Uncovering Markers for Honey Production & Defensive Behaviour using Pooled Genome-Wide Data with the Honeybee (*Apis mellifera*)

Stephen Anthony Rose

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

August 2018

ABSTRACT

The honeybee (*Apis mellifera*) has been an important insect for both the study of social insect behaviour and agriculture. Honey production and defensive behaviour are honeybee's two notable and economically valuable traits. Here we perform a genome-wide association study on 925 honeybee colonies from across Canada to elucidate the genetics of these two traits. We find that 168 SNPs for honey production and 41 SNPs for defensive behaviour are significantly associated with their respective phenotypes. Moreover, using genome-wide data, we achieved a predictive performance for honey production of R^2= 27.1% and for defensive behaviour an accuracy of 77.5%. My research shows how genome-wide data can be used both for understanding the genetics of honey production and defensive behaviour in honeybees and for predicting the phenotypes of individual colonies using machine learning techniques.

DEDICATION

I dedicate this thesis to computers, and the +50,000 little fuzzy insects with wings we have killed and the sea of das.

ACKNOWLEDGMENTS

I am very grateful to my supervisor Dr. Amro Zayed without whom this thesis would not be possible. I would like to thank Dr. Sapna Sharma for her statistical expertise. I would also like to thank Matt Betti for helping me explore and understand important concepts in mathematics. I would like to acknowledge Tanushree Tiwari, Harshil Patel, Clement Kent, Isabel Bestard Lorigados, Denis Adigamov and Sarah Wheeler for discussion and patience for all things I did not easily grasp. Finally, I must thank everyone that did all the field work, lab work and the broader Beeomics team that made this possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: Introduction

Honeybees have high economic value as they produce honey and are a critical pollinator in agriculture. Globally, honeybees produce more than 1.8 million tonne theys of honey per year (FAOSTAT 2017). In Canada, honeybees' direct economic contribution comes from producing 37 million kg of honey which amounts to $201 million (number from 2014) (Darrach and Page 2016). Indirectly, via pollination services, the economic value of honeybees is estimated at $4.39 billion per year (Darrach and Page 2016). The direct and indirect benefits that honeybees provide makes them an indispensable component of agriculture and the economy. Salient traits of honeybees are their production and long-term storage of honey (for which they get their name) and their propensity to sting perceived intruders. Honeybee stings are the most unpleasant and undesirable trait they possess, from a beekeepers' perspective. When European honeybees (*Apis mellifera mellifera*) hybridize with the African subspecies *Apis mellifera scutellata,* the hybrids become more likely to exhibit defensive behaviour (Guzmán-Novoa and Page Jr 1994; Hunt et al. 1998). Therefore, the demand to improve honey yield and reduce the concern of a potential increase in honeybees' defensive behaviour is apparent. This could be done with selective breeding programs.

Traditional breeding methods involve selectively breeding animals with the traits that are valued the most (i.e., only the most productive honeybees are bred). In honeybees, selective breeding is done differently than typical livestock and plants since colony traits tend to arise from the actions and traits of many worker bees. Honeybee selective breeding is done by selecting a few of the top performing colonies and grafting virgin queens from them and allowing them to mate with drones openly (Laidlaw and Page 1997). This is not ideal as only the maternal line is controlled. Another significant inefficiency arises from the fact that the queens and drones are an amalgamation of genetically high-quality and low-quality individuals. The uncertainty of genetic quality stems from the phenotypic variance observed being a composite of genetic effects (what is

attempted to be optimized), environmental effects, and random effects (Visscher et al. 2008). Therefore the introduction of genetic markers considerably improved selective breeding by using marker-assisted selection (Oldroyd and Thompson 2006; Mrode 2014). By using genetic markers, marker-assisted selection allows breeders to infer the breeding values (Clark and van der Werf 2013; Mrode 2014). Doing the same, honeybee breeders could infer the breeding values of their colonies directly. The breeding value is the mean additive effects that the queen and drones genes are able to pass down to their offspring (Mrode 2014). The genetic markers used in marker-assisted selection are those shown to be associated with the phenotype of interest (Collard et al. 2005). The first studies employed to discover informative markers for breeding were quantitative trait loci (QTL) studies (Collard et al. 2005).

QTL studies are a way of uncovering associations between genotypes and complex phenotypes (Collard et al. 2005; Miles and Wayne 2008). Typically, two pure-breeding lines for the extremes of a phenotype are generated via selective breeding (i.e. highly defensive vs docile colonies). These lines are then crossed to make an F1 hybrid and, depending on the study's design, that F1 is crossed again to form an F2 generation, or the F1 hybrid is backcrossed with one of the parental lines (Hunt et al. 1998; Mackay et al. 2009). Assays are performed on the offspring of these crosses to measure the phenotype, and then genotyped on many neutral genetic markers. Creating purebred lines requires a level of inbreeding that leaves high levels of linkage disequilibrium (long range genetic correlation) between the genotyped neutral markers and the causal marker. This linkage disequilibrium is exploited to gain knowledge about the ungenotyped causal marker by knowing the degree of association genotyped neutral markers that are linked to the causal mutation have with the phenotype (Collard et al. 2005; Mackay et al. 2009). These QTL analyses produce maps where peaks represent vast stretches of the genome that are associated with the phenotype (Hunt et al. 1998; Miles and Wayne 2008). These maps are low resolution, and therefore significant QTL regions would span many genes (Hunt et al. 1998; Korte and Farlow 2013). As the

cost of sequencing and computing has fallen substantially over the years, this has opened the possibility of sequencing and analysis of whole genomes (Wetterstrand 2018). The current trend is to perform a genome-wide association (GWAS) of all single nucleotide variations or polymorphisms (SNPs) at particular positions in the genome (Miles and Wayne 2008).

Genome-wide association studies have become the best way of uncovering candidate regions of a genome associated with a measured phenotype (Visscher et al. 2017). In many respects, genome-wide association studies and QTL studies are similar as both exploit linkage disequilibrium between the causal mutation and adjacent mutations to associate the phenotype of interest with narrow regions of the chromosome (Korte and Farlow 2013). The two techniques differ importantly. GWAS studies can be used in samples of natural populations and therefore, due to recombination having more time to break linkage disequilibrium, results in much higher resolution maps (Hayes 2013; Korte and Farlow 2013). These maps often have resolutions high enough to identify single genes (Speliotes et al. 2010; Visscher et al. 2017).

The last 10 years of genome-wide association studies have strongly suggested that common SNPs almost entirely govern complex and quantitative phenotypes from a wide range of taxa. Accordingly, I strongly hypothesize that mutations associated with honey production and defensive behaviour will be common variants (Yang et al. 2010b; Korte and Farlow 2013; Visscher et al. 2017). These common genetic variants tend to be additive (Korte and Farlow 2013; Yang et al. 2013). The utility of marker-assisted selection (Collard et al. 2005) combined with the resolution of GWAS (Visscher et al. 2017) are expected to lead to a marked improvement in honeybee breeding.

Honey production and defensive behaviour are colony-level traits that have a genetic basis (Hunt et al. 1995; Breed et al. 2004; Oldroyd and Thompson 2006; Hunt et al. 2007; Koffler et al. 2017). In the following two chapters I used statistical techniques to identify genes potentially

involved in these two traits and how genome-wide genetic variants influence colony defensive behaviour and honey production.

In Chapter 2, I explore the genetics of honey production. Honey is both chemically complex and economically valuable. I found statistically significant overlap with previously discovered pollen foraging QTLs. I found that genes associated with honey production were genes related to neurexin bindings, odorant receptors, acetylcholine receptors, dopamine receptors, and many other neural transmitter related functions. I found that using both genome-wide SNP data and a small number of selected markers one could predict the quantity of human a colony would produce.

Next in Chapter 3, I explored the genetics of defensive behaviour. Honeybees' inclination to sting would-be intruders is remarkable because of its variability, where some colonies do not sting at all, others sting hundreds of times and pursue targets for 100s of meters (Michener 1975; Hunt et al. 1998). I discovered that some North American honeybee genes that were associated with defensive behaviour were also in regions of the genome that segregate in africanised honeybees. I found that genes that were associated with defensive behaviour were related to sensory perception, nervous system regulation and regulation of transcription. Further, I found that using both genome-wide SNP data and a small number of selected markers one could predict with high accuracy whether a colony would sting or not.

CHAPTER 2: Honey Production

2.1 Background

All of the colony's nutritional needs are met by utilizing only two plant products, pollen and nectar. From this, honeybees are able to produce their most valued substance to beekeepers, honey (Bixby 2015). Honeybees, however, cannot live off of honey reserves alone; they need a protein source. The only two food products honeybees will derive all their nutritional needs from are pollen and nectar (Winston 1991). Pollen is the male gametophytes of seed-bearing plants and are honeybees' primary source of protein (Hunt et al. 1995; Scheiner et al. 2004). Nectar is a sugar-rich fluid that plants release to entice pollinating animals like honeybees to assist in transferring pollen to other plants (Winston 1991). Nectar is not directly suitable for long-term storage and therefore honeybees preserve it in the form of honey, which is considerably more stable. To transform nectar into honey, honeybees first regurgitate the collected nectar into a honeycomb cell. The honeybee then reduces the water content of the nectar's by fanning their wings. Once the regurgitated nectar solution becomes dehydrated sufficiently the honeybee seals that honeycomb cell to prevent further changes to the honey. To produce honey honeybees must first forage for nectar. When honeybees take to the wing on their foraging trips, they will encounter a wide variety of potential flowers. The flowers themselves can vary massively in the amount and quality of the nectar they present (Scheiner et al. 2004). Depending on the location a single colony is capable of collecting 5kg of nectar per day and produce over 100kg of honey in a season (Winston 1991). To collect such a quantity of nectar and pollen requires the orchestration and cooperation of many individual worker honeybees, with each honeybee performing their own complex behaviour. Many of these behaviours such as, age onset of foraging, the number of flowers explored, choosing flower sources, the amount of pollen collected, the volume of nectar collected, responsiveness to sucrose quality,

have all been found to be modulated by the bee's genetics (Hunt et al. 1995; Scheiner et al. 2004; Hunt et al. 2007; Koffler et al. 2017).

Honey production and its related traits (pollen and nectar foraging) have been shown to have a heritable genetic component (Hunt et al. 2007; Koffler et al. 2017). Previously held estimates for the heritability of honey production was approximately 25-60% (Table 1)(Koffler et al. 2017), where heritability is the proportion of the phenotype that is passed on to the offspring (Visscher et al. 2008). Although honey being the most valuable product honeybees produce directly, no QTL studies have ever been performed on honey production. QTL studies have been done on its related traits: pollen and nectar foraging. These QTL studies found 4 QTLs, *pln-1*, *pln-2*, *pln-3* and *pln-4* that were found to be associated with foraging behaviour (Hunt et al. 1995; Page Jr et al. 2000; Hunt et al. 2007). Specifically, *pln-1* and *pln-2* QTLs were associated with the amount of pollen collected by workers (Hunt et al. 2007). QTLs *pln-2* and *pln-3* were associated with honeybees ability to detect sugar concentration of collected nectar (Hunt et al. 1995; Hunt et al. 2007). Following studies have found a candidate gene associated with the *pln-4* QTL, *AmFor*. This gene is believed to be associated with the age onset of foraging behaviour in honeybees (Rueppell et al. 2004). Previous studies in this area have set the foundation for subsequent research into exposing unidentified genetic markers associated with honey production.

Are there SNPs that explain the variance of honey production observed across Canada? We aim to discover whether there are markers that are significantly associated with honey production and whether those markers could be used to predict the quantity of honey colonies will produce. I hypothesize there will be at least one SNP associated with honey production since this trait has already been demonstrated to be heritable in honeybees. This hypothesis will be tested for candidate SNPs using genome-wide association techniques. The hypothesis explicitly being tested is as follows.

H_0= The SNP is not associated with honey production.

H_a= The SNP is significantly associated with honey production.

2.2 Methods

2.2.1 Phenotype Collection

Phenotypes and genotypes were collected from a total of 925 colony samples from across Canada (Figure 1; Figure 2; Figure 3). 204 of the samples came from British Columbia, 231 from Alberta, 176 from Manitoba, 155 from Ontario and 158 from Quebec. Collaborators on the Beeomics project collected the phenotypic data. Honey production was estimated in the field by weighing the colonies twice two weeks apart during peak honey flow (July). The difference between the first and second measurement is considered the weight gain and is the target variable in our statistical models.

2.2.2 Sample Dissection

50 worker bees were sampled from each of the 925 colonies across Canada (total of 46300 samples). These samples were stored in 95% ethanol at -80C until they were processed. One leg from each honeybee sample was dissected using fine forceps. The left foreleg was preferentially dissected, however, if that was not possible then the next preferred possible dissection in order was – right foreleg, left midleg and right midleg. The forelegs were preferred to the hind legs due to the risk of pollen particle contamination. Each of colony's 50 corresponding legs was stored in the same tube for DNA extraction.

2.2.3 DNA Extraction and Sequencing

For each of the colony's 50 legs were dipped in liquid nitrogen then subsequently crushed and ground using a pestle. Next, we performed a sample tissue lysis. Tissue lysis was done by adding 350µl of Tissue Lysis Buffer, 20µl of Proteinase K, and heated the samples overnight. DNA extraction was then performed using Mag-Bind Blood and Tissue DNA HDQ 96 Kit (Omega Bio-tek Inc, USA) which was optimized for the KingFisher Flex Purification System (Thermo Fisher Scientific Inc, USA). 75µl of eluent volume was obtained from each of the 925 colony samples. The DNA was quantified using a NanoDrop 2000 Spectrophotometer (Thermo Fisher Scientific Inc, USA). Finally, 0.8% agarose gel electrophoresis assessed the quality of the DNA samples.

After all the colony samples were prepared in-house, the DNA collected was sequenced by McGill University using Illumina HiSeq X. The sequences are then aligned to an *Apis mellifera* reference genome. Finally, an automated pipeline identified single nucleotide polymorphisms (SNPs).

2.2.4 Bioinformatics Pipeline and Quality Filtration

The raw data generated by the sequencing machine must be processed into a form usable for analysis. Collaborators (Harshill Patel and Tanushree Tiwari) and I designed and implemented an automated pipeline in Python (described below) that controls the preprocessing of a batch of samples from the raw sequence data (FASTQ) to the final variant file. Quality control is implemented automatically by the pipeline with built-in checks, assertions, and programs. It takes 8-10 hours to run a sample manually through the pipeline. The automated pipeline takes about 4.9-5.2 hours to process a sample. Moreover, the samples are processed in parallel over four servers at an average rate of 20 samples a day. The median number of SNPs is approximately 1.5 million per colony after all filtration steps. Figure 4 shows the distribution of the number of SNPs in a sample.

The steps in our bioinformatics pipeline are focused on identifying true variants while filtering out spurious variants caused by sequencing or alignment errors.

The first step of the pipeline is Trimmomatic. Trimmomatic uses metadata from the sequenced reads to either trim off the ends of the reads or deletes a read (Bolger et al. 2014). These modifications to the reads were necessary because the Illumina adaptors added for sequencing the genome may be either later called as variants or hamper read alignment.

The next step was NextGenMap (NGM). NGM aligns the reads from the previous step to the correct location on in the honeybee genome by using a honeybee reference genome (Honeybee Sequencing Consortium v4.5) (Sedlazeck et al. 2013). NGM results in a BAM file (compressed sequence alignment map file) that contains are the aligned read data (Li et al. 2009).

Now duplicate reads are identified and removed from the BAM file using Picard's Mark Duplicates (BroadInstitute 2018). A duplicate read is one that was believed to have come from the same unique DNA fragment. These duplicated reads usually arise from the amplification process in sample preparation (BroadInstitute 2018). After the duplicated reads have been removed the quality scores of the read's bases need to be adjusted.

The next step was the Genome Analysis Tool Kit's Base Quality Score Recalibration (BQSR) to recalibrate the base pair quality scores. BQSR uses the quality scores reported by the sequencer, the position of the base in the read, and other statistics to recalculate a quality of the bases (A. et al. 2013). Recalibration is essential because the reported quality for a variety of physical or chemical reasons could overestimate or underestimate the actual quality.

The next step was LoFreq (Wilm et al. 2012). LoFreq does two things simultaneously. First, it identifies variants between the observed data and the reference genome. Second, it automatically filters out low-quality variants without using heuristics or approximations which achieves a variant identifying precision of ~100%. LoFreq yields a variant call file (VCF) that contains a list SNPs and Indels and their locations in the genome for a sample.

Ambiguous SNP calls are removed from the VCF by removing SNPs that are five base pairs up or downstream from an ambiguous position in the genome. Ambiguous positions in the genome are found by genotype calling sequenced honeybee drones as if it were diploid. Since honeybee drones are haploid, SNPs that are called as heterozygous in drones may point to something ambiguous about how the reads are aligned, and SNPs at and near those positions may not be trustworthy.

The final step in the pipeline was the 'Final VCF Filter' which filters SNPs that the other steps of the pipeline missed. Ambiguously aligned reads can create regions of the genome where 10,000's of reads are wrongly mapped to the same location. These incorrect mapping locations results in SNPs with anomalously high variant statistics (depth or quality). The Final VCF Filter removes these SNPs in particular. Lowest quality SNP calls need not be filtered as LoFreq has internal filtering that ensures all SNPs calls pass minimum quality thresholds.

2.2.5 Single Marker Association (GWAS)

In single marker regression, each SNP is tested for association with the phenotype using a Generalised Linear Model (GLM) with the following form (equation 1).

$$y = X\beta + \mu \qquad (1)$$

Where y is the target phenotype; μ is the mean phenotype; X is a matrix of both the SNP's allele frequency data at the position being tested and honeybee colony's yard ID as a covariate to control to environmental differences between yards; β is a vector of the estimate SNP and covariate effects. The SNP's allele frequency data were normalised using the following equation (equation 2) where x is the SNP's allele frequency data.

$$\frac{x_i - \bar{x}}{\sqrt{\bar{x}(1 - \bar{x})}} \qquad (2)$$

After all the SNPs are tested, the resulting p-values were adjusted for multiple testing using local FDR (Efron and Tibshirani 2002). Other *post-hoc* procedures such as the Benjamini-Hochberg procedure (for FDR control) and the Benjamini-Yekutieli procedure (for FDR control under positive regression dependence) were not used because they can be very conservative (Benjamini and Yekutieli 2001; Yi et al. 2015).

2.2.6 Penalised Multi-Marker Association (GWAS)

Standard GWAS methods only test SNPs one at a time. Such methods are standard due to legacy computational constraints which no longer apply (Korte and Farlow 2013). We test all candidate SNPs simultaneously using a standard linear model (assuming is center on 0) where values of β are chosen by minimizing the following elastic net objective function (equation 3 and 4).

$$J(\beta) = \frac{1}{2N} \left|\left|y - X\beta\right|\right|_2^2 + \lambda P(\beta) \qquad (3)$$

$$P(\beta) = \alpha\left|\left|\beta\right|\right|_1 + (1-\alpha)\frac{1}{2}\left|\left|\beta\right|\right|_2^2 \qquad (4)$$

In this case X is a matrix comprised of SNP and covariate data (Li et al. 2011; Yi et al. 2015). The model's coefficients β were subject to both L_1 and L_2 penalties (Tibshirani 1996; Zou and Hastie 2005). This combination of penalties allows for the L_1 term to perform SNP selection while the L_2 allows the model to deal with multicollinearity created from linkage disequilibrium (Yi et al. 2015). 10 fold cross validation was used to determine the magnitude of λ which optimises predictive performance (Friedman et al. 2001; Clarke et al. 2009). The proportion of λ partitioned to the L_1 and L_2 penalty is controlled by α. The entire model's trustworthiness was determined by assessing the significance of the entire model (Barla et al. 2008). Afterward, permutation testing computed p values for all the coefficients in the model so false discoveries are not considered (Yi et al. 2015; Arbet et al. 2017). Modelling all SNPs simultaneously leads to more accurate inference as this decreases the residual variance of the target phenotype (Yi et al. 2015; Arbet et al. 2017). A caveat

when using L_1 regression is that it has been empirically shown that when SNPs are strongly correlated with each other L_1 regression will select only one SNP (Yi et al. 2015). This differs from single marker regression as all markers in linkage disequilibrium will be significant (Yi et al. 2015; Arbet et al. 2017).

2.2.7 Whole Genome Prediction Model

Honey production was predicted using supervised principal components. This technique allows the creation of a regression model that can predict the phenotype using only subsets of the SNPs in the genome that have the highest estimated effect (Bair et al. 2006). The top 2 principal components were used in a linear regression model to predict honey production.

2.2.8 Candidate SNP Prediction Model

Since acquiring whole-genome data may be impractical, models that can infer the phenotype using only a subset of the SNPs are advantageous. The significant SNPs found by penalized multi-marker association were used as the regression features in a GLM model to predict honey production.

2.2.9 Previous QTL and Gene Ontology Analysis

We performed hypergeometric testing to calculate whether the intersection with our putative genes and previous putative genes from past QTL studies were statistically significant or not ($\alpha = 0.05$). We used HymenopteraMine to identify both the *Drosophila melanogaster* homologs and the 1 to 1 orthologs (Elsik et al. 2016). For gene ontology analyses, we again used HymenopteraMine.

2.3 Results

2.3.1 Heritability

Narrow sense heritability (or just heritability) is the proportion of the phenotypic variance that can be explained by the additive genetic effects (equation 5) (Visscher et al. 2008). We used Supervised Principal Components Analysis and 'LD Score Regression' to estimate the heritability. The estimated heritability using Supervised PCA was 27.1% (Table 1). LD Score Regression could not be used to estimate the heritability for honey production because of the median value of β_{SNP} gotten from single marker regression was not 0, which was a program requirement. The heritability estimate matched well with previously published estimates (Table 1) (Koffler et al. 2017).

$$h^2 = \frac{\sigma^2_{Additive}}{\sigma^2_{Phenotype}} \qquad (5)$$

2.3.2 Single Marker Association

No significant nor suggestive results were found for honey production when using single marker association after controlling for the false discovery rate. The most likely explanation for why all associations were non-significant was the lack of power due to low sample size (Figure 5). The lack of power is because there is an *a priori* belief that complex traits like honey production and defensive behaviour are governed by many common small effect variants (Gibson 2010; Yang et al. 2010a; Gibson 2011; Korte and Farlow 2013).

2.3.3 Penalized Multi-Marker Association

Preliminary analysis using this method found 168 SNP positions that were significantly associated with honey production. 284 genes found at and near those SNPs were analyzed with

HymenopteraMine (Table 12) (Elsik et al. 2016). To compare our results with past QTL studies, we converted our genes' new gene identifier codes (amel_OGSv3.2) to the legacy version (amel_OGSv1.0). We found four genes were within previously described pollen hoarding QTLs, which was a statistically significant overlap (p=0.026).

2.3.4 Putative Genes

We mined our putative gene list and their one-to-one orthologues to better understand honey production. Three of the genes that were on the *pln-1* QTL (*GB47783, GB47788, GB47791*) and the last was on *pln-3* (*GB43005*) (Hunt et al. 2007). The genes fruit fly (*Drosophila melanogaster*) orthologues were obtained to further understand how these genes could potentially relate to the phenotype (Table 10). Only the three genes found in pln-1 had one-to-one orthologues, and they were *GB47783* to *dpn*, *GB47788* to *pdm3* and *GB47791* to *CG2121*. Two orthologues *dpr7 (GB40061)* and *dpr9 (GB51181)* are genes that are associated with proboscis extension in response to sugar and chemical sensory perception (Table 10; Figure 6). The implications of these genes will be elaborated in the discussion section.

Next, we explored the one-to-one orthologues to understand more about the novel putative genes. Orthologue *nAChRalpha1* (*GB42850*) is Nicotinic Acetylcholine Receptor α1, which has been shown to confer neonicotinoids resistance in both fruit flies (Perry et al. 2008) and honeybees (Christen and Fent 2017). The implications of how these genes could affect foraging will be elaborated in the discussion section. In fruit fly orthologue *Appl* (*GB48454*) is β amyloid protein precursor-like and was found to be involved in learning, olfactory memory, long-term memory and short-term memory (Goguel et al. 2011; Bourdet et al. 2015). *Neur* (*GB55094*) in flies is involved in the Notch signaling pathway, nervous system development, sensory organ development and long-term memory (Lyne et al. 2007; Tweedie et al. 2008). Phosphofructokinase or *Pfk* (GB50943) is the

enzyme involved in performing the first committing step of glycolysis where it catalyzes the phosphorylation of D-fructose 6-phosphate to fructose 1,6-bisphosphate with ATP. Schippers et al. found that *Pfk* in honeybees correlated with the maturity of foraging worker bees (Schippers et al. 2006; Schippers et al. 2009). Turtle or *tutl* (GB53012) is a gene that in fruit flies is related to flight coordination (Bodily et al. 2001). Furthermore, another orthologue *bdl* or borderless (*GB53013*) a gene involved in photoreceptor activity was shown to be down-regulated by turtle (Chen et al. 2017). Neuroligin 2 or *Nlg2* (GB45987) in fruit flies is a transmembrane protein that interfaces with Neurexin-1, in honeybees it has been shown to be possibly related to foraging behaviour (Biswas et al. 2010). How exactly neuroligin 2 is related to foraging behaviour will be expanded in the discussion section. Finally, Ets96B, Atg16, Bruce and pudgy are a collection of orthologues that are related to starvation, food deprivation and fat metabolism (Tweedie et al. 2008).

To understand the novel putative genes, we performed gene ontology and publication enrichment analyses. HymenopteraMine flagged the 4 genes for having a publication enrichment, *GB53051, GB53052, GB53053, GB53055* (Benjamini Hochberg test FDR correction at p-value 0.05). In adult honeybees, these genes affected the expression of nicotinic subunits $\alpha 2$, $\alpha 7$, $\alpha 8$ and $\beta 1$ in their olfactory neuropiles (Thany et al. 2005; Dupuis et al. 2011). 32 other genes were implicated by 22 separate experimental studies. Next, the genes' fruit fly 82 one-to-one orthologues were analyzed with HymenopteraMine for gene ontology (Elsik et al. 2016). Gene ontology for biological processes revealed a plethora of gene ontology terms related to neuronal and synaptic regulation (Table 4; Table 5). Moreover, gene ontology for molecular function revealed many gene ontology terms related to neurotransmitters (Table 6). For full SNP positions and gene list, see table 2 and table 12.

2.3.5 Predicting Phenotype from Genetic Information

The phenotype was predicted directly from the genetic data in 2 different ways. The first method used to infer honey production was supervised principal components of the whole genome data with linear regression. This supervised principal components whole genome model achieved an R^2 of 27.1% (25.8%-28.3%) (Figure 7). The second method used the significantly associated SNPs found by the penalized multi-marker association model in a GLM with all non-significant SNPs removed. This GLM of the candidate SNPs achieved an R^2 of 35.1%. This high performance is most likely due to either overfitting or overestimating the effect of individual SNPs since the GLM model should not have been able to obtain an R^2 score greater than the supervised principal components model.

2.4 Discussion

2.4.1 Putative Genes

Using penalized multi-marker association to perform genome-wide association on honeybee colonies we found 168 SNP positions and 284 genes that are potentially associated with honey production. Many of the genes could be put into two general categories, memory and sensory genes and metabolic genes.

The set of olfactory, gustatory, long-term and short-term memory related putative genes found suggests the foraging component of honey production is dependent on these characteristics. Associative learning likely links all these characteristics together. Foraging bees learn from features (i.e., olfactory, gustatory and visual cues) from their foraging site (Menzel 2012). Many of our putative genes (GB47788/pdm3, GB47791/CG2121, GB42850/nAChalpha1, GB53051, GB53051, GB53053, GB53055/nAChbeta1) have been shown to expressed in the honeybees' olfactory centers of the brain, antennal lobes and the mushroom bodies (Tichy et al. 2008; Tweedie et al. 2008;

Dupuis et al. 2011). Mushroom bodies and antennal lobes are deeply involved in learning and memory in honeybees (Menzel 2012) and experiments have shown that foraging worker bees undergo an expansion of dendritic branching in these brain regions (Farris et al. 2001). Numerous experiments were able to decrease foraging performance in honeybees by using drugs known to impair learning and memory (such as imidacloprid (Williamson and Wright 2013; Wright et al. 2015)) (Decourtye et al. 2004; Yang et al. 2008; Henry et al. 2012). Some putative genes such as *GB48454/Appl* and *GB45987/Nlg2* could be involved in memory more directly. In fruit flies, *Appl* (β Amyloid Precursor Protein-Like) is required for associative learning and olfactory memory (Goguel et al. 2011; Bourdet et al. 2015) and is highly expressed in the mushroom bodies (Goguel et al. 2011). Even in humans, amyloid precursor protein is critically involved in Alzheimer's disease, a disease marked by both short-term and long-term memory loss (Burns and Iliffe 2009). *Nlg2* (neuroligin 2) has been experimentally shown to participate in learning and memory (Biswas et al. 2010). Results of Biswas et al. experiments suggest that *Nlg2* is involved in keeping the honeybee brain sensitive to afferent sensory input (Biswas et al. 2010). A potential source of pleiotropy exists between gene *Nlg2* and our putative defensive behaviour gene *Nrx-1* since *Nlg2* protein interacts with presynaptic *Nrx-1* in honeybees (Ramírez et al. 2016). Pleiotropy in general between honey production and defensive behaviour is not impossible as there is a phenotypic correlation between the two traits in our study (Figure 8).

The collection of metabolic related genes potentially influences honey production in complex way. Honeybee foraging performance increases as they approach the end of life (Schippers et al. 2006). The proposed explanation for this has been learning from experience and metabolism (Schippers et al. 2006). As with many socials insects, honeybees fat reserves are slowly depleted as they age (Toth and Robinson 2005). Transitioning from a nest-working honeybee to a forager appears to be dictated by internal signals about their own fat reserves (Toth et al. 2005). Many putative genes (*Ets96B, Atg16, Bruce, pdgy*) found were directly related to fatty acid metabolism

and starvation response (Lyne et al. 2007; Tweedie et al. 2008). The paradox arises from the fact the food collected by the forager is not primarily consumed for its own nutritional needs (Hunt et al. 2007). Therefore, it appears that in the evolution of honeybees, genes that trigger responses to starvation (like foraging) have been co-opted for foraging for the colony nutritional needs. Another way metabolism could affect honey bee foraging performance was suggested by one of our putative genes, phosphofructokinase (*pfk*), which is the enzyme that commits hexose molecules such as glucose to glycolysis. Honeybees require glycolysis for aerobic respiration, and mutations in phosphofructokinase potentially modulate the efficiency of glycolysis, this would affect flight performance (Schippers et al. 2006). In fact, Schippers et al. found that phosphofructokinase (along with citrate synthase, hexokinase, pyruvate kinase and cytochrome C oxidase) increases to foraging levels 5-10 days (Schippers et al. 2009), which is around the age honeybees begin to forage (Seeley 1982). Finally, our putative gene *GB47791* fruit fly orthologue (*CG2121*) is involved in skeletal muscle contraction (Tweedie et al. 2008). The interaction between metabolic genes and genes modulating muscle contraction further suggests how flight efficiency could influence honey production.

Overall our results suggest that honey production utilizes a complex gamut of sensory, metabolic, memory and learning related genes. Further research should be done on the metabolic aspect of honey production, and future honeybee GWAS studies should use larger sample sizes and measure precisely how much honey a hive produced in a season rather than the change of colony weight during peak honey flow.

2.4.2 Phenotype Prediction

Predicting honey production with whole-genome and candidate SNP data have great utility for apiarists and bee breeders. The ability to predict an organism's phenotype with an R^2 of 27.1%

is high enough to allow scientists, apiarists and bee breeders to make in breeding or selection decisions before needing to observe the phenotype directly. More economically, the ability to predict the colony's phenotype with an R^2 of 35.1% whilst only using 78 SNPs means that beekeepers and bee breeders could afford to perform marker-assisted selection. Finally, the ability to have the model explain its individual predictions allows scientists to move from a statistical approach to an individualized approach where scientists can come to understand why how two genetically different organisms could end up phenotypically similar.

CHAPTER 3: Defensive Behaviour

3.1 Background

The European honeybee *Apis mellifera mellifera* is widely known to sting a perceived intruder or threat. Compared to its relative the African honeybee *Apis mellifera scutellata*, it is a great deal more docile (Winston 1991). When researchers intended to create a hybrid honeybee that had the warm weather tolerance of the African honeybee and the productivity and gentle disposition of the European honeybee the resultant cross was nothing more than a failure (Winston et al. 1983). The hybridized honeybees (known as Africanised honeybees) were extremely defensive, absconded and swarmed more frequently. Swarming and absconsion may be a nuisance, and the Africanized honeybees' highly defensive behaviour has proven to be lethal (Michener 1975; Franca et al. 1994). Consequently, partically all research on aggression and defensive behaviour in honeybees have focused on Africanized hoenybees.

Honeybee defensive behaviour is a social behaviour involved in the defence of their colony by guarding the hive and stinging (Hunt et al. 2007). This trait is the chief concern of both beekeepers and members of the public (Bixby 2015). This behaviour is most robust near the hive, however, perceived threats can be stung at a distance from the hive. Africanised honeybees have even been shown to chase targets 16 times farther than European honeybees from the hive (Michener 1975; Hunt et al. 2007). In a particular incident, Africanised honeybees chased an experimenter from the colony for over 1 kilometer! In the northern half of North America, honeybees are a mix of predominantly 80% *Apis mellifera ligustica*—the italian honeybee and ~20% *Apis mellifera mellifera* (Harpur et al. 2012) however that could change into the future if the Africanised hybrids expands its range northward.

Our study will analyze only one aspect of defensive behaviour: stinging behaviour measured by the number of stings. Focusing on stinging alone may appear superficially narrow, but in fact, it

requires the integration of many behaviours believed to be involved in defensiveness; production of alarm pheromone, response to alarm pheromone, sensitivity to vibration, reactivity to moving stimuli, etc (Michener 1975; Breed et al. 2004). Previous crosses with highly defensive African-derived honeybees and less defensive European honeybees found at least 15 putative QTLs associated with stinging behaviour (Hunt et al. 1998; Hunt et al. 1999). Further crosses confirmed that 3 of the putative QTLs previously found were indeed associated with defensive behaviour (Hunt et al. 2007). QTL *sting-1* on chromosome 4 was found to be associated with an individual's overall likelihood to sting (Hunt et al. 1998; Hunt et al. 2007). Many of the putative genes found for honeybee defensive behaviour are related to metabolic genes, CNS development, sensory tuning and neural signaling pathways (Hunt 2007). QTLs *sting-2* and *sting-3* contain genes that in *Drosophila melanogaster* are involved in vision and olfaction which suggest that genes are involved in alarm pheromone detection and reacting to moving targets in honeybees (Hunt et al. 2007). Gene expression studies found that metabolic genes were linked to defensive behaviour, that is, when alarm pheromone was presented to honeybees, the more defensive honeybees down-regulated oxidative-phosphorylation genes and up-regulated glycolytic pathway genes (Alaux et al. 2009; Rittschof and Robinson 2013; Chandrasekaran et al. 2015). More recent studies have even shown that parent-specific gene expression potentially modulates honeybee defensive behaviour, that is, when European honeybee queens were mated with Africanized drones the resultant offspring were considerably more defensive (Breed et al. 2004). The modulated regions found were associated with the previously discovered *sting-1* and *sting-2* QTLs (Breed et al. 2004; Hunt et al. 2007). While all these studies give great insight into the genetics of honeybee defensive behaviour, they all suffer from a similar oversight; they all tacitly assume that highly defensive European honeybees are highly defensive due to Africanisation.

Are there SNPs explain the propensity of stinging behaviour observed across Canada? We aim to discover whether there are markers that are significantly associated with defensive

behaviour in a population that is *not* Africanised. This should be able to uncover whether highly defensive European honeybees and highly defensive Africanised honeybees are defensive for similar genetic reasons. We also will uncover whether those markers could be used to predict if a given honeybee colony will be defensive or not. We hypothesize there will be at least one SNP associated with defensive behaviour since this trait has already been demonstrated to be heritable in honeybees. This hypothesis will be tested for candidate SNPs using genome-wide association techniques. The hypothesis specifically being tested is as follows.

H_0= The SNP is not associated with defensive behaviour.

H_a= The SNP is significantly associated with defensive behaviour.

3.2 Methods

3.2.1 Phenotype Collection

See section 2.2.1 before continuing.

Defensive behaviour was measured per colony using a *defensive behaviour assay*. A defensive behaviour assay is performed by suspending a 3 by 3-inch black leather patch in the brood chamber of a bee colony then proceeding to swing the patch for 2 minutes. After 2 minutes the leather patch is removed, and the number of stingers embedded in the leather patch is counted (Hunt et al. 1998). This procedure was repeated and the average of the two measurements were taken. 60% of the colonies did not sting on the defensive behaviour assay and were label as docile. The reminding 40% stung up to 128 times per minute; these colonies were labeled defensive (Figure 9; Figure 10).

3.2.2 Sample Dissection

See section 2.2.2.

3.2.3 DNA Extraction and Sequencing

See section 2.2.3.

3.2.4 Bioinformatics Pipeline and Quality Filtration

See section 2.2.4.

3.2.5 Single Marker Association (GWAS)

See section 2.2.5 before continuing.

Since defensive behaviour is encoded as a categorical trait (0 = docile colony, 1 = defensive colony) a logit link function was used for the GLM rather than a linear link function.

3.2.6 Penalized Multi-Marker Association (GWAS)

See section 2.2.6 before reading this section.

As with single marker association of categorical traits, penalized multi-marker association required using a logit link function. Nothing else about this method needed to be modified for handling categorical features.

3.2.7 Whole Genome Prediction Model

See section 2.2.7 before continuing.

The only modification to the supervised principal components procedure was the top 2 principal components were put in a logistic regression model to predict colony defensive behaviour instead of linear regression.

3.2.8 Candidate SNP Prediction Model

See section 2.2.8 before continuing.

The only modification from the method described in section 2.2.8 was that a GLM with a logit link function was used to predict colony defensive behaviour rather than a linear link function.

3.2.9 Previous QTL and Gene Ontology Analysis

See section 2.2.9.

3.3 Results

3.3.1 Single Marker Association

No significant nor suggestive results were found for either defensive behaviour when using single marker association after controlling for the false discovery rate. The most likely explanation for why all associations were non-significant because of the lack of power due to low sample size (Figure 5). The lack of power is because there is an *a priori* belief that complex traits like defensive behaviour are governed by many common small effect variants (Gibson 2010; Yang et al. 2010a; Gibson 2011; Korte and Farlow 2013).

3.3.2 Penalized Multi-Marker Association

Preliminary analysis using this method for defensive behaviour found 31 SNP positions that were significantly associated with defensive behaviour. 145 genes found at and near those SNP positions were analyzed with HymenopteraMine (Table 13). Our gene list did not have any matches with gene lists from past QTL studies.

3.3.3 Putative Genes

We mined the one-to-one orthologues to understand more about the novel putative genes. HymenopteraMine found 44 one-to-one orthologues. Orthologue *5-HT7* (*GB40005*) is 5-hydroxytryptamine serotonin receptor 7 which is a kind of G-protein coupled receptor that is involved in learning and memory (Tweedie et al. 2008) and in mammals modulate mood disorders (Hayley et al. 2005). CG9747 (GB40659) is desaturase and an oxidoreductase and is therefore potentially involved in pheromone production (Lyne et al. 2007). Many of the orthologues (*GB46757/chn*, *GB49684/bi*, and *GB54477/Egfr*) were involved in eye and wing development (Tweedie et al. 2008). Other orthologues (*GB41523/CG11360*, *GB48636/Rrp46*, *GB49901/bowl*, *GB51608/TAF1C-like*, *GB54174/Sce*, *GB54796/PHDP* and *GB55498/Mitf*) were related with regulating transcription (Tweedie et al. 2008). How all these genes potentially affect defensive behaviour will the elaborated in the discussion section. For the full one-to-one orthologue list see table 11.

We next determined which genes were enriched for publications (sets of genes that appear in the publication at rates greater than you would expect from chance alone). HymenopteraMine flagged two genes, *GB52279/Nrx-1* and *GB52280,* for having publication enrichment (Benjamini Hochberg test FDR correction at p-value 0.05) (Figure 11). The expression levels of *Nrx-1* protein Neurexin-1 was shown to be associated with honeybee sensory processing and associative scent

learning (Biswas et al. 2010). Lastly, gene *GB52317* was found to be involved in honeybee venom and venom glands (GAULDIE et al. 1978; Hider and Ragnarsson 1981; Reinhard and Günther 1984). However how this gene would be related to defensive behaviour is unclear.

Lastly, we used gene ontology and pathway enrichment analysis to investigate whether any of our 145 putative genes share some functions in common. As we did not have enough one-to-one orthologues to do this analysis, we instead used the best homologues. HymenopteraMine found 144 fruit fly homologues using the previous 145 putative honeybee genes. Gene ontology was enriched for many biological processes some nervous system and sensory related terms were for G-protein coupled receptor signaling pathway, serotonin receptor signaling pathway, dopamine receptor signaling pathway and synaptic transmission (Table 7). More nervous system related gene ontology terms were found using molecular function gene ontology; instances were, neurotransmitter receptor activity, dopamine neurotransmitter receptor activity, signaling receptor activity (Table 9). For all significant SNP positions and full putative gene list, see table 3 and table 13 respectively.

3.3.4 Predicting Phenotype from Genetic Information

Defensive behaviour was predicted using two different methods. The first method used to infer defensive behaviour was supervised principal components of the whole genome data with logistic regression. This model achieved a median accuracy of 77.5% on 10-fold cross-validation. The second method used the significantly associated SNPs found by the penalized multi-marker association model in a logistic regression model. There were 2 logistic regression models one had access to the colony's location, and the other did not. The logistic regression model of only the candidate SNPs achieved an accuracy of 72.4% and a receiver operating characteristic area under the curve score (ROC AUC) of 0.948 (Figure 12). When the model had access to location information, an accuracy of 89.9% and a ROC AUC of 0.773 was achieved (Figure 12). Confusion

matrices of both logistic regression models show that both have an easier time confidently classifying honeybee colonies as docile as opposed to stinging (Figure 13; Figure 14). This can be seen by how far to the right prediction probabilities of the correctly classified docile colonies were compared to stinging colonies.

3.4 Discussion

3.4.1 Putative Genes

Using penalized multi-marker association to perform genome-wide association on 832 honeybee colonies we found 41 SNP positions and 145 genes that are potentially associated with defensive behaviour. Most of our putative genes could be partitioned into 4 main categories, genes that could modulate olfactory and visual stimuli, genes involved with eye and wing development, genes involved with DNA binding and regulation of transcription, and genes related to oxidative stress.

The ensemble of genes possibly modulates olfactory and visual signaling suggests that sensitivity to sensory input is a component of colony-level defensive behaviour. 5-hydroxytryptamine (serotonin) receptor 7 (GB40005/5-HT7) is a G-protein coupled receptor associated with mood disorders in mammals (Hayley et al. 2005). It is also known that 5-HT7 activates PKA and adenylyl cyclase which itself causes an increase in cAMP (Hunt et al. 2007). Supporting the involvement of serotonin in defensive behaviour is another of our putative genes *Dhit* (*GB46720*) or double-hit as it inhibits Gαo proteins, which are critical proteins for signaling activation to adenylyl cyclase (Lin et al. 2014). Others have suggested that serotonin levels could affect defensive behaviour (Hunt et al. 2007) this perhaps is unlikely as no studies have been able to correlate serotonin levels with aggression in insects (Kravitz and Huber 2003). Also, whether evolution would have co-opted aggressive behaviour for defensive behaviour is unknown. More

likely serotonin modulates the responsive of honeybees to possible threats. Experiments where serotonin was added to honeybee optic lobes found that honeybees reacted less to moving visual stimuli (Erber and Kloppenburg 1995). Neurexin-1 (*GB52279/Nrx-1*) is a transmembrane synaptic molecule that is associated with visual function and locomotion (Tweedie et al. 2008). Biswas et al. experiments suggest that neurexin-1 is important for visual and olfactory sensory integration (Biswas et al. 2010).

The collection of genes related to the regulation of transcription and DNA binding are surprising as it is not apparent how these are related to defensive behaviour. Some genes such as GB54174/Sce regulate transcription by silencing chromatin via histone H2A ubiquitination (Fritsch et al. 2003; Gutiérrez et al. 2011). Other genes such as GB53852/Sin3A regulate transcription via chromatin and histone binding (Dobi et al. 2014). Studies have suggested that gene regulation can affect colony defensive behaviour by modulating honeybees' sensitivity to alarm pheromone or age or environment (social) cues (Alaux et al. 2009).

The collection of oxidative stress related genes adds on the suggestions of past studies that oxidative stress is related to defensive behaviour (Alaux et al. 2009). *GB55784/mthl1, GB55499/Alp4*, *GB40659/CG9747* and *GB55811/CG4610* are orthologues that were found to reduce oxidative stress in fruit flies (Lyne et al. 2007; Alaux et al. 2009; Gimenez et al. 2013). Methuselah (*mthl1*) is a G-protein coupled receptor whose levels of expression confers varying levels of oxidative stress resistance (Gimenez et al. 2013). In fruit flies, *Alp4* (*GB55499*) expression is anticorrelated with the age of flies (Landis et al. 2004; Radyuk et al. 2012). If their oxidative stress increases with age, then it could explain why soldier honeybees tend to be old (Winston 1991).

Overall, our results suggestions that defensive behaviour is a complex colony level trait that is perhaps more governed by regulating transcription of pathways that potentially modulate honeybee's propensity to sting.

3.4.2 Phenotype Prediction

Predicting defensive behaviour with the candidate SNP data could have great utility for the bee breeders. This is because bee breeders desire docile colonies (Figure 15). One of the reasons driving efforts to control the introgression of Africanized honeybees across North America is the fear of extremely defensive Africanized honeybees. As shown, our ability to predict defensive behaviour accurately shows that we could potentially address the issue of highly defensive colonies directly rather than tangentially by avoiding importing or breeding with potentially Africanized honeybees.

CONCLUSION

This research represents the first genome-wide association analysis on 2 important colony level phenotypes, honey production and defensive behaviour. After gathering novel insights on the genetics of honey production in honeybees, we found that genes involved are likely related to learning, memory, and sensory response. This is also the first genome-wide analysis performed on defensive behaviour where the focus was on European honeybees only. We have also demonstrated that is possible to confidently infer the level of colony defensive behaviour and the quantity of honey produced from genetic data alone. This colony level inference on a subset of genome-wide SNPs opens the possibility for bee breeders to employ marker-assisted selection. Ultimately, the methods presented here, and results found here can be heavily built upon to further our understanding of honeybee genetics and improve the favourability of honeybee colonies.

LIST OF TABLES

Table 1. Comparison between previously estimated narrow-sense heritability and the Beeomics' project measured heritability estimates for *Apis mellifera* traits. 95% credible intervals are in parentheses. Literature estimates were sourced from Koffler et al. (Koffler et al. 2017).

Traits	Literature Heritability (h^2)	Supervised PCA Estimated Heritability (h^2)
Honey Production	25-60%	27.1% (25.9 – 28.3)
Defensive Behaviour	13-43%	NA

NA. Data was not computable.

Table 2. List of all the significantly associated SNP positions for honey production found by penalised multi-marker association ($\alpha = 0.05$).

Scaffold	Position	Reference Allele	Alternate Allele	Estimated Effect Size	p-value
1.2	147812	G	T	-0.33	0.037
1.3	190277	T	G	0.46	0.023
1.6	137232	G	T	-1.35	0.037
1.7	77703	T	C	0.32	0.038
1.9	191614	T	C	-0.38	0.034
1.14	5620	T	C	-0.34	0.019
1.14	368524	A	G	0.11	0.042
1.16	441711	T	C	-0.72	0.017
1.23	1222969	C	T	0.38	0.013
1.23	1283382	A	G	0.31	0.047
1.29	1748146	A	G	0.59	0.009
1.31	491247	A	G	0.56	0.020
1.33	507577	C	T	-0.74	0.019

1.35	684432	T	A	0.07	0.037
1.35	684434	T	C	0.22	0.022
1.37	504443	A	C	-0.18	0.037
1.37	1538246	T	C	-0.14	0.043
1.37	1623205	C	T	-0.46	0.022
1.37	1625098	G	A	-0.14	0.026
1.37	2072867	A	G	0.73	0.006
1.4	88364	C	A	-0.49	0.019
1.4	611345	T	C	-0.54	0.018
1.41	198771	C	T	-0.33	0.029
2.3	16749	A	C	1.74	0.032
2.4	68226	A	G	0.46	0.039
2.6	103790	T	C	0.17	0.037
2.7	408460	G	A	-0.75	0.011
2.9	224844	C	T	0.11	0.014
2.9	226125	T	C	0.01	0.026
2.11	16678	C	T	0.16	0.026
2.11	74102	T	C	-0.30	0.049
2.11	813649	G	A	-0.23	0.047
2.11	1344747	C	G	0.53	0.017
2.11	1858896	C	T	-0.48	0.049
2.15	693169	T	C	0.16	0.018
2.15	693734	C	T	-0.39	0.036
2.15	989394	G	A	0.27	0.043
2.17	193278	T	C	0.26	0.043
2.19	514154	T	C	-0.53	0.032
2.19	1376735	A	G	0.37	0.019
2.19	1376757	G	A	0.26	0.043
3.4	826917	T	C	0.33	0.047
3.5	632022	T	G	0.53	0.038
3.5	832537	C	T	0.57	0.050
3.8	101969	A	G	-0.43	0.012
3.8	154367	T	C	0.43	0.033

3.8	827937	T	C	0.53	0.030
3.9	432050	T	C	-0.19	0.045
3.9	714780	G	A	-0.36	0.039
3.9	1977656	A	G	0.26	0.024
3.9	1998126	A	C	2.45	0.020
3.15	94222	G	A	0.56	0.026
4.9	1439389	C	T	0.33	0.024
4.16	375653	A	G	0.33	0.017
5.2	54012	A	G	-0.07	0.049
5.2	453825	C	G	0.55	0.035
5.2	1250584	A	G	-0.47	0.024
5.8	690875	A	G	0.52	0.021
5.9	15358	A	G	-0.15	0.028
5.9	20723	A	G	-0.74	0.002
5.9	20757	A	G	-0.01	0.042
5.9	430959	C	T	0.26	0.029
5.12	882449	T	C	0.46	0.021
5.14	302704	C	A	0.29	0.017
5.14	820316	G	A	0.53	0.032
5.18	44266	T	C	1.12	0.032
5.2	8942	G	A	0.11	0.023
6.8	17709	T	C	0.11	0.049
6.1	3770	G	T	0.34	0.021
6.1	3783	A	G	0.03	0.044
6.14	47410	C	T	0.06	0.043
6.14	268998	T	C	0.28	0.032
6.15	166706	G	A	-1.25	0.034
6.23	169525	G	A	-0.32	0.039
6.23	613961	T	A	0.25	0.041
6.32	54698	G	A	-0.43	0.029
6.36	316855	G	A	0.37	0.049
6.38	89947	G	A	0.32	0.024
6.38	90817	A	G	0.08	0.017

6.38	91013	G	A	0.00	0.039
6.38	337704	A	G	0.33	0.038
6.38	474734	T	C	-1.05	0.013
7.5	639269	C	T	0.40	0.036
7.9	462932	G	C	-0.49	0.039
7.1	497660	G	A	0.44	0.020
7.17	311757	T	C	0.19	0.048
7.21	536683	G	A	0.25	0.025
7.24	1080159	T	A	0.33	0.043
8.2	207213	G	T	0.25	0.050
8.6	832849	T	G	0.50	0.011
8.7	723144	A	G	0.64	0.024
8.8	534361	T	C	-0.27	0.049
8.8	825803	C	T	0.50	0.038
8.8	887017	G	A	-0.23	0.047
8.9	756278	T	A	0.39	0.036
8.17	398212	C	T	-0.70	0.017
9.4	421562	G	T	-1.55	0.028
9.5	421210	C	G	-0.26	0.023
9.5	421213	G	A	-0.26	0.030
9.1	650453	G	A	0.47	0.033
9.1	695256	A	G	-0.66	0.017
9.1	2344793	T	G	0.42	0.037
9.1	3849263	A	G	0.34	0.040
9.1	3887593	G	A	0.06	0.041
9.12	885422	T	G	1.00	0.004
10.5	43067	A	G	-0.32	0.049
10.7	69684	A	G	-1.71	0.038
10.11	154129	G	A	0.42	0.049
10.2	205110	T	A	0.25	0.018
10.23	1319817	A	G	-0.69	0.009
10.23	1370492	G	A	-0.35	0.029
10.24	179792	C	T	0.84	0.017

10.26	1655994	G	A	0.84	0.005
11.6	1132440	A	G	-1.12	0.043
11.16	488620	G	A	-0.25	0.036
11.18	756269	A	G	-0.04	0.049
11.18	1517914	G	A	0.52	0.019
11.18	4624536	C	G	0.48	0.036
12.4	121448	C	T	0.01	0.050
12.8	100339	G	A	-3.47	0.010
12.8	148250	A	G	-0.13	0.050
12.8	148252	A	G	-0.28	0.027
12.8	158042	T	A	0.26	0.005
12.8	532764	T	C	0.20	0.032
12.16	517128	T	C	-0.49	0.049
12.16	677178	C	T	0.50	0.006
12.17	1505874	T	C	-0.34	0.019
12.17	1857396	C	T	0.30	0.046
13.7	619288	C	T	0.13	0.050
13.7	1222000	A	G	0.57	0.031
13.7	1225531	G	A	-0.19	0.035
13.7	1962974	A	G	-0.36	0.048
13.9	99704	G	A	0.39	0.027
13.1	82552	A	G	-0.62	0.016
13.1	469273	C	T	-0.60	0.024
13.12	1604586	C	T	-0.56	0.039
13.12	1969330	A	G	-0.33	0.039
14.1	475140	C	T	1.33	0.035
14.3	112955	T	A	0.25	0.026
14.3	134035	T	C	0.20	0.040
14.3	177179	A	G	-0.47	0.018
14.5	6665	A	G	0.28	0.046
14.8	144886	C	T	-0.27	0.040
14.9	177534	T	C	-0.50	0.038
14.9	384349	C	A	-0.30	0.031

14.1	402680	C	T	-0.25	0.032
14.1	815528	A	G	-0.25	0.019
14.1	841752	A	G	-0.11	0.046
14.13	729046	A	G	0.73	0.011
14.13	1077281	G	A	0.36	0.021
14.14	7833	T	C	-0.55	0.011
14.14	212010	A	G	-0.41	0.047
14.15	1440470	A	G	-2.89	0.023
15.2	340772	C	T	-0.28	0.023
15.5	541355	G	T	1.17	0.042
15.11	265118	G	A	0.68	0.020
15.14	725546	T	C	0.55	0.021
15.19	344452	T	C	-0.61	0.017
15.19	869469	T	C	0.33	0.033
15.19	1649916	A	G	-0.21	0.042
15.19	2303580	C	T	0.56	0.025
16.6	219317	C	G	0.39	0.040
16.8	337585	C	T	-0.32	0.048
17.3	8326	A	G	-0.07	0.034
17.77	43231	G	T	1.60	0.022
17.144	24255	A	G	-1.65	0.047
17.333	13773	A	G	0.32	0.034
17.1814	2400	T	A	-0.82	0.046

Table 3. List of all the significantly associated SNP positions for defensive behaviour found by penalised multi-marker association ($\alpha = 0.05$).

Scaffold	Position	Reference Allele	Alternate Allele	Estimated Effect Size	p-value
1.3	188872	G	A	0.19	0.0002
1.37	561379	G	A	-0.03	0.0002
1.41	296237	G	A	-0.06	0.0002

1.43	531151	T	G	0.00	0.0004
2.11	452395	G	C	0.08	0.0002
2.18	220479	C	T	0.03	0.0002
2.19	48759	G	A	0.15	0.0002
3.8	545485	T	C	0.07	0.0002
4.13	1727720	C	T	0.04	0.0002
5.2	1261134	T	C	0.22	0.0002
5.9	818558	T	C	0.01	0.0004
5.9	1025099	T	C	-0.08	0.0002
6.2	326438	T	C	-0.08	0.0002
6.13	107426	A	G	-0.13	0.0002
6.32	420686	G	A	0.09	0.0002
6.37	1011828	A	G	0.00	0.0002
7.1	17047	G	A	-0.06	0.0002
7.5	580480	T	C	0.00	0.0004
7.5	641219	G	C	0.17	0.0002
7.12	407752	C	T	-0.03	0.0002
7.21	699558	C	G	0.04	0.0002
7.21	948847	C	T	0.04	0.0002
8.12	294086	G	C	-0.05	0.0002
8.15	62380	T	C	-0.05	0.0002
9.8	622117	T	A	-0.02	0.0002
9.12	1771570	A	G	-0.01	0.0002
10.16	27778	G	A	-0.09	0.0002
10.26	730512	A	G	-0.01	0.0002
11.18	2512832	C	T	0.06	0.0002
11.18	3894811	C	G	0.10	0.0002
12.13	446481	A	T	-0.05	0.0002
12.17	504630	A	C	0.13	0.0002
12.17	555041	C	T	0.02	0.0004
13.2	205784	C	G	0.08	0.0002
13.5	562000	A	G	0.21	0.0002
13.6	93537	T	C	-0.04	0.0002

13.7	25370	C	T	0.00	0.0008
13.12	260995	A	G	0.02	0.0006
14.9	515380	A	G	0.04	0.0002
14.13	498446	T	C	0.07	0.0002
15.16	129573	T	C	0.06	0.0002

Table 4. Honey production gene ontology results for biological processes from the *D. melanogaster* one-to-one orthologues performed using HymenopteraMine. Only the significant terms after multiple test correction using Benjamini-Hochberg procedure are shown ($\alpha = 0.05$). P-values reported as 0 are less than 10^{-5}.

Gene Ontology Term	**p-value**	**ID**
Response to external stimulus	0.00484	GO:0009605
Biological regulation	0.00699	GO:0065007
Cell-cell adhesion	0.00900	GO:0098609
Single organismal cell-cell adhesion	0.01021	GO:0016337
Single organism cell adhesion	0.01067	GO:0098602
Homophilic cell adhesion via plasma membrane adhesion molecules	0.01105	GO:0007156

Table 5. Honey production gene ontology results for cellular processes from the *D. melanogaster* one-to-one orthologues performed using HymenopteraMine. Only the significant terms after multiple test correction using Benjamini-Hochberg procedure are shown ($\alpha = 0.05$).

Gene Ontology Term	**p-value**	**ID**
Intrinsic component of plasma membrane	$<10^{-5}$	GO:0031226
Cell periphery	$<10^{-5}$	GO:0071944
Plasma membrane	$<10^{-5}$	GO:0005886
Integral component of plasma membrane	$<10^{-5}$	GO:0005887
Plasma membrane part	0.000187	GO:0044459

Table 6. Honey production gene ontology results for molecular processes from the D. melanogaster homologues performed using HymenopteraMine. Only the significant terms after multiple test correction using Holm-Bonferroni procedure are shown ($\alpha = 0.05$). P-values reported as 0 are less than 10^{-5}.

Gene Ontology Term	**p-value**
glucuronosyltransferase activity	0.0000
UDP-glycosyltransferase activity	0.0000
neurotransmitter binding	0.0000
neurotransmitter receptor activity	0.0000
ammonium ion binding	0.0000
heme binding	0.0000
oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	0.0000
tetrapyrrole binding	0.0000
transferase activity, transferring hexosyl groups	0.0000
iron ion binding	0.0000
inorganic cation transmembrane transporter activity	0.0000
excitatory extracellular ligand-gated ion channel activity	0.0000
transferase activity, transferring glycosyl groups	0.0000
metal ion transmembrane transporter activity	0.0000
sodium ion transmembrane transporter activity	0.0000
cation transmembrane transporter activity	0.0000
cation channel activity	0.0000
ion binding	0.0000
ligand-gated cation channel activity	0.0000
monovalent inorganic cation transmembrane transporter activity	0.0000
G-protein coupled amine receptor activity	0.0000
transmitter-gated ion channel activity	0.0000
transmitter-gated channel activity	0.0000
cation binding	0.0000
neurotransmitter:sodium symporter activity	0.0000
secondary active transmembrane transporter activity	0.0000
oxidoreductase activity	0.0000
ion transmembrane transporter activity	0.0000
neurotransmitter transporter activity	0.0000
extracellular ligand-gated ion channel activity	0.0000
potassium ion transmembrane transporter activity	0.0000
solute:sodium symporter activity	0.0000
acetylcholine receptor activity	0.0000
acetylcholine binding	0.0000

substrate-specific transmembrane transporter activity	0.0000
ion channel activity	0.0000
kainate selective glutamate receptor activity	0.0000
substrate-specific channel activity	0.0000
gated channel activity	0.0000
transmembrane transporter activity	0.0000
channel activity	0.0000
passive transmembrane transporter activity	0.0000
potassium channel activity	0.0000
substrate-specific transporter activity	0.0000
acetylcholine-gated cation-selective channel activity	0.0000
postsynaptic neurotransmitter receptor activity	0.0000
neurotransmitter receptor activity involved in regulation of postsynaptic membrane potential	0.0000
transmitter-gated ion channel activity involved in regulation of postsynaptic membrane potential	0.0000
cation:amino acid symporter activity	0.0000
signaling receptor activity	0.0000
transmembrane signaling receptor activity	0.0000
extracellular-glutamate-gated ion channel activity	0.0000
receptor activity	0.0000
molecular transducer activity	0.0000
flavin adenine dinucleotide binding	0.0000
transporter activity	0.0000
calcium:cation antiporter activity	0.0000
solute:cation symporter activity	0.0000
Gi/o-coupled serotonin receptor activity	0.0000
symporter activity	0.0000
calcium ion transmembrane transporter activity	0.0001
ligand-gated ion channel activity	0.0001
ligand-gated channel activity	0.0001
oxidoreductase activity, acting on CH-OH group of donors	0.0001
transmembrane receptor activity	0.0001
active transmembrane transporter activity	0.0002
calcium, potassium:sodium antiporter activity	0.0002
sodium ion binding	0.0002
organic acid transmembrane transporter activity	0.0003
carboxylic acid transmembrane transporter activity	0.0003
ligand-gated calcium channel activity	0.0003
signal transducer activity	0.0003
metal ion binding	0.0005
calcium channel activity	0.0005
neurexin family protein binding	0.0005
dopamine neurotransmitter receptor activity	0.0006
dopamine binding	0.0006
catecholamine binding	0.0006
G-protein coupled serotonin receptor activity	0.0006
serotonin binding	0.0006

serotonin receptor activity	0.0006
divalent inorganic cation transmembrane transporter activity	0.0008
solute:cation antiporter activity	0.0009
cation:cation antiporter activity	0.0009
antiporter activity	0.0010
Gq/11-coupled serotonin receptor activity	0.0011
oxoglutarate:malate antiporter activity	0.0011
oxidative phosphorylation uncoupler activity	0.0011
sodium channel activity	0.0012
thioredoxin-disulfide reductase activity	0.0012
protein-disulfide reductase activity	0.0012
ionotropic glutamate receptor activity	0.0012
amino acid transmembrane transporter activity	0.0016
coenzyme binding	0.0018
glutamate-gated calcium ion channel activity	0.0018
monooxygenase activity	0.0019
oxidoreductase activity, acting on a sulfur group of donors, disulfide as acceptor	0.0022
cofactor binding	0.0024
amine binding	0.0024
transition metal ion binding	0.0026
carboxylic acid binding	0.0026
organic acid binding	0.0026
glutamate binding	0.0030
malate transmembrane transporter activity	0.0037
adrenergic receptor activity	0.0039
oxaloacetate transmembrane transporter activity	0.0039
neutral amino acid transmembrane transporter activity	0.0039
glutamate receptor activity	0.0039
transcription regulatory region sequence-specific DNA binding	0.0054
oxidoreductase activity, acting on a sulfur group of donors, NAD(P) as acceptor	0.0059
potassium ion antiporter activity	0.0059
amino acid binding	0.0060
transcription factor activity, RNA polymerase II distal enhancer sequence-specific binding	0.0066
dicarboxylic acid transmembrane transporter activity	0.0088
sodium:amino acid symporter activity	0.0088
organic acid:sodium symporter activity	0.0088
thiosulfate transmembrane transporter activity	0.0088
potassium ion binding	0.0088
alkali metal ion binding	0.0088
sequence-specific double-stranded DNA binding	0.0090
anion transmembrane transporter activity	0.0099
small molecule binding	0.0110
RNA polymerase II regulatory region sequence-specific DNA binding	0.0141
dipeptidase activity	0.0161
RNA polymerase II regulatory region DNA binding	0.0162
G-protein coupled receptor activity	0.0166

sequence-specific DNA binding	0.0168
transcription regulatory region DNA binding	0.0190
regulatory region DNA binding	0.0195
regulatory region nucleic acid binding	0.0195
voltage-gated cation channel activity	0.0196
double-stranded DNA binding	0.0205
RNA polymerase II distal enhancer sequence-specific DNA binding	0.0223
acid phosphatase activity	0.0228
phosphorelay sensor kinase activity	0.0233
N,N-dimethylaniline monooxygenase activity	0.0233
protein histidine kinase activity	0.0233
myosin light chain kinase activity	0.0233
somatostatin receptor activity	0.0233
dopamine transmembrane transporter activity	0.0233
dopamine:sodium symporter activity	0.0233
tyramine receptor activity	0.0233
glycine transmembrane transporter activity	0.0233
neuroligin family protein binding	0.0233
protein binding involved in cell adhesion	0.0233
protein binding involved in cell-cell adhesion	0.0233
anion binding	0.0308
enhancer sequence-specific DNA binding	0.0308
ion antiporter activity	0.0316
nucleic acid binding transcription factor activity	0.0330
transcription factor activity, sequence-specific DNA binding	0.0330
enhancer binding	0.0334
succinate transmembrane transporter activity	0.0352
organic anion transmembrane transporter activity	0.0405
protein disulfide oxidoreductase activity	0.0407
C4-dicarboxylate transmembrane transporter activity	0.0495

Table 7. Defensive behaviour gene ontology results for biological processes from the D. melanogaster homologues performed using HymenopteraMine. Only the significant terms after multiple test correction using Holm-Bonferroni procedure are shown ($\alpha = 0.05$). P-values reported as 0 are less than 10^{-5}.

Gene Ontology Term	**Corrected p-value**
adenylate cyclase-modulating G-protein coupled receptor signaling pathway	0.0000
G-protein coupled receptor signaling pathway, coupled to cyclic nucleotide second messenger	0.0000
unsaturated fatty acid biosynthetic process	0.0000
phospholipase C-activating G-protein coupled receptor signaling pathway	0.0000

unsaturated fatty acid metabolic process	0.0000
long-chain fatty acid biosynthetic process	0.0000
serotonin receptor signaling pathway	0.0000
G-protein coupled receptor signaling pathway	0.0000
long-chain fatty acid metabolic process	0.0000
dopamine receptor signaling pathway	0.0000
adenylate cyclase-inhibiting serotonin receptor signaling pathway	0.0001
G-protein coupled serotonin receptor signaling pathway	0.0001
adenylate cyclase-activating G-protein coupled receptor signaling pathway	0.0001
adenylate cyclase-inhibiting G-protein coupled receptor signaling pathway	0.0005
cAMP-mediated signaling	0.0009
dephosphorylation	0.0012
octopamine or tyramine signaling pathway	0.0014
adenylate cyclase-activating dopamine receptor signaling pathway	0.0032
synaptic transmission, dopaminergic	0.0038
negative regulation of transcription from RNA polymerase II promoter	0.0047
cyclic-nucleotide-mediated signaling	0.0078
cardiocyte differentiation	0.0086
adrenergic receptor signaling pathway	0.0090
adenylate cyclase-activating adrenergic receptor signaling pathway	0.0090

Table 8. Defensive behaviour gene ontology results for cellular processes from the D. melanogaster homologues performed using HymenopteraMine. Multiple test correction was done using the Holm-Bonferroni procedure. P-values reported as 0 are less than 10^{-5}.

Gene Ontology Term	**Corrected p-value**
cell surface	0.0007
integral component of plasma membrane	0.0422

Table 9. Defensive behaviour gene ontology results for molecular processes from the D. melanogaster homologues performed using HymenopteraMine. Only the significant terms after multiple test correction using Holm-Bonferroni procedure are shown ($\alpha = 0.05$). P-values reported as 0 are less than 10^{-5}.

Gene Ontology Term	**Corrected p-value**
G-protein coupled amine receptor activity	0.0000
alkaline phosphatase activity	0.0000

ammonium ion binding	0.0000
maltose alpha-glucosidase activity	0.0000
alpha-1,4-glucosidase activity	0.0000
G-protein coupled receptor activity	0.0000
alpha-glucosidase activity	0.0000
glucosidase activity	0.0000
neurotransmitter receptor activity	0.0000
stearoyl-CoA 9-desaturase activity	0.0000
acyl-CoA desaturase activity	0.0000
oxidoreductase activity, acting on paired donors, with oxidation of a pair of donors resulting in the reduction of molecular oxygen to two molecules of water	0.0000
dopamine neurotransmitter receptor activity	0.0000
G-protein coupled serotonin receptor activity	0.0000
serotonin receptor activity	0.0000
dopamine binding	0.0000
catecholamine binding	0.0000
dopamine neurotransmitter receptor activity, coupled via Gs	0.0000
serotonin binding	0.0001
amine binding	0.0003
phosphatase activity	0.0014
phosphoric ester hydrolase activity	0.0015
transmembrane signaling receptor activity	0.0048
tyramine receptor activity	0.0057
signaling receptor activity	0.0193
adrenergic receptor activity	0.0226
transmembrane receptor activity	0.0307

Table 10. Honey production one-to-one *Drosophila melanogaster* orthologue list. N=82.

Gene ID	**Orthologue**	**Orthologue ID**
GB40061	dpr7	FBgn0053481
GB40077	pdgy	FBgn0027601
GB40118	Gad1	FBgn0004516
GB40119	Faa	FBgn0016013
GB40907	CG8399	FBgn0034067
GB40963	Sox15	FBgn0005613
GB41562	DIP-zeta	FBgn0051708
GB41786	grn	FBgn0001138
GB42592	Vsx2	FBgn0263512
GB42728	para	FBgn0264255
GB42850	nAChRalpha1	FBgn0000036

GB42894	DAAM	FBgn0025641
GB43015	Mrtf	FBgn0052296
GB43304	Cad87A	FBgn0037963
GB43429	icln	FBgn0029079
GB43470	Upf2	FBgn0029992
GB43602	CG32082	FBgn0052082
GB43759	Atg16	FBgn0039705
GB43909	PMCA	FBgn0259214
GB44588	Bruce	FBgn0266717
GB45047	Prosbeta2	FBgn0023174
GB45120	qua	FBgn0003187
GB45486	Nckx30C	FBgn0028704
GB45530	kon	FBgn0032683
GB45618	mib1	FBgn0263601
GB45970	CadN	FBgn0015609
GB46022	Coq2	FBgn0037574
GB46051	CG9288	FBgn0260464
GB46404	Syx17	FBgn0035540
GB46739	mtd	FBgn0013576
GB47082	CG8097	FBgn0030660
GB47118	Cad89D	FBgn0038439
GB47783	dpn	FBgn0010109
GB47788	pdm3	FBgn0261588
GB47791	CG2121	FBgn0033289
GB48331	CG42541	FBgn0260658
GB48453	CG30022	FBgn0050022
GB48454	Appl	FBgn0000108
GB48665	Calx	FBgn0013995
GB48699	RpL11	FBgn0013325
GB49391	CG3328	FBgn0034985
GB49684	bi	FBgn0000179
GB49924	Samuel	FBgn0032330
GB50045	CG13436	FBgn0034532

GB50100	CG6136	FBgn0038332
GB50101	CG9990	FBgn0039594
GB50156	Ets96B	FBgn0039225
GB50157	CG34404	FBgn0085433
GB50184	Meltrin	FBgn0265140
GB50527	Dip-B	FBgn0000454
GB50734	CG13698	FBgn0036773
GB50769	a	FBgn0000008
GB50800	DIP-beta	FBgn0259245
GB50837	RhoGAP102A	FBgn0259216
GB50943	Pfk	FBgn0003071
GB50944	TRAM	FBgn0040340
GB51181	dpr9	FBgn0038282
GB51630	fas	FBgn0000633
GB51809	Mnt	FBgn0023215
GB51836	CG2144	FBgn0033187
GB51838	CG42361	FBgn0259707
GB52082	inaE	FBgn0261244
GB52091	su(f)	FBgn0003559
GB52266	Fur2	FBgn0004598
GB52742	pio	FBgn0020521
GB52743	enc	FBgn0004875
GB52756	CG42249	FBgn0259101
GB52910	Oamb	FBgn0024944
GB53012	tutl	FBgn0010473
GB53013	bdl	FBgn0028482
GB53055	nAChRbeta1	FBgn0000038
GB53340	beta-Spec	FBgn0250788
GB53682	CG10019	FBgn0031568
GB53701	Dscam3	FBgn0261046
GB54127	sei	FBgn0003353
GB54537	sli	FBgn0264089
GB54987	Nlg2	FBgn0031866

GB55094	neur	FBgn0002932
GB55539	sstn	FBgn0036476
GB55591	CG30069	FBgn0050069
GB55704	mbo	FBgn0026207
GB55715	ct	FBgn0004198

Table 11. Defensive behaviour one-to-one *Drosophila melanogaster* orthologues. N = 43.

Gene ID	**Orthologue**	**Orthologue ID**
GB40659	CG9747	FBgn0039754
GB40670	ko	FBgn0020294
GB41522	MBD-like	FBgn0027950
GB41523	CG11360	FBgn0039920
GB42271	CG12502	FBgn0035171
GB42465	CG10050	FBgn0037492
GB43245	Pde1c	FBgn0264815
GB45015	nesd	FBgn0032848
GB46206	Orc2	FBgn0015270
GB46720	Dhit	FBgn0028743
GB46757	chn	FBgn0015371
GB48005	5-HT7	FBgn0004573
GB48007	tn	FBgn0265356
GB48028	Clk	FBgn0023076
GB48636	Rrp46	FBgn0037815
GB49684	bi	FBgn0000179
GB49901	bowl	FBgn0004893
GB49902	SpdS	FBgn0037723
GB50563	Osi5	FBgn0037413
GB50564	Osi6	FBgn0027527
GB50565	Osi7	FBgn0037414
GB50652	Cct5	FBgn0010621
GB50690	didum	FBgn0261397
GB51607	CG15111	FBgn0034419

GB51608	TAF1C-like	FBgn0034631
GB52021	pico	FBgn0261811
GB52279	Nrx-1	FBgn0038975
GB53531	loh	FBgn0032252
GB53849	dsh	FBgn0000499
GB53852	Sin3A	FBgn0022764
GB53861	CG3021	FBgn0040337
GB53862	Nep4	FBgn0038818
GB54133	CG13857	FBgn0038958
GB54174	Sce	FBgn0003330
GB54175	CG12880	FBgn0046258
GB54469	CG4447	FBgn0035980
GB54477	Egfr	FBgn0003731
GB54478	CG6321	FBgn0036117
GB54796	PHDP	FBgn0025334
GB55498	Mitf	FBgn0263112
GB55499	Alp4	FBgn0016123
GB55509	CG13458	FBgn0036479
GB55784	mthl1	FBgn0030766
GB55810	CG9967	FBgn0031413
GB55811	CG4610	FBgn0034735

Table 12. Putative honey production genes. N = 284.

Gene ID	Gene Length (bp)	Chromosome Location (Start)	Chromosome Location (End)	Chromosome
GB40004	141	20365036	20365176	1
GB40005	2237	20352728	20354964	1
GB40061	122225	5134793	5257017	13
GB40077	10370	4529831	4540200	13
GB40118	20864	3868453	3889316	13
GB40119	2540	3889670	3892209	13
GB40144	3893	4453077	4456969	13
GB40157	5189	5240409	5245597	13

GB40520	74274	3004786	3079059	8
GB40521	53042	3101667	3154708	8
GB40907	17252	4669410	4686661	5
GB40931	4147	4689837	4693983	5
GB40963	4213	5845640	5849852	1
GB40964	311	5850202	5850512	1
GB41319	11773	13516945	13528717	5
GB41326	2027	13529945	13531971	5
GB41388	8600	1638651	1647250	9
GB41389	1326	1636026	1637351	9
GB41558	1482	3634084	3635565	2
GB41562	21779	3684320	3706098	2
GB41738	4328	10123432	10127759	14
GB41739	8995	10129982	10138976	14
GB41786	18879	8094758	8113636	8
GB41787	943	8218598	8219540	8
GB41991	240	11888449	11888688	8
GB41992	38102	11924809	11962910	8
GB42253	66569	1451399	1517967	10
GB42256	351	1497120	1497470	10
GB42268	70923	2990493	3061415	14
GB42269	30674	2946456	2977129	14
GB42349	1549	16112286	16113834	6
GB42386	5157	16123028	16128184	6
GB42443	405	10000040	10000444	7
GB42461	8935	9961371	9970305	7
GB42523	6374	2615337	2621710	7
GB42524	1671	2622660	2624330	7
GB42592	52982	8289621	8342602	9
GB42643	4432	6815828	6820259	9
GB42727	10451	5113546	5123996	9
GB42728	40431	5069614	5110044	9
GB42729	292	5068730	5069021	9

GB42850	65008	6702798	6767805	9
GB42893	90	8260480	8260569	9
GB42894	16076	8272485	8288560	9
GB42938	240	2623011	2623250	10
GB42947	3467	2611327	2614793	10
GB43005	83312	9648150	9731461	1
GB43015	97749	9540547	9638295	1
GB43288	1206	5753553	5754758	7
GB43304	98222	5656537	5754758	7
GB43429	3351	2849623	2852973	2
GB43470	4089	2852374	2856462	2
GB43596	8395	3390477	3398871	14
GB43597	141	3384854	3384994	14
GB43601	20582	3545471	3566052	14
GB43602	35093	3588079	3623171	14
GB43696	4260	4887447	4891706	14
GB43709	6715	5317882	5324596	14
GB43710	1548	5325181	5326728	14
GB43740	3868	4910936	4914803	14
GB43741	153	4909673	4909825	14
GB43743	13463	4886714	4900176	14
GB43751	68418	2016396	2084813	9
GB43759	147831	1833897	1981727	9
GB43818	3801	6355383	6359183	8
GB43885	8288	6694314	6702601	8
GB43886	334	6693809	6694142	8
GB43909	19130	6414684	6433813	8
GB43910	1421	6405203	6406623	8
GB43917	3804	6350400	6354203	8
GB44548	8872	9564141	9573012	5
GB44549	6144	9557376	9563519	5
GB44588	17812	10015173	10032984	5
GB44589	3969	10042050	10046018	5

GB45047	1863	10425028	10426890	11
GB45072	3891	9674207	9678097	11
GB45073	22785	9643253	9666037	11
GB45120	5002	10421306	10426307	11
GB45289	1414	13521379	13522792	11
GB45290	460	13532235	13532694	11
GB45486	44681	21982437	22027117	1
GB45530	28539	22047320	22075858	1
GB45609	5924	5678354	5684277	16
GB45618	103614	5571948	5675561	16
GB45681	1960	14321807	14323766	5
GB45682	1564	14312826	14314389	5
GB45970	13239	6233104	6246342	16
GB45971	19313	6247101	6266413	16
GB46022	2168	6203689	6205856	6
GB46051	614	6206632	6207245	6
GB46164	25649	1513681	1539329	15
GB46165	5454	1507373	1512826	15
GB46342	149860	702135	851994	8
GB46404	10290	13252816	13263105	1
GB46405	118635	13360018	13478652	1
GB46584	225067	7616931	7841997	2
GB46587	405	7875741	7876145	2
GB46588	2209	7879989	7882197	2
GB46593	166126	7541042	7707167	2
GB46721	1968	5229336	5231303	5
GB46722	2661	5212633	5215293	5
GB46738	941	4797427	4798367	5
GB46739	92520	4803871	4896390	5
GB46956	31280	2958502	2989781	3
GB46957	29501	3001502	3031002	3
GB46964	4752	3115151	3119902	3
GB46965	40510	3120223	3160732	3

GB47032	273	11477678	11477950	3
GB47035	60413	11481424	11541836	3
GB47082	1824	434746	436569	15
GB47083	61396	371102	432497	15
GB47118	21412	906714	928125	12
GB47119	576	885330	885905	12
GB47241	200659	7768218	7968876	11
GB47242	49352	7692763	7742114	11
GB47371	143915	16180480	16324394	1
GB47519	168	16216783	16216950	1
GB47714	371	2177313	2177683	1
GB47715	383	2174287	2174669	1
GB47782	615	6826125	6826739	13
GB47783	5514	6728416	6733929	13
GB47788	174630	6380834	6555463	13
GB47791	8291	6366923	6375213	13
GB47915	222	1015281	1015502	10
GB47917	120	984119	984238	10
GB48096	17886	12169487	12187372	7
GB48182	8394	12086723	12095116	7
GB48331	40579	7130781	7171359	10
GB48453	9448	7059681	7069128	10
GB48454	44055	7070570	7114624	10
GB48458	1010	7127367	7128376	10
GB48478	3438	1561497	1564934	14
GB48479	660	1644817	1645476	14
GB48665	89973	7698059	7788031	5
GB48666	204	7682681	7682884	5
GB48699	8870	369813	378682	2
GB48700	1602	390457	392058	2
GB49045	2868	431795	434662	14
GB49046	141429	210742	352170	14
GB49053	37528	2241922	2279449	3

GB49140	84	2205688	2205771	3
GB49259	5893	25745072	25750964	1
GB49268	39788	25179493	25219280	1
GB49272	48747	25230477	25279223	1
GB49290	11053	25730293	25741345	1
GB49390	4210	8493536	8497745	12
GB49391	17168	8443499	8460666	12
GB49395	1725	8340891	8342615	12
GB49396	3853	8332031	8335883	12
GB49683	10769	904630	915398	7
GB49684	92044	917607	1009650	7
GB49740	2570	1708206	1710775	2
GB49741	1348	1596291	1597638	2
GB49851	354	7588428	7588781	13
GB49855	333	8053649	8053981	13
GB49924	161512	7940150	8101661	13
GB49925	186218	7606995	7793212	13
GB49991	3274	8314620	8317893	15
GB50045	2774	7638693	7641466	15
GB50100	723	6358266	6358988	15
GB50101	32664	6319367	6352030	15
GB50156	16609	6856486	6873094	15
GB50157	36777	6874047	6910823	15
GB50184	38440	7642194	7680633	15
GB50238	2158	8312732	8314889	15
GB50341	16228	2933563	2949790	1
GB50342	99219	2969698	3068916	1
GB50402	149099	6922028	7071126	4
GB50483	138	6994753	6994890	4
GB50527	18436	4714348	4732783	15
GB50528	387	4713516	4713902	15
GB50702	552	869647	870198	2
GB50703	288	809332	809619	2

GB50721	207334	5687223	5894556	2
GB50734	39949	5335105	5375053	2
GB50755	501	4826559	4827059	2
GB50769	18698	4017139	4035836	2
GB50772	8489	4009263	4017751	2
GB50773	320	4104115	4104434	2
GB50800	107551	4699370	4806920	2
GB50837	28075	5302458	5330532	2
GB50860	10491	5999241	6009731	2
GB50943	9007	9910517	9919523	10
GB50944	2490	9919797	9922286	10
GB51180	192	2297822	2298013	6
GB51181	15411	2300645	2316055	6
GB51538	210	24414834	24415043	1
GB51548	57211	23968964	24026174	1
GB51549	300	23927038	23927337	1
GB51550	8096	23908741	23916836	1
GB51576	518	22990644	22991161	1
GB51609	44633	22809222	22853854	1
GB51630	312317	24367201	24679517	1
GB51808	7355	7616109	7623463	1
GB51809	25083	7625586	7650668	1
GB51836	3844	7992580	7996423	1
GB51838	2609	7989326	7991934	1
GB51934	208	18833007	18833214	1
GB51947	149429	18774109	18923537	1
GB51988	69434	10873067	10942500	12
GB52081	99	10538863	10538961	12
GB52082	42099	10540936	10583034	12
GB52091	6030	10877098	10883127	12
GB52179	3621	4172881	4176501	6
GB52180	357	4163862	4164218	6
GB52184	2776	4377537	4380312	6

GB52185	4022	4434960	4438981	6
GB52266	29829	4304860	4334688	1
GB52269	10302	4345346	4355647	1
GB52279	328254	1755808	2084061	5
GB52295	360	1077571	1077930	5
GB52296	772	1036419	1037190	5
GB52299	19328	712100	731427	5
GB52301	585	605353	605937	5
GB52319	147	1840022	1840168	5
GB52450	2162	8845045	8847206	2
GB52451	48313	8863178	8911490	2
GB52742	68462	7538528	7606989	14
GB52743	15086	7523120	7538205	14
GB52756	4707	7714473	7719179	14
GB52757	17311	7721267	7738577	14
GB52910	14802	3160697	3175498	15
GB52920	8610	3148270	3156879	15
GB52934	228	11297763	11297990	6
GB52938	3006	11680263	11683268	6
GB52940	1292	11711256	11712547	6
GB52945	128581	11232040	11360620	6
GB53012	49470	12119085	12168554	4
GB53013	8917	12171684	12180600	4
GB53051	9814	1234258	1244071	14
GB53052	1926	1258826	1260751	14
GB53053	21691	1285511	1307201	14
GB53055	2302	1326905	1329206	14
GB53214	83359	4989522	5072880	1
GB53215	222	5097019	5097240	1
GB53340	18906	10124166	10143071	9
GB53408	830	10132242	10133071	9
GB53545	2998	14596313	14599310	6
GB53547	6472	14587968	14594439	6

GB53673	21581	6963245	6984825	3
GB53682	64138	6669322	6733459	3
GB53691	585	6669411	6669995	3
GB53701	100962	6997112	7098073	3
GB53744	2925	8205154	8208078	3
GB53745	237	8231221	8231457	3
GB53772	201	3623916	3624116	7
GB53773	89551	3530688	3620238	7
GB53854	44785	5678755	5723539	13
GB53856	17209	5731765	5748973	13
GB54115	414	6357383	6357796	14
GB54116	5864	6411411	6417274	14
GB54126	9948	6717598	6727545	14
GB54127	28536	6738494	6767029	14
GB54274	1095	7414177	7415271	10
GB54275	978	7421688	7422665	10
GB54471	13369	2833122	2846490	6
GB54472	1515	2849007	2850521	6
GB54535	30782	4763257	4794038	8
GB54537	31778	5030449	5062226	8
GB54575	590	5155869	5156458	10
GB54612	5648	5166709	5172356	10
GB54944	189	26961606	26961794	1
GB54987	45656	26897328	26942983	1
GB55078	4140	2701557	2705696	12
GB55084	3713	2325658	2329370	12
GB55085	282	2200805	2201086	12
GB55086	352	2181198	2181549	12
GB55087	219	2322275	2322493	12
GB55094	13395	2681614	2695008	12
GB55227	2504	2869693	2872196	11
GB55228	13180	2874663	2887842	11
GB55479	1390	14982113	14983502	2

GB55539	3150	14978528	14981677	2
GB55591	26297	14103015	14129311	2
GB55592	3005	14098293	14101297	2
GB55704	57787	17671382	17729168	6
GB55711	2107	17671148	17673254	6
GB55713	1385	17897558	17898942	6
GB55714	4763	17915375	17920137	6
GB55715	109931	18202130	18312060	6
GB55791	145530	4056937	4202466	3
GB55792	369	4056343	4056711	3
GB55818	1272	4775723	4776994	3
GB55819	449	4791211	4791659	3

Table 13. Putative defensive behaviour genes. N = 145.

Gene ID	Gene Length (bp)	Chromosome Location (Start)	Chromosome Location (End)	Chromosome
GB40658	4396	7195714	7200109	12
GB40659	1853	7212393	7214245	12
GB40660	1518	7227217	7228734	12
GB40670	55048	7233412	7288459	12
GB40671	16786	7165018	7181803	12
GB41415	46346	8291864	8338209	6
GB41416	1189	8597968	8599156	6
GB41521	120997	8347937	8468933	6
GB41522	2273	8344209	8346481	6
GB41523	49623	8280050	8329672	6
GB42267	282	3153176	3153457	14
GB42268	70923	2990493	3061415	14
GB42269	30674	2946456	2977129	14
GB42270	162	2938145	2938306	14
GB42271	15952	2920445	2936396	14
GB42437	120	9605863	9605982	7
GB42438	4480	9612391	9616870	7

GB42463	131291	9558530	9689820	7
GB42464	1379	9539872	9541250	7
GB42465	3263	9517529	9520791	7
GB43245	120420	3956415	4076834	6
GB43247	11658	4035127	4046784	6
GB43248	8587	4021529	4030115	6
GB43249	8123	4000683	4008805	6
GB43250	2430	3981786	3984215	6
GB44919	6195	12799172	12805366	11
GB44920	474	12798188	12798661	11
GB44921	281	12797627	12797907	11
GB44922	219	12796851	12797069	11
GB44923	6356	12789195	12795550	11
GB45014	3609	11461914	11465522	11
GB45015	2511	11458327	11460837	11
GB45016	1942	11455361	11457302	11
GB45017	128930	11291423	11420352	11
GB45154	1852	11461002	11462853	11
GB46193	8363	1160170	1168532	15
GB46194	980	1098203	1099182	15
GB46196	8799	1086943	1095741	15
GB46205	2082	1091158	1093239	15
GB46206	2798	1099739	1102536	15
GB46718	34578	5329813	5364390	5
GB46720	9844	5234035	5243878	5
GB46721	1968	5229336	5231303	5
GB46757	16726	5250853	5267578	5
GB46758	324	5292238	5292561	5
GB46874	267	2792729	2792995	3
GB46939	27912	2723465	2751376	3
GB46940	6491	2757403	2763893	3
GB46941	456	2777468	2777923	3
GB46942	1065	2783005	2784069	3

GB48003	10699	3028108	3038806	13
GB48004	201	2948157	2948357	13
GB48005	33658	2905512	2939169	13
GB48006	5046	2900213	2905258	13
GB48007	23032	3063418	3086449	13
GB48024	5380	3424537	3429916	1
GB48025	3817	3407532	3411348	1
GB48026	8026	3394442	3402467	1
GB48027	8839	3372158	3380996	1
GB48028	5861	3344276	3350136	1
GB48196	2544	423425	425968	10
GB48197	6179	446311	452489	10
GB48635	1324	7829825	7831148	5
GB48636	1272	7846371	7847642	5
GB48637	164252	7862320	8026571	5
GB48656	2402	7843061	7845462	5
GB48657	745	7838704	7839448	5
GB49680	6523	770181	776703	7
GB49681	450	784610	785059	7
GB49682	1394	803444	804837	7
GB49683	10769	904630	915398	7
GB49684	92044	917607	1009650	7
GB49685	300	1043148	1043447	7
GB49686	62403	1078911	1141313	7
GB49710	3286	1083616	1086901	7
GB49874	249	9400070	9400318	13
GB49899	62513	9333608	9396120	13
GB49900	17059	9313591	9330649	13
GB49901	6210	9291851	9298060	13
GB49902	1889	9283364	9285252	13
GB50403	186	6866345	6866530	4
GB50404	1628	6853177	6854804	4
GB50405	35169	6797133	6832301	4

GB50406	112325	6675060	6787384	4
GB50477	1079	6882582	6883660	4
GB50561	3264	4536296	4539559	15
GB50562	2895	4544599	4547493	15
GB50563	1092	4551244	4552335	15
GB50564	1010	4559834	4560843	15
GB50565	1770	4566633	4568402	15
GB50652	3525	2778630	2782154	13
GB50653	6316	2748791	2755106	13
GB50654	7948	2740368	2748315	13
GB50690	14220	2762519	2776738	13
GB50691	627	2776898	2777524	13
GB51576	518	22990644	22991161	1
GB51578	2901	22790212	22793112	1
GB51607	4962	22781890	22786851	1
GB51608	3523	22787457	22790979	1
GB51609	44633	22809222	22853854	1
GB52020	29132	9595210	9624341	12
GB52021	11931	9574256	9586186	12
GB52022	9528	9561943	9571470	12
GB52049	261	9560851	9561111	12
GB52050	546	9590042	9590587	12
GB52279	328254	1755808	2084061	5
GB52280	7583	1741774	1749356	5
GB52317	823	1748544	1749366	5
GB52318	423	1752546	1752968	5
GB52319	147	1840022	1840168	5
GB53531	40191	14909103	14949293	6
GB53567	3700	14902686	14906385	6
GB53568	951	14907666	14908616	6
GB53569	643	14976466	14977108	6
GB53570	16179	15003657	15019835	6
GB53849	5691	5942658	5948348	13

GB53850	826	5940878	5941703	13
GB53852	12365	5911517	5923881	13
GB53861	1708	5928973	5930680	13
GB53862	5722	5932185	5937906	13
GB54133	999	6944331	6945329	14
GB54134	3297	6974894	6978190	14
GB54135	404	6989423	6989826	14
GB54174	3104	6979159	6982262	14
GB54175	22769	6948567	6971335	14
GB54469	855	3057554	3058408	6
GB54470	4170	3004756	3008925	6
GB54476	1779	2968484	2970262	6
GB54477	31369	3019642	3051010	6
GB54478	5732	3051394	3057125	6
GB54794	5354	9850696	9856049	8
GB54795	1386	9849003	9850388	8
GB54796	16899	9830543	9847441	8
GB54817	1988	9788464	9790451	8
GB54818	25133	9800255	9825387	8
GB55498	8346	15432053	15440398	2
GB55499	3468	15443160	15446627	2
GB55500	2506	15449847	15452352	2
GB55501	2023	15452668	15454690	2
GB55509	1209	15447673	15448881	2
GB55784	142723	4467864	4610586	3
GB55785	126	4460751	4460876	3
GB55810	603	4456737	4457339	3
GB55811	2719	4457568	4460286	3
GB55812	27170	4572580	4599749	3

LIST OF FIGURES

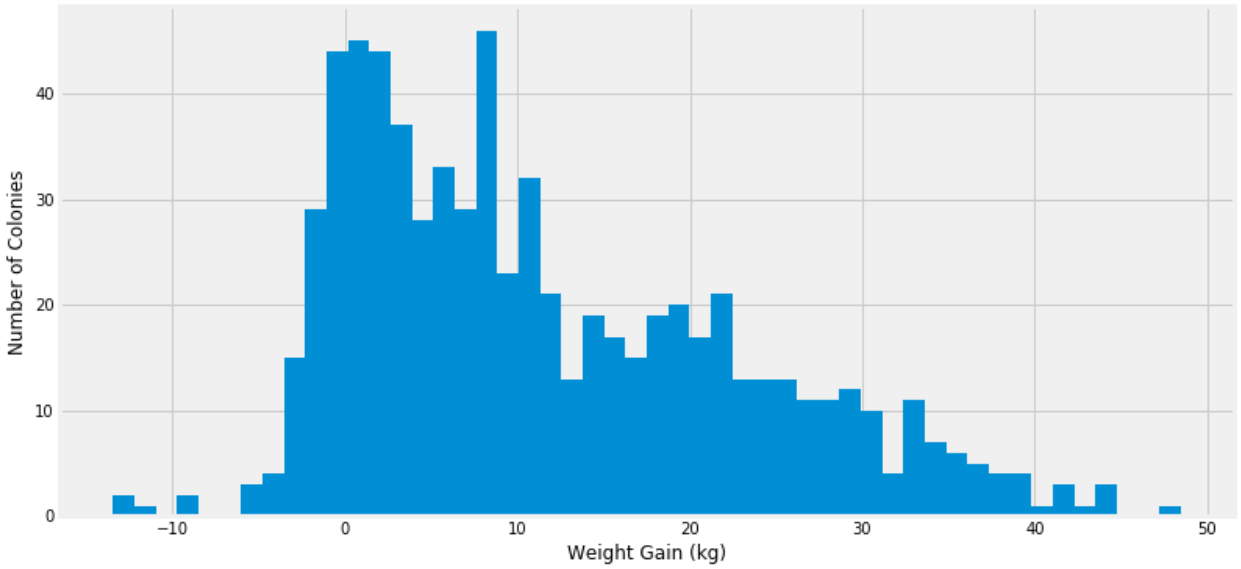

Figure 1. Histogram of honey production in honeybee colonies measured by weight gained during peak honey flow. N=712.

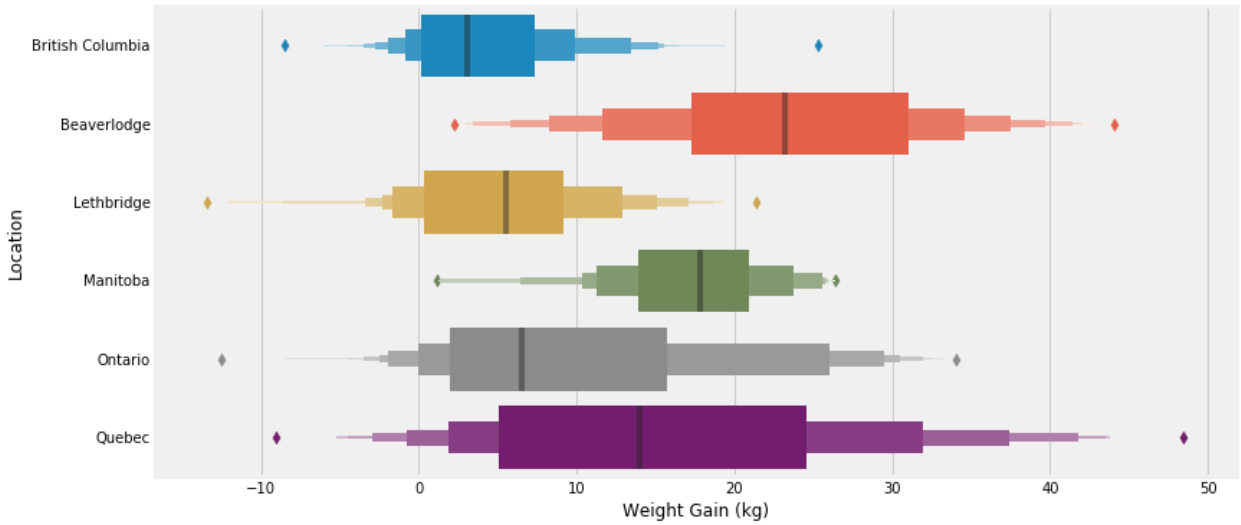

Figure 2. Letter-value plots of the Beeomics project's honeybee colonies honey production by location measured by weight gained during peak honey flow. N=712.

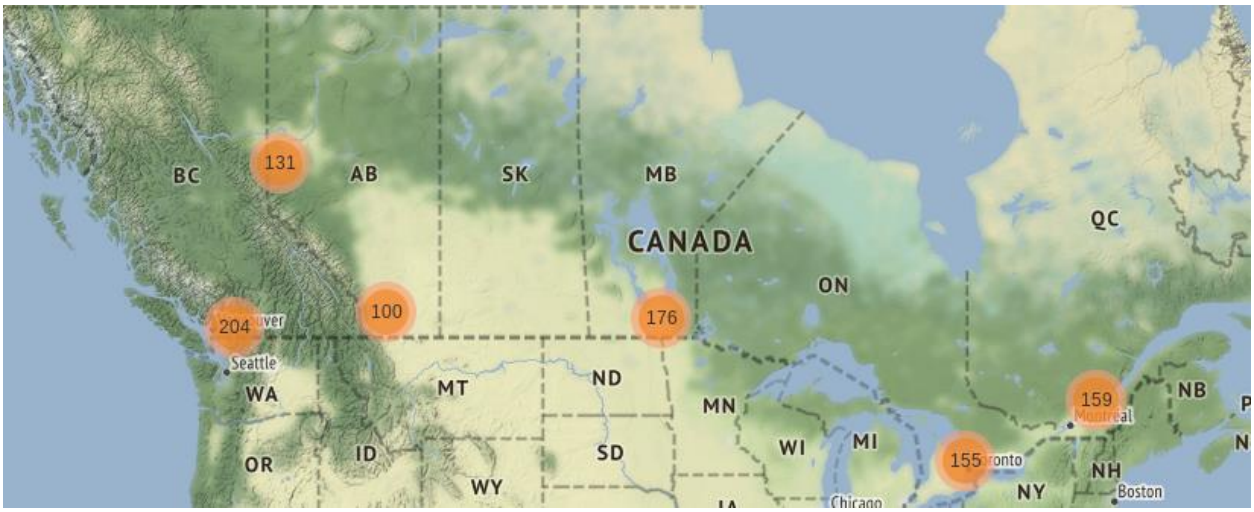

Figure 3. Map displaying the distribution of all 925 colonies in the Beeomics project.

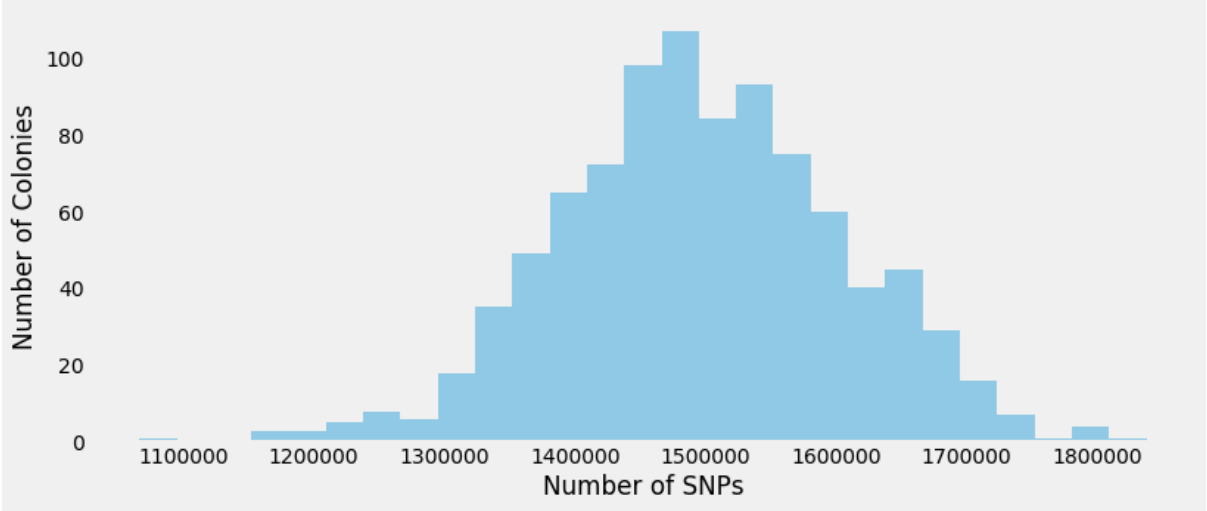

Figure 4. Distribution of the number of SNPs in each colony after running through the pipeline.

(Minimum = 1,066,264, Median = 1,491,197, Maximum = 1,837,944).

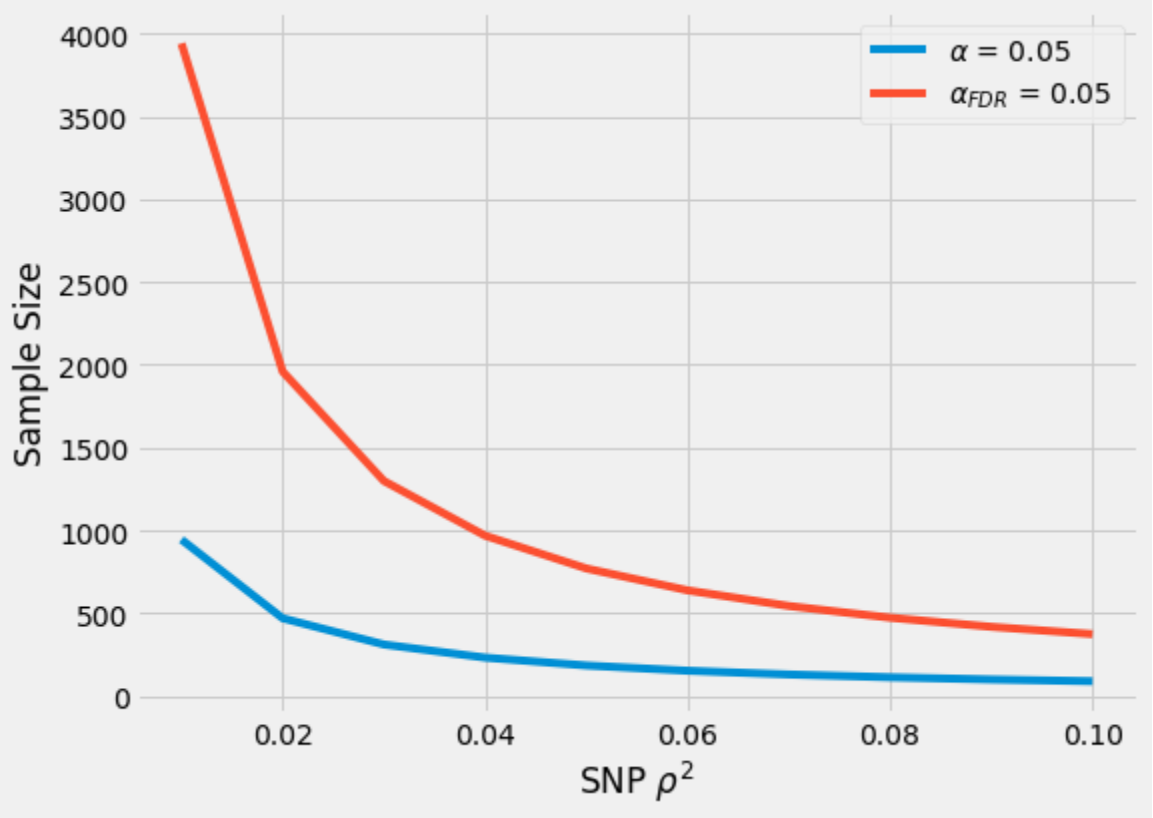

Figure 5. Regression power analysis is depicting the required sample sizes for detected a SNP 80% of the time at a given estimated R^2. The blue line shows the sample sizes required for $\alpha = 0.05$. The red line shows the sample sizes required for an FDR adjusted α. Power analysis was performed with G* Power (Faul et al. 2007).

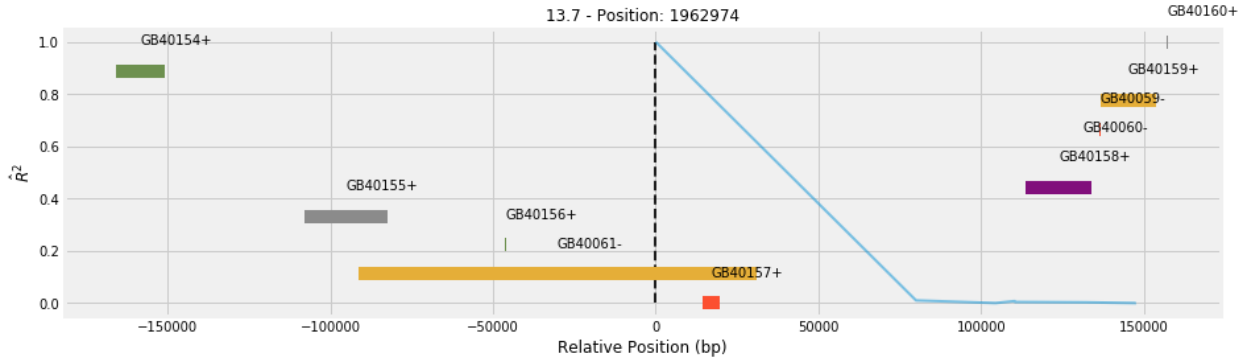

Figure 6. This is a gene plot centred on a SNP that explains honey production. Gene GB40061 that the SNP lies within is orthologue *dpr7.* The SNP is located on Scaffold 13.7 at position 1,962,974. The coloured horizontal bars depict genes. The blue line shows the degree of linkage disequilibrium of the model's selected SNP with nearby SNPs, $\hat{R}^2$ of 1 are SNPs in perfect linkage disequilibrium with the selected SNP and 0 means no linkage.

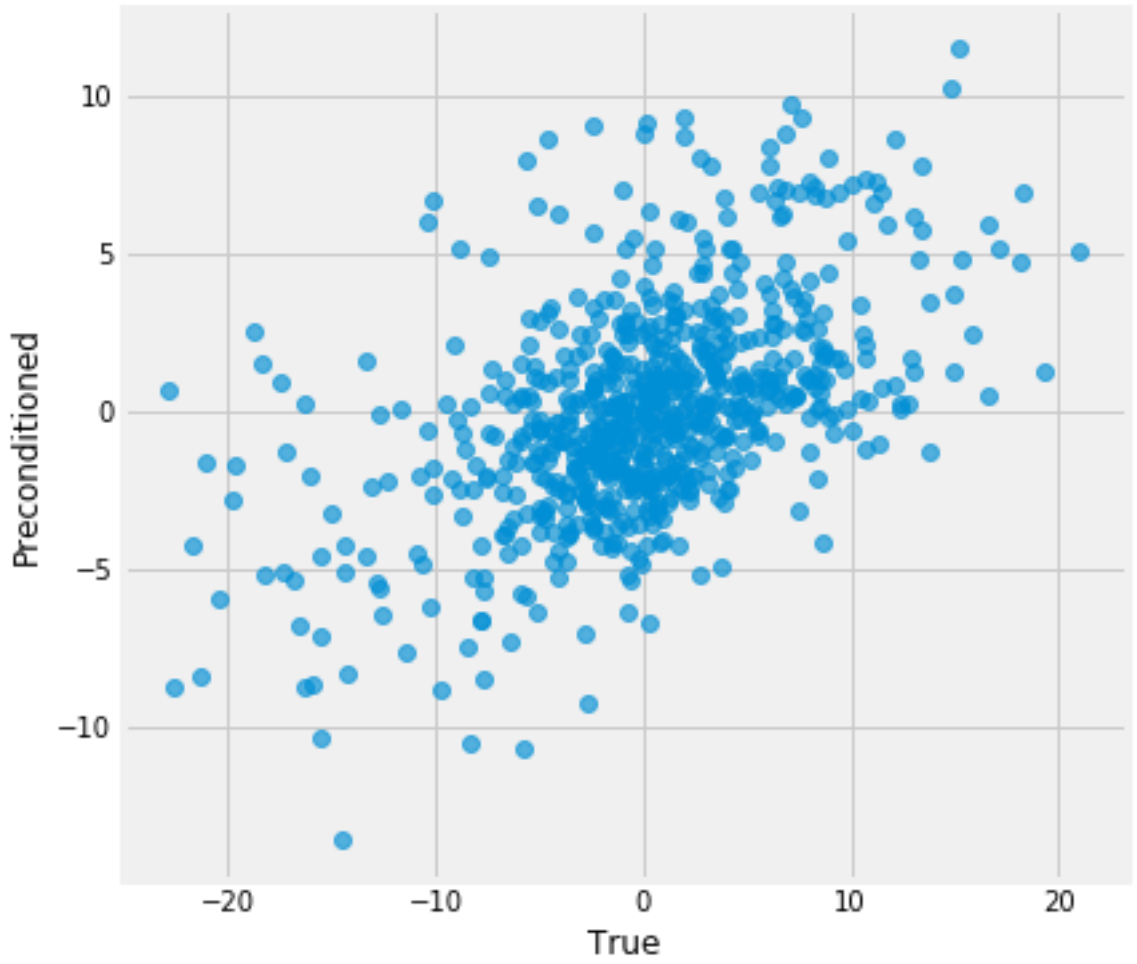

Figure 7. A scatterplot showing the predictive performance of the whole genome supervised principal components honey production. The y axis represents the estimated value for weight gain. Every bee yard's mean is centred on zero. $R^2 = 0.2711\ (0.2588, 0.2834)$.

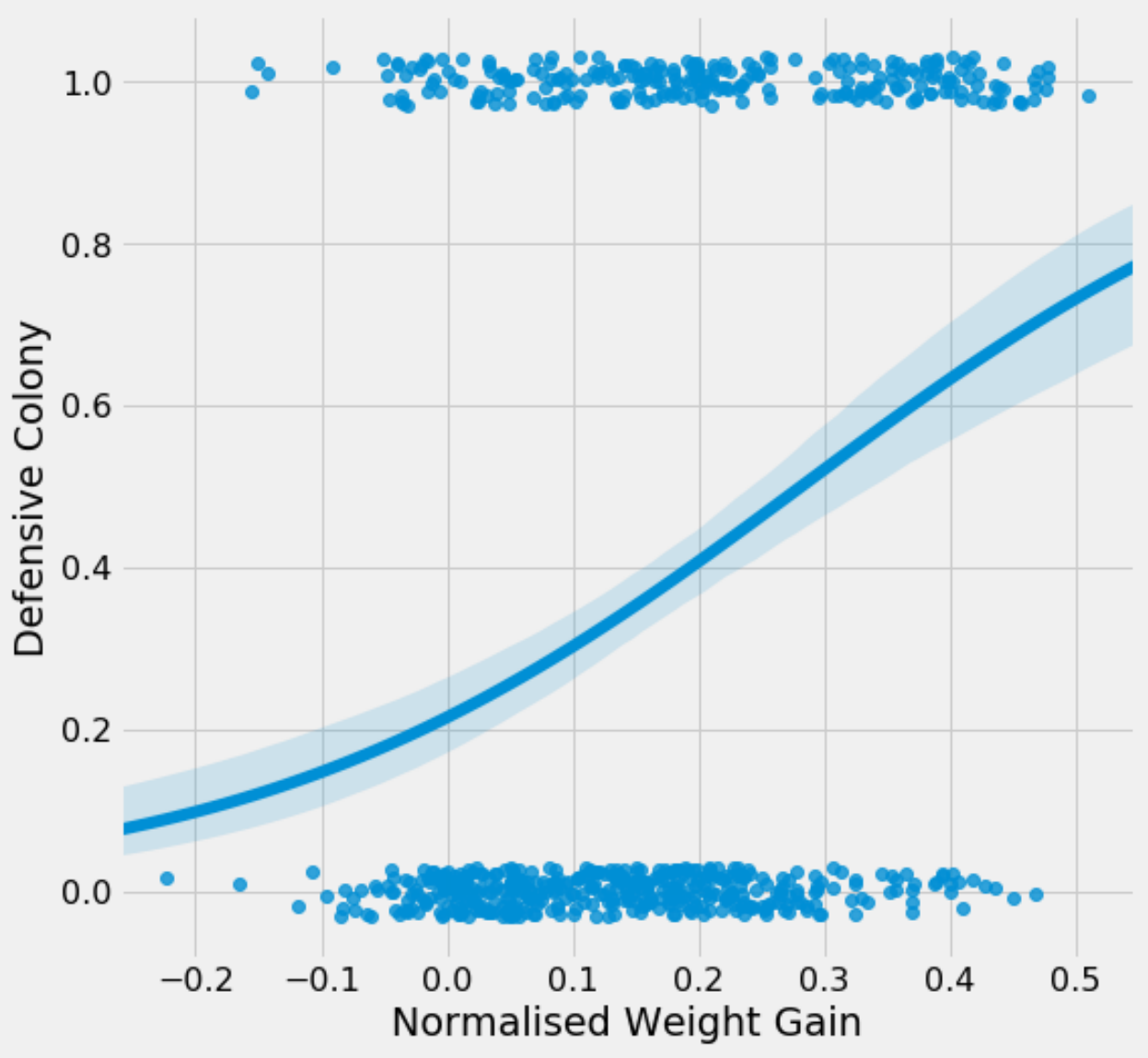

Figure 8. Logistic regression plot between colony defensive behaviour and normalized weight gain (honey production). Spearman Rank correlation coefficient = 0.27 (p-value = 8×10^{-13}).

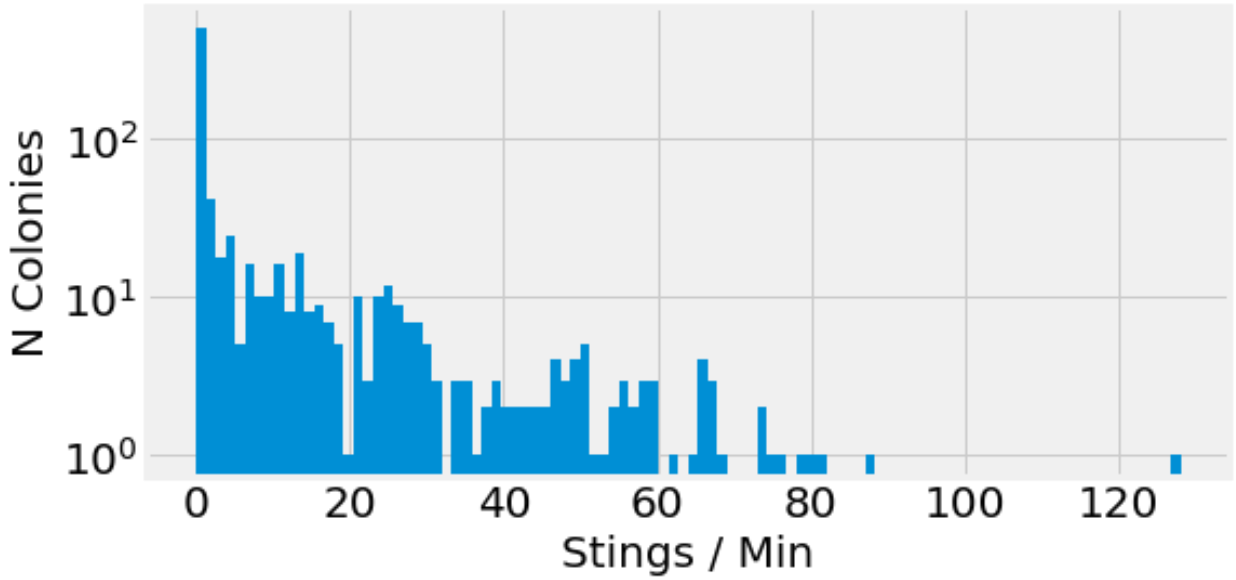

Figure 9. A histogram showing the distribution of stings per minute our honeybee colonies did during the defensive behaviour assay. N = 840.

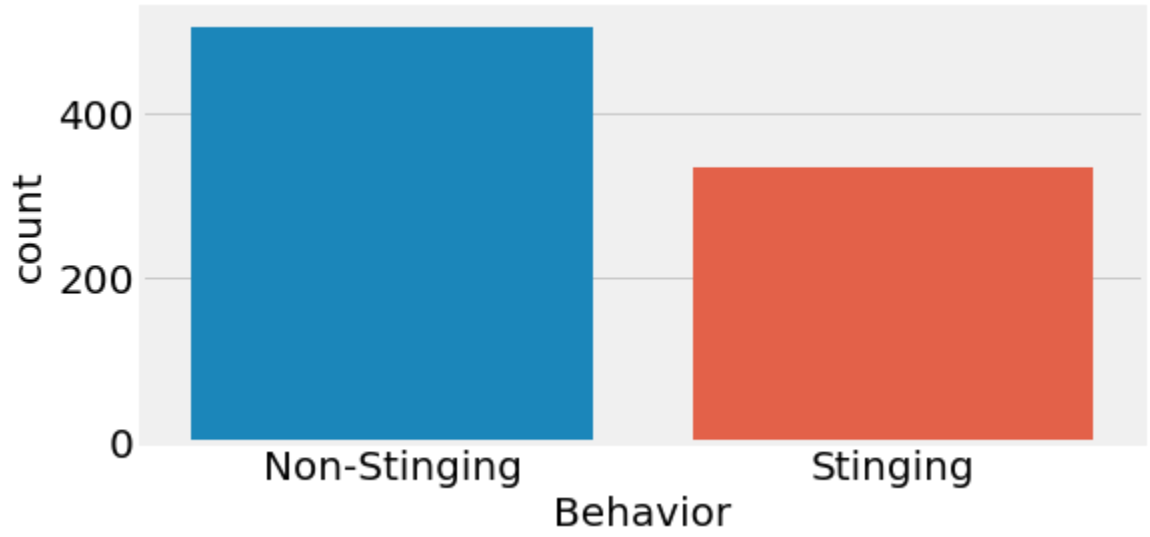

Figure 10. Count plot is representing the number of colonies that did not sting during the defensive behaviour assay and stinging (stung at least once per minute). N=840.

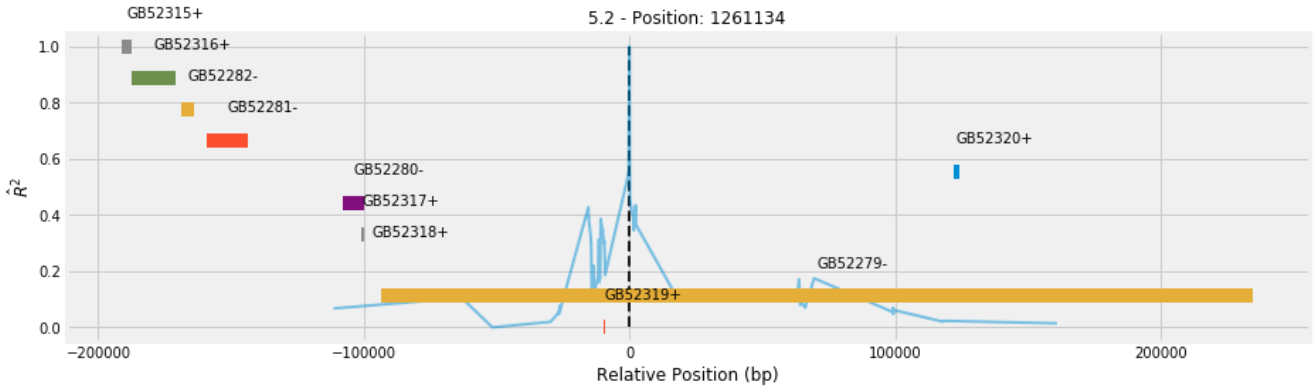

Figure 11. This is a gene plot centred on 1 of the two genes flagged by HymenopteraMine for publication enrichment on Scaffold 5.2 at position 1,261,134. The coloured horizontal bars depict genes. The blue line shows the degree of linkage disequilibrium of the model's selected SNP with nearby SNPs, $\hat{R}^2$ of 1 are SNPs in perfect linkage disequilibrium with the selected SNP and 0 means no linkage.

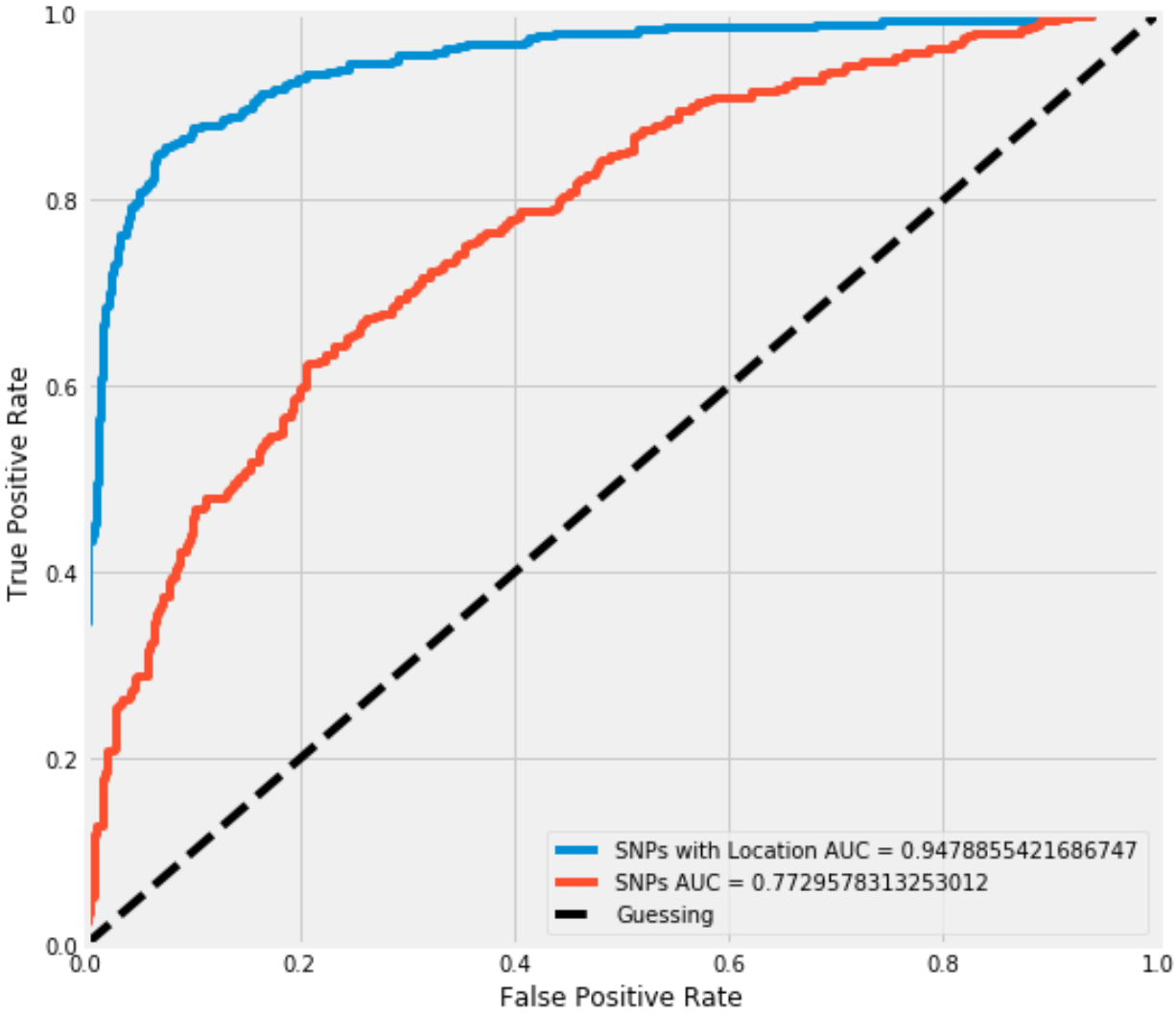

Figure 12. A plot is showing the predictive performance of models for classifying whether a colony will be defensive using receiver operating characteristic curves. Baseline (guessing) area under the curve score (AUC) is 0.5, and a perfect AUC is 1.0. The model's accuracy with access to location = 89.9%. Accuracy with only SNP information = 72.4%. Baseline accuracy = 60.1%.

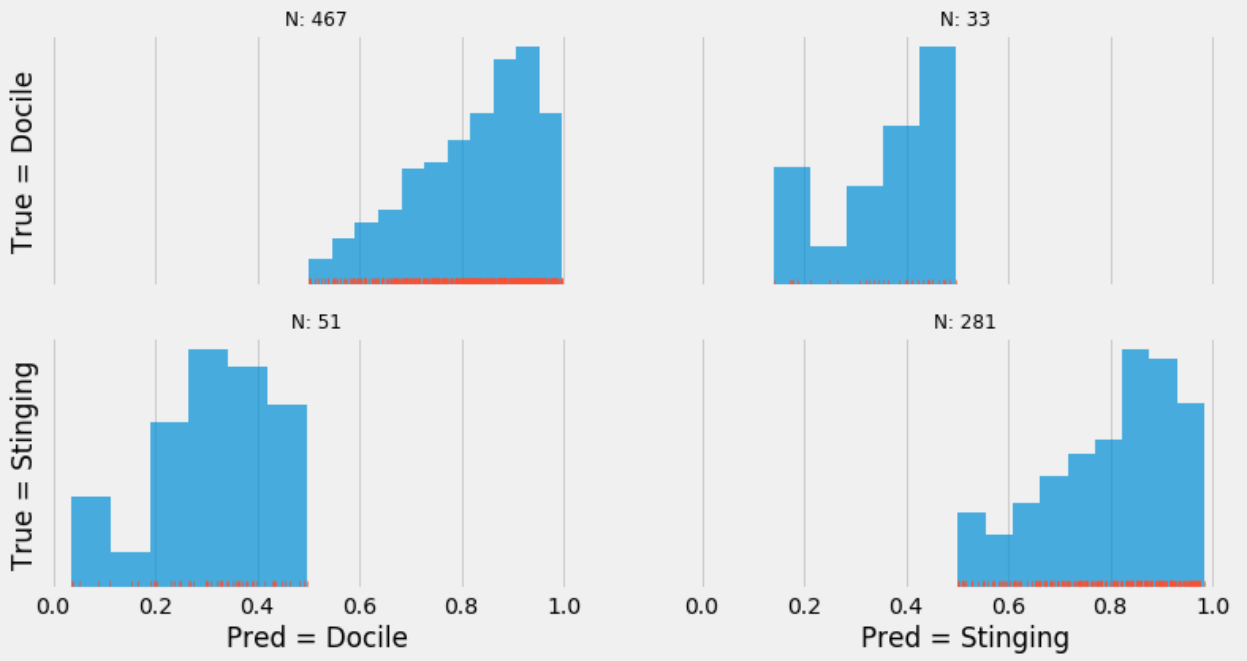

Figure 13. A confusion matrix where the cells show the distribution of the model's prediction probabilities for all the samples when it has access to location information and 41 SNP positions.

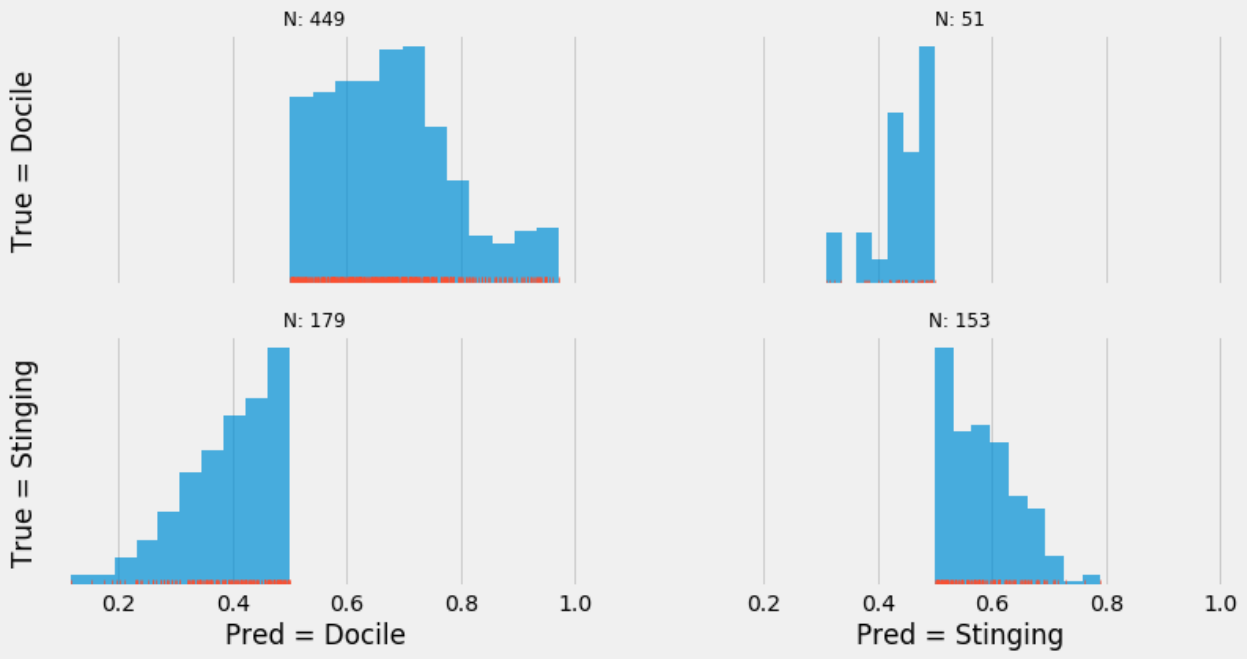

Figure 14. A confusion matrix where the cells show the distribution of the model's prediction probabilities for all the samples with only SNP information from 41 SNP positions.

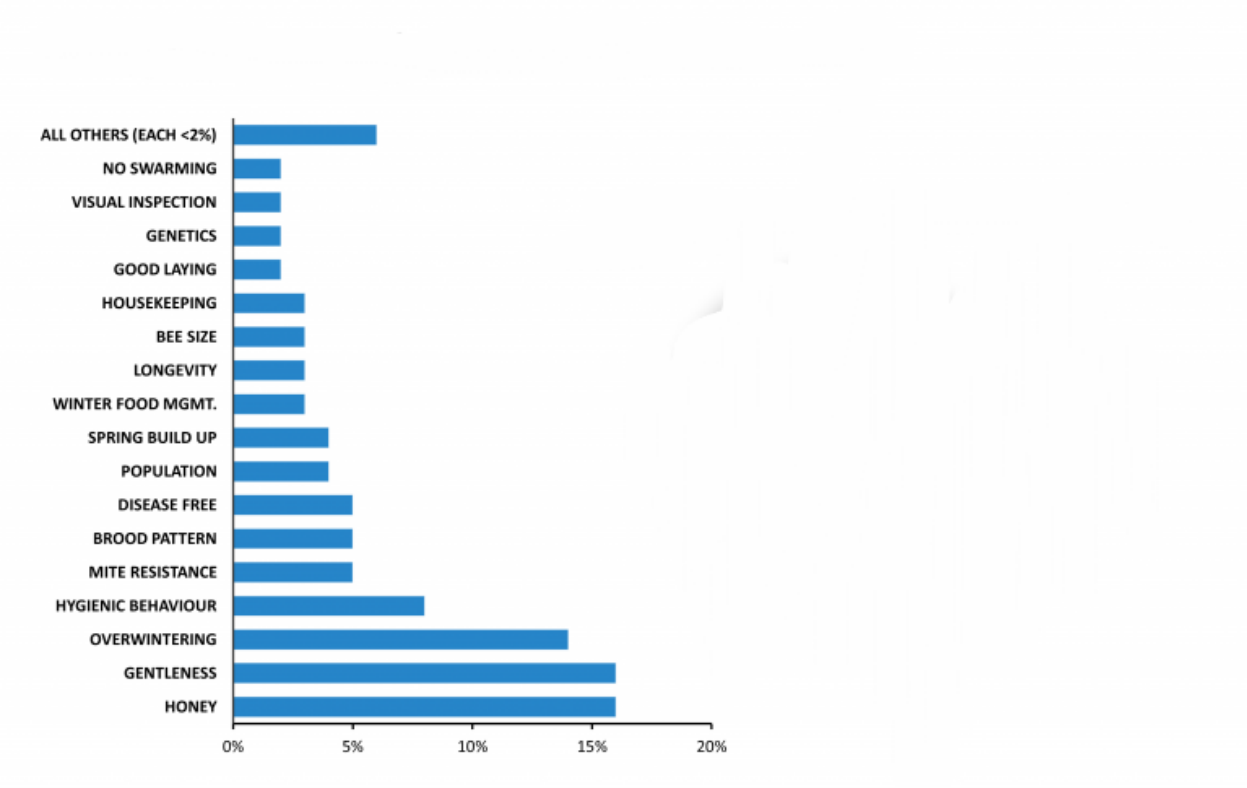

Figure 15. A bar plot showing survey results of questioning beekeepers 'what queen traits do you want?'. The results show that beekeepers highly value both honey production (honey) and defensive behaviour (gentleness) equally. N = 123.

REFERENCES

A., A. G., C. M. O., H. Christopher, P. Ryan, d. A. Guillermo, L. M. Ami, J. Tadeusz, S. Khalid, R. David, T. Joel, B. Eric, G. K. V., A. David, G. Stacey, and D. M. A. 2013. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Current Protocols in Bioinformatics 43:11.10.11-11.10.33.

Alaux, C., S. Sinha, L. Hasadsri, G. J. Hunt, E. Guzmán-Novoa, G. DeGrandi-Hoffman, J. L. Uribe-Rubio, B. R. Southey, S. Rodriguez-Zas, and G. E. Robinson. 2009. Honey bee aggression supports a link between gene regulation and behavioral evolution. Proceedings of the National Academy of Sciences 106:15400-15405.

Arbet, J., M. McGue, S. Chatterjee, and S. Basu. 2017. Resampling-based tests for Lasso in genome-wide association studies. BMC Genetics 18:70.

Bair, E., T. Hastie, D. Paul, and R. Tibshirani. 2006. Prediction by supervised principal components. Journal of the American Statistical Association 101:119-137.

Barla, A., S. Mosci, L. Rosasco, and A. Verri. 2008. A method for robust variable selection with significance assessment. Pp. 83-88. ESANN. Citeseer.

Benjamini, Y. and D. Yekutieli. 2001. The Control of the False Discovery Rate in Multiple Testing under Dependency. The Annals of Statistics 29:1165-1188.

Biswas, S., J. Reinhard, J. Oakeshott, R. Russell, M. V. Srinivasan, and C. Claudianos. 2010. Sensory Regulation of Neuroligins and Neurexin I in the Honeybee Brain. PLOS ONE 5:e9133.

Bixby, M. 2015. Beekeeper Survey: Bee 'Omics Research Project 2015.

Bodily, K. D., C. M. Morrison, R. B. Renden, and K. Broadie. 2001. A novel member of the Ig superfamily, turtle, is a CNS-specific protein required for coordinated motor control. Journal of Neuroscience 21:3113-3125.

Bolger, A. M., M. Lohse, and B. Usadel. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114-2120.

Bourdet, I., T. Preat, and V. Goguel. 2015. The full-length form of the Drosophila amyloid precursor protein is involved in memory formation. Journal of Neuroscience 35:1043-1051.

Breed, M. D., E. Guzmán-Novoa, and G. J. Hunt. 2004. Defensive behavior of honey bees: organization, genetics, and comparisons with other bees. Annual Reviews in Entomology 49:271-298.

BroadInstitute. 2018. Picard Tools.

Burns, A. and S. Iliffe. 2009. Alzheimer's disease. BMJ 338.

Chandrasekaran, S., C. Rittschof, D. Djukovic, H. Gu, D. Raftery, N. Price, and G. Robinson. 2015. Aggression is associated with aerobic glycolysis in the honey bee brain1. Genes, Brain and Behavior 14:158-166.

Chen, Y., S. Cameron, W.-T. Chang, and Y. Rao. 2017. Turtle interacts with borderless in regulating glial extension and axon ensheathment. Molecular Brain 10:17.

Christen, V. and K. Fent. 2017. Exposure of honey bees (Apis mellifera) to different classes of insecticides exhibit distinct molecular effect patterns at concentrations that mimic environmental contamination. Environmental Pollution 226:48-59.

Clark, S. A. and J. van der Werf. 2013. Genomic Best Linear Unbiased Prediction (gBLUP) for the Estimation of Genomic Breeding Values. Pp. 321-330 *in* C. Gondro, J. van der Werf, and B. Hayes, eds. Genome-Wide Association Studies and Genomic Prediction. Humana Press, Totowa, NJ.

Clarke, B., E. Fokoue, and H. H. Zhang. 2009. Principles and theory for data mining and machine learning. Springer Science & Business Media.

Collard, B. C. Y., M. Z. Z. Jahufer, J. B. Brouwer, and E. C. K. Pang. 2005. An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. Euphytica 142:169-196.

Darrach, M. and S. Page. 2016. Statistical Overview of the Canadian Honey and Bee Industry and the Economic Contribution of Honey Bee Pollination *in* H. a. C. S. D. A. a. A.-F. Canada, ed.

Decourtye, A., J. Devillers, S. Cluzeau, M. Charreton, and M.-H. Pham-Delègue. 2004. Effects of imidacloprid and deltamethrin on associative learning in honeybees under semi-field and laboratory conditions. Ecotoxicology and environmental safety 57:410-419.

Dobi, K. C., M. S. Halfon, and M. K. Baylies. 2014. Whole-genome analysis of muscle founder cells implicates the chromatin regulator Sin3A in muscle identity. Cell reports 8:858-870.

Dupuis, J. P., M. Gauthier, and V. Raymond-Delpech. 2011. Expression patterns of nicotinic subunits α2, α7, α8, and β1 affect the kinetics and pharmacology of ACh-induced currents in adult bee olfactory neuropiles. Journal of neurophysiology 106:1604-1613.

Efron, B. and R. Tibshirani. 2002. Empirical Bayes methods and false discovery rates for microarrays. Genetic epidemiology 23:70-86.

Elsik, C. G., A. Tayal, C. M. Diesh, D. R. Unni, M. L. Emery, H. N. Nguyen, and D. E. Hagen. 2016. Hymenoptera Genome Database: integrating genome annotations in HymenopteraMine. Nucleic Acids Research 44:D793-D800.

Erber, J. and P. Kloppenburg. 1995. The modulatory effects of serotonin and octopamine in the visual system of the honey bee (Apis mellifera L.). Journal of Comparative Physiology A 176:111-118.

FAOSTAT. 2017. Production quanity of honey (natural) in 20016, Livestock Primary/World Regions/Production Quantity from pick lists. United Nations.

Farris, S. M., G. E. Robinson, and S. E. Fahrbach. 2001. Experience- and Age-Related Outgrowth of Intrinsic Neurons in the Mushroom Bodies of the Adult Worker Honeybee. The Journal of Neuroscience 21:6395-6404.

Faul, F., E. Erdfelder, A.-G. Lang, and A. Buchner. 2007. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior research methods 39:175-191.

Franca, F. O. S., L. A. Benvenuti, H. W. Fan, D. R. D. Santos, S. H. Hain, F. R. Picchi-Martins, J. L. C. Cardoso, A. S. Kamiguti, R. D. G. Theakston, and D. A. Warrell. 1994. Severe and fatal mass attacks by 'killer' bees (Africanized honey bees—Apis mellifera scutellata) in Brazil: clinicopathological studies with measurement of serum venom concentrations. QJM: An International Journal of Medicine 87:269-282.

Friedman, J., T. Hastie, and R. Tibshirani. 2001. The elements of statistical learning. Springer series in statistics New York, NY, USA:.

Fritsch, C., D. Beuchle, and J. Müller. 2003. Molecular and genetic analysis of the Polycomb group gene Sex combs extra/Ring in Drosophila. Mechanisms of development 120:949-954.

GAULDIE, J., J. M. HANSON, R. A. SHIPOLINI, and C. A. VERNON. 1978. The structures of some peptides from bee venom. European journal of biochemistry 83:405-410.

Gibson, G. 2010. Hints of hidden heritability in GWAS. Nat Genet 42:558-560.

Gibson, G. 2011. Rare and Common Variants: Twenty arguments. Nature reviews. Genetics 13:135-145.

Gimenez, L. E., P. Ghildyal, K. E. Fischer, H. Hu, W. W. Ja, B. A. Eaton, Y. Wu, S. N. Austad, and R. Ranjan. 2013. Modulation of methuselah expression targeted to D rosophila insulin-producing cells extends life and enhances oxidative stress resistance. Aging cell 12:121-129.

Goguel, V., A.-L. Belair, D. Ayaz, A. Lampin-Saint-Amaux, N. Scaplehorn, B. A. Hassan, and T. Preat. 2011. Drosophila Amyloid Precursor Protein-Like Is Required for Long-Term Memory. The Journal of Neuroscience 31:1032-1037.

Gutiérrez, L., K. Oktaba, J. C. Scheuermann, M. C. Gambetta, N. Ly-Hartig, and J. Müller. 2011. The role of the histone H2A ubiquitinase Sce in Polycomb repression. Development:dev. 074450.

Guzmán-Novoa, E. and R. E. Page Jr. 1994. Genetic dominance and worker interactions affect honeybee colony defense. Behavioral Ecology 5:91-97.

Harpur, B. A., S. Minaei, C. F. Kent, and A. Zayed. 2012. Management increases genetic diversity of honey bees via admixture. Molecular Ecology 21:4414-4421.

Hayes, B. 2013. Overview of Statistical Methods for Genome-Wide Association Studies (GWAS). Pp. 149-169 *in* C. Gondro, J. van der Werf, and B. Hayes, eds. Genome-Wide Association Studies and Genomic Prediction. Humana Press, Totowa, NJ.

Hayley, S., M. O. Poulter, Z. Merali, and H. Anisman. 2005. The pathogenesis of clinical depression: Stressor- and cytokine-induced alterations of neuroplasticity. Neuroscience 135:659-678.

Henry, M., M. Béguin, F. Requier, O. Rollin, J.-F. Odoux, P. Aupinel, J. Aptel, S. Tchamitchian, and A. Decourtye. 2012. A Common Pesticide Decreases Foraging Success and Survival in Honey Bees. Science 336:348-350.

Hider, R. and U. Ragnarsson. 1981. A comparative structural study of apamin and related bee venom peptides. Biochimica et Biophysica Acta (BBA)-Protein Structure 667:197-208.

Hunt, G. J. 2007. Flight and fight: A comparative view of the neurophysiology and genetics of honey bee defensive behavior. Journal of Insect Physiology 53:399-410.

Hunt, G. J., G. V. Amdam, D. Schlipalius, C. Emore, N. Sardesai, C. E. Williams, O. Rueppell, E. Guzmán-Novoa, M. Arechavaleta-Velasco, S. Chandra, M. K. Fondrk, M. Beye, and R. E. Page. 2007. Behavioral genomics of honeybee foraging and nest defense. Naturwissenschaften 94:247-267.

Hunt, G. J., A. M. Collins, R. Rivera, J. R. E. Page, and E. Guzmán-Novoa. 1999. Brief communication. Quantitative trait loci influencing honeybee alarm pheromone levels. Journal of Heredity 90:585-589.

Hunt, G. J., E. Guzmán-Novoa, M. K. Fondrk, and R. E. Page. 1998. Quantitative Trait Loci for Honey Bee Stinging Behavior and Body Size. Genetics 148:1203-1213.

Hunt, G. J., R. E. Page, M. K. Fondrk, and C. J. Dullum. 1995. Major quantitative trait loci affecting honey bee foraging behavior. Genetics 141:1537-1545.

Koffler, S., A. de Matos Peixoto Kleinert, and R. Jaffé. 2017. Quantitative conservation genetics of wild and managed bees. Conservation Genetics 18:689-700.

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

Kravitz, E. A. and R. Huber. 2003. Aggression in invertebrates. Current Opinion in Neurobiology 13:736-743.

Laidlaw, H. H. and R. E. Page. 1997. Queen rearing and bee breeding.

Landis, G. N., D. Abdueva, D. Skvortsov, J. Yang, B. E. Rabin, J. Carrick, S. Tavaré, and J. Tower. 2004. Similar gene expression patterns characterize aging and oxidative stress in Drosophila melanogaster. Proceedings of the National Academy of Sciences of the United States of America 101:7663-7668.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078-2079.

Li, J., K. Das, G. Fu, R. Li, and R. Wu. 2011. The Bayesian lasso for genome-wide association studies. Bioinformatics 27:516-523.

Lin, C., A. Koval, S. Tishchenko, A. Gabdulkhakov, U. Tin, G. P. Solis, and V. L. Katanaev. 2014. Double suppression of the Gα protein activity by RGS proteins. Molecular cell 53:663-671.

Lyne, R., R. Smith, K. Rutherford, M. Wakeling, A. Varley, F. Guillier, H. Janssens, W. Ji, P. Mclaren, P. North, D. Rana, T. Riley, J. Sullivan, X. Watkins, M. Woodbridge, K. Lilley, S. Russell, M. Ashburner, K. Mizuguchi, and G. Micklem. 2007. FlyMine: an integrated database for Drosophila and Anopheles genomics. Genome Biology 8:R129.

Mackay, T. F., E. A. Stone, and J. F. Ayroles. 2009. The genetics of quantitative traits: challenges and prospects. Nat Rev Genet 10:565-577.

Menzel, R. 2012. The honeybee as a model for understanding the basis of cognition. Nature Reviews Neuroscience 13:758.

Michener, C. D. 1975. The Brazilian bee problem. Annual Review of Entomology 20:399-416.

Miles, C. and M. Wayne. 2008. Quantitative Trait Locus (QTL) Analysis. Nature Education 1:208.

Mrode, R. A. 2014. Linear models for the prediction of animal breeding values. Cabi.

Oldroyd, B. P. and G. J. Thompson. 2006. Behavioural genetics of the honey bee Apis mellifera. Advances in insect physiology 33:1-49.

Page Jr, R., M. Fondrk, G. Hunt, E. Guzman-Novoa, M. Humphries, K. Nguyen, and A. Greene. 2000. Genetic dissection of honeybee (Apis mellifera L.) foraging behavior. Journal of Heredity 91:474-479.

Perry, T., D. G. Heckel, J. A. McKenzie, and P. Batterham. 2008. Mutations in Dα1 or Dβ2 nicotinic acetylcholine receptor subunits can confer resistance to neonicotinoids in Drosophila melanogaster. Insect Biochemistry and Molecular Biology 38:520-528.

Radyuk, S. N., J. Gambini, C. Borras, E. Serna, V. I. Klichko, J. Viña, and W. C. Orr. 2012. Age-dependent changes in the transcription profile of long-lived Drosophila over-expressing glutamate cysteine ligase. Mechanisms of Ageing and Development 133:401-413.

Ramírez, G., C. Fagundez, J. P. Grosso, P. Argibay, A. Arenas, and W. M. Farina. 2016. Odor Experiences during Preimaginal Stages Cause Behavioral and Neural Plasticity in Adult Honeybees. Frontiers in Behavioral Neuroscience 10.

Reinhard, V. and K. Günther. 1984. Nucleotide sequence of cloned cDNAs coding for preprosecapin, a major product of queen-bee venom glands. European Journal of Biochemistry 145:279-282.

Rittschof, C. C. and G. E. Robinson. 2013. Manipulation of colony environment modulates honey bee aggression and brain gene expression. Genes, Brain and Behavior 12:802-811.

Rueppell, O., T. Pankiw, and R. Page Jr. 2004. Pleiotropy, epistasis and new QTL: the genetic architecture of honey bee foraging behavior. Journal of heredity 95:481-491.

Scheiner, R., R. E. Page, and J. Erber. 2004. Sucrose responsiveness and behavioral plasticity in honey bees (Apis mellifera). Apidologie 35:133-142.

Schippers, M.-P., R. Dukas, and G. B. McClelland. 2009. Lifetime- and caste-specific changes in flight metabolic rate and muscle biochemistry of honeybees, Apis mellifera. Journal of Comparative Physiology B 180:45.

Schippers, M.-P., R. Dukas, R. Smith, J. Wang, K. Smolen, and G. McClelland. 2006. Lifetime performance in foraging honeybees: behaviour and physiology. Journal of Experimental Biology 209:3828-3836.

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

Seeley, T. D. 1982. Adaptive significance of the age polyethism schedule in honeybee colonies. Behavioral Ecology and Sociobiology 11:287-293.

Speliotes, E. K., C. J. Willer, S. I. Berndt, K. L. Monda, G. Thorleifsson, A. U. Jackson, H. L. Allen, C. M. Lindgren, J. a. Luan, and R. Mägi. 2010. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nature genetics 42:937.

Thany, S. H., M. Crozatier, V. Raymond-Delpech, M. Gauthier, and G. Lenaers. 2005. Apisα2, Apisα7-1 and Apisα7-2: three new neuronal nicotinic acetylcholine receptor α-subunits in the honeybee brain. Gene 344:125-132.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological):267-288.

Tichy, A. L., A. Ray, and J. R. Carlson. 2008. A new Drosophila POU gene, pdm3, acts in odor receptor expression and axon targeting of olfactory neurons. Journal of Neuroscience 28:7121-7129.

Toth, A. L., S. Kantarovich, A. F. Meisel, and G. E. Robinson. 2005. Nutritional status influences socially regulated foraging ontogeny in honey bees. Journal of Experimental Biology 208:4641-4649.

Toth, A. L. and G. E. Robinson. 2005. Worker nutrition and division of labour in honeybees. Animal behaviour 69:427-435.

Tweedie, S., M. Ashburner, K. Falls, P. Leyland, P. McQuilton, S. Marygold, G. Millburn, D. Osumi-Sutherland, A. Schroeder, and R. Seal. 2008. FlyBase: enhancing Drosophila gene ontology annotations. Nucleic acids research 37:D555-D559.

Visscher, P. M., W. G. Hill, and N. R. Wray. 2008. Heritability in the genomics era—concepts and misconceptions. Nature reviews genetics 9:255.

Visscher, P. M., N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, and J. Yang. 2017. 10 Years of GWAS Discovery: Biology, Function, and Translation. The American Journal of Human Genetics 101:5-22.

Wetterstrand, K. 2018. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). National Human Genome Research Institute.

Williamson, S. M. and G. A. Wright. 2013. Exposure to multiple cholinergic pesticides impairs olfactory learning and memory in honeybees. The Journal of Experimental Biology.

Wilm, A., P. P. K. Aw, D. Bertrand, G. H. T. Yeo, S. H. Ong, C. H. Wong, C. C. Khor, R. Petric, M. L. Hibberd, and N. Nagarajan. 2012. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Research 40:11189-11201.

Winston, M. L. 1991. The biology of the honey bee. harvard university press.

Winston, M. L., O. R. Taylor, and G. W. Otis. 1983. Some Differences between Temperate European and Tropical African and South American Honeybees. Bee World 64:12-21.

Wright, G. A., S. Softley, and H. Earnshaw. 2015. Low doses of neonicotinoid pesticides in food rewards impair short-term olfactory memory in foraging-age honeybees. Scientific Reports 5:15322.

Yang, E., Y. Chuang, Y. Chen, and L. Chang. 2008. Abnormal foraging behavior induced by sublethal dosage of imidacloprid in the honey bee (Hymenoptera: Apidae). Journal of economic entomology 101:1743-1748.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher. 2010a. Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42:565-569.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher. 2010b. Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42:565-569.

Yang, J., S. H. Lee, M. E. Goddard, and P. M. Visscher. 2013. Genome-Wide Complex Trait Analysis (GCTA): Methods, Data Analyses, and Interpretations. Pp. 215-236 *in* C. Gondro, J. van der Werf, and B. Hayes, eds. Genome-Wide Association Studies and Genomic Prediction. Humana Press, Totowa, NJ.

Yi, H., P. Breheny, N. Imam, Y. Liu, and I. Hoeschele. 2015. Penalized Multimarker vs. Single-Marker Regression Methods for Genome-Wide Association Studies of Quantitative Traits. Genetics 199:205-222.

Zou, H. and T. Hastie. 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67:301-320.