MILLER, KATELYN JO, M.S. Construction and Fine-Scale Analysis of a High-Density, Genome-Wide Linkage Map to Examine Meiotic Recombination in the Honey Bee, *Apis mellifera*. (2014)
Directed by Dr. Olav Rueppell. 94 pp.

The western honey bee, *A. mellifera*, is an important biological model organism in research for ecological and behavioral studies in addition to molecular studies. Honey bees are also imperative in nature for reproduction and diversification of plants via pollination. A unique feature of honey bees is that they have the highest recombination rate of all metazoans. This gives rise to the important question: what causes honey bees to have such a high rate of recombination? The honey bee genome has already been sequenced, but the available linkage maps are not detailed enough to characterize individual recombination events at the genome level. High recombination rates in honey bees may be caused by abundant recombination hotspots found throughout the genome. Resequencing the honey bee genome with next-generation sequencing and using over 900,000 markers genome-wide to identify recombination events showed that recombination rate in honey bees may be underestimated. This study calculated the average recombination rate to be 178.7 cM/Mb as opposed to the second most recent average of 22 cM/Mb. These high recombination rates in this study could be explained by mistakes in the current assembly of the reference genome. Further analyses are necessary to verify proper assembly of the current reference genome, genome-wide recombination events, and recombination rates. Based on the verified data set it will then be possible to confirm whether hotspots are present in honey bees and to correctly correlate recombination hotspots to sequence motifs.

CONSTRUCTION AND FINE-SCALE ANALYSIS OF A HIGH-DENSITY, GENOME-WIDE LINKAGE MAP TO EXAMINE MEIOTIC RECOMBINATION IN THE HONEY BEE, *APIS MELLIFERA*

by

Katelyn Jo Miller

A Thesis Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Greensboro
2014

Approved by

_____
Committee Chair

APPROVAL PAGE

This thesis has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair _____

Committee Members _____

_____

_____
Date of Acceptance by Committee

_____
Date of Final Oral Examination

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Olav Rueppell for providing continual guidance, encouragement, and mentorship throughout my project and first two years of graduate school. I also thank my committee members Dr. David Remington and Dr. Malcolm Schug for their support and suggestions throughout the project.

A big thank you goes to Caitlin Ross for using her expertise in computer programming, assisting with my project, and teaching me the basics of computer programming. I would also like to thank our collaborator Dr. Corbin Jones at UNC-Chapel Hill and everyone at the High Throughput Sequencing Facility for allowing me to work in their facility and utilize their tools to better my project, and also thank you to Dr. Ed Vargo and his lab at NCSU for allowing me to use their equipment when ours were out of commission.

This project could not have been completed without the support of my family, friends, all members of the Rueppell lab, UNCG Biology Department, and all the graduate students. Lastly, I would like to thank the NIH for funding this project.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

Honey Bee Importance

The western honey bee, *Apis mellifera,* is a unique organism and useful experimental model. Honey bees have become of great interest due to their significant contributions to pollination services and science through behavior and molecular studies. Most recently bees have sparked worldwide attention due to the alarming decline of bee populations. Honey bees play a major role in agriculture because they pollinate almost 1/3 of all crops (Losey & Vaughan, 2006). Certain crops, specifically almonds, rely solely on honey bees for pollination; other crops also utilize honey bee pollination such as blueberries, apples, cherries, broccoli, and melons (Klein et al., 2007). Pollination by bees contribute up to 14 million dollars in crop value per year in the United States as of 2000 and up to 200 billion dollars worldwide as of 2005 (Gallai, 2008). Pollination has become of such necessity to farmers that services of honey bee distributors are created to ship hundreds of colonies across the country (Potts et al., 2010). Honey bees also play an important role for human genome studies since the 2006 sequencing of the honey bee genome. Sequencing revealed that honey bees, on a genomic level, are more similar to vertebrates than *Drosophila* with regards to DNA methylation, RNA interference, and circadian rhythm (Honeybee Genome Sequencing Consortium, 2006). Therefore,

genomic studies of the honey bee will not only allow further understanding of general bee biology, but also the human genome.

Eusociality in Honey Bees

A. mellifera is categorized as eusocial. Members of eusocial insects are from the order Isoptera (termites) and Hymenoptera including species in the family Vespidae (wasps), Formicidae (ants), and Apidae (bees) (Hermann, 1979). By definition, eusociality involves the division of labor, overlapping generations, and cooperative behavior (Wilson, 1971). Division of labor can be categorized through castes. Castes are sets of individuals within a colony that have a specialized function (Oster & Wilson, 1978). The reproductive castes consist of drones who provide sperm for mating to a queen outside their colony and queens who produce eggs for future offspring (Oster & Wilson, 1978). The non-reproductive caste of bees consists of workers who have the important jobs of maintaining the hive and taking care of both brood and the queen (Winston, 1987). For every honey bee colony, there is one queen who will produce both sterile workers and fertile drones. Drones are the only males in the colony and do not contribute work to the hive (Free & Williams, 1975). Unlike the female workers and queens, male drones are produced from unfertilized eggs. Thus, they only have one set of chromosomes and are genetically identical to their mother's gamete (Hamilton, 1964; Kerr, 1962; Whiting, 1945). Honey bees are organisms that are able to have viable offspring without fertilization of the eggs because of a type of asexual reproduction known as parthenogenesis. Honey bees and some other insects undergo a form of

parthenogenesis called arrhenotoky, or haplodiploidy, where unfertilized eggs become males (Whiting, 1945). Haploid males are thus useful in recombination studies because any genotypic variation among them is due to the independent assortment of chromosomes and meiotic recombination, while variation in workers may be due to different paternity.

The evolution of the eusociality of insects can ultimately be explained by an individual's inclusive fitness contribution for future offspring or the theory of kin selection and altruism (Hamilton, 1964). Kin selection can lead to an individual's sacrifice of its personal reproduction and even its life to increase the fitness of its relatives (Hamilton, 1964). This model uses the coefficient of relatedness, or *r,* which is the probability of two individual's homologous alleles are identical by descent (Hamilton, 1964) to weigh the cost and benefit of a potentially altruistic behavior. In animal behavior, altruism is the act of an individual, the altruist, which benefits others at an apparent fitness cost for the altruist. There has been much debate over the evolution of social insects and its connection to kin selection and inclusive fitness (Trivers, 1971; Trivers & Hare, 1976; Wilson, 1998). Eusocial Hymenoptera queens are known to mate multiple times or have a high degree of polyandry (Page, 1986). Honey bee queens, depending on the species, mate on average 8 to 27 patrilines (Estoup, 1994; Oldroyd et al., 1997; Tarpy & Page, 2002). Thus, female workers' relatedness to their sisters is more distant than previously believed which leads to the argument that kin selection cannot explain the evolution of eusociality (Trivers, 1971). However, single mating is

presumably ancestral to all eusocial insects lineages, which indicates the importance of kin selection (Hughes et al., 2008).

Meiotic Recombination

Meiotic recombination is an important biological function for organisms that undergo sexual reproduction. Recombination involves the exchange of genetic information of both maternal and paternal alleles in order to produce genetically diverse gametes and progeny. Meiotic recombination is also important for facilitating proper segregation of chromosomes and maintaining stability of the genome (Coop & Przeworski, 2007). Understanding meiotic recombination function and mechanisms is also important for understanding the evolution of sexual reproduction, adaptive evolution, and genetic selection (Otto & Barton, 2001; Rice, 1983). *A. mellifera* has now become a unique and useful model organism to study the evolution of recombination because they have the highest rate of recombination in all metazoans with average rates ranging from 19-22 centimorgans/megabase (cM/Mb) (The Honeybee Genome Sequencing Consortium, 2006). Other social insects, especially within Hymenoptera, also have high recombination rates (Wilfert, Gadau & Schmid-Hempel, 2007). Taxa with a refined division of labor system or those who are highly eusocial have higher rates of recombination while social insects that are more primitive in division of labor have somewhat lower recombination rates (Wilfert et al., 2007).

Evolution of Recombination in Social Insects

Division of labor in social insect colonies may benefit from genetic diversity among colony members, which may explain the evolution of high recombination within social insects because recombination increases diversity of future progeny (Sirvio et al., 2011). Genetic diversity is also important in the evolution of disease resistance in social insects and has been used to explain high recombination rates in eusocial insects (Hughes & Boomsma, 2004; Tarpy & Seeley, 2006). In an opposing view, simulations showed that recombination did not increase genotypic variance of quantitative traits in social insect colonies (Rueppell, Johnson & Rychtar, 2008; Rueppell, Meier & Deutsch, 2012).

Genome Features of *Apis mellifera*

Sequencing the honey bee genome has revealed other interesting genome characteristics that may be related to sociality and to honey bee biology. Compared to other insects such as *Drosophila melanogaster* (fruit fly) and *Anopheles gambiae* (mosquito), honey bees have a high A+T content(67%) and a relatively high frequency of CpG motifs (Honeybee Genome Sequencing Consortium, 2006). The Honeybee Genome Sequencing Consortium (HGSC) also found that in many aspects, honey bees are more similar to vertebrates than fruit flies and mosquitoes (2006). Honey bees and vertebrates have similar genes that encode for RNA interference, DNA methylation, and circadian rhythm (Honeybee Genome Sequencing Consortium, 2006). The mean sequence identity between the human genome and the honey bee genome is more similar at 47.5% in comparison to the mean sequence identity between the human and fly genome at 44.5% and the human and mosquito genome at 46.6% (Honeybee Genome Sequencing

Consortium, 2006). Honey bees have also maintained 80% of ancient introns from their common ancestors with vertebrates (Honeybee Genome Sequencing Consortium, 2006). Honey bees are also similar to humans because of the presence of telomerase (Honeybee Genome Sequencing Consortium, 2006). The repeat sequence of TTAGG is found on both distal and proximal chromosomes and the honey bee genome contains a gene that has a 23% similarity to the human *TERT* protein, a subunit for telomerase (Honeybee Genome Sequencing Consortium, 2006).

Sequence Motifs and Meiotic Recombination

Studies using different models, most importantly *Drosophila*, have discovered that specific sequence motifs influence recombination events. These motifs could benefit and translate towards honey bee recombination studies. Specific nucleotide sequences were discovered that correspond to recombination events in *Drosophila melanogaster* (Comeron, Ratnappan & Bailin, 2012). The study was able to distinguish between cross-over and gene conversion rates in genomic and population variation and revealed that in *D. melanogaster* there are more sequence motifs associated with recombination events than in humans and mice (Comeron et al., 2012). Several studies in humans have examined potential recombination hotspots at a fine scale level (Myers et al., 2005) with approaches as suggested here for my study. Recombination hotspots or areas where high rates of recombination events have been of great interest in human genome studies to understand patterns in linkage disequilibrium (LD) and create disease-associated linkage maps relating to recombination (Zheng et al., 2010; Zhou et al., 2013).

In humans, more than 25,000 recombination hotspots have been identified (Myers et al., 2005) while 3,604 hotspots have been indentified in yeast (Pan et al., 2011). In contrast these high frequency recombination hotspots, *Drosophila melanogaster* appear to have a low frequency of recombination hotspots (Comeron et al., 2012) with 10 putative hotspots (Chan, Jenkins & Song, 2012). Oligonucleotide motifs such as CCTCCCT and CCCCACCCC are believed to be associated with hotspot locations and the promotion of hotspots in humans (Myers et al., 2005; Stevison & Noor, 2010). The *prdm9* gene and the PRDM9 protein is also of interest to understand hotspots since the zinc fingers of the protein could bind to regions of the DNA to initiate recombination (Baudat et al., 2010). Since honey bees have similarities to both humans and fruit flies in other regards, this current knowledge about hotspots may be applied to the honey bee genome. However, since fruit flies and humans have low rates of recombination in comparison to the honey bee, more research is necessary to determine whether the concept of hotspots and motifs can be applied to bees, whether the number of such hotspots is increased in honey bees relative to humans and fruit flies or whether the intensity of the hotspots is higher in honey bees.

Honey Bee Recombination Mapping

Recombination in bees has been heavily researched since the construction of the first honey bee genetic linkage map using random amplified polymorphic DNA (RAPD) markers, leading to initial indications of high recombination rates (Hunt and Page, 1995). High rates of recombination events are seen particularly within the sex determining locus

of the genome (Beye et al., 1999). Microsatellite linkage maps with hundreds of annotated loci have also been an advancement in understanding recombination and genetic linkage in honey bees (Solignac et al., 2004). Construction and analysis of detailed linkage maps have led to the discovery that honey bee recombination events are genome wide, occurring similarly across all chromosomes regardless of size (Beye et al., 2006; Solignac et al., 2007). However, the recombination rate is not constant and hotspots may exist in certain genome regions, which could not be ascertained due to the limited resolution of the existing linkage maps (Solignac et al., 2007). Recombination rate has also been associated with genes that are important for the evolution of the behavior of worker bees (Hunt et al., 2007). Recombination rate and GC content are correlated at a fine scale and both show strong correlations to the rate of molecular evolution (Kent et al., 2012). It has also been proposed that eusocial insect genome structure and recombination may co-evolve via a feedback loop influencing queen and worker behavior phenotypes by natural selection and recombination (Kent & Zayed, 2013).

Potential sequence motifs for recombination hotspots (CGCA, GCCGC, and CCGCA) have been identified in honey bees (Bessoltane et al., 2012). However, a higher resolution of the honey bee recombination map assembly is necessary to confirm that previously identified motifs for recombination signals also occur in honey bees (Bessoltane et al., 2012). All of the recent analyses of honey bee recombination rely on the sequencing data generated by the genome project (Honeybee Genome Sequencing Consortium, 2006). However, sample size for the sequencing project is unclear because a

combination of haploid drones ranging from 20 to 100 individuals was used. The empirical data were also limited by genome coverage. For example, only 3.5% of the genome was screened when searching for crossover events (Bessoltane et al., 2012). My project sought out to generate data of the entire genome to improve our understanding of honey bee recombination patterns.

Single nucleotide polymorphisms (SNPs) are differences in single base pairs that are found throughout a genome. SNPs can be caused by insertions or deletions of single nucleotides and by DNA transversions, the substitution of a purine to a pyrimidine or a pyrimidine to a purine, or transitions, purine/purine and pyrimidine/pyrimidine substitutions. Individuals may have varying alleles or sequence alternatives (Brookes, 1999). SNPs also exist as differences between homologous chromosomes within individuals and thus can be used to identify recombination events throughout a genome. SNPs are relatively common, making them good markers for high-resolution genotyping across the genome. Before my study, no genome-wide, high resolution recombination map for honey bees was available.

Hypothesis, Aims, and Rationale

I hypothesized that high recombination rates in honey bees are due to frequent recombination hotspots found throughout the genome. I predicted that I should observe thousands of recombination hotspots genome-wide with higher frequency, in comparison to humans. Alternatively, more intense hotspots at a lower frequency, closer to *Drosophila* estimates, may also explain high recombination in honey bees. Therefore, if

hotspots are present, I should see regions with very high recombination rate occurring only a few times throughout the genome. I completed three specific aims to test these hypotheses:

1) Generate a mapping population of high quality genomic DNA.

2) Sequence drone samples to identify SNPs and created a fine scale, high density map.

3) Analyze recombination patterns to determine potential hotspots and to characterize the associated sequence features.

My project differs from other recent projects since it did not use the existing, pooled data from the honey bee consortium, but resequenced the genome using extracted DNA from drones collected from one experimental mapping population. My approach was the first genome-wide evaluation of recombination patterns, aiming at single-nucleotide resolution. The project relied on the newest genome assembly (Amel_4.5: http://hymenopteragenome.org/beebase/). My project provides next generation sequencing data from the Illumina™ (San Diego, CA) platform by shallow whole genome resequencing.

CHAPTER II

METHODS

Sample Collection: Aim 1

During the summer of 2012, drone pupae were removed from their cells of a comb harvested from a single, unselected colony in the UNCG bee yard. All individuals were collected from the same colony, suggesting that all drones came from the same mother. Drone cells were distinguished from worker cells by the size of the cell. Since drones were larger in size they had a larger cell, and the capping of the cell was dome-like while capped worker cells were relatively level or flat (Snodgrass, 1984). The drones were placed in 1.7 mL centrifuge tubes and stored in a -20°C freezer. In total, ~1300 drones were collected.

Extraction of Genomic DNA: Aim 1

Different protocols for extracting genomic DNA from the collected samples were tested to compare the yield and determine the best method. Cell DNA, mouse tail DNA, and tissue DNA protocols from Maxwell® automated extraction kits (Promega™, Lincoln, NE) were tested. Manual protocols for DNA extractions from tissue and mouse tail tissue were also tested with Wizard® Genomic DNA purification Kit (Promega™, Lincoln, NE) and with a Puregene® kit (Qiagen,Inc., USA). Among the tested

procedures, the mouse tail tissue protocol from Qiagen Puregene® kit yielded the most DNA and was therefore used for all DNA extractions.

Genomic DNA from 800 drone individuals was extracted following the Qiagen protocol for mouse tail tissue. Individuals chosen for extractions were in the pupa stage of life. Most were white in color, while other, more mature pupae were more yellow or tan in color and had a purple pigmentation to their eyes. These samples were kept frozen at -20°C until the extractions. Only the thoraces of the drones were used in the DNA extraction to avoid contaminants from the abdomen and the eyes. The abdomen contains the gut including waste and enzymes that may be problematic for DNA extractions. The eyes also contained eye pigment that may inhibit downstream reactions (Boncristiani et al., 2011). The thoraces were separated from the other body parts while still frozen using two forceps. The forceps were sterilized using a Bunsen burner between each use. Each thorax was placed in a new, autoclaved centrifuge tube. While still frozen, each thorax was ground up using a melted pipet tip that had been molded to fit the shape of the centrifuge tube. Approximately100 µL of the tissue was taken and placed in a new centrifuge tube with 1.5 µL of proteinase K and 300 µL of cell lysis solution. ProteinaseK helped in the breakdown of protein found within the tissue and the cell lysis solution helped with the breakdown of the cell in order to retrieve genomic DNA from the nucleus. The mixture of the reagents was left in a 55°C incubator overnight to increase breakdown of both proteins and cells. Subsequently, 1.0 µL RNase was added to the mixture and incubated at 37°C for 30 minutes to break down RNA within the thorax. The samples were cooled on ice and 100 µL of protein precipitation solution was added

to the tube before vortexing vigorously. Each tube was centrifuged at high speed (10,000 to 14, 000 RPM) to collect the solid protein material at the bottom of the tube ensuring the pellet was very tight. The supernatant was added to a new tube containing 300 µL isopropanol, avoiding the protein pellet and the top layer of potential lipid matter. The new mixture was inverted several times so the DNA strands would precipitate. Again, the mixture was centrifuged at high speeds, creating a pellet containing the DNA. The supernatant was discarded and 300 µL of 70% ethanol was added to wash the pellet and then centrifuged again. Extra ethanol was drained and the tube was left to air dry. 100 µL of Tris-EDTA (TE) buffer was added to rehydrate the DNA pellet and resuspend the DNA. Incubation at 65° C for one hour helped dissolve the pellet. The DNA solution was left to incubate in the refrigerator at 4°C overnight and then stored at -20 °C until further use.

Quality Control of Extracted DNA Samples: Aim 1

Once the DNA was extracted from the thorax using the Qiagen kit, each sample was tested to ensure both high-quality and yield. A Nanodrop™ spectrophotometer was used to quantify the DNA in the sample and to verify its purity. Two microliters of each sample were tested. Purity of the sample was measured by the deviation from the ideal 1.8 260/280 absorption ratio. Based on the absorption ratio, quantity was measured in nanograms per microliter, and samples selected with minimum yield of at least 100 ng/µL for a 100 µL sample.

Samples that met the quantity standards mentioned above were then tested for quality by gel electrophoresis. Quality standards of the samples included high-molecular weight, non-degraded DNA, and absence of RNA. Five microliters of gDNA were loaded on a 0.5% TBE agarose gel and underwent electrophoresis at 90 V for two hours. The gel was stained with ethidium bromide to verify the presence and the quality of the DNA. Of the 800 drones, the 192 individuals with highest quality and quantity of DNA were used for subsequent procedures.

Verification of Sample Identity through Polymerase Chain Reaction and Microsatellite Genotyping: Aim 1

Each of the 192 individuals was tested at three microsatellite loci to verify that each drone came from the same mother and that each individual was indeed haploid. Seven sets of primers were tested to determine the best primer pairs for genotyping three microsatellites. Polymerase chain reaction (PCR) was the first step for sample identification and also to determine if the gDNA was amplifiable. Each 10X master mix included 89.5 µL of molecular grade water, 15 µL of 10x buffer, 15 µL of 200 µM dNTPs (deoxynucleotide triphosphates), 3.5 µL of 0.25 µM forward primer, 7.5 µL of 0.5µM reverse primer, 7.5 µL of LI-COR 700 IRDye, and 2.0 µL of 0.2 µTaq polymerase (Hayworth et al., 2009). The mixture was vortexted vigorously to mix. Each PCR reaction consisted of 14 µL of the master mix and 1 µL of DNA template. After going through the necessary temperature cycles (Hayworth et al., 2009), the PCR products were tested by electrophoresis on a 1.0% agarose gel running for approximately 1.5 hours at

120 V. Ethidium bromide stained the gel for 30 to 60 minutes and the gel was observed under ultra violet light for the presence of amplified product. If product was present, samples were genotyped to verify that they were derived from the same queen.

Genotyping of the samples was done by gel electrophoresis with the LI-COR 4300 DNA analyzer (Licor Inc., Lincoln, NB) using the IRD700 color label for microsatellite alleles (Dixon et al., 2012; Schuelke, 2000). Loci with different sizes can be distinguished from other loci when run on the same gel. From the seven tested primer sets, three were chosen for genotyping: K0907, K05128, and AP174 (Solignac et al., 2007). The loci were chosen based on their amplification, varied sizes, and level of polymorphism in the preliminary samples. PCR was performed using the three chosen primers on all 192 samples. Based on the relative amplification strength of the K0907 product, it was diluted 1:10 with water; 10 µL of the diluted product was combined with 2.5 µL of loading buffer and 1 µL of the mixture was loaded on the gel. The other loci, K05128 and AP174, were combined and used undiluted. Before loading on the gel, samples were denatured in a thermocycler at 95°C for five minutes. Denatured samples were loaded into the wells of the polyacrylamide gel using glass syringes. Two size markers were loaded at both ends of the gel. The samples underwent electrophoresis using a polyacrylamide gel where the DNA was detected via a laser and detector. The IRD (infrared dye) within the prepped samples produced fluorescent data and a real-time image of the gel containing the genotype of the drones was produced. The genotypes verified that all drones came from the same mother because maximally two alleles were present at each locus.

DNA Library Preparation and Quality Control: Aim 2

For each individual to be sequenced, 96 samples were combined one multiplexed sequencing library. All 192 samples were taken to the High Throughput Sequencing Facility (HTSF) at UNC-Chapel Hill for library preparation, construction, and sequencing. I went to the HTSF and assisted with the creation of the libraries where human preparation was necessary for the semi-automated protocol, performed by a G3 by Caliper Life Sciences ™ robot.

The genomic DNA was first sonicated, meaning the DNA was sheared into small, workable fragments for library construction and sequencing. Genomic DNA was sheared by the E220 focused-ultrasonicator (Covaris Inc., Woburn, MA) which used Adaptive Focused Acoustics™ (AFA) technology. For each sample, 55 µL of genomic DNA was pipetted into a 96- well sonication plate. These plates have glass wells with tubes containing AFA fibers that aid in the sonication step. Once samples were transferred into the new plate, it was sealed with an aluminum film and placed into the ultrasonicator. A preset computerized program controlled the robot to take the plate and submerged it into the cool water bath. For two hours, the ultrasonicator individually fragmented the DNA in each of the 96 wells. After sonication, the samples were transferred to a standard 96-well plate and sealed in preparation of quantifying the amount of DNA in the 55 µL samples. The majority of samples had quantities lower than the target 1 µg of total DNA. Many samples had less than 0.5 µg of total DNA with the lowest quantity at 100 ng of total DNA.

Libraries were constructed by using a high throughput (HTP) library preparation kit (KAPA BioSystems, Woburn, MA) for the HiSeq 2000 platform (Illumina Inc., San Diego, CA). This kit was specifically designed for high throughput projects using four main reactions: 1) end repair, 2) A-tailing, 3) adapter ligation, and 4) library amplification. The end repair reaction creates blunt ends with fragments that are 5'-phosphorylated. A-tailing reactions entail the addition of dAMP (deoxyadenosine monophosphate) to double-stranded DNA (dsDNA) at the 3' end. The adapter ligation attaches the adapter to the fragmented DNA. The dsDNA adapters attach to 3'-dTMP (deoxythymidine monophosphate) overhangs to correspond with the A-tailed fragments. Library amplification increased the amount of library fragments that have the adapter sequences using PCR (KAPA BIOSYSTEMS, Woburn, MA).

In the first step of each library preparation, 50 µL of double stranded, sheared DNA was incubated with 20 µL end repair mix for 30 minutes at 20 ºC. Components of each end repair mix included 10x repair buffer (7 µL), end repair enzyme (5 µL), and water (8 µL). To the mixture, 120 µL of paramagnetic SPRI (solid phase reversible immobilization) beads were added and incubated at room temperature (RT) for 15 minutes for proper binding of DNA to the beads. The sample plate was placed on a magnet to collect the beads at the bottom of the well and the supernatant discarded. The beads that contained the DNA were washed with 200 µL of ethanol twice at RT for 60 seconds. Dried beads with the end repaired DNA were added to 50 µL of A-tailing master mix. Reagents in the A-tailing master mix include 10x buffer (5 µL), A-tailing enzyme (3 µL), and water (42 µL). The procedure was repeated as before with the mixing

of the reagents, incubation temperature at 30° C for 30 minutes. To each 50 µL A-tailed sample, 90 µL of PEG (polyethylene glycol) was added. Dependent on specific concentration, PEG allows for size-selective binding of DNA to magnetic beads (Lundin et al., 2010). The beads were washed twice with 200 µL of ethanol, incubated at RT for 60 seconds, and dried. To the A-tailed DNA, 45 µL of ligation mixture and 5 µL of adapters were added and incubated at 20°C for 15 minutes. Adapter ligation master mix contained 5x buffer (10 µL), DNA ligase (5 µL), water (30 µL) and the adapters (5 µL). For each 50 µL of adapter ligation reaction, 50 µL of PEG solution was added and incubated at RT for 15 minutes. The supernatant was removed; the beads were washed twice with 200 µL of ethanol, incubated at RT for 60 seconds, and let dry. The cleanup was done a second time before the beads were transferred into100 µL of resuspension buffer. After thorough mixing, the supernatant containing the DNA was removed and used for size selection.

PEG solution and magnetic beads were also used to size select adapter ligated DNA at approximately 300 bp. The first size selection was to select DNA fragments smaller than ~450bp. Small volumes of PEG (60 µL) allowed for fragments 450 bp or greater attach to the beads, leaving smaller fragments in the supernatant. The supernatant with fragments of less than 450 bp was transferred to a new plate and 20 µL of beads were added to the supernatant, mixed thoroughly, and incubated for 15 minutes at RT. Using higher concentrations of PEG, DNA fragments larger than 250 bp bound to the beads, while smaller fragments remained in solution. The supernatant was discarded and the beads were washed twice with 200 µL of 80% ethanol for 60 seconds, and incubated

at RT. The beads containing the target sized DNA were added to 25 µL of resuspension buffer. The supernatant that contained the DNA libraries was combined with 25 µL of PCR master mix to amplify the libraries of appropriate size fragments. The master mix contained 2x HiFiHot Start Ready Mix and Trueseq primers (5 µL). The PCR was performed according to manufacturer directions (KAPA BioSystems, Moburn, MA) 25 total cycles and shorter overall time due to the short fragments and to avoid amplification of repeats and remainder adapters.

After PCR, 50 µL of additional beads were added to 50 µL of amplified libraries, mixed, and incubated at RT for 15 minutes. The supernatant was discarded and the beads were washed twice with 200 µL of ethanol. Ethanol was removed to allow the beads to dry and the beads were added to 45 µL of elution buffer (10 mM Tris-HCl, pH 8.0) containing 0.1% tween. The plate was placed on the magnet to collect the magnetic beads and the clear supernatant was transferred to a new plate. Completed libraries were tested on a micro-gel to verify purity, quantity, and proper size selection.

Each library was tagged with 2D, indexed adapters or barcodes. Indexed adapters contain specific primers (oligonucleotides) and are tagged to the sample sequence in order to be identified after sequencing and be able to demultiplex in pooled samples (Meyer et al., 2007; Peterson et al., 2012). Dual adapter (2D) indexes can help with inaccuracies seen in multiplexed samples since each adapter incorporating an index increases specificity to the sample (Kircher, Sawyer & Meyer, 2012). Adapters were diluted 1.5x from their standard concentration to account for low total DNA quantities.

Samples were multiplexed into two separate pools, each containing 96 individual libraries. Each of the two pools should have a final concentration of approximately 15 nM. The pooled samples were quality control tested; concentration measured on a Qubit Fluorometer using the broad range setting. Pools were stored at -20°C until sequencing.

Illumina™ Sequencing: Aim 2

Sequencing of the libraries was done on an Illumina™ HiSeq 2000 (San Diego, CA). Each of the two pools was placed on one lane of the sequencer with 100 bp single end read parameters. Illumina™ (San Diego, CA) sequencing uses the sequencing by synthesis (SBS) technology. Two runs of paired-ended 100bp runs were added to the output to obtain better coverage and detail. Thus, a total of three runs per individual were used for alignment and analysis.

Sequence Alignment to the Reference Genome: Aim 2

The genome sequence data for each of the 16 chromosomes for honey bees were downloaded from the NCBI website (ftp://ftp.ncbi.nlm.nih.gov/genomes/Apis_mellifera/) and concatenated together to form a complete genome. Each sequence read was aligned back to the reference genome using the Burrows-Wheeler Alignment (BWA) tool (Li & Durbin, 2009). Since the genome of the honey bee has been sequenced and annotated, sequence realignment back to the reference genome gave an accurate idea of the physical location of each. Realignment was done for each of the 192 individuals separately for the paired-end reads and the single end reads. The sequencing read data were in the FASTA format; BWA used these files and the reference genome file to output .sai files which

contained the aligned reads. All files were then converted to the SAM format which allowed me to visualize the data (Figure 1). To facilitate subsequent computing, the SAM files were then converted to BAM files, the binary form of the file. For all processes, I collaborated with the undergraduate student Caitlin Ross to create scripts to automate the process for 192 individuals.

Detection of Single Nucleotide Polymorphisms (SNPs): Aim 2

Once alignment was complete, each individual underwent a filtering step within SAMtools (Li et al., 2009) that only kept reads that properly aligned to the reference genome. Reads that did not align properly may have been due to sequencing errors or contamination. The filtered reads from the three separate run per individual were merged together with the merge function in SAMtools. This resulted in 192 files that were further processed using the sorting option in SAMtools to order all the aligned reads according to their genomic position. Individuals were analyzed for detection of single nucleotide polymorphisms (SNPs) with the SAMtools mpileup and BCFtools option (Li et. al., 2009). The program identified the sequenced reads that carried deviations from the reference genome, including SNPs and INDELS. The results for each individual were stored in a file in the variant call format (VCF) for each individual contained a list of data for each discovered SNP, including contig location of the SNP, physical position of the SNP on the contig, the reference base call, and actual SNP call.

SNP Matrices: Aim 2

Individual VCF files from all individuals were combined into one SNP matrix file for each chromosome. Each chromosome file contained the genotype information for all 192 individuals and the reference genotype, contig identifiers and position in rows, for all SNPs ordered according to their physical position in the reference genome (Amel4.5). Individuals that did not have a record for a particular marker were assigned a "Missing" genotype. Another script was created to count all reported genotypes and "Missing" data. To account for sequencing errors, markers with less than 20 counts or more than 172 counts for the major allele were removed from the matrix. Keeping SNPs that had genotype counts between 20 and 172 allowed for loci with >20% minor allele frequency. Due to our previous procedure, loci that included the reference genotype as one of the segregating variant were coded as "Missing" in individuals that were identical to the reference genome. Therefore, a new reference genome had to be generated that incorporated the alternative alleles in the mapping population to determine whether "Missing" data was not detected due to missing coverage or due to reads being identical to the original reference genome.

Creation of an Alternative Reference Genome: Aim 2

Creation of a new reference genome was necessary to account for individuals in our mapping population that had variants that were identical to the reference. In order to determine "Missing" data were not called because of missing coverage or having identical reads to the original reference, a program was created to take the matrix data

from each chromosome and compare it to the original reference genome. The program went into the original reference genome file and compared the SNP calls and "Missing" data and determined if the calls were the same as the original or if it was a new call. The new reference genome did not include INDELs.

All raw sequencing data was processed again with BWA and SAMtools using the new reference genome. The VCF files that contained the SNPs for all individuals were processed again and new SNP matrices were created. The same scripts were used to count the genotype frequencies of each SNP. A script generated quality scores based on the frequency of the genotypes for each SNP. Scores ranged from 1 to 5; SNPs with a 1 quality score had a minor allele frequency of greater than 80 counts or more. SNPs with quality scores 2 to 4 were assigned based on minor allele frequency of $\geq$60, $\geq$40, and $\geq$20 respectively. A quality score of 5 was assigned if a third allele was greater than half of the 2nd genotype. The program counted the number of SNPs for each quality score and removed SNPs that had genotypes with less than a frequency of 20 counts for a genotype.

Additional analysis steps were included to remove low quality markers. For each chromosome and marker, a script compared the linkage of each locus at four different intervals: it determined linkage to the adjacent maker and the following three markers. Markers were eliminated that showed a recombination frequency of larger than 5% because this was considered unrealistic, given an average marker spacing of <250 base pairs.

Subsequently, marker genotypes were converted to the alternative phases based on the linkage patterns, starting with each genome contig by assigning an arbitrary phase designation to the two alternative alleles. Genotype calls that did not correspond to the two segregating alleles were converted to "Missing" data. Thus, individual data points were coded as 0, 1, or 2:" 0"'s represented missing data, and "1"'s and "2"s corresponded to the two alternative phases. Phase switches between 1's and 2's were used to identify either gene conversions or crossover events. Intervals between two different genomic contigs were ignored because they lacked exact physical length estimates. In some instances, our data suggested errors in the contig order orientation. Tract lengths were used to categorize whether a phase switch was categorized as a recombination or gene conversion event. If the tract length between two phase switches (must be considered a 1 to 2 switch or a 2 to 1 switch) was 15 kilobases (kb) or smaller, the phase switch was considered a gene conversion. If the tract lengths on both sides of a phase switch was >25 kb it was considered a recombination event. Phase switches that occurred between intervals that were 15 kb to 25kb in size ranged were excluded from the analysis. These tract lengths were based on a *Drosophila* study (Comeron et al., 2012) that identified and distinguished GC and CO events. The number of total switches, number of recombination events, and number of gene conversions across all individuals between each marker pair, and for each individual were computed. Five individuals were removed from all the matrices due to abnormally high number of total switches (approximately 10 times the average), indicating poor genotype call quality.

Based on the results for each marker interval and the physical location of each marker, the chromosome files were processed to compute the number of recombination and gene conversion events within 100 kb windows.

Calculating Genetic Length and Recombination Rate: Aim 3

The genetic length for each chromosome (centimorgans) was calculated by taking the total number of recombination events for a single chromosome, dividing by N (n=187) and multiplied by 100. Recombination rate was calculated by taking the genetic length (centimorgan) and dividing by the physical length (megabase pairs) of the chromosome using the Amel_4.5 assembly (http://www.ncbi.nlm.nih.gov/assembly/GCF_000002195.4#/st).

Correlation Analysis with DNA Features: Aim 3

The number of recombination events was additionally quantified for each 1000 bp window to statistically correlate this number to DNA features, using SPSS (v.16). Eleven features were analyzed with the number of recombination events for each chromosome (Table 1). Pearson correlation coefficients of recombination frequency to various a-priori features (Table 2) were computed for each chromosome. The selected features came from studies that investigated potential DNA motifs that are associated with recombination (Badis et al., 2008; Bessoltane et al., 2012; Beye et al., 2006; Brandstrom et al., 2008; Dou et al., 1994; Lyko et al., 2010; Myers et al., 2005; Stevison & Noor, 2010). An existing file of DNA features at this resolution was used. The file used for the DNA

features was created by Caitlin Ross and used by her, and fellow past UNCG student Dominick Defelice, and Olav Rueppell for another project.

CHAPTER III

RESULTS

DNA Extractions and Quality Control

Thoraces of 800 drones were initially used for the DNA extractions. Of those individuals, the best samples based on DNA purity, quality, and quantity were used for the study. Overall quantities of the DNA extracted ranged from 2.1 ng/µL to 2033.3 ng/µL when the DNA was rehydrated in 100 µL of TE buffer. DNA samples less than 50 ng/µL were discarded. Of the remaining individuals, 100 were not usable in the study due to DNA degradation and were discarded. Gel electrophoresis showed significant smearing and no distinct bands were present (Figure 2). Of those samples tested, the 192 best samples were used for sequencing. These samples showed a distinct, high-molecular weight DNA band after gel electrophoresis (Figure 3) and varied in quantity ranging from 53.6 ng/µL to 291.2 ng/µL and 260/280 ratios ranging from 1.76 to 2.19.

Polymerase Chain Reaction and Genotyping

Nine sets of primers were used to for PCR and genotyping; AP174, SP167, K05128, K0957, K0905, K0958, AT064, and K0907 (Solignac et al., 2007). Amplification of these primers was tested on a 0.5% agarose gel using gel

electrophoresis. AT064 was the only locus that did not amplify during PCR and therefore was not used in genotyping.

Genotyping on a DNAnalyzer (Licor) was done to verify all drone samples came from the same queen. The seven loci were tested for their level of polymorphism with eight DNA samples that were not used in the project. Of the seven loci, three polymorphic ones (AP174, K05128, and K0907) were chosen for genotyping all 192 samples to be included in the project. The three loci were also chosen because they could be multiplexed on a single gel due to different size. AP174 was approximately 204 bp, K05128 was approximately 290 bp, and K0907 was between 160 bp and 165 bp in length.

After genotyping all 192 samples using the three loci, the results were scored (Figure 4). All lanes showed appropriate banding for the loci selected. Banding patterns of the chosen loci appeared at the appropriate locations of 290, 160-165, and 204 base pairs. For a few individuals, one of the three loci did not amplify but we had never two or more loci missing. Based on the genotyping, all 192 samples were accepted for use in sequencing because their genotypes confirmed they were all derived from the same mother.

Library Construction

Library construction was done at the High Throughput Sequencing Facility at UNC-Chapel Hill, according to the above described protocol. Samples of the raw genomic DNA went through sonication and the library prep, resulting in 192 separate

libraries. Of those 192 samples, 96 samples were pooled together in one tube at a concentration of 15 nM. The other 96 libraries were also pooled together in another tube with a concentration of 15 nM. When tested on the Qubit fluorometer, the concentrations of the pools were as followed: 13.7 ng/μL for HCOMB_1 pool and 17.0 ng/μL for HCOMB_2 pool. Sequencing of the two pools was completed using a HiSeq 2000 (Illumina Inc., San Diego, CA) sequencer.

Initial sequencing only obtained one 100 bp single ended run, but in order to increase sequencing coverage, two additional runs of 100 bp paired-ends were sequenced. Sequencing coverage for 192 individuals ranged from 0.457x to 6.08x.

Single Nucleotide Polymorphisms/Genetic Marker Identification

The number of markers for all 16 chromosomes before removal of unlinked markers totaled 1,252,481 (Figure 5). The final count of markers that were properly linked and used for the calculation of recombination was 939,342, after eliminating approximately 300,000 markers. This is nearly a 450 fold difference in comparison to the 2,008 markers in the microsatellite linkage map (Solignac et al., 2007).

Recombination Events and Gene Conversion Events at 100 kb Windows

Windows at 100 kb intervals were used based on of a similar study in *Drosophila* (Comeron et al., 2012). This relatively large interval allowed for recombination frequency to be clearly visualized. Smaller intervals of 1kb and 10kb were also visually

inspected, but showed a more even distribution without clear evidence for hotspots (Figure 7 B-C).

A large number of recombination events and gene conversion events were observed in all chromosomes. Recombination ranged from 2057 events to 9547 events (Table 3). Gene conversion ranged from 7915 events to 26823 events (Table 3). The ratio of gene conversion events to recombination events ranged from 2:1 to 4:1.

For all chromosomes, recombination events occurred throughout the chromosome and not in (a) specific region(s) of the chromosomes (Figure 6 A-P). At the scale of 100 kb, distinct peaks of recombination rates and regions with very low rates of recombination were observed. This result could be confirmed by plotting the frequency distribution of recombination rate across all chromosomes, which showed a bimodal distribution (Figure 7A). In smaller windows, 1000 bp and 10 kb, the bimodality is less pronounced (Figure 7B-C). The distribution of recombination events across all 16 chromosomes was biased towards less than 50 recombination events per 100kb window. However, there were windows with 150 to 250 recombination events that had a higher frequency than counts above 50 (Figure 7A). GC events occurred at high frequency within each 100kb window across the entirety of each chromosome (Figure 8 A-P). GC events did not show regions of very high occurrences and very low occurrences like recombination events illustrated.

Genetic Length and Recombination Rate of the Resequenced Genome of the Honey Bee

The genetic lengths across all chromosomes were remarkably long even in comparison to the current size of the honey bee at 40 Morgans (M) (Solignac et al., 2007). The total genetic length for this mapping population was 354.2 M, almost 9x the size of Solignac's estimate (Table 4). Even the shortest genetic length at 1038.8 cM at chromosome 4 was still 3.5x longer than the Solignac's length for that chromosome. Correspondingly, recombination rates were also higher than previous estimates. The genome-wide average rate was 178.7 cM/Mb, where the 2,008 marker linkage map had an average recombination rate of 22 cM/Mb (Solignac et al., 2007). The average recombination rate of chromosomes ranged from 87.8 cM/Mb to 248.2 cM/Mb and was not correlated to chromosome size (r=-0.146, n=16, *P*=0.589).

Recombination Hotspots

At the 100 kb window level, potential hotspots were identified based on the high peaks where specific windows throughout the chromosomes had exceptionally high numbers of recombination events. All 16 chromosomes had regions where many recombination events occurred within the 100 kb windows and regions where little to no recombination occurred (Figure 6 A-P). The number of markers per chromosome and the number of recombination events were compared to ensure that this effect was not an artifact of uneven marker coverage at regions where there was little or no recombination.

Genome-wide, recombination rate and the number of markers in each chromosome was correlated (r=0.067, n=2204, *P*=0.002).

DNA Features and Recombination

Window size for evaluating recombination events was changed to 1000 bp to analyze recombination and DNA features. Each feature was correlated with recombination using a separate Pearson's correlation analysis for each chromosome. Several chromosome-specific correlations were indicated (Table 2), but no consistent results emerged and correlations were weak and mostly non-significant after Bonferroni correction.

CHAPTER IV

DISCUSSION

My estimate of the genome-wide recombination rate for *Apis melleifera* was extremely high, nearly 9 times higher than all previous estimates. A likely explanation for these high recombination rates would be a mis-assembly of the honey bee genome. Although we used the most recent version of the honey bee genome assembly (Amel 4.5) our results may be interpreted as evidence for multiple locations throughout the genome here the assembly is incorrect. Alternatively, large blocks of SNP markers may have been misaligned through BWA. However, this seems less likely.

Genome mis-assembly can leave regions of sequences rearranged or discarded and are caused by either repeat collapse or sequence rearrangements, both involving sequence repeats (Phillippy, Schatz & Pop, 2008). Repeat collapse involves the incorrect estimation of repeat copies while sequence rearrangements occurs when repeat sequence copies are shuffled, disrupting any distinctive sequences. Despite the latest reports of the honey bee having low quantities of repetitive sequences (Elsik et al., 2014), it is possible mis-assembly is occurring near transposable elements like *Mariner* elements, which have been discovered in the honey bee genome (Elsik et al., 2014).

Indications that the mis-assembly of the genome was a possible explanation of the high honey bee recombination in this study arose when evaluating the number of

recombination rates for all 16 chromosomes. Observed recombination maxima were somewhat variable but very distinct from the background. We would expect to observe heterogeneity in honey bee recombination rate according to past studies in honey bees (Beye et al., 2006) and other fine-scale studies in humans (Lien et al., 2000) and *Drosophila (Cirulli, Kliman & Noor, 2007)*. However, most peaks consisted of about 175 recombination events per window across all chromosomes which was approximately the number of individuals in this study, accounting for missing data. Within the 100 kb windows with a high recombination rate, recombination events were occurring at distinct regions several hundred markers apart. The number of recombination events in these peaks was similar to 50% recombination, which would be expected by chance for two unlinked markers. These regions may be sections of the genome that belong elsewhere on other chromosomes or tandem repeats of sequences (Phillippy et al., 2008). Proper assembly or removal of the regions with low recombination would most likely disperse the regions with high recombination in a more heterogeneous distribution. In each chromosome it is also possible we observed double crossovers that are close together due to regions or fragments of the assembly that should belong in another chromosome.

If the genome of the honey bee is mis-assembled our results indicate that more improvements for the assembly are necessary. This may be especially needed for bridging any discrepancies between shotgun sequencing and next-generation sequencing. However, we cannot exclude the possibility that the current genome assembly is correct. In that case, the following interpretation would explain our current results.

Potential Underestimation of Recombination Rate in Honey Bees

Resequencing the honey bee genome and analysis at a fine-scale level revealed the current estimation of the honey bee recombination rate may be severely underestimated. Next generation sequencing (sequencing by synthesis) and the use of the most recent version of the reference genome (Amel_4.5, http://www.ncbi.nlm.nih.gov/assembly/GCF_000002195.4#/st) assisted with better genome-wide sequence coverage, thus a better representation of the honey bee genome. We are confident in our estimations of the newly calculated recombination rates because of the high number of markers used in the study and the steps taken to avoid overestimation for mapping. Elimination of poorly linked markers and individuals with abnormally high recombination and gene conversion events contributed to avoiding overestimation of recombination. Removing markers that were due to sequencing errors also played a major role in properly calling phase shifts and recombination rate. Single marker phase shifts between two phases were considered sequencing errors because long tract lengths between shifts should be indication for recombination or gene conversion events. Our conservative threshold of phase shift tract lengths for determining recombination events (greater than 25kb) or gene conversions (less than 15 kb) was also important to avoid miscalling the events and not having an over representation of recombination events.

The current average estimate of maximally 22 cM/Mb from the microsatellite map only used 2,008 markers while this study had over 450 times the amount of markers

which could explain the higher number of identified recombination events (Solignac et al., 2007). This large set of markers increased the coverage of each chromosome. The number of markers per 100 kb window (marker density) was also improved in comparison to the microsatellite linkage map. 2,008 markers would on average be less than one marker per 100 kb, but with over 900,000 markers, hundreds of markers per 100 kb were observed in all chromosomes. Recombination events may have been missed with only 2,008 markers for the entire genome. Our criterion of >25 kb phase tracks would theoretically allow three recombination events in a stretch of DNA that was covered in previous maps only by a single marker. Many of our "hotspot" intervals show more than 187 recombination events, which indicates that there were more than one recombination event in that interval per individual on average. Fine-scale studies on chromosome 15 and 16 in honey bees used 322 and 242 markers respectively in their analysis (Mougel et al., 2014). Looking at different sized windows ranging from 35 kb to 500 kb, their study calculated "hot" recombination rates to be 3-6 times higher than the current average ranging between 60 cM/Mb and 130 cM/Mb and within a 35 kb window recombination rate of 100 cM/Mb (Mougel et al., 2014).

In agreement to the previous studies in honey bees (Bessoltane et al., 2012) and *Drosophila* (Comeron et al., 2012; Gay, Myers & McVean, 2007), we identified more gene conversion events than recombination events. Gene conversion has also been seen to occur more often than recombination events in humans (Jeffreys & May, 2004). In part, this may be a result of our generous threshold to call any phase track of <15 kb a gene conversion event. Empirical evidence for the length of gene conversion tracts

supports humans, yeast, and *Drosophila* (Table 5). Gene conversion tracts have been identified for honey bees, but the lengths are not as well established as those from other model organisms (Bessoltane et al., 2012). A previous study looked at crossovers and non-crossovers/gene conversion events in honey bees and found that gene conversion may also play a role in explaining high recombination rates in honey bees (Bessoltane et al., 2012). High gene conversion in meiotic recombination may have a similar role as gene conversions have in non-allelic homologous recombination by preserving hotspots in humans (Fawcett & Innan, 2013). Gene conversion events in this study appeared to be occurring at a more consistent rate across the chromosomes than to recombination. The ratio of gene conversion events to recombination events was similar to the lower range of the crossover to non-crossover ratio reported by Bessoltane et al. (2012) and fell within the range of ratios for plants and metazoans in general. At 1000 bp intervals there was a positive, yet weak correlation between recombination and gene conversion for all 16 chromosomes (Table 6). Analysis at larger intervals still needs to be done using both the recombination events and gene conversion events to see if there is a relationship between the two at a fine-scale level.

Recombination Hotspots

Each chromosome has regions of very high recombination and regions with very low recombination across the entire chromosome. Regions that have more than 187 recombination events per 100 kb window have an average of at least one recombination event in each individual in this interval. There are 22 such windows out of the 2204 total

windows across the genome. Thus at the 100 kb scale these peaks may represent recombination hotspots. At other scales, the pattern is less distinct, as can be seen by the distribution of recombination events at different scales (Figure 7A-C). The distribution of recombination events (Figure 7A) showed that 0 to 49 recombination events per window were most prevalent. The identified "hotspots" represented only 15-16% of all intervals. In humans, between 60% and 70% of crossovers take place in approximately 10% of the genome (Myers et al., 2005) . The peaks and regions of high recombination could be a result of hotspots being "hotter" or the intensity of recombination is higher. This is in contrast to my hypothesis that high recombination in honey bees is caused by frequent recombination hotspots throughout the genome. Although the occurrence of recombination events genome-wide is very high, when looking at 100 kb windows, hotspots only occur if the number of recombination events per window is above 100 (Figure 7A). At this threshold, chromosome 14 (Figure 6N) only has 12 hotspots, while chromosome 1 (Figure 6A) has over double the amount of hotspots and yet they still have very high and comparable recombination rates (Table 4).

In order to accurately investigate recombination hotspots at a fine scale level in comparison to traditional linkage maps number of individuals sampled would have to increase (Arnheim, Calabrese & Nordborg, 2003). Fewer individuals are needed if there is high recombination rate because more recombination events happen per individual. In this study we have increased the marker density from 2,008 markers to over 900,000 markers, but the sample size could be increased since the number of individuals (n=187) was within the range used for the existing linkage map (n=100-200); (Solignac et al.,

2007). It is also necessary to define hotspots within various sized windows, whether hotspots are a single window where high recombination is occurring or if hotspots are defined as clusters of windows adjacent to each other with high rates of recombination.

Recombination Rate and DNA Features

At 1000 bp intervals, recombination was not consistently correlated to the selected DNA features such as sequence motifs, microsatellites, GC content, CpG islands and low complexity sequence. Single correlations were significant in specific chromosomes but no overall pattern for these features was identified. Seven chromosomes showed no correlates of recombination at all. It is difficult to interpret these finding because a mechanism of chromosome-specific effects on recombination is hard to envision. Intervals at this fine scale may be too small to identify high correlation with sequence features. However, in other studies such analyses have yielded significant results (Mougel et al., 2014). Larger intervals of 100 kb size may have produced clearer results because recombination rates vary distinctly at this scale. The selected DNA features have been correlated in other studies with recombination (Bessoltane et al., 2012; Beye et al., 2006; Mougel et al., 2014; Myers et al., 2005). Among these studies, only Mougel et al. (2014) is directly comparable because it deals with *A. mellifera* at a map resolution that is similar to my study. The results, limited to chromosomes 15 and 16, found evidence that DNA methylation, CCAAT, CGCA, and GCCCGC regulate the initiation of recombination localization (Mougel et al., 2014). The features just listed were also features that were correlated with recombination in our study (Table 2).

Chromosomes 15 and 16 both had weak correlations with GCCGC and DNA methylation (CpG islands). Chromosome 16 also had a weak correlation with CGCA. Further investigation of DNA features need to be done at larger interval sizes. Such information would lead to a better comparison to previous studies. Currently, our results suggest that no specific motifs are responsible for hotspots in the honey bee genome: the large scale of hotspots rather suggests that some more global DNA features that vary across the chromosomes in terms of kbp.

We may not be able to see any correlation between recombination and any sequence motifs even at larger intervals. This may be explained by the Red Queen Hypothesis, such that honey bees are expected to have a high hotspot turnover rate, meaning they are quickly created and quickly destroyed by biased gene conversion (Ubeda & Wilkins, 2011). With a high hotspot turnover rate, we then would not be able to identify any features. Honey bees may rely on other mechanisms or features that would allow them to maintain a high recombination rate. High gene conversion rate may be one of those features to help maintain recombination rate.

Conclusions

It has been well established through linkage mapping that the honey bee has the highest recombination rate of all metazoans. However, my study is the first investigation at a genome-wide and fine-scale level. Results suggested that the reference genome has numerous, major assembly errors. Before these errors are corrected, no genome-wide fine-scale analysis of recombination patterns would be meaningful. Alternatively, the

assembly is correct and the results presented here are valid. This would imply that previous recombination maps have omitted important patterns due to low coverage, resulting in a severe underestimation of honey bee recombination rates. Recombination may occur at an almost 9 times higher rate than previously reported. The large set of markers used for the study was made possible because of next generation sequencing and the use of the most recent honey bee genome assembly to create a better representation of the genome. The average genome wide recombination rate from this study was calculated at 178.7 cM/Mb and at 100 kb windows, hotspots could be seen in all 16 chromosomes.

Further analyses need to be done in order to better understand the characteristics of these recombination hotspots and the features and mechanisms that initiate recombination hotspots in the exceptional case of the honey bee genome. Confirmation of proper genome assembly is essential for verifying the presented results. Additional research also needs to be done to better understand gene conversion events in the honey bee genome and their association with high recombination. It may be necessary to reevaluate the tract lengths used in the study with the possibility of gene conversions were counted as recombination events. Since tract lengths tend to vary from organism to organism, tract lengths to use with honey bee studies need to be defined to avoid possible overlap between gene conversion and recombination events. Additional evaluations of hotspots are also necessary to characterize and confirm their presence in honey bees. By better understanding these high recombination rates in honey bees at a fine scale level, especially if the current recombination rate is underestimated, we can better understand the overall mechanisms of recombination.

REFERENCES

ARNHEIM, N., CALABRESE, P. & NORDBORG, M. (2003). Hot and cold spots of recombination in the human genome: the reason we should find them and how this can be achieved. *Am J Hum Genet* 73, 5-16.

BADIS, G., CHAN, E. T., VAN BAKEL, H., PENA-CASTILLO, L., TILLO, D., TSUI, K., CARLSON, C. D., GOSSETT, A. J., HASINOFF, M. J., WARREN, C. L., GEBBIA, M., TALUKDER, S., YANG, A., MNAIMNEH, S., TERTEROV, D., COBURN, D., LI YEO, A., YEO, Z. X., CLARKE, N. D., LIEB, J. D., ANSARI, A. Z., NISLOW, C. & HUGHES, T. R. (2008). A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. *Mol Cell* 32, 878-87.

BAUDAT, F., BUARD, J., GREY, C., FLEDEL-ALON, A., OBER, C., PRZEWORSKI, M., COOP, G. & DE MASSY, B. (2010). PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* 327, 836-40.

BESSOLTANE, N., TOFFANO-NIOCHE, C., SOLIGNAC, M. & MOUGEL, F. (2012). Fine scale analysis of crossover and non-crossover and detection of recombination sequence motifs in the honeybee (Apis mellifera). *PLoS One* 7, e36229.

BEYE, M., GATTERMEIER, I., HASSELMANN, M., GEMPE, T., SCHIOETT, M., BAINES, J. F., SCHLIPALIUS, D., MOUGEL, F., EMORE, C., RUEPPELL, O., SIRVIO, A., GUZMAN-NOVOA, E., HUNT, G., SOLIGNAC, M. & PAGE, R. E., JR. (2006). Exceptionally high levels of recombination across the honey bee genome. *Genome Res* 16, 1339-44.

BEYE, M., HUNT, G. J., PAGE, R. E., FONDRK, M. K., GROHMANN, L. & MORITZ, R. F. (1999). Unusually high recombination rate detected in the sex locus region of the honey bee (Apis mellifera). ***Genetics*** **153, 1701-8.**

BONCRISTIANI, H., LI, J. L., EVANS, J. D., PETTIS, J. & CHEN, Y. P. (2011). Scientific note on PCR inhibitors in the compound eyes of honey bees, Apis mellifera. ***Apidologie*** **42, 457-460.**

BRANDSTROM, M., BAGSHAW, A. T., GEMMELL, N. J. & ELLEGREN, H. (2008). The relationship between microsatellite polymorphism and recombination hot spots in the human genome. ***Mol Biol Evol*** **25, 2579-87.**

BROOKES, A. J. (1999). The essence of SNPs. ***Gene*** **234, 177-86.**

CHAN, A. H., JENKINS, P. A. & SONG, Y. S. (2012). Genome-wide fine-scale recombination rate variation in Drosophila melanogaster. ***PLoS Genet*** **8, e1003090.**

CIRULLI, E. T., KLIMAN, R. M. & NOOR, M. A. (2007). Fine-scale crossover rate heterogeneity in Drosophila pseudoobscura. ***J Mol Evol*** **64, 129-35.**

COMERON, J. M., RATNAPPAN, R. & BAILIN, S. (2012). The Many Landscapes of Recombination in Drosophila melanogaster. ***Plos Genetics*** **8.**

COOP, G. & PRZEWORSKI, M. (2007). An evolutionary view of human recombination. ***Nat Rev Genet*** **8, 23-34.**

DIXON, L. R., MCQUAGE, M. R., LONON, E. J., BUEHLER, D., SECK, O. & RUEPPELL, O. (2012). Pleiotropy of segregating genetic variants

that affect honey bee worker life expectancy. *Exp Gerontol* 47, 631-7.

DOU, S., ZENG, X., CORTES, P., ERDJUMENT-BROMAGE, H., TEMPST, P., HONJO, T. & VALES, L. D. (1994). The recombination signal sequence-binding protein RBP-2N functions as a transcriptional repressor. *Mol Cell Biol* 14, 3310-9.

ELSIK, C. G., WORLEY, K. C., BENNETT, A. K., BEYE, M., CAMARA, F., CHILDERS, C. P., DE GRAAF, D. C., DEBYSER, G., DENG, J., DEVREESE, B., ELHAIK, E., EVANS, J. D., FOSTER, L. J., GRAUR, D., GUIGO, R., TEAMS, H. P., HOFF, K. J., HOLDER, M. E., HUDSON, M. E., HUNT, G. J., JIANG, H., JOSHI, V., KHETANI, R. S., KOSAREV, P., KOVAR, C. L., MA, J., MALESZKA, R., MORITZ, R. F., MUNOZ-TORRES, M. C., MURPHY, T. D., MUZNY, D. M., NEWSHAM, I. F., REESE, J. T., ROBERTSON, H. M., ROBINSON, G. E., RUEPPELL, O., SOLOVYEV, V., STANKE, M., STOLLE, E., TSURUDA, J. M., VAERENBERGH, M. V., WATERHOUSE, R. M., WEAVER, D. B., WHITFIELD, C. W., WU, Y., ZDOBNOV, E. M., ZHANG, L., ZHU, D., GIBBS, R. A. & HONEY BEE GENOME SEQUENCING, C. (2014). Finding the missing honey bee genes: lessons learned from a genome upgrade. *BMC Genomics* 15, 86.

ESTOUP, A., SOLIGNAC, M., AND CORNUET, JM. (1994). Precise Assessment of the Number of Patrilines and of Genetic Relatedness in Honeybee Colonies. *The Royal Society* 258, 8.

FAWCETT, J. A. & INNAN, H. (2013). The role of gene conversion in preserving rearrangement hotspots in the human genome. *Trends Genet* 29, 561-8.

FREE, J. B. & WILLIAMS, I. H. (1975). Factors determining the rearing and rejection of drones by the honey bee colony.

GALLAI, N., SALLES, J.M., SETTELE, J., AND VAISSIERE, B.E. (2008). Economic evaluation of the vulnerability of world agriculture confronted with pollinator decline. *Ecological Economics* 68, 810-821.

GAY, J., MYERS, S. & MCVEAN, G. (2007). Estimating meiotic gene conversion rates from population genetic data. *Genetics* 177, 881-94.

HAMILTON, W. D. (1964). The genetical evolution of social behaviour. II. *J Theor Biol* 7, 17-52.

HAYWORTH, M. K., JOHNSON, N. G., WILHELM, M. E., GOVE, R. P., METHENY, J. D. & RUEPPELL, O. (2009). Added Weights Lead to Reduced Flight Behavior and Mating Success in Polyandrous Honey Bee Queens (Apis mellifera). *Ethology* 115, 698-706.

HERMANN, H. R. (1979). *Social insects*. Academic Press, New York.

HONEYBEE GENOME SEQUENCING, C. (2006). Insights into social insects from the genome of the honeybee Apis mellifera. *Nature* 443, 931-49.

HUGHES, W. O. & BOOMSMA, J. J. (2004). Genetic diversity and disease resistance in leaf-cutting ant societies. *Evolution* 58, 1251-60.

HUGHES, W. O., OLDROYD, B. P., BEEKMAN, M. & RATNIEKS, F. L. (2008). Ancestral monogamy shows kin selection is key to the evolution of eusociality. *Science* 320, 1213-6.

HUNT, G. J., AMDAM, G. V., SCHLIPALIUS, D., EMORE, C., SARDESAI, N., WILLIAMS, C. E., RUEPPELL, O., GUZMAN-NOVOA, E., ARECHAVALETA-VELASCO, M., CHANDRA, S., FONDRK, M. K., BEYE, M. & PAGE, R. E., JR.

(2007). Behavioral genomics of honeybee foraging and nest defense. *Naturwissenschaften* 94, 247-67.

HUNT, G. J. A. P. J., R.E. (1995). Linkage map of the honey bee, *Apis mellifera,* based on RAPD markers. *The Genetics Society of America* 139, 1371-1382.

JEFFREYS, A. J. & MAY, C. A. (2004). Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nat Genet* 36, 151-6.

KENT, C. F., MINAEI, S., HARPUR, B. A. & ZAYED, A. (2012). Recombination is associated with the evolution of genome structure and worker behavior in honey bees. *Proc Natl Acad Sci U S A* 109, 18012-7.

KENT, C. F. & ZAYED, A. (2013). Evolution of recombination and genome structure in eusocial insects. *Commun Integr Biol* 6, e22919.

KERR, W. E. (1962). Genetics of sex determination. *Annu Rev Entomol* 7, 157-76.

KIRCHER, M., SAWYER, S. & MEYER, M. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res* 40, e3.

KLEIN, A. M., VAISSIERE, B. E., CANE, J. H., STEFFAN-DEWENTER, I., CUNNINGHAM, S. A., KREMEN, C. & TSCHARNTKE, T. (2007). Importance of pollinators in changing landscapes for world crops. *Proc Biol Sci* 274, 303-13.

LI, H. & DURBIN, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754-60.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G., DURBIN, R. & GENOME PROJECT DATA PROCESSING, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078-9.

LIEN, S., SZYDA, J., SCHECHINGER, B., RAPPOLD, G. & ARNHEIM, N. (2000). Evidence for heterogeneity in recombination in the human pseudoautosomal region: high resolution analysis by sperm typing and radiation-hybrid mapping. *Am J Hum Genet* 66, 557-66.

LOSEY, J. E. & VAUGHAN, M. (2006). The economic value of ecological services provided by insects. *Bioscience* 56, 311-323.

LUNDIN, S., STRANNEHEIM, H., PETTERSSON, E., KLEVEBRING, D. & LUNDEBERG, J. (2010). Increased throughput by parallelization of library preparation for massive sequencing. *PLoS One* 5, e10029.

LYKO, F., FORET, S., KUCHARSKI, R., WOLF, S., FALCKENHAYN, C. & MALESZKA, R. (2010). The honey bee epigenomes: differential methylation of brain DNA in queens and workers. *PLoS Biol* 8, e1000506.

MEYER, M., STENZEL, U., MYLES, S., PRUFER, K. & HOFREITER, M. (2007). Targeted high-throughput sequencing of tagged nucleic acid samples. *Nucleic Acids Res* 35, e97.

MOUGEL, F., POURSAT, M. A., BEAUME, N., VAUTRIN, D. & SOLIGNAC, M. (2014). High-resolution linkage map for two honeybee

chromosomes: the hotspot quest. *Mol Genet Genomics* 289, 11-24.

MYERS, S., BOTTOLO, L., FREEMAN, C., MCVEAN, G. & DONNELLY, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science* 310, 321-4.

MYERS, S., FREEMAN, C., AUTON, A., DONNELLY, P. & MCVEAN, G. (2008). A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat Genet* 40, 1124-9.

OLDROYD, B. P., CLIFTON, M. J., WONGSIRI, S., RINDERER, T. E., SYLVESTER, H. A. & CROZIER, R. H. (1997). Polyandry in the genus Apis, particularly Apis andreniformis. *Behavioral Ecology and Sociobiology* 40, 17-26.

OSTER, G. F. & WILSON, E. O. (1978). *Caste and ecology in the social insects*. Princeton University Press, Princeton, N.J.

OTTO, S. P. & BARTON, N. H. (2001). Selection for recombination in small populations. *Evolution* 55, 1921-1931.

PAGE, R. E. (1986). Sperm Utilization in Social Insects. *Annual Review of Entomology* 31, 297-320.

PAN, J., SASAKI, M., KNIEWEL, R., MURAKAMI, H., BLITZBLAU, H. G., TISCHFIELD, S. E., ZHU, X., NEALE, M. J., JASIN, M., SOCCI, N. D., HOCHWAGEN, A. & KEENEY, S. (2011). A hierarchical combination of factors shapes the genome-wide topography of yeast meiotic recombination initiation. *Cell* 144, 719-31.

PETERSON, B. K., WEBER, J. N., KAY, E. H., FISHER, H. S. & HOEKSTRA, H. E. (2012). Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One* 7, e37135.

PHILLIPPY, A. M., SCHATZ, M. C. & POP, M. (2008). Genome assembly forensics: finding the elusive mis-assembly. *Genome Biol* 9, R55.

POTTS, S. G., BIESMEIJER, J. C., KREMEN, C., NEUMANN, P., SCHWEIGER, O. & KUNIN, W. E. (2010). Global pollinator declines: trends, impacts and drivers. *Trends Ecol Evol* 25, 345-53.

RICE, W. R. (1983). Parent-Offspring Pathogen Transmission - a Selective Agent Promoting Sexual Reproduction. *American Naturalist* 121, 187-203.

RUEPPELL, O., JOHNSON, N. & RYCHTAR, J. (2008). Variance-based selection may explain general mating patterns in social insects. *Biol Lett* 4, 270-3.

RUEPPELL, O., MEIER, S. & DEUTSCH, R. (2012). Multiple mating but not recombination causes quantitative increase in offspring genetic diversity for varying genetic architectures. *PLoS One* 7, e47220.

SCHUELKE, M. (2000). An economic method for the fluorescent labeling of PCR fragments. *Nat Biotechnol* 18, 233-4.

SIRVIO, A., JOHNSTON, J. S., WENSELEERS, T. & PAMILO, P. (2011). A high recombination rate in eusocial Hymenoptera: evidence from the common wasp Vespula vulgaris. *BMC Genet* 12, 95.

SNODGRASS, R. E. (1984). *Anatomy of the honey bee*. Comstock Pub. Associates, Ithaca.

SOLIGNAC, M., MOUGEL, F., VAUTRIN, D., MONNEROT, M. & CORNUET, J. M. (2007). A third-generation microsatellite-based linkage map of the honey bee, Apis mellifera, and its comparison with the sequence-based physical map. *Genome Biol* 8, R66.

SOLIGNAC, M., VAUTRIN, D., BAUDRY, E., MOUGEL, F., LOISEAU, A. & CORNUET, J. M. (2004). A microsatellite-based linkage map of the honeybee, Apis mellifera L. *Genetics* 167, 253-62.

STEINER, W. W., DAVIDOW, P. A. & BAGSHAW, A. T. (2011). Important characteristics of sequence-specific recombination hotspots in Schizosaccharomyces pombe. *Genetics* 187, 385-96.

STEVISON, L. S. & NOOR, M. A. (2010). Genetic and evolutionary correlates of fine-scale recombination rate variation in Drosophila persimilis. *J Mol Evol* 71, 332-45.

TARPY, D. R. & PAGE, R. E. (2002). Sex determination and the evolution of polyandry in honey bees (Apis mellifera). *Behavioral Ecology and Sociobiology* 52, 143-150.

TARPY, D. R. & SEELEY, T. D. (2006). Lower disease infections in honeybee (Apis mellifera) colonies headed by polyandrous vs monandrous queens. *Naturwissenschaften* 93, 195-9.

TRIVERS, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology* 46, 22.

TRIVERS, R. L. & HARE, H. (1976). Haploidploidy and the evolution of the social insect. *Science* 191, 249-63.

UBEDA, F. & WILKINS, J. F. (2011). The Red Queen theory of recombination hotspots. *J Evol Biol* 24, 541-53.

WHITING, P. W. (1945). The evolution of male haploidy. *Q Rev Biol* 20, 231-60.

WILFERT, L., GADAU, J. & SCHMID-HEMPEL, P. (2007). Variation in genomic recombination rates among animal taxa and the case of social insects. *Heredity (Edinb)* 98, 189-97.

WILSON, E. O. (1971). *The insect societies*. Belknap Press of Harvard University Press, Cambridge, Mass.,.

WILSON, E. O. (1998). Kin selection as the key to altruism: its rise and fall. *Social Research* 72.

WINSTON, M. L. (1987). *The biology of the honey bee*. Harvard University Press, Cambridge, Mass.

ZHENG, J., KHIL, P. P., CAMERINI-OTERO, R. D. & PRZYTYCKA, T. M. (2010). Detecting sequence polymorphisms associated with meiotic recombination hotspots in the human genome. *Genome Biol* 11, R103.

ZHOU, T., HU, Z., ZHOU, Z., GUO, X. & SHA, J. (2013). Genome-wide analysis of human hotspot intersected genes highlights the roles of meiotic recombination in evolution and disease. *BMC Genomics* 14, 67.

APPENDIX A

TABLES

Table 1. Sequence Features Associated with Recombination. These features were analyzed to see if they correlated with recombination. (Badis et al., 2008; Bessoltane et al., 2012; Beye et al., 2006; Brandstrom et al., 2008; Dou et al., 1994; Lyko et al., 2010; Myers et al., 2005; Myers et al., 2008; Steiner, Davidow & Bagshaw, 2011; Stevison & Noor, 2010).

Sequence Feature	Source
GC Content	Beye et al., 2006
CGCA	Bessoltane et al., 2012
GCCGC	Bessoltane et al., 2012
CCGCA	Bessoltane et al., 2012
CCTCCT	Stevison and Noor, 2010 and Myers et al., 2005
CCCCGCA	Badis et al., 2008
CCAATCA	Steiner et al., 2011
TGGGAAA	Dou et al., 1994
CpG Islands	Lyko et al., 2010
Low Complexity Sequences	Myers et al., 2008
All Microsatellites	Brandstrom et al., 2008

Table 2. Pearson Correlation between Recombination and DNA Features. Only features that had a p-value of <0.05 were included in the table. At 1000 bp intervals, there is no correlation between recombination and the listed DNA features.

Chrom.	Variable	Correlation	p-value	Adjusted p-value	n
2	TGGGAAA	0.029	<0.001	0.016	15495
2	All Microsatellites	0.017	0.032	0.512	15495
3	CGCA	0.026	0.003	0.048	13229
4	GCCGC	0.041	<0.001	0.016	12718
6	CGCA	0.018	0.013	0.208	18473
6	CCGCA	0.026	<0.001	0.016	18473
7	CCTCCT	0.021	0.015	0.24	13217
7	TGGGAAA	0.028	<0.001	0.016	13217
7	Low Complexity	-0.024	0.006	0.096	13217
8	GC Content	0.018	0.038	0.608	13546
8	TGGGAAA	0.032	<0.001	0.016	13546
8	CpG	0.017	0.048	0.768	13546
12	CCAATCA	0.019	0.035	0.56	11902
13	CCCCGCA	0.043	<0.001	0.016	10289
15	GC Content	0.027	0.007	0.112	10115
15	GCCGC	0.029	0.004	0.064	10115
15	CCAATCA	0.027	0.008	0.128	10115
15	CpG	0.028	0.005	0.08	10115
16	GC Content	0.029	0.013	0.208	7207
16	CGCA	0.030	0.011	0.176	7207
16	GCCGC	0.038	0.001	0.016	7207
16	CpG	0.025	0.034	0.544	7207

Table 3. Number of Recombination Events and Gene Conversion Events for each Chromosome.

Chromosome	Recombination Events	GC events
1	9547	26823
2	2998	9694
3	5586	12331
4	2057	9608
5	4163	11288
6	4837	18325
7	4855	14025
8	4343	13135
9	3060	10827
10	3657	12242
11	4959	11678
12	2913	7915
13	2866	10818
14	2784	7264
15	3986	7728
16	3055	9376

Table 4. Genetic Lengths, Physical Lengths, Recombination Rate, and Number of Recombination Events across All 16 Chromosomes.

Chromosome	Genetic length (cM)	Physical Length (Mb)	Recombination Rate (cM/Mb)	Recombination Events
1	5105.3	27.3	186.8	9547
2	1603.2	14.4	111.3	2998
3	2987.1	12.0	248.2	5586
4	1038.8	11.8	87.8	2057
5	2226.2	13.1	168.7	4163
6	2586.6	16.2	159.6	4837
7	2596.2	11.8	219.7	4855
8	2510.1	12.3	202.7	4343
9	1636.3	10.4	157.2	3060
10	1955.6	11.3	171.9	3657
11	2651.8	13.5	195.3	4959
12	1557.7	10.9	142.5	2913
13	1532.6	9.3	164.4	2866
14	1488.7	9.3	159.5	2784
15	2131.5	8.9	238.9	3986
16	1633.6	6.6	244.2	3055
Total	35424.1	199.7	-	65666
Average	-	-	178.7	-

Table 5. Gene Conversion Tract Lengths for Model Organisms.

Organism	GC Tract Length Ranges	Source
Honey bee	~1 to 1000bp	Bessoltane et al., 2012
Human	300-1091 bp	Jeffreys and May, 2004
Drosophila	up to 15kb	Comeron et. al., 2012
Yeast	1109-7575 bp	Qi et. al., 2009 and Paques and Harper, 1999

Table 6. Correlation between Gene Conversions and Recombination Rates across All 16 Chromosomes.

Chromosome	Correlation	p-Value	Corrected p-Value	n
1	0.117	<0.001	<0.016	29891
2	0.12	<0.001	<0.016	15495
3	0.183	<0.001	<0.016	13229
4	0.112	<0.001	<0.016	12718
5	0.179	<0.001	<0.016	14363
6	0.134	<0.001	<0.016	18473
7	0.155	<0.001	<0.016	13127
8	0.103	<0.001	<0.016	13546
9	0.17	<0.001	<0.016	11112
10	0.145	<0.001	<0.016	12914
11	0.115	<0.001	<0.016	14726
12	0.119	<0.001	<0.016	11902
13	0.117	<0.001	<0.016	10289
14	0.136	<0.001	<0.016	10244
15	0.236	<0.001	<0.016	10115
16	0.12	<0.001	<0.016	7207

APPENDIX B

FIGURES

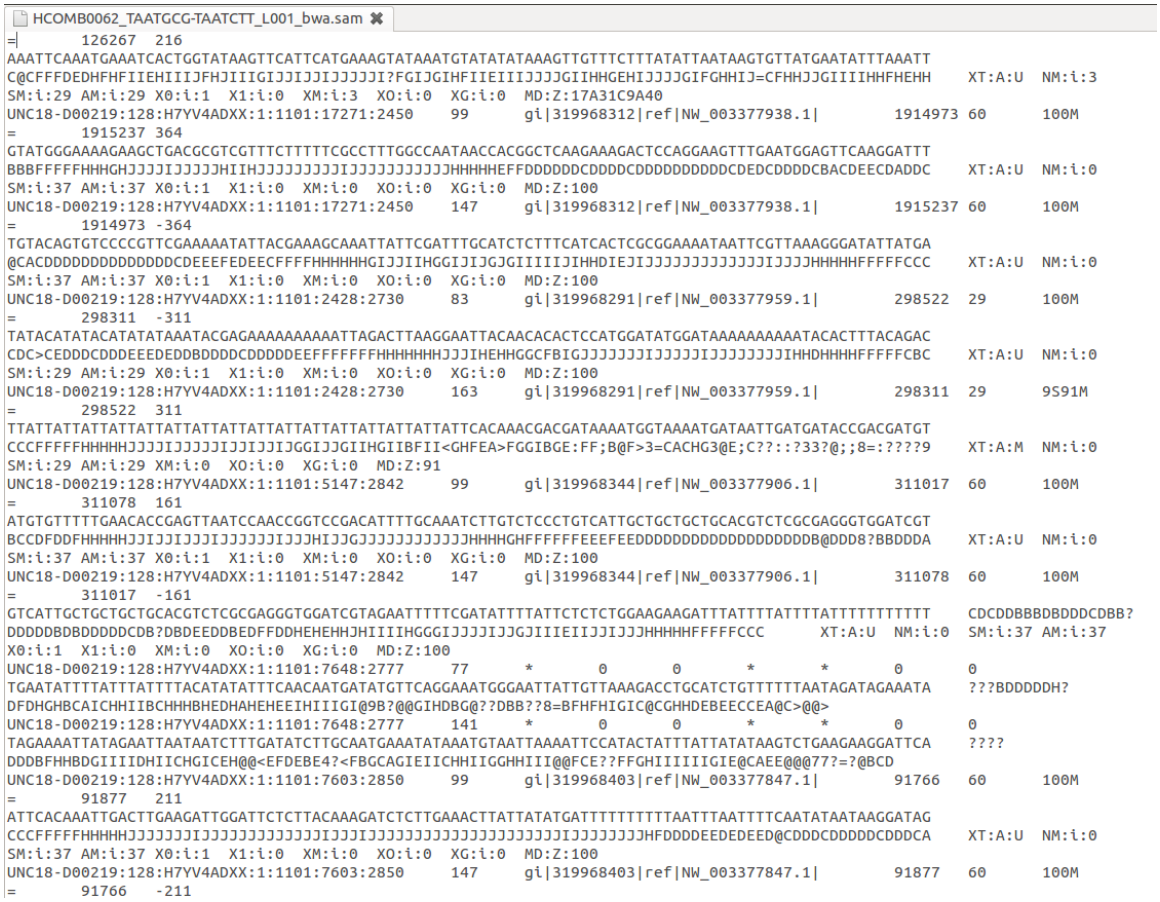

Figure 1. Example of a SAM Format File. In the SAM format, files can be visualized outside the command line.

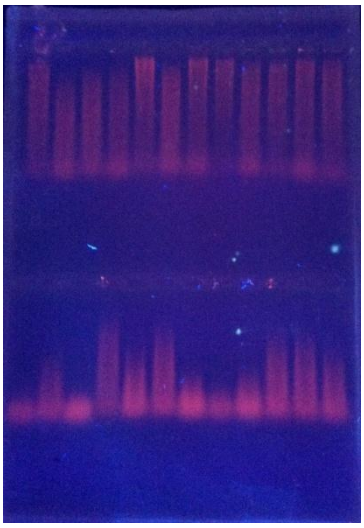

Figure 2. Agarose Gel with Smeared DNA Samples and did not Show Distinct Bands from Drone Honey Bees.

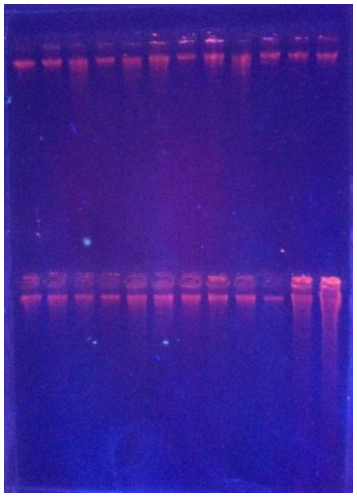

Figure 3. Agarose Gel with Clear, Distinct Genomic DNA Bands from Drone Honey Bees.

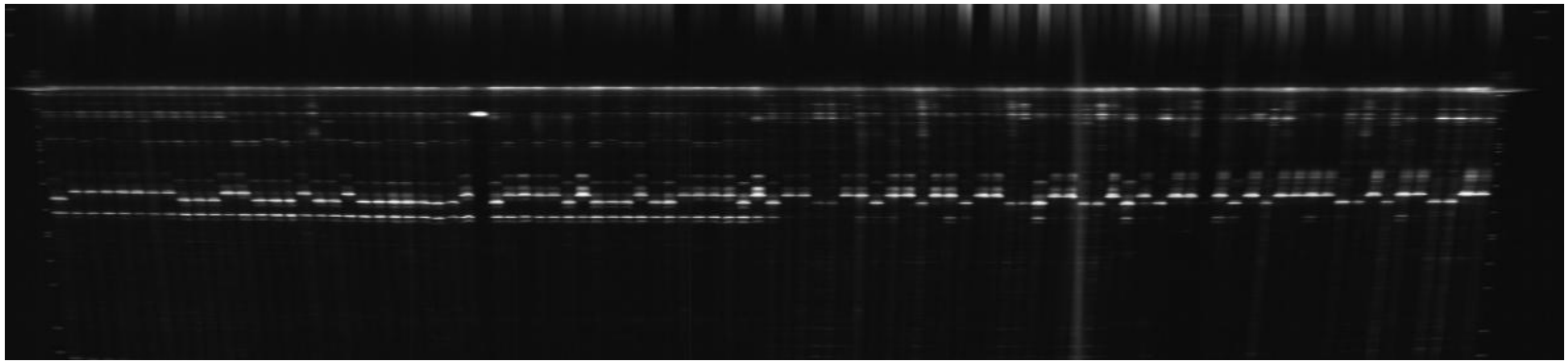

Figure 4. Licor Gel Illustrating Genotyping for Drone Samples. Genotyping was performed to ensure all drones used in the project were true brothers, coming from the same queen. Gels were scored to verify all individuals displayed the chosen amplified loci.

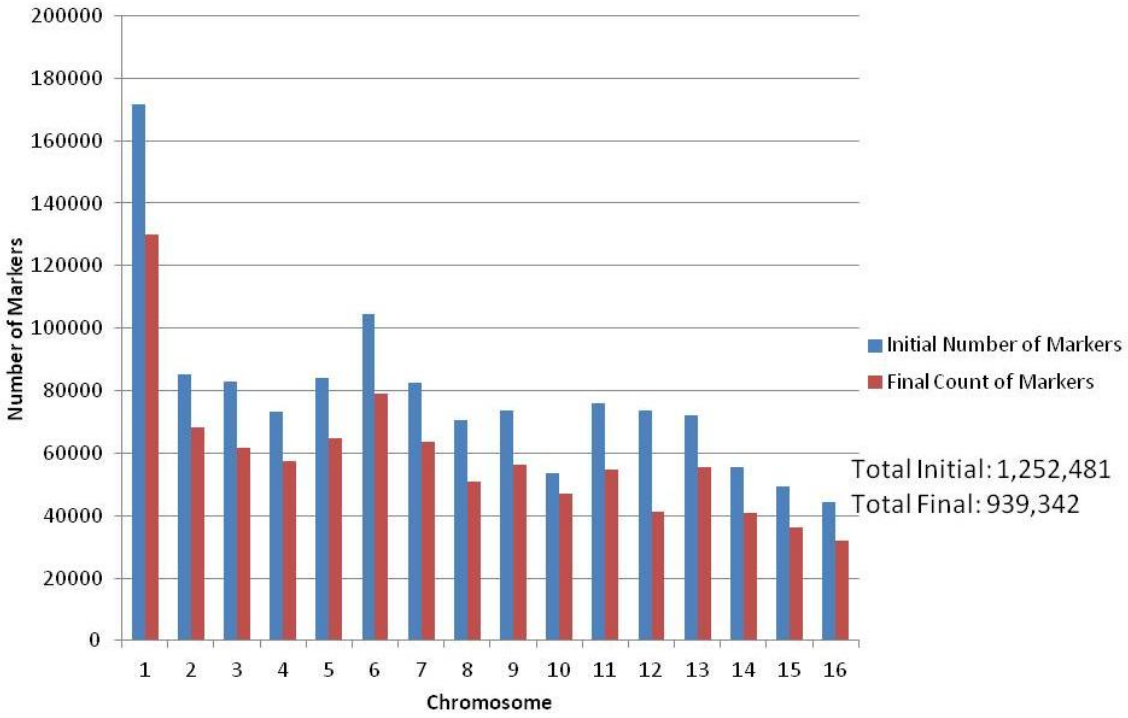

Figure 5. Bar Graph Comparing the Initial Amount of Markers and Final Amount of Markers used for Linkage in All 16 Chromosomes of the Honey Bee. Along the x-axis is the chromosome number and along the y-axis is number of markers. Before filtering out markers, there were approximately 1.2 million markers. Approximately 300,000 markers with poor linkage to adjacent markers were removed.

A

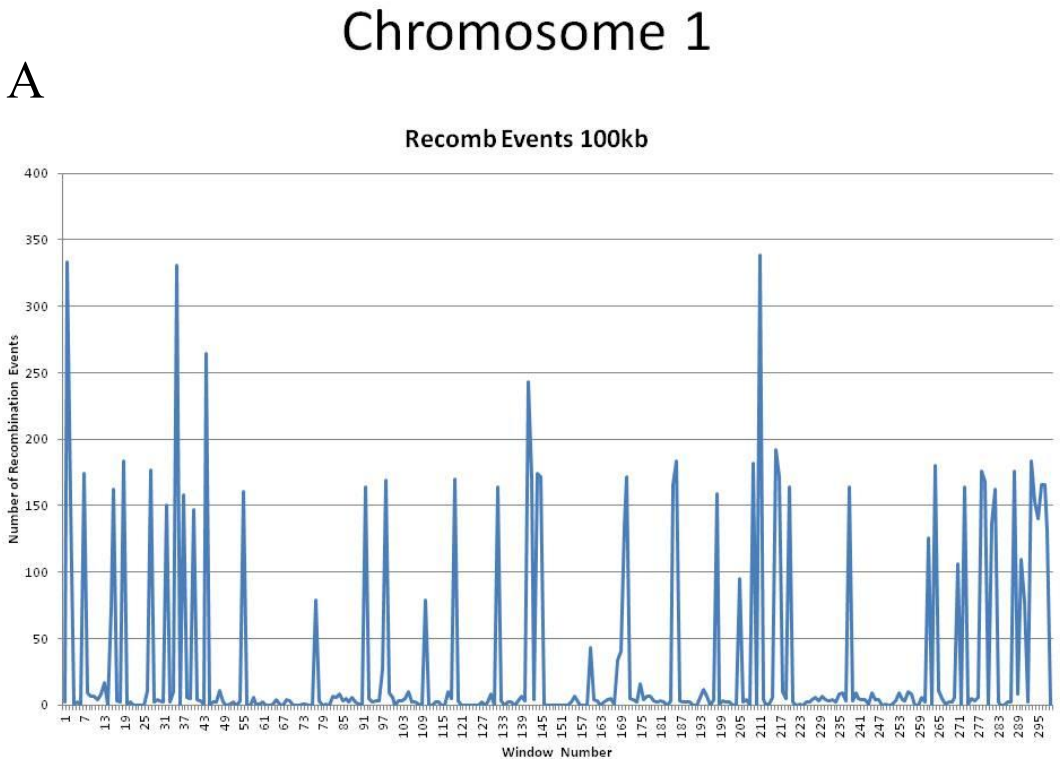

Figure 6A. Recombination Events for All 16 Chromosomes of the Honey Bee in 100 kb Windows. Each graph represents each chromosome and the number of recombination events. The x-axis represents each 100 kb window and the y-axis is the total count of recombination events. Windows with high number counts of recombination events are seen as peaks on the graph and are potential recombination hotspots.

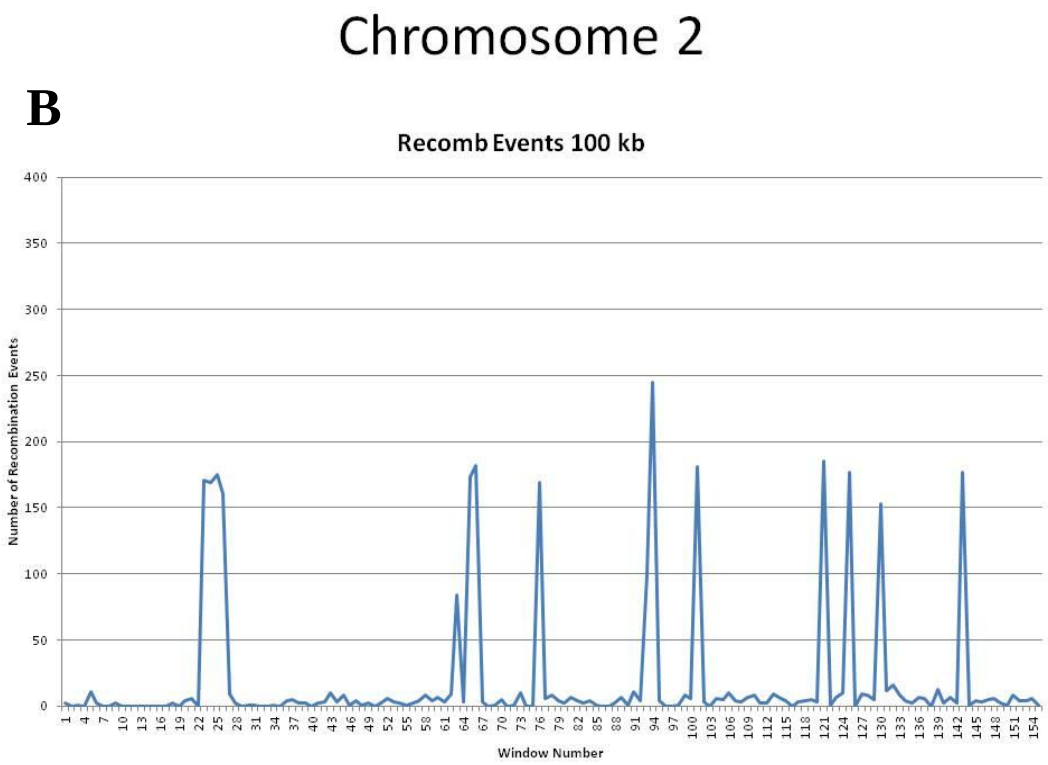

Figure 6B. Recombination Events for All 16 Chromosomes of the Honey Bee in 100 kb Windows. Each graph represents each chromosome and the number of recombination events. The x-axis represents each 100 kb window and the y-axis is the total count of recombination events. Windows with high number counts of recombination events are seen as peaks on the graph and are potential recombination hotspots.

Chromosome 3

C

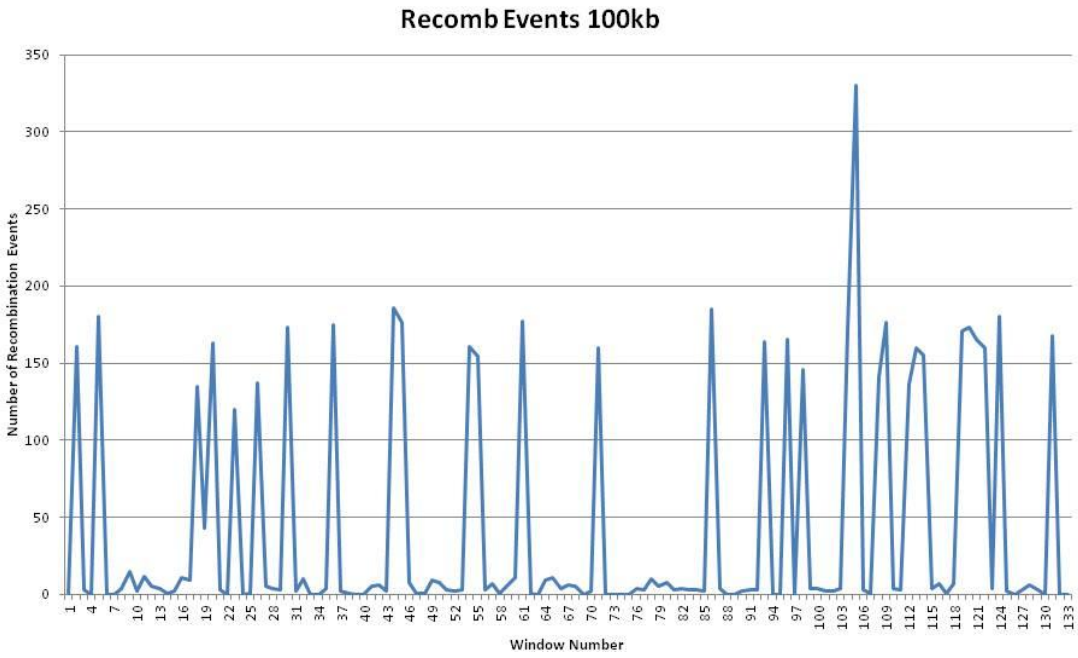

Figure 6C. Recombination Events for All 16 Chromosomes of the Honey Bee in 100 kb Windows. Each graph represents each chromosome and the number of recombination events. The x-axis represents each 100 kb window and the y-axis is the total count of recombination events. Windows with high number counts of recombination events are seen as peaks on the graph and are potential recombination hotspots.

D

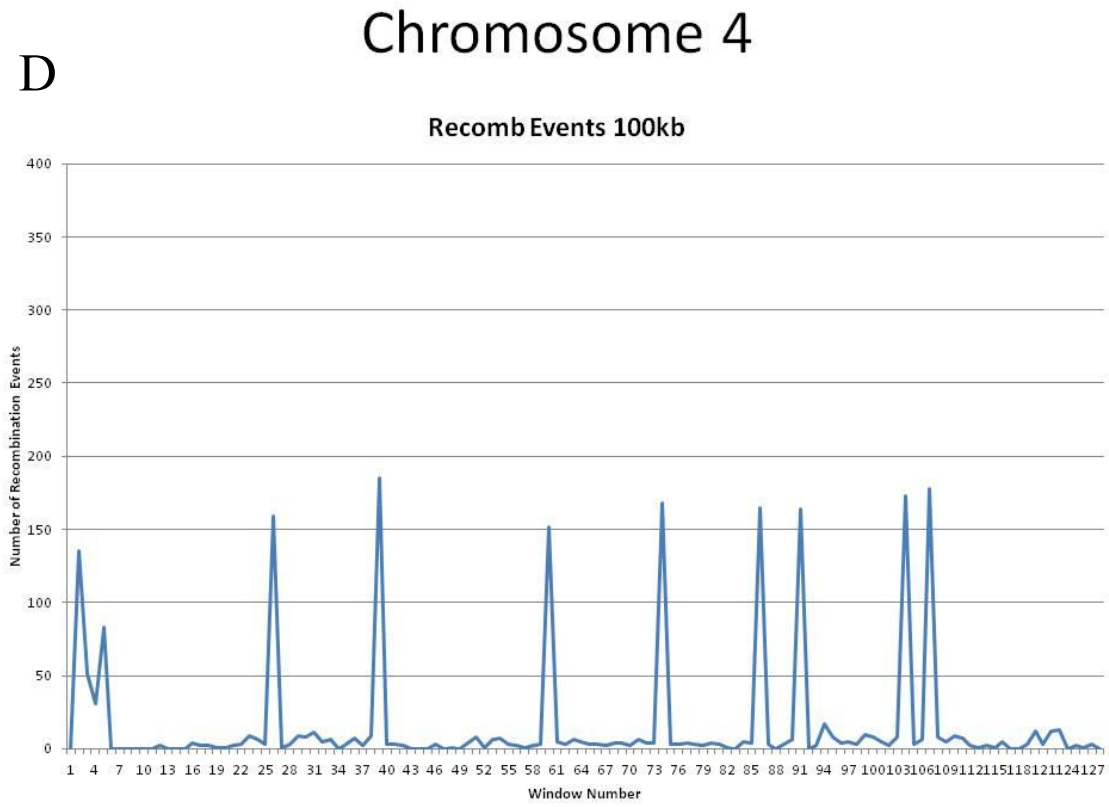

Figure 6D. Recombination Events for All 16 Chromosomes of the Honey Bee in 100 kb Windows. Each graph represents each chromosome and the number of recombination events. The x-axis represents each 100 kb window and the y-axis is the total count of recombination events. Windows with high number counts of recombination events are seen as peaks on the graph and are potential recombination hotspots.

E

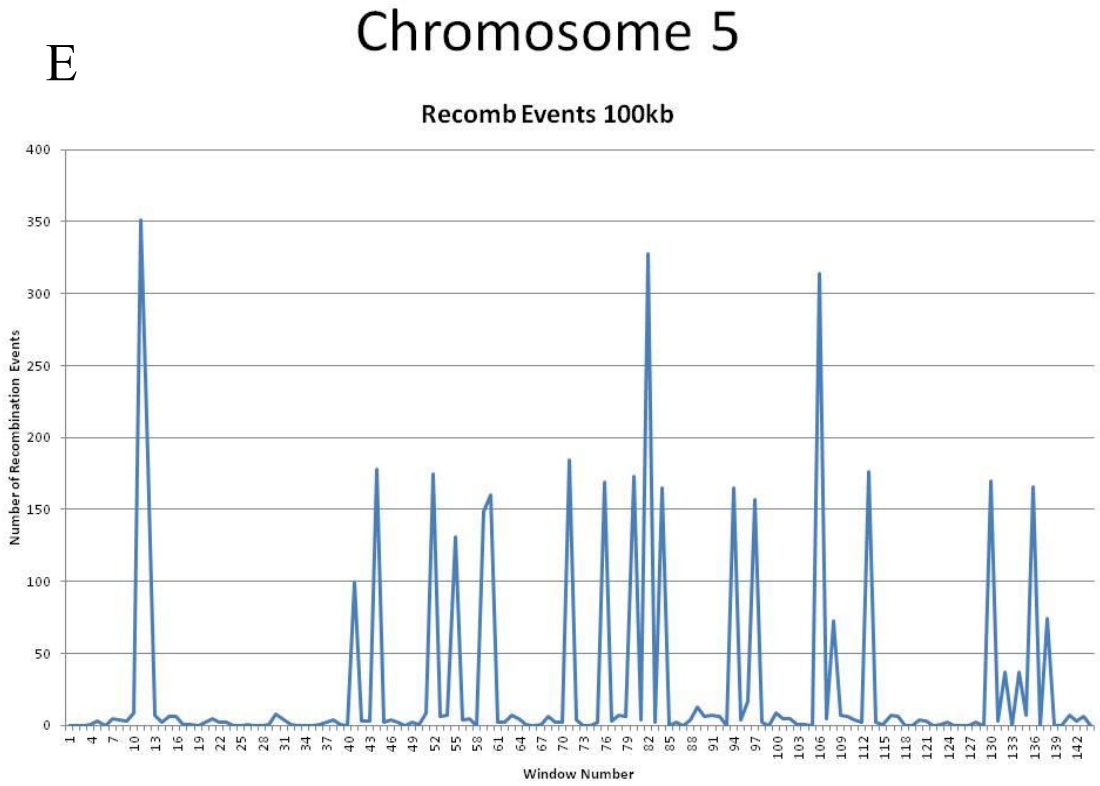

Figure 6E. Recombination Events for All 16 Chromosomes of the Honey Bee in 100 kb Windows. Each graph represents each chromosome and the number of recombination events. The x-axis represents each 100 kb window and the y-axis is the total count of recombination events. Windows with high number counts of recombination events are seen as peaks on the graph and are potential recombination hotspots.

F

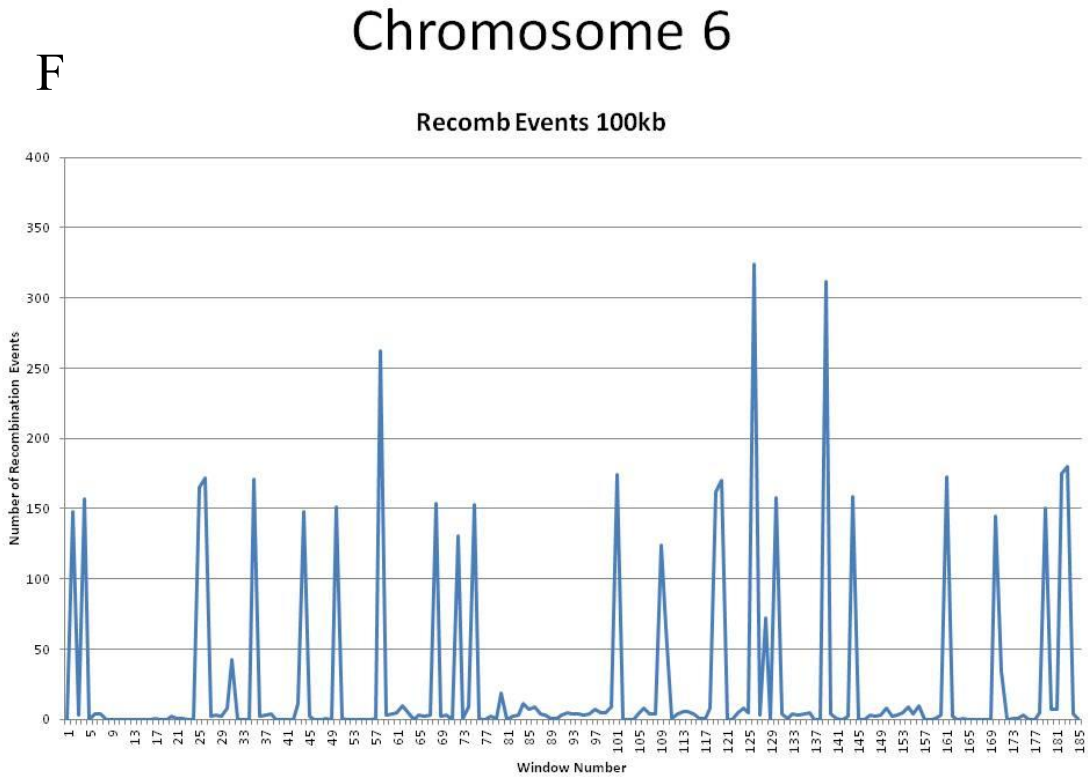

Figure 6F. Recombination Events for All 16 Chromosomes of the Honey Bee in 100 kb Windows. Each graph represents each chromosome and the number of recombination events. The x-axis represents each 100 kb window and the y-axis is the total count of recombination events. Windows with high number counts of recombination events are seen as peaks on the graph and are potential recombination hotspots.

G

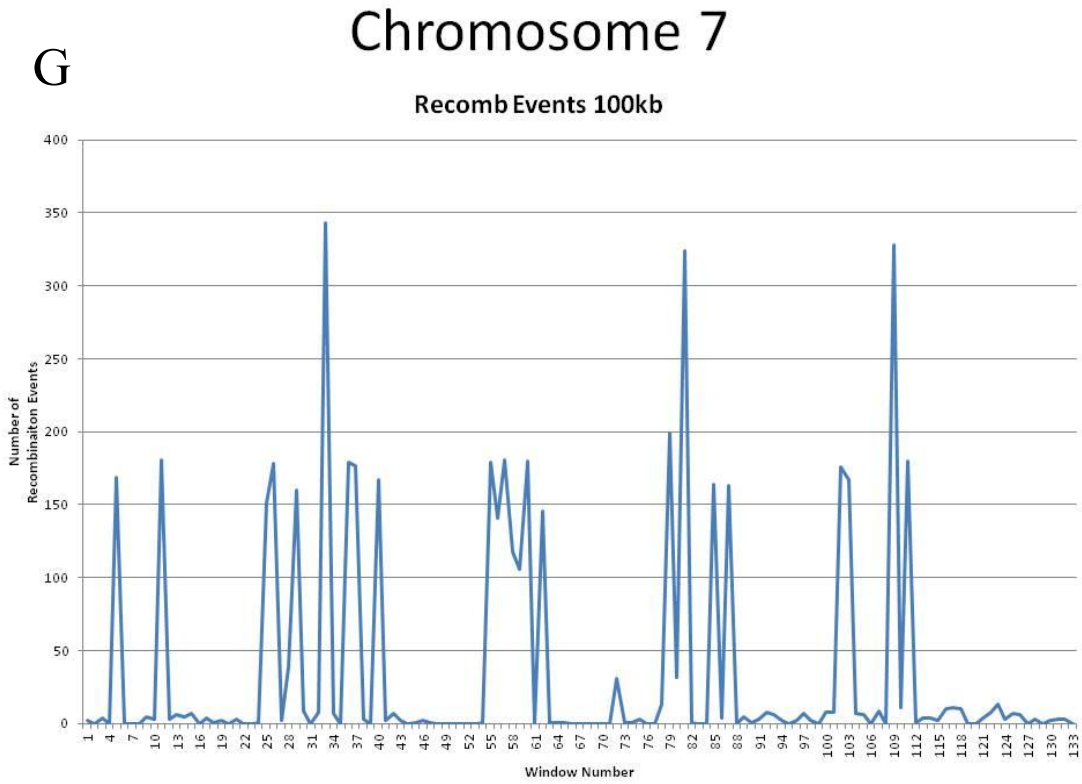

Figure 6G. Recombination Events for All 16 Chromosomes of the Honey Bee in 100 kb Windows. Each graph represents each chromosome and the number of recombination events. The x-axis represents each 100 kb window and the y-axis is the total count of recombination events. Windows with high number counts of recombination events are seen as peaks on the graph and are potential recombination hotspots.

H

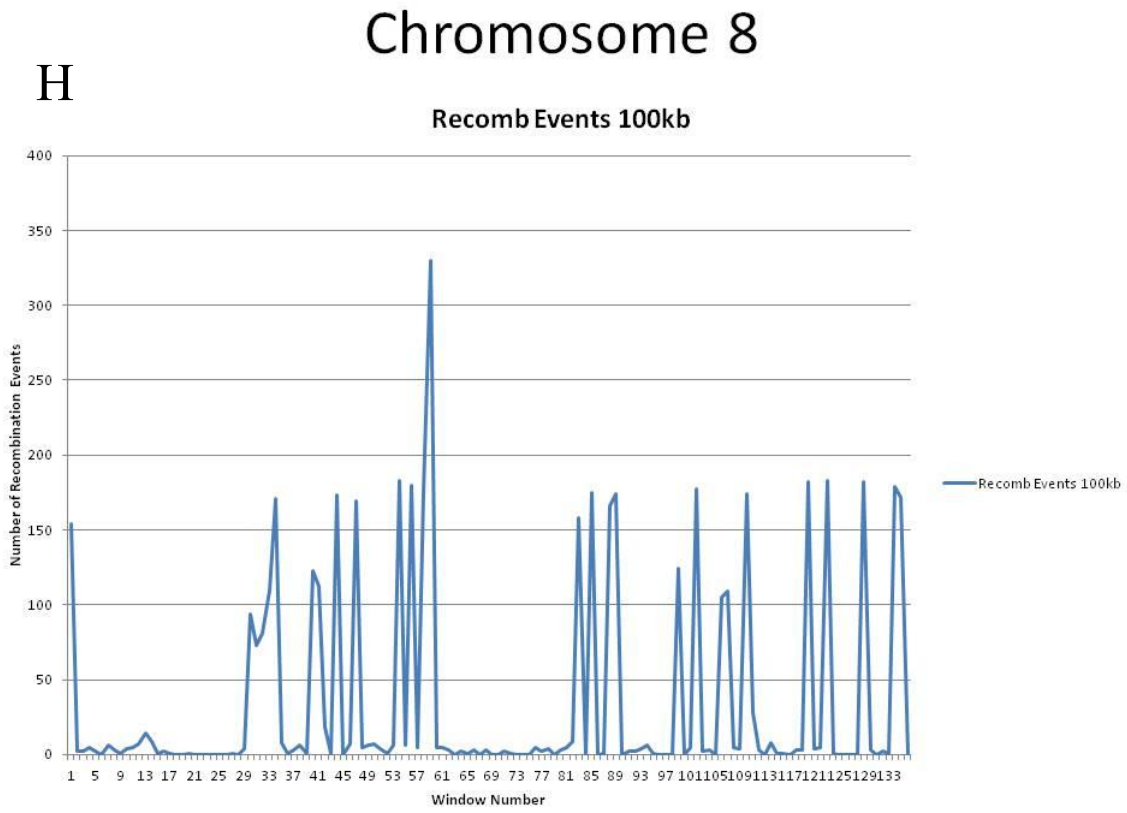

Figure 6H. Recombination Events for All 16 Chromosomes of the Honey Bee in 100 kb Windows. Each graph represents each chromosome and the number of recombination events. The x-axis represents each 100 kb window and the y-axis is the total count of recombination events. Windows with high number counts of recombination events are seen as peaks on the graph and are potential recombination hotspots.

I

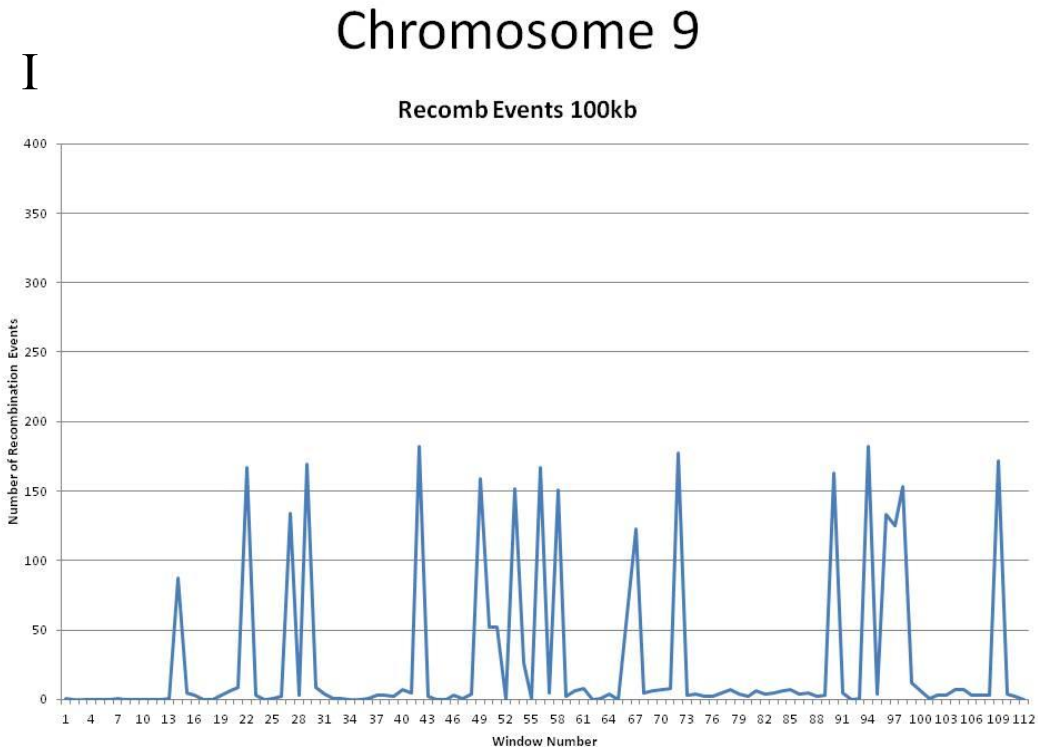

Figure 6I. Recombination Events for All 16 Chromosomes of the Honey Bee in 100 kb Windows. Each graph represents each chromosome and the number of recombination events. The x-axis represents each 100 kb window and the y-axis is the total count of recombination events. Windows with high number counts of recombination events are seen as peaks on the graph and are potential recombination hotspots.

J

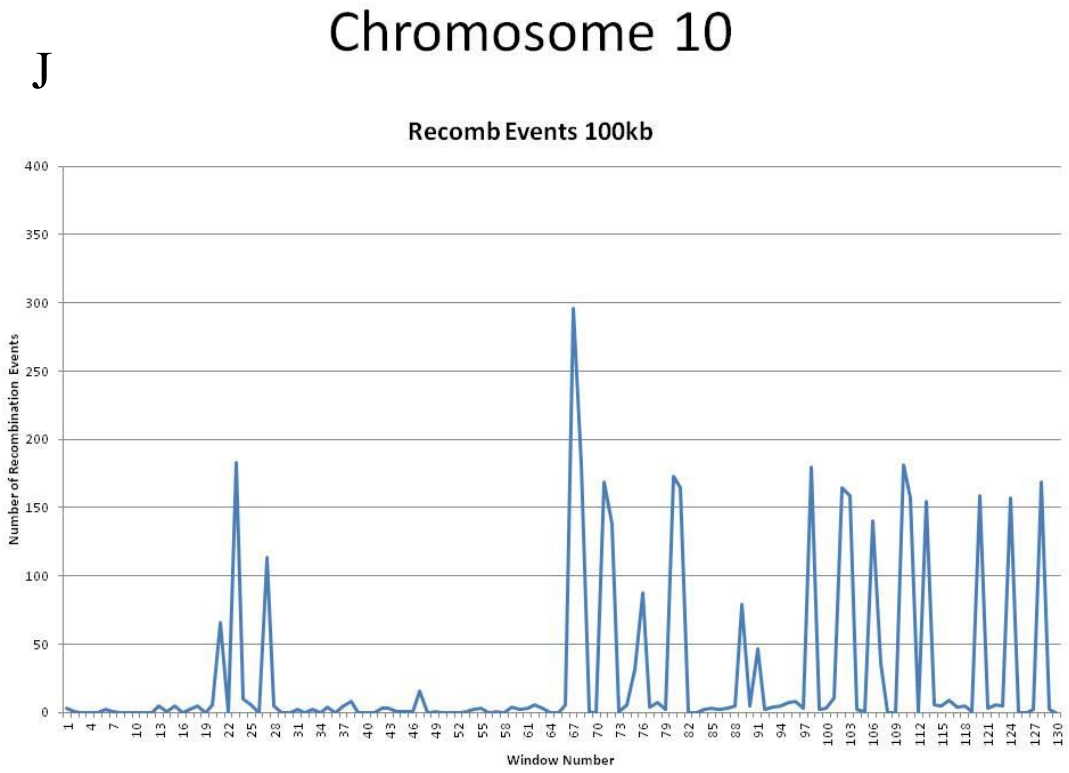

Figure 6J. Recombination Events for All 16 Chromosomes of the Honey Bee in 100 kb Windows. Each graph represents each chromosome and the number of recombination events. The x-axis represents each 100 kb window and the y-axis is the total count of recombination events. Windows with high number counts of recombination events are seen as peaks on the graph and are potential recombination hotspots.

K

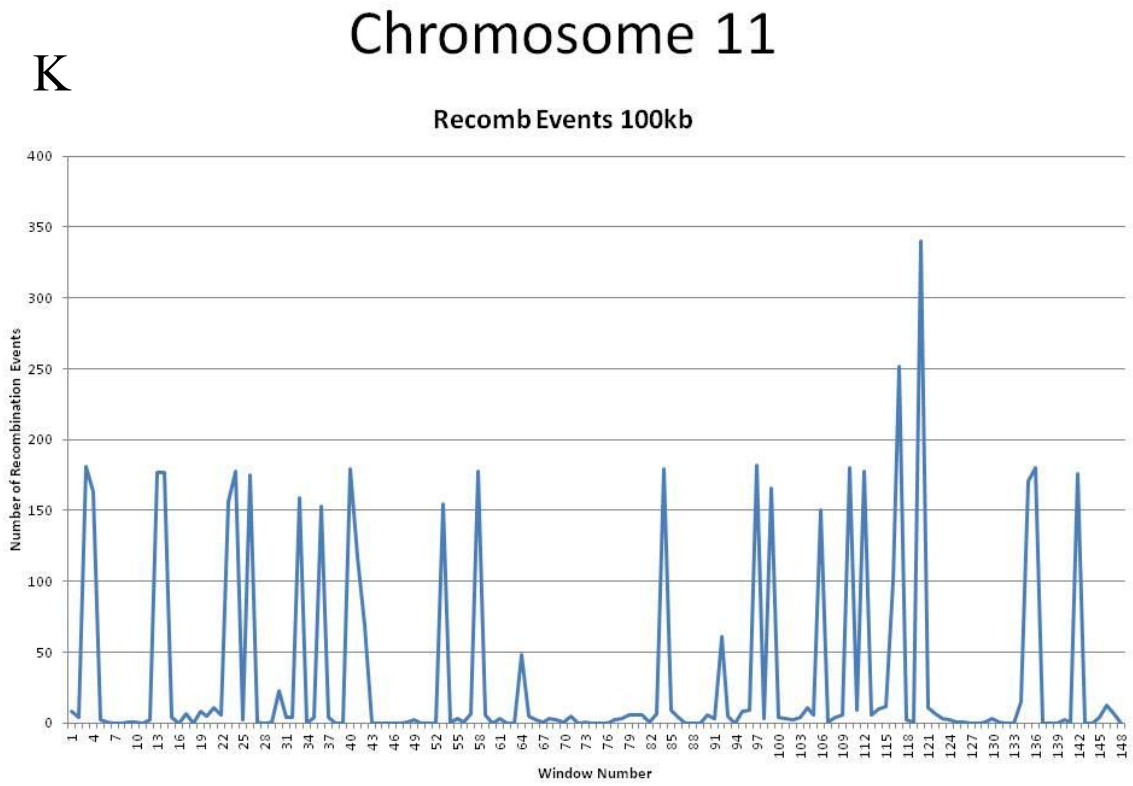

Figure 6K. Recombination Events for All 16 Chromosomes of the Honey Bee in 100 kb Windows. Each graph represents each chromosome and the number of recombination events. The x-axis represents each 100 kb window and the y-axis is the total count of recombination events. Windows with high number counts of recombination events are seen as peaks on the graph and are potential recombination hotspots.

L

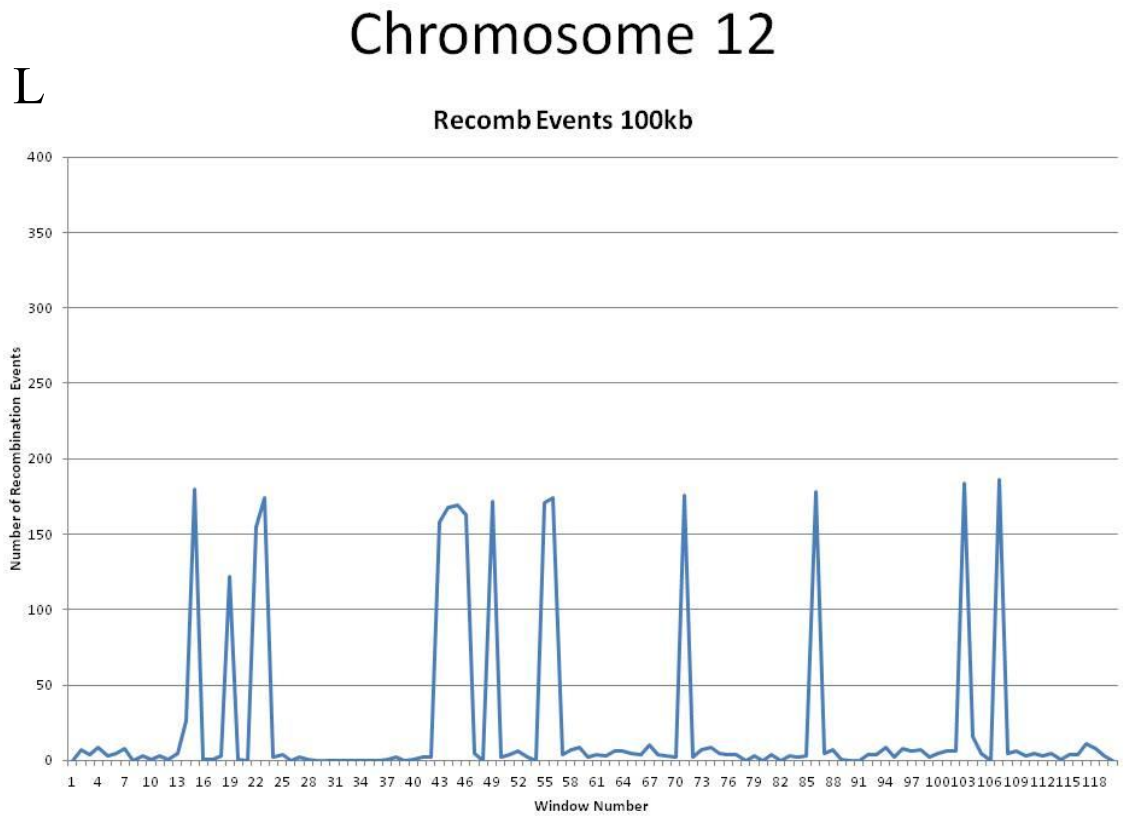

Figure 6L. Recombination Events for All 16 Chromosomes of the Honey Bee in 100 kb Windows. Each graph represents each chromosome and the number of recombination events. The x-axis represents each 100 kb window and the y-axis is the total count of recombination events. Windows with high number counts of recombination events are seen as peaks on the graph and are potential recombination hotspots.

M

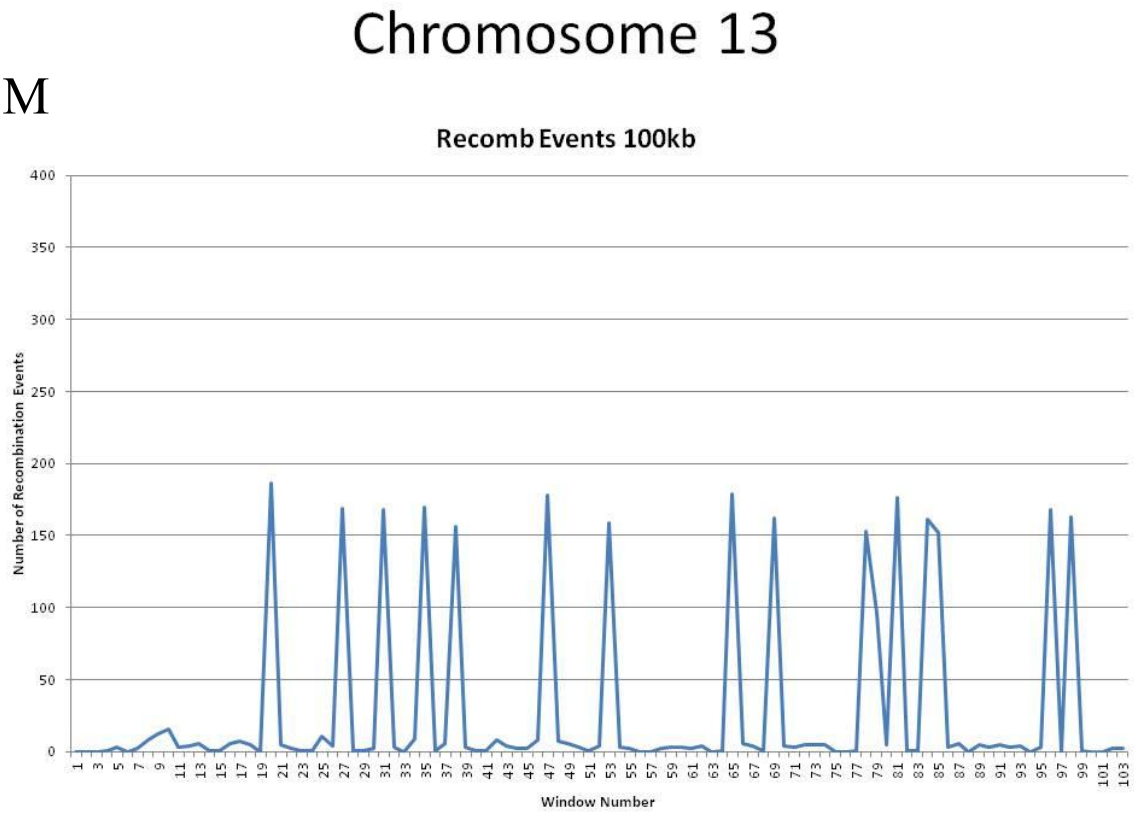

Figure 6M. Recombination Events for All 16 Chromosomes of the Honey Bee in 100 kb Windows. Each graph represents each chromosome and the number of recombination events. The x-axis represents each 100 kb window and the y-axis is the total count of recombination events. Windows with high number counts of recombination events are seen as peaks on the graph and are potential recombination hotspots.

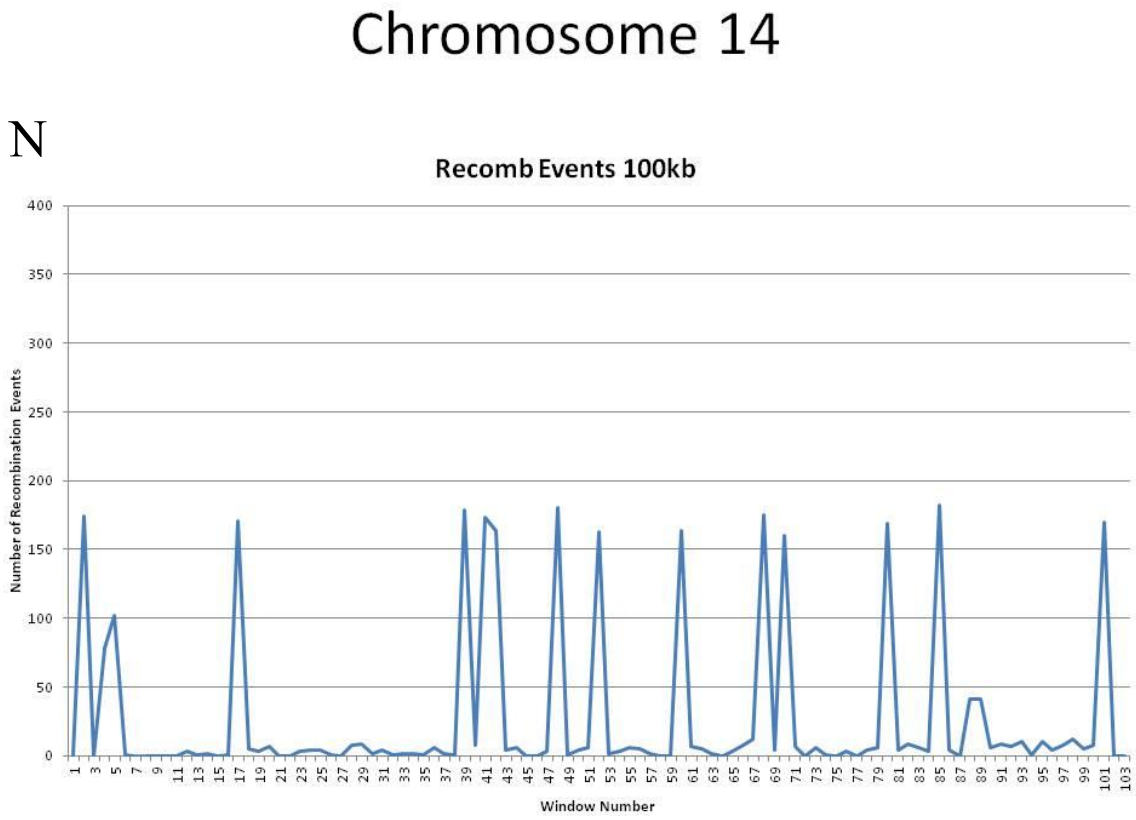

Figure 6N. Recombination Events for All 16 Chromosomes of the Honey Bee in 100 kb Windows. Each graph represents each chromosome and the number of recombination events. The x-axis represents each 100 kb window and the y-axis is the total count of recombination events. Windows with high number counts of recombination events are seen as peaks on the graph and are potential recombination hotspots.

O

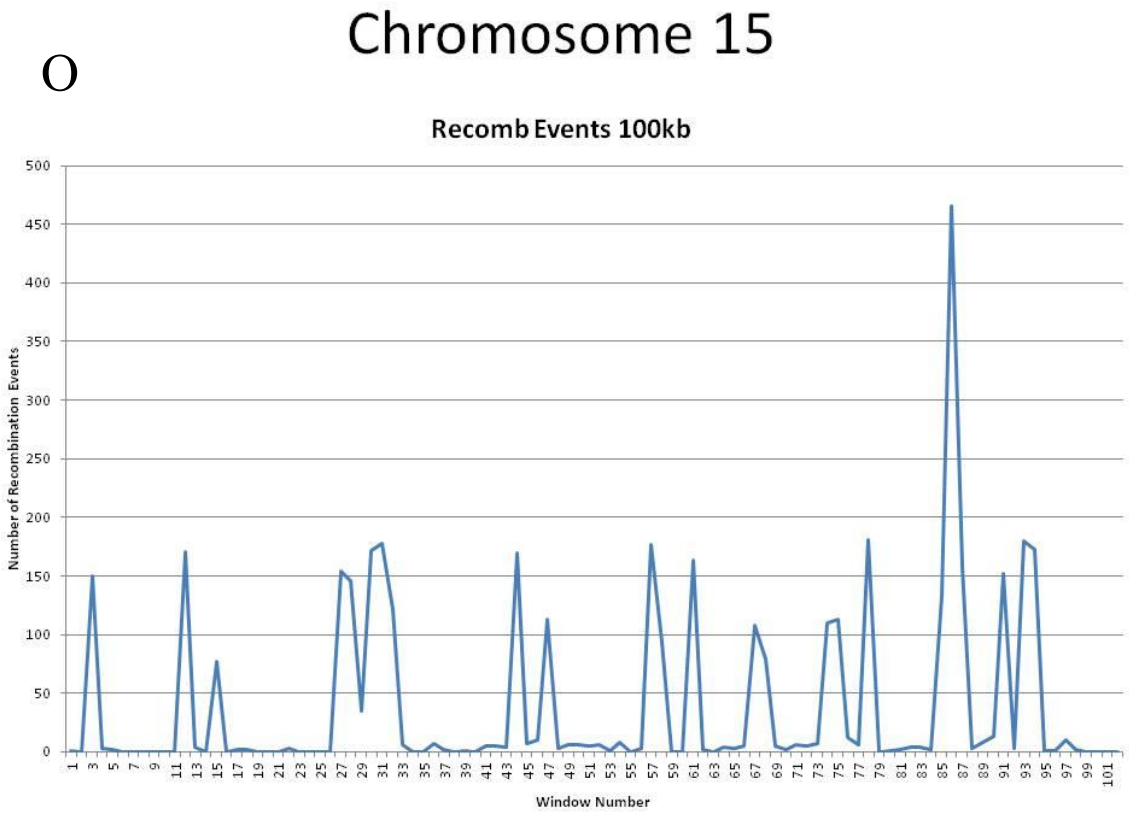

Figure 6O. Recombination Events for All 16 Chromosomes of the Honey Bee in 100 kb Windows. Each graph represents each chromosome and the number of recombination events. The x-axis represents each 100 kb window and the y-axis is the total count of recombination events. Windows with high number counts of recombination events are seen as peaks on the graph and are potential recombination hotspots.

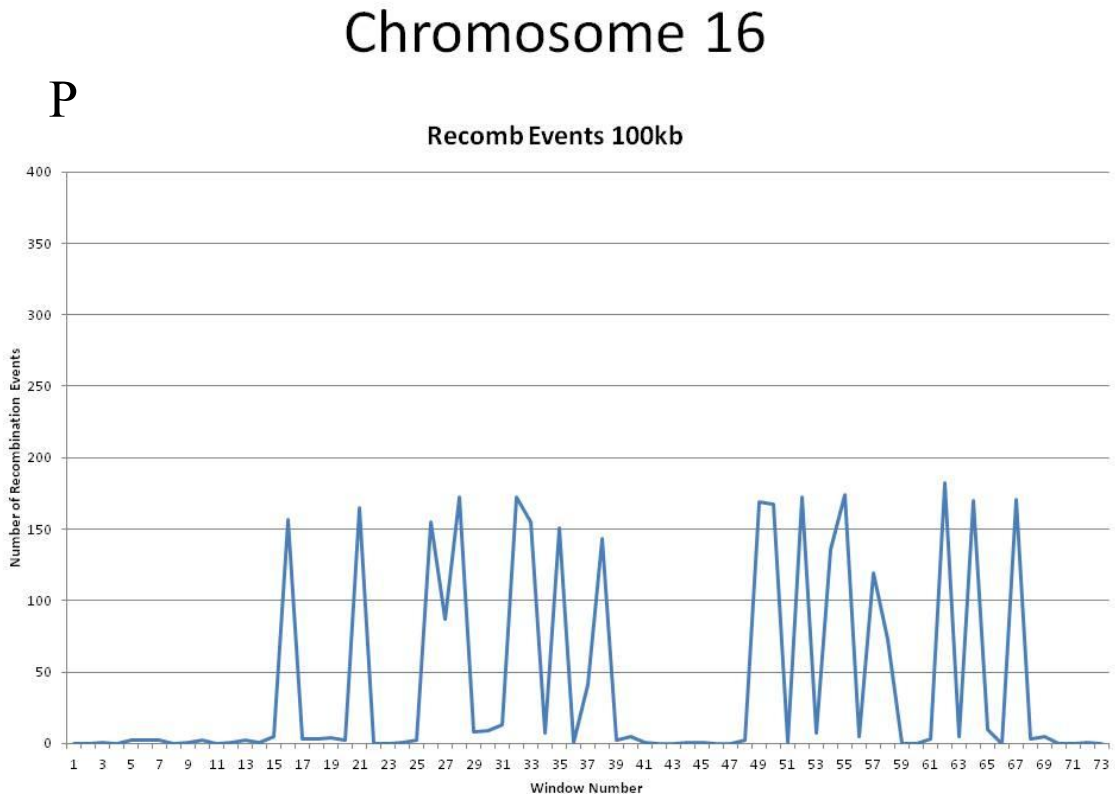

Figure 6P. Recombination Events for All 16 Chromosomes of the Honey Bee in 100 kb Windows. Each graph represents each chromosome and the number of recombination events. The x-axis represents each 100 kb window and the y-axis is the total count of recombination events. Windows with high number counts of recombination events are seen as peaks on the graph and are potential recombination hotspots.

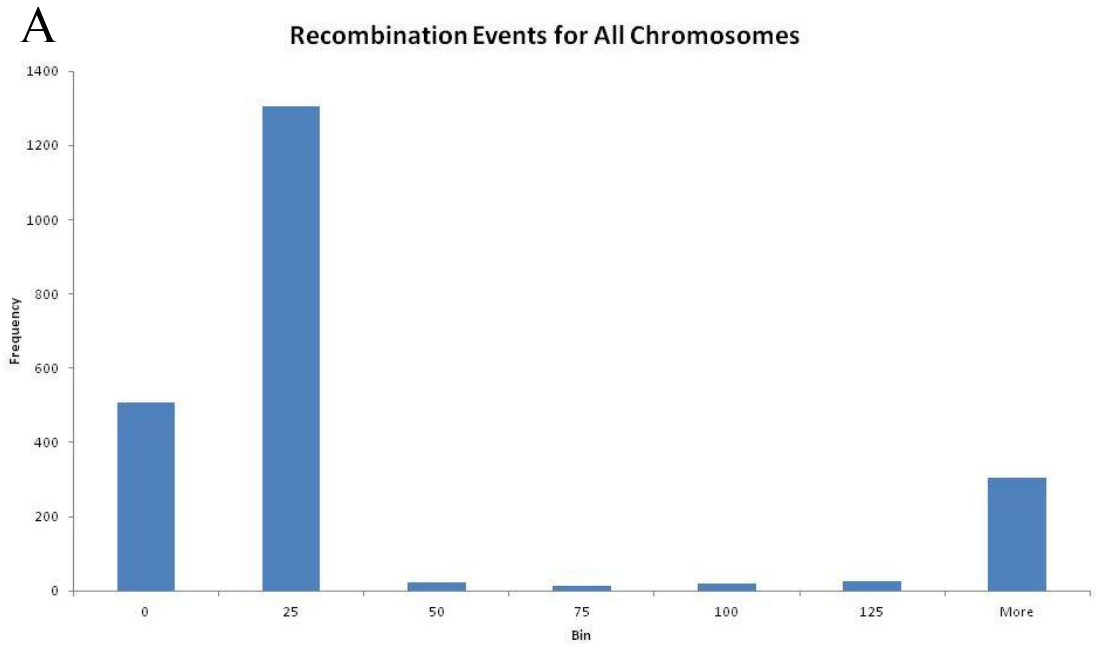

Figure 7A. Distribution of Recombination Markers across All Chromosomes. Distribution of Recombination Events are Distributed per 100 kb Window. Bin number across the x-axis represents the number of counts in a given window; frequency on the y-axis represents the number of times a recombination count for a given bin occurs. Each bin increased by 25. The histogram shows recombination events 50 counts and below per 100 kb window is most prevalent.

B

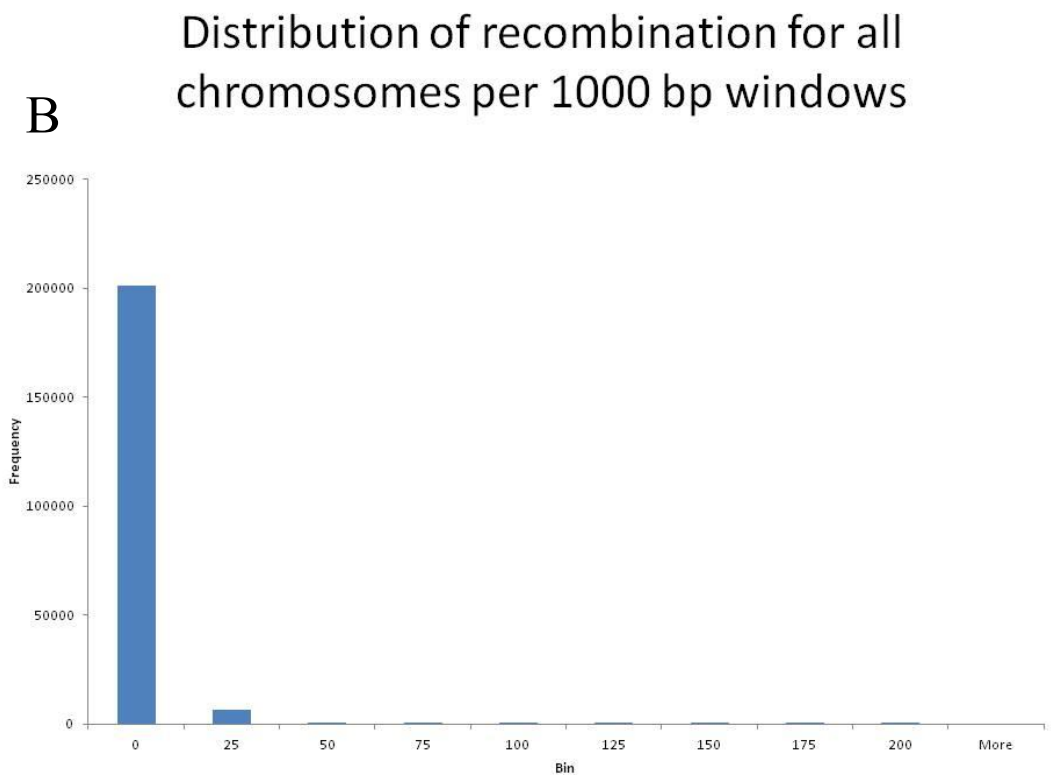

Figure 7B. Distribution of Recombination Markers across All Chromosomes. Distribution of recombination events per 1000 bp. Bin number across the x-axis represents the number of counts in a given window; frequency on the y-axis represents the number of times a recombination count for a given bin occurs. Each bin increased by 25. The histogram shows recombination events 50 counts and below per 100 kb window is most prevalent.

C

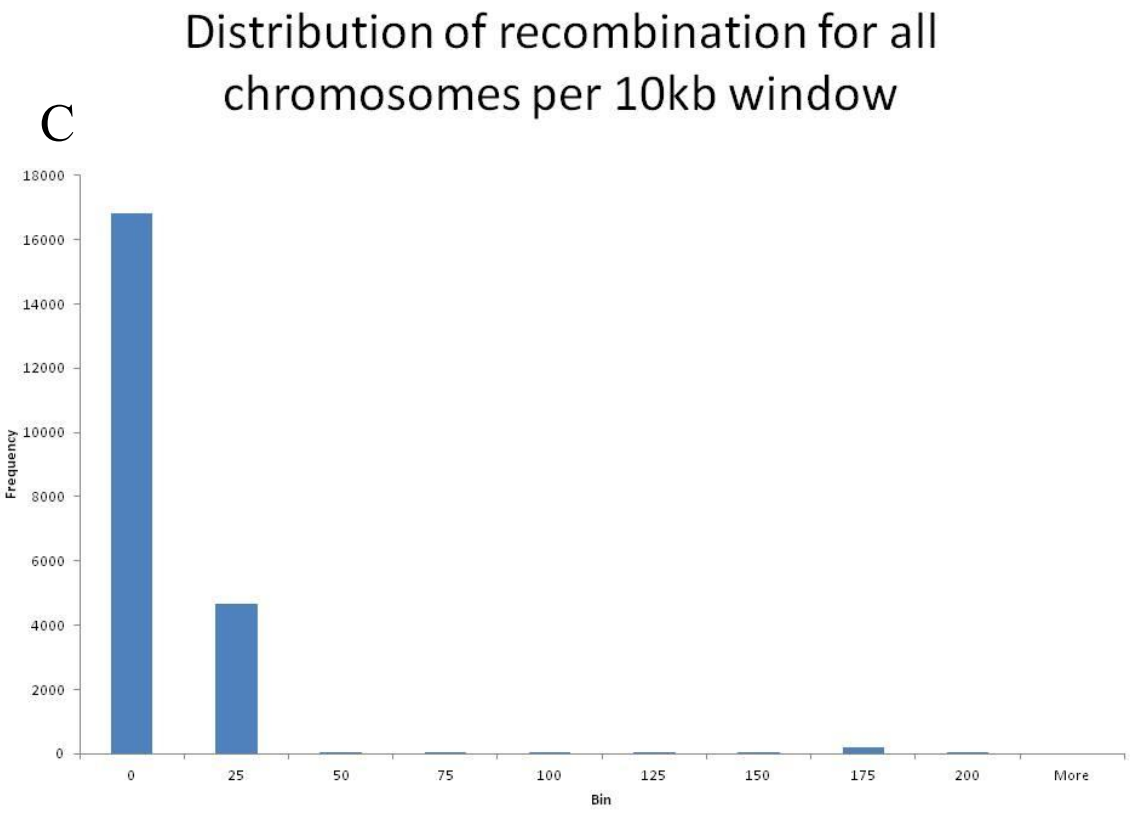

Figure 7C. Distribution of Recombination Markers across All Chromosomes. Distribution of recombination events per 10 kb. Bin number across the x-axis represents the number of counts in a given window; frequency on the y-axis represents the number of times a recombination count for a given bin occurs. Each bin increased by 25. The histogram shows recombination events 50 counts and below per 100 kb window is most prevalent.

A

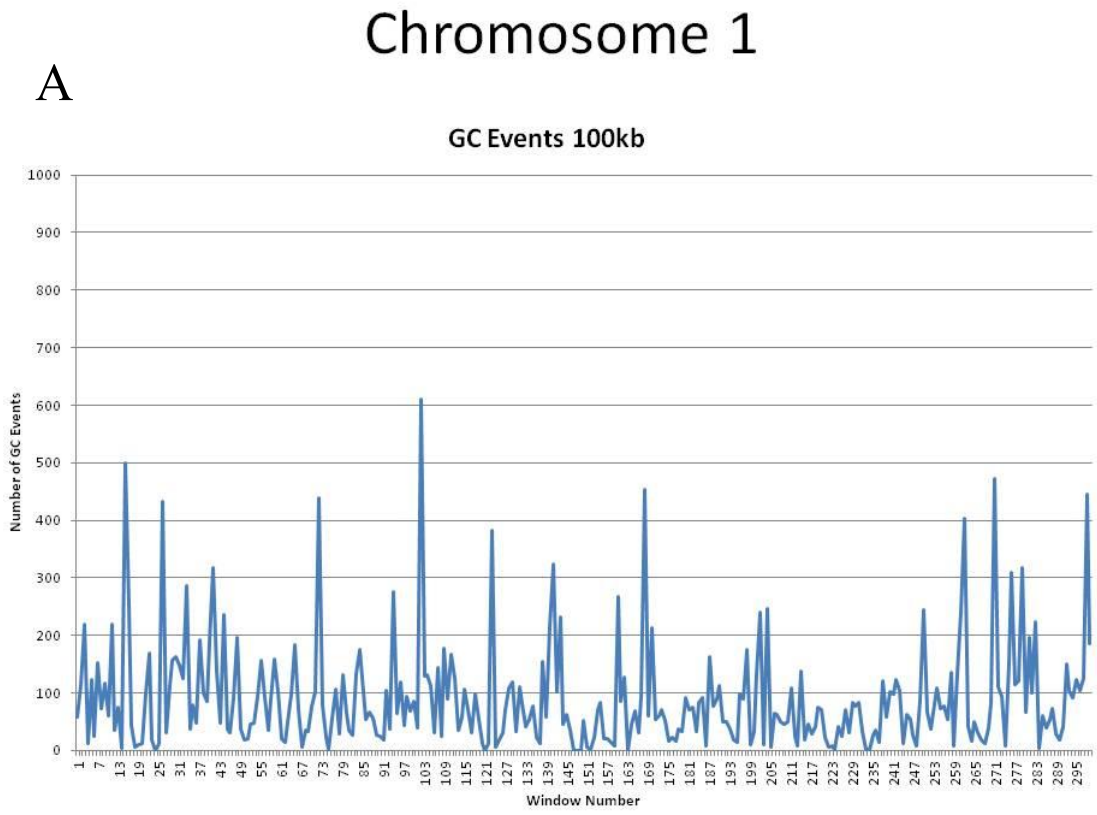

Figure 8A. Gene Conversion Events for All 16 Chromosomes of the Honey Bee in 100 kb Windows. Each graph represents each chromosome and the number of GC events occurring in each window. The x-axis represents each of the 100 kb windows in the chromosome and the y-axis represents the total count of GC events.

B

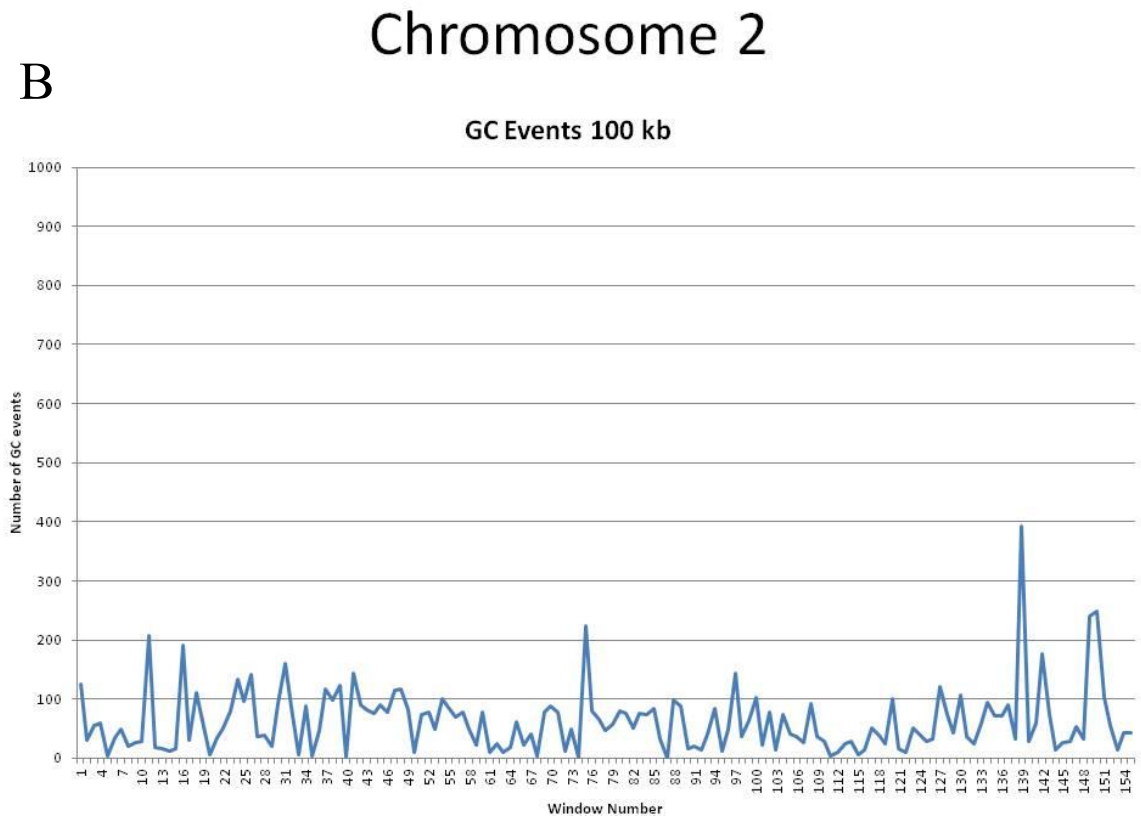

Figure 8B. Gene Conversion Events for All 16 Chromosomes of the Honey Bee in 100 kb Windows. Each graph represents each chromosome and the number of GC events occurring in each window. The x-axis represents each of the 100 kb windows in the chromosome and the y-axis represents the total count of GC events.

C

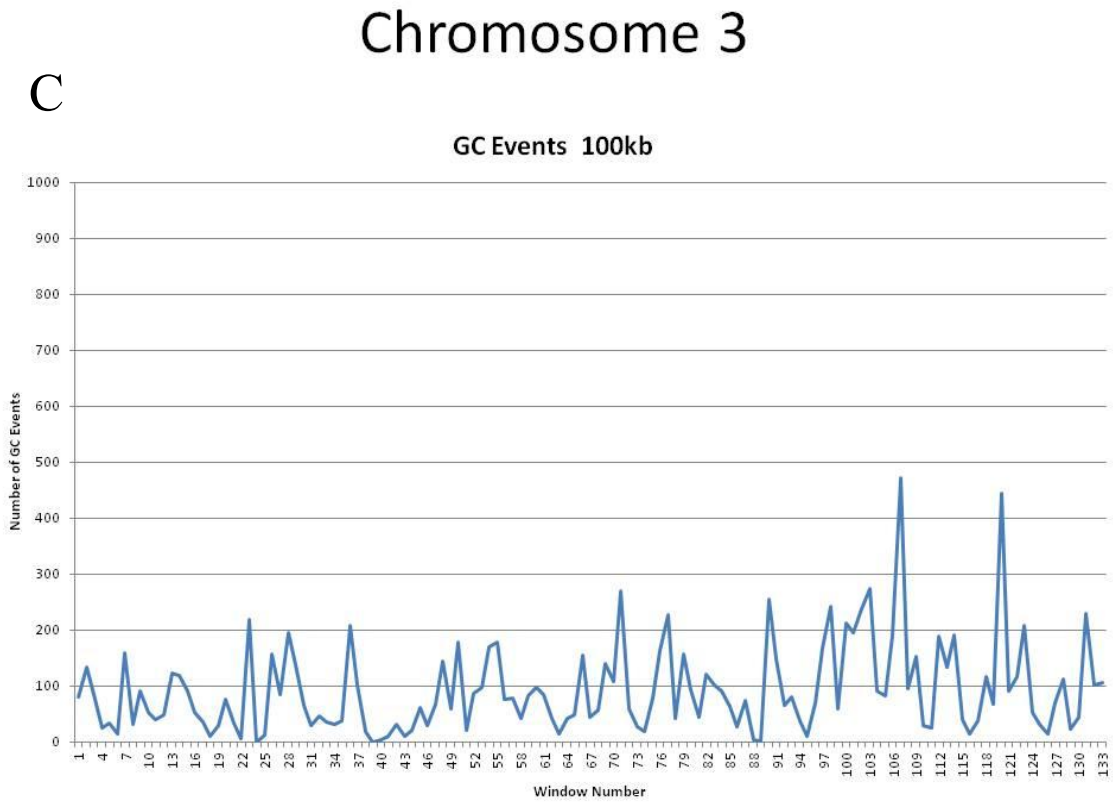

Figure 8C. Gene Conversion Events for All 16 Chromosomes of the Honey Bee in 100 kb Windows. Each graph represents each chromosome and the number of GC events occurring in each window. The x-axis represents each of the 100 kb windows in the chromosome and the y-axis represents the total count of GC events.

Chromosome 4

D

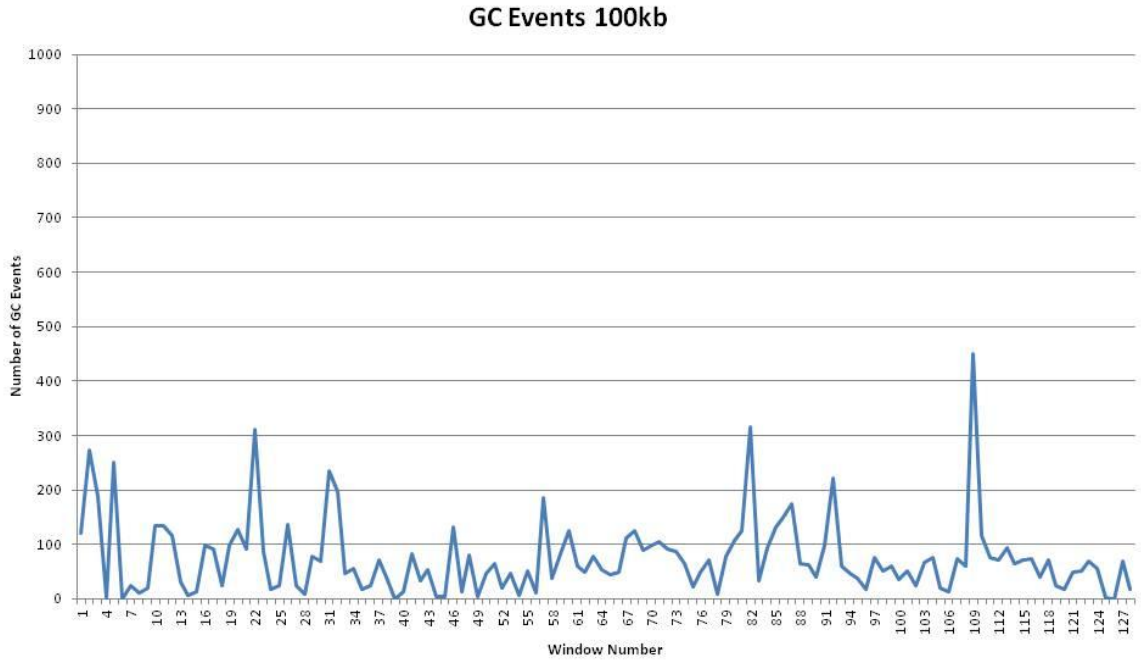

Figure 8D. Gene Conversion Events for All 16 Chromosomes of the Honey Bee in 100 kb Windows. Each graph represents each chromosome and the number of GC events occurring in each window. The x-axis represents each of the 100 kb windows in the chromosome and the y-axis represents the total count of GC events.

E

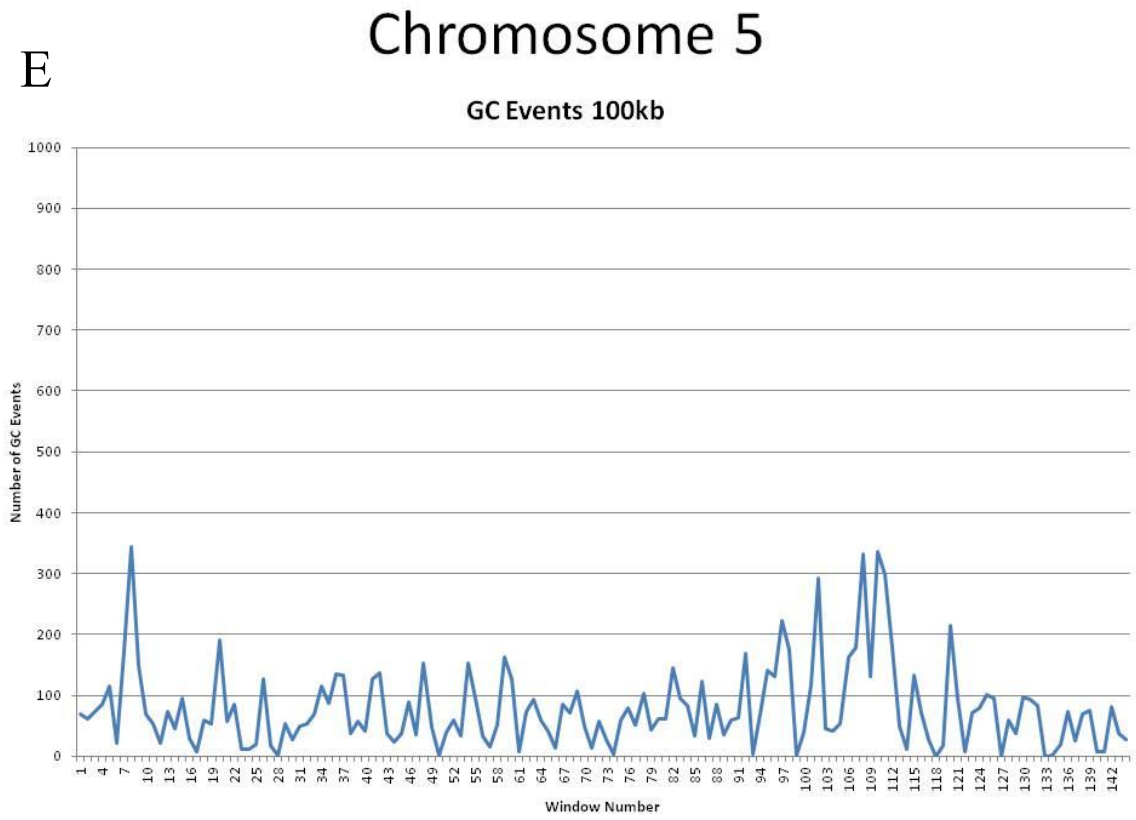

Figure 8E. Gene Conversion Events for All 16 Chromosomes of the Honey Bee in 100 kb Windows. Each graph represents each chromosome and the number of GC events occurring in each window. The x-axis represents each of the 100 kb windows in the chromosome and the y-axis represents the total count of GC events.

F

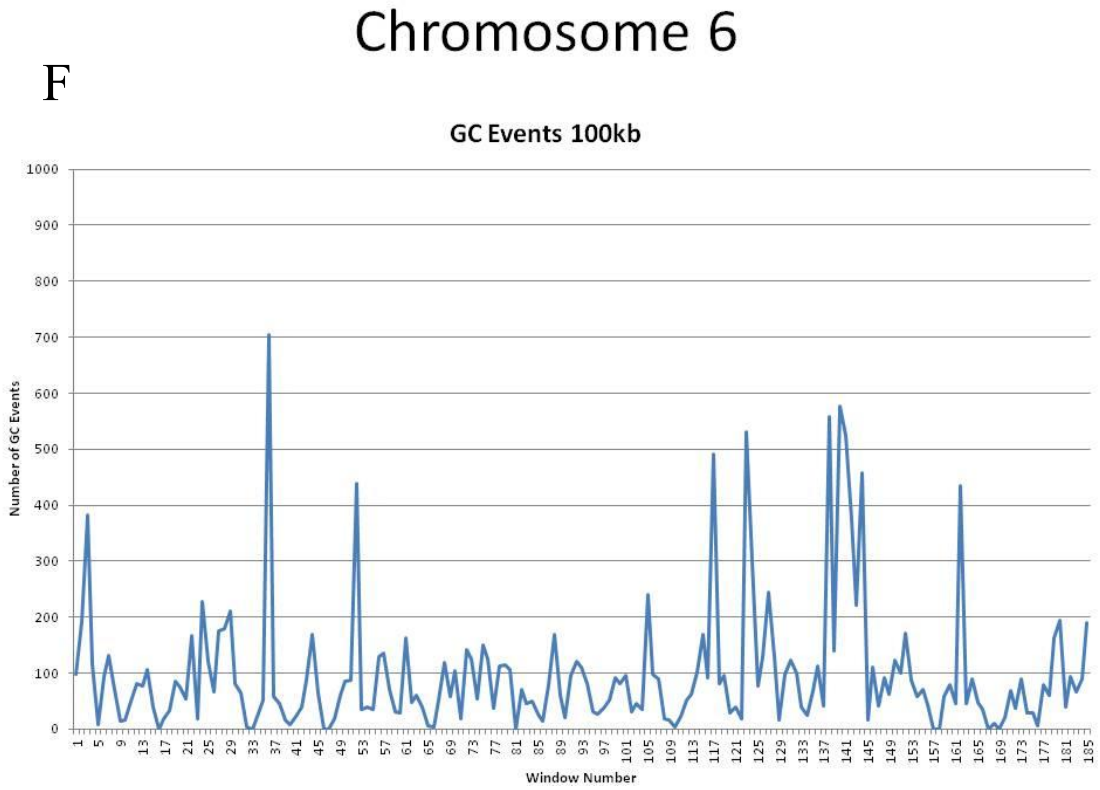

Figure 8F. Gene Conversion Events for All 16 Chromosomes of the Honey Bee in 100 kb Windows. Each graph represents each chromosome and the number of GC events occurring in each window. The x-axis represents each of the 100 kb windows in the chromosome and the y-axis represents the total count of GC events.

G

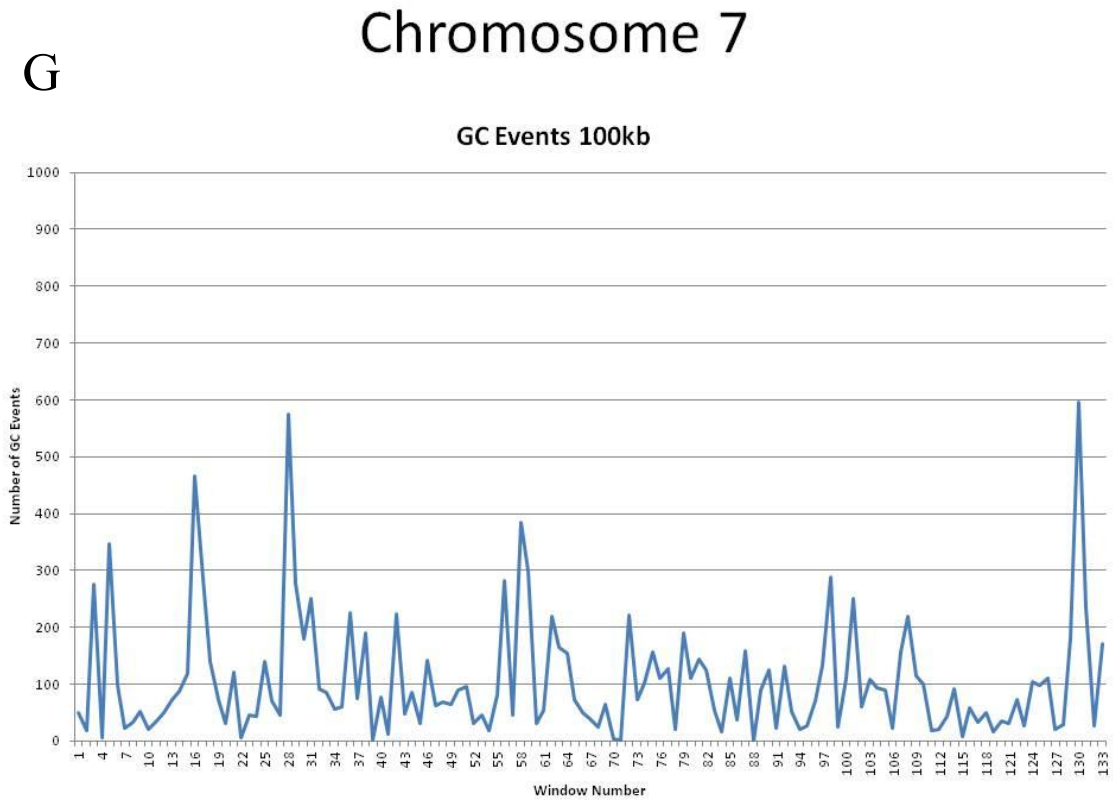

Figure 8G. Gene Conversion Events for All 16 Chromosomes of the Honey Bee in 100 kb Windows. Each graph represents each chromosome and the number of GC events occurring in each window. The x-axis represents each of the 100 kb windows in the chromosome and the y-axis represents the total count of GC events.

H

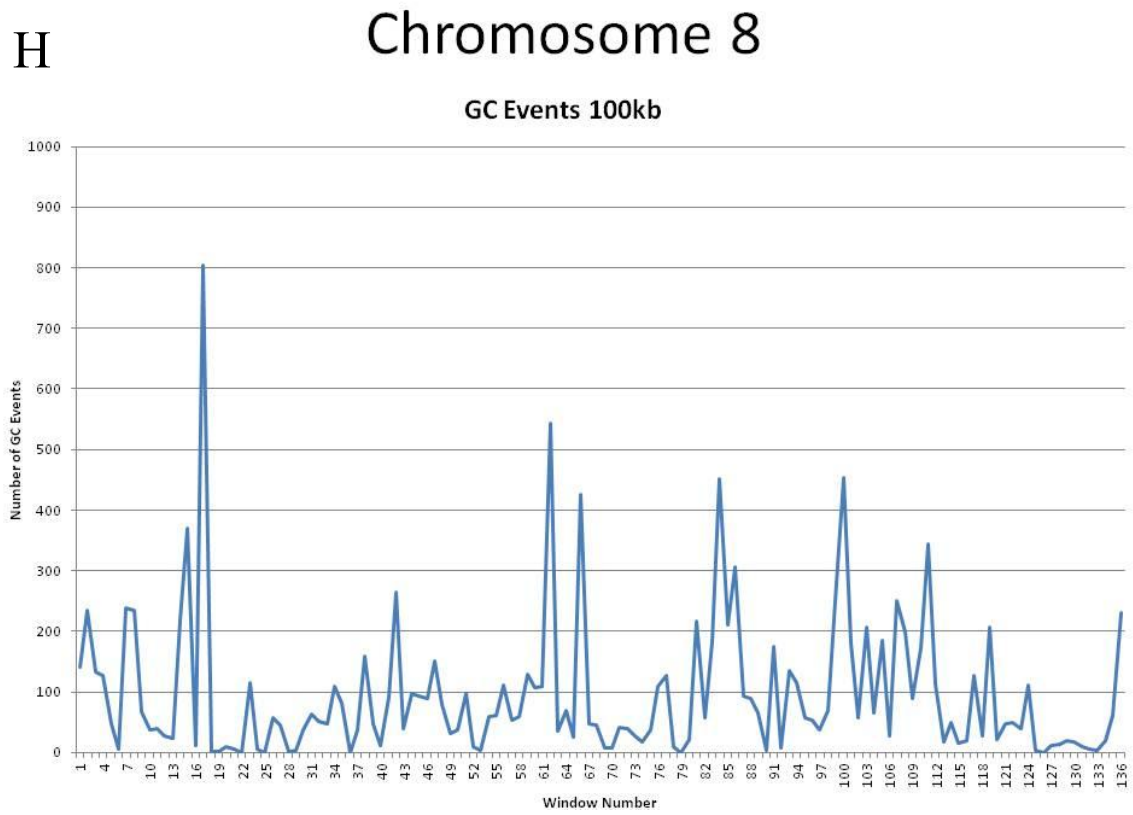

Figure 8H. Gene Conversion Events for All 16 Chromosomes of the Honey Bee in 100 kb Windows. Each graph represents each chromosome and the number of GC events occurring in each window. The x-axis represents each of the 100 kb windows in the chromosome and the y-axis represents the total count of GC events.

I

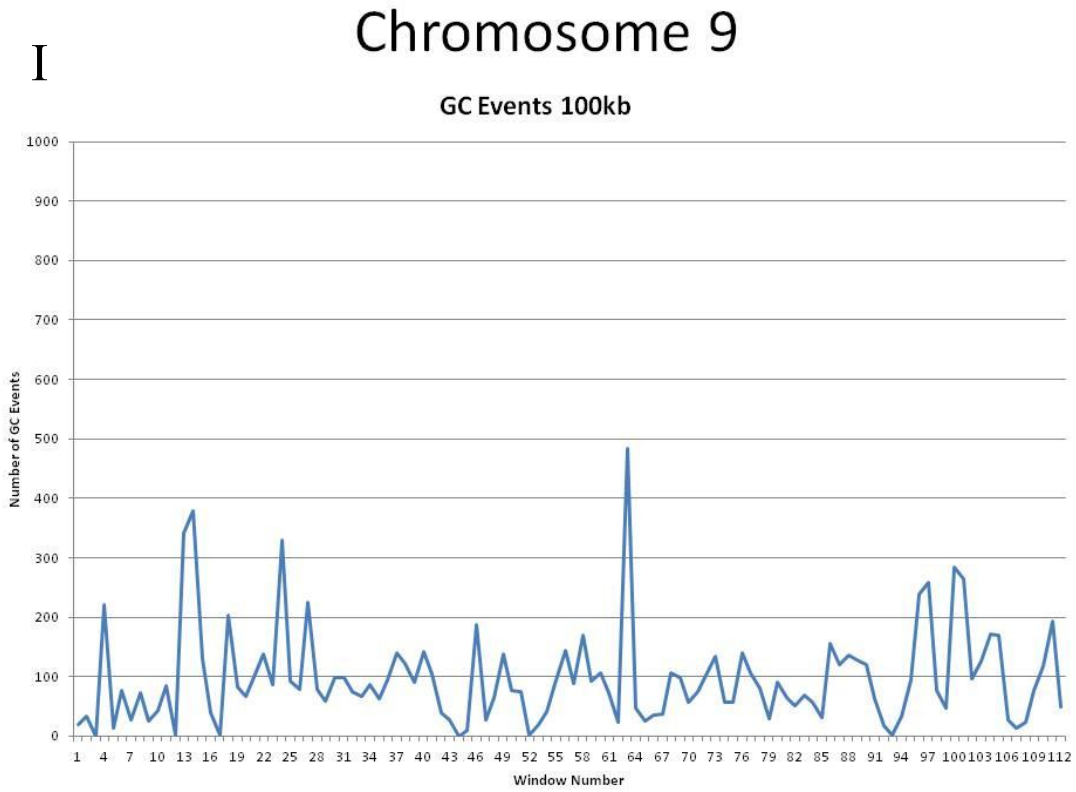

Figure 8I. Gene Conversion Events for All 16 Chromosomes of the Honey Bee in 100 kb Windows. Each graph represents each chromosome and the number of GC events occurring in each window. The x-axis represents each of the 100 kb windows in the chromosome and the y-axis represents the total count of GC events.

J

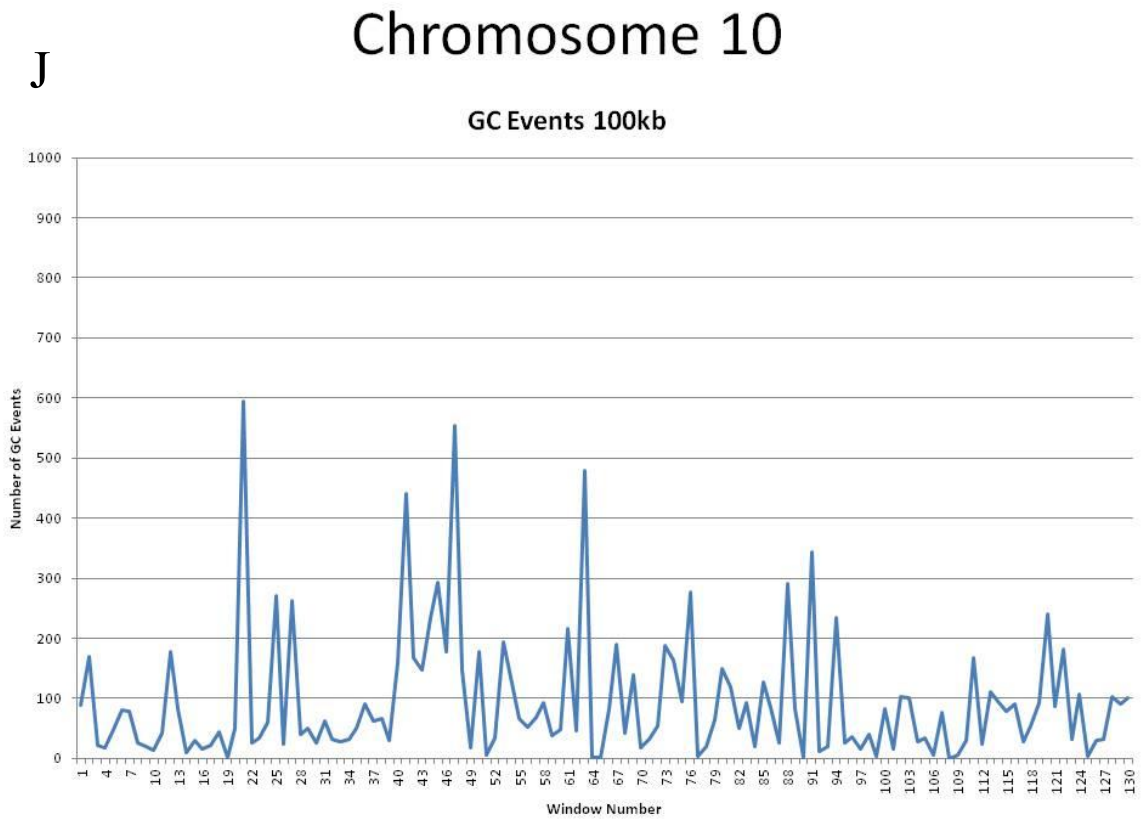

Figure 8J. Gene Conversion Events for All 16 Chromosomes of the Honey Bee in 100 kb Windows. Each graph represents each chromosome and the number of GC events occurring in each window. The x-axis represents each of the 100 kb windows in the chromosome and the y-axis represents the total count of GC events.

K

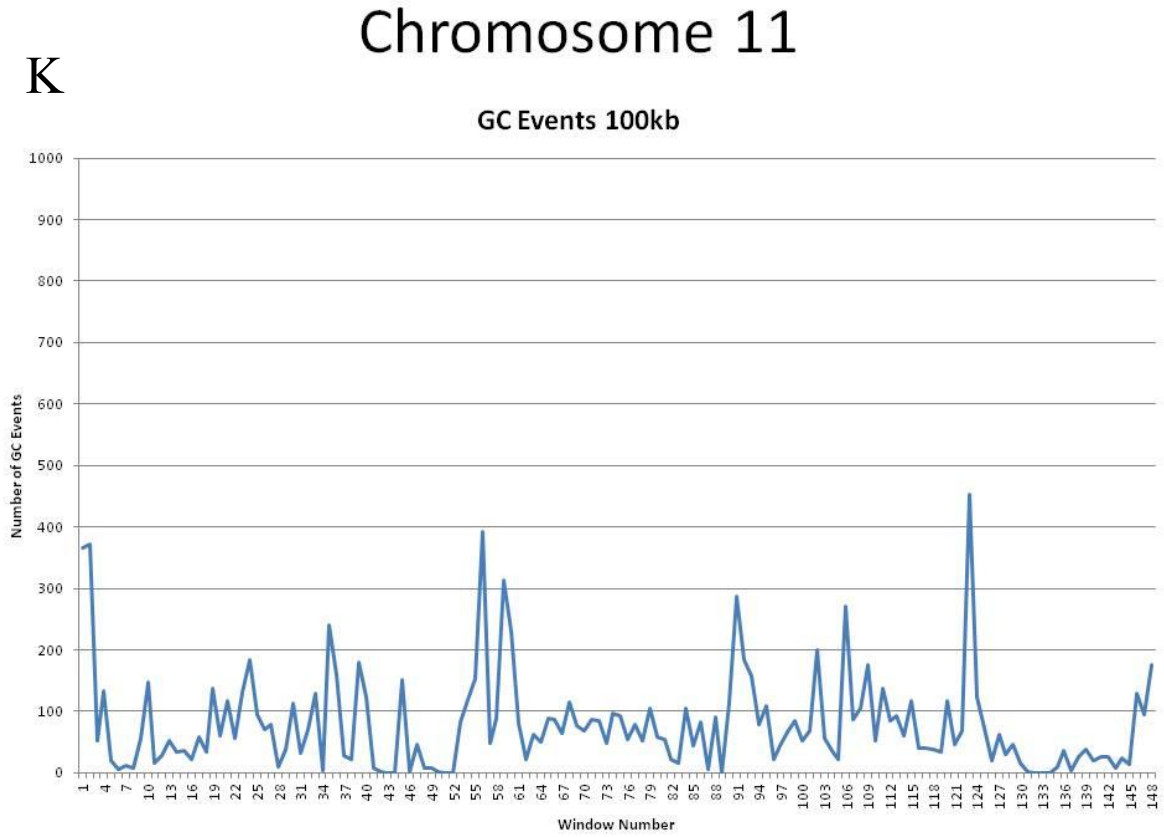

Figure 8K. Gene Conversion Events for All 16 Chromosomes of the Honey Bee in 100 kb Windows. Each graph represents each chromosome and the number of GC events occurring in each window. The x-axis represents each of the 100 kb windows in the chromosome and the y-axis represents the total count of GC events.

L

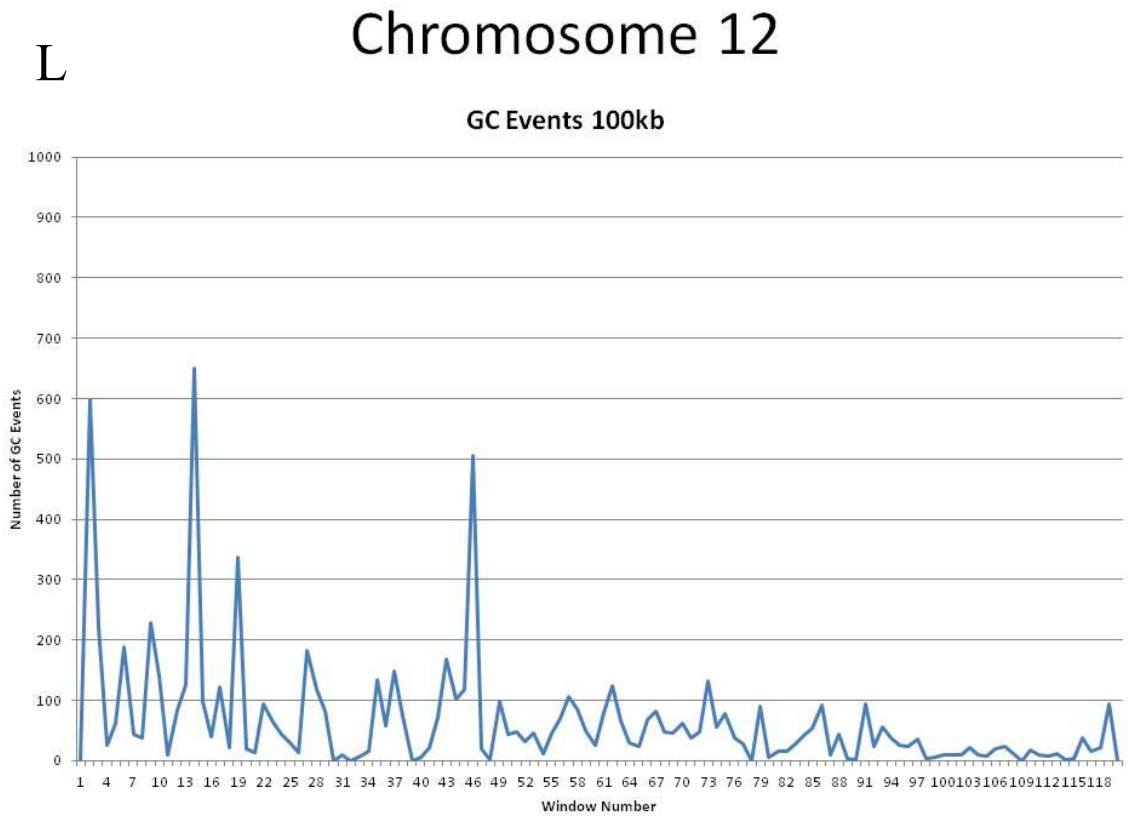

Figure 8L. Gene Conversion Events for All 16 Chromosomes of the Honey Bee in 100 kb Windows. Each graph represents each chromosome and the number of GC events occurring in each window. The x-axis represents each of the 100 kb windows in the chromosome and the y-axis represents the total count of GC events.

M

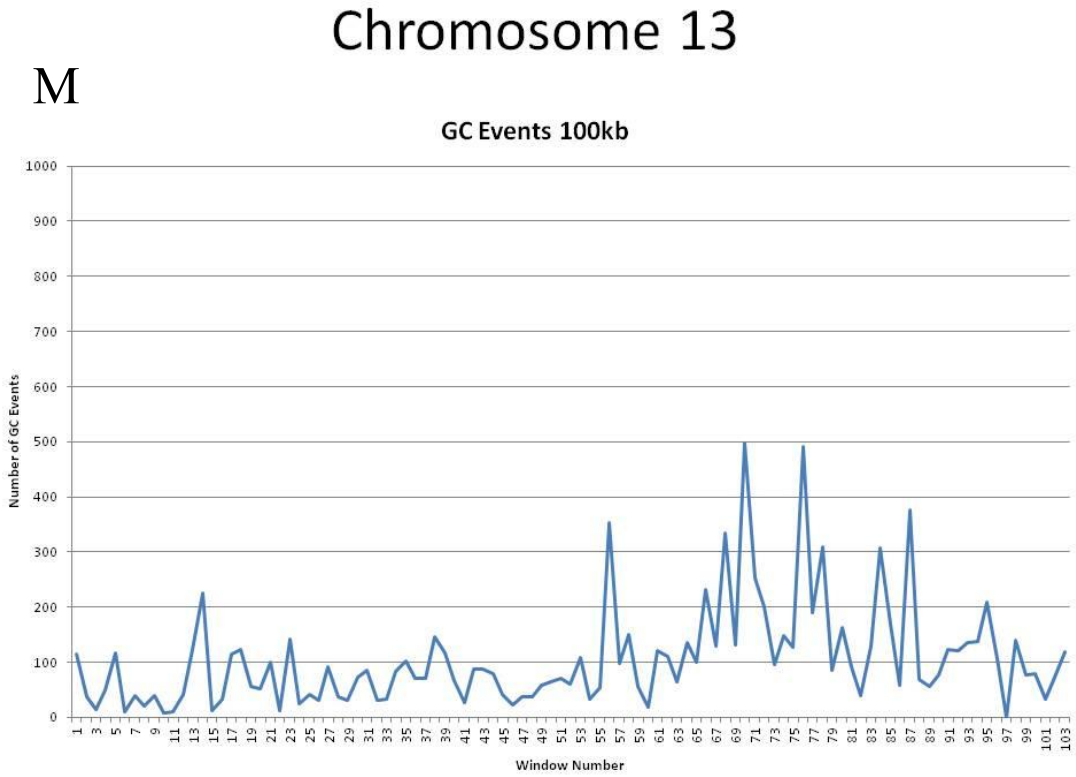

Figure 8M. Gene Conversion Events for All 16 Chromosomes of the Honey Bee in 100 kb Windows. Each graph represents each chromosome and the number of GC events occurring in each window. The x-axis represents each of the 100 kb windows in the chromosome and the y-axis represents the total count of GC events.

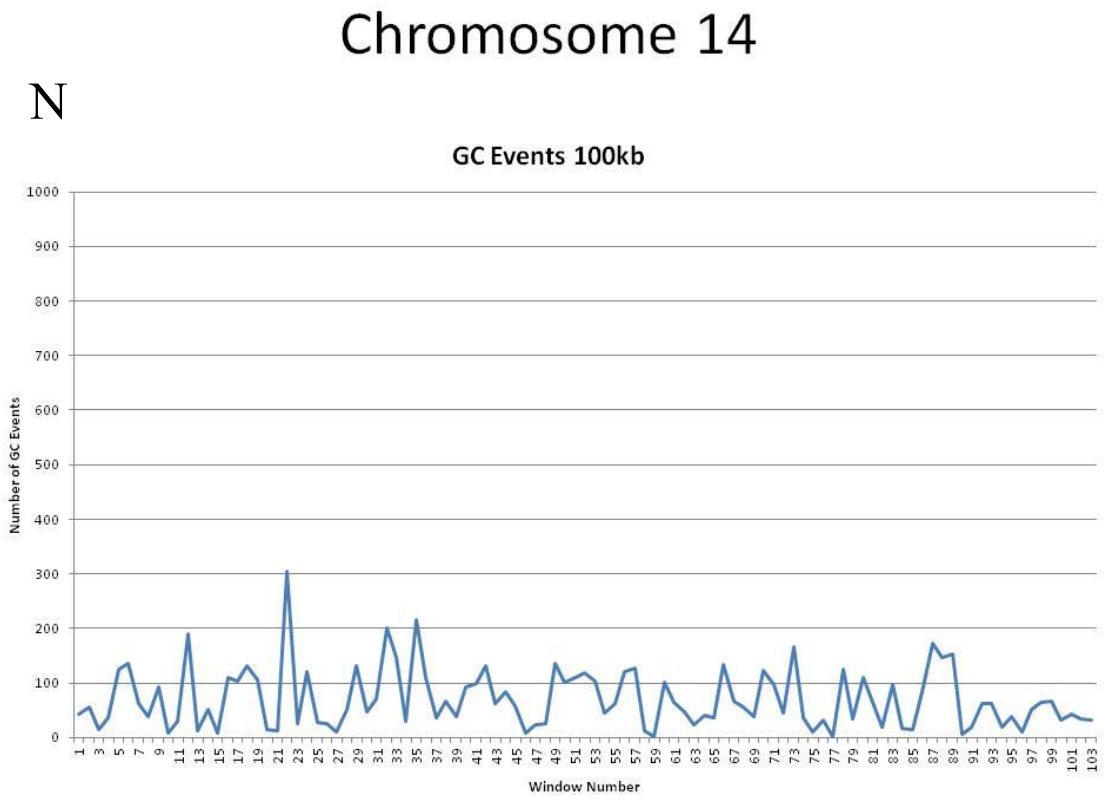

Figure 8N. Gene Conversion Events for All 16 Chromosomes of the Honey Bee in 100 kb Windows. Each graph represents each chromosome and the number of GC events occurring in each window. The x-axis represents each of the 100 kb windows in the chromosome and the y-axis represents the total count of GC events.

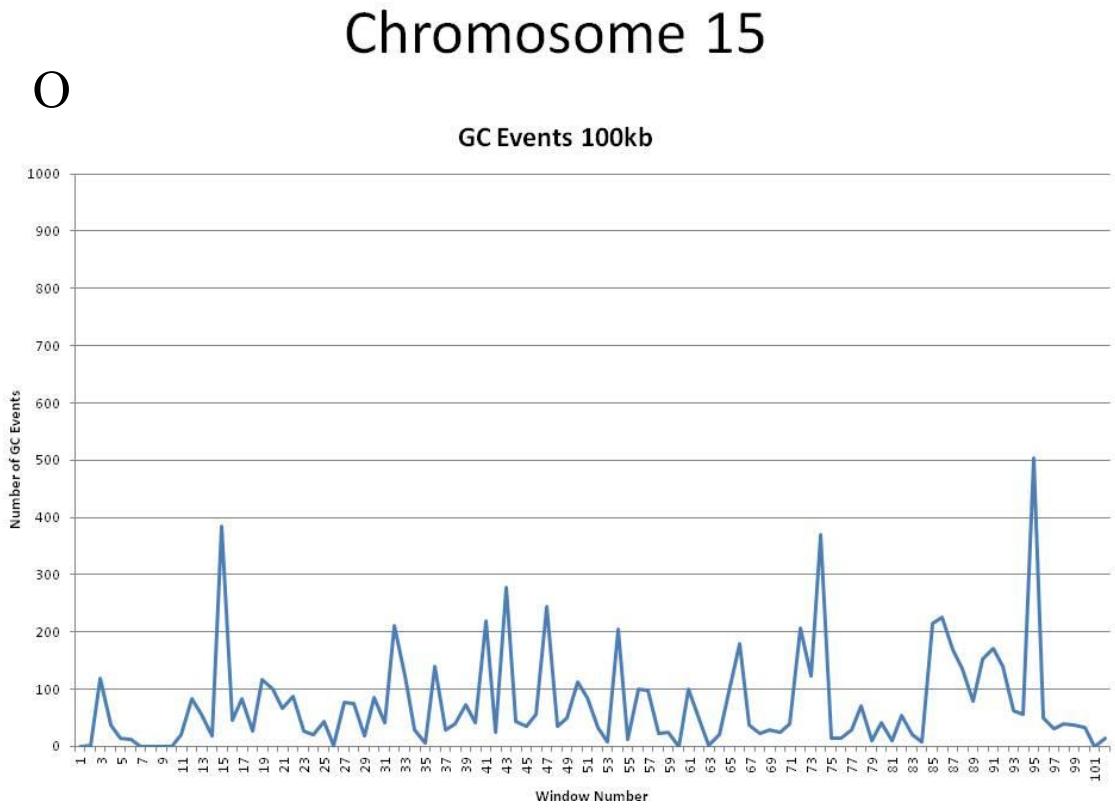

Figure 8O. Gene Conversion Events for All 16 Chromosomes of the Honey Bee in 100 kb Windows. Each graph represents each chromosome and the number of GC events occurring in each window. The x-axis represents each of the 100 kb windows in the chromosome and the y-axis represents the total count of GC events.

Chromosome 16

P

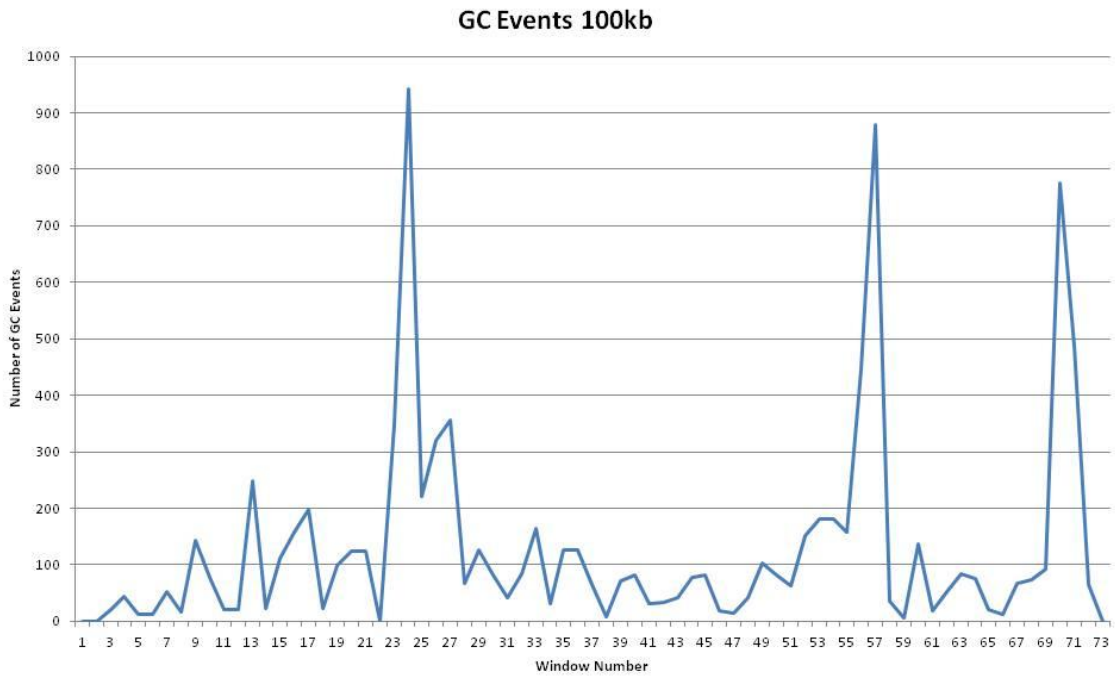

Figure 8P. Gene Conversion Events for All 16 Chromosomes of the Honey Bee in 100 kb Windows. Each graph represents each chromosome and the number of GC events occurring in each window. The x-axis represents each of the 100 kb windows in the chromosome and the y-axis represents the total count of GC events.