Searching for the genetic basis of hygienic behavior and overwintering in the honeybee (*Apis mellifera*)

Harshilkumar Patel

A Thesis submitted to the Faculty of Graduate Studies in Partial Fulfillment of the Requirements for the Degree of Master of Science

Graduate Program in Biology,
York University,
Toronto,
Ontario

September 2018

Abstract

The recent decline in honeybee populations can be mitigated through genomics and marker-assisted selection. The current techniques, such as chemical treatment to prevent disease, are only short-term solutions. The ability to breed honeybees that are disease and winter resistant would be ideal. Current breeding techniques lack knowledge of predictive markers that may improve these traits. Here we perform a genome-wide association study on 925 colonies by measuring hygienic and overwintering behavior of the colonies, followed by sequencing their genomes. L1 regression is a technique developed to pick the best Single Nucleotide Polymorphisms that explain the variance in the phenotype. Using L1 regression, we found 27 Single Nucleotide Polymorphisms for hygiene and 32 Single Nucleotide Polymorphisms for overwintering behaviour that could be used to breed for healthier and winter hardy honeybees.

Acknowledgements

This project would not be possible without my supervisor Dr. Amro Zayed. I would like to thank him for giving me the opportunity to work in his lab under his guidance. He helped shape my understanding on genomics and improved my scientific thinking and writing skills. Special thanks to Stephen Rose and Tanushree Tiwari who really helped me understand the nuances of genomic data and taught me about the tools that allowed me to work with this data. I also appreciate the discussions we had either during lunch or in the office about the statistical challenges the project entails. I will forever remember our time here. I would like to thank Alivia Dey and her team of super undergrads for handling the extraction of DNA for sequencing. I also thank Dr. Matt Betti for his mathematical input and knowledge drops, and for spending much of his time reading math equations at my behest. I would like to thank Dr. Clement Kent for his guidance on the biological aspects of the project. I thank Renata and the Beeomics field teams, without whom we would have no phenotypic data to use. Honorary mention to all the das that crossed our path. Finally, I thank the Zayed lab members for the discussions that led to both fun and insight.

Table of Contents

Abstract…………………………………………………………………………………ii

Acknowledgements..…………………………………………………………………..iii

Table of Contents………………………………………………………………………iv

List of Tables…………………………………………………………………………...v

List of Figures…………………………………………………………………………vi

Introduction……………………………………………………………………………..1

Chapter 1: Hygienic Behavior………………………………………………………. 6

1.1 Introduction…………………………………………………………………..…..6

1.4 Methods…………………………………………………………………………. 8

1.4.1 Colony and Phenotype……………………………………………………… 8

1.4.2 Extraction and Sequencing………………………………………………… 9

1.4.3 Alignment and SNP calling………………………………………………….. 11

1.4.4 Genome-Wide Association (GWA) analysis………………………………… 14

1.5 Results……………………………………………………………………………..17

1.6 Discussion……………………………………………………………………… 21

1.7 Tables and Figures……………………………………………………………... 26

Chapter 2: The Genetics of Overwintering Behavior…………………………………38

2.1 Introduction………………………………………………………………………..38

2.4 Methods……………………………………………………………………………40

2.4.1 Colony and Phenotype………………………………………………………... 40

2.4.2 Extraction and Sequencing…………………………………………………... 40

2.4.3 Alignment and SNP calling………………………………………………….. 40

2.4.4 GWAS analysis……………………………………………………………….41

2.5 Results……………………………………………………………………………..41

2.6 Discussion……………………………………………………………………… 44

2.7 Tables and Figures……………………………………………………………... 47

Conclusion…………………………………………………………………………. 54

References……………………………………………………………………………..55

Appendix A: Sequencing Pipeline……………………………………………………63

Appendix B: Abbreviations…………………………………………………………... 63

List of Tables

Table 1.1: L1 regression results for hygienic behavior ……………………………………...…36

Table 1.2: Honeybee genes closest to the markers found by L1 regression for hygiene ……….37

Table 2.1: L1 regression results for overwintering behavior …………………………………...52

Table 2.2: Honeybee genes closest to the markers found by L1 regression for overwintering ...53

List of Figures

Figure 1.1: Number and location of sampled honeybee colonies ………….................................26
Figure 1.2: Freeze kill brood assay ………………………………………………………...26
Figure 1.3: Distribution of hygienic behavior ……………………………..............................27
Figure 1.4: Distribution of allele frequency ………………………………….……….......29
Figure 1.5: Location of single nucleotide polymorphisms across the honeybee genome ……...30
Figure 1.6: Correlation matrix between allele frequencies of hygienic behavior markers ……...31
Figure 1.7: Scatter plots of allele frequency and hygienic behavior ……………………………32
Figure 1.8: Manhattan plot of chromosome 2 for hygienic behavior …………………………...33
Figure 1.9: Manhattan plot of chromosome 5 for hygienic behavior …………………….........34
Figure 1.10: Genes near one hygienic behavior marker ……………………………………...35

Figure 2.1 Percentage of honeybee colonies lost during the winter in Canada ………...........47
Figure 2.2: Distribution of Overwintering behaviour …………………….............................47
Figure 2.3: Correlation matrix between allele frequencies of hygienic behavior markers ……...48
Figure 2.4: Distribution of correlations between the allele frequency of overwintering snps…...49
Figure 2.5: Relationship between regularization strength and coefficients ……………………..49
Figure 2.6: Distribution of allele frequency between alive and dead colonies ...…..……………50
Figure 2.7: Receiver operator curve showing the prediction of colony survival ………………..51
Figure 2.8: Fst between two populations of Drosophila ……………………………………...51

Introduction

Studying the honeybee is becoming increasingly important due to their immense value for humans. They contribute directly to our economy by producing honey and wax, which we use for food and medicinal purposes, with honey alone valued at $232 million in Canada in 2015 (AAFC, 2013). By far, though, their main contribution is pollination of crops, which promotes both food diversity and quantity (Klein et al. 2007). The honeybee is the most important managed pollinator, employed to pollinate several crops; fruits such as apples, peaches, plums, blueberries, and pumpkins; and seeds such as canola, sunflower, and soybeans (Morse et al. 2000). Honeybees contributed an estimated $177 million of a total $197 million in harvest value of apple production in Canada for 2013 (AAFC, 2013). For the same year, they also contributed to an estimated $168 million out of $177 million in blueberry production (AAFC, 2013). In total, their contribution to fruit and vegetable production for that year was estimated to be $562 million. Furthermore, the production of Canola, of which Canada is the world's top producer, relies heavily on honeybees, as 50% of the yield is attributed directly to their activity. An estimate of their contribution to Canola production in Canada ranges from $3.15 billion to $4.39 billion (AAFC, 2013). Based on these estimates, honeybees contribute between 3 and 5 billion dollars in additional crop value through their pollination services to the Canadian economy.

Clearly the honeybee is an indispensable organism to humans, but unfortunately, beekeepers are experiencing substantial colony loss, threatening all the services we value them for. In fact, cultivation of pollinator-dependent crops has increased relative to non-pollinator crops globally for the past 45 years. Thus, the demand for managed pollinators like the honeybee is on the rise, yet, their health is declining (Klein et al. 2007). During the winter of 2009-2010, Beekeepers experienced a 7-30% wintering colony loss across Europe and a 16-25% wintering colony loss in six Canadian provinces (Zee et al. 2012). Several causes have been proposed such as pesticide use, climate change, and pathogens, which are often found at high levels in dead colonies (vanEngelsdorp and Meixner, 2010, Tsvetkov et al. 2017). In Canada, harsh winters and disease lead to significant colony losses throughout the year.

Despite knowing the causes of the decline, we generally lack tools to improve bee health and circumvent colony losses. Previous attempts to address the issue do not solve the problem entirely. For example, to fight against pathogens and pests of their colonies, beekeepers use various chemical treatments. Though effective, this is inherently problematic. For one, it is only a short-term solution because disease and mites become resistant to the treatments. American Foul Brood Disease is a devastating honeybee pathogen that was treated with Terramycin for decades until the disease developed resistance to the drug (Murray et al. 2009). Furthermore, antibiotics, fungicides, and pesticides can all contaminate honey, beeswax, and potentially have side effects on the bees. Several international studies have reported above regulatory standards levels of antibiotics and pesticides in honey products (Al-Waili et al. 2012). Finally, bees should not require our periodic treatments for survival; it would be ideal to permanently select for traits such as natural disease resistance. This approach has the added benefit of saving beekeepers time and money as they would not have to tend to diseased colonies or spend money on chemical treatments. Several lines of honeybees have already been bred that are resistant to American Foulbrood and Chalkbrood, which are devastating honeybee diseases (Spivak and Reuter, 2001).

Therefore, breeding bees with desired traits is potentially advantageous because the honey and wax produced will be free from chemical contaminants, the bees will be healthy without constant upkeep from the beekeeper, and resistance to treatment by the pathogens will be limited.

Additionally, breeding healthier bees can be used to simultaneously select for other desired traits, such as winter survival, which would be very useful to Canadian beekeepers who experience significant colony loss during the winter. The previous breeding techniques were time consuming because measuring colony level traits in the honeybee is labour intensive (Oldroyd and Thompson, 2007). The process required manual inspections of the colony with multi-step measurements of the desired traits. The manual attention needed for each colony meant breeding programs were small therefore unlikely to have a significant impact on honeybee health. The biggest hurdle, which is yet to be crossed, for breeding healthier bees is the identification of well-known genomic markers for the various valuable traits that, when selected for, improve the trait. The most common genomic markers are single nucleotide polymorphisms (SNPs), which are a single base pair mutation within the genome. If a SNP is near a region that controls a trait, it may become associated with the trait. Therefore, these SNPs act as markers that can be selected for to vary the trait in the organism. Finding markers for selection of important honeybee traits such as disease resistance and winter hardiness would potentially allow improved breeding practices.

Previous attempts at understanding the genetics of colony traits have provided little knowledge about actual mutations or markers that can be used for selection. One technique used in the past to find markers for traits is called Quantitative Trait Loci analysis (QTL analysis). QTL analysis works by selecting for the trait of interest over many generations through inbreeding to create two homozygous lines, one with the desired trait, and the other with the undesired trait (Kent et al., 2018). After breeding the two lines to produce an F1, then backcrossing the F1 to produce an F2, the number of markers from the desired trait parent will differ among the progeny. These markers can thus be correlated to the trait and be used to narrow down areas of the genome that may control the trait (Broman, 2001). Although the idea is sound, there are drawbacks to QTL analysis. First, they cannot pinpoint markers precisely since the typical resolution of QTL analysis is low, pointing to a vast region composed of thousands of genes (Korte and Farlow, 2013). It is therefore difficult to select for traits from such a vast region without knowing the most informative region since recombination would soon separate most of the correlated markers from the causal mutation. Second, QTL analysis relies on the genetic diversity between two individuals, the original parents, therefore they likely lack all the variants that contribute to the phenotype (Borevitz and Nordborg, 2003). Selection for the trait without access to all the variants that control the phenotype will be subpar, since an individual with the right alleles across all the causal markers is likely to have a better phenotype, yet a QTL analysis would fail to identify the other markers.

Next-generation sequencing and Genome Wide Association (GWA) studies show great promise for understanding the genetics of economically valuable traits by improving the resolution of identified regions and using many more variants than a typical QTL study (Kent et al. 2018). GWA works by phenotyping many individuals for the trait of interest, genotyping them by sequencing their genomes, then correlating the genotypes with phenotypes (Korte and Farlow, 2013). In principle, GWA is like a QTL analysis; a heritable phenotype is partially caused by a genetic component, so the individuals that have a particular allele will express the phenotype, and for a complex trait controlled by many sites in the genome, individuals with more of the causal alleles will show even higher levels of the phenotype. Therefore, alleles that significantly correlate with the phenotype are likely causal or near causal regions of the genome. There are two major differences between GWA studies and QTL analysis. First, A GWA study uses the variation present within the studied population which is genetically more diverse than the parental individuals used in a QTL study. If the parental individuals in the QTL study do not have some of the causal alleles for the trait, those alleles will not be captured. Thus, the genetic diversity in a

GWA study improves the chances of detecting a causal site within the genome because more individuals and thus more markers are observed. Second, and most important, is QTL have a lower resolution than GWA studies. QTL studies increase the genetic linkage within the parental lines, hence, nearby markers become associated with each other. The lack of independence between the markers reduces the resolution of the study, as it prevents the causal marker from being identified. Whereas for a GWA study, a natural population is observed, which has had generations to decrease the genetic linkage between nearby SNPs through recombination. Therefore, a causal SNP within the genome will only be linked to very close SNPs. There are also other advantages to performing a GWA study in the honeybee. The honeybee genome is small in size which makes sequencing it significantly cheaper. Furthermore, the honeybee genome has a high recombination rate (Beye et al. 2006), further increasing the resolution such that only SNPs near a causal SNP will be associated with the phenotype.

Here I will take advantage of these features and the principles of GWA studies to identify genomic markers that will improve selective breeding of healthier bees, specifically for hygienic behavior (Chapter 1) and overwintering behavior (Chapter 2). Hygienic behaviour and overwintering behaviour are two important traits when it comes to honeybee health. Hygienic behaviour is the removal of sick and dead individuals from the hive as a means of disease prevention and management. Overwintering behaviour is the clustering and shivering to generate heat that allows the colony to survive in cold climates. Both traits are likely heritable, therefore, finding markers associated with the traits for selection will improve breeding practices and potentially produce resilient honeybees.

Chapter 1: Hygienic Behavior

1.1 Introduction

The honeybee is known for its sizeable social structure, typically composed of 20000-100000 bees, and although there is strength in numbers, the sheer volume of individuals allows diseases to spread quickly within the colony. In addition to the individual density, the constant interactions like the workers having to regurgitate food for the larva, each other, and the queen, further exacerbates this effect. To combat the spread of disease, the honeybee has evolved multiple lines of defence (Evans et al. 2006)**.** The first is called the innate immune system which involves the molecular pathways that target pathogens, and the other is a series of behavioral traits collectively called social immunity which involves actively cleaning the hive, removing sick or dead bees, defecating away from the colony, grooming each other and social fever to heat kill bacteria (Evans et al. 2006; Cremer et al. 2007).

Despite the demonstrable need to fight disease, the honeybee has a reduced set of complement innate immunity genes relative to other insects, emphasizing their reliance on social immunity (Evans et al. 2006). The impact of social immunity in combating disease through the honeybee's life history may have been enough to cause relaxed selection on innate immunity genes (Harpur and Zayed, 2013). The likely reduced effectiveness of the honeybee innate immunity system is worrisome since the death of honeybee colonies has been attributed to disease. Though many factors, such as pesticides and climate change are likely at play, pathogens are often found at high levels in dead colonies (vanEngelsdorp and Meixner, 2010) and can directly cause colony loss (Higes et al. 2008). The multiple stressors can further impair their immune system; exposure to the pesticide clothianidin leads to rapid replication of deformed wing virus, one of the honeybees most devastating disease (di Prisco et al. 2013). Therefore, the ability to deal with disease is becoming more critical, and bees with better social immunity are of immense value. Hygienic behaviour is a subset of social immunity that refers to the removal or social-exclusion of dead and infected bees to prevent the spread of disease. Hygienic behaviour is very important to having healthy disease resistant bees, and so understanding its genetic basis could lead to improved breeding of the trait.

The genetics of hygienic behaviour has been previously studied in the honeybee but the identified regions are too wide to be helpful. There is considerable variability in this behaviour, for example, some colonies remove American Foulbrood killed larvae and pupae quickly, while others allow the dead individuals to remain for weeks or longer (Rothenbuhler and Thompson 1955). Hygienic behaviour is also heritable, with a narrow sense heritability estimated to be between 0.20-0.65, which is the proportion of variance in the trait that can be attributed to additive genetic factors (Koffler et al. 2016; Harbo and Harris, 1999). Hygienic bees uncap the brood chamber and remove infected pupae from the colony. Rothenbuhler (1964) developed a two-locus hypothesis for hygienic behaviour, one locus that controls the uncapping and another locus that controls the removal of the infected individual. To do so, he crossed highly hygienic bees with non-hygienic bees, then backcrossed the F1 generation with the highly hygienic parental line. He found four possible groups from the offspring of this backcross; uncappers and removers, uncappers only, removers only, and neither uncappers nor removers. Moritz et al. 2000 first challenged the proposed two-locus hypothesis, by suggesting the data is better explained by three loci. Lapidge et al. (2002) did a follow-up QTL study to test the model but found seven loci that segregated with the hygienic behaviour phenotype each explaining 9-15% of the phenotypic variance. Oxley et al. (2010) on the other hand genotyped bees segregating in the behaviour within a colony and found six loci that together account for 30% of the phenotypic variance. From the studies, it is unclear how many loci contribute to hygienic behaviour. Since modern studies suggest multiple loci, hygienic behaviour is likely a complex trait with multiple loci contributing to its effect. Although multiple loci were detected by the QTL studies, the loci are large portions of the genome, leading to ambiguity in the actual genes that may be involved. The low resolution of QTL studies makes it difficult to use the information, since without pinpointing the causal variant recombination would cause disassociation from the marker within a few generations. As mentioned earlier, with the power and promise of genome-wide association studies, the target area can be extremely narrowed down, making marker-assisted selection for hygienic behaviour much more reliable.

My aim is to find genetic markers that associate with hygienic behavior by conducting a Genome-Wide Association study on 925 honeybee colonies across Canada. The colonies will be measured for the hygienic behaviour phenotype and have the genome of a sample of workers from the colony sequenced. I will then associate the genetic information to the phenotype to identify mutations that may explain hygienic behaviour. Since hygienic behaviour is heritable, I will find one or more Single Nucleotide Polymorphism (SNP), which is a single base pair change in the genome, that associates with hygienic behaviour. Therefore, for each SNP I will test the hypothesis that the SNP is associated with hygienic behaviour.

1.4 Methods

1.4.1 Colony and Phenotype

925 honeybee colonies were sampled across Canada; from each of Alberta, Ontario, Manitoba, Quebec, and British Columbia. The approximate location and the number of colonies

at each location is shown in Figure 1. From these colonies, hygienic behaviour and overwintering behaviour were measured, and a sample of 50 workers from each colony was obtained for sequencing. The collected bees were frozen immediately on dry ice, shipped in 50ml centrifuge tubes, then stored at - 80C in the laboratory until processed.

Hygienic behaviour was measured with a freeze-kill assay which involves freezing a small portion of the brood cells with liquid nitrogen to induce death (Spivak and Downey, 1998). The brood cells are placed into the colony, and after a 24-hour period, the proportion of uncapped and cleared cells is measured. Figure 2 shows the brood cells from a colony during one assay, first before the 24-hour period (Figure 2A), then after the 24-hour period (Figure 2B). Each capped cell contains a pupa that would develop into an adult bee, however after the freeze-kill procedure, the nurse bees find the cells, uncap the cells, and remove the dead pupae from the hive – the result is what we call hygienic behaviour. Hygienic behaviour should be correlated with the proportion of cleared cells; therefore, by measuring the proportion of cleared cells, this assay can be used to measure the propensity of hygiene for colonies.

1.4.2 Extraction and Sequencing

Since we are interested in the genetic factors that may affect how hygienic a colony is, we sequenced the DNA of workers within a colony. To do so, the front left legs of fifty bees from each colony were pooled together and crushed for DNA extraction. If the front left leg was unavailable, substitute legs were used in the following order - right foreleg, then left or right midleg. The hind legs were avoided due to possible cross-contamination with pollen as foragers store pollen on their hind legs. The act of pooling bees then sequencing, tells us, for each base pair in the genome, what proportion of bees have a particular allele. For example, if, at the first base pair of the honeybee genome, 45 of the 50 bees had the reference allele A and 5 of the 50 bees had the alternate allele T, the measured allele frequency at that colony would be 5/50 = 0.10 for the allele T. Thus, pooled sequencing allows us to determine the allele frequency for each colony at each site of the genome. The reason that as few as 5 of the 50 workers have the allele T is due to the reproduction system of the honeybee. The queen of each colony mates with around 20 males and stores their sperm in a specialized organ for the rest of her life (Winston, 1987). To make a worker she fertilizes an egg with the stored sperm, therefore if only a few of the drones have the allele T only a few of the workers will as well, as only a small proportion of the stored sperm will have the allele T. This mating system adds diversity to the pool of alleles available at each locus, therefore, multiple alleles are likely present in the workers at the segregating loci. By sequencing 50 bees, we are ensuring that for each locus we have a large enough sample to catch all the possible variants from the multiple fathers and get an accurate representation of the frequency of alleles segregating within the colony.

To extract the DNA, the tubes containing bee legs were dipped in liquid nitrogen to freeze the tissues, then finely ground using a pestle. Proper grinding ensured that the maximum amount of pooled bee tissue was obtained for the following steps. For tissue lysis, we added 350µl of Tissue Lysis Buffer and 20µl of 20mg/ml of Proteinase K, and incubated samples overnight at 50°C. DNA extraction was performed using Mag-Bind® Blood &amp; Tissue DNA HDQ 96 Kit (Omega Bio-tek Inc., USA) optimized for KingFisher™ Flex Purification System (Thermo Fisher Scientific Inc., USA). A final eluent volume of approximately 75µl was obtained for each set of 50 bee legs. DNA was quantified using NanoDrop™ 2000 Spectrophotometer (Thermo Fisher Scientific Inc., USA). DNA quality was assessed with 0.8% agarose gel electrophoresis.

After extraction, the pooled DNA was sent for sequencing to Genome Quebec Innovation Centre's Next Generation Sequencing team at McGill University. Sequencing was done to ~90x coverage on the Illumina HiSeqX 150bp paired-end sequencing platform.

1.4.3 Alignment and SNP calling

After sequencing, the reads were passed through our bioinformatics pipeline to determine the most accurate genotypes for the colonies (Appendix A). The first step, Trimmomatic, uses the metadata from reads to either clip reads appropriately or remove untrustworthy reads altogether (Bolger et al. 2014). Trimmomatic identifies and removes Illumina specific adaptors from the reads to prevent misinterpreting these regions as SNPs, as they are not sequences from the original sample. Trimmomatic also clips the leading and trailing ends of reads by performing a quality check; if the first and last basepairs of the read are below a quality of 20 they are removed, and this process is repeated until the new first and last basepairs have a quality above 20. Trimmomatic then traverses the read with a sliding window of 20 basepairs, and if the average quality of the basepairs within that window falls below 25, the read is clipped further from the start of the window. Finally, after the clippings have occurred, trimmomatic drops all reads that are below a length of 50bp. The entire trimmomatic process then ensures higher quality reads with sufficient length to be used for SNP calling.

The second step is to align the reads to the reference genome using the NextGenMap (NGM) software (Sedlazeck et al. 2013). The reference genome used is from the honeybee sequencing consortium, version 4.5. The purpose of this step is to determine which part of the genome each read belongs to. Doing so would allow the SNP caller to identify SNPs by recording differences between the reference genome and the read that mapped there. After the reads are mapped, we sort and index them using samtools then remove duplicate reads using Picard tools as they are likely to be PCR artifacts.

The third step uses Base Quality Score Recalibration (BQSR) from the Genome Analysis Toolkit (GATK) to systematically improve the per-base quality scores. The per-base quality scores are provided by the sequencer, showing the confidence the sequencer had in calling the base (Auwera et al. 2013). This quality score could be an over or underestimate due to the physics or the chemistry of how the sequencing reaction works or due to manufacturing flaws in the equipment. BQSR builds a model using the reported quality score, the position of the base in the read, and the preceding basepairs within the read then adjusts the per-base quality scores for each base in a read.

The fourth step is the SNP caller Lofreq, which uses the now filtered and mapped reads, to compresses the information into locations where the sequence differs from the reference (Wilm et al. 2012). The Lofreq algorithm determines which differences between the read and the genome to label as a true SNP using various criteria including depth. Lofreq outputs a Variant Call File (VCF) that shows the positions of the genome that differ from the reference along with per SNP quality and read depth values.

The following steps apply multiple filters to the vcf file, removing untrustworthy SNPs. The first filter removes any SNP that is within five basepairs of known honeybee indel sites since aligners have a difficult time near indels. To do so, we used data from previously sequenced honeybees and 24 samples from our 925 colonies to identify honey bee indel sites. The next filters use quality, depth, and strand bias information to remove SNPs. The Upper limits for all three criteria are determined for each sample by looking for outlier values. In a histogram showing the distribution of the criterion, an outlier causes a large separation between the previous value and the outlier value. Therefore, finding areas of large changes in the distribution can be used to identify when an outlier is first observed and any values as extreme or more extreme than the outlier are discarded. Therefore, after identifying which values of each criterion are outliers, any SNP with that value or more for that criterion is considered an outlier and subsequently filtered. A lower limit for depth was not used since the lofreq algorithm already determines when a low depth SNP is a false SNP and removes them, but a lower limit of 100 for quality was used such that any SNP with quality less than 100 was filtered from the vcf.

The final filter uses known ambiguous sites of the genome to remove nearby SNPs. These areas are determined by sequencing a haploid genome like that of honeybee males. The haploid nature of the genome means that there should never exist a heterozygous site. However, due to the presence of repetitive sequences in the genome, reads belonging to similar sequences may map where they do not belong. This may erroneously lead to disagreement on which base belongs to that position between the reads that belong there and the reads that do not, creating false heterozygous calls. To this end, we sequenced ten drones using Illumina HiSeqX and combined this dataset with 40 previously sequenced drones (Harpur et al. 2014) which produced a comprehensive set of honeybee heterozygous sites. This process identifies sites of the honeybee genome that are ambiguous, and since these ambiguous sites are likely to cause errors in SNP calling of diploid workers, we remove any SNPs that are within five base pairs of these sites.

Although these techniques will filter out some true SNPs, by and large, they will target false SNPs, increasing our confidence in the data. Overall these filtration steps removed a total of 2 million SNPs across all samples, resulting in 5 million SNPs among all 925 colony samples. We further validated the removal of these SNPs by annotating them with SnpEff, which labels SNPs by how they affect genes. SnpEff can label a SNP that is within an exon as nonsynonymous, meaning the protein is altered, or synonymous, meaning the protein is unaltered. If these steps removed erroneous SNPs, the removed SNPs should have a higher nonsynonymous to synonymous ratio than the filtered SNPs since true SNPs are less likely to be nonsynonymous. The nonsynonymous to synonymous ratio is 0.94 for the removed SNPs compared to 0.60 for the SNPs that were kept, therefore the removed SNPs are likely false calls, and we are confident in their exclusion.

1.4.4 Genome-Wide Association (GWA) analysis

With the genotypes and phenotypes measured, we can check for associations that occur between them since the most significantly associated SNPs could be causal or near causal mutations. GWAS is typically performed using the single marker regression approach whereby a single SNP is tested in a linear model for an effect on the phenotype (Cantor et al. 2010). This approach has two disadvantages. The first is the large number of hypothesis testing that is conducted, once for every variant. In each case we usually would have a 5% chance of rejecting the null hypothesis, therefore conducting many tests leads to many false positives. To correct for this, typically a multiple testing correction procedure is used, however, this often leads to many false negatives (Waldmann et al., 2013). A standard approach is a Benjamini-Hochberg-Yekutieli technique, which involves ranking your p-values from most significant to least, then using the rank and a predetermined false discovery rate to determine which p-values will remain significant (Benjamini and Yekutieli, 2001). The second disadvantage of single marker regression is the lack of the other SNPs in the model may overestimate the effect of the SNP, since the other SNPs could explain the phenotype (Qianchuan and Dan-yu, 2011). For example, a SNP that is not related to the trait but is correlated with a SNP that is will be associated with the trait further leading to false positives.

These statistical challenges are due to the large number of features relative to the small sample size, that is, the large number of SNPs to check compared to the small number of colonies sampled. To overcome these challenges several techniques have been developed, such as the L1 regression (Tibshirani, 1996). L1 regression can accept many SNPs in one model, thus, multiple testing correction is not required. Furthermore, when a SNP explains some of the variation in the phenotype, the other SNPs have less variation to explain. L1 regression allows multiple SNPs and does feature selection using a regularization parameter that controls the strength and number of non-zero coefficients in the model. Therefore, if a SNP is not useful in modelling the phenotype, its coefficient will be set to zero. The net result of a model that accounts for multiple SNPs is weakened false positive signals and an increased chance of causal SNPs being selected (Qianchuan and Dan-yu, 2011). L1 Regression works by adjusting the effect size of each SNP to minimize the cost function:

$$C(\beta) = \frac{1}{2}\sum_{i=1}^{N}(y_i - \sum_j \beta_j x_{ij})^2 + \lambda \sum_j ||\beta_j||$$

Where β_j is the effect size of SNP j, y_i is the true hygiene level of sample i, $\sum_j \beta_j x_{ij}$ is the model's prediction of the hygiene level of sample i and $\lambda \sum_j ||\beta_j||$ is the regularization term of L1 regression. The first term in this summation is just the squared error of the models' prediction, a model with a low error generally performs better. The second term in the summation adds a cost to the model by summing up every effect size the model chooses for each SNP. The more non-zero effect sizes and the larger the effect sizes the model uses, the more it increases the cost of the model. L1 regression aims to find the lowest squared error by using the least number of SNPs. The λ term before the regularization term controls the regularization strength, the higher the λ value is, the more the model is penalized for using additional SNPs. The regularization strength λ was chosen with care to account for false positives. If lambda is too strong all SNPs will be given non-zero coefficients, if it's too weak many SNPs, including false positives will be given non-zero coefficients. Here, we chose λ by shuffling the phenotype values, thus break any true association between the genetics and the phenotype. Therefore, if L1 regresson is run with shuffled phenotype values, any SNP with a non-zero coefficient is a false positive. This allows us to calculate the false positive rate for every λ value by dividng the number of false positives by the number of SNPs tested. For every tested λ we ran this false positive test then chose a lambda value where the false positive rate was less than 1 / 5000. Therefore, one out of 5000 of the nonzero coefficient SNPs will be a false positive, which is very conservative. Despite doing this, L1 regression does not directly provide p-values or true effect sizes of the SNPs due to regularization. To estimate the effect sizes and show the associated p-value single marker regression was done for each SNP.

Before performing L1 regression on hygienic behaviour we first controlled for the environmental effect on the phenotypic data. Variation in hygienic behaviour can occur due to both genetics and the environment. The colonies used in this study were organized into yards, so the colonies within a yard share the same environment. To account for the environment effect, hygienic behaviour was standardized within a yard, by calculating a z-score of hygiene within each yard, since the differences between colonies within a yard are more likely to be due to genetics than environment as they share similar environments. Before performing L1 regression, we also dealt with missing allele frequency values by mean imputation then normalizing for each SNP.

L1 regression was done using the glmnet package (Friedman, 2010). The strength of regularization in L1 can be controlled; zero regularization is equal to a normal ordinary least squares regression, while a high regularization strength would set most coefficients to zero. The regularization parameter was determined using cross-validation as described by Yi et al. 2014. The final L1 regression model was trained using the regularization parameter this method found.

1.5 Results

The freeze-kill brood assay was performed on our colonies to determine the proportion of cleaned out dead brood after 24 hours. There is considerable variation in hygiene among our colonies with quite a few colonies showing high hygiene ability (Figure 3A). The distribution is skewed to the left, with about 200 colonies clearing close to 100% of the dead brood. The goal of the project is to explore what in the genetics of the bees causes this variation essentially asking are there genetic differences between colonies that are highly hygienic and colonies that are not as hygienic.

Hygiene levels also varies between yards, with some yards having most colonies with high hygiene levels and others having most colonies with lower hygiene levels (Figure 3B). To account for the effect of the yard on the phenotype, we standardized the hygiene levels of colonies within the yards they were in. Performing the standardization within the yard makes the distribution of hygiene levels between yards centred, removing the effect of the yard on the phenotype (Figure 3C). Although some yards still have more variance than others, the distributions are more centred when comparing across different yards which will prevent the model from just choosing SNPs that have different allele frequencies between the yards. The transformation also makes the phenotype closer to a normal distribution, which better fits the assumptions of a linear model (Figure 3D).

Next-generation sequencing was performed on 50 bees from each of the colonies. The result is a measure of allele frequencies of SNP's across the genome for each colony. The allele frequency distribution is shown in Figure 4. There are very few SNPs between the range of 0 and 0.1 due to a combination of filtration and sampling. The filtration steps tend to remove low allele frequency SNPs due to their lower depth. A low depth is the result of only a few reads having the alternate allele, making it hard to distinguish between a true SNP and a sequencing error since sequencing errors also produce a small number of reads with a different allele from the reference. Sampling also prevents the measurement of low allele frequency SNPs because when collecting 50 bees from a colony to sequence, a bee with a rare allele (since the allele is at a low frequency) is less likely to be selected for extraction. As a result, the bee with the rare allele will not be in the pool that is sequenced hence the allele will not be measured. The lack of allele frequencies in this region is not too much of a concern since Genome-Wide Association studies are best used for identified common variants as opposed to rare ones (Korte and Farlow, 2013). The distribution of the SNPs found on the chromosomes is shown in figure 5. The SNPs are well spread across the 16 chromosomes of the honeybee. SNPs are found nearly on all parts of the chromosome, though there are some areas with low SNP density. The lowest density regions are seen on chromosome 8 and 11 though these regions are only a small part of the chromosome.

Feature selection by L1 Regression on hygienic behaviour resulted in 27 SNPs (Table 1). Single marker regression was performed to estimate the unregularized coefficient and p-value of the SNPs, although the chances of any of the SNP being a false positive is 1 in 5000. The locations of the SNPs are evenly spread throughout the genome with most chromosomes having 2 SNPs, and a few having 1 or 3 SNPs. Most of the SNPs are reasonably distanced from each other, however, the three SNPs on chromosome 5 appear to be very close to each other. The correlation between allele frequencies of the SNPs found by L1 is shown in figure 6. We see weak correlations between SNPs on different chromosomes. This makes sense because any observed correlation is due to chance since chromosomes segregate independently during meiosis. We also see weak correlations between SNPs within the same chromosome but separated by a large distance. After a separation of 20000 basepairs between two SNPs on the same chromosome, we get correlations lower than 0.25 showing the resolution GWA studies have since the different markers behave nearly independently due to recombination. Correlations of physically proximate SNPs have the highest correlations since they are less likely to be separated by a recombination event. The largest correlation is observed between SNP 5.9-457762 and 5.9-461612 at 0.63, which are about 4000 basepairs apart. It is interesting that there lies a SNP between them that has a weaker correlation with either one of the two, an indication that some of the correlation observed between these two SNPs is due to chance.

We can see a scatter plot between allele frequency and hygiene levels for some of the SNPs found by L1 regression in figure 7. The line drawn is determined by ordinary least squares with bootstrapping used to determine the confidence intervals. The entire width of the line shows the 95% confidence interval for the slope of the line. The plots again show the small effect of the allele frequencies of these SNPs on hygiene level. However, these scatter plots are just independent effects of the SNP on the phenotype; the true model would be a multidimensional plot that cannot be visualized.

The top panel of Figure 8 shows a single marker regression performed using all SNPs on chromosome 2. The y-axis represents the negative log p-value of the association between a SNP and hygiene, and the x-axis is the position along chromosome 2. The higher the point on the y-axis the higher its association with hygiene, and the position on the x-axis localizes the SNP to show where on the chromosome the strongest associations are. The bottom panel of the figure shows the L1 regression coefficients using SNPs found on chromosome 2. The y-axis shows the regularized coefficient for that SNP, and the x-axis again shows where on the chromosome the SNP is located. The non-zero coefficient chosen by L1 regression near the middle of chromosome 2 is located where the single marker regression found the largest associations between hygiene and the SNPs there. A similar figure is shown for chromosome 5 in Figure 9. Here we can see four SNPs selected by L1 regression to have non-zero coefficients. These SNPs also match with peaks of high association sites from single marker regression. The width of the peaks observed in the single marker regression plots are small, encompassing less than 1 million base pairs. The narrow peaks show the advantage of GWA studies over QTL studies. Essentially, recombination over time reduces the linkage between regions of the same chromosome, improving our ability to pin point causal regions.

A big advantage to having narrower associated regions is the smaller number of genes within that region that need to be explored as potential candidates for controlling hygienic behavior. The SNP found using L1 regression on chromosome 2 is in the middle of a gene cluster (Figure 10). The black dotted line is the position of the SNP, and each coloured rectangle is the position of a nearby gene, with the x-axis representing the relative position of the genes to the SNP. The 10 genes displayed are all odorant receptor genes and are used for identifying smells by the honeybee. The nearest honeybee genes to the SNPs found by L1 regression are shown in Table 2 along with their *Drosophila* orthologues. Most of the genes have orthologues that are also involved with odor detection or memory, which may be an important aspect of detecting sick individuals (see discussion).

1.6 Discussion

The genetics of hygienic behaviour remains contentious due to the lack of resolution in the methods used by previous studies. The low resolution prevents identification of markers that associate with causal loci to be used for marker-assisted selection. The need for marker-assisted selection for hygienic behaviour is important because hygiene improves the survivability of honeybee colonies. The nurse bees use hygienic behaviour to detect and evict sick or dead bees from the hive, as a preventative measure against disease. Current measures of dealing with disease, such as chemical treatment, are inadequate for long-term honeybee health. Here, I use modern genetic techniques to search for genetic markers of hygiene that can potentially be used for breeding hygienic bees. I also explore potential gene candidates that require validating in future research.

The genetics of hygienic behaviour was determined by performing a genome-wide association study. Colonies were measured for hygienic behavior using the freeze-kill assay, then 50 workers from each were sequenced in a pool to determine the genetics of the colony. L1 regression was used to select the SNPs that strongly affect the trait. The major finding of this study are the 27 SNPs selected by L1 regression to be associated with hygienic behavior. The effect sizes of each SNP are small, suggesting tiny influences from many loci affect overall hygienic behavior in a honeybee colony.

Previous studies have looked for the number of sites or the locations within the honeybee genome that are associated with hygiene. These studies used techniques of low resolution and low genetic variability, which likely underestimated the number of involved SNPs. For example, if two independent loci cause hygienic behaviour, the linkage produced by a QTL study may link the two loci. The results would then show only one region of the genome as associated with hygiene. The first studies ever done simply aimed to determine the number of regions but not their location. These experiments suggested two or three loci control hygienic behaviour (Rothenbuhler, 1964; Moritz et al. 2000). Later studies attempted to find the number of loci and their location. Lapidge et al. 2002 found one significant and six suggestive loci associated with hygienic behaviour. Unfortunately, at the time this study was performed the honeybee genome was yet to be sequenced, so the loci found can only be narrowed down to chromosomes; 2, 4, 5, 6, 13, and 16. Oxley et al. 2010 performed a follow up QTL analysis on hygienic behaviour after the honeybee genome was published. They found 6 significant loci on chromosomes 2, 5, 9, 10, 16. As predicted, we have identified more sites than previous studies due to the low resolution and genetic variability of those studies. Here we found 27 SNPs of the honeybee genome that associate with hygienic behaviour. Many of these SNPs are within the QTL regions identified in previous studies, but because QTL regions are so large it is hard to say whether the results are replicated. The QTLs found on chromosome 16 in previous studies, however, were not replicated in this study. All of The SNPs for hygienic behaviour on chromosome 2, chromosome 5, and some of the SNPs on chromosome 9 are encompassed by three of the six QTL regions found significant by Oxley et al. (2010).

The SNP on chromosome 2 is located within a cluster of odorant receptor genes (Figure 10). Odorant receptors translate odours experienced by the bee into neuronal information (Galizia and Menzel, 2000). Many of the *Drosophila* orthologues of the genes near the SNPs found also have odor related functions. For example, the orthologue for the gene GB46114 is *Frl* and is involved in axon extension and mushroom body development (Dollar et al. 2016). The mushroom body is a center for olfactory learning (Heisenberg et al. 2003) and olfactory learning is inhibited in mushroom body mutants in *Drosophila* (Heisenberg et al. 1984). Furthermore, thermal stress in *Drosophila* larvae and pupae disrupts mushroom body development while having minimal effect on other brain structures, and as a result impairs odor learning in the adult flies without affecting sensory perception (Wang et al. 2007). Similarly, other orthologues found are involved in olfactory related neuronal regions. The *Drosophila* orthologue for the gene GB1005 is the N-alpha-acetyltransferase 20A (*Naa20A*) and is part of the N-acetyltransferase complex B (NatB). The NatB complex has been shown both in vivo and in vitro to affect odorant receptor neuron formation (Stephan et al. 2012). Psidin is a subunit of the NatB complex that prevents developing olfactory neurons from undergoing apoptosis. As a result, flies with a null mutant of Psidin have a reduced number of axons of certain classes of odorant receptor neurons (Stephan et al. 2012). The orthologue ara functions in the development of the eye-antennal imaginal disc in *Drosophila* (Cavodeassi, 2000), which is important since odorant receptor neurons develop from the eye-antennal disc (Rodrigues and Humme, 2008). Furthermore, the orthologue chn is a transcription factor that regulates the achaete/scute complex which is necessary for the formation of external sensory organs (Escudero et al. 2005). Under-expression of chn results in fewer sensory organs whereas over-expression results in more sensory organs. The use of these genes to process odor makes the next example striking. The *Drosophila* orthologues for the genes GB55285, GB45092, and GB53422 were associated with an odorant response to the odorants 2-Phenylethanol, ethyl acetate, and hexanal respectively (Arya et al. 2015). It turns out, 2-phenylethanol and phenethyl acetate are present in volatile collections from chalkbrood infected larvae but are absent from volatile collections from healthy larvae in the honeybee (Swanson et al. 2009). The remaining orthologues either have unknown function or are involved in memory formation. Sba like ara is expressed in the eye/antennal disc, and perhaps is also involved in odorant receptor neuron formation (Zeidler and Mlodzik, 1997). 2mit is expressed in mushroom bodies of the adult fly and is involved in short term memory affecting behavioural plasticity (Baggio et al. 2013). Nep2 and CG7231 were tested for their effect on metabolism and toxicity of beta amyloid peptides, which are found in plaques that form in individuals with Alzheimer's disease. While Nep2 lowered the total beta amyloid protein by 70%, CG7231 had no effect (Cao et al. 2008). Knockdown of dhit improves olfactory memory in the fly and Orc3 mutants have reduced ability for learning and remembering odors in the fly (Walkinshaw et al. 2015, Boynton and Tully, 1992). The orthologue timeout is involved in circadian rhythm (Benna et al. 2000), which has been shown to affect the olfactory response of antennal chemosensory cells in *Drosophila* (Krishnan et al, 1999).

The importance of finding odor related genes could mean hygienic bees are better at detecting sick individuals or dead individuals through odor. This is in line with the hypothesis that all honeybees can perform hygienic behavior, but the difference between hygienic bees and non-hygienic bees is the propensity of detection to initiate the behavior in the first place (Wilson-Rich et al. 2009). Honeybees bred for hygienic behavior better discriminate between odors of healthy versus diseased brood, even though both non-hygienic and hygienic bees discriminated odors from two different flowers equally well (Masterman et al. 2000). Furthermore, the more hygienic a colony the more they detected and removed dummy larvae that were treated with 2-phenylalcohol and phenethyl acetate, which are volatile compounds found in chalkbrood infected larvae of which we found orthologues that associated with response to the odor in *Drosophila* (Swanson et al. 2009).

Using just the 27 SNPs we have found here, we can potentially breed for honeybee colonies that are hygienic. However, further validation is required of these SNPs. The *Drosophila* orthologues give some indication as to the possible mechanisms involved but they are just hypotheses that need to be formally tested. Many follow up experiments can be conducted, such as a common garden experiment where we know the genotype at some of these SNPs of the colonies within the yard and can make predictions about their hygiene levels. Molecular work will also help identify the pathways that these SNPs are involved in to manipulate the behaviour of these bees. The work done here can also be used to predict the state of colonies within a beekeeper's yard, so they can make smarter decisions on how to treat their bees.

1.7 Tables and Figures

Fig. 1. The number of honeybee colonies used in this study and where they are located across Canada.

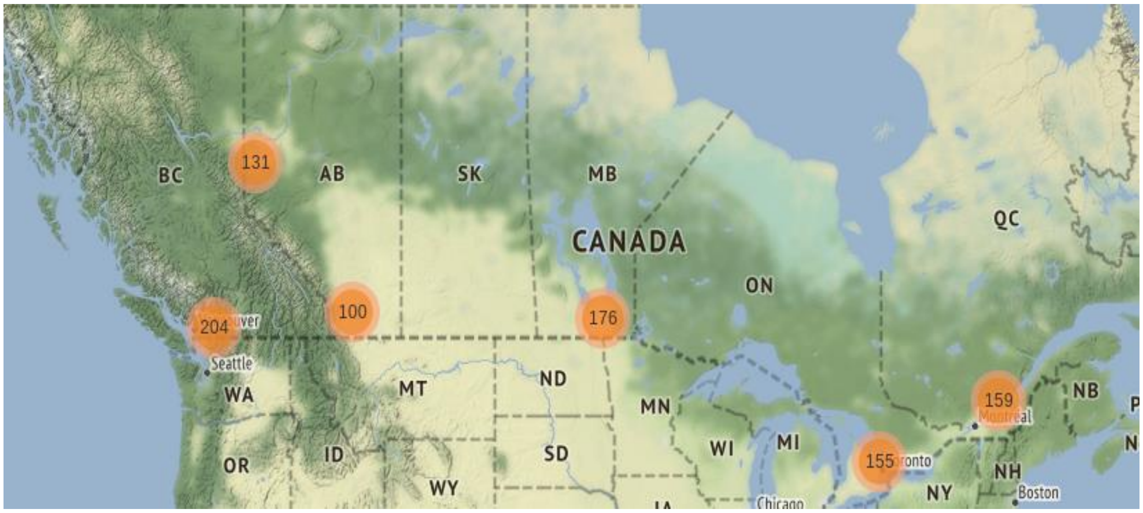


Fig. 2. Freeze kill brood assay measurement of hygienic behaviour for a single colony. A) A small section of capped brood is selected from the colony then killed using liquid nitrogen. B) The same brood section 24 hours after freezing, the proportion of cleared out larvae is used as the measurement for hygienic behaviour.

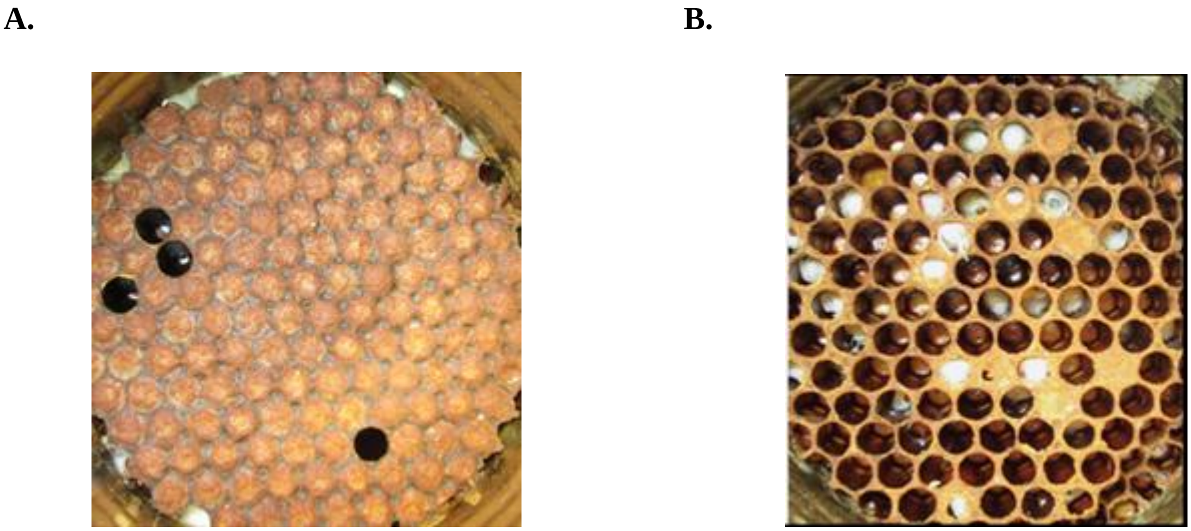
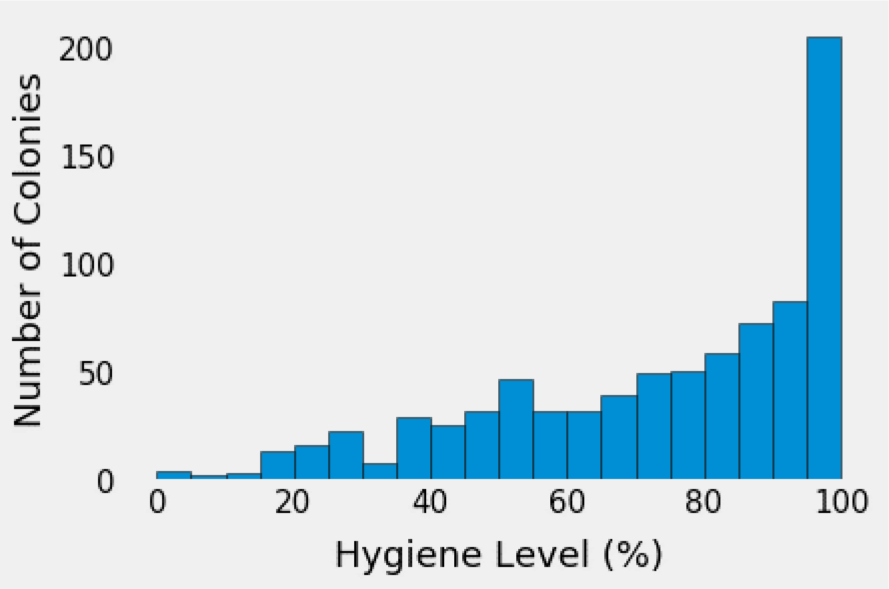
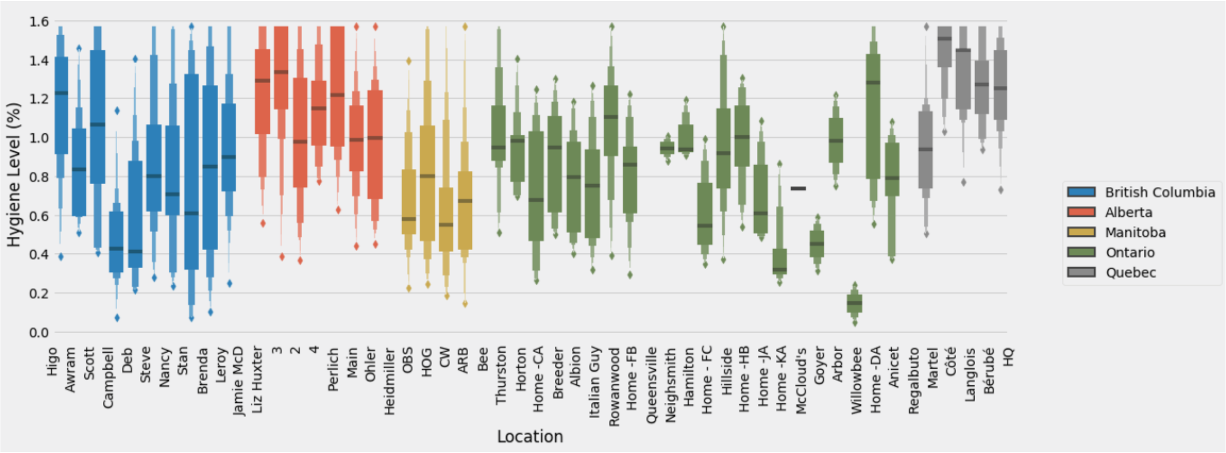


Fig. 3. The distribution of hygienic behavior for the honeybee colonies used in this study. Hygienic behavior is measured as the proportion of dead brood cleared away after 24 hours. A) The distribution of hygiene before transformation for our 925 colonies. B) The distribution of hygiene before transformation for our 925 colonies separated by yard. C) The distribution of hygiene levels after the colonies were standardized within yards. D) The distribution of hygiene levels after transformation for our 925 colonies.

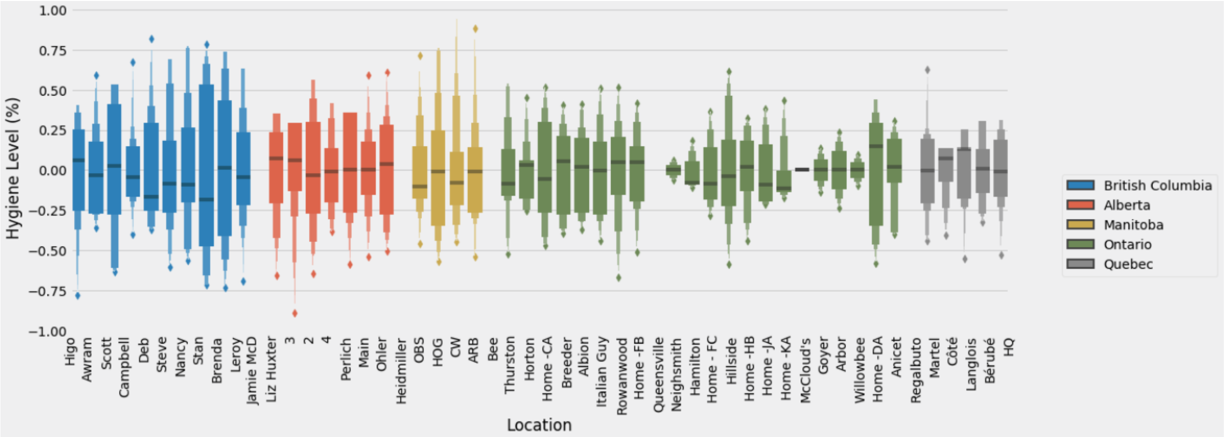
A)



B)


Hygiene Level (%)
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Higo
Awram
Scott
Campbell
Deb
Steve
Nancy
Stan
Brenda
Leroy
Jamie McD
Liz Huxter
3
2
4
Perlich
Main
Ohler
Heidmiller
OBS
HOG
CW
ARB
Bee
Thurston
Horton
Home -CA
Breeder
Albion
Italian Guy
Rowanwood
Home -FB
Queensville
Neighsmith
Hamilton
Home - FC
Hillside
Home -HB
Home -JA
Home -KA
McCloud's
Goyer
Arbor
Willowbee
Home -DA
Anicet
Regalbuto
Martel
Côté
Langlois
Bérubé
HQ
Location
British Columbia
Alberta
Manitoba
Ontario
Quebec

C)


Hygiene Level (%)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Higo
Awram
Scott
Campbell
Deb
Steve
Nancy
Stan
Brenda
Leroy
Jamie McD
Liz Huxter
3
2
4
Perlich
Main
Ohler
Heidmiller
OBS
HOG
CW
ARB
Bee
Thurston
Horton
Home -CA
Breeder
Albion
Italian Guy
Rowanwood
Home -FB
Queensville
Neighsmith
Hamilton
Home - FC
Hillside
Home -HB
Home -JA
Home -KA
McCloud's
Goyer
Arbor
Willowbee
Home -DA
Anicet
Regalbuto
Martel
Côté
Langlois
Bérubé
HQ
Location
British Columbia
Alberta
Manitoba
Ontario
Quebec

D)

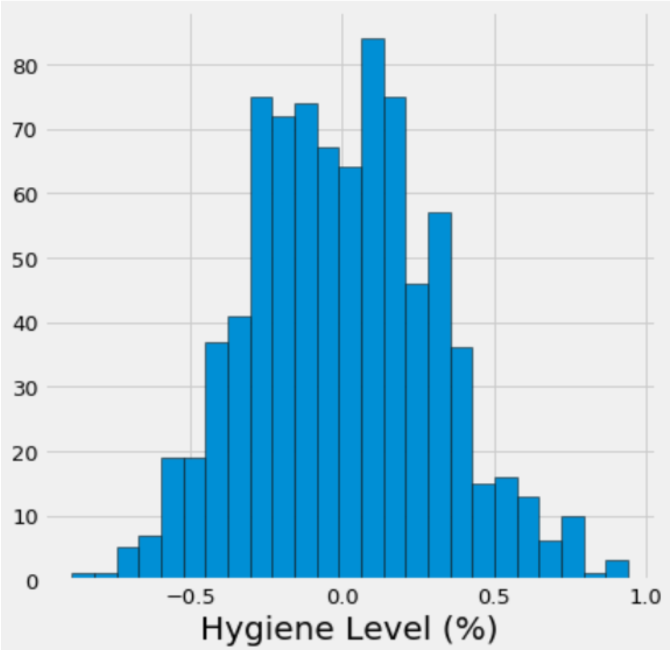


Figure 4: The distribution of allele frequencies measured at all SNPs of the genome of all 925 honeybee colonies.

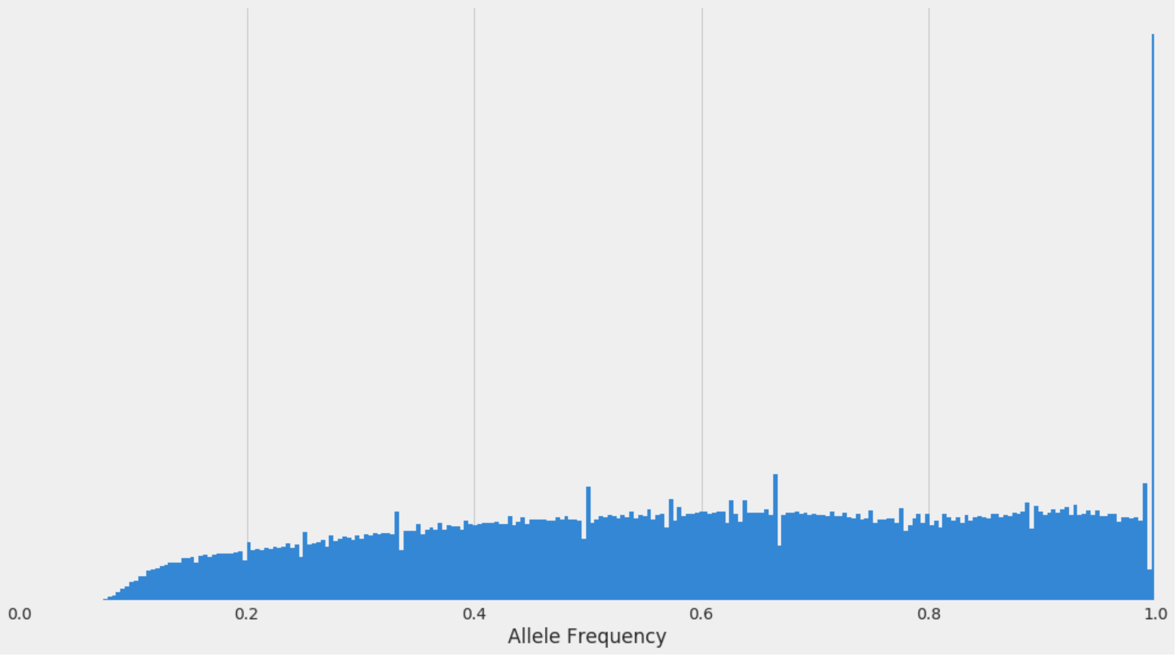


Fig. 5. The spread and location on each honey bee chromosome of the Single Nucleotide Polymorphisms found in all the 925 colonies of this study,

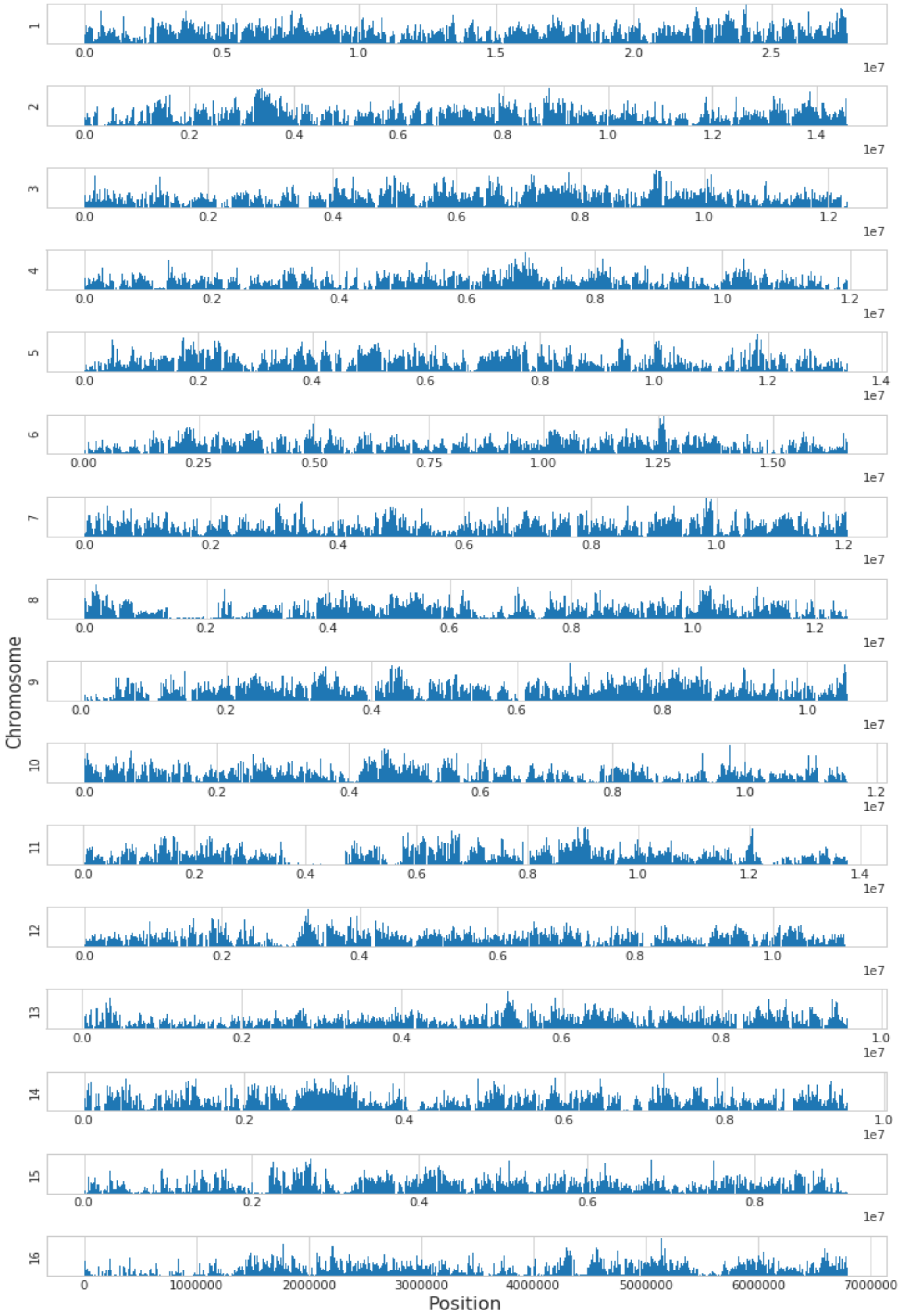


Fig. 6. A correlation matrix between the allele frequencies of all 925 colonies for the 27 SNPs found associated with hygiene by L1 regression analysis.

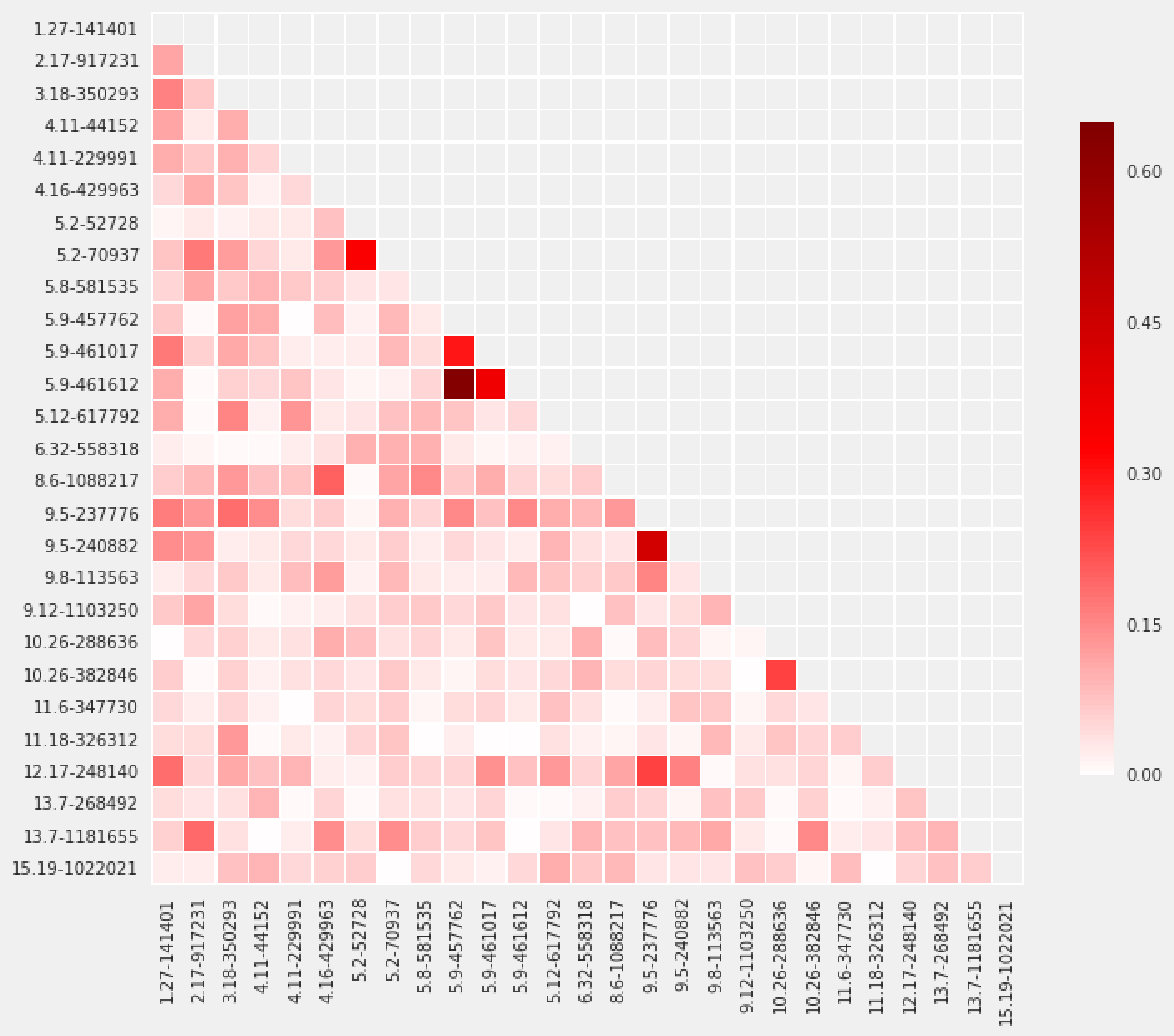


Fig. 7. Scatter plots of allele frequency and the transformed hygiene score on the some of the Single Nucleotide Polymorphisms discovered by L1 regression as predictors of hygiene. The slope is determined by ordinary least squares regression. The width of the line represents the 95% confidence interval for the slope determined by bootstrapping.

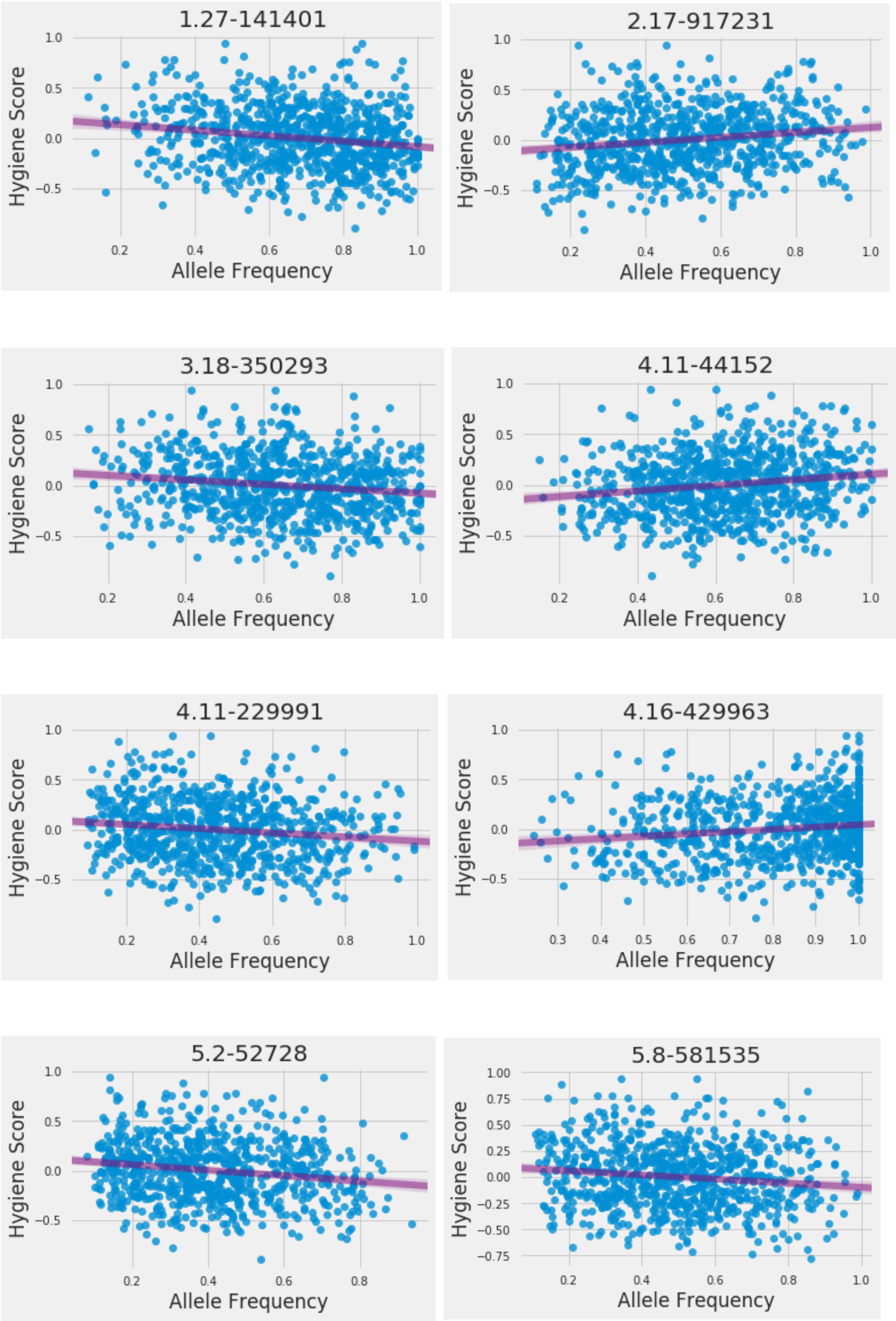


Fig. 8. SNP locations on chromosome 2 and their associations with hygienic behavior. The top panel shows the negative log p-values for single marker regression performed on the SNPs found on chromosome 2. The bottom panel shows the coefficient from an L1 regression for SNPs found on chromosome 2.

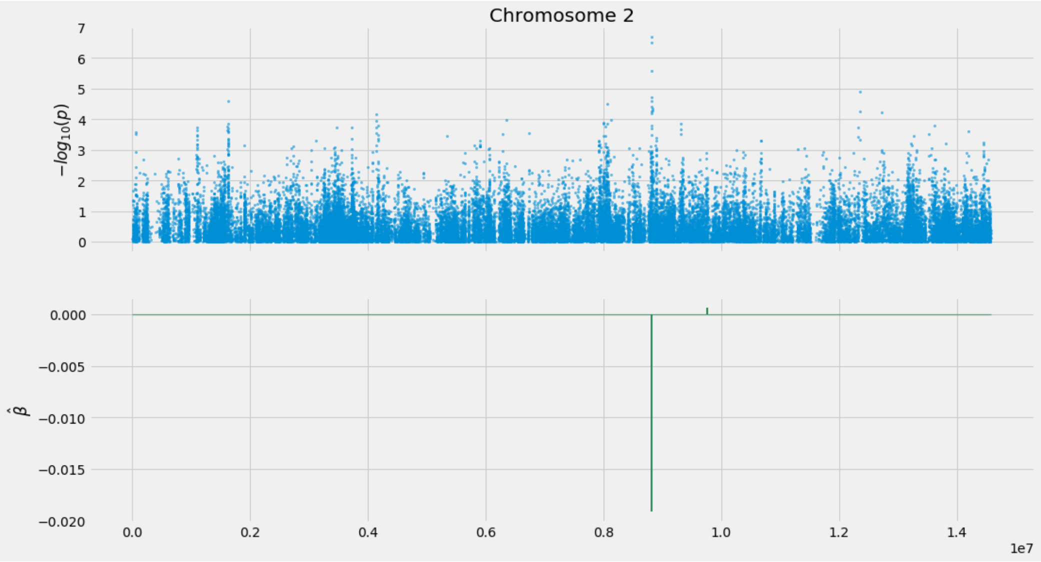


Fig. 9. SNP locations on chromosome 5 and their associations with hygienic behavior. The top panel shows the negative log p-values for single marker regression performed on the SNPs found on chromosome 5. The bottom panel shows the coefficient from an L1 regression for SNPs and their location along chromosome 5.

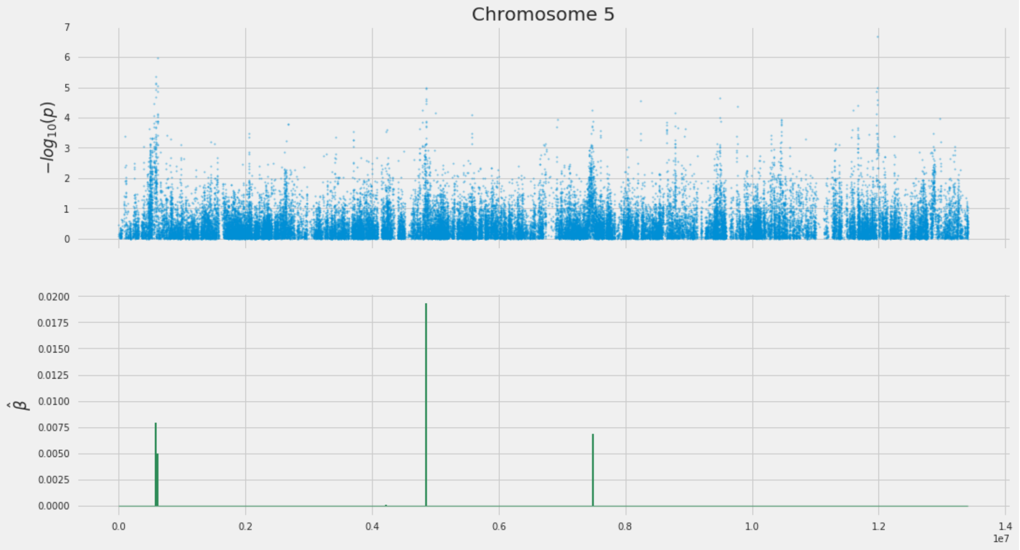


Fig. 10. The relative location and identity of genes nearby a SNP found by L1-regression. The x-axis is the relative position away from the SNP of interest. Each colored rectangle shows the genes location and size. Above the rectangle the gene ID is displayed.

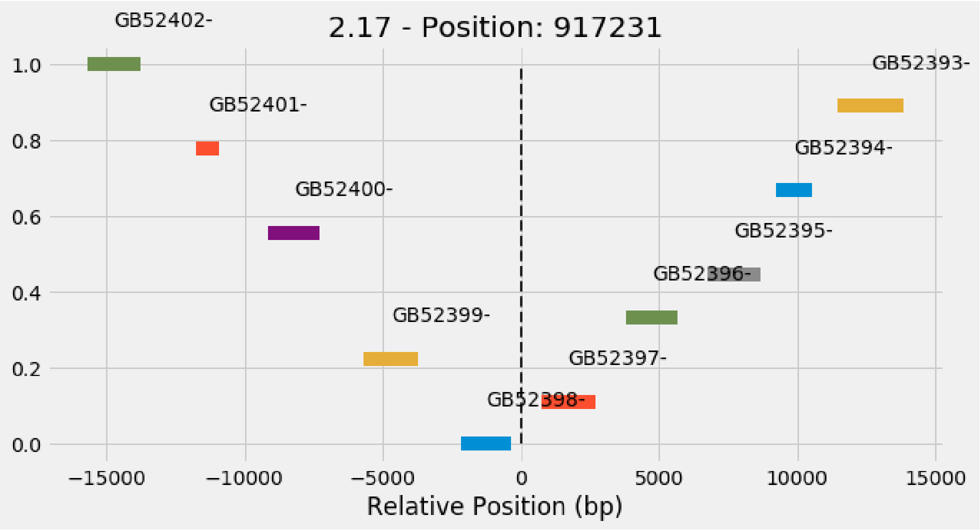


Table 1: L1 regression results for hygienic behavior. The SNPs were first selected by L1 regression then a single marker regression as used to estimate the effect size and the p-value.

Chromosome	**Position**	**Coefficient**	**p-value**
1.27	141401	-0.125	0.0000015
2.17	917231	0.124	0.0000046
3.18	350293	-0.106	0.0000567
4.11	44152	0.130	0.0000059
4.11	229991	-0.104	0.0001497
4.16	429963	0.102	0.0001053
5.2	52728	-0.135	0.0000043
5.2	70937	-0.128	0.0000011
5.8	581535	-0.099	0.0002918
5.9	457762	-0.119	0.0000355
5.9	461017	-0.101	0.0001383
5.9	461612	-0.119	0.0000281
5.12	617792	-0.098	0.0002474
6.32	558318	-0.113	0.0000225
8.6	1088217	-0.102	0.0000633
9.5	237776	-0.114	0.0000072
9.5	240882	-0.129	0.0000019
9.8	113563	-0.110	0.0000327
9.12	1103250	-0.103	0.0000862
10.26	288636	0.122	0.0000228
10.26	382846	0.095	0.0001543
11.6	347730	0.111	0.0000716
11.18	326312	0.121	0.0000361
12.17	248140	-0.119	0.0000054
13.7	268492	0.123	0.0000257
13.7	1181655	-0.108	0.0000204
15.19	1022021	0.083	0.0017951

Table 2: Honeybee genes closest to the Single Nucleotide Polymorphisms found by L1 regression analysis and their fly orthologues.

Chromosome	Start	Gene	Orthologue
1	15476132	GB55285	CG32225
3	12962146	GB46114	Frl
4	12195806	GB53015	CG7231
4	8656629	GB40999	sba
4	8494687	GB40996	2mit
4	8531049	GB41002	timeout
5	5250853	GB46757	chn
5	5234035	GB46720	Dhit
9	3615040	GB44850	Orc3
9	10339556	GB53422	CG30157
10	11260332	GB51005	NAA20
11	2032842	GB55196	ara
11	9011264	GB45092	CG1402
12	9282426	GB52025	Nep2
13	3442086	GB40109	mdy

Chapter 2: The Genetics of Overwintering Behavior

2.1 Introduction

Canadian beekeepers experience significant colony loss over the winter. Figure 1 shows the results of reported colony loss due to winter for the five provinces in this study over the past decade. Within each province, between 20% to 60% of the honeybee colonies die each winter. Counter intuitively the number of colonies within these provinces have increased within the same period. The increase is explained by an increase in imported bees by Canadian beekeepers to compensate for their winter losses (Gallai et al. 2009). Although there seems to be a slight decrease in the percentage of colony loss over the recent years, the increased number of colonies implies that the total number of honeybee colonies dying is larger. Therefore, importing foreign honeybees is likely to continue. Apart from the cost, two major concern for importing honeybees is the introduction of new diseases or accidental importing of Africanized honey bees. In fact, the honeybees' greatest parasite, the varroa mite, was a parasite to the Asian honeybee *Apis cerana*, and has spread across the world in the last few decades (Rosenkranz et al. 2010). The introduction is so severe that even with miticide treatment and breeding of varroa resistant bees, it is difficult to avoid a mite infestation. Africanized bees, which are more aggressive and produce less honey, were also the result of importing bees from Africa to Brazil (Winston, 1987).

The more colonies that are imported the higher the chances of introducing foreign pathogens and pests, and the higher the chances of importing Africanized bees. Reducing the number of imported bees, and thus reducing this risk, can be done by preventing winter colony loss. To improve winter survival of Canadian honeybees, we can determine the genetics of overwintering behavior in honeybees and find markers to breed winter hardy bees.

Overwintering behavior refers to the behavioral and physiological changes to honeybee workers that permit them to remain dormant during the winter. They do so by remaining within the colony, reducing their activity, and increasing their longevity (Doke et al, 2015). To survive the cold, the bees form a thermoregulating cluster by huddling and vibrating their flight muscles constantly to generate heat (Heinrich and Esch, 1994). The clustering begins when outside temperatures reach 18˚C, and as the temperature drops, the cluster contracts, which conserves heat by reducing the surface area over which heat is lost (Winston, 1987). Contraction stops when the temperature reaches below -5˚C, at which point the workers themselves generate heat (Free, 1977). Honeybees consume honey to expend energy to generate heat, however if the conditions are too cold the colony may die since individual bees are unable to move to honey resources to feed (Haydak, 1958). Brood rearing is also significantly reduced due to the temperature constraints required to develop healthy brood (Winston, 1987).

We know overwintering behavior is heritable indirectly because certain lineages of the honeybee survive the winter better (Southwick and Heldmaier, 1987). The natural habitat of the honeybee ranges from South Africa, through the savannahs, rain forest, desert, and the Mediterranean to northern Europe and Southern Scandinavia (Winston, 1987). The vast range all with differing environments combined with the movement of honeybees due to apiculture has produced honeybees with varying characteristics adapted to each region (Winston, 1987). The European honeybees *Apis mellifera mellifera*. originated in Europe and west-central Russia and winter well but are aggressive so have been used less. *Apis mellifera ligustica* originated in Italy and they too winter well in large numbers though they are docile. *Apis mellifera carnica* originated in the Australian alps but they overwinter in small colonies though grow rapidly in the spring. *Apis mellifera caucasica* on the other hand are poor at overwintering due to susceptibility to disease.

African races of the honeybee, such as *intermissa, lamarckii, scutellate, adansonii, monticola*, and *capensis* do not survive the winter at all. Additionally, Buchler et al. (2015) showed that colonies with a local queen survived the winter an average of 83 +/- 23 days relative to colonies without a local queen, suggesting a genotypic effect on overwintering.

My aim is to find genetic factors behind the overwintering abilities of the M lineage of the honeybee. I will perform Genome Wide Association on 925 honeybee colonies across Canada, which will be measured for the overwintering behavior phenotype and have their genome sequenced. I will then associate the genotype information to the phenotype to identify mutations that may explain overwintering behavior. Since overwintering behavior is likely heritable, I will find one or more Single Nucleotide Polymorphism (SNP) that associates with the trait. Therefore, for each SNP, I will be test the hypothesis that the SNP is associated with overwintering behavior.

2.4 Methods

2.4.1 Colony and Phenotype

Since overwintering behavior is crucial to colony winter survival it will be proxied by observing which of the 925 colonies survived the winter. Colonies that perform the behavior better, either by having a larger cluster or the ability to shiver more, the more likely they are to survive.

2.4.2 Extraction and Sequencing

See 1.4.2 for in-depth details, See Appendix A for summary of steps

2.4.3 Alignment and SNP calling

See 1.4.3

2.4.4 GWAS analysis

See 1.4.4 for general information regarding GWAS analysis via L1 Regression.

L1 regression can also be performed on categorical phenotype data using the glmnet package. Logistic regression is used to model a binary categorical dependent variable. The principles for L1 regression for a logistic regression are the same, but instead of modeling the phenotype, logistic regression models the log odds of belonging to one of the classes (survived or died from the winter).

To account for differences in survival between provinces the province was used as a cofactor in the model. The province cofactor was unregularized when L1 regression was performed to ensure the effect of the province is accounted for. The regularization strength λ was also chosen when the false positive rate was less than 1/5000.

2.5 Results

Overwintering mortality was measured as the number of colonies that survived after the winter of 2016. The total number of colonies that survived were 601 while 139 died, the remaining colonies died before winter and were excluded from the analysis. This translates to a 23% winter survival rate, which is in the range of winter survival observed across Canada (Figure 1). Differences in weather and management between provinces also led to drastic differences in overwintering mortality between the provinces. For each province, the percentage of colonies that survived within each yard was calculated. The distribution of yard survival rates for each province is shown in Figure 2. There is extreme variation in winter survival for yards within Ontario, Quebec, and British Columbia. There are yards in these provinces where all colonies survived, but there are also yards within these provinces where all colonies died. For most yards however, there were more colonies that survived than died. Alberta and Manitoba on the other hand have almost no variation in winter survival between the yards in these provinces. Both Manitoba and Alberta have high rates of colony survival and less variability, with most yards having a greater than 80% colony survival rate through the winter.

Feature selection by L1 Regression on overwintering behavior resulted in 32 SNPs (Table 1). Unlike for hygienic behavior the locations of the SNPs are not evenly spread among chromosomes. Chromosome 1, 7, and 11 have the most associated SNP's while the rest have 1 or 2 SNPs. A correlation matrix between the SNPs found by L1 regression is shown in figure 3. Like hygienic behavior, most SNPs are uncorrelated with each other. This is because the separation of SNPs into different chromosomes allows them to be independently inherited. Similarly, separation of SNPs on the same chromosome by a large distance allows recombination to separate them. The distribution of correlation values between the SNPs is shown in figure 4. Overall most of the SNPs appear to be uncorrelated or weakly correlated with each other. However, when SNPs are nearby, the chance of a recombination event is low which causes high linkage and thus high correlation. The highest correlation observed is 0.93 and it happens between SNPs 11.12-487386 and 11.2-487546 which are only 160 basepairs apart.

The regularization strength, λ, path can be seen in figure 5. The x axis shows the different λ values attempted, the y axis shows the coefficient value for a SNP, each line is a different SNP, and the black vertical line is the optimal λ chosen. At high λ values almost all SNPs have zero coefficients, and at low λ values many SNPs have non-zero coefficients. At the optimal value chosen 32 of the SNPs have none zero coefficients, and the chance of any of them being a false positive is 1 in 5000. Since L1 regression coefficients are regularized, they don't accurately reflect the true effect of the SNP. Therefore, single marker regression was performed on the SNPs to estimate the true coefficient. All the SNPs have similar effect sizes, though some are negative, meaning having the reference allele leads to an increase in the odds of the colony dying. The SNPs found for overwintering have a larger effect size than those found for hygienic behavior. Figure 6 shows how allele frequency is distributed between colonies that lived and colonies that died for some of the SNPs found by L1 regression. It is hard to visualize the true model which had both the other SNPs and the location of the colonies as cofactors. However, there seems to be some difference between the allele frequencies of these two groups which the model used.

Finally, the regions within which the SNPs are found can be explored for genes. The number of genes to explore are small since recombination disassociates nearby regions from each other. The nearest honeybee genes and their *Drosophila* orthologues are shown in Table 2. The orthologues are involved in cold acclimation, muscle development, transcription, olfaction, and metabolism (see discussion).

The ability of the model to predict whether a colony will die is shown in figure 7 using a receiver operator curve. The graph shows, for different thresholds, what the true positive rate and false positive rate is. The dashed black line represents a model that randomly labels a colony as alive or dead while the blue line represents the L1 regression model. In a receiver operator curve, a model with a line above the dashed line performs better than a model that is randomly guessing. Therefore, the area under the curve numerically represents the prediction ability of the model, since a line above the dashed line will contain more area. The area covered by a perfect predictor is 1. The L1 regression model has an area of 0.74, thus performs better than chance but does not make perfect predictions.

2.6 Discussion

The genetics of overwintering behavior is unknown, though there is some evidence of genetic factors at play. The different races of the honeybee have different adaptations to winter survival, with some races able to survive the winter and others unable to. In the races that do, overwintering behavior is performed by the remaining bees during the winter season to generate heat and survive the cold. Unfortunately, there is considerable colony loss over the winter in Canada (Figure 1). The current approach to recouping winter colony losses, such as importing more bees, pose a risk of introducing foreign pathogens or unwanted genetics of the Africanized honeybee. Here, we search for genetic markers for these overwintering to potentially use marker assisted selection to breed winter hardy honeybees.

The genetics of overwintering was determined by performing a genome wide association study. Colonies were scored for winter survival and 50 workers from each were sequenced in a pool to determine the genetics of the colony. L1 regression was used to select the SNPs that strongly affect the trait. The major finding here are the 32 SNPs that are significantly associated with winter survival.

There are no previous studies that aimed to identify the genetics of overwintering behavior in the honeybee. QTL analysis for example would be hard to perform on this trait since it is difficult to breed colonies that did not survive the winter. The model presented here uses 32 SNPs and does better than chance at predicting whether a colony would live or die. However, depending on the threshold used to choose whether a colony will die, the model has differing true positive and false positive rates (Figure 7). This is important, because different parts of Canada have different needs when it comes to overwintering bees. For example, Alberta hardly loses colonies over the winter, whereas Ontario loses many colonies. Beekeepers within each of these provinces can then choose different thresholds of the model that is appropriate with how much they care about overwintering. A beekeeper in Alberta would aim for a low false positive rate, since the model would tell them many of their colonies that lived would die. Whereas a beekeeper in Ontario would want a higher true positive rate and wouldn't mind some false positives if it meant that they identify the colonies that did die. The significance of this is that by just genotyping 32 SNPs from a few bees within their colony, they can know how much effort to put for that colony in ensuring they survive the winter before winter is observed.

The genes nearest the SNPs found by L1 regression is shown in Table 2 along with their *Drosophila* orthologues, which are involved in cold acclimation and muscle development. The orthologue *Sfmbt* had two fixed substitutions in the middle of the gene between temperate and tropical populations of *Drosophila* in Australia (Kolaczkowski et al. 2011). The authors measured genetic differentiation using Fst with 1kb windows, which primarily had single genes in them. The region containing Sfmbt was in the top 2.5% of Fst values (Fst=0.45) between the tropical and temperate population (Figure 8). Two of the genes Sfmbt interacts with, CG33275 and CG17018 (Yu et al. 2008), were also in regions with high Fst values, 0.28 and 0.33 respectively. The orthologues CG1402 and CG9259, are also significantly differentially expressed between tropical and temperate populations of *Drosophila* in Australia and in the United States (Juneja et al. 2016). The orthologues CG31751 and CG14022 were found to be significantly downregulated (1.29-fold and 1.97-fold respectively) by cold acclimated flies compared to warm acclimated flies (MacMillan et al. 2016). The orthologue CG14022 was also found upregulated in hypercontraction mutants (Mhc mutants), which are flightless *Drosophila* mutants and were found with mutations in the myosin heavy chain gene (Montana and Littleton, 2006). To acclimate in cold temperatures, poikilothermic fish change the abundance and type of muscle fibers by differential expression of proteins involved in muscle contraction, including myosin light and heavy chains (Gracey et al. 2004). Moreover, both myosin heavy and light chains were significantly upregulated 3.5-fold in cold acclimated flies (MacMillan et al. 2016). The honeybee vibrates its wing muscles to generate heat over the winter. This is significant because, not only are *Drosophila* myosin heavy chain mutants unable to fly, in both *Drosophila* and fish myosin heavy chain is differentially expressed in cold acclimation studies. On the same note, the orthologue *nrm* is expressed in developing muscle cells and may be a cell adhesion molecule, binding nearby muscle cells together (Kania et al. 1993). The orthologue CG14446 is differentially expressed between nociceptor neurons and non-nociceptor neurons (Honjo et al. 2016). Nociceptor neurons are sensory neurons that respond to harmful or potentially harmful stimuli, in this case temperature was the stimulus. CG14446 was knocked down in flies which made them hypersensitive to temperature. Finally, the remaining orthologues are involved in calcium transport, gene regulation as transcription factors, metabolism, olfaction, locomotion, and wing development.

Using just the 32 SNPs we have found here, we can potentially breed for honeybee colonies that are strong for the winter. However, once again, the orthologues present new hypotheses to test and further validation is required of these SNPs. Many follow up experiments can be conducted, such as a common garden experiments where we know the genotype at some of these SNPs of the colonies within the yard. Molecular work will also help identify the pathways that these SNPs are involved in to manipulate the behavior of these bees. Finally, the markers used here can help beekeepers predict the fate of their colony come winter, and they can choose whether to intervene by insulating the colony, moving it indoors, or giving it more resources.

2.7 Tables and Figures

Fig. 1. The percent of honeybee colonies lost during the winter for the five provinces in this study over the past decade. Data from Canadian Association of Professional Apiculturists.

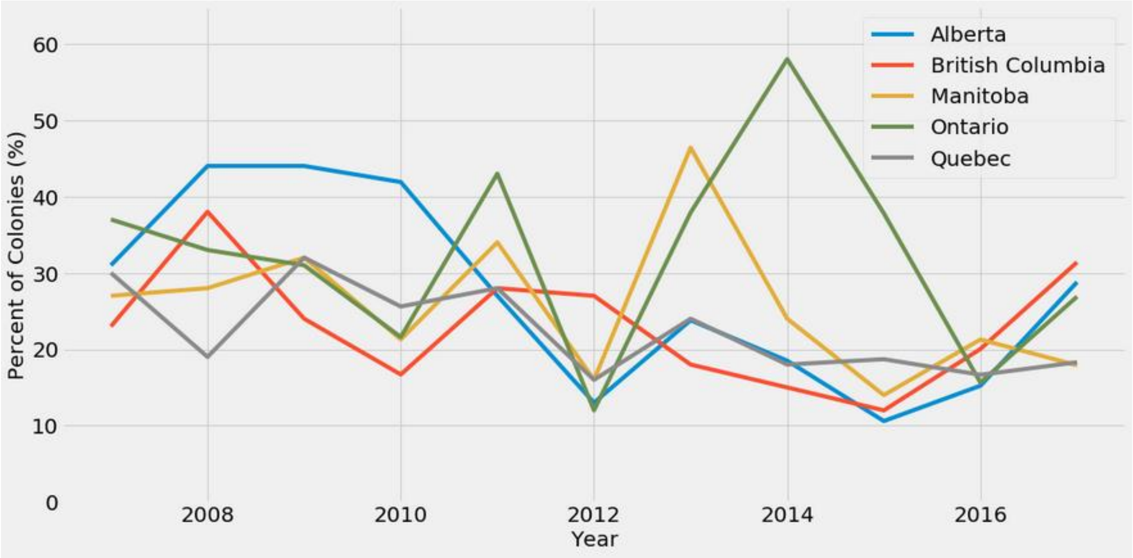


Fig. 2. Distribution of percentage of colonies survived within a yard of each province.

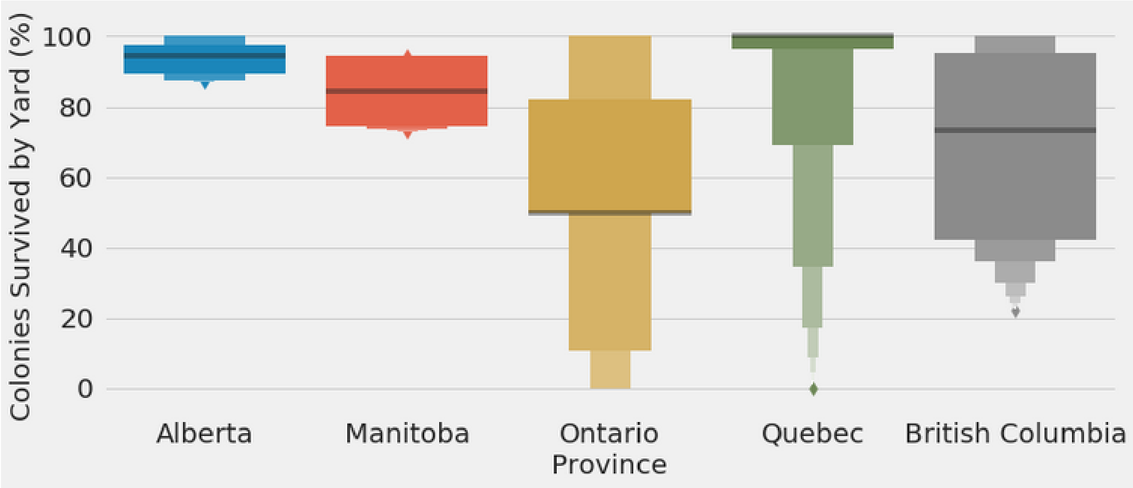


Fig. 3. A correlation matrix between the SNPs found by L1 regression that associate with winter survival in the honeybee.

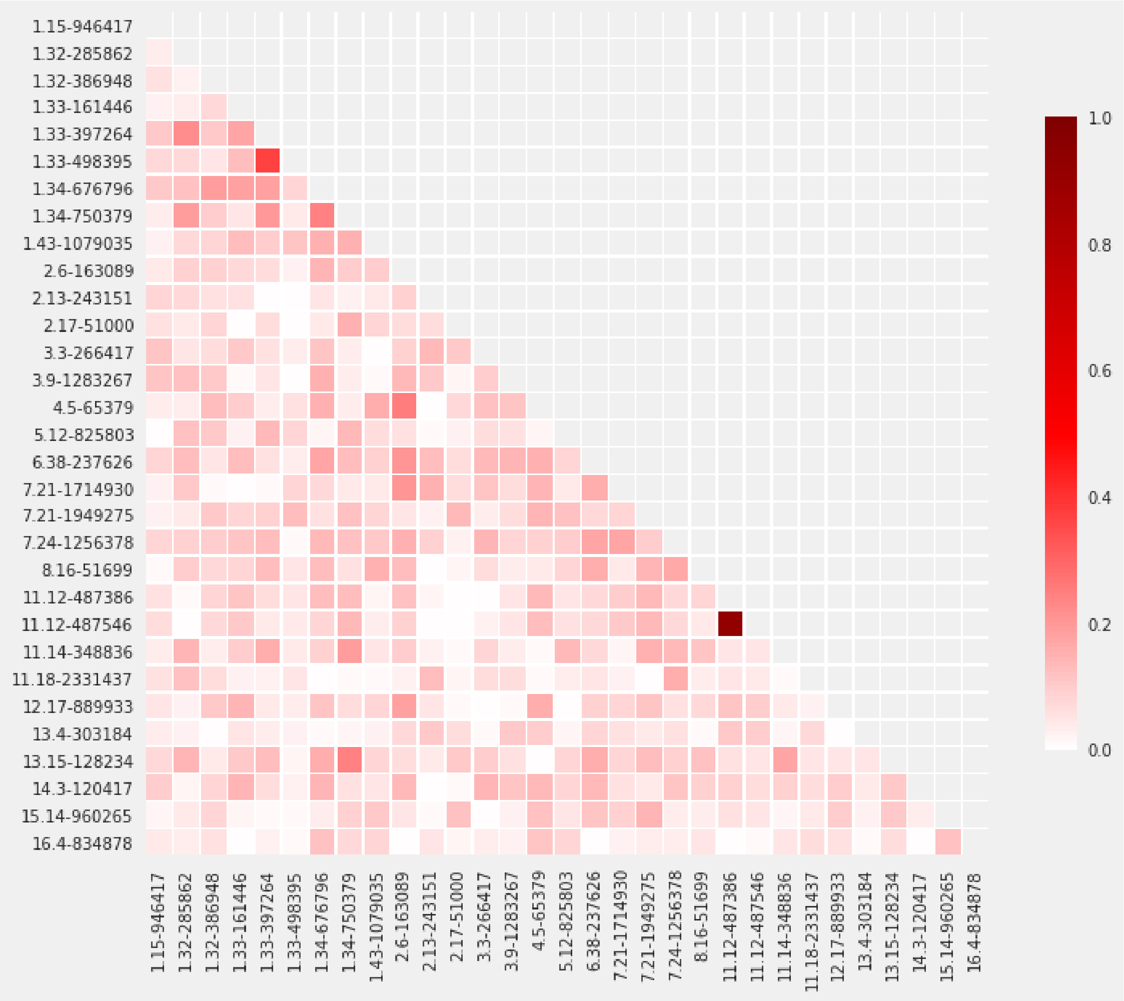


Fig. 4. The distribution of correlations between the SNPs found by L1 regression.

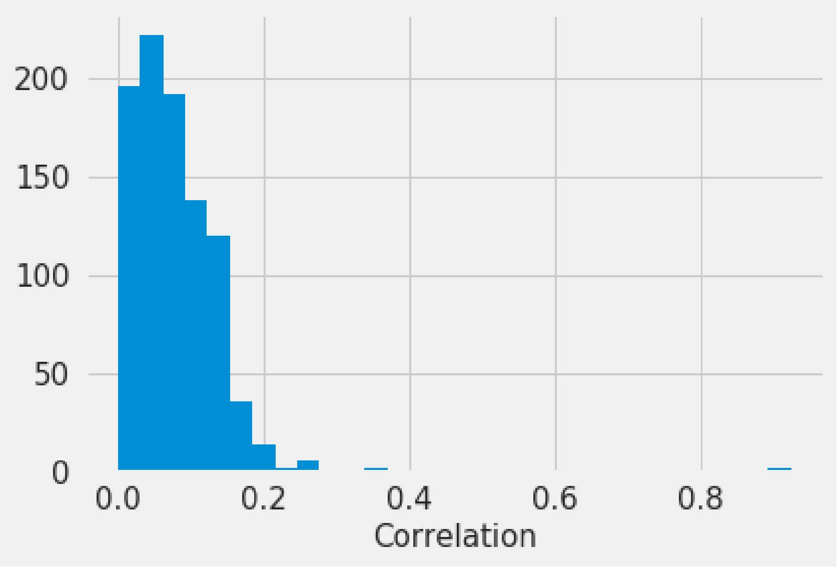


Fig. 5. The relationship between regularization strength and the coefficients of the SNPs effect on overwintering. The vertical black line is the lambda value chosen while optimizing for false positives.

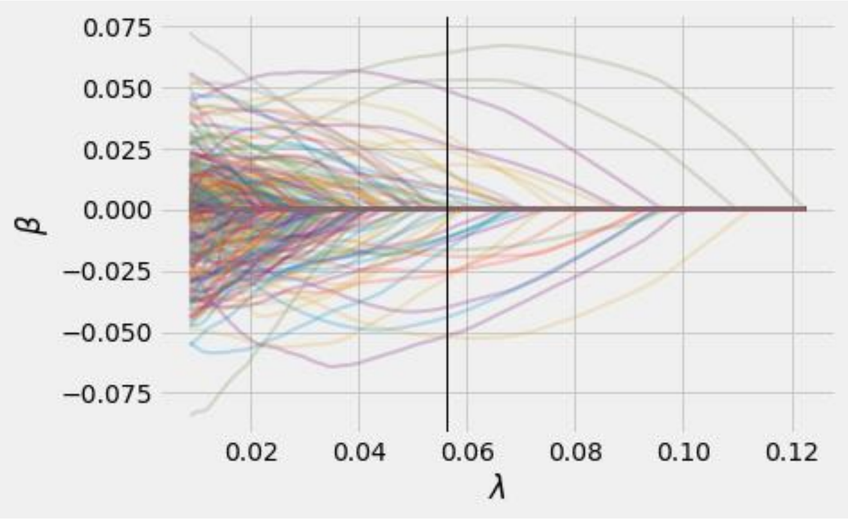


Fig. 6. Distribution of allele frequencies of the SNPs found by L1 regression between honeybee colonies that survived the winter and those that lived.

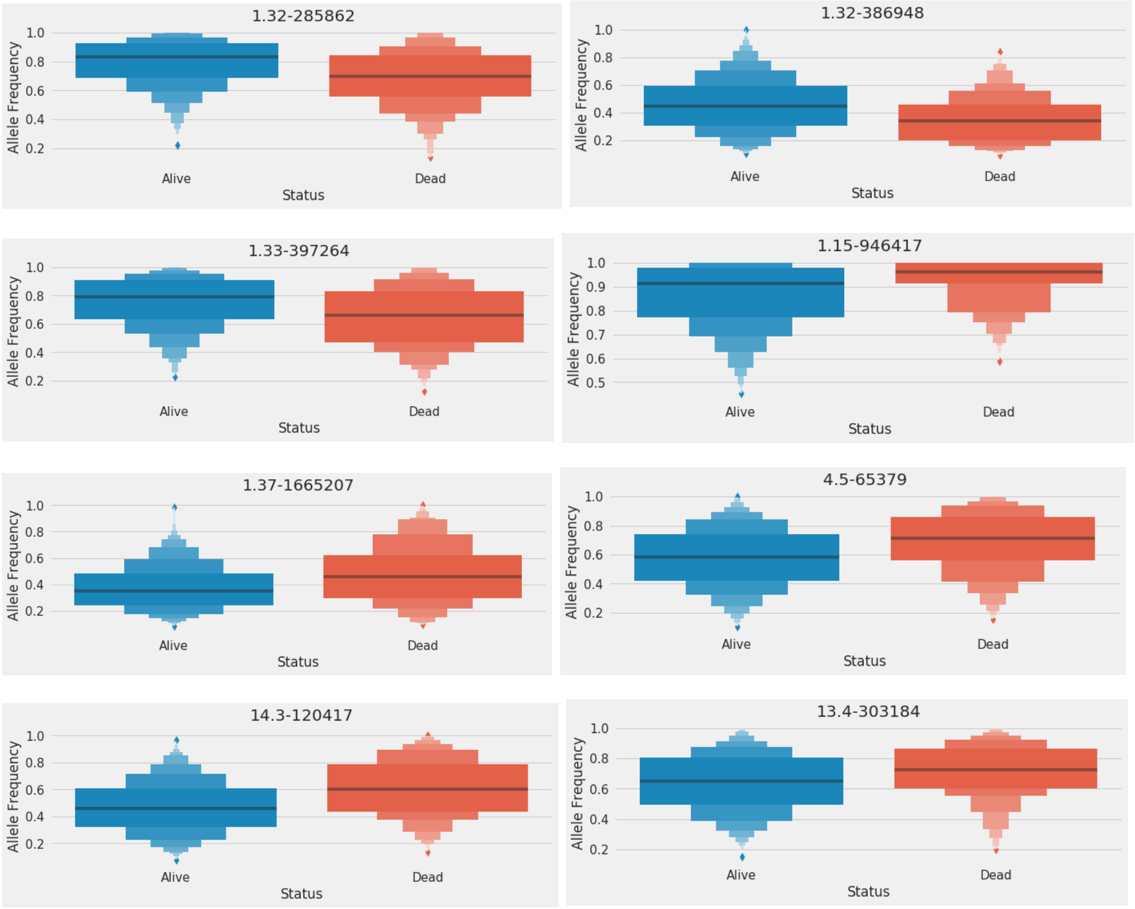


Fig. 7. A receiver operator curve showing the prediction of colony survival after winter. The dashed black line is a model that guesses whether the colony lives or dies. The blue line is the models ability under different thresholds of determining survivability.

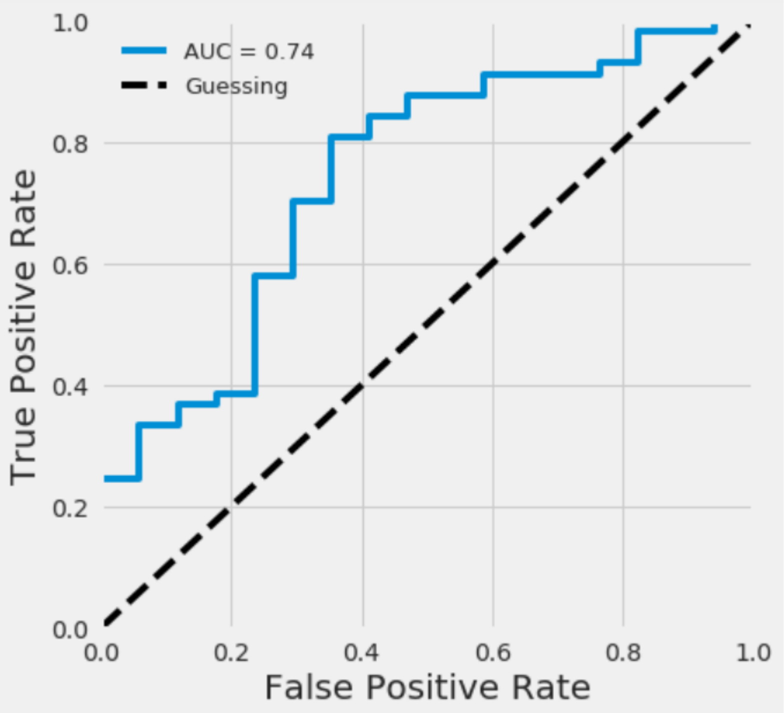


Fig. 8. Fst calculated over 1kb windows between temperate and tropical populations of *Drosophila* (Kolaczkowski et al. 2011).

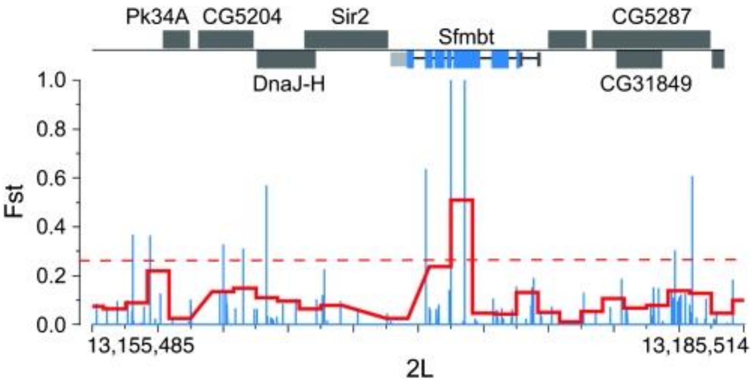


Table 1: L1 regression results for overwintering behavior. The SNPs were first selected by L1 regression then a single marker regression was used to estimate the effect size and the p-value.

Chromosome	Position	Coefficient	P-value
1.15	946417	-1.82077	0.00166
1.32	285862	1.578596	0.00002
1.32	386948	1.728478	0.00009
1.33	161446	1.061101	0.00157
1.33	397264	1.114577	0.00087
1.33	498395	1.052584	0.00265
1.34	676796	1.246062	0.00137
1.34	750379	1.287128	0.00019
1.37	1665207	-1.23827	0.00182
1.43	1079035	0.910622	0.00306
2.6	163089	-1.15747	0.00301
2.13	243151	0.932249	0.00336
2.17	51000	1.008364	0.00314
3.3	266417	1.258365	0.00028
3.9	1283267	-1.1098	0.00211
4.5	65379	-1.53947	0.00008
5.12	825803	1.136059	0.00139
6.38	237626	-1.18236	0.00095
7.21	1714930	-1.43108	0.00055
7.21	1949275	-1.29037	0.00100
7.24	1256378	1.013474	0.00239
8.16	51699	1.016585	0.00316
11.12	487386	-1.21489	0.00114
11.12	487546	-1.29515	0.00055
11.14	348836	1.100246	0.00273
11.18	2331437	-1.16208	0.00349
12.17	889933	-1.34706	0.00023
13.4	303184	-1.51995	0.00009
13.15	128234	-1.09088	0.00269
14.3	120417	-1.34054	0.00011
15.14	960265	1.057922	0.00255
16.4	834878	1.477864	0.00007

Table 2: The nearest genes found to some of the SNPs selected by L1 regression for winter survival along with their *Drosophila* orthologues.

chromosome	**start**	**gene**	**orthologue**
1	8964454	GB48998	Sfmbt
1	20055611	GB40022	CG31751
1	20216040	GB40013	CG14022
2	7049105	GB43389	CG10898
2	8766705	GB52446	geko
2	7044563	GB43388	TfIIA-S
3	7438614	GB53732	CG9259
5	7698059	GB48665	Calx
7	8904172	GB42411	twz
7	12239992	GB48097	Nmdar2
11	11223500	GB45020	CG14082
11	11205076	GB45150	Ac3
11	6420214	GB43497	nrm
11	5458139	GB54030	zfh2
13	1711578	GB48229	CG14446

Conclusion

In this thesis I explore the genetics of hygienic behavior and overwintering behavior in the honeybee using a genome wide association study. The project led to the development of scripts that can process large amounts of pooled sequencing data and perform downstream statistical analyses. Researchers have attempted to determine the genetics of hygienic behavior since the 1960's but have only managed to identify areas of the genome that span millions of base pairs and contain hundreds of genes. No previous attempts have been made to determine the genetic basis of overwintering behavior. Here, I find 27 and 32 Single Nucleotide Polymorphism for hygienic behavior and overwintering behavior respectively. The associated regions that contain these SNPs span 50000 base pairs representing a tremendous increase in the resolution over previous results. Genomic regions associated with these phenotypes only contain 5-10 genes which can easily be explored and lead to testable hypotheses about how these genes could potentially be involved in these behaviors. The knowledge of these SNPs also allows for informed breeding programs that can select for healthier bees using these markers. To do so, we would sample bees from a colony and determine their genotypes at the SNPs found. From this, a prediction can be made on how hygienic and winter hardy the colony is. If this is repeated for many colonies, it becomes possible to choose desirable colonies and breed them to propagate their genetics, and ultimately these traits. The result would be an increase in disease resistance and winter survival of the selected honeybee colonies. The methods and results of this project will lead to new knowledge about the genetics of complex social traits in the honeybee and potentially improve the health of honeybees while reducing the risks and costs of having to import honeybee colonies.

References

1. AAFC. 2015. Statistical Overview of the Canadian Honey and Bee Industry.
2. Al-Waili N, Salom K, Al-Ghamdi A, Ansari MJ. 2012. Antibiotic, pesticide, and microbial contaminants of honey: human health hazards. Scientific World Journal.
3. Arya G.H, Magwire M.M, Huang W. Serrano-Negron Y.L. Mackay T.F.C. Anholt R.R.H. 2015. The genetic basis for variation in olfactory behavior in *Drosophila* melanogaster. Chem Senses. 40(4): 233-243
4. Auwera G.A, Carneiro M, Hartl C, Poplin R, Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella K, Altshuler D, Gabriel S, DePristo M. 2013. From fastQ Data to high-confidence variant calls: The Genome Analysis Toolkit Best Practices. Current protocols in Bioinformatics.
5. Baggio F, Bozzato A, Benna C, Leonardi E, Romoli O, Cognolato M, Tosatto S.C.E, Costa R, Sandrelli F. 2013. 2mit, an intronic gene of *Drosophila* melanogaster timeless2, is involved in behavioral plasticity. PLos One. 8(9):
6. Benjamini Y, Yekutieli D. 2001. The control of the false discovery rate in multiple testing under dependency. The annals of Statistics. 29(4): 1165-1188
7. Benna C, Scannapieco P, Piccin A, Sandrelli F, Zordan M, Rosato E, Kyriacou C.P. Valle G, Costa R. 2000. A second timeless gene in *Drosophila* shares greater sequence similarity with mammalian tim. Current Biology. 10(14): 512-513
8. Beye M, Gattermeier I, Hasselmann M, Gempe T, Schioett M, Baines JF, Schlipalius D, Mougel F, Emore C, Rueppell O, Sirviö A. 2006. Exceptionally high levels of recombination across the honey bee genome. *Genome Research*. 16(11):1339-1344
9. Bolger, A. M., Lohse, M., & Usadel, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 30(15), 2114–2120.

10. Borevitz J. O, Nordborg M. 2003. The impact of genomics on the study of natural variation in Arabidopsis. Plant Physiol. 132 (2): 718-725
11. Boynton S, Tully T. 1992. Latheo, a new gene involved in associative learning and memory in *Drosophila melanogaster*, identified from P element mutagenesis. Genetics: 131(3): 665-672
12. Broman K.W. 2001. Review of Statistical methods for QTL mapping in experimental crosses. Lab animal. 30 (7)
13. Buchler R.B, Costa C, Hatjina F, Andonov S, Meixner M.D, Le Conte Y, Uzunov A, Berg S, Binkowska M, Bouga M, Drazic M, Dyrba W, Kryger P, Panasiuk B, Pechhacker H, Petrov P, Kezi N, Korpela S, Wild J. 2014. The influence of genetic origin and its interaction with environmental effects on the survival of Apis mellifera L. colonies in Europe. Journal of Apicultural Research. 53(2): 205-214
14. Cantor R. M, Lange K, Sinsheimer J.S. 2010. Prioritizing GWAS results: A review of statistical methods and recommendations for their application. Am j Hum Genet. 86: 6-22
15. Cao W, Song H, Gangi T, Kelkar A, Antani I, Garza D, Konsolaki M. 2008. Identification of novel genes that modify phenotypes induced by Alzheimer's beta amyloid overexpression in *Drosophila*.
16. Cavodeassi F. Modolell J. Campuzano S. 2000. The Iroquois homeobox genes function as dorsal selectors in the *Drosophila* head. Development. 197:1921-1929
17. Cremer S, Armitage A.O, Schmid-Hempel P. 2007. Social Immunity. Current Biology. 17:693-702
18. Doke M.A, Frazier M, Grozinger C.M. 2015. Overwintering honey bees: biology and management. Insect Science. 10: 185-193
19. Dollar G. Gombos R. Barnett A.A. Hernandez D.S. Maung S.M.T. Mihaly J. Jenny A. Unique and overlapping functions of formins Frl and DAAM during ommatidial rotation and neuronal development in *Drosophila*. Genetics. 209(3):

20. Evans J.D, Aronstein K, Chen Y.P, Hetru C, Imler J, Jiang H, Kanost M, Thompson G.J., Zou Z, and Hultmark D. 2006. Immune pathways and defense mechanisms in honeybees Apis mellifera. Insect Mol Biol. 15(5):645-656.
21. Escudero L.M, Caminero E, Schulze K.L, Bellen H.J, Modolell J. Charlatan, a Zn-finger transcription factor, establishes a novel level of regulation of the proneural achaete/scute genes of *Drosophila*. Development. 132(6): 1211-1222
22. Free J.B. 1977. The social organization of the honeybee. Northern Bee Books.
23. Gallai et al. 2009. Economic valuation of the vulnerability of world agriculture confronted to pollinato decline. Ecological Economics. 68(3): 810-821
24. Galizia G.C, Menze R. 2000. Odour perception in honeybees: coding information in glomerular patterns. Current opinions in Neurobiology. 10: 504-510
25. Gracey A.Y, Fraser E.J. Li W. Fang Y. Taylor RR. Rogers J. Brass A. Cossins A.R. 2004. Coping with cold: An integrative, multitissue analysis of the transcriptome of poikilothermic vertebrate. Proc Natl Acad Sci. 101(48): 16970-16975
26. Harbo JR, Harris HA. 1999. Heritability in Honey Bees (Hymenoptera: Apidae) of Characteristics Associated with resistance to Varroa jacobsoni (MesostrigmataL Varroidae). Apiculture and Social Insects. 90(2): 261-265
27. Haydak M.H. 1958.Wintering of bees in Minnesota. J. Econ. Entomol. 51:332-334.
28. Harpur B.A, Kent C.F, Molodtsova D, Lebon J.M.D, Alqarni A.S, Owayss A.A, Zayed A. 2014. Population genomics of the honeybee reveals strong signatures of positive selection on worker traits. PNAS. 111(7): 2614-2619
29. Harpur B.A, Zayed A. 2013. Accelerated evolution of innate immunity proteins in social insects: adaptive evolution or relaxed constraint? Mol. Biol. Evol. 30(7): 1665-1674
30. Heinrich B, Esch H. 1994. Thermoregulation in Bees. American Scientist. 82(2): 164-170
31. Heisenberg M. 2003. Mushroom body memoir: from maps to models. Nature reviews neuroscience. 4:

32. Heisenberg M, Borst A, Wagner S, Byers D. 1984. *Drosophila* mushroom body mutants are deficient in olfactory learning. Journal of Neurogenetics. 2(1): 1-30

33. Higes M, Martín-Hernández R, Botías C, Bailón E.G, González-Porto, A.V, Barrios L, del Nozal M.J, Bernal J.L, Jiménez J.J, Palencia P.G, Meana A. (2008), How natural infection by *Nosema ceranae* causes honeybee colony collapse. Environmental Microbiology, 10: 2659-2669.

34. Honey Bee Genome Sequencing Consortium. 2014. Finding the missing honey bee genes: lessons learned from a genome upgrade. BMC Genomics 15:86.

35. Honey Bee Genome Sequencing Consortium. 2006. Insights into social insects from the genome of the honey bee Apis mellifera. Nature 443:931-949.

36. Honjo K, Mauthner SE, Wang Y, Skene JP, Tracey Jr WD. 2016. Nociceptor-enriched genes required for normal thermal nociception. Cell reports. 16(2): 295-303.

37. Friedman J, Hastie T, Tibshirani R. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software. 33(1), 1-22.

38. Juneja P. Quinn A. Jiggins F. 2016. Latitudinal clines in gene expression and cis-regulatory element variation in *Drosophila* melanogaster. BMC genomics. 17: 981

39. Kania A, Han P, Kim Y, Bellen H. 1993. Neuromusculin, a *Drosophila* gene expressed in peripheral neuronal precursors and muscles, encodes a cell adhesion molecule. Neuron. 11: 637-687

40. Kent CK, Tiwari T, Rose S, Patel H, Conflitti IM, and Zayed A. In press. Studying the genetics of behaviour in the genomics era. Encyclopedia of Animal Behavior, 2nd Edition, Elsevier.

41. Klein A, Vaissiere B, Cane J.H, Steffan-Dewenter I, Cunningham S.A, Kremen C, and Tscharntke T. 2007. Importance of pollinators in changing landscapes for world crops. Proc R Soc. 274: 303-313.

42. Koffler S, de Matos Peixoto Kleinert A, Jaffe R. 2016. Quantitative Conservation genetics of wild and managed bees. Conservation genetics.

43. Kolaczkowski B. Kern A.D. Holloway A.K. Behun D.J. 2011. Genomic differentiation between temperate and tropical Australian populations of *Drosophila* melanogaster. Genetics. 187(1): 245-260

44. Korte A, Farlow A. 2013. The advantages and limitations of trait analysis with GWAS: a review. Plant Methods. 9(29)

45. Krishnan B, Dryer S, Hardin P.E. 1999. Circadian rhythms in olfactory responses of *Drosophila* melanogaster. Nature. 400. 375-8.

46. Lapidge K.L, Oldroyd B.P, Spivak M. 2002. Seven suggestive quantitative trait loci influence hygienic behavior of honeybees. Naturwissenschaften. 89:565-568

47. MacMillan H.A. Knee J.M. Dennis A.B. Udaka H. Marshall K.E. Merritt T.J.S. Sinclair B.J. 2016. Cold acclimation wholly reorganizes the *Drosophila* melanogaster transcriptome and metabolome.

48. Masterman R, Smith B.H, Spivak M. 2000. Brood odor discrimination abilities in hygienic bees (*Apis mellifera L.)* using proboscis extension reflex conditioning. Journal of Insect Behaviour. 13(1):

49. Montana ES, Littleton JT. 2006. Expression profiling of a hypercontraction-induced myopathy in Drosophila suggests a compensatory cytoskeletal remodeling response. Journal of Biological Chemistry.

50. Moritz R.F.A, 1988. A reevaluation of the two-locus model for hygienic behavior in honeybees (*Apis mellifera L.*). Journal of Heredity. *79*(4), pp.257-262.

51. Morse R.A, and Calderone N.W. The Value of Honeybees As Pollinators of U.S. Crops in 2000. Pollination.

52. Murray, K.D., Aronstein, K.A., Eischen, F.A. 2009. Promiscuous DNA and terramycin resistance in American Foulbrood bacteria. American Bee Journal. 149(6):577-581.

53. Oldroyd B.P, Thompson G.J. 2007. Behavioural Genetics of the Honey Bee *Apis mellifera*. Advances in Insect Physiology. 33:

54. Oxley P.R, Spivak M; Oldroyd B.P. 2010. Six quantitative trait loci influence task thresholds for hygienic behavior in honeybees (Apis mellifera). Molecular Ecology. 19(7): 1452-1461

55. Di Prisco G.D, Cavaliereb V, Annosciac D, Varricchioa P, Caprioa E, Nazzic F, Gargiulob G, and Pennacchioa F. 2013. Neonicotinoid clothianidin adversely affects insect immunity and promotes replication of a viral pathogen in honeybees. Proc Natl Acad Sci. 110: 18466-18471

56. Qianchuan H, Dan-yu L. 2011. A variable selection method for genome-wide association studies. Bioinformatics. 27(1): 1–8

57. Rodrigues V, Hummel T. 2008. Development of the *Drosophila* Olfactory System. Adv Exp Med Biol. 628:82-101

58. Rosenkranz P, Aumeier, Ziegelmann B. Biology and control of *Varooa destructor*. Journal of invertebrate physiology. 103: 96-119

59. Rothenbuhler W.C, Thompson V.C. 1955. Resistance to American foulbrood in honeybees. I. Differential survival of larvae of different genetic lines. J. Econ. Entomol. 49:470-475

60. Rothenbuhler W.C. 1964. Behavior genetics of nest cleaning in honeybees. IV. Responses of F1 and backcross generations to disease killed brood. American Zoologist. 4:111-123

61. Sedlazeck F.J, Rescheneder P, Haeseler A.V. 2013. NextGenMap: fast and accurate read mapping in highly polymorphic genomes. Bioinformatics. 29(21): 2790-2791

62. Southwick EE, Heldmaier G. 1987. Temperature control in honey bee colonies. Bioscience. 37(6): 395-9.

63. Spivak M, Downey D.L. 1998. Field assays for hygienic behavior in honey bees. Journal of economic entomology. 91: 64-70

64. Spivak M, Gary S.R. 2001. Resistance to American foulbrood disease by honey bee colonies Apis mellifera bred for hygienic behavior. *Apidologie* 32(6): 555-565.

65. Stephan D. Sanchez-Soriano N. Loschek L.F. Gerhards R. Gutmann S. Storchova Z. Prokop A. Ilona C. Kadow G. 2012. *Drosophila* Psidin regulates olfactory neuron number and axon

targeting through two distinct molecular mechanism. Journal of Neuroscience. 32(46): 16080-16094

66. Swanson J.A.I. Torto B. Kells S.A. Mesce K.A. Tumlinson J.h. Spivak M. 2009. Odorants that induce Hygienic Behaviour in Honeybees: Identification of volatile compounds in chalkbrood-infected honeybee larvae. Journal of Chemical Ecology. 35(9):1108-1116
67. Tibshirani R. 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. 267-288
68. Tsuruda J. M, Harris J. W, Bourgeois L, Danka R. G, Hunt G. J. 2012. High-resolution linkage analyses to identify genes that influence Varroa sensitive hygiene behavior in honey bees. PLos One. 7(11):
69. Tsvetkov N, Samson-Robert O, Sood K, Patel H.S, Malena D.A, Gajiwala P.H, Maciukiewicz P, Fournier V, Zayed A. (2017). Chronic exposure to neonicotinoids reduces honeybee health near corn crops. Science. 356:1395-1397
70. vanEngelsdorp D, and Meixner D. 2010. A historical review of managed honeybees populations in Europe and the United States and the factors that may affect them. J Invertebr Pathol. 103: s80-s95.
71. Waldmann P, Meszaros G, Gredler B, Fuerst C, Solkner J. Evaluation of the lasso and the elastic net in genome-wide association studies. Frontiers in genetics. 4, 270.
72. Walkinshaw E, Gai Y, Farkas C, Richter D, Nicholas E, Keleman K, Davis RL. 2015. Identification of genes that promote or inhibit olfactory memory formation in *Drosophila*. Genetics. 199(4): 1173-1182
73. Wang X, Green D.S. Roberts S.P. de Belle S.J. 2007. Thermal disruption of mushroom body development and odor learning in *Drosophila*. PLos One. 2(11):
74. Wilson-Rich N, Spivak M, Fefferman N.H, Starks P.T. 2009. Genetic, individual, and group facilitation of disease resistance in insect societies.

75. Wilm et al. 2012. A sequence quality aware, ultra sensitive variant caller for uncovering cell population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res. 40(22): 11189-201.

76. Winston M.L. 1987. The biology of the honeybee. Harvard University Press.

77. Yi H, Breheny P, Imam N, Lio Y, Ina H. 2014. Penalized multi-marker regression methods for genome-wide association studies of quantitative traits. Genetics. 209(3):

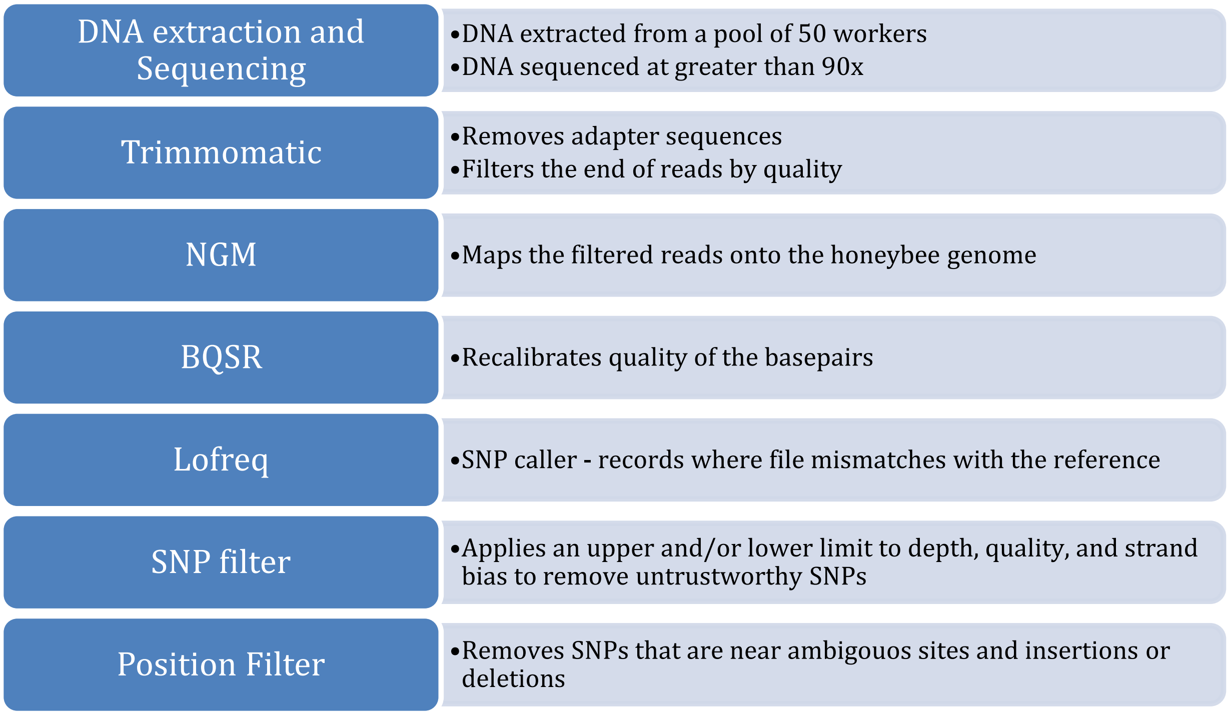
78. Yu J.S. Pacifico S. Liu G. Finley R.L. 2008. The *Drosophila* interactions database, a comprehensive resource for annotated gene and protein interactions. BMC genomics 9: 461

79. Zee R.V.D, Pisa L, Andonov S, Brodschneider R, Charrière J, Chlebo R, Coffey M.F, Crailsheim K, Dahle B, Gajda A, Gray A, Drazi M.M, Higes M, Kauko L, Kence A, Meral Kence M, Kezic N, Kiprijanovska H, Jasna Kralj J, Kristiansen P, Hernandez R.M, Mutinelli F, Nguyen B.K, Otten C, Özkırım A, Pernal S.F, Peterson M, Ramsay G, Santrac V, Soroker V, Topolska G, Uzunov A, Vejsnæs F, Wei S, and Wilkins S. 2012. Managed honeybees colony losses in Canada, China, Europe, Israel and Turkey, for the winters of 2008-9 and 2009-10. AGR RES. 51(1): 100-114

80. Zeidler M.P, Mlodzik M. 1997. Six-banded, a novel *Drosophila* gene, is expressed in 6 segmental stripes during embryonic development and in the eye imaginal disc.

Appendix A: Sequencing Pipeline

Fig. 1. Steps used to collect and process the genomic data from the honeybee colonies used in this experiment.



DNA extraction and Sequencing	•DNA extracted from a pool of 50 workers •DNA sequenced at greater than 90x
Trimmomatic	•Removes adapter sequences •Filters the end of reads by quality
NGM	•Maps the filtered reads onto the honeybee genome
BQSR	•Recalibrates quality of the basepairs
Lofreq	•SNP caller - records where file mismatches with the reference
SNP filter	•Applies an upper and/or lower limit to depth, quality, and strand bias to remove untrustworthy SNPs
Position Filter	•Removes SNPs that are near ambigouos sites and insertions or deletions

Appendix B: Abbreviations

SNP	Single Nucleotide Polymorphism
QTL	Quantitative Trait Loci
GWA	Genome Wide Association
DNA	Deoxyribonucleic Acid
NGM	Next Gen Map
BQSR	Base Quality Score Recalibration
GATK	Genome Analysis Tool Kit
VCF	Variant Call File