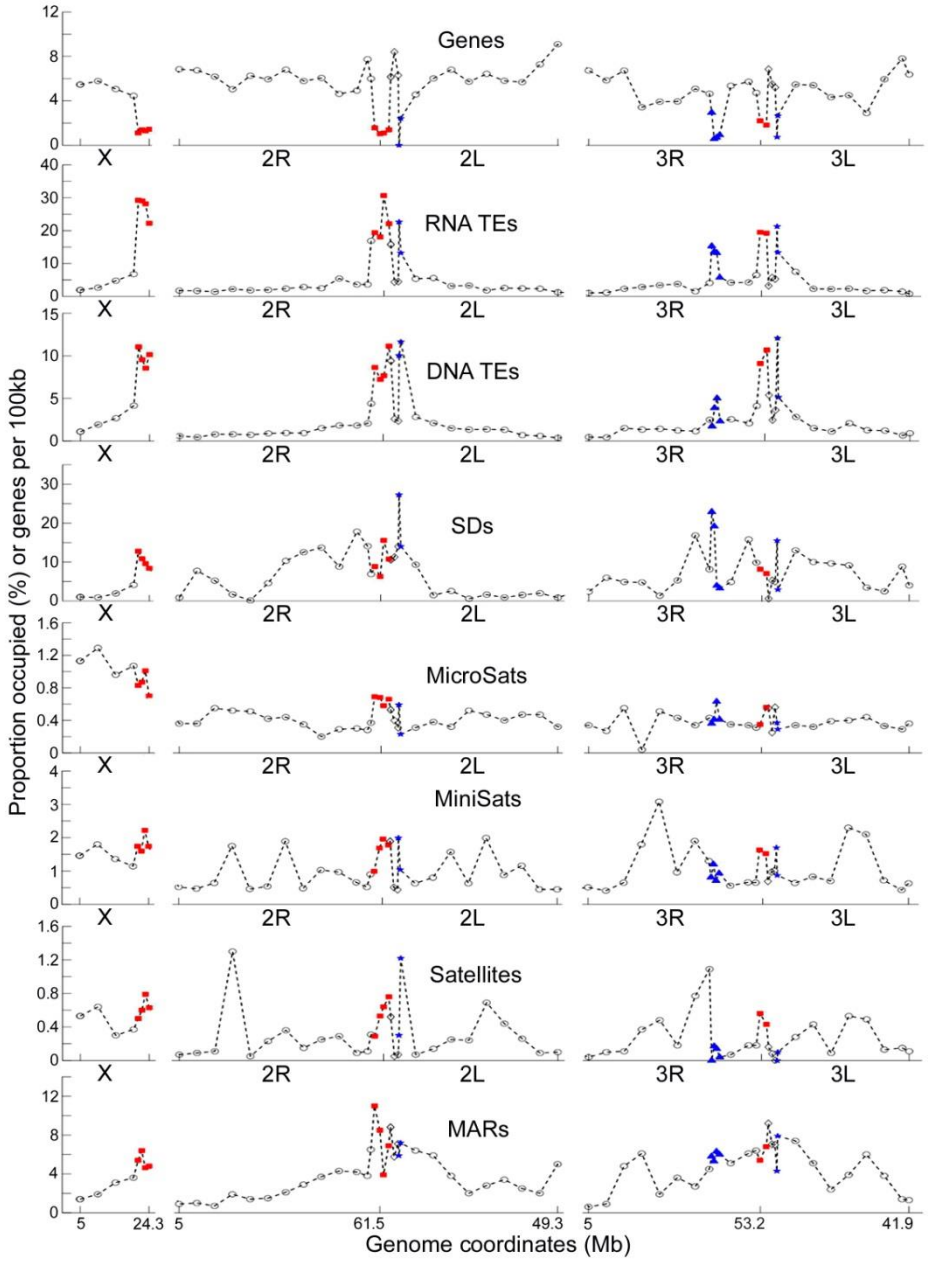


Figure 2.14. Genome landscapes of the *An. gambiae* heterochromatin and euchromatin. Median values of coverage of molecular features are displayed as 5-Mb intervals in euchromatin (open circles) and < 1-Mb intervals in heterochromatin. Red squares--pericentric heterochromatin, open diamonds--proximal euchromatin, blue stars--diffuse intercalary heterochromatin, blue triangles--compact intercalary heterochromatin.

Overall, this analysis confirmed morphological predictions of heterochromatin. All types of heterochromatin in the *An. gambiae* genome had typical heterochromatic molecular features: low gene density and high coverage of TEs and SDs. However, because TEs are significantly underannotated in the *An. gambiae* genome, meaningful comparisons of the TE content of heterochromatin between mosquito and fruit fly are difficult.

Unmapped genome assembly of *An. gambiae*

The unmapped portion of the AgamP3 *An. gambiae* genome assembly comprises 42 Mb [365], i.e.~16% of the genome, and has 491 protein coding genes **(Figure 2.11A)**. The analysis of the genomic content of this "unknown chromosome" (http://www.vectorbase.org/ webcite) revealed that the density of genes and the coverage of TEs and microsatellites were similar to that of the heterochromatin **(Figure 2.15)**. The highest coverage of minisatellites and satellites was detected in the "unknown chromosome" suggesting that the majority of these scaffolds belong to heterochromatin. Two satellites, AgY477 and Ag53C, were mapped to the most proximal heterochromatin of the *An. gambiae* polytene chromosomes [366]. The location of satellite DNA in the proximal pericentric heterochromatin has also been demonstrated in *An. stephensi* [367]. An enrichment with highly repetitive DNA has been found in the compact heterochromatin of the *An. macullipennis* subgroup [346]. Telomeres of the *An. gambiae* chromosomes do not display heterochromatic morphology. Subtelomeric regions possess typical euchromatic banding patterns. However, molecular analysis of the telomeric end of the 2L arm demonstrated the presence of satellites and minisatellites [368,369]. Therefore, the unmapped portion of the *An. gambiae* genome assembly likely contains sequences from the most proximal pericentric, most distal telomeric ends of chromosomes, and intercalary diffuse heterochromatin. In *D.*

melanogaster, 10 Mb of the unmapped portion of the genome was also enriched in tandem repeats and satellites [302].

Bayesian statistical analysis of molecular features in the chromatin types

We have developed a model and procedure for discerning differences in molecular features between chromatin types. For this analysis, we incorporated data which distinguishes both the counts for each molecular feature and the overall coverage of each feature in subdivided regions of each of the five chromatin types of interest $\xi_j \in A$ = {EU, PH, IHc, PEU, IHd}, where PH–pericentric heterochromatin, IHc–compact intercalary heterochromatin, IHd–diffuse intercalary heterochromatin, PEU– proximal euchromatin, EU–euchromatin. Since each region of the genome where these chromatin types are located is closely independent of each other, the likelihood follows as:

$$\prod_{\xi_j \in A} \prod_{i \in \xi_j} \Pr\left(C_i, \xi_j \middle| \text{Data}, \Theta\right),$$

where, C_i, ξ_j are the counts associated with arm ξ_j for chromatin type *i* and Θ are the unknown model parameters that must be estimated.

For our application, we used a Poisson random effects model for explaining the counts, but included information about the coverage in each region as well. To make this connection, we parameterized the mean effect λ_i, ξ_j, through the log-link function as:

$$\log \lambda_i, \xi_j = \mu_{\xi_i} + \beta_{\xi_i} \log(L_i) + \zeta_{\xi_i} \log K_i$$

where L_i is the total length and K_i is the coverage length for chromatin type i. For each chromatin type β_{ξ_i} and ζ_{ξ_i} are random effects relating to the effect each length has on distinguishing the number of the molecular feature. μ_{ξ_i} relates to the overall density of the counts for each chromatin region. Hence in our case, the model unknowns are

$$\Theta = \{\mu_{\xi_i}, \beta_{\xi_i}, \zeta_{\xi_i}\} \text{ for each } \xi_i \in A = \{\text{EU, PH, IH c, PEU, IH d}\}$$

Our ultimate goal was to determine if random effects $\{\mu_{\xi_i}, \beta_{\xi_i}, \zeta_{\xi_i}\}$ can be statistically distinguished between chromatin types. Dominant model selection procedures have the ability to compare all possible competing models and also to compensate for the number of parameters involved in each model. That is, if model fit is the objective, then all procedures will determine optimality by utilizing as many parameters as is possible. In our case, these could correspond to 125 possible parameter configurations. Since models selected this way are generally suboptimal in terms of prediction, likelihood penalization schemes are common practice. For instance, the Bayesian information criterion (BIC) and the Akaike information criterion (AIC) are commonly used devices for selecting between models. We relied on the former, since this criterion closely aligns with Bayes Factor computation. Explicitly, BIC, under model Mk, is computed as:

$$-BIC_{M_k} = 2\log(p\left(\text{Data}\middle|\widehat{M}\right)\Big) - \log(N)p$$

where $(p\left(\text{Data}\middle|\widehat{M}\right)\Big)$ is the maximum likelihood estimate (MLE), under model *k*, *N* is the number of observations, and *p* is the number of parameters in model *k*. Bayes factors select between models through the ratio

$$BF = \frac{\int p(\text{Data}|\Theta, M_k)p(\Theta|M_k)d\Theta}{\int p(\text{Data}|\Theta, M_{\tilde{k}})p(\Theta|M_{\tilde{k}})d\Theta}$$

which can be interpreted as the level of support model M_k has in favor of the data over model $M_{\tilde{k}}$. As an approximation, we have

$$e^{-\frac{1}{2}\Delta BIC} = e^{-\frac{1}{2}\left(BIC_{M_k} - BIC_{M_{\tilde{k}}}\right)} \approx BF$$

which is a measure that decides between models and accounts for high degrees of observational variation. In order to compute the MLEs used in BIC calculations, we relied on an annealing algorithm. Specifically, given multiple locations in model space, state values in Θ and model configurations are simultaneously maximized to provide the MLE estimates for each data set. This procedure was repeated 1,000,000 times to ensure global optimization was achieved and that the best models (MAX models) were selected. The MAX models for each feature are given below.

RNA TEs: MAX model is (PEU = EU)–BIC = -2471.87, (PEU = EU, IHc = IHd)–BIC = -2474.12, and all different–(-BIC = -2475.64). So, PEU = EU has strong support (MAX model) over the model with all distinguishing chromatin types ΔBIC = 3.77, and ΔBIC = 1.52 for distinguishing models with all distinct from (PEU = EU, IHc = IHd), which is moderate support that euchromatin and intercalary heterochromatin types can be considered the same for retro-elements. All other models have negligible support (ΔBIC > 10).

DNA TEs: MAX model is (IHc = PEU)–BIC = -1032.5, (IHc = PEU, IHd = PH)–BIC = -1034.0, so there is support for (IHc = PEU, IHd = PH) ΔBIC = 1.5. All other hypotheses ΔBIC > 9.

SDs: MAX model is (IHc = IHd)–BIC = -1540.2, all other hypotheses have ΔBIC > 8.

MARs: MAX model is (PH = IHc)–BIC = -1305.21, (PH = IHc = IHd)–BIC = -1306.44, (PH = IHc = IHd = PEU)–BIC = -1309.39. So, distinguishing IHd from (PH = IHc) has support ΔBIC = 1.23, which is mild. PEU is sufficiently different from each of the other candidate hypotheses, so we deem (PH = IHc = IHd). Differentiation from the all distinguishable model has ΔBIC > 10.

Genes: MAX model is (EU = PEU, PH = IH c = IHd)–BIC 468.96, ΔBIC > 10 for all non-nested hypotheses.

Microsatellites: MAX model is (IHc = PEU)–BIC = -1408.22, (IHc = PEU = IHd)–BIC = -1407.98, ΔBIC =1.76. Supported hypothesis is (IHc = PEU = IHd).

Minisatellites: MAX model is (PEU = IHc)–BIC = -1887.03, (EU = PEU = IHc)–BIC = -1890.15 ΔBIC = 3.12, so supported hypothesis is EU = PEU = IHc, and less parsimoniously PEU = IHc. All other hypotheses have ΔBIC > 10.

Satellites: MAX model is (IHc = PEU)–BIC = -656.78, all other hypotheses have ΔBIC > 10.

List of abbreviations

AIC: Akaike information criterion; BIC: Bayesian information criterion; BSA: bovine serum albumin; DABCO:

1,4-diazabicyclo[2.2.2]octane; EU: euchromatin; FISH: fluorescent in situ hybridization; GO: gene ontology;

HP1: heterochromatin protein 1; IHc: compact intercalary heterochromatin; IHd: diffuse intercalary heterochromatin; lhr: lethal hybrid rescue; MAR: matrix attachment region; MLE: maximum likelihood estimate; MR4: Malaria Research and Reference Reagent Resource Center; NE: nuclear envelope; PBS: phosphate buffered saline; PH: pericentric heterochromatin; PEU: proximal euchromatin; SD: segmental duplication; SSC: saline-sodium citrate; TE: transposable element.

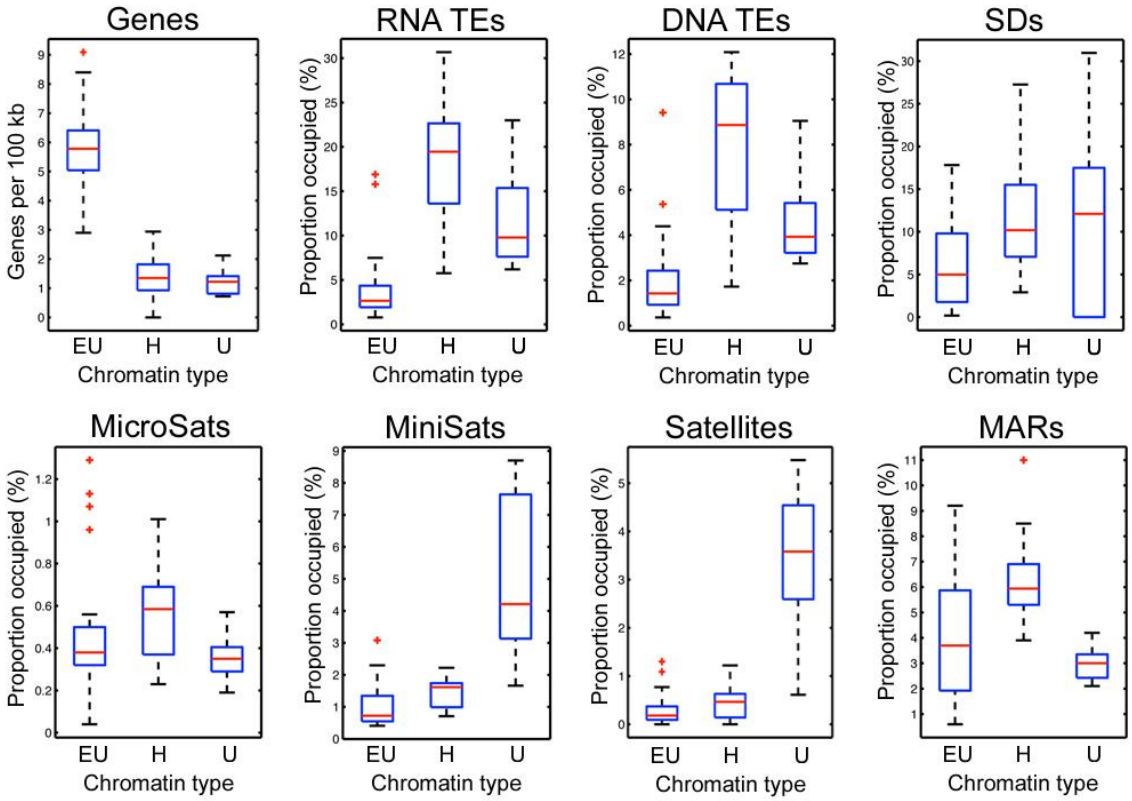


Figure 2.15. Median values of gene density and repetitive element coverage in "unknown chromosome" of *An. gambiae*. EU--total euchromatin, H--total heterochromatin, U--"unknown chromosome."

Conclusions

Morphological identification and detailed physical mapping allowed us to define an expanded compartment of recognizable heterochromatin with distinct molecular features within the *An. gambiae* genome assembly. Now about 16.6 Mb of mapped heterochromatin with 232 protein-coding genes is available for further characterization. GO analysis revealed that heterochromatin is enriched in genes that encode for proteins that may be involved in epigenetic regulation of chromatin. This study described the large regions of intercalary heterochromatin with a morphology not seen in *D. melanogaster*. We also provided evidence that heterochromatin and euchromatin significantly differ in gene density and the coverage of RNA TEs and SDs. The sequence composition, in terms of DNA TEs, RNA TEs, minisatellites, and satellites, can differentiate between the diffuse and compact types of intercalary heterochromatin. Conversely, MARs are distributed regardless of the chromatin type. The results of immunostaining with HP1 and lamin confirmed the general principle of nuclear organization--that the gene-poor regions of the genome reside at the nuclear periphery. Future investigations of *An. gambiae* heterochromatin need to show whether specific molecular composition can actually lead to chromosome-NE interactions. Given that the 42-Mb-long "unknown chromosome" has the molecular characteristics of heterochromatin, it is possible that only one third of heterochromatic sequences in the *An. gambiae* genome assembly have been placed to chromosomes. Finally, we found that pericentric islands of genomic divergence between M and S incipient species of *An. gambiae* are almost completely heterochromatic, demonstrating the elevated evolutionary plasticity of the mosquito heterochromatin.

2.4.4 Methods

Mosquito strain and chromosome preparation

A laboratory SUA strain of *An. gambiae* was used in this study. Mosquitoes were reared at 28°C at 80% humidity. Mosquitoes were grown at a low density (500-750 mosquitoes per 4 liter pan) to obtain better quality chromosomes. Larvae were fed ad libitum. Adults were given sugar water through dampened cotton balls that were removed at least 2 hours preblood feeding to ensure that most mosquitoes would take a blood meal. To obtain the chromosomal preparations, females were blood fed twice with a Guinea pig. Chromosomal slides for the morphological analysis were prepared as described previously >Biessmann, 1998 #4769<. Images were recorded with an Olympus Q-color5 digital cooled 5 megapixel camera and the Olympus CX41 light microscope using 1000× magnification (Olympus America Inc., Melville, NY, USA).

Probe preparation and FISH

Genomic DNA from *An. gambiae* mosquitoes was isolated via a DNeasy Blood and Tissue Kit (Qiagen Inc., Valencia, CA, USA). PCR probes were chosen from the euchromatin--heterochromatin transition zones of the *An. gambiae* genome. Many of these probes were based on genes located near expected heterochromatin-euchromatin boundaries on each chromosome arm. Primers were designed using the Primer3 program [370]. PCR products ranged from 400-600 bp in size. The in situ hybridization procedure was done as previously described [371]. PCR products were gel purified using the Geneclean kit (Qbiogene, Inc., Irvine, CA). The DNA was labeled with Cy3-AP3-dUTP (GE Healthcare UK Ltd., Buckinghamshire, England) using the Random Primer DNA Labeling System (Invitrogen Corporation, Carlsbad, CA, USA). DNA

probes were hybridized to the chromosomes at 39°C overnight in hybridization solution (Invitrogen Corporation, Carlsbad, CA, USA). Then the chromosomes were washed in 0.2 × SSC, (Saline-Sodium Citrate: 0.03 M Sodium Chloride, 0.003 M Sodium Citrate) counterstained with YOYO-1, and mounted in DABCO. Fluorescent signals were detected and recorded using a Zeiss LSM 510 Laser Scanning Microscope (Carl Zeiss MicroImaging Inc., Thornwood, NY, USA).

HP1 and lamin antibodies immunolocalization

The original method of chromosome immunostaining was slightly modified for application to ovarian nurse cell polytene chromosomes [353,372]. In order to obtain polytene chromosomes from ovarian nurse cells, we blood fed female mosquitoes and kept them at regular conditions (temperature 26°C, humidity 80%) over night for 25 hours. Then half gravid females were placed on ice, and their ovaries were dissected. Every ovary was divided into two parts; each part was placed in fixative solution (47% water, 45% acetic acid, and 8% formaldehyde) separately; and follicles were spread on the slide by needles. Afterwards, the fixative solution was removed by filter paper, and follicles were placed in a fresh drop of the solution. Follicles were squashed under a cover slip and frozen in liquid nitrogen. Then cover slips were removed, and slides were kept in 70% cold ethanol at -20°C for several hours. Just before immunohybridization, slides were washed in PBS saline buffer (Boston Bioproduct, Worcester, MA, USA) with 0.1% Nonidet P40 and incubated for 20 minutes in blocking solution (1% BSA in PBS).

Primary mouse monoclonal antibodies C1A9 for Heterochromatin Protein 1 of *D. melanogaster* and ADL67.10 for Drosophila lamin Dm0 (Developmental Studies Hybridoma

Bank, The University of Iowa, USA) were used for immunostaining of *An. gambiae* polytene chromosomes. Primary antibodies were diluted in 1:50 ratio and incubated overnight with the chromosomes in a humid chamber at 4°C. Secondary goat antibodies to mouse were Cy3 labeled (KPL, Guildford, UK) and diluted in 1:200 ratio. Slides were incubated with secondary antibodies for 40 minutes at room temperature. Chromosomes were counterstained with YOYO-1 (Invitrogen, Way Carlsbad, CA 92008 USA) and mounted in DABCO antifade solution (0.233 g DABCO, 800 µl H2O, 200 µl 1 M trisHCl pH 8.0, 9 ml glycerol). Slides were examined using a Zeiss LSM 510 Laser Scanning Microscope (Carl Zeiss MicroImaging Inc., Thornwood, NY, USA).

GO annotation of heterochromatin and unmapped genome assembly

The *An. gambiae* AgamP4 annotated peptide set was analyzed using a locally installed copy of Interproscan 4.4.1 [373]. A GO [364] annotation file was generated using Interproscan-assigned GO terms and custom Perl scripts. Go-Term-Finder [374] version 0.86 was used to search for significantly overrepresented (i.e., $p < 0.05$) GO terms assigned to genes in heterochromatin relative to frequencies for all GO-annotated genes in the peptide dataset. All scores reported have been Bonferroni corrected to account for multiple comparisons. Genes within the euchromatin--heterochromatin transition zones were considered euchromatic for this analysis. Bar graphs were generated with Microsoft Excel and labeled using Adobe Illustrator CS4.

Gene and repetitive element databases

Counts and length of coverage of all molecular features were identified in 5-Mb intervals in euchromatin and < 1-Mb intervals in heterochromatin of the *An. gambiae* AgamP3 genome

assembly [365]. Gene density and TE coverage were analyzed using the Biomart [375] and RepeatMasker [376] (http://www.repeatmasker.org/webcite) programs, respectively. Micro- and minisatellites were analyzed by Tandem Repeats Finder [377]. Only tandem repeats with 80% matches and a copy number of 2 or more (8 or more for microsatellites) were included in the analysis. Microsatellites, minisatellites, and satellites had period sizes ranging from 2 to 6, from 7 to 99, and from 100 or more, respectively. SDs were detected using BLAST-based whole-genome assembly comparison [378] limited to putative SDs represented by pair-wise alignments with ≥2.5-kb and >90% sequence identity. The alignment length was specifically chosen to avoid the vast majority of incompletely masked repetitive elements. SD counts are not discrete duplication events, but indicate the number of regions that have been involved in duplications within a given interval. Putative MARs in the *An. gambiae* genome sequence were predicted using the SMARTest bioinformatic tool [379].

2.5. A molecular feature based model for the prediction of chromatin types in *An. gambiae*

2.5.1. Abstract

The chromatin of eukaryotic organisms persists in different states of compaction that provide different environments responsible for regulation of gene expression. These states exhibit different hallmark features including abundance of repetitive elements and genes, different epigenetic markers, and various chromosomally bound protein interactions. Here, we have identified molecular features that can successfully be used to characterize and differentiate heterochromatin from euchromatin. We present a tool that can predict the likelihood that a segment of DNA is a chromatin type based on these molecular features associated with that

region. As well, we outline an updated landscape of the *Anopheles gambiae* genome based on predictions from this model. The prediction tool has identified seven novel intercalary heterochromatin regions that were previously assigned as euchromatin.

2.5.2. Introduction

Chromatin, the physical genetic material of an organism, is folded and packed into small chromosomes that fit within eukaryotic cell nuclei. The state of chromatin, or degree of condensation, provides different environments that can lead to fluctuations in chromosome interactions including transcription, recombination, and DNA repair [66]. These different compaction states, influenced by epigenetic post-translational histone modifications, are traditionally classified into two primary classes: euchromatin and heterochromatin. Euchromatin is characterized by histone modifications that loosen the chromatin and allow transcription factors to access the DNA, resulting in an environment that aids gene expression. Heterochromatin domains are characterized as repressive transcription environments resulting from compact chromatin that inhibits transcription factors from completing transcription. Heterochromatin is classically defined as late-replicating, highly condensed chromatin that stained more intensely than its euchromatic counterpart. Heterochromatin inhibits transcription factors from reaching the DNA, creating a repressive environment associated with gene silencing. However, heterochromatin does not just provide an environment that stops the transcription of DNA. Heterochromatin has been shown to have biological significance of heterochromatin in regulating genes, providing chromosomal stability, affecting nuclear organization [69], even providing an environment necessary for specific gene expression.

The technological advancement of high-throughput sequencing has dramatically increased the understanding of genome organization, evolution, and dynamics. However, the link between genome content and chromatin dynamics is still relatively unclear, especially in less studied organisms. This issue is compounded by the fact that the repeat rich nature of heterochromatin makes assembly and analysis of this important chromatin type very difficult. *Drosophila melanogaster* is a model organism that has been used for many studies in chromosome function and structure. Although many genetic studies primarily focus on euchromatin, a focus has been placed on annotating the heterochromatin of *D. melanogaster* [295]. Regions of heterochromatin have been identified through various different analyses including morphological characterization [308], replication timing [380], and histone modification patterns [381,382].

D. melanogaster, as well as some other species from the order Diptera, have giant chromosomes called polytene chromosomes that exhibit unique banding patterns and morphologically identifiable landmarks. These chromosomes, a result of replication without cell division, provide a unique opportunity to morphologically characterize chromatin without the need for staining. In *D. melanogaster* polytene chromosomes, heterochromatin is under-replicated and poorly banded. It can be morphologically divided into two classes: α-heterochromatin and β-heterochromatin [308,383]. α-heterochromatin, found at the proximal, central region of the centromeres, is morphologically compact and enriched in satellite DNA. β-heterochromatin, characterized by a diffuse, mesh-like structure, contains an abundance of transposable elements (TEs) and comprises the distal region of the centromeres. β-heterochromatin has been linked to organizing chromosomes within the nucleus. Diffuse

heterochromatin anchors chromosomes to the nuclear periphery in *D. melanogaster* [320,324] as well as in various species of Anopheline mosquitoes [325,326].

Within the euchromatin, some bands exhibit similar features to the peri-centromeric heterochromatin, and thus were identified as intercalary heterochromatin (IH) [384]. These regions appear as bands that are very compact and induce chromosomal breakage in polytene chromosomes. 52 regions located within the euchromatin of the *D. melanogaster* genome were found to be late replicating as well as under-replicated [307]. Many of these regions were described as silent chromatin types enriched in heterochromatin-associating proteins. However, the IH in *D. melanogaster* lacks an abundance of transposable elements and repetitive sequences, a hallmark of the peri-centromeric heterochromatin. Genic content is depressed, but more closely resembles euchromatin, as well [307,385]. More recently, the use of protein-binding profiles resulted in an expansion of the genomic landscape. A secondary active chromatin type, as well as a Polycomb related heterochromatin were identified, resulting in five different states of chromatin [386]. Histone modification analyses have provided a secondary method for identifying potential boundaries between heterochromatin and euchromatin. Peri-centromeric heterochromatin exhibits epigenetic histone modifications including hypoacetylation and tri-methylation of histone 3 lysine 9 (H3K9) and tri-methylation of histone 4 at lysine 20 [70,71]. These modifications are hallmarks of heterochromatin. These modifications, along with a host of other hallmark modifications have been used to create chromatin landscapes for species including *D. melanogaster* [381,382], *Caenorhabditis elegans* [387], and *Homo sapiens* [388,389]. Combined, these data supply a wealth of information regarding the composition and structure of chromatin types in *D. melanogaster*.

In *Anopheles gambiae sensu stricto,* an overall landscape has been described, outlining heterochromatin-euchromatin boundaries and some molecular features that are representative of these different chromatin types. *An. gambiae s.s.* belongs to the *Anopheles gambiae* complex, seven morphologically indistinguishable mosquito species that exhibit varying levels of competence towards transmitting malaria. These mosquitoes, too, have polytene chromosomes that aid studying chromosome structure and function at a higher resolution than standard interphase chromosomes.

Morphologically, the polytene chromosomes from ovarian nurse cells resemble their salivary gland counterparts from *D. melanogaster*. In polytene chromosomes of *Anopheles gambiae* ovarian nurse cells (ONCs), heterochromatin is characterized by either a diffuse granulated structure lacking banding pattern that is comparable to β-heterochromatin or as a compact, condensed block [257]. Euchromatin is characterized by bands and interbands unique to individual species. However, intercalary heterochromatin in *D. melanogaster* is quite different than that in the *An. gambiae* mosquito. As mentioned previously, intercalary heterochromatin of *D. melanogaster* looks like dark bands that can be found dispersed within euchromatin. The intercalary heterochromatin is late-replicating and results in under-replicated regions that have a reduced number of DNA sequence copies. *D. melanogaster* intercalary heterochromatin is not enriched in repetitive sequences, and contains a number of unique genes dispersed across the various regions [307,385]. This is in contrast to the copious amount of TEs found within the IH of *An.gambiae* [343,344]. Likewise, the morphological features of mosquito IH is different than in the fruit fly. *An. gambiae* heterochromatin is characterized by diffuse granulated chromatin that resembles the peri-centromeric heterochromatin, or the extremely condensed block of heterochromatin found on the 3R chromosome.

To better understand the chromatin dynamics of the major malaria vector, *Anopheles gambiae*, we analyzed a variety of molecular features to determine if they can be used to classify different chromatin types across the genomic landscape. Here, we present a model capable of predicting whether a region of DNA is likely to be heterochromatin or euchromatin based on a group of molecular features. This model has been used to successfully predict previously identified heterochromatin, as well as identify new regions of heterochromatin that were labeled as euchromatin. The model was also applied to the unknown chromosome, an accumulation of unassigned scaffolds from the *Anopheles gambiae* genome assembly, to predict chromatin types. First, we partitioned the genome into 100 kilobase DNA sequence windows to provide a higher-resolution genomic landscape. Second, we tabulated values for molecular features that we presumed may be representative of a particular chromosome type. Finally, these features were inserted into an algorithm that allowed for chromatin classification based on values representative of each 100 kilobase window.

2.5.3. Results and Discussion

Identification of molecular characteristics representative of *An. gambiae* genome

Previously, the genomic landscape of *An. gambiae* was monitored over a broad scale. Eight molecular features: genes, DNA TEs, RNA TEs, segmental duplications, micro- and mini-satellites, satellites, and matrix attachment regions (MAR) were observed over five megabase (euchromatin) and one megabase (heterochromatin) windows. This study provides an excellent basis for understanding how molecular features associate with chromatin state, but the lack of resolution can misrepresent smaller islands of intercalary heterochromatin hidden within euchromatin.

To overcome the likelihood of misidentifying regions of heterochromatin, this study partitioned the current *An. gambiae* assembly into 100 kilobase windows. For each 100 kb window, values for fifteen molecular features were recorded and tabulated for all five chromosomal arms (2R, 2L, 3R, 3L, and X), as well as the unknown chromosome. The initial predictive model included values and percentages for the following fifteen features **(Table 2.7)**:

Values	Repeats Genes Recent TE fragments TE fragments MAR sequences Sense piRNAs Anti-sense piRNAs
Percentages	Recent TE content (compared ONLY with TEs) Recent TE content (compared to window size) Total TE content Average TE fragment length Average recent TE fragment length MAR content MAR fragment length TE strand bias

Table 2.7. Molecular features used in initial chromatin prediction model.

The values for these features were tabulated for each chromosomal arm as well as the unknown chromosome, and graphed to identify potential correlations with currently known chromatin types.

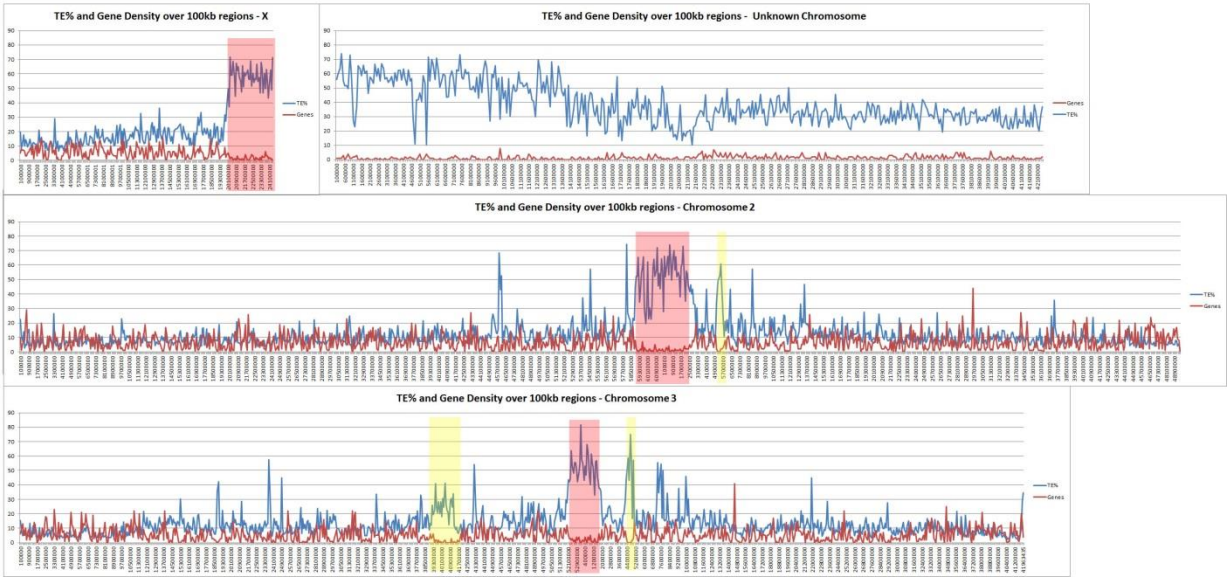


Figure 2.16. Transposable element and gene content across the *An. gambiae* landscape. Blue lines represent the percentage of TEs in a single 100kb window; red lines represent the number of genes in a single window. Red highlight indicates previously identified peri-centromeric heterochromatin. Yellow highlight indicates previously identified intercalary heterochromatin.

Figure 2.17. Recent TE content across the *An. gambiae* landscape. Blue lines indicate the number of recent TE fragments per 100 kb window. Red lines indicate the percent content of recent TEs compared to total TE fragments, while green lines indicate percent content of recent TEs compared to total window length. Red highlights represent previously identified peri-centromeric heterochromatin, while yellow highlights represent identified intercalary heterochromatin.

Five of the features are represented in **figures 2.16 and 2.17**, above. As shown in the Sharakhova *et al.* study [257], heterochromatic regions are gene-poor and contain an abundance of TEs. The abundance of repeats is elevated in heterochromatic regions. Further investigation of transposable elements, specifically recent TEs based on sequence similarity to consensus TE sequences, shows different patterns in regions of heterochromatin as compared to euchromatin. Two trends can be seen in regards to these transposable elements. First, heterochromatin appears to contain larger individual TE fragments. This data suggests that heterochromatin is capable of

preserving TEs better than the euchromatin and is likely corroborated by the fact that heterochromatin, although rapidly evolving, is the likely final resting place of most TEs. Larger fragments that have undergone mutations are present in the heterochromatin, providing vestigial elements that can be used within the piRNA pathway for TE mobilization protection. Second, the number of recent element fragments, recent element fragments being those fragments that exhibit at least a 98% sequence similarity to consensus sequences of known TEs, is also higher in heterochromatin. It has been suggested that most recently inserted transposable elements only persist in heterochromatic regions. The mechanisms behind TE regulation would actively remove any novel insertion sites within the euchromatin. Matrix attachment regions (MARs) represent anchor sites that attach DNA to the nuclear matrix. These MARs are typically enriched in repetitive elements [331], and are comprised of AT-rich stretches. Predictive tools, including SMARtest, MARfinder, and SIDD [390], identify potential MAR sites using complex calculations and sequence assumptions. Using SMARtest, we see an increased abundance of MAR sites on the centromeric half of chromosomes. However, the highest concentration of MAR regions appears to be in the peri-centromeric heterochromatin. The same could be seen for repetitive elements. Satellites, minisatellites, and microsatellites were all most abundant in most heterochromatin regions, with diffuse intercalary and peri-centromeric heterochromatin containing the largest quantities.

A model that predicts chromatin type

After tabulation of the values for each molecular feature, an identifier was also set to predetermine the chromatin type associated with every 100 kilobase window. Euchromatic regions as defined by Sharakhova *et al.* [257] were assigned a zero, while all forms of

heterochromatin were assigned a one. This allowed for rapid identification and comparison of heterochromatic regions in the model when analyzing the genomic landscape.

Identifying the best features for predicting chromatin types

In order to provide a model that best incorporates the molecular differences between euchromatin and heterochromatin, features were adjusted to find the best combination indicative of a chromatin type. The initial list of fifteen features was reduced to nine as shown below **(Table 2.8)**:

Values	Genes Ribosomal DNA genes Microsatellites Minisatellites Satellites
Percentages	Recent TE content (compared ONLY with TEs) Total TE content Average TE fragment length MAR content

Table 2.8. Molecular features used in final chromatin prediction model.

Compare prediction results to previously identified heterochromatic regions

The model was used to predict two types of heterochromatin, "chromatin type 1" and "chromatin type 2." "Chromatin type 1" embodies all of the peri-centromeric heterochromatin, as well as the intercalary heterochromatin regions found on 2L and 3L that resemble the *D. melanogaster* β-heterochromatin. "Chromatin type 2" represents the condensed α–heterochromatin-like 35B region of the 3R chromosomal arm. All of the remaining genomic

windows were assigned the designation of euchromatin. The "unknown chromosome," a concatenation of the unassembled scaffolds remaining from sequencing, was also run through the model.

A majority of the windows that were previously classified as heterochromatin were successfully identified by the predictive model at an 80% probability. The peri-centromeric heterochromatin boundary of the X and 3R chromosomes were predicted to be approximately 200,000 bp less than the previously identified junction. The X and 3R peri-centromeric heterochromatin is predicted to be nearly 4,193,107 bp and 900,683 bp, respectively. The previous 3R boundary (52,161,877 bp) can be resolved when using a 70% probability. The 100 kb windows approximating the boundaries for the 2R and 3L chromosomes match the previously identified boundaries. The 2R peri-centromeric heterochromatin is predicted to be approximately 2,545,104 bp, while the 3L peri-centromeric heterochromatin is approximately 1,800,000 bp in size. The 2L peri-centromeric boundary was predicted to expand approximately 300,000 kb further than was previously identified. The resulting heterochromatic region is roughly 2,700,000 bp in size. The intercalary heterochromatin boundaries of 3L were accurate; both boundary windows were within the range that was previously identified. However, the 2L intercalary heterochromatin boundaries were slightly different. The predicted boundaries suggest that the 2L intercalary heterochromatin is approximately 100 kb smaller on both ends.

Identification of new islands of heterochromatin and expansion of current heterochromatin boundaries

This model provides the user the ability to change the power with which a region will be assigned a chromatin type. Using a 80% likelihood prediction, six new type 1 chromatin sites and a secondary type 2 chromatin site were identified **(Table 2.9)**. These regions can be seen in table 3. As mentioned before, peri-centromeric heterochromatin boundaries were also both contracted and expanded. Previously, morphological characterization of *An. gambiae* heterochromatin was limited to the diffuse, granulated structure of the β-like heterochromatin or the condensed block nature of the α-like heterochromatin. Currently, only the centromeres and intercalary sites located on chromosomes 2L, 3R, and 3L have been labeled as heterochromatic. Sites predicted by this model, however, exhibit euchromatic morphology. These regions appear as bands and interbands that look much like their neighboring euchromatin.

"Type 1 Chromatin"			
Arm	**Start Pos.**	**End Pos.**	**Region Size**
2R	45800001	45900000	100000
	54500001	54600000	100000
	58000001	58100000	100000
2L	8400001	8500000	100000
3R	23600001	23700000	100000
3L	7200001	7300000	100000
	7500001	7600000	100000
"Type 2 Chromatin"			
Arm	**Start Pos.**	**End Pos.**	**Region Size**
3R	26600001	26700000	100000

Table 2.9. Novel intercalary heterochromatin regions found on the autosomes. Chromatin type 1 corresponds to peri-centromeric and diffuse intercalary heterochromatin. Chromatin type 2 corresponds to compact intercalary heterochromatin.

The unknown chromosome

Sharakhova *et al.* suggested that the "unknown chromosome," a collection of unmapped scaffolds, exhibited heterochromatic features [257]. Our model supports this notion, predicting almost the first 15 megabases of the unknown chromosome to be heterochromatic. This prediction would likely be larger if the scaffolds did not get progressively smaller toward the end of the chromosome. Molecular features were consistent with peri-centromeric heterochromatin, having an abundance of repeats, transposable elements, MARs, and a dearth of genes.

2.5.4. Methods

Quantification of molecular features used in predictive model

Genes and ribosomal DNA sequences were calculated using BioMart software [391]. The AgamP3.7 VectorBase gene set was used to calculate these values. Micro-satellites (period size of 2-6), mini-satellites (period size of 7-99) and satellites (period size of 100+) were calculated using Tandem Repeat Finder (TRF) [377]. Transposable element data were collected using the Genetic Information Research Institute (GIRI) available Repeat Masker. All sequence sources were selected to cover the widest possibility of transposable elements available. Matrix attachment regions (MARs) were calculated by using SMARtest [379,392].

Multinomial Logistic Chromatin Analysis

Using identified, known chromatic regions, we use a supervised learning algorithm in order to associate 11 covariates with chromatin types I, II, and the euchromatic re- gions, which we denote as Y. Hence, using the identified chromatic regions on the X, 2L, 2R, 3L, and 3R chromosomes, we train our classifier based on the 11-dimensional feature space composed of: {Repeats, Genes, Recent TEs (counts), TE (total), Mar counts, sense piRNA counts, anti-sense counts, Recent TE% (all), Recent TE% (to- tal), TE% (length), AFL, ARFL,%MAR, and MAR Ave. Frag, C%}, denoted by Z.

For our purposes, each observation consists of a sequence of $Z_{i,j}$s, where the indices $\{i, j\}$ denotes the ith partition of the jth chromome, $j \in \{X, 2L, 2R, 3L, 3R\}$, and $Y_{i,j}$ represents an indicator function for whether the observed class is chromatin type

I, II, or euchromatin. Using the 3 class designation, we use the Multinomial-Logistic regression classifier in order to learn the relationships between Z and Y, as well as predict regions in which Chromatin Types I and II are likely to occur. Multivariate multinomial-logistic regression is applied when the response variable (Y) is a multi- categorical nominal variable and there is more than one independent variable. This is analogous to the standard logistic regression model, when the response variable consists of only two classes. This technique is also commonly referred to as soft-max regression [393]. Given observation of $Z_{i,j}$, the probability of belonging to the classes: {I, II, eu}, respectively, are modeled by:

$$p_I(\beta_1, \beta_2) = \frac{\exp(Z'_{i,j}\beta_1)}{1 + exp(Z'_{i,j}\beta_1) + exp(Z'_{i,j}\beta_2)}$$

$$p_I(\beta_1, \beta_2) = \frac{\exp(Z'_{i,j}\beta_2)}{1 + exp(Z'_{i,j}\beta_1) + exp(Z'_{i,j}\beta_2)}$$

$$p_{eu}(\beta_1, \beta_2) = 1 - (p_I(\beta_1, \beta_2) + (p_{II}(\beta_1, \beta_2)),$$

where the vectors $\beta 1 = (1, b_1^{(1)}, b_2^{(1)}, \ldots, b_1^{(1)}1)'$, and $\beta 2 = (1, b_1^{(2)}, b_2^{(2)}, \ldots, b_1^{(2)}1)'$ represent the measured dependencies between the observed features (Z) and the response class (Y). These vectors (β_1, β_2) are found by maximizing the multinomial likelihood function:

$$L(\beta_1, \beta_2 | Z, Y) = \prod_i \prod_j p_I(\beta_1, \beta_2)^{\delta_I(Y_{i,j})} p_{II}(\beta_1, \beta_2)^{\delta_{II}(Y_{i,j})} p_{eu}(\beta_1, \beta_2)^{\delta_{eu}(Y_{i,j})}$$

where $\delta_c(Y_{i,j}) = 1$ if $Y_{i,j}$ is of class $c \in \{eu, I, II\}$, and 0 otherwise.

Chapter 3. Organization of piRNAs and evolution of piRNA clusters in the African malaria vector *Anopheles gambiae*

3.1. Summary

Background

The Piwi-interacting RNA (piRNA) pathway is an important mechanism in the defense against transposable element (TE) mobilization in many species including *Drosophila melanogaster* (the fruit fly), *Aedes aegypti* (the Yellow Fever mosquito), and *Mus musculus* (the mouse). These piRNAs localize in specific genomic loci, termed piRNA clusters that produce long single-stranded piRNA pre-cursors. These pre-cursors are later processed into piRNAs that allow predominantly for targeted inactivation of complementary TEs in species with small genomes, such as *D. melanogaster*. Although vestigial TEs are responsible for a large proportion of piRNAs of in many species, a shift toward gene derived piRNAs can be seen in species with larger genomes including *Ae. aegypti*. This study represents that first attempt to characterize the piRNA pathway in *An. gambiae*, a species with an intermediate-sized genome.

Methods

To better understand how this mechanism has diverged, we have sequenced and created libraries of small-RNAs from ovarian tissues of both the M (Mali) and S (Zanu) forms and mapped potential piRNAs to the *Anopheles gambiae* PEST reference genome. piRNA libraries including uniquely mapping small RNAs ranging from 24-30 nucleotides in size were mapped using a short-read mapper, NucBase. Gene ontology (GO) terms and expression enrichment were

analyzed to determine potential function of 40 piRNA enriched genes. Finally, locations of a chromosomally bound Piwi homolog (AGAP009209) were identified using immunostaining.

Results

Here, we identify a probable pool of piRNAs sequenced from ovarian nurse cells of *An. gambiae* M and S molecular forms, classify likely genomic loci that likely generate these piRNAs, and analyze sequences enriched in piRNAs. We identified potential piRNA clusters and their spatial organization within the *An. gambiae* genome, resulting in 81 and 77 unique genomic loci with piRNA enrichment. We found that compared with *D. melanogaster* and *A. aegypti, An. gambiae* has intermediate values for % of genome occupied by piRNA clusters (6%), and % of major intercalary piRNA clusters (53%). piRNAs mostly localize to LTR retrotransposons (Gypsy, Pao) and non-LTR (LINE elements) TEs, but DNA TEs make up the majority in the *An. gambiae* genome. Although only 0.5% piRNAs map to genes in the top 15 clusters (vs 13.6% to TEs), ~17% of total unique piRNAs mapped to the sense strand of gene exons. The piRNA targeted genes are enriched with germ-line cell/reproduction/development and translation initiation factors GO terms. These genes significantly change expression profile during individual development, blood meal development, and in reproductive organs of males and females. Unlike in *D. melanogaster*, PIWI1 has no significant enrichment in centromeres and shows very little co-localization with HP1 in *An. gambiae.*

Conclusions

piRNAs in *An. gambiae* represent a shift from primarily transposon derived sequences located in sub-telomeric and peri-centromeric heterochromatin to a uniform dispersion of clusters throughout the genome. As the genome size increases mapping of piRNAs to TEs

decreases and clusters spread from pericentric/telomeric heterochromatin to intercalary heterochromatin and euchromatin occupying larger portion of the genome. However, a large fraction of Anopheline TEs is not targeted by the piRNA pathway and a large portion of piRNAs outside clusters targets genes. Our data suggest that the piRNA pathway may play a role in fertility, reproduction, development, and maternal inheritance of *An. gambiae*. Results from this study also suggest evolution of the piRNA pathway as a result of increasing genomic complexity in Dipteran species.

3.2. Introduction

Transposable elements (TEs) comprise a significant part of many eukaryotic genomes. TEs are DNA elements capable of mobilizing to new sites within a host's genome, resulting in potentially mutagenic or disruptive consequences that can have a negative effect on the host [107]. The abundance of TEs in genomes tends to be high in many species, especially in higher eukaryotes; however, genomic TE content varies significantly from species to species. Estimations range from approximately 0.3% in *Escherichia coli* up to 45% in *Homo sapiens* and 77% in *Rana esculenta* [104]. When properly regulated, these elements can have positive benefits toward the host, resulting in positive genome evolution. However, the de-repression of TEs can lead to deleterious consequences within the host genome. Thus, it is vital that the genome has protection against mass mobilization of TEs.

Small RNA silencing mechanisms, which are present in many metazoan species, provide post-transcriptional silencing and degradation of complementary messenger RNA (mRNA). One of these mechanisms, the Piwi-interacting small RNA (piRNA) pathway, plays an important role

in regulating TEs by silencing active elements through slicer-mediated cleavage. The piRNA pathway is a dicer independent small RNA silencing mechanism that utilizes RNA induced silencing complexes (RISCs) [169,394] which target selfish mobile elements. Like the other small RNA pathways, the piRNA pathway involves members of the Argonaute family of proteins in order to perform its role in TE regulation. Members of the Piwi clade, a subfamily of Argonaute, interact with short RNA sequences to effectively create a RISC that can target and silence complementary mRNA sequences. A mutation of any of the three key Piwi proteins- Piwi, Aubergine, and Argonaute 3- result in de-repression of TEs in the germline [147-149], indicating the necessity of these proteins in functional TE silencing.

Originally described in *Drosophila melanogaster* [144], the primary function of the piRNA pathway is silencing transposons within the germline. piRNA pathways were also discovered in other model organisms including Zebrafish [169] and mice [395], as well as one of the more studied blood feeding organisms, *Aedes aegypti* [161]. In addition to TE silencing, the piRNA pathway has been implicated to have other functions. piRNAs have been shown to interact with various protein coding genes including *Stellate* [157], *Fascicilin 3* [158,159], and *vasa* [160] in *D. melanogaster*, functionally silencing the mRNA transcripts. Two of these genes, *Stellate* and *Traffic jam* (which silences *Fascicilin 3*), show piRNA enriched in *Aedes aegypti*, as well [161]. The piRNA pathway has been connected to anti-viral immune responses in *Ae. aegypti*. Elevated quantities of 23-30 nucleotide small RNAs have been documented as a result of viral infection in both *Ae. aegypti* and *Ae. albopictus* [170-173]. The piRNA pathway has also been implicated in germline development and function. piRNA interactions occur during gametogenesis, and the Piwi protein is essential to fertility in *Caenorhabditis elegans* [167,168], *Danio rerio* [169], *and Mus*

musculus [153]. Germline stem cell loss or sterility has also been documented as a result of piRNA pathway mutation [153,155,162,163,396].

Two mechanisms for piRNA production have been identified in *D. melanogaster*. Both mechanisms stem from long single stranded piRNA precursors that originate from vestigial transposons in *D. melanogaster* [144] and *M. musculus* [397] as well as genes in the *A. aegypti* genome [161]. In all cases, however, piRNAs appear to originate from clusters, small genomic regions ranging in size from approximately 1 kb to 250 kb [144,161]. In *D. melanogaster*, these clusters are almost exclusively located in heterochromatic regions, namely the sub-telomeric and peri-centromeric regions that have an abundance of TEs [144]. *A. aegypti* clusters are much more pervasive, occupying all regions of the genome [161]. piRNA clusters do not have an explicit strand bias; however, in some cases do exhibit high percentages of TEs in one orientation or the other [144,397].

In *D. melanogaster*, piRNA clusters produce long, single-stranded RNA transcripts that are processed into primary piRNAs that are cleaved into smaller primary piRNAs that can be loaded onto one of the three piwi proteins. This process has been referred to as primary piRNA biogenesis. Trimming of the piRNA to the 24-30 nucleotide characteristic of these small RNAs requires the cytoplasmic endonuclease Zucchini [182,183]. Once trimmed to the proper size, the piRNA can be loaded onto the Piwi protein. Secondary piRNAs, responsible for a large portion of the total piRNA pool in the germline, are generated through an amplification cycle referred to as the ping-pong cycle [144]. Fellow piwi proteins Aubergine and Argonaute 3 are able to bind anti-sense and sense piRNAs, respectively. Similar to primary biogenesis, piRNA clusters can produce transcripts that feed the ping-pong cycle with precursor transcripts that are cleaved and trimmed by the piwi proteins. Alternatively, transposon RNAs can be targeted by the protein-

RNA complexes, resulting in secondary piRNAs that can be loaded on partner piwi proteins. Many of the Aubergine and Piwi associated piRNAs are derived from antisense TEs, while Argonaute 3 piRNAs come from sense oriented TEs.

piRNAs appear to be somewhat conserved through species, sharing similar features including length (approximately 24-31 nucleotides) and a typical motif of Uridine at position one of anti-sense transposon derived piRNAs and Adenine at position ten of sense strand transposon derived piRNAs [144]. A ten base pair overlap can be seen between complementary piRNAs [144,184]. The clusters where the piRNAs are produced also appear to be well conserved spatially, but not in regards to sequence [398].

It is of note that there are similarities between the piRNA pathways of various species, but in some respects are very different. Using the two insect species mentioned previously, *Ae. aegypti* and *D. melanogaster* have very different genomes. TEs account for approximately 47% of the *Ae. aegypti* genome, while the *Drosophila* genome is composed of approximately 15-22% TEs [104]. The proteins involved in the piRNA pathway have expanded from the single Argonaute 3, Aubergine, and Piwi in *D. melanogaster* to six Piwi genes and a single Argonaute 3 in *Ae. aegypti*. There is also a noticeable shift in piRNA cluster location. piRNA clusters are primarily confined to the peri-centromeric heterochromatin of *D. melanogaster*. piRNA clusters are found dispersed across the genome, with many piRNAs being generated from genes, rather than TEs. The *Anopheles gambiae* genome assembly is currently 260 Mb and has approximately an 18% TE load. This genome represents an intermediate between *D. melanogaster* (175 Mb, 15.8% TE load) [399] and *Ae. aegypti* (1,310 Mb, 47% TE load) [99]. Studies regarding piRNAs in the *An. gambiae* genome may provide a link between compact genomes with smaller TE loads and larger genomes containing a higher proportion of TEs.

3.3. Results

3.3.1 Small RNA libraries of *An. gambiae*

To identify potential piRNAs in *An. gambiae*, we isolated and sequenced small RNAs from the M and S molecular forms of *Anopheles gambiae*. A single library was generated for each form, with a unimodal RNA distribution that had peaks occurring at 22 nt, likely indicative of miRNAs, 27 nt, as well a smaller peak at 24 nt for both forms. A broad peak spanning 24-30 nt was present in both libraries. This range is in close concordance with *Aedes aegypti* (24-31 nt peak)[161], *Drosophila melanogaster* (23-29 nt)[144], *Bombyx mori* (26-31 nt)[400], and *Danio rerio* (24-30 nt)[169].

Libraries were modified to only include unique sequences that mapped a single time to the *An. gambiae* reference genome. A read mapped called NucBase [401] was used to map the sequences to the *An. gambiae* PEST genome (Release AGamP3.7), a reference strain that serves as an admixture for both the M and S forms. All non-unique mappers were removed from the data set, and re-plotted with NucBase. The final libraries resulted in 585,366 and 496,393 small RNA sequences for the M and S forms, respectively.

Approximately 12% of the total sequenced piRNA population mapped to annotated TE fragments in the *An. gambiae* genome **(Figure 3.1)**. TEs from the long terminal repeat (LTR) families Gypsy and Pao-Bel, as well as long interspersed elements (LINEs) CR1 and R1 had the highest piRNA enrichment as shown in (**Figure 3.2)**. Roughly 17% of the unique piRNA population mapped to exons of genes.

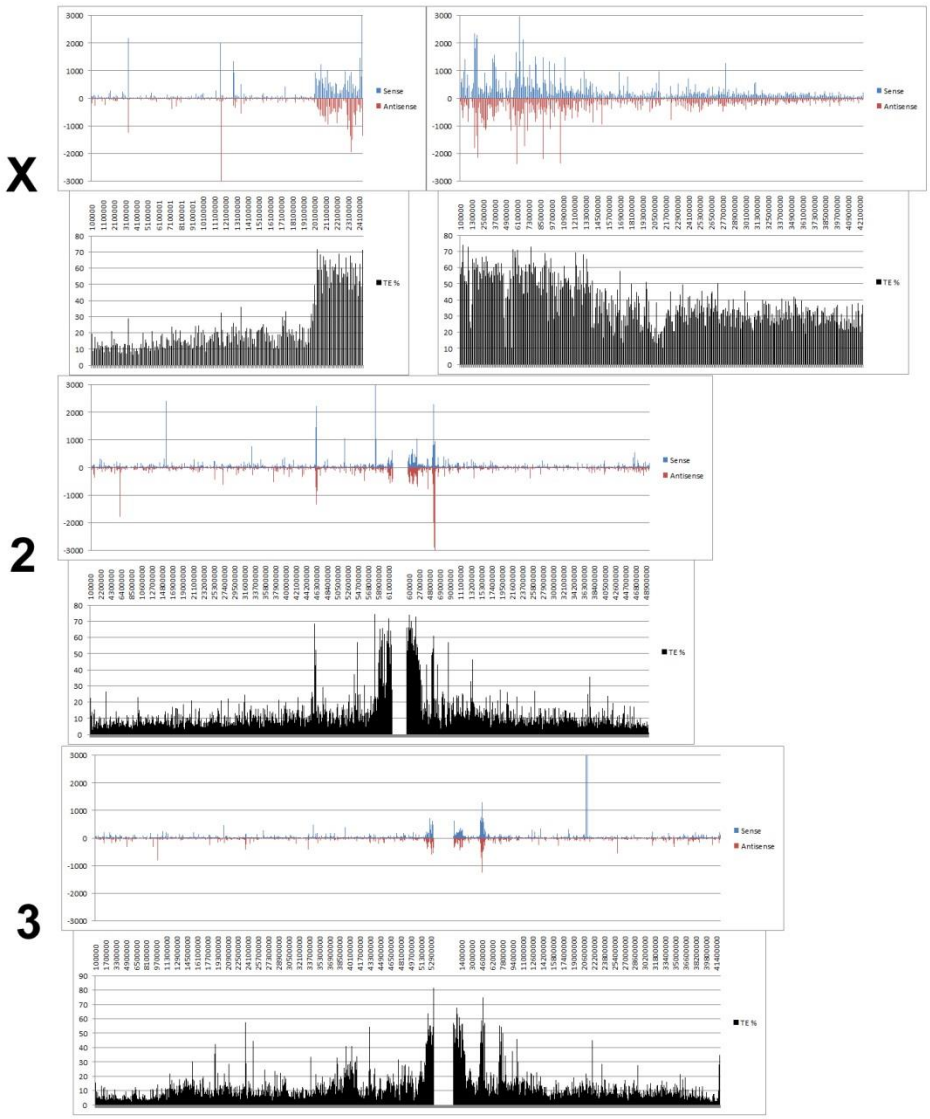


Figure 3.1. piRNA and TE distribution across *An. gambiae* genome assembly. Red and blue lines in upper graphs indicate opposite orientation piRNAs; black line graph indicates transposable element content. Arms are divided at the centromere (blank space).

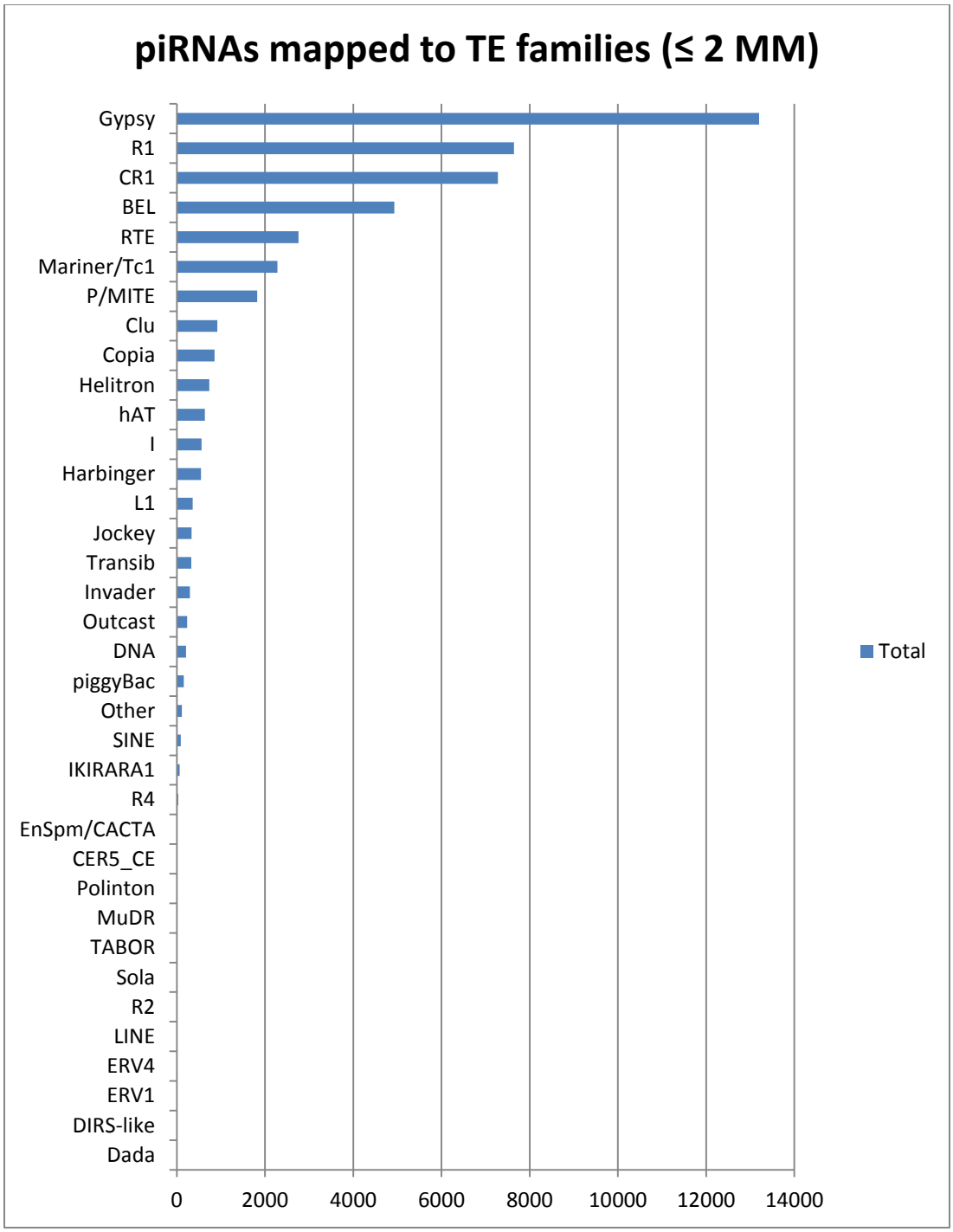


Figure 3.2. Distribution of transposable elements families in the M form of *An. gambiae*. Retrotransposon families represent the most highly enriched annotated transposable elements.

3.3.2 piRNA clusters

piRNA clusters have been shown to be one of the primary sources of piRNA production in the genome. The location of these clusters is different in various species, with clusters primarily being restricted to sub-telomeric and peri-centromeric heterochromatin [144] in *D. melanogaster*, yet dispersed throughout the entire genome in *Ae. Aegypti* [161]. In *An. gambiae*, piRNA clusters are predominantly found in the peri-centromeric heterochromatin, as well as the diffuse intercalary heterochromatin found on chromosomes 2L and 3L **(Figure 3.3).** Regions with an abundance of piRNAs mapping to them were typically found to be enriched in TEs.

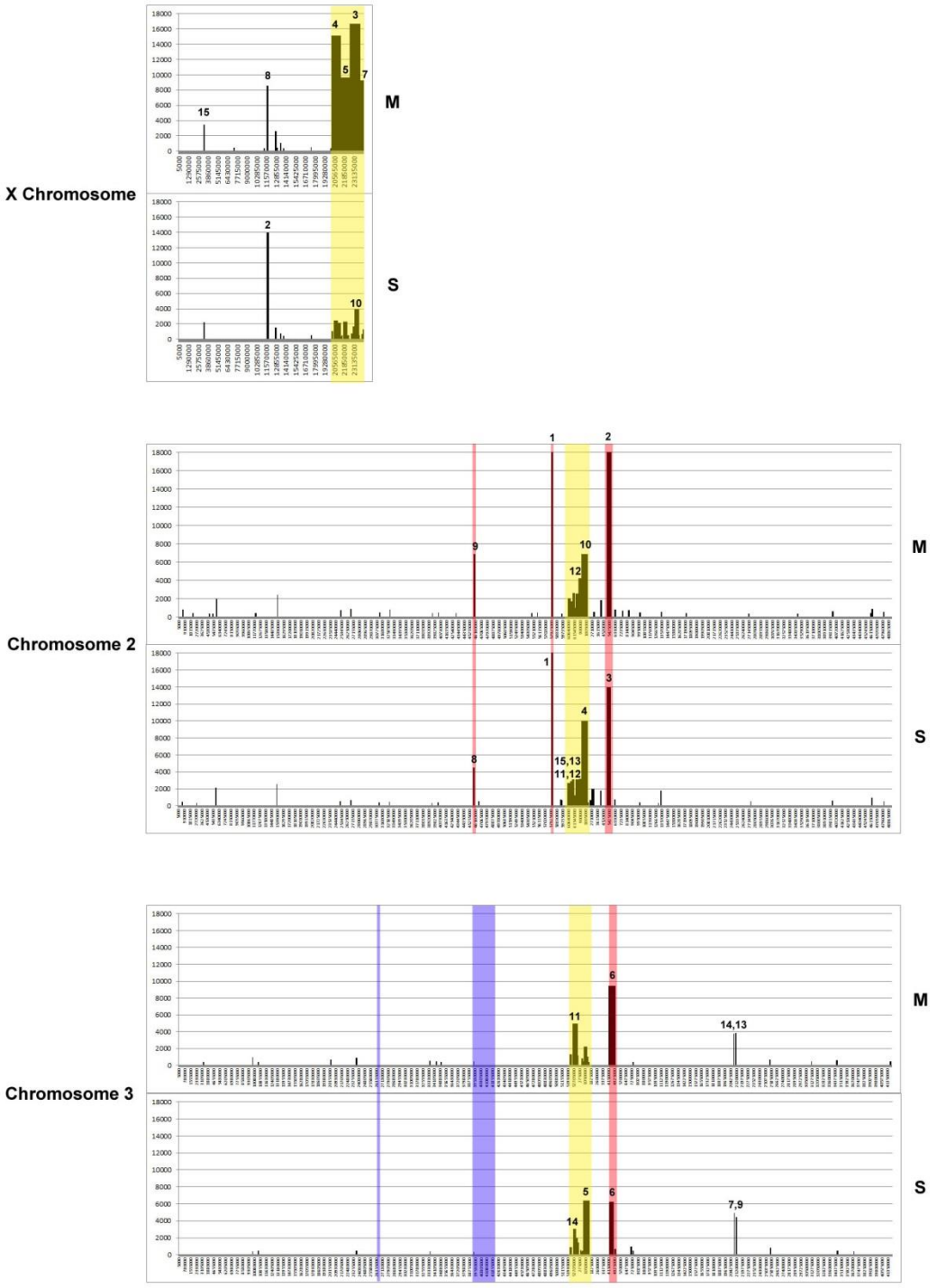


Figure 3.3. Distribution of piRNA clusters across *An. gambiae* M and S form chromosomes. Yellow highlighted regions represent peri-centromeric heterochromatin, red highlight regions represent diffuse intercalary heterochromatin, and purple highlighted regions represent compact intercalary heterochromatin. The top 15 clusters are also denoted in descending order based on total number of piRNAs mapped to the cluster.

To assign a minimum cutoff value for a designated cluster, we chose to use 359 piRNAs, which was .05% of the total original piRNA library from the M form. Clusters were assigned with two restrictions: consecutive windows had to have ten or greater piRNAs, with a maximum of five adjacent windows totaling 25 kb falling below this ten piRNA threshold. A basic Perl script was generated to identify consecutive windows with greater than 10 piRNAs, as well as concatenated windows that met the 359 piRNA requirement. A total of 82 individual clusters were identified in the M form, while 78 clusters were identified in the S form. These clusters ranged in size from 10 kb to 1.3 Mb, with members of the top 15 ranging from 35 kb to 1.28 Mb in total size. In general, these clusters were larger in size than the observed clusters in both *Aedes* and *Drosophila*, which range in size from 6 to 184 kb [161] and 2 to 242 kb [144], respectively. When creating clusters, we opted to exclude the "unknown" chromosome, which comprises 42.4 Mb of the genome and is believed to be primarily heterochromatin. The unknown chromosome is comprised of unmapped scaffolds placed in sequential order based on size. Although the potential for piRNA clusters being found on the unknown chromosome is high, the lack of ordering ability and assembly make identifying individual clusters within the "chromosome" increasingly difficult. Thus, we chose to only analyze the chromosomally assembled portion of the genome that would allow for prediction of clusters.

According to the obtained mapping data, approximately 6% of the genomic landscape of *An. gambiae* is predicted to be clusters that produce piRNAs. Many of the largest piRNA clusters were found in heterochromatic regions. Of our identified top 15 clusters, nine were found to be located in heterochromatin. Of those nine, seven were identified as peri-centromeric, while the other two belonged to the intercalary heterochromatin of 2L and 3L. Interestingly, there is a noticeable shift from clusters located in the peri-centromeric heterochromatin to the

intercalary heterochromatin of 2L and 3L. These regions, unlike intercalary heterochromatin in *D. melanogaster*, are enriched in TEs, so it not much of a surprise that these regions represent significant piRNA production sites. Clusters located in intercalary heterochromatin produce approximately 42% of the piRNAs from the sequenced library, while the peri-centromeric heterochromatin yields approximately 45%. Telomeric regions of *An. gambiae* chromosomes did not show any abundance of piRNAs. This is likely attributed to poor assembly of telomeres or to the fact that there is not an enrichment of TEs found in the telomeres in *An. gambiae*.

We mapped both the M and S forms to the *An. gambiae* PEST reference strain, as genetic similarity is high between the two forms. Much of the genetic diversity between these forms can be seen in the peri-centromeric heterochromatin of the X, 2L, and 3L arms. Thus, major differences in piRNA localization were expected in these areas. Strikingly, little difference was observed in spatial positioning of clusters between the two forms. 12 of the top 15 piRNA enriched clusters were present in both the M and S forms. The other 3 clusters rounding out the top 15 in the M form were displaced in S due to the X peri-centromeric clusters being split into 4 clusters that produced an abundance of piRNAs. A majority of the remaining clusters were present in both forms, with minor differences in predicted cluster boundaries. In most cases, when a cluster was predicted in one form and not the other, the form lacking the cluster had piRNA enrichment in the same region, but not abundant enough to meet the minimum 359 piRNA requirement to be classified as a cluster. The X chromosome showed the largest numerical difference in clusters. The M form's X chromosome had 5 clusters that accounted for the entire peri-centromeric heterochromatin, while 13 smaller clusters comprised the entirety of the same heterochromatic region.

3.3.3. TE composition of clusters

Brennecke *et al.* described piRNA clusters that were composed of between 70-99% TEs [144], supporting the theory that these clusters are potentially involved in regulating transposons. However, when analyzing the top 15 clusters of the two *Anopheles gambiae* molecular forms, these clusters show TE content ranging from approximately 15% to 64%, with an average of approximately 48% TEs **(Figure 3.2)**. TE content was also markedly lower than the average in over half of the euchromatic clusters within the top 15 (48.41%, 27.13%, 46.51%, 14.51%, 28.23%, 32.92%); yet these loci account for 37.5% of the piRNAs that mapped to top 15 clusters. These numbers fall in line with the observed TE content of piRNA clusters in *Aedes aegypti,* which displayed a near 50% TE composition for its top 30 clusters [161]. Of the top 15 clusters, 12 had piRNAs transcribed from both strands; the other three clusters all were uni-directionally transcribed from the sense strand **(Figure 3.4)**.

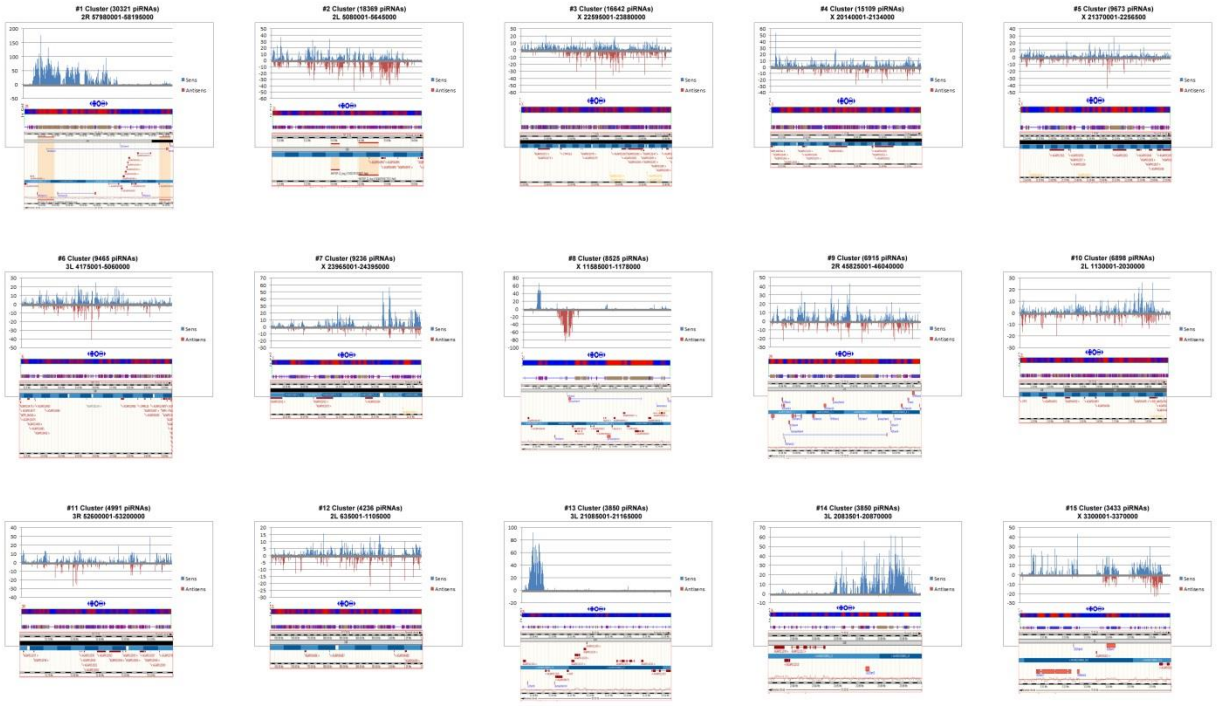


Figure 3.4. The composition of the top 15 piRNA producing clusters. piRNA mapping is shown as a bar graph, with transposable element heat maps and gene locations within the cluster shown below.

3.3.4. piRNAs mapping to genes

As mentioned previously, almost 18% of the piRNA pool mapped to gene exons. We identified genes that had more than 100 piRNAs mapped to them and explored their functions, as well as differential expression patterns. A total of 79 genes were included in this bioinformatics study; the list of genes can be found in **table 3.1**.

AGAP004203	AGAP004896	AGAP008482	AGAP008578
AGAP012447	AGAP010929	AGAP007075	AGAP004827
AGAP003387	AGAP009441	AGAP002803	AGAP012888
AGAP007406	AGAP009031	AGAP006023	AGAP012334
AGAP007365	AGAP008879	AGAP005947	AGAP001780
AGAP002076	AGAP001157	AGAP009990	AGAP007388
AGAP004038	AGAP002667	AGAP001279	AGAP010173
AGAP000203	AGAP004212	AGAP004251	AGAP012701
AGAP004895	AGAP003663	AGAP006958	AGAP002919
AGAP000427	AGAP002306	AGAP009057	AGAP001928
AGAP005005	AGAP000029	AGAP009863	AGAP003592
AGAP013330	AGAP004192	AGAP000837	AGAP004251
AGAP009135	AGAP000167	AGAP003519	AGAP003290
AGAP002655	AGAP002100	AGAP004672	AGAP006609
AGAP028112	AGAP001296	AGAP005134	AGAP001826
AGAP004750	AGAP006023	AGAP000883	AGAP008738
AGAP002654	AGAP011092	AGAP007647	AGAP004344
AGAP003556	AGAP007493	AGAP009929	AGAP011872
AGAP009441	AGAP004896	AGAP010257	AGAP003532
AGAP004671	AGAP007867	AGAP003769	

Table 3.1. List of *An. gambiae* annotated genes with more than 150 piRNAs mapping to them.

Gene ontology (GO) terms were assigned and clustered for genes found within this list. Three primary clusters were identified, with primary functions relating to ribosomal proteins, reproductive development, and nucleotide binding. These clusters can be found in **figure 3.5**. The identification of a cluster of fives genes associated with reproductive development and embryogenesis is not all that surprising. As mentioned above, the piRNA pathway has been implicated in germline stability and is essential to fertility in multiple species. *Vasa*, one of the

genes found to be enriched in piRNAs in *Drosophila melanogaster*, encodes a required protein for embryonic patterning and germ cell specification [402]. This protein is homologous to translation initiation factor eIF-4a in the mouse [403,404]. That these genes are involved in development and translation helps to support the piRNA pool that we have identified.

Annotation Cluster 1	Enrichment Score: 3.9	Count	P_Value
SP_PIR_KEYWORDS	ribosomal protein	7	0.00000
SP_PIR_KEYWORDS	ribonucleoprotein	6	0.00001
GOTERM_BP_FAT	translation	10	0.00001
GOTERM_MF_FAT	structural constituent of ribosome	8	0.00002
GOTERM_CC_FAT	ribosome	8	0.00014
KEGG_PATHWAY	Ribosome	7	0.00024
GOTERM_CC_FAT	ribonucleoprotein complex	8	0.00053
GOTERM_MF_FAT	structural molecule activity	9	0.00081
GOTERM_CC_FAT	non-membrane-bounded organelle	9	0.01800
GOTERM_CC_FAT	intracellular non-membrane-bounded organelle	9	0.01800
Annotation Cluster 2	Enrichment Score: 2.05	Count	P_Value
GOTERM_BP_FAT	reproductive developmental process	5	0.00300
GOTERM_BP_FAT	gamete generation	5	0.00570
GOTERM_BP_FAT	sexual reproduction	5	0.00670
GOTERM_BP_FAT	multicellular organism reproduction	5	0.00690
GOTERM_BP_FAT	reproductive process in a multicellular organism	5	0.00690
GOTERM_BP_FAT	spermatogenesis	3	0.00920
GOTERM_BP_FAT	male gamete generation	3	0.00920
GOTERM_BP_FAT	reproductive cellular process	4	0.02100
GOTERM_BP_FAT	germ cell development	3	0.03600
Annotation Cluster 3	Enrichment Score: 1.05	Count	P_Value
SP_PIR_KEYWORDS	nucleotide-binding	8	0.01300
GOTERM_MF_FAT	GTPase activity	3	0.02600

Figure 3.5. Gene ontology (GO) term clustering analyses of gene subset with more than 100 piRNA enrichment. All GO terms are significant, as shown in P-value side bar. Biological processes (BP) showing significance in cluster two were reproduction and germline development related. Molecular functions (MF) with the most enrichment were translation initiation related.

We took the five genes that comprised the second annotation cluster that corresponds to reproductive development and gametogenesis. Expression patterns were taken for this gene subset, with the top 10 most differentially expressed experiments and subsequent changes in expression profiles being taken for each gene. These expression results can be seen in **figure 3.6**.

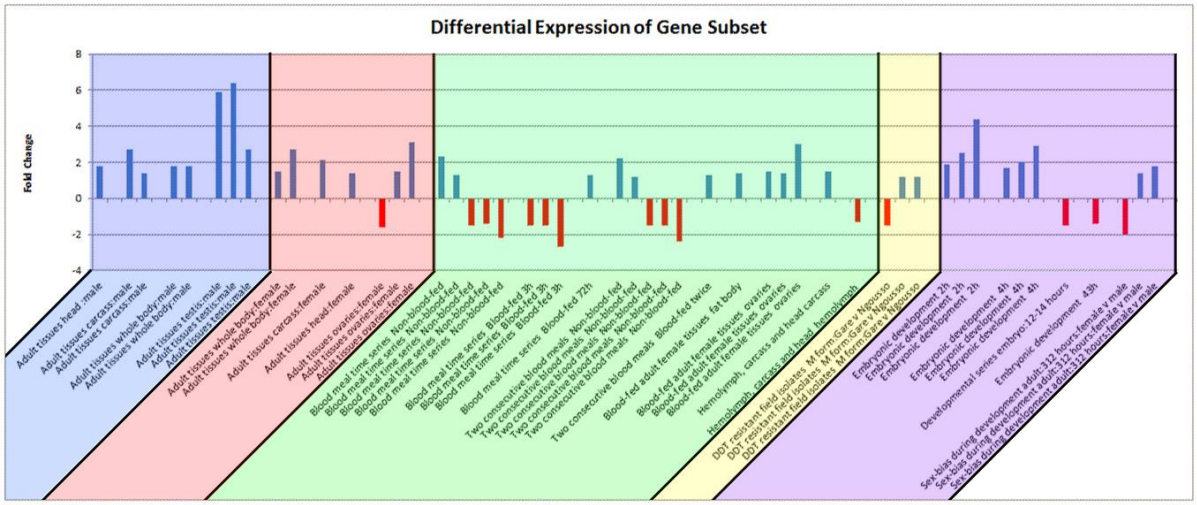


Figure 3.6. Expression analysis of reproduction related GO term annotation cluster genes. The graph indicates the number of genes (out of 5) that are either up- or down-regulated in a specific expression assay. A cluster of three genes (AGAP003290, AGAP008578, and AGAP001780) exhibit a similar trend where genes are down-regulated prior to blood meal and up-regulated after successful blood meal.

Two observations become apparent when looking at expression profiles of these genes. Many of the genes are up-regulated in reproductive tissues for both male and female, as well as in experiments regarding embryonic development. Finding an increased expression in female tissues can be expected, as RNAs were initially extracted from ovarian tissue. However, seeing these same genes up-regulated in male tissues may highlight the importance of these genes in reproductive function.

Second, we see a developing pattern amongst a subset of these five genes, where AGAP003290, AGAP008578, and AGAP001780 all have similar expression profiles before taking a blood meal and up to three hours post-blood meal. These genes are down-regulated, as shown in red of **figure 3.6** for the cluster of three genes in the blood meal time series. After completion of the blood meal, however, these genes are up-regulated throughout embryonic development (also shown in **figure 3.6**). The comparable expression profiles from these three genes may show their importance in facilitating the development of embryos in *An. gambiae.*

3.3.5. Piwi immunostaining

Homologs of the Piwi protein have been previously identified in *An. gambiae* [405,406]. A gene in *An. gambiae,* AGAP009509 (Piwi 1), most closely resemble Piwi genes from *D. melanogaster*, having 47.5% to Piwi. We had antibodies produced from synthesized peptides derived from mostly unique regions of Piwi 1. These antibodies were used in immunostaining experiments to determine chromosomal bound locations of the hypothetical Piwi 1. In *D. melanogaster*, Piwi co-localizes with Heterochromatin Protein 1 (HP1) [407]. We expected to see the same pattern of binding, as much of the piRNA pool associated with heterochromatic regions in *An. gambiae*. Very little co-localization was seen when co-immunostaining using HP1 and Piwi as seen in **figure 3.7**. However, multiple regions were characterized by Piwi surrounded by HP1. Refer to **table 3.2** for all of the localization sites for both HP1 and Piwi in *An. gambiae*.

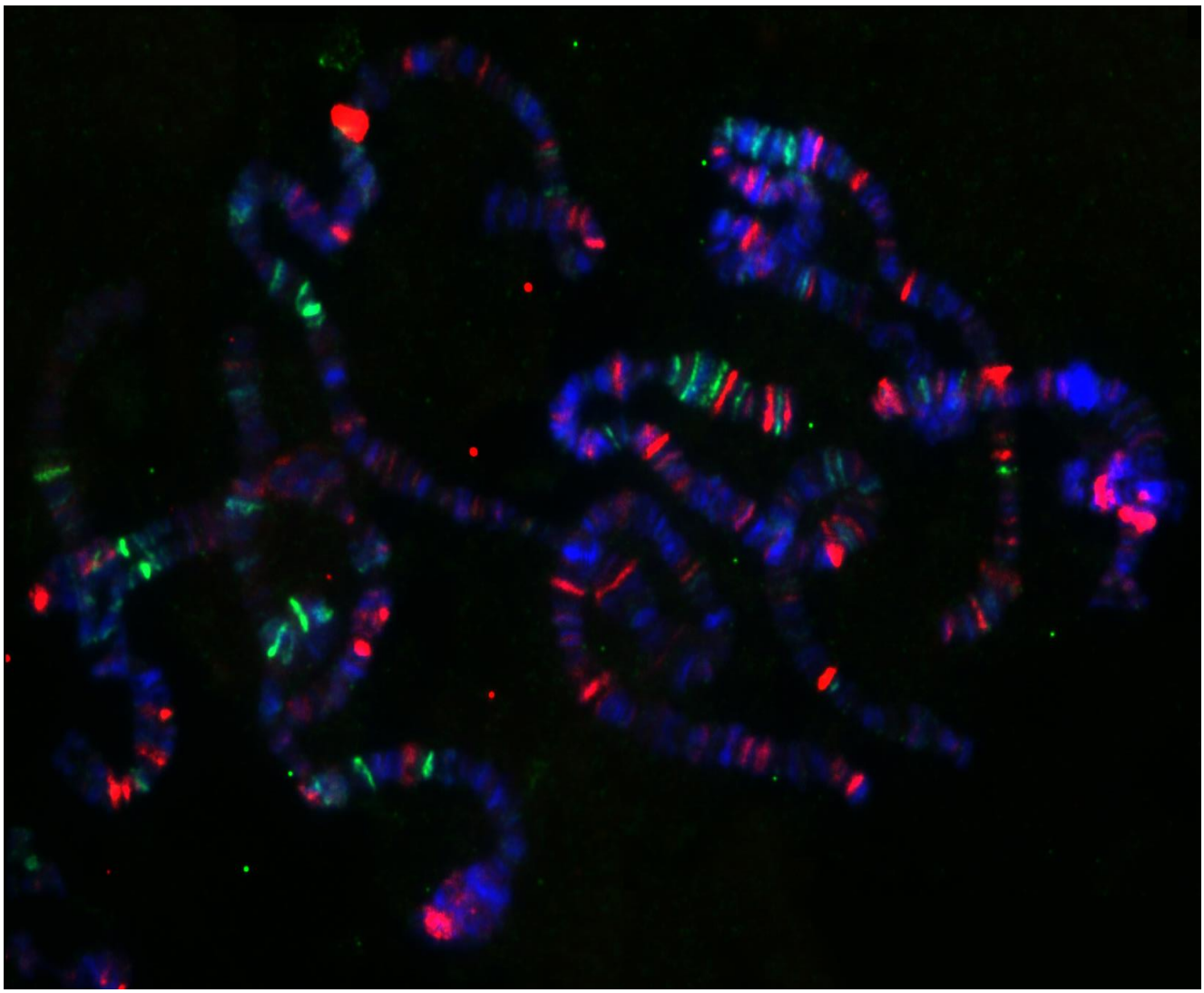


Figure 3.7 Immunolocalization of Piwi1 and HP1 antibodies to chromosomal squashes. Antibodies for Piwi1 (green) and HP1 (red) were localized to polytene chromosomes (labeled in blue) from *An. gambiae* mosquitoes.

	PIWI	HP1
2R		
Total Sites	34	57
Multiple Sites	13	26
2L		
Total Sites	26	43
Multiple Sites	10	11
3R		
Total Sites	29	38
Multiple Sites	16	19
3L		
Total Sites	17	15
Multiple Sites		
X		
Total Sites	22	10
Multiple Sites	14	5

Table 3.2 Identification of PIWI and HP1 localization sites in *An. gambiae* via immunolocalization. Identified localization sites were counted to identify permanent (multiple sites), as well as sites not found across all slides (total sites).

3.4. Discussion

3.4.1. Evolution of piRNA spatial organization

With the identification of piRNA clusters in both *D. melanogaster* and *Ae. aegypti*, it has become clear that the piRNA pathway has evolved within Dipteran species. Surprisingly, although the *Ae. aegypti* genome has such an abundance of TEs in the genome, many of the piRNAs do not map to these elements, but rather genes. *An. gambiae* represents a genome that, in terms of genomic content, can be considered an intermediary between these Dipteran species

in terms of genome size and TE load. Our results show that the piRNA pathway is expanding, and we suggest two potential reasons. Not only have the proteins involved in the pathway expanded from one Piwi protein in *D. melanogaster*, to two in *An. gambiae*, to six in *Ae. aegypti*, but piRNA clusters also have begun to occupy a larger portion of the genome. There is certainly a correlation that can be observed in regards to genome size, piRNA cluster occupancy, and expansion of Piwi proteins. This would suggest that the piRNA pathway is compensating for the increased genomic load, and as a result has requires more genetic resources to maintain stability. A recent study in *Drosophila* has shown that novel piRNA clusters can be generated *de novo* by transposon containing transgene insertion [408]. Intercalary heterochromatin in *An. gambiae* is highly enriched in TEs and is similar to peri-centeromeric heterochromatin, while TEs are dispersed throughout the *Ae. aegypti* genome. It can be construed, then, that novel clusters appearing outside of the peri-centromeric heterochromatin would form, creating these new regions that may contain protein coding genes and other non-repetitive DNA.

A secondary viewpoint would implicate the piRNA pathway in secondary roles that extend past TE inactivation, and thus requiring expansion. piRNAs have already been identified in anti-viral immunity and gene silencing roles. Our results show that many of the genes associated with piRNAs have similar functions in development, reproduction, and translation initiation. These same functions are evident in piRNA-associating genes in both *Ae. aegypti* and *D. melanogaster*. It is possible that the piRNA machinery has begun to extend to more genes within the mosquito genomes as a result of clusters located outside of gene-poor heterochromatic regions.

3.4.2. Spatial organization within closely related forms

Although this study does not attempt to quantitatively analyze the two piRNA libraries, we were able to make conclusions on the localization of the RNAs to a reference genome. When the study was initiated, it could be expected that the M and S forms would likely show different piRNA localization in the three peri-centromeric regions that showed a high level of genetic diversity, termed "speciation islands" [51]. Differentiation in TEs in these two forms was already identified [409,410] and at least some of this differentiation occurs within heterochromatic regions. However, we saw a high level of spatial conservation between the two forms. The greatest difference between the two forms was noticed in the X chromosome. Where five clusters comprise a majority of the peri-centromeric heterochromatin in the M form, 13 smaller clusters encapsulate the same region in the S form. Although these two datasets cannot be used to quantitatively analyze the differences between the two piRNA libraries, it is of note that this region represented one of the few regions that had a very large decrease in piRNA enrichment, having almost a 3 fold difference in total piRNA enrichment. These data would be in correspondence with genetic divergence being isolated in specific regions of the genome.

3.4.3. Likely biological functions of genes enriched in piRNAs

Genes identified to be enriched in piRNAs appear to have similar functions within the organism. Gene ontology analysis shows that many of the genes are development related. There was also noted enrichment in germ cell development. Both of these processes are not all that surprising when compared with current knowledge of genes associated with piRNAs. The piRNA pathway has been established as important to germline stability, and development

[153,155,162,163,396]. Implications in reproductive development also reflect previously identified piRNA roles, as the pathway is essential to fertility [167-169] and gonad development. These data suggest a similar role in both germline and reproductive organ development in *An. gambiae*. The identification of a cluster of three genes that follow similar expression profiles further hints at the importance of some piRNA related genes in embryonic development. These genes only become up-regulated after blood meal intake and continue through at least embryonic development stages. These genes could be candidate targets for vector control, using knowledge about their potential importance in embryogenesis only after blood meals as a focal point.

3.5. Methods

3.5.1. Purification of small RNAs and library construction

Christopher's Stage III ovaries were dissected from 25 hour gravid females of the Mali strain (M form) and Zanu strain (S form) of *A. gambiae* and preserved in Trizol to prevent RNA degradation. Total RNA was extracted from ovaries preserved in Trizol reagent of approximately 40-50 individual M and S mosquitoes. RNA was precipitated using a standard phenol-chloroform extraction, solubilized in RNase-free water and stored at -80°C. Total RNA was isolated for both forms and Illumina sequencing was performed on the RNAs ranging from 20-32 nucleotides in length in Fasteris, Inc. Libraries containing RNAs ranging in size from 20-32 nt were created from a single RNA sample for the M and S forms. The two libraries had a sharp peak at 22 nt, with a secondary broad peak that was greatest at 26-27 nt.

3.5.2. Library processing

The two small RNA datasets were edited to best represent the piRNAs in the library. Repetitive sequences were initially removed, leaving only unique sequences. After identical sentences were filtered out, the library was pruned to only contain RNA sequences ranging in size from 24-30 nt, the predicted size of piRNAs for both *D. melanogaster* and *A. aegypti*. The NucBase software was run to determine the number of times each RNA sequence mapped within the *A. gambiae* genome[401]. Any non-unique sequences were discarded from the datasets, resulting in a library of non-repetitive, unique mapping RNAs that would be used for subsequent analysis.

3.5.3. Mapping sequences to reference genomes

The AgamP3.7 assembly of the *A. gambiae* PEST genome was used for the mapping of two small RNA libraries using a short-read mapper, NucBase [401]. Default settings of zero mismatches and a minimum seed size of zero (the minimum number of nucleotides required to create a match) were used within the program. NucBase counts the number of times a specific RNA sequence maps to the reference genome, allowing the user to filter out all repetitive sequences from the library. Unique mapping piRNAs were parsed into a secondary, final library that was used to discover the location of piRNA clusters.

3.5.4. TE, Gene ontology (GO), and expression data analysis

All TE analysis was performed by using a Repeat Masker (girinst.org) search against all available annotated elements. We elected to use the "all TE sources" option when performing RepeatMasker searches, as the complete annotation of TEs in *Anopheles gambiae* is still lacking. piRNAs were localized to gene exon DNA sequences extracted using BioMART [375,411,412]. Genes that had more than 150 piRNAs mapping to them were used in gene ontology (GO) term analysis, as well as differential expression analysis. Gene ontology terms were identified using DAVID v6.7 [413,414]. The default settings in DAVID were used with the addition of Bonferroni correction for P-values. The same subset of genes was used when looking at expression assays. An expression browser, available through VectorBase, was used to identify expression patterns of genes in different assay conditions [415]. The ten highest expression differentials were taken for each node that contained a gene from the identified subset.

3.5.5. PIWI antibody generation and immunostaining

Polyclonal antibodies were generated for the proteins transcribed from the *A. gambiae* genes AGAP009509 (PIWI1) and AGAP011204 (PIWI2). 16 amino acid synthetic peptide sequences were generated for the translated protein sequences of the two genes and submitted to Eurogentec to create polyclonal antibodies.

We used a slightly modified version of an immunostaining protocol [353,372] adapted to immunolocalize antibodies to ovarian nurse cells. Polytene chromosome squashes were obtained from half gravid females that were dissected 25 hours post-blood meal. Ovaries were divided and placed in a fixative solution (47% water, 45% acetic acid, and 8% formaldehyde) and

individual follicles were separated. Old fixative solution was removed and replaced with a fresh drop. The follicles were squashed under a coverslip and frozen in liquid nitrogen. The coverslip was immediately removed, and the slide was placed in 70% cold ethanol (-20°C). Slides were washed in PBS saline buffer (0.1% Nonidet P40 in PBS) followed by blocking solution (1% BSA in PBS) for 20 minutes. Primary polyclonal antibodies raised in rat (AGAP009509) and guinea pig (AGAP011204) were left on the slide overnight at 4°C, and replaced with secondary labeled antibodies for two hours. Slides were washed using PBS and Prolong Anti-fade with DAPI was added to stain DNA and preserve fluorescence. Slides were then viewed under an ACCORD™ PLUS Automated Scanning System. Images taken from the scanning system were then processed through Adobe Photoshop.

4. Chapter Summary

4.1. General review

The goal of these studies is to better characterize the heterochromatin of *Anopheles gambiae s.s.* through cytogenetic and bioinformatics approaches. Through chromatin characterization, we attempt to understand how the organism compensates for an increase in genome size. We take *Drosophila melanogaster* and *Aedes aegypti* as insect relatives with more extreme genome sizes. The compact *D. melanogaster* has a very strict organization of repetitive elements, confining them to peri-centromeric and sub-telomeric regions of the chromosome. *Ae. aegypti*, on the other hand, has its large transposable element load dispersed throughout the genome with no noticeable enrichment in the peri-centromeric regions of the chromosomes.

Anopheles gambiae represents an intermediate in both genome size and transposable element load.

The compartmentalization of heterochromatin has important implications in genome function, as these regions provide a very different environment for gene regulation and genome stability. The expansion of these heterochromatic sequences outside of a single centralized location can potentially lead to a change in genome dynamics.

However, in order to understand how genome size may reflect upon heterochromatin organization, it is important to fully characterize the genomic landscape. Here, we identify a trend where, as genome size increases, heterochromatin becomes less centralized in order to compensate for increased repetitive element load.

4.2. Review of chapter 2

Traditional cytogenetic techniques are both tedious and difficult to master. Many genomic assemblies rely on physical mapping of scaffolds to assemble available sequences. We have developed protocols that help facilitate genome assembly and annotation. We have developed an efficient method for producing high-resolution chromosomal squashes using polytene chromosomes. This technique increases efficiency, while producing slides with better banding quality for physical mapping purposes. The chromosomal complement was also given genomic coordinates that link DNA sequences to specific regions along the chromosome. 302 unique markers available on VectorBase were used to assign genomic coordinates to loci across the genome. These coordinates are helpful in quickly identifying approximate genetic coordinates of regions of interest found on the chromosome.

This technique was paired with a protocol that incorporates automated fluorescent *in situ* hybridization (FISH) method to efficiently physically map labeled DNA probes. The protocol allows for the automated hybridization of labeled probes to many slides that then can be automatically visualized using an automated stage and software attached to a fluorescence microscope. These techniques add to the cytogenetic toolbox, providing more efficient techniques to produce chromosomal squash preparations used in many physical mapping experiments.

To understand how heterochromatin is organized within *An. gambiae*, we characterized the chromatin of the *Anopheles gambiae* mosquito, assessing the genomic landscape using molecular features including repetitive elements and genic content. Boundaries between the heterochromatin and euchromatin were identified using FISH and immunolocalization. 82 labeled PCR generated DNA probes thought to be near the heterochromatin-euchromatin boundary were hybridized to chromosomal squash preparations and mapped. Heterochromatin protein 1 (HP1) and lamin antibodies were used to detect chromosome regions associated with heterochromatin. A total of 16.6 Mb of the mapped genome assembly was determined to be heterochromatin, with 1.5 Mb being diffuse intercalary heterochromatin, 2.9 Mb being condensed intercalary heterochromatin, and 12.2 Mb being peri-centromeric heterochromatin. These regions were characterized by high abundance of repetitive elements, with a low genic count (232 or ~1.8% of the 13,000 predicted genes in *An. gambiae*).

Finally, we developed a model that can effectively predict a DNA sequence's chromatin type based on molecular features. The model incorporates features defined by the previous heterochromatin study that include repetitive elements and genic content. However, the genome was divided into 100 kb windows to potentially discover novel islands of heterochromatin that

were previously classified as euchromatin. We have identified seven new heterochromatic regions, six being diffuse heterochromatin and one compact heterochromatin, at an 80% probability rate. Morphologically, these regions appear to be euchromatic, but exhibit features including high repeat abundance and low gene content that are characteristic of heterochromatin.

Through morphological and genetic analyses, we see compartmentalization of heterochromatin in regions of euchromatin. These intercalary heterochromatin regions mostly resemble peri-centromeric heterochromatin, and serve as secondary partitions housing repetitive elements. These data would suggest de-centralization of heterochromatin from a single or a few particular regions within the genome, toward an even dispersion of these repetitive elements across the genome. In *An. gambiae*, we still see organization of these elements into distinct intercalary clusters.

4.3. Review of chapter 3

We found that the piRNA pathway is present and active in *Anopheles gambiae* mosquitoes. Small RNAs isolated from ovaries of M and S forms resulted in a large population of 24-30 nucleotide small RNAs. These RNAs exhibit similar motifs to previously identified piRNAs, having 1-U and 10T nucleotide biases, as well as a 10 basepair 5' overlap from complementary piRNAs. We have identified piRNA clusters, likely production sites for piRNA precursors in both the M and S forms of *An. gambiae*, and conclude that spatially, piRNA clusters are well conserved. Most clusters are present in both the M and S forms, with these clusters exhibiting similar piRNA enrichment profiles. Genes enriched in piRNAs are involved in developmental and sex-bias differentiation, which support implications that the piRNA pathway is important in germline development and fertility. Particularly, a group of three genes

has been identified that could have an important role in post-blood feed development of the embryo.

piRNA mapping suggests genome size and content are related to the expansion of piRNA clusters into intercalary and euchromatic regions in the genome. Many of the highest piRNA producing regions are still heterochromatic in nature, even though only 12% of the total piRNA sequences map to TEs. As these heterochromatic regions confine much of the transposable element sequences found in the *An gambiae* genome, it is not surprising that we see a similar expansion of clusters into intercalary heterochromatin. These data hint at the possibility that the increasing presence of transposable elements within the genome may support the evolution of function within the piRNA pathway, where long, piRNA clusters begin to incorporate genes and other non-TE related sequences into their long, single-stranded piRNA precursors. These non-TE related sequences would provide the target necessary for the Piwi protein-piRNA complex to actively seek out complementary sequences.

Bibliography

1 WHO. Malaria, Fact Sheets. (2011).
2 WHO. World Malaria Report 2012. 195 (2012).
3 Guerra, C. A., Snow, R. W. & Hay, S. I. Mapping the global extent of malaria in 2005. *Trends in parasitology* **22**, 353-358, doi:10.1016/j.pt.2006.06.006 (2006).
4 Guerra, C. A., Snow, R. W. & Hay, S. I. Defining the global spatial limits of malaria transmission in 2005. *Advances in parasitology* **62**, 157-179, doi:10.1016/S0065-308X(05)62005-2 (2006).
5 Guerra, C. A. *et al.* The international limits and population at risk of Plasmodium vivax transmission in 2009. *PLoS neglected tropical diseases* **4**, e774, doi:10.1371/journal.pntd.0000774 (2010).
6 *Malaria: Obstacles and Opportunities*. (The National Academies Press, 1991).
7 Bronner, U., Divis, P. C., Farnert, A. & Singh, B. Swedish traveller with Plasmodium knowlesi malaria after visiting Malaysian Borneo. *Malaria journal* **8**, 15, doi:10.1186/1475-2875-8-15 (2009).
8 Daneshvar, C. *et al.* Clinical and laboratory features of human Plasmodium knowlesi infection. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* **49**, 852-860, doi:10.1086/605439 (2009).
9 RollBackMalaria. *Frequently Asked Questions About Malaria*, <http://www.rbm.who.int/malariaFAQ.html> (2012).
10 Guerra, C. A. *et al.* The limits and intensity of Plasmodium falciparum transmission: implications for malaria control and elimination worldwide. *PLoS medicine* **5**, e38, doi:10.1371/journal.pmed.0050038 (2008).
11 Sutherland, C. J. *et al.* Two nonrecombining sympatric forms of the human malaria parasite Plasmodium ovale occur globally. *The Journal of infectious diseases* **201**, 1544-1550, doi:10.1086/652240 (2010).
12 Collins, W. E. & Jeffery, G. M. Plasmodium malariae: parasite and disease. *Clinical microbiology reviews* **20**, 579-592, doi:10.1128/CMR.00027-07 (2007).
13 Stothard, J. R. Essential Malariology, 4th edition. David A. Warrell & Herbert M. Gilles (editors). London: Arnold, 2002. xii + 348 pp. Price £65.00. ISBN 0-340-74064-7. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **97**, 199, doi:10.1016/s0035-9203(03)90118-0 (2003).
14 Kiszewski, A. *et al.* A global index representing the stability of malaria transmission. *The American journal of tropical medicine and hygiene* **70**, 486-498 (2004).
15 Hay, S. I., Guerra, C. A., Tatem, A. J., Noor, A. M. & Snow, R. W. The global distribution and population at risk of malaria: past, present, and future. *The Lancet infectious diseases* **4**, 327-336, doi:10.1016/S1473-3099(04)01043-6 (2004).
16 WHO. Global malaria control and elimination: report of a technical review. 54 (2008).
17 Carter, R. & Mendis, K. N. Evolutionary and historical aspects of the burden of malaria. *Clinical microbiology reviews* **15**, 564-594 (2002).
18 Raghavendra, K., Barik, T. K., Reddy, B. P., Sharma, P. & Dash, A. P. Malaria vector control: from past to future. *Parasitology research* **108**, 757-779, doi:10.1007/s00436-010-2232-0 (2011).

19 Pennetier, C. *et al.* Mixture for controlling insecticide-resistant malaria vectors. *Emerging infectious diseases* **14**, 1707-1714, doi:10.3201/eid1411.071575 (2008).
20 Kroeger, A., Horstick, O., Riedl, C., Kaiser, A. & Becker, N. The potential for malaria control with the biological larvicide Bacillus thuringiensis israelensis (Bti) in Peru and Ecuador. *Acta tropica* **60**, 47-57 (1995).
21 Mittal, P. K. Biolarvicides in vector control: challenges and prospects. *Journal of vector borne diseases* **40**, 20-32 (2003).
22 Couch, J. N., Umphlett, C. J. & Iyengar, M. O. Infection experiments with a fungus (coelomomyces) which kills malarial mosquitoes. *Science* **158**, 526, doi:10.1126/science.158.3800.526 (1967).
23 Petersen, J. J. & Willis, O. R. Diximermis peterseni (Nematoda: Mermithidae): a potential biocontrol agent of Anopheles mosquito larvae. *Journal of invertebrate pathology* **24**, 20-23 (1974).
24 Perez Pacheco, R., Santamarina Mijares, A., Martinez, S. H. & Flores Ambrosio, G. The susceptibility of Anopheles pseudopunctipennis larvae to parasitism by the nematode Romanomermis iyengari (Nematoda: Mermithidae), the state of Oaxaca, Mexico. *Revista cubana de medicina tropical* **50**, 199-202 (1998).
25 Schenker, W., Maier, W. A. & Seitz, H. M. The effects of Nosema algerae on the development of Plasmodium yoelii nigeriensis in Anopheles stephensi. *Parasitology research* **78**, 56-59 (1992).
26 Wijesinghe, W. M. G. S., Wickramasinghe, M. B., Kusumawathie, P. H. D., Jayasooriya, G. A. J. S. K. & De Silva, B. G. D. N. K. Studies on the efficacy of Toxorhynchites larvae and three larvivorous fish species for the control of Aedes larval populations in water-storage tanks in the Matale district of Sri Lanka. *Dengue Bulletin* **33**, 140-147 (2009).
27 Dyck, V. A., Hendrichs, J. & Robinson, A. S. *The Sterile Insect Technique: Principles and Practice in Area-Wide Integrated Pest Management*. (Springer, 2005).
28 Weidhaas, D. E., Schmidt, C. H. & Seabrook, E. L. Field studies on the release of sterile males for the control of Anopheles quadrimaculatus. *Journal of the American Mosquito Control Association* **22**, 283-291 (1962).
29 Morlan, H., McCray, E. & Kilpatrick, J. Field tests with sexually sterile males for control of Aedes aegypti. *Mosquito News* **22**, 295-300 (1962).
30 Patterson, R. S. *et al.* Use of radiosterilized males to control indigenous populations of Culex pipiens quinquefasciatus Say: laboratory and field studies. *Journal of the American Mosquito Control Association* **35**, 1-7 (1975).
31 Reisen, W., Milby, M., Asman, S., Bock, M. & Meyer, R. Attempted suppression of a semi-isolated Culex tarsalis population by the release of irradiated males: a second experiment using males from a recently colonized strain. *Mosquito News* **42**, 565-575 (1982).
32 Thomas, D. D., Donnelly, C. A., Wood, R. J. & Alphey, L. S. Insect population control using a dominant, repressible, lethal genetic system. *Science* **287**, 2474-2476 (2000).
33 Wilke, A. B., Gomes Ade, C., Natal, D. & Marrelli, M. T. Control of vector populations using genetically modified mosquitoes. *Revista de saude publica* **43**, 869-874 (2009).
34 Robinson, A. S. Genetic sexing strains in medfly, Ceratitis capitata, sterile insect technique programmes. *Genetica* **116**, 5-13 (2002).

35 Black, W. C. t., Alphey, L. & James, A. A. Why RIDL is not SIT. *Trends in parasitology* **27**, 362-370, doi:10.1016/j.pt.2011.04.004 (2011).
36 Coluzzi, M., Sabatini, A., Petrarca, V. & Angela Di Deco, M. Behavioural divergences between mosquitoes with different inversion karyotypes in polymorphic populations of the Anopheles gambiae complex. *Nature* **266**, 832-833 (1977).
37 Marshall, J. M. & Taylor, C. E. Malaria control with transgenic mosquitoes. *PLoS medicine* **6**, e20, doi:10.1371/journal.pmed.1000020 (2009).
38 Blandin, S. *et al.* Complement-like protein TEP1 is a determinant of vectorial capacity in the malaria vector Anopheles gambiae. *Cell* **116**, 661-670 (2004).
39 Mitri, C. *et al.* Fine pathogen discrimination within the APL1 gene family protects Anopheles gambiae against human and rodent malaria species. *PLoS pathogens* **5**, e1000576, doi:10.1371/journal.ppat.1000576 (2009).
40 Povelones, M., Waterhouse, R. M., Kafatos, F. C. & Christophides, G. K. Leucine-rich repeat protein complex activates mosquito complement in defense against Plasmodium parasites. *Science* **324**, 258-261, doi:10.1126/science.1171400 (2009).
41 Dong, Y. & Dimopoulos, G. Anopheles fibrinogen-related proteins provide expanded pattern recognition capacity against bacteria and malaria parasites. *The Journal of biological chemistry* **284**, 9835-9844, doi:10.1074/jbc.M807084200 (2009).
42 Coetzee, M., Craig, M. & le Sueur, D. Distribution of African malaria mosquitoes belonging to the Anopheles gambiae complex. *Parasitol Today* **16**, 74-77 (2000).
43 Lindsay, S. W., Parson, L. & Thomas, C. J. Mapping the ranges and relative abundance of the two principal African malaria vectors, Anopheles gambiae sensu stricto and An. arabiensis, using climate data. *Proceedings. Biological sciences / The Royal Society* **265**, 847-854, doi:10.1098/rspb.1998.0369 (1998).
44 White, G. B. Anopheles gambiae complex and disease transmission in Africa. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **68**, 278-301 (1974).
45 White, G. B. Anopheles-Bwambae Sp-N, a Malaria Vector in the Semliki Valley, Uganda, and Its Relationships with Other Sibling Species of the an-Gambiae Complex (Diptera, Culicidae). *Syst Entomol* **10**, 501-522, doi:DOI 10.1111/j.1365-3113.1985.tb00155.x (1985).
46 Hunt, R. H., Coetzee, M. & Fettene, M. The Anopheles gambiae complex: a new species from Ethiopia. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **92**, 231-235 (1998).
47 White, G. B. Anopheles gambiae complex and disease transmission in Africa. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **68**, 278-301 (1974).
48 Dekker, T. & Takken, W. Differential responses of mosquito sibling species Anopheles arabiensis and An. quadriannulatus to carbon dioxide, a man or a calf. *Medical and veterinary entomology* **12**, 136-140 (1998).
49 Prior, A. & Torr, S. J. Host selection by Anopheles arabiensis and An. quadriannulatus feeding on cattle in Zimbabwe. *Medical and veterinary entomology* **16**, 207-213 (2002).
50 della Torre, A. *et al.* Molecular evidence of incipient speciation within Anopheles gambiae s.s. in West Africa. *Insect molecular biology* **10**, 9-18 (2001).
51 Turner, T., Hahn, M. & Nuzhdin, S. Genomic islands of speciation in Anopheles gambiae. *PLoS biology* **3**, e285 (2005).
52 Zhimulev, I. F. Morphology and structure of polytene chromosomes. *Advances in genetics* **34**, 1-497 (1996).

53 Coluzzi, M. Osservazioni comparative sul cromosoma X nella specie A e B del complesso Anopheles gambiae. *Rend Accad Naz Lincei* **40**, 671-678 (1966).
54 Coluzzi, M. a. & Sabatini, A. Cytogenetic observations on species A and B of the Anopheles gambiae complex. *Parassitologia* **9**, 73-88 (1967).
55 Coluzzi, M. Cromosomi politenici delle cellule nutrici ovariche nel complesso gambiae del genere Anopheles. *Parassitologia* **10**, 179-184 (1968).
56 Coluzzi, M., Sabatini, A., della Torre, A., Di Deco, M. A. & Petrarca, V. A polytene chromosome analysis of the Anopheles gambiae species complex. *Science* **298**, 1415-1418, doi:10.1126/science.1077769 (2002).
57 Coluzzi, M. The polytene chromosomes of Anopheles superpictus and relationships with Anopheles stephensi. *Parassitologia* **12**, 101-112 (1970).
58 Zheng, L. *et al.* Low-resolution genome map of the malaria mosquito Anopheles gambiae. *Proceedings of the National Academy of Sciences of the United States of America* **88**, 11187-11191 (1991).
59 Zheng, L., Benedict, M. Q., Cornel, A. J., Collins, F. H. & Kafatos, F. C. An integrated genetic map of the African human malaria vector mosquito, Anopheles gambiae. *Genetics* **143**, 941-952 (1996).
60 Dimopoulos, G. *et al.* Integrated genetic map of Anopheles gambiae: use of RAPD polymorphisms for genetic, cytogenetic and STS landmarks. *Genetics* **143**, 953-960 (1996).
61 della Torre, A., Favia, G., Mariotti, G., Coluzzi, M. & Mathiopoulos, K. D. Physical map of the malaria vector Anopheles gambiae. *Genetics* **143**, 1307-1311 (1996).
62 Holt, R. *et al.* The genome sequence of the malaria mosquito Anopheles gambiae. *Science* **298**, 129 - 149 (2002).
63 Sharakhova, M. *et al.* Update of the Anopheles gambiae PEST genome assembly. *Genome biology* **8**, R5 (2007).
64 Lawniczak, M. K. *et al.* Widespread divergence between incipient Anopheles gambiae species revealed by whole genome sequences. *Science* **330**, 512-514, doi:10.1126/science.1195755 (2010).
65 Neafsey, D. E. *et al.* The evolution of the Anopheles 16 genomes project. *G3 (Bethesda)* **3**, 1191-1194, doi:10.1534/g3.113.006247 (2013).
66 Kosak, S. T. & Groudine, M. Gene order and dynamic domains. *Science* **306**, 644-647, doi:10.1126/science.1103864 (2004).
67 Heitz, E. Das Heterochromatin der Moose. *Jahrb Wiss Botanik* **69**, 762-818 (1928).
68 Ohno, S. So much "junk" DNA in our genome. *Brookhaven symposia in biology* **23**, 366-370 (1972).
69 Grewal, S. I. & Jia, S. Heterochromatin revisited. *Nature reviews. Genetics* **8**, 35-46, doi:10.1038/nrg2008 (2007).
70 Lohe, A. R., Hilliker, A. J. & Roberts, P. A. Mapping simple repeated DNA sequences in heterochromatin of Drosophila melanogaster. *Genetics* **134**, 1149-1174 (1993).
71 Dimitri, P. Constitutive heterochromatin and transposable elements in Drosophila melanogaster. *Genetica* **100**, 85-93 (1997).
72 Trojer, P. & Reinberg, D. Facultative heterochromatin: is there a distinctive molecular signature? *Molecular cell* **28**, 1-13, doi:10.1016/j.molcel.2007.09.011 (2007).
73 Jenuwein, T. & Allis, C. D. Translating the histone code. *Science* **293**, 1074-1080, doi:10.1126/science.1063127 (2001).

74 Allfrey, V. G., Faulkner, R. & Mirsky, A. E. Acetylation and Methylation of Histones and Their Possible Role in the Regulation of Rna Synthesis. *Proceedings of the National Academy of Sciences of the United States of America* **51**, 786-794 (1964).
75 Pogo, B. G., Allfrey, V. G. & Mirsky, A. E. RNA synthesis and histone acetylation during the course of gene activation in lymphocytes. *Proceedings of the National Academy of Sciences of the United States of America* **55**, 805-812 (1966).
76 Hebbes, T. R., Thorne, A. W. & Crane-Robinson, C. A direct link between core histone acetylation and transcriptionally active chromatin. *The EMBO journal* **7**, 1395-1402 (1988).
77 Vettese-Dadey, M. *et al.* Acetylation of histone H4 plays a primary role in enhancing transcription factor binding to nucleosomal DNA in vitro. *The EMBO journal* **15**, 2508-2518 (1996).
78 Taverna, S. D., Li, H., Ruthenburg, A. J., Allis, C. D. & Patel, D. J. How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers. *Nature structural & molecular biology* **14**, 1025-1040, doi:10.1038/nsmb1338 (2007).
79 Fanti, L., Giovinazzo, G., Berloco, M. & Pimpinelli, S. The heterochromatin protein 1 prevents telomere fusions in Drosophila. *Molecular cell* **2**, 527-538 (1998).
80 Cheutin, T. *et al.* Maintenance of stable heterochromatin domains by dynamic HP1 binding. *Science* **299**, 721-725, doi:10.1126/science.1078572 (2003).
81 Nielsen, A. L. *et al.* Selective interaction between the chromatin-remodeling factor BRG1 and the heterochromatin-associated protein HP1alpha. *The EMBO journal* **21**, 5797-5806 (2002).
82 James, T. & Elgin, S. Identification of a nonhistone chromosomal protein associated with heterochromatin in Drosophila melanogaster and its gene. *Molecular and cellular biology* **6**, 3862 - 3872 (1986).
83 Muller, H. J. & Altenburg, E. The Frequency of Translocations Produced by X-Rays in Drosophila. *Genetics* **15**, 283-311 (1930).
84 Wustmann, G., Szidonya, J., Taubert, H. & Reuter, G. The genetics of position-effect variegation modifying loci in Drosophila melanogaster. *Molecular & general genetics : MGG* **217**, 520-527 (1989).
85 Wallrath, L. L. & Elgin, S. C. Position effect variegation in Drosophila is associated with an altered chromatin structure. *Genes & development* **9**, 1263-1277 (1995).
86 Fanti, L., Berloco, M., Piacentini, L. & Pimpinelli, S. Chromosomal distribution of heterochromatin protein 1 (HP1) in Drosophila: a cytological map of euchromatic HP1 binding sites. *Genetica* **117**, 135-147 (2003).
87 James, T. *et al.* Distribution patterns of HP1, a heterochromatin-associated nonhistone chromosomal protein of Drosophila. *European journal of cell biology* **50**, 170 - 180 (1989).
88 Hwang, K. K., Eissenberg, J. C. & Worman, H. J. Transcriptional repression of euchromatic genes by Drosophila heterochromatin protein 1 and histone modifiers. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 11423-11427, doi:10.1073/pnas.211303598 (2001).
89 Belyaeva, E. S., Andreyeva, E. N., Belyakin, S. N., Volkova, E. I. & Zhimulev, I. F. Intercalary heterochromatin in polytene chromosomes of Drosophila melanogaster. *Chromosoma* **117**, 411-418, doi:10.1007/s00412-008-0163-7 (2008).

90 Bridges, C. Salivary chromosome map with a key to the banding of the chromosomes of Drosophila melanogaster. *Journal of Heredity* **26**, 60-64 (1935).
91 Hochstrasser, M. & Sedat, J. Three-dimensional organization of Drosophila melanogaster interphase nuclei. I. Tissue-specific aspects of polytene nuclear architecture. *The Journal of cell biology* **104**, 1455 - 1470 (1987).
92 Sharakhov, I., Sharakhova, M., Mbogo, C., Koekemoer, L. & Yan, G. Linear and spatial organization of polytene chromosomes of the African malaria mosquito Anopheles funestus. *Genetics* **159**, 211 - 218 (2001).
93 Stegnii, V. & Sharakhova, M. Systemic reorganization of the architechtonics of polytene chromosomes in onto- and phylogenesis of malaria mosquitoes. Structural features regional of chromosomal adhesion to the nuclear membrane. *Genetika* **27**, 828 - 835 (1991).
94 McClintock, B. The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences of the United States of America* **36**, 344-355 (1950).
95 Honeybee Genome Sequencing Consortium, G. M. Insights into social insects from the genome of the honeybee Apis mellifera. *Nature* **443**, 931-949, doi:10.1038/nature05260 (10/26/2006).
96 Consortium, T. I. S. G. The genome of a lepidopteran model insect, the silkworm Bombyx mori. *Insect biochemistry and molecular biology* **38**, 1036-1045, doi:10.1016/j.ibmb.2008.11.004 (2008).
97 Osanai-Futahashi, M., Suetsugu, Y., Mita, K. & Fujiwara, H. Genome-wide screening and characterization of transposable elements and their distribution analysis in the silkworm, Bombyx mori. *Insect biochemistry and molecular biology* **38**, 1046-1057 (2008).
98 Schnable, P. S. *et al.* The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112-1115, doi:10.1126/science.1178534 (2009).
99 Nene, V. *et al.* Genome sequence of Aedes aegypti, a major arbovirus vector. *Science* **316**, 1718-1723, doi:10.1126/science.1138878 (2007).
100 Clark, A. G. *et al.* Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **450**, 203-218, doi:10.1038/nature06341 (2007).
101 Capy, P. Classification and nomenclature of retrotransposable elements. *Cytogenetic and genome research* **110**, 457-461, doi:10.1159/000084978 (2005).
102 Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921, doi:10.1038/35057062 (2001).
103 Fernandez-Medina, R. D., Struchiner, C. J. & Ribeiro, J. M. Novel transposable elements from Anopheles gambiae. *BMC genomics* **12**, 260, doi:10.1186/1471-2164-12-260 (2011).
104 Biemont, C. & Vieira, C. Genetics: junk DNA as an evolutionary force. *Nature* **443**, 521-524, doi:10.1038/443521a (2006).
105 Miller, W. J. & Capy, P. Mobile genetic elements as natural tools for genome evolution. *Methods Mol Biol* **260**, 1-20, doi:10.1385/1-59259-755-6:001 (2004).
106 Orgel, L. E. & Crick, F. H. Selfish DNA: the ultimate parasite. *Nature* **284**, 604-607 (1980).
107 Doolittle, W. F. & Sapienza, C. Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284**, 601-603 (1980).
108 Dawkins, R. *The Selfish Gene*. (Oxford University Press, 1976).

109 Lim, J. K. & Simmons, M. J. Gross chromosome rearrangements mediated by transposable elements in Drosophila melanogaster. *BioEssays : news and reviews in molecular, cellular and developmental biology* **16**, 269-275, doi:10.1002/bies.950160410 (1994).
110 Mieczkowski, P. A., Lemoine, F. J. & Petes, T. D. Recombination between retrotransposons as a source of chromosome rearrangements in the yeast Saccharomyces cerevisiae. *DNA repair* **5**, 1010-1020, doi:10.1016/j.dnarep.2006.05.027 (2006).
111 Delprat, A., Negre, B., Puig, M. & Ruiz, A. The transposon Galileo generates natural chromosomal inversions in Drosophila by ectopic recombination. *PloS one* **4**, e7883, doi:10.1371/journal.pone.0007883 (2009).
112 Sen, S. K. *et al.* Human genomic deletions mediated by recombination between Alu elements. *American journal of human genetics* **79**, 41-53, doi:10.1086/504600 (2006).
113 Lim, J. K. Intrachromosomal rearrangements mediated by hobo transposons in Drosophila melanogaster. *Proceedings of the National Academy of Sciences of the United States of America* **85**, 9153-9157 (1988).
114 Mathiopoulos, K. D., della Torre, A., Predazzi, V., Petrarca, V. & Coluzzi, M. Cloning of inversion breakpoints in the Anopheles gambiae complex traces a transposable element at the inversion junction. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 12444-12449 (1998).
115 Kidwell, M. G., Kidwell, J. F. & Sved, J. A. Hybrid Dysgenesis in DROSOPHILA MELANOGASTER: A Syndrome of Aberrant Traits Including Mutation, Sterility and Male Recombination. *Genetics* **86**, 813-833 (1977).
116 Kidwell, M. G. Reciprocal differences in female recombination associated with hybrid dysgenesis in Drosophila melanogaster. *Genetical research* **30**, 77-88 (1977).
117 Rubin, G. M., Kidwell, M. G. & Bingham, P. M. The molecular basis of P-M hybrid dysgenesis: the nature of induced mutations. *Cell* **29**, 987-994 (1982).
118 Britten, R. J. & Davidson, E. H. Gene regulation for higher cells: a theory. *Science* **165**, 349-357 (1969).
119 Britten, R. J. Mobile elements inserted in the distant past have taken on important functions. *Gene* **205**, 177-182 (1997).
120 McClintock, B. Chromosome organization and genic expression. *Cold Spring Harbor symposia on quantitative biology* **16**, 13-47 (1951).
121 Maksakova, I. A. *et al.* Retroviral elements and their hosts: insertional mutagenesis in the mouse germ line. *PLoS genetics* **2**, e2, doi:10.1371/journal.pgen.0020002 (2006).
122 Conley, A. B., Piriyapongsa, J. & Jordan, I. K. Retroviral promoters in the human genome. *Bioinformatics* **24**, 1563-1567, doi:10.1093/bioinformatics/btn243 (2008).
123 van de Lagemaat, L. N., Landry, J. R., Mager, D. L. & Medstrand, P. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends in genetics : TIG* **19**, 530-536, doi:10.1016/j.tig.2003.08.004 (2003).
124 Pardue, M. L. & DeBaryshe, P. G. Retrotransposons provide an evolutionarily robust non-telomerase mechanism to maintain telomeres. *Annual review of genetics* **37**, 485-511, doi:10.1146/annurev.genet.38.072902.093115 (2003).
125 Wong, L. H. & Choo, K. H. Evolutionary dynamics of transposable elements at the centromere. *Trends in genetics : TIG* **20**, 611-616, doi:10.1016/j.tig.2004.09.011 (2004).

126 Pasyukova, E. G., Nuzhdin, S. V., Morozova, T. V. & Mackay, T. F. Accumulation of transposable elements in the genome of Drosophila melanogaster is associated with a decrease in fitness. *The Journal of heredity* **95**, 284-290, doi:10.1093/jhered/esh050 (2004).
127 Mackay, T. F., Lyman, R. F. & Jackson, M. S. Effects of P element insertions on quantitative traits in Drosophila melanogaster. *Genetics* **130**, 315-332 (1992).
128 Charlesworth, B., Jarne, P. & Assimacopoulos, S. The distribution of transposable elements within and between chromosomes in a population of Drosophila melanogaster. III. Element abundances in heterochromatin. *Genetical research* **64**, 183-197 (1994).
129 Gao, L., McCarthy, E. M., Ganko, E. W. & McDonald, J. F. Evolutionary history of Oryza sativa LTR retrotransposons: a preliminary survey of the rice genome sequences. *BMC genomics* **5**, 18, doi:10.1186/1471-2164-5-18 (2004).
130 Bartolome, C., Maside, X. & Charlesworth, B. On the abundance and distribution of transposable elements in the genome of Drosophila melanogaster. *Molecular biology and evolution* **19**, 926-937 (2002).
131 Duret, L., Marais, G. & Biemont, C. Transposons but not retrotransposons are located preferentially in regions of high recombination rate in Caenorhabditis elegans. *Genetics* **156**, 1661-1669 (2000).
132 Dimitri, P. & Junakovic, N. Revising the selfish DNA hypothesis: new evidence on accumulation of transposable elements in heterochromatin. *Trends in genetics : TIG* **15**, 123-124 (1999).
133 Langley, C. H., Montgomery, E., Hudson, R., Kaplan, N. & Charlesworth, B. On the role of unequal exchange in the containment of transposable element copy number. *Genetical research* **52**, 223-235 (1988).
134 Charlesworth, B. & Langley, C. H. The population genetics of Drosophila transposable elements. *Annual review of genetics* **23**, 251-287, doi:10.1146/annurev.ge.23.120189.001343 (1989).
135 Finnegan, D. J. Transposable elements. *Current opinion in genetics & development* **2**, 861-867 (1992).
136 Nuzhdin, S. V. Sure facts, speculations, and open questions about the evolution of transposable element copy number. *Genetica* **107**, 129-137 (1999).
137 Jurka, J. Repeats in genomic DNA: mining and meaning. *Current opinion in structural biology* **8**, 333-337 (1998).
138 Jurka, J. Repbase update: a database and an electronic journal of repetitive elements. *Trends in genetics : TIG* **16**, 418-420 (2000).
139 Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research* **110**, 462-467, doi:10.1159/000084979 (2005).
140 Carmell, M. A., Xuan, Z., Zhang, M. Q. & Hannon, G. J. The Argonaute family: tentacles that reach into RNAi, developmental control, stem cell maintenance, and tumorigenesis. *Genes & development* **16**, 2733-2742, doi:10.1101/gad.1026102 (2002).
141 Cerutti, H. & Casas-Mollano, J. A. On the origin and functions of RNA-mediated silencing: from protists to man. *Current genetics* **50**, 81-99, doi:10.1007/s00294-006-0078-x (2006).
142 Cox, D. N., Chao, A. & Lin, H. piwi encodes a nucleoplasmic factor whose activity modulates the number and division rate of germline stem cells. *Development* **127**, 503-514 (2000).

143 Harris, A. N. & Macdonald, P. M. Aubergine encodes a Drosophila polar granule component required for pole cell formation and related to eIF2C. *Development* **128**, 2823-2832 (2001).
144 Brennecke, J. *et al.* Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell* **128**, 1089-1103, doi:10.1016/j.cell.2007.01.043 (2007).
145 Kidwell, M. G. & Novy, J. B. Hybrid Dysgenesis in DROSOPHILA MELANOGASTER: Sterility Resulting from Gonadal Dysgenesis in the P-M System. *Genetics* **92**, 1127-1140 (1979).
146 Schaefer, R. E., Kidwell, M. G. & Fausto-Sterling, A. Hybrid Dysgenesis in DROSOPHILA MELANOGASTER: Morphological and Cytological Studies of Ovarian Dysgenesis. *Genetics* **92**, 1141-1152 (1979).
147 Reiss, D., Josse, T., Anxolabehere, D. & Ronsseray, S. aubergine mutations in Drosophila melanogaster impair P cytotype determination by telomeric P elements inserted in heterochromatin. *Molecular genetics and genomics : MGG* **272**, 336-343, doi:10.1007/s00438-004-1061-1 (2004).
148 Sarot, E., Payen-Groschene, G., Bucheton, A. & Pelisson, A. Evidence for a piwi-dependent RNA silencing of the gypsy endogenous retrovirus by the Drosophila melanogaster flamenco gene. *Genetics* **166**, 1313-1321 (2004).
149 Savitsky, M., Kwon, D., Georgiev, P., Kalmykova, A. & Gvozdev, V. Telomere elongation is under the control of the RNAi-based mechanism in the Drosophila germline. *Genes & development* **20**, 345-354, doi:10.1101/gad.370206 (2006).
150 Vagin, V. V. *et al.* The RNA interference proteins and vasa locus are involved in the silencing of retrotransposons in the female germline of Drosophila melanogaster. *RNA biology* **1**, 54-58 (2004).
151 Kalmykova, A. I., Klenov, M. S. & Gvozdev, V. A. Argonaute protein PIWI controls mobilization of retrotransposons in the Drosophila male germline. *Nucleic acids research* **33**, 2052-2059, doi:10.1093/nar/gki323 (2005).
152 Li, C. *et al.* Collapse of germline piRNAs in the absence of Argonaute3 reveals somatic piRNAs in flies. *Cell* **137**, 509-521, doi:10.1016/j.cell.2009.04.027 (2009).
153 Deng, W. & Lin, H. miwi, a murine homolog of piwi, encodes a cytoplasmic protein essential for spermatogenesis. *Developmental cell* **2**, 819-830 (2002).
154 Kuramochi-Miyagawa, S. *et al.* Mili, a mammalian member of piwi family gene, is essential for spermatogenesis. *Development* **131**, 839-849, doi:10.1242/dev.00973 (2004).
155 Carmell, M. A. *et al.* MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline. *Developmental cell* **12**, 503-514, doi:10.1016/j.devcel.2007.03.001 (2007).
156 Aravin, A. A., Sachidanandam, R., Girard, A., Fejes-Toth, K. & Hannon, G. J. Developmentally regulated piRNA clusters implicate MILI in transposon control. *Science* **316**, 744-747, doi:10.1126/science.1142612 (2007).
157 Aravin, A. A. *et al.* Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the D. melanogaster germline. *Current biology : CB* **11**, 1017-1027 (2001).

158 Li, M. A., Alls, J. D., Avancini, R. M., Koo, K. & Godt, D. The large Maf factor Traffic Jam controls gonad morphogenesis in Drosophila. *Nature cell biology* **5**, 994-1000, doi:10.1038/ncb1058 (2003).
159 Saito, K. *et al.* A regulatory circuit for piwi by the large Maf gene traffic jam in Drosophila. *Nature* **461**, 1296-1299, doi:10.1038/nature08501 (2009).
160 Nishida, K. M. *et al.* Gene silencing mechanisms mediated by Aubergine piRNA complexes in Drosophila male gonad. *RNA* **13**, 1911-1922, doi:10.1261/rna.744307 (2007).
161 Arensburger, P., Hice, R. H., Wright, J. A., Craig, N. L. & Atkinson, P. W. The mosquito Aedes aegypti has a large genome size and high transposable element load but contains a low proportion of transposon-specific piRNAs. *BMC genomics* **12**, 606, doi:10.1186/1471-2164-12-606 (2011).
162 Gonzalez-Reyes, A., Elliott, H. & St Johnston, D. Oocyte determination and the origin of polarity in Drosophila: the role of the spindle genes. *Development* **124**, 4927-4937 (1997).
163 Lin, H. & Spradling, A. C. A novel group of pumilio mutations affects the asymmetric division of germline stem cells in the Drosophila ovary. *Development* **124**, 2463-2476 (1997).
164 Juliano, C., Wang, J. & Lin, H. Uniting germline and stem cells: the function of Piwi proteins and the piRNA pathway in diverse organisms. *Annual review of genetics* **45**, 447-469, doi:10.1146/annurev-genet-110410-132541 (2011).
165 Siomi, M. C., Sato, K., Pezic, D. & Aravin, A. A. PIWI-interacting small RNAs: the vanguard of genome defence. *Nature reviews. Molecular cell biology* **12**, 246-258, doi:10.1038/nrm3089 (2011).
166 Klenov, M. S. *et al.* Separation of stem cell maintenance and transposon silencing functions of Piwi protein. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 18760-18765, doi:10.1073/pnas.1106676108 (2011).
167 Wang, G. & Reinke, V. A C. elegans Piwi, PRG-1, regulates 21U-RNAs during spermatogenesis. *Current biology : CB* **18**, 861-867, doi:10.1016/j.cub.2008.05.009 (2008).
168 Batista, P. J. *et al.* PRG-1 and 21U-RNAs interact to form the piRNA complex required for fertility in C. elegans. *Molecular cell* **31**, 67-78, doi:10.1016/j.molcel.2008.06.002 (2008).
169 Houwing, S. *et al.* A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. *Cell* **129**, 69-82, doi:10.1016/j.cell.2007.03.026 (2007).
170 Brackney, D. E. *et al.* C6/36 Aedes albopictus cells have a dysfunctional antiviral RNA interference response. *PLoS neglected tropical diseases* **4**, e856, doi:10.1371/journal.pntd.0000856 (2010).
171 Hess, A. M. *et al.* Small RNA profiling of Dengue virus-mosquito interactions implicates the PIWI RNA pathway in anti-viral defense. *BMC microbiology* **11**, 45, doi:10.1186/1471-2180-11-45 (2011).
172 Scott, J. C. *et al.* Comparison of dengue virus type 2-specific small RNAs from RNA interference-competent and -incompetent mosquito cells. *PLoS neglected tropical diseases* **4**, e848, doi:10.1371/journal.pntd.0000848 (2010).

173 Morazzani, E. M., Wiley, M. R., Murreddu, M. G., Adelman, Z. N. & Myles, K. M. Production of virus-derived ping-pong-dependent piRNA-like small RNAs in the mosquito soma. *PLoS pathogens* **8**, e1002470, doi:10.1371/journal.ppat.1002470 (2012).
174 Vodovar, N. *et al.* Arbovirus-derived piRNAs exhibit a ping-pong signature in mosquito cells. *PloS one* **7**, e30861, doi:10.1371/journal.pone.0030861 (2012).
175 Girard, A., Sachidanandam, R., Hannon, G. J. & Carmell, M. A. A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature* **442**, 199-202, doi:10.1038/nature04917 (2006).
176 Grivna, S. T., Beyret, E., Wang, Z. & Lin, H. A novel class of small RNAs in mouse spermatogenic cells. *Genes & development* **20**, 1709-1714, doi:10.1101/gad.1434406 (2006).
177 Li, X. Z. *et al.* An ancient transcription factor initiates the burst of piRNA production during early meiosis in mouse testes. *Molecular cell* **50**, 67-81, doi:10.1016/j.molcel.2013.02.016 (2013).
178 Malone, C. D. *et al.* Specialized piRNA pathways act in germline and somatic tissues of the Drosophila ovary. *Cell* **137**, 522-535, doi:10.1016/j.cell.2009.03.040 (2009).
179 Gu, W. *et al.* CapSeq and CIP-TAP identify Pol II start sites and reveal capped small RNAs as C. elegans piRNA precursors. *Cell* **151**, 1488-1500, doi:10.1016/j.cell.2012.11.023 (2012).
180 Klattenhoff, C. *et al.* The Drosophila HP1 homolog Rhino is required for transposon silencing and piRNA production by dual-strand clusters. *Cell* **138**, 1137-1149, doi:10.1016/j.cell.2009.07.014 (2009).
181 Pane, A., Jiang, P., Zhao, D. Y., Singh, M. & Schupbach, T. The Cutoff protein regulates piRNA cluster expression and piRNA production in the Drosophila germline. *The EMBO journal* **30**, 4601-4615, doi:10.1038/emboj.2011.334 (2011).
182 Pane, A., Wehr, K. & Schupbach, T. zucchini and squash encode two putative nucleases required for rasiRNA production in the Drosophila germline. *Developmental cell* **12**, 851-862, doi:10.1016/j.devcel.2007.03.022 (2007).
183 Nishimasu, H. *et al.* Structure and function of Zucchini endoribonuclease in piRNA biogenesis. *Nature* **491**, 284-287, doi:10.1038/nature11509 (2012).
184 Gunawardane, L. S. *et al.* A slicer-mediated mechanism for repeat-associated siRNA 5' end formation in Drosophila. *Science* **315**, 1587-1590, doi:10.1126/science.1140494 (2007).
185 Scalenghe, F., Turco, E., Edstrom, J. E., Pirrotta, V. & Melli, M. Microdissection and cloning of DNA from a specific region of Drosophila melanogaster polytene chromosomes. *Chromosoma* **82**, 205-216 (1981).
186 Bates, G. P., Wainwright, B. J., Williamson, R. & Brown, S. D. Microdissection of and microcloning from the short arm of human chromosome 2. *Molecular and cellular biology* **6**, 3826-3830 (1986).
187 Rohme, D. *et al.* Molecular clones of the mouse t complex derived from microdissected metaphase chromosomes. *Cell* **36**, 783-788 (1984).
188 Monajembashi, S., Cremer, C., Cremer, T., Wolfrum, J. & Greulich, K. O. Microdissection of human chromosomes by a laser microbeam. *Experimental cell research* **167**, 262-265 (1986).
189 Suarez-Quian, C. A. *et al.* Laser capture microdissection of single cells from complex tissues. *BioTechniques* **26**, 328-335 (1999).

190 Hofbauer, M. *et al.* Clonal tracking of autoaggressive T cells in polymyositis by combining laser microdissection, single-cell PCR, and CDR3-spectratype analysis. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 4090-4095, doi:10.1073/pnas.0236183100 (2003).
191 Schutze, K. & Lahr, G. Identification of expressed genes by laser-mediated manipulation of single cells. *Nature biotechnology* **16**, 737-742, doi:10.1038/nbt0898-737 (1998).
192 Standaert, A. R. *et al.* Obtaining single cells: analysis and evaluation of an experimental protocol by means of a simulation model. *International journal of food microbiology* **100**, 55-66, doi:10.1016/j.ijfoodmicro.2004.10.033 (2005).
193 Gillespie, J. W. *et al.* The role of tissue microdissection in cancer research. *Cancer J* **7**, 32-39 (2001).
194 Maitra, A., Wistuba, II & Gazdar, A. F. Microdissection and the study of cancer pathways. *Current molecular medicine* **1**, 153-162 (2001).
195 Braakman, R. B., Luider, T. M., Martens, J. W., Foekens, J. A. & Umar, A. Laser capture microdissection applications in breast cancer proteomics. *Methods Mol Biol* **755**, 143-154, doi:10.1007/978-1-61779-163-5_11 (2011).
196 Drosopoulou, E. *et al.* Sex chromosomes and associated rDNA form a heterochromatic network in the polytene nuclei of Bactrocera oleae (Diptera: Tephritidae). *Genetica* **140**, 169-180, doi:10.1007/s10709-012-9668-3 (2012).
197 Thalhammer, S., Langer, S., Speicher, M. R., Heckl, W. M. & Geigl, J. B. Generation of chromosome painting probes from single chromosomes by laser microdissection and linker-adaptor PCR. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology* **12**, 337-343, doi:10.1023/B:CHRO.0000034132.77192.5f (2004).
198 Rutten, K. B. *et al.* Chromosomal anchoring of linkage groups and identification of wing size QTL using markers and FISH probes derived from microdissected chromosomes in Nasonia (Pteromalidae: Hymenoptera). *Cytogenetic and genome research* **105**, 126-133, doi:10.1159/000078019 (2004).
199 Deng, H. X. *et al.* Chromosome-band-specific painting: chromosome in situ suppression hybridization using PCR products from a microdissected chromosome band as a probe pool. *Human genetics* **89**, 13-17 (1992).
200 Guan, X. Y., Trent, J. M. & Meltzer, P. S. Generation of band-specific painting probes from a single microdissected chromosome. *Human molecular genetics* **2**, 1117-1121 (1993).
201 Meltzer, P. S., Guan, X. Y., Burgess, A. & Trent, J. M. Rapid generation of region specific probes by chromosome microdissection and their application. *Nature genetics* **1**, 24-28, doi:10.1038/ng0492-24 (1992).
202 Ponelies, N., Bautz, E. K., Monajembashi, S., Wolfrum, J. & Greulich, K. O. Telomeric sequences derived from laser-microdissected polytene chromosomes. *Chromosoma* **98**, 351-357 (1989).
203 Wesley, C. S., Ben, M., Kreitman, M., Hagag, N. & Eanes, W. F. Cloning regions of the Drosophila genome by microdissection of polytene chromosome DNA and PCR with nonspecific primer. *Nucleic acids research* **18**, 599-603 (1990).
204 Wei, J. *et al.* Linkage mapping of microdissected clones from distal mouse chromosome 16. *Somatic cell and molecular genetics* **22**, 227-232 (1996).

205 Bahary, N., McGraw, D. E., Shilling, R. & Friedman, J. M. Microdissection and microcloning of mid-chromosome 4: genetic mapping of 41 microdissection clones. *Genomics* **16**, 113-122, doi:10.1006/geno.1993.1148 (1993).
206 Davis, L. M. *et al.* Somatic cell hybrid and long-range physical mapping of 11p13 microdissected genomic clones. *Proceedings of the National Academy of Sciences of the United States of America* **87**, 7005-7009 (1990).
207 Tremblay, A., Li, S., Scheffler, B. E. & Matthews, B. F. Laser capture microdissection and expressed sequence tag analysis of uredinia formed by Phakopsora pachyrhizi, the causal agent of Asian soybean rust. *Physiological and Molecular Plant Pathology* **73**, 163-174, doi:http://dx.doi.org/10.1016/j.pmpp.2009.06.001 (2008).
208 Zhou, R. N. *et al.* Rapid EST isolation from chromosome 1R of rye. *BMC plant biology* **8**, 28, doi:10.1186/1471-2229-8-28 (2008).
209 Gracia, E. *et al.* Isolation of chromosome-specific ESTs by microdissection-mediated cDNA capture. *Genome research* **7**, 100-107 (1997).
210 Telenius, H. *et al.* Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. *Genomics* **13**, 718-725 (1992).
211 Zhang, L. *et al.* Whole genome amplification from a single cell: implications for genetic analysis. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 5847-5851 (1992).
212 Ludecke, H. J., Senger, G., Claussen, U. & Horsthemke, B. Cloning defined regions of the human genome by microdissection of banded chromosomes and enzymatic amplification. *Nature* **338**, 348-350, doi:10.1038/338348a0 (1989).
213 Foucault, F., Praz, F., Jaulin, C. & Amor-Gueret, M. Experimental limits of PCR analysis of (CA)n repeat alterations. *Trends in genetics : TIG* **12**, 450-452 (1996).
214 Wells, D., Sherlock, J. K., Handyside, A. H. & Delhanty, J. D. Detailed chromosomal and molecular genetic analysis of single cells by whole genome amplification and comparative genomic hybridisation. *Nucleic acids research* **27**, 1214-1218 (1999).
215 Snabes, M. C. *et al.* Preimplantation single-cell analysis of multiple genetic loci by whole-genome amplification. *Proceedings of the National Academy of Sciences of the United States of America* **91**, 6181-6185 (1994).
216 Barrett, M. T., Reid, B. J. & Joslyn, G. Genotypic analysis of multiple loci in somatic cells by whole genome amplification. *Nucleic acids research* **23**, 3488-3492 (1995).
217 Schaaff, F., Wedemann, H. & Schwinger, E. Analysis of sex and delta F508 in single amniocytes using primer extension preamplification. *Human genetics* **98**, 158-161 (1996).
218 Gribble, S., Ng, B. L., Prigmore, E., Burford, D. C. & Carter, N. P. Chromosome paints from single copies of chromosomes. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology* **12**, 143-151 (2004).
219 Zheng, Y. M., Wang, N., Li, L. & Jin, F. Whole genome amplification in preimplantation genetic diagnosis. *Journal of Zhejiang University. Science. B* **12**, 1-11, doi:10.1631/jzus.B1000196 (2011).
220 Kittler, R., Stoneking, M. & Kayser, M. A whole genome amplification method to generate long fragments from low quantities of genomic DNA. *Analytical biochemistry* **300**, 237-244, doi:10.1006/abio.2001.5460 (2002).

221 Dietmaier, W. *et al.* Multiple mutation analyses in single tumor cells with improved whole genome amplification. *The American journal of pathology* **154**, 83-95, doi:10.1016/S0002-9440(10)65254-6 (1999).
222 Fiegler, H. *et al.* High resolution array-CGH analysis of single cells. *Nucleic acids research* **35**, e15, doi:10.1093/nar/gkl1030 (2007).
223 O'Geen, H., Nicolet, C. M., Blahnik, K., Green, R. & Farnham, P. J. Comparison of sample preparation methods for ChIP-chip assays. *BioTechniques* **41**, 577-580 (2006).
224 Wasserstrom, A. *et al.* Reconstruction of cell lineage trees in mice. *PloS one* **3**, e1939, doi:10.1371/journal.pone.0001939 (2008).
225 Blanco, L. *et al.* Highly efficient DNA synthesis by the phage phi 29 DNA polymerase. Symmetrical mode of DNA replication. *The Journal of biological chemistry* **264**, 8935-8940 (1989).
226 Dean, F. B. *et al.* Comprehensive human genome amplification using multiple displacement amplification. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 5261-5266, doi:10.1073/pnas.082089499 (2002).
227 Hosono, S. *et al.* Unbiased whole-genome amplification directly from clinical samples. *Genome research* **13**, 954-964, doi:10.1101/gr.816903 (2003).
228 Renwick, P. J. *et al.* Proof of principle and first cases using preimplantation genetic haplotyping--a paradigm shift for embryo diagnosis. *Reproductive biomedicine online* **13**, 110-119 (2006).
229 Burlet, P. *et al.* Multiple displacement amplification improves PGD for fragile X syndrome. *Molecular human reproduction* **12**, 647-652, doi:10.1093/molehr/gal069 (2006).
230 Hellani, A. *et al.* Multiple displacement amplification on single cell and possible PGD applications. *Molecular human reproduction* **10**, 847-852, doi:10.1093/molehr/gah114 (2004).
231 Kumar, G., Garnova, E., Reagin, M. & Vidali, A. Improved multiple displacement amplification with phi29 DNA polymerase for genotyping of single human cells. *BioTechniques* **44**, 879-890, doi:10.2144/000112755 (2008).
232 Raghunathan, A. *et al.* Genomic DNA amplification from a single bacterium. *Applied and environmental microbiology* **71**, 3342-3347, doi:10.1128/AEM.71.6.3342-3347.2005 (2005).
233 Handyside, A. H. *et al.* Isothermal whole genome amplification from single and small numbers of cells: a new era for preimplantation genetic diagnosis of inherited disease. *Molecular human reproduction* **10**, 767-772, doi:10.1093/molehr/gah101 (2004).
234 Pinkel, D. *et al.* Fluorescence in situ hybridization with human chromosome-specific libraries: detection of trisomy 21 and translocations of chromosome 4. *Proceedings of the National Academy of Sciences of the United States of America* **85**, 9138-9142 (1988).
235 Lichter, P., Cremer, T., Borden, J., Manuelidis, L. & Ward, D. C. Delineation of individual human chromosomes in metaphase and interphase cells by in situ suppression hybridization using recombinant DNA libraries. *Human genetics* **80**, 224-234 (1988).
236 Cremer, T., Lichter, P., Borden, J., Ward, D. C. & Manuelidis, L. Detection of chromosome aberrations in metaphase and interphase tumor cells by in situ hybridization using chromosome-specific library probes. *Human genetics* **80**, 235-246 (1988).

237 Jauch, A. *et al.* Reconstruction of genomic rearrangements in great apes and gibbons by chromosome painting. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 8611-8615 (1992).
238 Lengauer, C. *et al.* Painting of defined chromosomal regions by in situ suppression hybridization of libraries from laser-microdissected chromosomes. *Cytogenetics and cell genetics* **56**, 27-30 (1991).
239 Collins, C. *et al.* Construction and characterization of plasmid libraries enriched in sequences from single human chromosomes. *Genomics* **11**, 997-1006 (1991).
240 Telenius, H. *et al.* Cytogenetic analysis by chromosome painting using DOP-PCR amplified flow-sorted chromosomes. *Genes, chromosomes & cancer* **4**, 257-263 (1992).
241 Carter, N. P. *et al.* Reverse chromosome painting: a method for the rapid analysis of aberrant chromosomes in clinical cytogenetics. *Journal of medical genetics* **29**, 299-307 (1992).
242 Vezuli, A., Hauwy, M., Warter, S. & Rumpler, Y. Chromosome painting: a method for testing chromosomal changes in lemur evolution. *Cytogenetics and cell genetics* **78**, 147-152 (1997).
243 Geyer, P. K., Vitalini, M. W. & Wallrath, L. L. Nuclear organization: taking a position on gene expression. *Current opinion in cell biology* **23**, 354-359, doi:10.1016/j.ceb.2011.03.002 (2011).
244 Sexton, T., Schober, H., Fraser, P. & Gasser, S. M. Gene regulation through nuclear organization. *Nature structural & molecular biology* **14**, 1049-1055, doi:10.1038/nsmb1324 (2007).
245 Schneider, R. & Grosschedl, R. Dynamics and interplay of nuclear architecture, genome organization, and gene expression. *Genes & development* **21**, 3027-3043, doi:10.1101/gad.1604607 (2007).
246 Schrock, E. *et al.* Multicolor spectral karyotyping of human chromosomes. *Science* **273**, 494-497 (1996).
247 Speicher, M. R., Gwyn Ballard, S. & Ward, D. C. Karyotyping human chromosomes by combinatorial multi-fluor FISH. *Nature genetics* **12**, 368-375, doi:10.1038/ng0496-368 (1996).
248 Liyanage, M. *et al.* Multicolour spectral karyotyping of mouse chromosomes. *Nature genetics* **14**, 312-315, doi:10.1038/ng1196-312 (1996).
249 Weaver, Z. A. *et al.* A recurring pattern of chromosomal aberrations in mammary gland tumors of MMTV-cmyc transgenic mice. *Genes, chromosomes & cancer* **25**, 251-260 (1999).
250 Rocca, K. A., Gray, E. M., Costantini, C. & Besansky, N. J. 2La chromosomal inversion enhances thermal tolerance of Anopheles gambiae larvae. *Malaria journal* **8**, 147, doi:10.1186/1475-2875-8-147 (2009).
251 Fouet, C., Gray, E., Besansky, N. J. & Costantini, C. Adaptation to aridity in the malaria mosquito Anopheles gambiae: chromosomal inversion polymorphism and body size influence resistance to desiccation. *PloS one* **7**, e34841, doi:10.1371/journal.pone.0034841 (2012).
252 Weise, A. *et al.* High-throughput sequencing of microdissected chromosomal regions. *European journal of human genetics : EJHG* **18**, 457-462, doi:10.1038/ejhg.2009.196 (2010).

253 Baird, D. M., Rowson, J., Wynford-Thomas, D. & Kipling, D. Extensive allelic variation and ultrashort telomeres in senescent human cells. *Nature genetics* **33**, 203-207, doi:10.1038/ng1084 (2003).
254 Hemann, M. T., Strong, M. A., Hao, L. Y. & Greider, C. W. The shortest telomere, not average telomere length, is critical for cell viability and chromosome stability. *Cell* **107**, 67-77 (2001).
255 Seifertova, E. *et al.* Efficient high-throughput sequencing of a laser microdissected chromosome arm. *BMC genomics* **14**, 357, doi:10.1186/1471-2164-14-357 (2013).
256 Holt, R. A. *et al.* The genome sequence of the malaria mosquito Anopheles gambiae. *Science* **298**, 129-149, doi:10.1126/science.1076181 (2002).
257 Sharakhova, M. V. *et al.* Genome mapping and characterization of the Anopheles gambiae heterochromatin. *BMC genomics* **11**, 459, doi:10.1186/1471-2164-11-459 (2010).
258 Hoskins, R. A. *et al.* Heterochromatic sequences in a Drosophila whole-genome shotgun assembly. *Genome biology* **3**, RESEARCH0085 (2002).
259 Genome, K. C. o. S. Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *The Journal of heredity* **100**, 659-674, doi:10.1093/jhered/esp086 (2009).
260 Robinson, G. E. *et al.* Creating a buzz about insect genomes. *Science* **331**, 1386, doi:10.1126/science.331.6023.1386 (2011).
261 Besansky, N. J. Genome Analysis Of Vectorial Capacity In Major Anopheles Vectors Of Malaria Parasites. (2008).
262 Alkan, C., Sajjadian, S. & Eichler, E. E. Limitations of next-generation genome sequence assembly. *Nature methods* **8**, 61-65, doi:10.1038/nmeth.1527 (2011).
263 Lewin, H. A., Larkin, D. M., Pontius, J. & O'Brien, S. J. Every genome sequence needs a good map. *Genome research* **19**, 1925-1928, doi:10.1101/gr.094557.109 (2009).
264 Bhutkar, A. *et al.* Chromosomal rearrangement inferred from comparisons of 12 Drosophila genomes. *Genetics* **179**, 1657-1680, doi:10.1534/genetics.107.086108 (2008).
265 Novikov, D. V., Kireev, I. & Belmont, A. S. High-pressure treatment of polytene chromosomes improves structural resolution. *Nature methods* **4**, 483-485, doi:10.1038/nmeth1049 (2007).
266 George, P., Sharakhova, M. & Sharakhov, I. High-resolution cytogenetic map for the African malaria vector Anopheles gambiae. *Insect molecular biology* (2010).
267 Molecular, A. *http://www.abbottmolecular.com/products/instrumentation-automation/fish/xmatrx.html*, <http://www.abbottmolecular.com/products/instrumentation-automation/fish/xmatrx.html> (
268 Bioview. *BIOVIEW ACCORD SCANNING STATION*, <http://www.bioview.co.il/htmls/article.aspx?c0=12510&bsp=12332&bss53=12332> (
269 Vectorbase.org. *BLAST @ VectorBase.org*, <http://agambiae.vectorbase.org/Tools/BLAST/> (2012).
270 Clements, A. N. *The Biology of Mosquitoes: Development, nutrition, and reproduction*. (Chapman & Hall, 1992).
271 Sharakhova, M. V. *et al.* Imaginal discs--a new source of chromosomes for genome mapping of the yellow fever mosquito Aedes aegypti. *PLoS neglected tropical diseases* **5**, e1335, doi:10.1371/journal.pntd.0001335 (2011).

272 Scriven, P. N., Kirby, T. L. & Ogilvie, C. M. FISH for pre-implantation genetic diagnosis. *Journal of visualized experiments : JoVE*, doi:10.3791/2570 (2011).
273 Cai, W., Jin, Y., Girton, J., Johansen, J. & Johansen, K. M. Preparation of Drosophila polytene chromosome squashes for antibody labeling. *Journal of visualized experiments : JoVE*, doi:10.3791/1748 (2010).
274 Vatolina, T. Y. *et al.* Identical functional organization of nonpolytene and polytene chromosomes in Drosophila melanogaster. *PloS one* **6**, e25960, doi:10.1371/journal.pone.0025960 (2011).
275 Tolle, M. A. Mosquito-borne diseases. *Current problems in pediatric and adolescent health care* **39**, 97-140, doi:10.1016/j.cppeds.2009.01.001 (2009).
276 Sharakhov, I. & Sharakhova, M. in *Chromosome mapping research developments* eds J.F. Verrity & L.E. Abbington) (Nova Science Publishers, Inc., 2008).
277 Cornel, A. J. & Collins, F. H. Maintenance of chromosome arm integrity between two Anopheles mosquito subgenera. *The Journal of heredity* **91**, 364-370 (2000).
278 Sharakhov, I. V. *et al.* Breakpoint structure reveals the unique origin of an interspecific chromosomal inversion (2La) in the Anopheles gambiae complex. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 6258-6262, doi:10.1073/pnas.0509683103 (2006).
279 Frizzi, G. & Holstein, M. Etude cytogénétique d'Anopheles gambiae. *Bulletin of the World Health Organization* **15**, 425-435 (1956).
280 Coluzzi, M. Chromosomi politenici cellule nutrici ovariche nel complesso gambiae del genere Anopheles. *Parassitologia* **10**, 179-183 (1968).
281 Coluzzi, M. a. & Sabatini, A. Cytogenetic observations on species C of the Anopheles gambiae complex. *Parassitologia* **10**, 155-165 (1968).
282 Coluzzi, M. Cytogenetic observations on the salt water species, Anopheles merus and Anopheles melas, of the gambiae complex. *Parassitologia* **11** (1969).
283 Bryan, J. H., Deco, M. A., Petrarca, V. & Coluzzi, M. Inversion polymorphism and incipient speciation in Anopheles gambiae s.str. in The Gambia, West Africa. *Genetica* **59**, 167-176, doi:10.1007/bf00056539 (1982).
284 Coluzzi, M., Petrarca, V. & di Deco, M. A. Chromosomal inversion intergradation and incipient speciation in Anopheles gambiae. *Bolletino di zoologia* **52**, 45-63, doi:10.1080/11250008509440343 (1985).
285 Ayala, F. J. & Coluzzi, M. Chromosome speciation: humans, Drosophila, and mosquitoes. *Proceedings of the National Academy of Sciences of the United States of America* **102 Suppl 1**, 6535-6542, doi:10.1073/pnas.0501847102 (2005).
286 Xia, A., Sharakhova, M. V. & Sharakhov, I. V. Reconstructing ancestral autosomal arrangements in the Anopheles gambiae complex. *Journal of computational biology : a journal of computational molecular cell biology* **15**, 965-980, doi:10.1089/cmb.2008.0076 (2008).
287 Campos, J., Andrade, C. F. & Recco-Pimentel, S. M. Malpighian tubule polytene chromosomes of Culex quinquefasciatus (Diptera, Culicinae). *Memorias do Instituto Oswaldo Cruz* **98**, 383-386 (2003).
288 McAbee, R. D., Christiansen, J. A. & Cornel, A. J. A detailed larval salivary gland polytene chromosome photomap for Culex quinquefasciatus (Diptera: Culicidae) from Johannesburg, South Africa. *Journal of medical entomology* **44**, 229-237 (2007).

289 Campos, J., Andrade, C. F. & Recco-Pimentel, S. M. A technique for preparing polytene chromosomes from Aedes aegypti (Diptera, Culicinae). *Memorias do Instituto Oswaldo Cruz* **98**, 387-390 (2003).
290 Lawson, D. *et al.* VectorBase: a data resource for invertebrate vector genomics. *Nucleic acids research* **37**, D583 - 587 (2009).
291 Sharakhova, M., Xia, A., McAlister, S. & Sharakhov, I. A standard cytogenetic photomap for the mosquito Anopheles stephensi (Diptera: Culicidae): application for physical mapping. *Journal of medical entomology* **43**, 861 - 866 (2006).
292 Bernard, P. *et al.* Requirement of heterochromatin for cohesion at centromeres. *Science* **294**, 2539 - 2542 (2001).
293 Dernburg, A., Sedat, J. & Hawley, R. Direct evidence of a role for heterochromatin in meiotic chromosome segregation. *Cell* **86**, 135 - 146 (1996).
294 Swedlow, J. & Lamond, A. Nuclear dynamics: where genes are and how they got there. *Genome biology* **2**, REVIEWS0002 (2001).
295 Zhimulev, I. Polytene chromosomes, heterochromatin, and position effect variegation. *Advances in genetics* **37**, 1 - 566 (1998).
296 Riddle, N. & Elgin, S. A role for RNAi in heterochromatin formation in Drosophila. *Current topics in microbiology and immunology* **320**, 185 - 209 (2008).
297 Vermaak, D. & Malik, H. Multiple roles for heterochromatin protein 1 genes in Drosophila. *Annual review of genetics* **43**, 467 - 492 (2009).
298 Girton, J. & Johansen, K. Chromatin structure and the regulation of gene expression: the lessons of PEV in Drosophila. *Advances in genetics* **61**, 1 - 43 (2008).
299 Abad, P. *et al.* A long interspersed repetitive element--the I factor of Drosophila teissieri--is able to transpose in different Drosophila species. *Proc Natl Acad Sci USA* **86**, 8887 - 8891 (1989).
300 Weiler, K. & Wakimoto, B. Heterochromatin and gene expression in Drosophila. *Annual review of genetics* **29**, 577 - 605 (1995).
301 Hoskins, R. *et al.* Sequence finishing and mapping of Drosophila melanogaster heterochromatin. *Science* **316**, 1625 - 1628 (2007).
302 Smith, C., Shu, S., Mungall, C. & Karpen, G. The Release 5.1 annotation of Drosophila melanogaster heterochromatin. *Science* **316**, 1586 - 1591 (2007).
303 Andreyeva, E. *et al.* High-resolution analysis of Drosophila heterochromatin organization using SuUR Su(var)3-9 double mutants. *Proc Natl Acad Sci USA* **104**, 12819 - 12824 (2007).
304 Fitzpatrick, K., Sinclair, D., Schulze, S., Syrzycka, M. & Honda, B. A genetic and molecular profile of third chromosome centric heterochromatin in Drosophila melanogaster. *Genome/National Research Council Canada = Genome/Conseil national de recherches Canada* **48**, 571 - 584 (2005).
305 Prokofyeva-Belgovskaya, A. Inert regions of internal parts in X-chromosome Drosophila melanogaster. *Proc Acad Sci USSR* **3**, 362 - 370 (1939).
306 Zhimulev, I. F. *et al.* Intercalary heterochromatin in Drosophila melanogaster polytene chromosomes and the problem of genetic silencing. *Genetica* **117**, 259-270 (2003).
307 Belyakin, S. N. *et al.* Genomic analysis of Drosophila chromosome underreplication reveals a link between replication control and transcriptional territories. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 8269-8274, doi:10.1073/pnas.0502702102 (2005).

308 Heitz, E. Uber alpha- und beta-Heterochromatin sowie Konstanz und Bau der Chromomeren bei Drosophila. *Biol Zbl* **54**, 588 - 609 (1934).
309 Vaury, C., Bucheton, A. & Pelisson, A. The beta heterochromatic sequences flanking the I elements are themselves defective transposable elements. *Chromosoma* **98**, 215 - 224 (1989).
310 Miklos, G., Yamamoto, M., Davies, J. & Pirrotta, V. Microcloning reveals a high frequency of repetitive sequences characteristic of chromosome 4 and the beta-heterochromatin of Drosophila melanogaster. *Proc Natl Acad Sci USA* **85**, 2051 - 2055 (1988).
311 Strahl, B. & Allis, C. The language of covalent histone modifications. *Nature* **403**, 41 - 45 (2000).
312 Bannister, A. *et al.* Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature* **410**, 120 - 124 (2001).
313 Lachner, M., O'Carroll, D., Rea, S., Mechtler, K. & Jenuwein, T. Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature* **410**, 116 - 120 (2001).
314 James, T. C. & Elgin, S. C. Identification of a nonhistone chromosomal protein associated with heterochromatin in Drosophila melanogaster and its gene. *Molecular and cellular biology* **6**, 3862-3872 (1986).
315 Riddle, N. & Elgin, S. The dot chromosome of Drosophila: insights into chromatin states and their change over evolutionary time. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology* **14**, 405 - 416 (2006).
316 Riddle, N. *et al.* An investigation of heterochromatin domains on the fourth chromosome of Drosophila melanogaster. *Genetics* **178**, 1177 - 1191 (2008).
317 Slawson, E. *et al.* Comparison of dot chromosome sequences from D. melanogaster and D. virilis reveals an enrichment of DNA transposon sequences in heterochromatic domains. *Genome biology* **7**, R15 (2006).
318 Baricheva, E. *et al.* DNA from Drosophila melanogaster beta-heterochromatin binds specifically to nuclear lamins in vitro and the nuclear envelope in situ. *Gene* **171**, 171 - 176 (1996).
319 Hochstrasser, M. & Sedat, J. W. Three-dimensional organization of Drosophila melanogaster interphase nuclei. II. Chromosome spatial organization and gene regulation. *The Journal of cell biology* **104**, 1471-1483 (1987).
320 Hochstrasser, M. & Sedat, J. W. Three-dimensional organization of Drosophila melanogaster interphase nuclei. I. Tissue-specific aspects of polytene nuclear architecture. *The Journal of cell biology* **104**, 1455-1470 (1987).
321 Akhtar, A. & Gasser, S. The nuclear envelope and transcriptional control. *Nature reviews. Genetics* **8**, 507 - 517 (2007).
322 Scherthan, H. Telomere attachment and clustering during meiosis. *Cellular and molecular life sciences : CMLS* **64**, 117 - 124 (2007).
323 Singh, P. & Georgatos, S. HP1: facts, open questions, and speculation. *Journal of structural biology* **140**, 10 - 16 (2002).
324 Hochstrasser, M., Mathog, D., Gruenbaum, Y., Saumweber, H. & Sedat, J. Spatial organization of chromosomes in the salivary gland nuclei of Drosophila melanogaster. *The Journal of cell biology* **102**, 112 - 123 (1986).

325 Stegnii, V. N. & Sharakhova, M. V. Systemic reorganization of the architechtonics of polytene chromosomes in onto- and phylogenesis of malaria mosquitoes. Structural features regional of chromosomal adhesion to the nuclear membrane. *Genetika* **27**, 828-835 (1991).
326 Sharakhov, I. V., Sharakhova, M. V., Mbogo, C. M., Koekemoer, L. L. & Yan, G. Linear and spatial organization of polytene chromosomes of the African malaria mosquito Anopheles funestus. *Genetics* **159**, 211-218 (2001).
327 Rzepecki, R., Bogachev, S., Kokoza, E., Stuurman, N. & Fisher, P. In vivo association of lamins with nucleic acids in Drosophila melanogaster. *Journal of cell science* **111**, 121 - 129 (1998).
328 Stierle, V. *et al.* The carboxyl-terminal region common to lamins A and C contains a DNA binding domain. *Biochemistry* **42**, 4819 - 4828 (2003).
329 Dechat, T. *et al.* Nuclear lamins: major factors in the structural organization and function of the nucleus and chromatin. *Genes & development* **22**, 832 - 853 (2008).
330 Luderus, M. *et al.* Binding of matrix attachment regions to lamin B1. *Cell* **70**, 949 - 959 (1992).
331 Strausbaugh, L. & Williams, S. High density of an SAR-associated motif differentiates heterochromatin from euchromatin. *Journal of theoretical biology* **183**, 159 - 167 (1996).
332 von Sternberg, R. & Shapiro, J. How repeated retroelements format genome function. *Cytogenetic and genome research* **110**, 108 - 116 (2005).
333 Shapiro, J. & von Sternberg, R. Why repetitive DNA is essential to genome function. *Biol Rev Camb Philos Soc* **80**, 227 - 250 (2005).
334 Gatti, M., Santini, G., Pimpinelli, S. & Coluzz, M. Fluorescence banding techniques in the identification of sibling species of the anopheles gambiae complex. *Heredity* **38**, 105 - 108 (1977).
335 Sharakhova, M., Stegnii, V. & Braginets, O. Interspecies differences in the ovarian trophocyte precentromere heterochromatin structure and evolution of the malaria mosquito complex Anopheles maculipennis. *Genetika* **33**, 1640 - 1648 (1997).
336 Sharakhova, M., Stegnii, V. & Timofeeva, O. Polymorphism of pericentromere heterochromatin of polytene chromosomes of ovarian trophocytes in natural populations of the malaria mosquito Anopheles messeae Fall. *Genetika* **33**, 281 - 283 (1997).
337 Bonaccorsi, S., Santini, G., Gatti, M., Pimpinelli, S. & Colluzzi, M. Intraspecific polymorphism of sex chromosome heterochromatin in two species of the Anopheles gambiae complex. *Chromosoma* **76**, 57 - 64 (1980).
338 Fraccaro, M., Tiepolo, L., Laudani, U., Marchi, A. & Jayakar, S. Y chromosome controls mating behaviour on Anopheles mosquitoes. *Nature* **265**, 326 - 328 (1977).
339 della Torre, A., Merzagora, L., Powell, J. R. & Coluzzi, M. Selective introgression of paracentric inversions between two sibling species of the Anopheles gambiae complex. *Genetics* **146**, 239-244 (1997).
340 Lehmann, T. & Diabate, A. The molecular forms of Anopheles gambiae: a phenotypic perspective. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases* **8**, 737 - 746 (2008).
341 Turner, T. L., Hahn, M. W. & Nuzhdin, S. V. Genomic islands of speciation in Anopheles gambiae. *PLoS biology* **3**, e285, doi:10.1371/journal.pbio.0030285 (2005).

342 White, B., Cheng, C., Simard, F., Costantini, C. & Besansky, N. Genetic association of physically unlinked islands of genomic divergence in incipient species of Anopheles gambiae. *Molecular ecology* **19**, 925 - 939 (2010).
343 Besansky, N. & Powell, J. Reassociation kinetics of Anopheles gambiae (Diptera: Culicidae) DNA. *Journal of medical entomology* **29**, 125 - 128 (1992).
344 Sharakhova, M. V. *et al.* Update of the Anopheles gambiae PEST genome assembly. *Genome biology* **8**, R5, doi:10.1186/gb-2007-8-1-r5 (2007).
345 George, P., Sharakhova, M. V. & Sharakhov, I. V. High-resolution cytogenetic map for the African malaria vector Anopheles gambiae. *Insect molecular biology* **19**, 675-682, doi:10.1111/j.1365-2583.2010.01025.x (2010).
346 Sharakhova, M., Sharakhov, I., Gosteva, L., Kotlikova, I. & Stegnii, V. Pericentromeric and intercalary alpha-heterochromatin of polytene chromosomes of the malaria mosquito. *Genetika* **36**, 175 - 181 (2000).
347 Belyaeva, E. S. *et al.* DNA underreplication in intercalary heterochromatin regions in polytene chromosomes of Drosophila melanogaster correlates with the formation of partial chromosomal aberrations and ectopic pairing. *Chromosoma* **115**, 355-366, doi:10.1007/s00412-006-0063-7 (2006).
348 Mal'ceva, N., Gyurkovics, H. & Zhimulev, I. General characteristics of the polytene chromosome from ovarian pseudonurse cells of the Drosophila melanogaster otu11 and fs(2)B mutants. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology* **3**, 191 - 200 (1995).
349 James, T. C. *et al.* Distribution patterns of HP1, a heterochromatin-associated nonhistone chromosomal protein of Drosophila. *European journal of cell biology* **50**, 170-180 (1989).
350 Buglia, G. & Ferraro, M. Germline cyst development and imprinting in male mealybug Planococcus citri. *Chromosoma* **113**, 284 - 294 (2004).
351 Piacentini, L., Fanti, L., Berloco, M., Perrini, B. & Pimpinelli, S. Heterochromatin protein 1 (HP1) is associated with induced gene expression in Drosophila euchromatin. *The Journal of cell biology* **161**, 707 - 714 (2003).
352 Cryderman, D. *et al.* Role of Drosophila HP1 in euchromatic gene expression. *Dev Dyn* **232**, 767 - 774 (2005).
353 Koryakov, D. E., Reuter, G., Dimitri, P. & Zhimulev, I. F. The SuUR gene influences the distribution of heterochromatic proteins HP1 and SU(VAR)3-9 on nurse cell polytene chromosomes of Drosophila melanogaster. *Chromosoma* **115**, 296-310, doi:10.1007/s00412-005-0044-2 (2006).
354 Aasland, R. & Stewart, A. The chromo shadow domain, a second chromo domain in heterochromatin-binding protein 1, HP1. *Nucleic acids research* **23**, 3168 - 3173 (1995).
355 Lomberk, G., Wallrath, L. & Urrutia, R. The Heterochromatin Protein 1 family. *Genome biology* **7**, 228 (2006).
356 Grewal, S. I. & Elgin, S. C. Transcription and RNA interference in the formation of heterochromatin. *Nature* **447**, 399-406, doi:10.1038/nature05914 (2007).
357 Ye, Q. & Worman, H. Interaction between an integral protein of the nuclear envelope inner membrane and human chromodomain proteins homologous to Drosophila HP1. *The Journal of biological chemistry* **271**, 14653 - 14656 (1996).

358 Ye, Q., Callebaut, I., Pezhman, A., Courvalin, J. & Worman, H. Domain-specific interactions of human HP1-type chromodomain proteins and inner nuclear membrane protein LBR. *The Journal of biological chemistry* **272**, 14983 - 14989 (1997).
359 Pickersgill, H. *et al.* Characterization of the Drosophila melanogaster genome at the nuclear lamina. *Nature genetics* **38**, 1005 - 1014 (2006).
360 Wang-Sattler, R. *et al.* Mosaic genome architecture of the Anopheles gambiae species complex. *PloS one* **2**, e1249 (2007).
361 Vermaak, D., Henikoff, S. & Malik, H. Positive selection drives the evolution of rhino, a member of the heterochromatin protein 1 family in Drosophila. *PLoS genetics* **1**, 96 - 108 (2005).
362 Talbert, P., Bryson, T. & Henikoff, S. Adaptive evolution of centromere proteins in plants and animals. *J Biol* **3**, 18 (2004).
363 Brideau, N. *et al.* Two Dobzhansky-Muller genes interact to cause hybrid lethality in Drosophila. *Science* **314**, 1292 - 1295 (2006).
364 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* **25**, 25 - 29 (2000).
365 Lawson, D. *et al.* VectorBase: a data resource for invertebrate vector genomics. *Nucleic acids research* **37**, D583-587, doi:10.1093/nar/gkn857 (2009).
366 Krzywinski, J., Sangare, D. & Besansky, N. Satellite DNA from the Y chromosome of the malaria vector Anopheles gambiae. *Genetics* **169**, 185 - 196 (2005).
367 Redfern, C. Satellite DNA of Anopheles stephensi Liston (Diptera: Culicidae). Chromosomal location and under-replication in polytene nuclei. *Chromosoma* **82**, 561 - 581 (1981).
368 Biessmann, H., Donath, J. & Walter, M. Molecular characterization of the Anopheles gambiae 2L telomeric region via an integrated transgene. *Insect molecular biology* **5**, 11 - 20 (1996).
369 Biessmann, H., Kobeski, F., Walter, M., Kasravi, A. & Roth, C. DNA organization and length polymorphism at the 2L telomeric region of Anopheles gambiae. *Insect molecular biology* **7**, 83 - 93 (1998).
370 Rozen, S. & Skaletsky, H. Primer3 on the www for general users and for biologist programmers. *Methods Mol Biol* **132**, 365 - 386 (2000).
371 Sharakhova, M. V., Xia, A., McAlister, S. I. & Sharakhov, I. V. A standard cytogenetic photomap for the mosquito Anopheles stephensi (Diptera: Culicidae): application for physical mapping. *Journal of medical entomology* **43**, 861-866 (2006).
372 Stephens, G. E., Craig, C. A., Li, Y., Wallrath, L. L. & Elgin, S. C. Immunofluorescent staining of polytene chromosomes: exploiting genetic tools. *Methods in enzymology* **376**, 372-393, doi:10.1016/S0076-6879(03)76025-X (2004).
373 Zdobnov, E. & Apweiler, R. InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847 - 848 (2001).
374 Boyle, E. *et al.* GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* **20**, 3710 - 3715 (2004).
375 Haider, S. *et al.* BioMart Central Portal--unified access to biological data. *Nucleic acids research* **37**, W23 - 27 (2009).
376 Smit, A., Hubley, R. & Green, P. (1996-2004) RepeatMasker Open-3.0.

377 Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research* **27**, 573 - 580 (1999).
378 Bailey, J., Yavor, A., Massa, H., Trask, B. & Eichler, E. Segmental duplications: organization and impact within the current human genome project assembly. *Genome research* **11**, 1005 - 1017 (2001).
379 Frisch, M. *et al.* In silico prediction of scaffold/matrix attachment regions in large genomic sequences. *Genome research* **12**, 349 - 354 (2002).
380 Andreyenkova, N. G. *et al.* Late replication domains are evolutionary conserved in the Drosophila genome. *PloS one* **8**, e83319, doi:10.1371/journal.pone.0083319 (2013).
381 Kharchenko, P. V. *et al.* Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature* **471**, 480-485, doi:10.1038/nature09725 (2011).
382 Riddle, N. C. *et al.* Plasticity in patterns of histone modifications and chromosomal proteins in Drosophila heterochromatin. *Genome research* **21**, 147-163, doi:10.1101/gr.110098.110 (2011).
383 Lakhotia, S. C. & Jacob, J. EM autoradiographic studies on polytene nuclei of Drosophila melanogaster. II. Organization and transcriptive activity of the chromocentre. *Experimental cell research* **86**, 253-263 (1974).
384 Kaufmann, B. P. Distribution of Induced Breaks along the X-Chromosome of Drosophila Melanogaster. *Proceedings of the National Academy of Sciences of the United States of America* **25**, 571-577 (1939).
385 Belyakin, S. N. *et al.* Gene density profile reveals the marking of late replicated domains in the Drosophila melanogaster genome. *Chromosoma* **119**, 589-600, doi:10.1007/s00412-010-0280-y (2010).
386 Filion, G. J. *et al.* Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell* **143**, 212-224, doi:10.1016/j.cell.2010.09.009 (2010).
387 Liu, T. *et al.* Broad chromosomal domains of histone modification patterns in C. elegans. *Genome research* **21**, 227-236, doi:10.1101/gr.115519.110 (2011).
388 Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature biotechnology* **28**, 817-825, doi:10.1038/nbt.1662 (2010).
389 Wang, J., Lunyak, V. V. & Jordan, I. K. Genome-wide prediction and analysis of human chromatin boundary elements. *Nucleic acids research* **40**, 511-529, doi:10.1093/nar/gkr750 (2012).
390 Evans, K., Ott, S., Hansen, A., Koentges, G. & Wernisch, L. A comparative study of S/MAR prediction tools. *BMC bioinformatics* **8**, 71, doi:10.1186/1471-2105-8-71 (2007).
391 Kasprzyk, A. BioMart: driving a paradigm change in biological data management. *Database : the journal of biological databases and curation* **2011**, bar049, doi:10.1093/database/bar049 (2011).
392 Liebich, I., Bode, J., Frisch, M. & Wingender, E. S/MARt DB: a database on scaffold/matrix attached regions. *Nucleic acids research* **30**, 372-374 (2002).
393 Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software* **33**, 1-22 (2010).
394 Vagin, V. V. *et al.* A distinct small RNA pathway silences selfish genetic elements in the germline. *Science* **313**, 320-324, doi:10.1126/science.1129333 (2006).

395 Aravin, A. A. *et al.* A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. *Molecular cell* **31**, 785-799, doi:10.1016/j.molcel.2008.09.003 (2008).
396 Cox, D. N., Chao, A. & Lin, H. F. piwi encodes a nucleoplasmic factor whose activity modulates the number and division rate of germline stem cells. *Development* **127**, 503-514 (2000).
397 Aravin, A. A., Hannon, G. J. & Brennecke, J. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science* **318**, 761-764, doi:10.1126/science.1146484 (2007).
398 Malone, C. D. & Hannon, G. J. Small RNAs as guardians of the genome. *Cell* **136**, 656-668, doi:10.1016/j.cell.2009.01.045 (2009).
399 Kaminker, J. S. *et al.* The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. *Genome biology* **3**, RESEARCH0084 (2002).
400 Kawaoka, S. *et al.* Bombyx small RNAs: genomic defense system against transposons in the silkworm, Bombyx mori. *Insect biochemistry and molecular biology* **38**, 1058-1065, doi:10.1016/j.ibmb.2008.03.007 (2008).
401 Dufourt, J., Pouchin, P., Peyret, P., Brasset, E. & Vaury, C. NucBase, an easy to use read mapper for small RNAs. *Mobile DNA* **4**, 1, doi:10.1186/1759-8753-4-1 (2013).
402 Carrera, P. *et al.* VASA mediates translation through interaction with a Drosophila yIF2 homolog. *Molecular cell* **5**, 181-187 (2000).
403 Lasko, P. F. & Ashburner, M. The product of the Drosophila gene vasa is very similar to eukaryotic initiation factor-4A. *Nature* **335**, 611-617, doi:10.1038/335611a0 (1988).
404 Dorn, R., Morawietz, H., Reuter, G. & Saumweber, H. Identification of an essential Drosophila gene that is homologous to the translation initiation factor eIF-4A of yeast and mouse. *Molecular & general genetics : MGG* **237**, 233-240 (1993).
405 Campbell, C. L., Black, W. C. t., Hess, A. M. & Foy, B. D. Comparative genomics of small RNA regulatory pathway components in vector mosquitoes. *BMC Genomics* **9**, 425, doi:10.1186/1471-2164-9-425 (2008).
406 Lu, H. L. *et al.* Expansion of genes encoding piRNA-associated argonaute proteins in the pea aphid: diversification of expression profiles in different plastic morphs. *PloS one* **6**, e28051, doi:10.1371/journal.pone.0028051 (2011).
407 Brower-Toland, B. *et al.* Drosophila PIWI associates with chromatin and interacts directly with HP1a. *Genes & development* **21**, 2300-2311, doi:10.1101/gad.1564307 (2007).
408 Olovnikov, I. *et al.* De novo piRNA cluster formation in the Drosophila germ line triggered by transgenes containing a transcribed transposon fragment. *Nucleic acids research* **41**, 5757-5768, doi:10.1093/nar/gkt310 (2013).
409 Esnault, C. *et al.* High genetic differentiation between the M and S molecular forms of Anopheles gambiae in Africa. *PloS one* **3**, e1968, doi:10.1371/journal.pone.0001968 (2008).
410 Boulesteix, M. *et al.* Insertion polymorphism of transposable elements and population structure of Anopheles gambiae M and S molecular forms in Cameroon. *Molecular ecology* **16**, 441-452, doi:10.1111/j.1365-294X.2006.03150.x (2007).
411 Durinck, S. *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**, 3439-3440, doi:10.1093/bioinformatics/bti525 (2005).

412 Guberman, J. M. *et al.* BioMart Central Portal: an open database network for the biological community. *Database : the journal of biological databases and curation* **2011**, bar041, doi:10.1093/database/bar041 (2011).
413 Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols* **4**, 44-57, doi:10.1038/nprot.2008.211 (2009).
414 Huang da, W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research* **37**, 1-13, doi:10.1093/nar/gkn923 (2009).
415 Maccallum, R. M., Redmond, S. N. & Christophides, G. K. An expression map for Anopheles gambiae. *BMC genomics* **12**, 620, doi:10.1186/1471-2164-12-620 (2011).

Appendix

Table S1. Genome coordinates of DNA markers mapped to polytene chromosomes of Anopheles gambiae.

X Chromosome				
Clone	Coordinates	Size	Subdivision	Approximate Location
25D14	81155-205563	124408 bps	4C	0.1 mb
25E 24	459415-587625	128210 bps	4C	0.5 mb
10F12	1036205-1037000	796 bps	4C	1 mb
03P02	1517289-1517667	379 bps	4B	1.5 mb
19D05	1997097-2153751	156654 bps	4B	2.1 mb
16M16	2217836-2218506	670 bps	4B	2.2 mb
20N08	2363190-2453423	90234 bps	4A	2.4 mb
27D16	2898137-3033190	135054 bps	4A	3 mb
24J01	3632055-3771323	139269 bps	3D	3.7 mb
04M21	4638348-4779467	141120 bps	3D	4.7 mb
26C03	5566115-5672612	106498 bps	3C	5.6 mb
28F08	6001832-6002087	256 bps	3B	6 mb
26I10	6956106-7055652	99546 bps	2C	7.0 mb
21P07	7585474-7691519	106046 bps	2C	7.6 mb
27E 10	8521267-8646883	125617 bps	2C	8.6 mb
05N05	9244024-9347480	103457 bps	2B	9.3 mb
20K19	10003137-10003540	404 bps	1D	10.0 mb
21O06	10978077-11129929	151853 bps	1D	11.1 mb
06M18	11640558-11743930	103373 bps	1D	11.7 mb
28B15	12122044-12216167	94124 bps	1C	12.2 mb
12A04	12937062-13030894	93833 bps	1C	13.0 mb
01N17	13291852-13380483	88632 bps	1B	13.3 mb
21F12	14068003-14190611	122609 bps	1B	14.1 mb
02B01	14953244-15028536	75293 bps	5A	15.0 mb
10B09	15755248-15891232	135985 bps	5A	15.8 mb
04B15	16278632-16400217	121586 bps	5B	16.3 mb
04O07	17186238-17187012	775 bps	5C	17.1 mb
08N21	18834806-18950393	115588 bps	5D	18.9 mb
02N07	19217001-19217441	441 bps	5D	19.2 mb
12A22	21861369-21861918	550 bps	6	21.8 mb
148J20	22073871-22076879	3009 bps	6	22.0 mb
AAAB01008861	24283498-24393108	109610 bps	6	24.3 mb
2R				
Clone	Coordinates	Size	Subdivision	Approximate Location

170B21	30628-77784	47156 bps	7A	0.1 mb
22B14	847272-847842	571 bps	7A	0.8 mb
14N04	1239097-1341746	102650 bps	7A	1.3 mb
17D11	1689790-1804980	115191 bps	7B	1.7 mb
21P13	2926282-3031624	105343 bps	7C	3.0 mb
03F17	4138321-4230808	92488 bps	8A	4.2 mb
29A16	5494193-5599055	104862 bps	8B	5.5 mb
25J11	5907152-6022776	115625 bps	8C	6.0 mb
14N21	6376280-6376943	664 bps	8C	6.3 mb
23I05	7021330-7021793	464 bps	8D	7.0 mb
28F19	8121684-8230477	108794 bps	8D	8.2 mb
20M22	8909576-9019429	109854 bps	8E	9.0 mb
04A11	9422798-9524853	102056 bps	9A	9.5 mb
26D07	10591666-10591844	179 bps	9B	10.6 mb
21E 12	10960707-11085995	125289 bps	9B	11.0 mb
27H11	12097915-12188454	90539 bps	9C	12.1 mb
20N10	12874430-12976648	102219 bps	9C	12.9 mb
23F12	13422101-13422258	158 bps	10A	13.4 mb
26D05	14102924-14103663	739 bps	10B	14.1 mb
08F22	15289656-15290437	782 bps	10C	15.3 mb
29M01	15813982-15933601	119620 bps	10C	15.9 mb
27A07	16990637-17099698	109062 bps	11A	17.0 mb
10G19	17405111-17539109	133999 bps	11A	17.5 mb
04A04	18463193-18463508	316 bps	11B	18.5 mb
10H09	18848634-18962390	113757 bps	11B	18.9 mb
26E 05	19962731-20055989	93259 bps	11C	20.0 mb
28O19	21130822-21220882	90060 bps	12B	21.2 mb
02H10	21721637-21816165	94529 bps	12B	21.8 mb
03P17	21860095-21959413	99319 bps	12B	21.9 mb
17M11	22473478-22577231	103754 bps	12C	22.5 mb
10A14	23320527-23435004	114478 bps	12C	23.4 mb
20M20	24308753-24309358	606 bps	12D	24.3 mb
26K01	25474326-25561165	86840 bps	12D	25.5 mb
02D19	26118607-26199369	80763 bps	12E	26.2 mb
28M08	27065015-27176517	111503 bps	13A	27.1 mb
02B20	28746411-28866225	119815 bps	13C	28.8 mb
10E 06	29490904-29598621	107718 bps	13C	29.5 mb
01C03	29974465-30077977	103513 bps	13D	30.0 mb
25D23	30307389-30403309	95921 bps	13D	30.4 mb
21A19	31115376-31228604	113229 bps	13E	31.2 mb
27D14	31850649-31980514	129865 bps	14A	31.9 mb
30F04	32848120-32966447	118328 bps	14B	32.9 mb
16G03	33081856-33082648	793 bps	14B	33.1 mb

25K03	33823740-33952438	128699 bps	14C	33.9 mb
28D21	34269163-34354638	85476 bps	14D	34.3 mb
27D09	34632975-34633740	766 bps	14D	34.6 mb
23O01	35770783-35790690	19907 bps	14E	35.7 mb
27B11	36523005-36523222	218 bps	15A	36.5 mb
26D18	37295601-37295830	230 bps	15B	37.3 mb
18K16	37806060-37906793	100734 bps	15C	37.9 mb
10I06	38516264-38516800	537 bps	15C	38.5 mb
28E 17	39730226-39851980	121755 bps	15D	39.8 mb
16D19	40198687-40292691	94005 bps	15D	40.2 mb
25P16	40318096-40433737	115642 bps	15D	40.4 mb
02F15	41215985-41341188	125204 bps	16A	41.3 mb
24L13	41886371-42012539	126169 bps	16A	42.0 mb
134C18	42670968-42800138	129171 bps	16B	42.7 mb
04P24	43342083-43443508	101426 bps	16B	43.4 mb
26F18	44249394-44414407	165014 bps	16D	44.3 mb
02D12	44916318-45014323	98006 bps	16D	45.0 mb
10A20	45292709-45293225	517 bps	16D	45.3 mb
10L15	46246786-46246992	207 bps	17A	46.2 mb
19G16	46247822-46341988	94167 bps	17A	46.3 mb
22F14	47697331-47697962	631 bps	17B	47.7 mb
02E 10	48098686-48210078	111393 bps	17C	48.2 mb
106D14	49419080-49540070	120990 bps	17C	49.5 mb
24J22	49885101-49923971	38871 bps	18A	49.9 mb
26E 04	50353027-50353954	928 bps	18A	50.3 mb
24D05	51713828-51797219	83392 bps	18B	51.7 mb
27F03	52012525-52012796	271 bps	18B	52.0 mb
26A09	53483379-53484123	745 bps	18C	53.5 mb
21H17	53985951-53986628	678 bps	18D	54.0 mb
10H17	55346482-55456897	110416 bps	19A	55.4 mb
12A12	56539815-56645102	105288 bps	19B	56.6 mb
03C20	57581772-57712731	130960 bps	19C	57.7 mb
17N16	58537341-58618071	80731 bps	19D	58.6 mb
AG4644	58955663-58969862	14199 bps	19D	58.9 mb
18M14	60797875-60881453	83579 bps	19D	60.8 mb
AAAB01002897	61285265-61545105	259840 bps	19E	61.4 mb

2L				
Clone	Coordinates	Size	Subdivision	Approximate Location
03C12	49180017-49180392	375 bps	28D	49.2 mb
03H03	48903738-49045784	142047 bps	28C	49.0 mb
22E 07	48282304-48283051	747 bps	28C	48.3 mb

27A15	47866905-47950880	83976 bps	28C	47.9 mb
07B16	47380027-47512326	132300 bps	28B	47.5 mb
20I11	46864317-47024697	160381 bps	28A	47.0 mb
21C23	46581675-46699000	117325 bps	28A	46.6 mb
30C19	45803169-45922407	119239 bps	27D	45.9 mb
25J06	45463046-45560094	97049 bps	27C	45.5 mb
01L03	45201201-45292054	90854 bps	27C	45.2 mb
10L02	44565994-44566735	742 bps	27B	44.5 mb
10L16	43731824-43820423	88600 bps	27B	43.8 mb
17K01	43171445-43292220	120775 bps	27A	43.2 mb
25I23	43077638-43171667	94030 bps	27A	43.1 mb
19P23	42027946-42150891	122946 bps	26D	42.1 mb
18C03	40866175-40867023	849 bps	26D	40.9 mb
30F11	40330482-40429255	98774 bps	26C	40.4 mb
20K10	39399062-39497969	98907 bps	26B	39.4 mb
25P19	39205770-39205951	182 bps	26B	39.2 mb
26A15	38391109-38518805	127696 bps	26A	38.4 mb
10D03	38025066-38103586	78521 bps	26A	38.1 mb
26D22	37376269-37455968	79699 bps	26A	37.4 mb
26I21	36772997-36887881	114885 bps	25D	36.8 mb
26H21	36262959-36325662	62704 bps	25C	36.3 mb
26D19	35047600-35138691	91092 bps	25C	35.1 mb
24E 18	34471078-34572118	101041 bps	25B	34.5 mb
26C12	33515028-33631083	116056 bps	25A	33.6 mb
17L24	33141060-33244353	103294 bps	25A	33.2 mb
19B01	32464131-32562330	98200 bps	25A	32.5 mb
29B16	31793773-31933685	139913 bps	24D	31.9 mb
24F06	31283028-31355655	72627 bps	24D	31.3 mb
21D04	30756235-30876679	120445 bps	24C	30.8 mb
16L07	30222685-30255418	32734 bps	24C	30.2 mb
28O12	29439277-29561250	131973 bps	24B	29.5 mb
21G06	28287907-28348216	60310 bps	24A	28.3 mb
10G23	27983129-28082937	99809 bps	24A	28.0 mb
30N11	26910489-27083474	172986 bps	23D	27.0 mb
10I22	25868124-25868757	634 bps	23D	25.8 mb
10C14	25122392-25202749	80358 bps	23C	25.2 mb
27F10	24493319-24585606	92288 bps	23C	24.5 mb
21I15	23653832-23785291	131460 bps	23B	23.7 mb
29J23	22727631-22794719	67089 bps	23B	22.7 mb
10A08	21916157-22003070	86914 bps	23A	22.0 mb
10H19	20724055-20724336	282 bps	23A	20.7 mb
06L01	20175680-20303395	127716 bps	22F	20.2 mb
29G01	19389088-19494735	105648 bps	22F	19.4 mb

10L10	18998750-19085684	86935 bps	22E	19.0 mb
04E 11	18745549-18845795	100247 bps	22E	18.8 mb
10A10	17846659-17943643	96985 bps	22D	17.9 mb
17F11	17424307-17517098	92792 bps	22D	17.5 mb
26E 18	17318003-17318294	291 bps	22D	17.3 mb
19D01	16535903-16661112	125210 bps	22C	16.6 mb
21G04	16140295-16141063	769 bps	22C	16.1 mb
28N13	15414400-15546992	132592 bps	22B	15.5 mb
26D01	14475811-14475967	157 bps	22A	14.5 mb
157I18	13513086-13564568	51482 bps	21F	13.5 mb
19K23	13022724-13088549	65826 bps	21F	13.0 mb
19E 23	11813482-11920675	107194 bps	21E	11.9 mb
10I10	11044526-11045054	529 bps	21E	11.0 mb
10E 14	10024766-10118285	93520 bps	21D	10.1 mb
02A19	8667457-8791152	123696 bps	21C	8.7 mb
10I01	8472251-8490508	18258 bps	21B	8.4 mb
17J22	7932810-8078933	146124 bps	21B	8.0 mb
03D18	6481155-6632781	151627 bps	21A	6.5 mb
AGAP004919	6015228-6016503	1275 bps	21A	6.0 mb
26D23	4833982-4987605	153624 bps	20D	4.9 mb
29K13	3929742-4008906	79165 bps	20D	4.0 mb
05B09	2076578-2103715	27138 bps	20C	2.1 mb
10I23	713431-713954	524 bps	20B	0.7 mb
AAAB01008942	1-373146	373145 bps	20A	0.2 mb

3R

Clone	Coordinates	Size	Subdivision	Approximate Location
31F12	403258-542070	138813 bps	29A	0.5 mb
21L08	1028113-1151127	123015 bps	29A	1.1 mb
16F07	1502553-1576633	74081 bps	29B	1.5 mb
27D21	1703994-1806113	102120 bps	29B	1.8 mb
26E 12	2823642-2939883	116242 bps	29C	2.9 mb
22E 11	3688636-3689245	610 bps	29C	3.7 mb
29C20	4350458-4470622	120165 bps	29D	4.4 mb
21G24	4717542-4718142	600 bps	29D	4.7 mb
27B18	5285103-5285877	774 bps	29D	5.3 mb
47N11	5654656-5655245	589 bps	29D	5.6 mb
24K22	6541607-6644135	102529 bps	30A	6.6 mb
18O09	7386082-7386689	608 bps	30A	7.4 mb
26D10	8316171-8432597	116426 bps	30B	8.4 mb
10B14	9231635-9330732	99098 bps	30C	9.3 mb
10B23	10035114-10035963	850 bps	30C	10.0 mb

17I13	10601831-10711262	109431 bps	30D	10.7 mb
30K06	11193135-11295446	102312 bps	30D	11.2 mb
08D06	11630201-11763014	132814 bps	30D	11.7 mb
03A13	11902025-12009143	107119 bps	30E	12.0 mb
28C17	12844438-12922860	78422 bps	30E	12.9 mb
19H09	13278257-13373944	95688 bps	30E	13.3 mb
28I07	14034080-14136840	102761 bps	31A	14.1 mb
03A07	14655464-14655679	216 bps	31B	14.6 mb
21K04	15582655-15701986	119332 bps	31B	15.6 mb
10F04	16038336-16118290	79954 bps	31C	16.1 mb
10J01	16420253-16516038	95786 bps	31C	16.5 mb
20G17	17203702-17203967	266 bps	31C	17.2 mb
30A12	18212628-18212994	367 bps	31D	18.2 mb
10C17	18326063-18427656	101594 bps	31D	18.4 mb
10D13	18995253-19069944	74692 bps	32A	19.0 mb
26H09	19929259-20029374	100115 bps	32B	20.0 mb
12K15	20602144-20714880	112737 bps	32B	20.7 mb
10E 05	21310055-21310697	643 bps	32C	21.3 mb
26K07	22474009-22474607	599 bps	32D	22.4 mb
10A02	23848710-23849261	552 bps	32D	23.8 mb
10A13	24910732-25051098	140367 bps	33A	25.0 mb
25M15	25546570-25694004	147434 bps	33B	25.6 mb
29J09	27123482-27237932	114451 bps	33B	27.2 mb
21P24	27712438-27827838	115401 bps	33C	27.8 mb
20N11	28278483-28398430	119948 bps	33C	28.3 mb
19F01	30050288-30145477	95190 bps	33D	30.1 mb
03K10	31398213-31398824	612 bps	33D	31.4 mb
24D07	32712805-32747359	34555 bps	34A	32.7 mb
22G20	33399839-33400526	688 bps	34A	33.4 mb
27D08	34036783-34181642	144860 bps	34B	34.1 mb
10E 16	34769414-34868489	99076 bps	34B	34.8 mb
17P09	35937380-36048686	111307 bps	34C	36.0 mb
17K02	36786896-36864853	77958 bps	34D	36.8 mb
19L19	37170326-37171099	774 bps	34D	37.2 mb
24J15	37740175-37855593	115419 bps	35A	37.8 mb
30M08	38737788-38737956	169 bps	35B	38.7 mb
10J15	40151038-40151873	836 bps	35B	40.2 mb
17F09	42064515-42182071	117557 bps	35B	42.1 mb
21L09	42364381-42478902	114522 bps	35C	42.4 mb
10L04	43322654-43417932	95279 bps	35C	43.4 mb
22A03	43969313-43969617	305 bps	35D	43.9 mb
29E 22	44336828-44436216	99389 bps	35D	44.4 mb
26B21	45129939-45130715	777 bps	36A	45.1 mb

25I16	45717226-45718021	796 bps	36B	45.7 mb
18M01	46669897-46777043	107147 bps	36B	46.7 mb
02O08	47339723-47433434	93712 bps	36C	47.4 mb
26J01	47794340-47931635	137296 bps	36C	47.8 mb
27D15	49112285-49204769	92485 bps	37A	49.2 mb
21N09	49461925-49607808	145884 bps	37A	49.6 mb
25J20	49896843-50027355	130513 bps	37B	50.0 mb
26D11	50901914-50968507	66594 bps	37C	50.9 mb
25H11	52037301-52131026	93726 bps	37D	52.1 mb
AAAB01008943	53027442-53200684	173242 bps	37D	52.3 mb
3L				
Clone	Coordinates	Size	Subdivision	Approximate Location
28G12	41538347-41662654	124308 bps	46D	41.6 mb
28J05	40982674-41158235	175562 bps	46C	41.1 mb
26C13	39947932-40027004	79073 bps	46C	40.0 mb
26E 14	39477560-39574860	97301 bps	46B	39.5 mb
19L20	38711515-38828618	117104 bps	46A	38.8 mb
26A18	38425601-38518561	92961 bps	45C	38.5 mb
30K01	37663495-37746990	83496 bps	45C	37.7 mb
03A05	36779452-36896807	117356 bps	45B	36.8 mb
04M03	36310592-36311033	442 bps	45A	36.3 mb
17B02	35989907-35990697	791 bps	45A	35.9 mb
26F21	35603550-35603750	201 bps	44D	35.6 mb
04J12	35198095-35198966	872 bps	44D	35.2 mb
24L05	34102108-34186748	84641 bps	44C	34.1 mb
25K09	33650483-33778541	128059 bps	44C	33.7 mb
16F10	32668248-32668459	212 bps	44B	32.6 mb
17M10	32182836-32269132	86297 bps	44B	32.2 mb
16J09	31994861-31995557	697 bps	44A	32.0 mb
21M06	31242758-31326722	83965 bps	44A	31.3 mb
26M18	30930157-30930693	537 bps	43D	30.9 mb
12A19	30310782-30311433	652 bps	43D	30.3 mb
10B19	28096581-28202451	105871 bps	43C	28.2 mb
27A19	27822263-27923021	100759 bps	43C	27.9 mb
24L10	26682367-26683169	803 bps	43B	26.7 mb
23O04	25560758-25561612	855 bps	43B	25.5 mb
03C01	24966909-24967087	179 bps	43A	24.9 mb
20B11	24542105-24673659	131555 bps	43A	24.6 mb
10K24	24035132-24035445	314 bps	43A	24.0 mb
10F21	23713939-23804027	90089 bps	42C	23.8 mb
27N14	22117399-22118054	655 bps	42C	22.1 mb

26D21	21715772-21834821	119050 bps	42B	21.8 mb
23G12	21230628-21335302	104675 bps	42B	21.3 mb
12D12	20619497-20747455	127959 bps	42A	20.7 mb
26B23	20094612-20207522	112911 bps	42A	20.2 mb
10D12	19690217-19806477	116261 bps	42A	19.7 mb
07D16	19098444-19201647	103204 bps	41D	19.2 mb
24E 23	18848877-18974610	125734 bps	41D	18.9 mb
27K08	18002187-18143178	140992 bps	41C	18.1 mb
24K17	16731209-16731948	740 bps	41B	16.7 mb
26H23	16177527-16177857	331 bps	41A	16.2 mb
08F18	14578788-14693933	115145 bps	41A	14.6 mb
16M11	13406089-13406948	860 bps	40C	13.4 mb
26C11	13275184-13356588	81404 bps	40B	13.3 mb
26I14	12208424-12298365	89942 bps	40B	12.2 mb
10G14	11977066-11977380	315 bps	40B	11.9 mb
17C24	10692973-10790986	98014 bps	40A	10.7 mb
06P09	7548199-7659413	111215 bps	39B	7.6 mb
25P21	6538752-6675283	136532 bps	39A	6.6 mb
26H17	5258364-5350971	92608 bps	38C	5.3 mb
AGAP010482	4264713-4266440	1727 bps	38C	4.2 mb
26L19	2920705-2921251	547 bps	38B	2.9 mb
28I19	2275699-2437989	162291 bps	38A	2.3 mb
AGAP010342	1807945-1815119	7174 bps	38A	1.8 mb
AAAB01008981	1-219224	219223 bps	38A	0.1 mb