Genomic Insights into Honey Bee Health: Genetics of Immunity and Colony-level Traits

Tanushree Tiwari

A DISSERTATION SUBMITTED TO THE FACULTY OF GRADUATE STUDIES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN BIOLOGY,
YORK UNIVERSITY, TORONTO, ONTARIO

April 2026

Abstract

Honey bees (Apis mellifera) are essential pollinators in natural and agricultural ecosystems and a key model for studying social behavior. Despite their ecological and economic importance, managed honey bee colonies have experienced persistently high mortality over the past two decades due to interacting stressors, including pests, pathogens, and environmental pressures. Recent national surveys in Canada by the Canadian Association of Professional Apiculturists and Provincial Apiarists reported winter colony losses of 39.3% during 2024–2025, substantially exceeding long-term averages, with Varroa destructor, associated viral infections, weak fall colonies, and poor queen quality identified as major contributors. At the same time, honey bees have substantial phenotypic and genetic diversity across behavioral, immune, and colony-level traits, many of which show measurable narrow-sense heritability. This standing variation provides an opportunity to investigate the genetic basis of colony resilience and immunity.

Advances in next-generation sequencing have enabled genome-wide analyses of complex traits at unprecedented scale and resolution. In this dissertation, Chapter 2 establishes the applied and conceptual foundation by reviewing genomic tools for honey bee health management, including stock certification, lineage monitoring, and marker-assisted selection. Building on this framework, Chapter 3 presents a large-scale genome-wide association study (GWAS) of 1,350 colonies spanning behavioral, social, and immune traits, revealing that colony-level traits are heritable yet highly polygenic, shaped by local adaptation, and influenced by extensive pleiotropy. To connect genotype to function further in Chapter 4, protein quantitative trait locus mapping was done to identify loci and genes regulating innate immune protein abundance and pathogen resistance, finding shared loci that influence both protein expression and pathogen load. Integrating these results, Chapter 5 examines the shared genetic architecture of innate immunity, social immunity, and pathogen-associated traits using network analyses, demonstrating that robust honey bee defense is underpinned by interconnected gene networks. Finally, Chapter 6 characterizes understudied short insertion-deletion (InDel) variants and their contribution to colony-level traits, providing complementary evidence to SNP-based analyses.

Overall, my dissertation identifies candidate genes, loci, and different genetic variants associated with honey bee health, immunity, and colony performance. These findings highlight key

targets for future functional validation and provide a foundation for exploring how natural genetic variation can inform breeding and management strategies aimed at enhancing colony resilience.

Acknowledgments

This journey has been long, at times challenging, rewarding, and full of memorable moments, and I could not have done it without the love and support of my family, especially my kids, Palvit and Krishiv, my husband, and my friends, who have cheered me on every step of the way. Your encouragement, kindness, and celebrations have kept me going, and I am blessed to have you all.

I am deeply thankful to Dr. Amro Zayed, my supervisor for the past nine years, for his unwavering guidance, encouragement, and for introducing me to the fascinating world of honey bees, beekeeping, and their genetics. Your support and constructive feedback have been invaluable, helping me grow as a researcher and as a person. I will always cherish the opportunities, experiences, and mentorship that have shaped this chapter of my life.

I am also grateful to the past and present members of the Zayed lab. It has been a pleasure getting to know everyone throughout my PhD journey. I would like to especially thank Clement, who patiently helped me navigate Canadian culture and guided me throughout this journey. A special mention to Kathleen, who always listened patiently to my questions, was my conference buddy, made it easy to talk about anything, and was there to support me during the tough moments. Thank you to Rodney, Stephen, and Harshil for working with me on this project and creating unforgettable memories. I am also grateful to Dova, with whom I navigated the final phase of my PhD through countless Slack chats and Mariana, with whom I shared the office space. The lab has been an incredible source of support, and I will miss you all dearly.

Finally, I would like to thank my past and present committee members and examination committee, Dr. Shore, Dr. Stutchbury, Dr. Sharma, Dr. Schott, Dr. Bucking, Dr. Heffernan and Dr. Richards, for their insightful guidance and valuable feedback. I am also thankful to the many professors at York for whom I served as a teaching assistant, as well as to the many others who contributed to a positive and enriching academic experience.

Table of Contents

List Of Figures

List Of Tables

Chapter one: Introduction - A genomic framework for studying honey bee genetics and health

Tanushree Tiwari

The biology and health of honey bees (*Apis mellifera*) have been of interest to human societies for centuries. They are a key model for studying social behavior and are among the most important managed pollinators in agriculture (Klein et al., 2007; Potts et al., 2010). Their ability to be transported as large colonies to meet pollination demands makes them indispensable to modern crop production. Despite this ecological and agricultural importance, honey bee colonies have experienced elevated mortality over the past two decades (Goulson et al., 2015a; VanEngelsdorp et al., 2009). These losses are often multifactorial, involving pesticides (Sanchez-Bayo & Goka, 2014), pathogens (Evans & Schwarz, 2011), and environmental stressors (Goulson et al., 2015a). Recent national surveys in Canada highlight the severity of this issue, reporting winter colony losses of approximately 39% in 2024–2025 (CAPA, 2025)(Claing et al.), which is substantially higher than the long-term national average of approximately 27–28% recorded since 2007 (CAPA, 2024; CAPA, 2025) (Claing et al.).

Widespread colony losses underscore an important opportunity: the substantial phenotypic variation present within and among honey bee colonies. Traits related to behaviour, colony productivity, pathogen resistance, and immune function show measurable variation, and many exhibit moderate-to-high narrow-sense heritability. The presence of heritable variation makes selective breeding a viable strategy for improving colony resilience. In other agricultural systems, marker-assisted selection (MAS) and genomic selection have substantially accelerated genetic gain by integrating molecular markers into breeding decisions (Moniruzzaman et al., 2014). However, their application in honey bees has been limited, primarily due to the incomplete identification of specific genetic variants underlying immune and health-related traits.

Recent advances in next-generation sequencing (NGS) and genome-wide association studies (GWAS) now provide the necessary tools to overcome this limitation. These approaches enable genome-wide screening of large populations and the identification of genetic variants associated

with phenotypic differences at increasingly accessible costs (Koboldt et al., 2013; Korte & Farlow, 2013). By linking naturally occurring genetic variation to measurable health and immune traits, genomic approaches offer a practical pathway to harness existing diversity for sustained improvement in colony health.

Chapter 2 of my dissertation provides an in-depth review of three major genomic applications that address different dimensions of the bee health crisis: genetic quality, disease resistance, and colony-level diagnostics. First, I explore how genomic data can be used to study the genetic background of bees, which can be used for stock certification, tracking, and to detect invasive bee lineages such as Africanized 'killer' bees. Second, I explore how marker-assisted selection (MAS) can be informed by genome-wide association studies (GWAS) to accelerate selective breeding. Finally, I discuss how genomics can be used for the development of biomarkers to assess colony health, offering new opportunities for the early detection of stressors and pathogens in the field.

Chapters 3 to 6 are based on a single large-scale experiment – BeeOMICs, pQTLs, Social Immunity vs Innate Immunity to InDels characterize the genetic architecture of colony health and immune-related traits in honey bees. The overarching goal of this experiment was to identify genomic variants associated with variation in pathogen resistance, innate immune function, social immunity, and other colony-level phenotypes, and to assess how these components interact at both molecular and colony scales. To achieve this, we generated a comprehensive dataset comprising 1,350 honey bee colonies sampled over two consecutive years across five Canadian provinces. For each colony, whole-genome sequencing was performed on pools of 50 worker bees to capture colony-level genetic variation. In parallel, extensive phenotypic data were collected, including colony performance metrics, pathogen loads, and multiple innate and social immune traits.

Eusocial animals exhibit emergent colony-level behaviors that arise from the cooperation of thousands of individuals (Seeley, 1997), yet the underlying genetics of these traits have remained largely unknown. In **Chapter 3,** we used the BeeOMICs dataset to carry out genome wide association studies (GWAS) to study the genetic basis of "superorganism" traits in honey bees. Specifically, we carried out a GWAS to identify the genes underlying Hygiene, Honey Production, Defense Behavior, Grooming, and the levels of *Varroa* mites, RNA viruses, *Nosema* and *Lotmaria* within colonies. The GWAS analysis revealed that while colony-level traits have substantial

genomic heritability, their architecture is highly polygenic, consisting of many small-effect loci. We discovered extensive pleiotropy, particularly within prominent neuronal genes such as neurexin-1, neuroligins, and nicotinic acetylcholine receptors which influence multiple colony-level traits simultaneously. Furthermore, these genes are overwhelmingly associated with the local adaptation of native honey bee subspecies in Europe and West Asia. This provides strong evidence that superorganism traits are key drivers of fitness and adaptive evolution.

In **Chapter 4**, I focused on the genetic basis of innate immune regulation by integrating whole-genome sequencing with quantitative proteomics of 64 immune-related proteins. Immune protein abundance showed narrow-sense heritability estimates averaging 0.152 and ranging from 0.005 to 0.485. pQTL association analysis identified 2,559 candidate loci across 16 linkage groups, including 20 *cis*-acting loci associated with 15 proteins and 2,539 *trans*-acting SNPs, indicating that most immune protein variation is regulated by distributed genetic networks rather than local regulatory variants. Functional annotation revealed that many variants mapped to protein-coding genes and regulatory regions, highlighting a complex genetic architecture underlying immune protein expression. Associations between immune protein abundance and pathogen traits further revealed shared genetic influences across multiple immune challenges. Together, these results demonstrate that variation in honey bee innate immunity is partly genetically determined and shaped largely by *trans*-acting regulatory variation.

Eusocial insects rely on both individual (innate) immunity and social (behavioral) immunity, yet the genetic coordination between these distinct layers has remained poorly understood. While the previous chapters map the genetic architecture of colony traits (Chapter 3) and the regulation of immune proteins (Chapter 4), **Chapter 5** uses these findings to investigate how honey bees coordinate their dual-layered defense system. Our analyses show that both immune systems contain largely distinct genetic components but also share a subset of common genetic influences**.** We identified 41 core pleiotropic genes, including several uncharacterized loci, that are shared between innate and social immunity and are associated with a diverse range of pathogens, such as *Varroa* mites, *Nosema, Lotmaria* and various RNA viruses. The enrichment of neural and synaptic regulators among these shared genes reinforces the concept that honey bee immunity is fundamentally embedded within neurobehavioral networks.

Finally, **Chapter 6** broadens the genomic research from single nucleotide polymorphisms (SNPs) to the second most common source of genetic variants - insertion/deletion polymorphism (InDels). While InDels are an important source of genetic variation, they are not well known in honey bees. To address this, we characterized short InDel variants in the BeeOMICs dataset and identified more than 125,000 short InDels**,** the majority of which were in non-coding regions**,** with intronic and intergenic variants predominating. We then tested their associations with five heritable colony-level traits previously analyzed using SNP markers. Although fewer InDels were associated with these traits in absolute number compared with SNPs**,** the proportion of suggestive associations relative to the number of variants tested was higher for InDels**.** However, the estimated genetic effect sizes of InDels were generally smaller than those of SNPs**,** indicating that their contributions are typically more subtle. Together, these results show that short InDels represent a widespread source of genetic variation in the honey bee genome and contribute to the genetic architecture of colony-level traits, with generally smaller individual effects than SNPs.

To summarize, my dissertation establishes a genomic framework for understanding the biological foundations of honey bee health and colony-level resilience. By integrating whole-genome sequencing, GWAS, proteomics, and extensive phenotyping across 1,350 colonies, it reveals that complex honey bee traits are highly polygenic and follow a pattern of many loci each contributing a small effect. Across chapters, neuronal and signaling molecules repeatedly emerge as central components linking behavior, immunity, and colony organization. These genes appear to form a regulatory backbone of superorganismal traits. Many of the same genes also show signatures of local adaptation in European and West Asian subspecies, highlighting that colony-level traits are critically important for the fitness of honey bee colony, as predicted by evolutionary theory. My research further demonstrates that short InDels contribute distinct and complementary variation beyond SNP-based signals, refining our understanding of honey bee genomic architecture. Together, these analyses generate a large catalogue of candidate genes and loci associated with immunity, pathogen resistance, productivity, and resilience. While functional validation remains essential, these loci provide a practical foundation for marker-assisted selection to improve honey bee health.

Chapter two: Practical applications of genomics in managing honey bee health [1]

Tanushree Tiwari, Amro Zayed, PhD*[1]

Introduction:

The honey bee *Apis mellifera* is a model organism for the study of social behavior and is essential to our agroeconomy because of its importance to pollination. The honey bee is the most widely used pollinator in agriculture because their large perennial colonies can be moved wherever they are needed, enabling beekeepers and farmers to easily deploy large numbers of bees for pollination whenever necessary. However, honey bee colonies have experienced declining health leading to colony losses over the past 2 decades. For example, the average winter mortality in Canada and the United States over the past decade is 24.3% (Ferland et al., 2020) and 27.3% (Bruckner et al., 2019), respectively. These averages are higher than the historic threshold of 15% mortality experienced by North American beekeepers before the introduction of *Varroa* mites. Some specific years have been extremely challenging for North American beekeepers. In 2020, the Canadian winter mortality averaged 30.2% (Ferland et al., 2020) with some provinces reporting even higher values. Similarly, beekeepers in the United States lost approximately 43.7% of their colonies in the winter of 2020. This loss is the second highest average annual loss in the United States, with the first occurring earlier this decade in 2013 when approximately 45% of colonies died during the winter (Bruckner et al., 2018). These losses are unsustainable and have the potential to negatively impact our agroeconomy and our food security.

Broadly speaking, several general factors have been implicated in the decline of honey bee health. These factors are commonly called the 4 “P’s”: Pathogens, Pests, Pesticides, and Poor nutrition (Goulson et al., 2015b). Stressors, however, do not act in isolation, and their interactions are difficult to predict. For example, pesticides used in agriculture can act

[1] [1]This published manuscript has been reprinted with the permission of its co-authors and publisher from the original manuscript: Tiwari, T., & Zayed, A. (2021). Practical Applications of Genomics in Managing Honey bee Health. *Veterinary Clinics: Food Animal Practice*, *37*(3), 535-543.

synergistically rather than having additive effects on honey bees (Sih et al., 2004). Pesticide exposure and nutritional stress in the form of limited forage on monoculture can compromise the honey making them more susceptible to parasites (Goulson et al., 2015b). It is clear that chronic exposure to multiple interacting stressors is an important factor leading to honey bee colony losses and declines of wild pollinators, but the combination and number of stressors differ from place to place (Neumann & Carreck, 2010) making universal diagnoses and cures for bee health very challenging.

Managed honey bee populations are also at risk from hybridization with invasive honey bee lineages, such as "killer" bees in South America and clonal bees in South Africa. "Killer bees," hybrid African and west European bees accidentally introduced to South America in 1956 (Jernelöv, 2017), are up to 6 times more defensive (i.e., aggressive) than European-derived bees. (Stort & Gonçalves, 1991). Their swarming frequency is also high, which, combined with their aggression, make them very difficult to manage commercially. Similarly, parasitic clonal bees from the cape of South Africa, *Apis mellifera capensis*, can wreak havoc on commercial managed honey bees, leading to dwindling colony syndrome (Allsopp & Crewe, 1993). These workers invade colonies and produce clonal female offspring through automixis (Verma & Ruttner, 1983), thereby establishing themselves as pseudo queens; once established, these clonal workers continue reproducing until the colony collapses (Hemmling et al., 1979).

The study of honey bee genetics and genomics offers many intriguing possibilities for improving honey bee health. The first honey bee genome was sequenced in 2006 (Consortium, 2006). This genome sequence was assembled with DNA extracted from multiple drones derived from a single queen (Consortium, 2006). A more contiguous genome assembly of *A mellifera* was released in 2019 (Wallberg et al., 2019) using a combination of data from long-read and short-read sequencing. The long reads played a significant role in filling the gaps and joining thousands of scaffolds into 16 superscaffolds. The newer assembly particularly improved the quality of the bee genome near repetitive regions (Wallberg et al., 2019). Overall, the honey bee genome contains 236 million base-pairs organized into 16 chromosomes with more than 10,000 predicted genes (Consortium, 2006). The size of the bee genome is one-tenth that of the human genome.

Several properties render honey bees different for most typical livestock from a genetics and breeding perspective. Honey bees are haplodiploid (Beye et al., 2003): females are diploid,

and males are haploid. The sex of a bee is determined by genotype at a variable autosomal gene known as complementary sex determiner, *csd*. (Beye et al., 2003). Queen bees can selectively fertilize their eggs. Unfertilized eggs are hemizygous at *csd*, and these eggs develop into haploid males (Beye, 2004). When producing daughters or future queens, females will fertilize their eggs and under most circumstances (because of the high genetic variability at *csd*), and these eggs will be heterozygous at *csd* and develop as female (Beye et al., 2003). However, under some circumstances, fertilized eggs will be homozygous at *csd*, leading to the development of diploid males, which are eaten by workers at the first larval instar stage (Woyke, 1963). Any population genetic process that leads to a reduction in genetic diversity at *csd*, such as low effective population size or inbreeding, will result in higher frequencies of diploid male production, leading to a phenotype called "shot brood", a very spotty brood pattern, as many of the developing larva are removed by workers (Beye et al., 2003); this phenomena drains the colony of workers and substantially reduces the colony's fitness (Zayed, 2004). As such, many of the classical selective breeding methods applied to livestock cannot be effectively applied to honey bees.

In addition to haplodiploidy, queen honey bees are highly polyandrous (Tarpy & Page, 2000). Queens naturally mate with a large number of haploid males (typically, 15-25 males) during the first week of their life as adults and store the sperm for use over the duration of their lifespan (2-4 years), introducing a high degree of genetic diversity within colonies, which has been linked to colony fitness (Mattila & Seeley, 2007). Polyandry also makes traditional breeding approaches common in other livestock very difficult because of our general inability to control mating. Although it is possible to use instrumental insemination to generate desired crosses (Cobey et al., 2013) the technique requires highly specialized skill sets and equipment and is not commonly used in beekeeping.

Last, a major difficulty with studying honey bee genetics is that many of the traits of interest are "collective" and are thus not directly amenable to genetic experiments aimed at studying traits expressed by individuals. A honey bee colony contains a single reproductive queen and thousands of sterile workers (i.e., nurses, foragers, guards) that build and maintain the nest, care for the brood, forage for food, and engage in nest defense (Michener, 1974). Worker bees exhibit many fascinating behaviors, such as dancing to communicate the location of new food sources (Von Frisch, 1967), removing sick or dead individuals from a hive to prevent the

spread of disease, (Park et al., 1937) attacking invaders (Breed et al., 2004), grooming each other to remove pests and parasites (Aumeier, 2001) and clustering together in the winter to keep the colony warm (Seeley & Visscher, 1985). These collective behaviors, performed by a large number of cooperating workers with different genetics, ultimately contribute to the survival of colonies. Elucidating the genetics of these colony level traits has been challenging.

Despite the difficulties introduced by haplodiploidy, *csd*, polyandry, and sociality, there is substantive potential for using genetics and genomics in managing bee health. Here, the authors review 3 practical applications of incorporating genomics into everyday beekeeping. First, they discuss the utility of using single-nucleotide polymorphisms (SNPs) to distinguish between the many European and Asian honey bee subspecies that are commonly used in apiculture and invasive strains that are undesirable for beekeeping. Second, many honey bee traits are highly heritable and can thus be improved by selective breeding. The authors discuss the potential for using genomics to identify causal mutations underlying economically valuable traits and using them in marker-assisted selection (MAS). Finally, when honey bee colonies perish, it is often very difficult to understand the specific causes that led to their demise. The authors discuss the potential of using expression biomarkers to rapidly identify stressors affecting honey bee colonies in the field.

Application 1: Stock certification and control of invasive lineages

The availability of honey bee genomes and SNP data sets has allowed us to identify genomic variation within and between populations or species of honey bees (Cornman et al., 2013; Harpur et al., 2014a; Parejo et al., 2016). These studies help in elucidating differences in genetics between native honey bee subspecies and the source and contribution of these native subspecies in the typically highly admixed managed colonies used in beekeeping (Chapman et al., 2016; Harpur et al., 2015; Harpur et al., 2012, 2013). Diagnostic mutations for specific honey bee races can be used in breeding programs aimed at maintaining pure local subspecies, which are at risk of extinction via introgression with managed honey bees (Henriques et al., 2018; Muñoz et al., 2017). Moreover, several studies have identified sets of SNPs that are successful in separating managed honey bees from invasive Africanized “killer” bees. For example, a diagnostic assay of 95 (Chapman et al., 2015) and 37 (Chapman et al., 2017) SNPs has been developed to accurately identify “killer bees” by measuring their African lineage ancestry. These diagnostic assays should be widely used to ascertain the genetics of imported queens or colonies,

especially in North America, where killer bees are spreading into areas with large queen breeding operations in the southern United States (Kono & Kohn, 2015; Lin et al., 2018).

Application 2: Marker assisted selection powered by genome- wide association studies.

Many of the colony level traits that influence the health and temperament of honey bee colonies are heritable. For example, narrow-sense heritability for hygienic behavior, a type of social immunity whereby hygienic bees uncap and remove dead or *Varroa*- infected pupae from the colony, is between 20% and 65% (Harris et al., 2003; Rothenbuhler, 1964). Similarly, honey production by a colony is an economically important trait that beekeepers wish to maximize and has a heritability ranging from 25% to 60% (Soller & Bar-Cohen, 1967). Defensive behavior is a colony-level heritable trait whereby guard bees detect intruders at the colony entrance and release alarm pheromones, which recruits older bees that sting intruders, sacrificing themselves in the process (Hunt, 2007). Heritability estimates for defensive behavior range between 13% and 43% (Bienefeld & Pirchner, 1990). The temperament of bees is important for the industry, as beekeepers generally prefer to work with docile bees (Breed et al., 1990; Hunt et al., 2007). In addition, at the individual level, honey bees have several lines of innate immune defense mechanisms against pathogens, and it is known that some of these defenses are heritable as well. For example, the antimicrobial peptide abaecin has a heritability estimate between 30% and 40% (Decanini et al., 2007).In theory, this heritability can be used to breed honey bees with superior characteristics for beekeeping.

Population genomics methods are useful in identifying genes and mutations with signatures of natural or artificial selection. Generally, population genomic studies sequence bees that show phenotypic differences caused either by natural selection in native bee races or by artificial selection acting to enhance specific traits in managed populations. Traditional analysis of positive selection (Fuller et al., 2015; Nielsen, 2005) thereby enables researchers to identify genes and mutations associated with phenotypic traits. For example, a recent study sequenced drones from populations selected for high royal jelly production and discovered 44 regions associated with this trait (Wragg et al., 2016). Sequencing haploid drones in bees is beneficial because it reduces the sequencing coverage needed to accurately identify SNPs. A similar study sequenced populations selected for high hygienic behavior and identified 73 protein-coding genes associated with this trait (Harpur et al., 2019)

Genome-wide association studies (GWAS) is an emerging approach for uncovering candidate regions of a genome associated with complex quantitative traits, such as behavior. Powered by next-generation sequencing data, GWAS approaches have made it economically practical and feasible to sequence thousands of individuals to discover millions of common SNP (Koboldt et al., 2013) and associate these genetic differences with phenotypic differences between individuals (Hayes, 2013). Genome-wide association mapping is suitable for identifying common mutations that have small to medium additive effects on a phenotype (Korte & Farlow, 2013). A major advantage of GWAS for identifying the loci that affect bee health is that it does not require crosses that are difficult to perform in the honey bee.

Another benefit of applying GWAS in honey bees is that honey bees have very high recombination rates, resulting in rapid decay of linkage disequilibrium (Wallberg et al., 2015). This typically results in small genomic regions showing a statistical association with traits studied via GWAS (tens of kilobases or smaller); however, adequately powered experiments are needed to achieve such results (Wu et al., 2017). If the causal variants affecting key honey bee traits can be identified, such as immunity, aggression, hygiene, honey production, and overwintering mortality, breeding efforts can be substantially aided by applying MAS.

Although still uncommon, a few GWAS-type studies have been carried out on honey bees. For example, using a 44,000 Affymetrix SNP array, researchers studied a social immune trait involving the detection and uncapping of a *Varroa*-affected brood by honey bee workers and discovered 6 SNP markers for the trait (Spötter et al., 2016). Recently, a GWAS study was performed to determine the genetics of defensive behavior in honey bees (A. Avalos et al., 2020). The investigators found 1172 SNPs for which colony allele frequency was significantly associated with colony defensive behavior (A. Avalos et al., 2020). The authors look forward to more GWAS studies on honey bee traits of ecological and economical importance.

Application 3: Biomarker for diagnosing Bee health

Beekeepers and conservation practitioners do not have adequate diagnostic tools to understand why honey bee health is declining, and this is hampering abilities to manage honey bees and circumvent colony losses. Beekeepers rarely have access to professional support to help them diagnose bee health issues, unlike in other livestock systems. Presently, beekeepers manually inspect colonies and treat them with miticides and antibiotics to manage a few known

pests and pathogens, often without knowing or testing their levels. For pathogens, quantitative polymerase chain reaction assays are available (Jay D Evans et al., 2013), but the testing methods are costly and low throughput. Similarly, analysis of agrochemicals found within bees and colonies is possible, but typically too expensive for regular use in bee health management. Often, beekeepers would only carry out postmortem tests for agrochemicals or pathogens if they experienced unusually high colony mortality, but obviously, postmortem tests are not effective for planning interventions. These shortcomings are creating problems for beekeepers, veterinarians, and government agencies who require real-time data on stressors within colonies to better understand how to manage and improve bee health.

There is a strong need for robust methods to identify stressors impacting the health of bee colonies. Here, biomarker-based assays can play a pivotal role in understanding the stressors affecting colonies. Biomarkers are commonly used in human health care and livestock management (Adnane et al., 2018) for diagnosing disease. A few studies have suggested that expression profiling can be used to identify stressor-specific biomarkers in bees, who have distinct genetic pathways for dealing with pathogenic (Brutscher et al., 2015), nutritional (Kunieda et al., 2006), and xenobiotic (Claudianos et al., 2006) challenges that are associated with specific and measurable changes in gene expression. For example, a large meta-analysis of transcriptomic studies involving honey bees exposed to different pests and pathogens discovered both common and pathogen-specific changes in gene expression associated with infections (Doublet et al., 2017b). For example, *Atg2* (LOC726497), *Metap2* (LOC551771), and *Dnr1* (LOC412897) are among genes that were differentially expressed when colonies were exposed to microsporidian parasite, *Nosema* (but not *Varroa* mites), whereas *Iap2* (LOC413374), *Rel* (LOC552247), *Tube* (LOC725368), and *Defensin2* were differentially expressed when colonies had *Varroa* and viral infections (but not *Nosema*). These genes are potential biomarkers for infection by *Nosema* or *Varroa*, respectively.

In theory, a comprehensive set of experiments that expose bees to a large number of relevant stressors, followed by expression profiling, can identify stressor-specific biomarkers that can be used for honey bee health diagnostics. Work on this front has started with the BeeCSI project (https://beecsi.ca/). This project involves assaying the transcriptome, proteome, and gut microbiome of many bees exposed to a variety of pathogenic, xenobiotic, and nutritional stressors in the laboratory and in the field. By comparing lists of differentially expressed markers

across studies, the BeeCSI team will strive to identify stressor-specific biomarkers (e.g., genes that are upregulated in bees exposed to the insecticide clothianidin) as well as general biomarkers (e.g., genes that are upregulated in bees exposed to a general class of insecticides, such as neonicotinoids). These tools can be used by beekeepers and veterinarians to understand and manage stressors affecting honey bee colonies, or by governmental regulators investigating general patterns affecting bee health at regional, provincial, or national scales.

Summary:

Several practical applications have been discussed that can help mitigate honey bee losses and improve bee health. The methods target different aspects of the bee health problem and thus collectively provide a powerful toolkit for improving the genetic stock of managed honey bees and for diagnosing or monitoring the health of colonies in the field. Although these applications have yet to reach maturity in honey bee diagnostics, similar methods and tools have been successfully applied in many livestock systems. With the reduced cost of next-generation sequencing, it is becoming increasingly feasible to perform the discussed analyses on a large number of colonies typically managed by beekeepers for crop pollination, honey production, and queen breeding. The wide adoption and use of these tools will result in improvement of health of managed honey bee colonies like never before.

Clinics Care Points:

- Distinguishing between commonly managed strains of honey bees and more aggressive, invasive lineages can be determined using single-nucleotide polymorphisms.
- Selective breeding and marker-assisted selection are developing fields and have remarkable potential to improve disease management and honey bee health.
- Once expression biomarkers were commercially available, and veterinarians and beekeepers may find expression biomarkers useful in the diagnostic process by rapidly identifying stressors affecting honey bee colonies in the field.
- Although the genomic solutions discussed are still in the developmental stage, veterinarians should keep themselves updated with developments in the ever-changing landscape of bee health and genomics

Chapter three: The genetics and evolution of the superorganism

Kathleen A Dogantzis[†], Tanushree Tiwari[†], Rodney T Richardson, Clement F Kent, Stephen Rose, Harshilkumar Patel, Alivia Dey, Ida M Conflitti, Renata B. Labuschagne, Shelley E Hoover, Rob W Currie, Pierre Giovenazzo, M Marta Guarna, Stephen F Pernal, Leonard J Foster, Amro Zayed[*]

[†] These authors are considered **co-first authors**

Summary:

Eusocial animals have emergent 'superorganism' traits that arise from the cooperation and coordination of large numbers of individuals. Despite decades of research, the genetics underlying superorganismal traits and how they evolved remain a mystery. Here, we sequenced the composite genomes of 1,350 honey bee colonies to identify the genes underlying a comprehensive set of ecologically and economically important colony-level traits. While genomic estimates of heritability were substantial for most colony-level traits, the underlying genetic architecture was highly polygenic with many small-effect loci. We found strong evidence of pleiotropy with several prominent neuronal genes, such as neurexin-1, neuroligins, and nicotinic acetylcholine receptors, affecting multiple colony-level traits. Finally, genes underlying colony-level traits were overwhelmingly associated with local adaptation of the honey bee's native subspecies in Europe and West Asia. Our study indicates that superorganismal traits are key for the fitness and adaptive evolution of eusocial lineages and provide a plethora of resources for improving the health of managed honey bees via genome-assisted breeding.

Introduction:

Evolution has been punctuated by several major events that fundamentally increased the complexity and biodiversity of life (Szathmáry & Smith, 1995), such as the evolution of sex, multicellularity, and eusociality. Eusociality has evolved multiple times in animals and involves division of labour between individuals that cooperate to benefit a group of related individuals (Batra, 1966; Wilson & Hölldobler, 2005). Eusocial animals have collective 'superorganismal' traits that are emergent properties of a complex system composed of hundreds to thousands of individuals interacting with one another. For example, consider a honey bee colony; the hive itself is constructed from wax that is simultaneously secreted and manipulated by thousands of

workers (Winston, 1991). Foragers collect food for the colony and are recruited to specific locations through socially transmitted information encoded in the waggle dance (Frisch, 1993). Finally, hundreds to thousands of worker bees act in concert to defend their nest against intruders, sacrificing their lives in the process (Winston, 1991). The fitness of a honey bee colony is clearly dependent on the collective action and interaction of thousands of bees.

Mapping the complex genetics of superorganismal traits has been challenging, largely because the tools typically used in solitary organisms, which link individual genomes to individual traits, are difficult to apply. However, in superorganisms, a single collective phenotype emerges from the contributions of thousands of unique genomes (Arián Avalos et al., 2020; Eynard et al., 2025; Walsh et al., 2022; Wang et al., 2013). This knowledge gap has hindered our ability to develop and test models for the evolution of eusociality, which in the absence of empirical data, often assume extremely simple genetic models for social traits (Hamilton, 1964; Osborne et al., 1997). This knowledge gap has also prevented us from effectively managing the declining health of beneficial superorganisms, such as honey bees and bumble bees, that provide indispensable pollination services to wild plants and agricultural crops (Grozinger & Zayed, 2020b). Here, we present findings from the BeeOMICs project, which used a novel method of sequencing collective ‘colony’ genomes (Arián Avalos et al., 2020) to identify the mutations and genes that underlie colony-level traits in the honey bee, the world’s most important managed pollinator and a model organism for the study of social behaviour and evolution.

The transition from solitary to colony living involves the aggregation of several individuals into a single cooperative superorganism (Hölldobler & Wilson, 2009; Szathmáry & Smith, 1995). It is postulated that selection can act at the level of the superorganism when phenotypes that benefit the group outweigh individual fitness benefits (Wilson & Wilson, 2007). In eusocial insects, such as the honey bee, workers do not typically reproduce, thus, selection is thought to act indirectly through the effects of workers on colony traits (Harpur et al., 2014b; Linksvayer, 2006; Linksvayer & Wade, 2009). Consequently, colony allele frequencies likely heavily influence social traits compared to the influence of individual worker bee genotypes (Arián Avalos et al., 2020). To study the genetics of colony level traits, our team studied 1,350 honey bee colonies from across Canada over two years and quantified 12 colony-level phenotypes related to foraging, defensive behaviour, and social and innate immunity (Figure 1). We then sequenced pools of 50 workers from these colonies at an average depth of 111.29 ± 37.16x and used these colony genomes to identify mutations and genes associated with the colony level traits.

Methods:

Experimental Design and Measurement of Phenotypes

This experiment was conducted across a two-year period within five provinces in Canada (British Columbia, Alberta, Manitoba, Ontario, and Québec), where genotypic and phenotypic data was collected for 1,350 colonies. Detailed methods were previously published (Borba et al., 2022; Richardson et al., 2023) and included in following sections. Year one of the experiment was conducted from May 2016 to April 2017 with 926 colonies established under an intensive (IM) or standard (SM) management system (IM = 380, SM = 546). Year two was conducted from May 2017 to April 2018, with 424 colonies established under IM. IM colonies were managed by research teams to allow for more accurate standardization and management protocols, and SM colonies were managed by cooperating beekeepers and management practices standardized within an apiary as per the beekeeper's discretion.

Hygienic behaviour

To assess the degree of hygienic behaviour of the colonies, we used the liquid nitrogen freeze-kill assay as previously described (M. Marta Guarna et al., 2017). This phenotype was measured twice before the honey flow in SM tier colonies (i.e., April to June) and twice after the honey flow in IM tier colonies (i.e., late August and early September). An average of the two

measurements was calculated and the phenotype was normalized using the bestNormalize function to select the best transformation according to the Pearson P statistic (Peterson, 2017; Peterson, 2021a), which in this case was ordered quantile normalization. In total there were 1,168 colonies that had both genotype and phenotype data across both years.

Defensive behaviour

To assess the degree of defensive behaviour (also known as colony aggression) of the colonies we used a flag stinging-test. This phenotype was measured twice before the honey flow in SM tier colonies (i.e., April to June) and twice after the honey flow in IM tier colonies (i.e., late August and early September). To model this trait, we focused on colonies that stung at least once and randomly selected 20 non-stinging colonies from yards that contained at least one stinging colony. An average of the two measurements was calculated and the phenotype was normalized using the bestNormalize function (Peterson, 2021a) (ordered quantile normalization). In total we used 637 colonies that had both phenotype and genotype information.

Honey Production

Honey production was measured as weight gain over a two-week period during peak honey flow (Szabo & Heikel, 1987). The total gross weight of the entire colony was recorded at the beginning of the two-week period during the early morning hours. After two weeks (13 -15 days), the colony was re-weighed. The net weight gain, calculated as the difference between the two measurements, was used for the association studies. The phenotype was normalized using the bestNormalize function (Peterson, 2021a) (ordered quantile normalization). In total, 758 colonies had both genotype and phenotype information.

Nosema spore load

Samples of 60 honey bees per colony were homogenized in 60 mL of 70% ethanol (1 mL/bee) and a 6 uL aliquot was loaded into a counting chamber with a Thoma ruling and cell depth of 0.02mm (Helber counting chamber, Hawksley, Sussex, UK). Samples were examined for *Nosema* spores twice under phase contrast 400X magnification (Eclipse Ci-L, Nikon, Tokyo, Japan) with spores counted from all 16 squares of the gridded area. The average number of spores were reported per bee per colony. The phenotype was normalized using the bestNormalize

function (Peterson, 2021a) (Log_b(x + a) normalization). In total, 1161 colonies had both genotype and phenotype information.

Trypanosomatid (Lotmaria passim) parasite load

Honey bee (60 per colony) samples were homogenized, and DNA was extracted. Levels of *L. passim* were quantified using SYBR Green real-time PCR with RPS5 as the reference gene. Standard curves generated from plasmid dilutions (10^7–10^2 copies) enabled calculation of copy numbers per bee, with amplification specificity confirmed by melt-curve analysis. Copy number per bee was normalized using the bestNormalize function (Peterson, 2021a) (ordered quantile normalization). In total, 1161 colonies had both genotype and phenotype information.

Quantification of honey bee viruses

Honey bees (60 per colony) were homogenized, and total RNA was extracted and converted to cDNA for qPCR analysis. Viral loads for BQCV, DWVA, DWVB, and SBV were quantified using SYBR Green real-time PCR with RP49 as the reference gene. Standard curves from plasmid dilutions (10^7–10^2 copies) enabled calculation of viral copy numbers, and melt-curve analysis confirmed amplification specificity. Viral copy numbers per bee was normalized using the bestNormalize function (Peterson, 2021a) applied to BQCV and SBV (asinh(x)), and DWVA and DWVB (ordered quantile normalization). In total, 1059 colonies had both genotype and phenotype information for each of the measured viruses.

Grooming behaviour

Grooming behavior was evaluated by determining the proportion of *Varroa* destructor mites collected from sticky boards that showed signs of damage. Injuries were categorized as mild (partial damage or parts of one leg missing), moderate (a whole leg or mouthparts missing), or severe (multiple legs or entire mouthparts missing). The proportion of damaged mites was calculated by dividing the number of mites showing any signs of damage by the total number of mites collected, referred to as "Grooming total" (Groom/Total; Groom/T). Additionally, the daily rate was determined by dividing the Grooming total by the number of days the sticky boards were used to collect the damaged mites, giving "Grooming per day" (Groom/Day; Groom/D). The phenotypes were normalized using the bestNormalize function

(Peterson, 2021a) (ordered quantile normalization). In total, 526 colonies had genotype and phenotype information

Mite population growth rate

Varroa mite resistance was measured using the Instantaneous Mite Population Growth Rate (IMPGR) using an established method (Harris et al., 2003). The phenotype was normalized using the bestNormalize function (Peterson, 2021a) (ordered quantile normalization). In total, 613 colonies had both genotype and phenotype information

DNA extraction and sequencing

Adult honey bees were collected from frames containing brood and immediately placed on dry ice in the field. Samples were shipped on dry ice and stored at -80°C. DNA was extracted from a pooled sample composed of the legs of 50 bees per colony. The legs were flash-frozen in liquid nitrogen and finely ground using a pestle. Tissue lysis was performed by adding 350µl of Tissue Lysis Buffer and 20µl of Proteinase K, which was heated overnight at 50°C. DNA extraction was performed using the Mag-Bind® Blood & Tissue DNA HDQ 96 Kit (Omega Bio-tek Inc., USA) optimized for the KingFisher™ Flex Purification System (Thermo Fisher Scientific Inc., USA). Samples were eluted using the Mag-Bind Kit Elution Solution (Thermo Fisher Scientific Inc., USA) to a final volume of 75µl. DNA was quantified using a NanoDrop™ 2000 Spectrophotometer (Thermo Fisher Scientific Inc., USA) and qualified using 0.8% agarose gel using electrophoresis. Library preparation and pooled genome sequencing was done on Illumina HiSeqX at Genome Quebec Innovation Centre at McGill University. Sequencing depth was an average of 111.29 ± 37.16x.

Sequence alignment and SNP discovery

Sequence reads of pooled Canadian colonies were trimmed of Illumina adaptors and low-quality bases (<20), retaining reads >50bps in length, using Trimmomatic v0.36 (Bolger et al., 2014a). Trimmed reads were aligned to the Apis mellifera reference genome (HAv3.1) (Wallberg et al., 2019) using NextGenMap v0.4.12 (Sedlazeck et al., 2013b). Bam files were sorted using SAMtools v1.3.1 (Li et al., 2009) and duplicate reads were marked using the Picard toolkit v2.1.0 (Picard toolkit, 2019). Base quality scores among BAM files were recalibrated using GATK v4.0.11.0 BaseRecalibrator (Van der Auwera et al., 2013b) using previously identified

variants. SNP and Indel detection among the pooled genomes was conducted with Lofreq v2.1.2 (Wilm et al., 2012), which uses mapped reads to determine frequencies for reference and alternative alleles. Because Lofreq can only provide information about variant sites, we used bcftools v1.16-90-g2191405 (Li, 2011) mpileup and call to verify homozygous reference genotypes, relative to missing data, among non-variant sites.

Variant filtration

Several filtration steps were used to ensure the retention of high-quality SNPs. First, we filtered individual VCF files based on the following criteria. We excluded variants located within 5-bps of an insertion-deletion site as identified through Lofreq and bcftools. We also removed variants within 5-bps of an ambiguous site, identified as heterozygous genotypes among 40 haploid drones. Additional filters based on quality and depth were determined using an outlier threshold, including SB > 50, depth > 500, depth < 10, and SNP quality < 90. Post-filtration, individual variant files were combined into a single VCF. A second set of filters were then applied to the entire dataset. We excluded loci located on the unplaced scaffolds, excluded loci with ≥90% missing data, excluded loci where <5% of samples were heterozygous, and exclude multiallelic sites. We also removed loci where the reference or alternative allele frequency distribution was zero at the 90th percentile, indicating a near fixed loci. Finally, we excluded variants that were not located among or within 5000-bps of a gene, as indicated in the annotation features (GFF) file for the *Apis mellifera* reference genome (HAv3.1) (Wallberg et al., 2019). We retained 1,742,882 SNPs for analysis, and final genotypes were represented by the frequency of the alternative allele.

Allele frequency imputation

Missing allele frequencies were imputed using the mean alterative allele frequency of closely related colonies. To calculate relatedness, we first subset loci that had no missing genotypes, and then pruned markers based on a distance threshold of 5000-bps and linkage disequilibrium (r2) threshold of 0.5, retaining loci with the greater minor allele frequency. This process identified 83,196 unlinked markers. We then calculated pairwise measures of the mean absolute genetic distance between colonies. Missing values were imputed per colony based on the average alternative allele frequency from the top ten colonies with the lowest genetic distance.

Genome wide association

Genome wide association studies (GWAS) were performed using two approaches. The first method conducted a per-loci association test between the phenotype and the alternative allele frequency using a linear mixed effects model (LMM) as implemented in GridLMM (Runcie & Crawford, 2019). The phenotype modeling using the following equation, where X is vector of the per colony alternative allele frequency. To account for population stratification between colonies, we included a genetic relatedness matrix (GRM) as a random effect. The GRM was constructed using the 83,196 unlinked markers from the imputation analysis, where a matrix organized as Colony x SNP, was mean centered and scaled, and its cross-product derived as per (Runcie & Crawford, 2019). In addition to the GRM, we corrected for additional measured and unmeasured confounders by including the apiary and sampling year (1|Yard/Year) as a random effect in the model. The phenotype was modeled using the following equation, where X is vector of the per colony alternative allele frequency and was fit using the fast algorithm and a REML estimation, which implements a Wald test for significance.

$$Phenotype \sim X + \left(1 \middle| \frac{Yard}{Year}\right) + GRM$$

For the second method we used a Bayesian sparse linear mixed model (BSLMM), as implement in GEMMA (Zhou et al., 2013), to estimate the genomic architecture of each phenotype. This model can identify major effect loci and can estimate the proportion of phenotypic variance explained by all SNPs and those with major effects. To run BSLMM we converted the allele frequency data to consensus genotypes based on the alternative allele frequency q: a consensus homozygous reference genotype was assigned when $q \leq 0.25$, a consensus heterozygous genotype was assigned when $0.25 < q < 0.75$, and a consensus homozygous alternative genotype was assigned when $q \geq 0.75$. Genotypes were converted to binary PED file format. To account for population stratification between colonies, we constructed a GRM from the 83,196 unlinked markers used in the imputation analysis. The matrix was constructed using the -gk 1 option in GEMMA which calculates the centered relatedness matrix (Zhou et al., 2013). We performed two independent runs of the model for each trait using the standard linear BSLMM, each with 20,000,000 iterations and a 10,000,000 burn-in, sampled every 10,000 iterations.

Power analysis

A regression power analysis was conducted using the using the pwrss.f.reg function in the pwrss package (Bulus, 2023) to estimate the sample size needed to determine if the predictors in the model significantly explain the variability in phenotype. In the model, the significance level was set to 0.05 and $5.0x10^{-8}$, the power threshold was set to 0.8, and the model included one predictor, representing the SNP. The analysis was carried out using the median estimated R-squared value (r2) for SNPs of interest, including SNPs with outlier associations (p-value lowest 1%) in the LMM and SNPs with a posterior inclusion probability (PIP) ≥0.01 in the BSLMM. The r2 value was estimated using a linear regression model where the alternative allele frequency of the target SNP was regressed against the normalized phenotype value using the lm function in R 3.6.3 (R Core Team, 2013).

Determining candidate SNPs

To identify candidate SNPs highly associated with our measured phenotypes, we combined the results from the per SNP linear mixed model and the Bayesian sparse regression and focused on targeting SNPs with large effects and outlier levels of association. SNPs were retained if they had a sparse effect (βγ) above the 99.9% quantile of the genome distribution, had a posterior inclusion probability (PIP) (gamma; γ) ≥0.01(Comeault et al., 2014) in the BSLMM analysis and had a significant association (p-value) below the 1% in the per SNP LMM analysis. In addition, we retained SNPs that had a PIP value ≥0.10 (Chaves et al., 2016) and a sparse effect (βγ) above the 99.9% quantile of the genome distribution, regardless of the significance of the association.

Genetic architecture

From the BSLMM models, we can extract estimates of heritability. Values of PVE estimate the proportion of phenotypic variance explained by all genotypes in the model, while PGE estimates the proportion of genetic variance explain by the major effect (sparse effect) loci (gamma). To estimate narrow sense heritability (*h2*), we derived the product of PVE and PGE to estimate the proportion of phenotypic variance explain by the major effect loci (Bresadola et al., 2019).

Functional Annotation

We used the program SNPeff v5.2a (Cingolani et al., 2012) to annotate SNPs highly associated with a phenotype among protein and non-coding gene regions, including exons, introns, and flanking sequence region. In addition, SNPeff v5.2a was used to predict the functional changes of mutations including synonymous and non-synonymous amino acid changes. The functional annotation analysis considers multiple SNP effects due to overlapping genes and differences in isoform annotation. Gene ontology (GO) enrichment was conducted with DAVID v2024q4 (Sherman et al., 2022) using *Apis mellifera* Entrez gene ID's. The target gene list included 204 pleiotropic genes, and the background list included any gene that had a SNP incorporated into the LMM and BSLMM models. GO functional annotation clusters with an enrichment score ≥1.3 and GO terms with $P < 0.05$ after Benjamini-Hochberg correction were of interest.

Phenotype predictions and cross validation

To estimate the accuracy and reliability of the candidate SNPs we conducted phenotype predictions based on alternative allele frequency values. First, to ensure variants at a locus were independently associated, we filtered loci using LD-pruning. Markers were pruned based on a distance threshold of 5000-bps and linkage disequilibrium (r2) threshold of 0.5, retaining loci with the higher minor allele frequency. With the remaining loci, we generated a multivariate generalized linear model to estimate colony phenotypes. We used 10-fold cross-validation, without replacement, to estimate model accuracy and evaluate model performance by calculating the model coefficient of determination (R2), Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the Pearson correlation (r) all calculated between the predicted values and observed values. The model was tested over 100 iterations.

Overlap with signatures of selection

We determined if genes with candidate SNPs, as determined via GWAS, were enriched for signatures of selection using two independently generated datasets. The first, used outlier measures of F_{ST} to identity high frequency, and lineage specific mutations, associated with genic regions (exons, introns, and regulatory regions) (Dogantzis et al., 2021). Comparisons were restricted to the M lineage, C lineage, and O lineage, as these lineages comprise the genetic ancestry of commercial colonies in Canada. The second study used a modified measure of

the McDonald–Kreitman (M-K) test calculated as gamma (γ), where $\gamma > 1$ indicates positive selection (Harpur et al., 2014b). Since the GWAS and selection analyses were performed on different genome versions, the analysis was restrained to comparable gene annotations. In total there were 8445 genes that had comparable GB annotations between the selection study using F_{ST} and the GWAS analysis, while there were 8080 genes that had comparable GB annotations between the selection study using γ and the GWAS analysis. Enrichment was calculated using a Fisher Exact test in R 3.6.3 (R Core Team, 2013). When needed, *P*-values for each trait were corrected for false discovery rate (FDR) using the Benjamini-Hochberg correction in R 3.6.3.

Results and Discussion:

Colony phenotypes are highly polygenic

We modeled the consensus genotype of honey bee colonies using a Bayesian sparse linear mixed model (BSLMM) to shed light on the genetic architecture of important social phenotypes (Table 3.1). Several phenotypes were found to be highly polygenic, governed by many loci with small effects. For example, hygienic behaviour was estimated to be influenced by approximately 181 loci, suggesting a complex genetic architecture. These variants account for 29% of the genetic variance (proportion of genetic variance; PGE) and influence 23% of the overall phenotypic variance (narrow sense heritability; h^2) (Figure 3.2A; Table 3.1). Phenotypes associated with viral and parasite load were estimated to be influenced by an average of 171 loci, that account for, on average, 52% of the genetic variance (PGE) and influence 38% of the phenotypic variance (h^2) (Figure 3.2A; Table 3.1). This includes grooming behaviour, where loci account for 51% of the phenotypic variance (h^2). Loci associated with defensive behaviour had the lowest effect among the measured traits, accounting for, on average, 20% of the phenotypic variance (h^2) (Figure 3.2A; Table 3.1).

While there is clear evidence for moderate heritability for colony-level traits, individual SNPs typically have very subtle effects on such phenotypes. We found no significant marker-trait associations using a per SNP linear mixed-effects model (LMM) at the genomic threshold ($\alpha < 5x10^{-8}$), or after applying a multiple test correction. The median R^2 value, representing the phenotypic variance explained by a SNP, was low, even among SNPs in the bottom 1% of p-values (i.e., those with the strongest associations), which ranged from $3.52x10^{-3}$ to $1.41x10^{-2}$ (Table 3.2). The complex nature of the phenotypes measured, along with the absence of strong

associations identified using LMMs was congruent with the BSLMM analysis, suggesting that colony traits are influenced by many variants with small effect. This pattern suggests the need for much larger sample sizes to detect significant association (likely n=6,000 colonies), though an average sample size of 1,188 provides sufficient power (80%) to identify associations at an alpha threshold of 0.05 (Table 3.2). Despite limited power, we were able to successfully model and predict the phenotype for some of the more heritable traits (Table 3.3). On average, candidate SNPs explained 55% of the phenotypic variants, meaning more the half of the differences among colonies could be predicted from just a few variants. Notably, the model for grooming performed especially well, accounting for 83% of the observed variance. These results provide confidence in the robustness of significant SNP associations identified from the BSLMM and LMM models

Phenotype associated markers are found in non-coding regions

The phenotypes measured in our study are complex and governed by numerous small effect loci, limiting our power to detect strongly associated loci. Thus, to identify SNPs that may be highly associated with our phenotypes, we combined the results from the per SNP linear mixed-effects model and the Bayesian sparse linear mixed model, focusing on SNPs with large effects and outlier levels of association (see methods). Using this approach, we refined our results to a subset of SNPs associated with the measured phenotypes (Figure 3.2B).

SNPs associated with phenotypes were linked to 145 genes related to grooming and hygienic behavior, an average of 110 genes related to viral load, and 39 genes associated with defensive behavior. On average, 82% of the genes were considered protein-coding, while a notable 16% were classified as non-coding, primarily long non-coding RNA (lncRNA) (Figure 3.2C). Mutations among genes were concentrated among introns, representing on average 67% of functional annotations, and upstream and downstream of genes, accounting for on average 18% and 12% of functional annotations, respectively (Figure 3.2C). We found significant enrichment of mutations within introns of genes associated with mite growth rate (IMPGR) and *Nosema* parasite loads ($P = 2.81\text{x}10^{-3}$), upstream of genes associated with DWVB ($P = 2.3\text{x}10^{-3}$), and upstream and downstream of genes associated with total grooming and hygienic behaviour ($P < 0.045$).

The identification of SNPs in non-coding regions, such as lncRNA and introns, is common in GWAS studies (Giral et al., 2018; Jo & Choi, 2015) and prevalent in honey bee

quantitative genetic studies, which have linked these markers to defensive behaviour (Arián Avalos et al., 2020; Harpur et al., 2020), calmness and gentleness (Guichard et al., 2021), *Varroa* resistance and recapping (Eynard et al., 2025; Guichard et al., 2022), hygienic behaviour (Harpur et al., 2019), and trophallaxis (Traniello et al., 2025). LncRNAs influence gene regulation through chromatin modification, direct transcription effects, and post-transcriptional processes like splicing, with demonstrated roles in honey bee reproduction (Chen & Shi, 2020; Humann et al., 2013) and heat stress (Zhang et al., 2024). Similarly, introns contribute to gene regulation via intron-mediated enhancement, alternative splicing, and often harbour non-coding RNA (Jo & Choi, 2015; Rose, 2019). This pattern may suggest that complex social phenotypes are likely to be modulated by intricate gene regulation mechanisms, as previously hypothesized (A L Toth & G E Robinson, 2007), though SNPs in non-coding regions might also be linked to causal loci in protein-coding regions.

Pleiotropic neuronal genes drive the evolution of superorganismal traits

Our analysis identified 204 genes with pleiotropic effects across two or more colony-level traits. Pairwise comparisons between phenotypes revealed an average of six shared genes, with 75% of comparisons showing statistically significant overlap ($P < 0.048$, FDR-adjusted) (Figure 3.3A; Table 3.4). Pleiotropic genes were enriched for gene ontology (GO) terms related to neurotransmission, including chemical synaptic transmission (GO:0007268, $P = 2.6\text{x}10^{-5}$), pre (GO:0097105, $P = 5.3\text{x}10^{-3}$) and post (GO:0097104, $P = 5.3\text{x}10^{-3}$) synaptic membrane assembly, and signalling receptor activity (GO:0038023, $P = 2.7\text{x}10^{-2}$) (Table 3.5).

Notable genes include four nicotinic acetylcholine receptors (*nAChRa1, nAChRa4, nAChRa6,* and *nAChRa7*) (Figure 3B). These receptors have been linked to maintaining immune homeostasis (Chavan et al., 2017). In *Drosophila* (Giordani et al., 2023) and *A. mellifera* (Brandt et al., 2016; Christen & Fent, 2017; Di Prisco et al., 2013), disruption of acetylcholine receptors has been linked to a suppression of innate immunity. Alternative alleles within these genes were associated with resistance to, BQCV DWVA, reduced *Varroa* mite population growth, and increased honey production and defensive behaviour. Interestingly, defensive behavior in honey bees has previously been linked to pesticide resilience and mite loads (Rittschof et al., 2015). These findings suggest that immune function and colony-level defensive behaviour may both be coordinated by nicotinic acetylcholine receptors.

Genes significantly associated with colony traits also included neurexin-1 (*Nrx-1*) and four neuroligins (*Nlg-1, Nlg-3, Nlg-4*, and *Nlg-5*) (Figure 3.3B), which are cell adhesion molecules and synaptic binding partners that regulate neural signaling (Biswas et al., 2008; Craig & Kang, 2007; Knight et al., 2011). In honey bees, these genes are highly expressed in the adult mushroom bodies (Biswas et al., 2008) and respond to sensory input (Biswas et al., 2010), influencing behaviors such as odor-reward learning and memory (Grosso et al., 2018), and grooming in response to *Varroa (Arechavaleta-Velasco et al., 2012; Hamiduzzaman et al., 2017; Russo et al., 2024)*. In our study, alternative alleles in these genes were associated with reduced levels of BQCV, DWVB, SBV, defensive behavior, and increased honey production and hygienic behavior. Sensitivity to olfactory cues has been shown to be a key mediator of hygienic behavior, enabling workers to detect and remove dead or diseased brood (Guarna et al., 2015; Masterman et al., 2001; Spivak & Danka, 2021). As such, these genes may underlie the neural mechanisms through which complex disease signals are interpreted and translated into coordinated social immune responses within the colony.

Additional genes of interest (Figure 3.3B) include *neurotrimin*, which encodes an immunoglobulin domain and has been found under selection in managed colonies (Parejo et al., 2020). Alternative alleles within these genes were associated with increased hygienic and grooming behavior and reduced DWVA prevalence. We also identified alternative alleles in *Dop3*, which encodes a dopamine receptor (Beggs et al., 2005), associated with reduced DWVB levels and mite population growth. *Dop3* has previously been linked to mite suppression through hygienic behavior (Tsuruda et al., 2012), likely due to its expression in the optic and antennal lobes (Beggs et al., 2005), where it may contribute to processing sensory cues. Finally, we found *serine proteinase stubble* to be associated with defensive behavior and *Lotmaria* levels, as well as hygienic behaviour. This gene was previously mapped to a QTL on chromosome 15 associated with *Varroa* resistance (Conlon et al., 2018).

Although we only highlight a small subset of genes here, our study highlights a complex interconnected genetic architecture that underlie multiple, often correlated (Borba et al., 2022), colony-level traits. While in some instances, genes with pleiotropic effect involve simultaneous positive effects on colony level traits, the majority of pleiotropic genes involve trade-offs (Figure 3.2B) that may limit that ability of natural or artificial selection to fully optimize colony level

traits. For example, some alternative alleles in the nicotinic acetylcholine receptors discussed previously, are associated with reduced hygiene, an increase in *Lotmaria* infection, and increased mite population growth. Our results underscore the importance of considering pleiotropy in honey bee breeding programs.

The evolution of superorganismal genes

In eusocial honey bees, signatures of selection have been shown to be enriched among the genes with worker-biased expression (Dogantzis, 2021; Dogantzis et al., 2021; Harpur et al., 2014b) despite their inability to lay fertilized eggs, suggesting there are indirect fitness benefits associated with worker phenotypes. Here, we directly test the hypothesis that natural selection acts on genes underlying colony level traits using two large population genomic studies (Dogantzis et al., 2021; Harpur et al., 2014b).

We find that genes associated with superorganismal traits are significantly enriched for signatures of positive selection associated with the honey bee's local adaptation, as evidenced by extreme lineage specific genetic differences (outlier F_{ST}, $P < 0.037$ FDR) (Figure 3.4; Table 3.6). Genes associated with superorganismal traits had, on average, a 144% ± 42% enrichment of signatures of positive selection, relative to expected values. However, we do not find enrichment of positive selection when comparing historical macroevolution occurring between *A. mellifera* and its sister species, *A. cerana* as measured by the M-K test ($P > 0.07$ FDR). This pattern is appropriate as genetic changes, identified by F_{ST}, likely highlight patterns of recent spatial and temporal adaptations (Vitti et al., 2013), whereas the M-K test identifies genetic changes associated with divergence from a common ancestor (McDonald & Kreitman, 1991; Vitti et al., 2013). Additionally, F_{ST} can identify differentiation among non-coding regions, which we have been shown to contain most of the phenotype associated SNPs in this study, while the M-K test only estimates selection acting on non-synonymous mutations.

Our work provides an important glimpse into the genetics and evolution of colony level traits. We demonstrate that colony-level traits are often controlled by a large number of loci with relatively small effects. As predicted by phenotypic studies, we also uncovered a large number of pleotropic genes that influence multiple traits; a finding that is important, not only for understanding the nature of genes affecting social traits, but also for attempting to improve the health of managed honey bees via genetically assisted breeding. Finally, we found that loci

affecting colony-level traits are enriched among genes that are associated with local adaptation of native honey bee lineages in the old world. To our knowledge, this represents the strongest evidence that superorganismal traits contribute to the adaptive evolution of eusocial animals a central tenant of sociobiology.

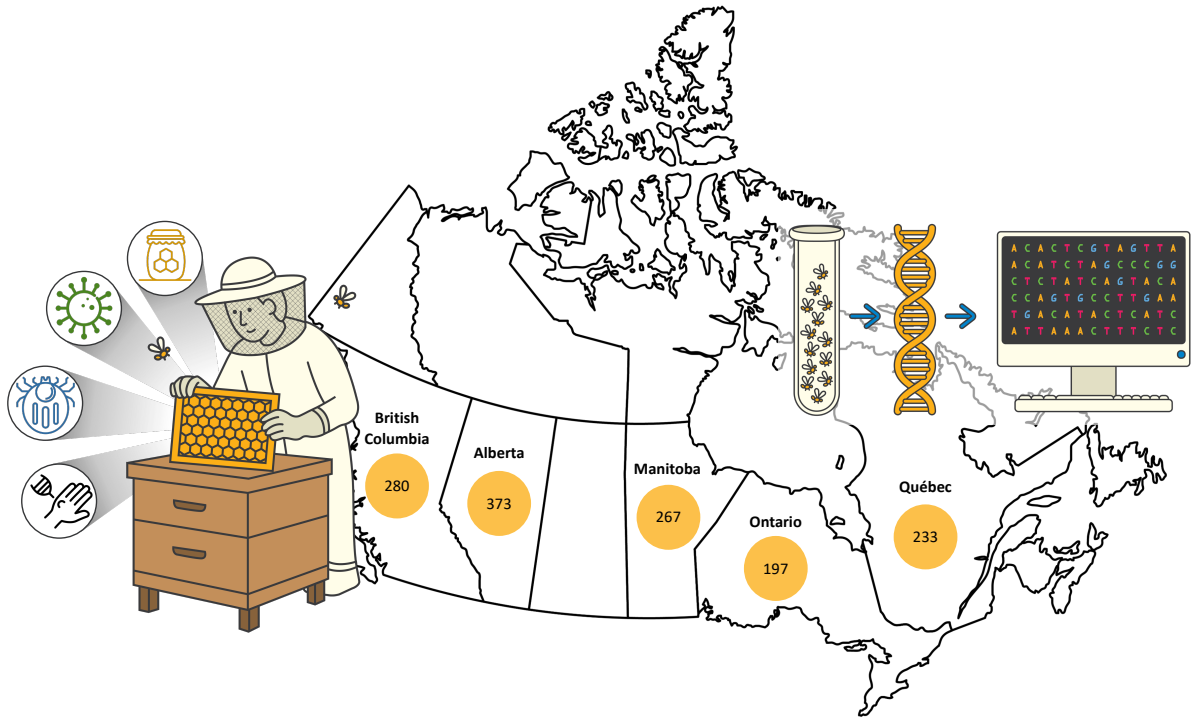


Figure 3.1: Study overview. Twelve phenotypes were measured in 1,350 Canadian honey bee colonies across five provinces over two years (N=926 year one, N=424 year two). Circles on the map indicate the number of colonies sampled per province. DNA was extracted from pooled samples of 50 bees per colony, yielding one whole genome per colony. SNP variants were analyzed using a Bayesian sparse linear mixed model and a linear mixed effect model to identify mutations and genes associated with colony-level traits.

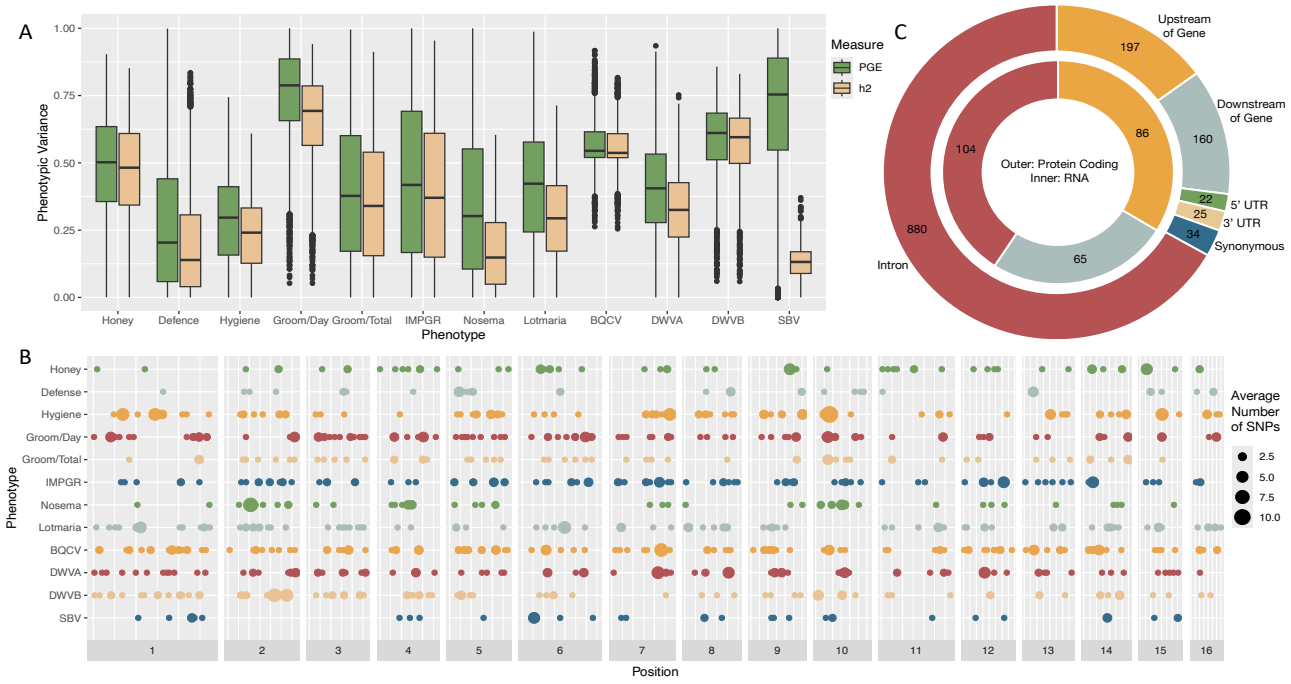


Figure 3.2: Genomic architecture underlying colony-level phenotypes in honey bees. **A)** Bayesian sparse linear mixed models (BSLMM) were used to infer the proportion of phenotypic variance (PGE) attributable to sparse SNP effects (i.e., those with the largest effects), and the narrow-sense heritability (h^2). Boxplots depict measures from the posterior distribution that was sampled every 10,000 iterations. **B)** Genomic positions of single nucleotide polymorphisms (SNPs) with strong associations to the phenotype, as identified through an integrative approach combining results from the per-SNP linear mixed models and the BSLMM model. The size of the data points indicates the average number of SNPs found within 500Kb intervals. **C)** Distribution of phenotype-associated SNPs within annotated genes, including introns, exons, and UTRs, and within 5,000 bp upstream or downstream of annotated genes. The outer ring displays the SNP counts within protein-coding genes, while the inner rings show SNP counts within non-coding RNAs. Coloured segment labels of the outer ring match those of the inner ring. All acronyms defined in methods.

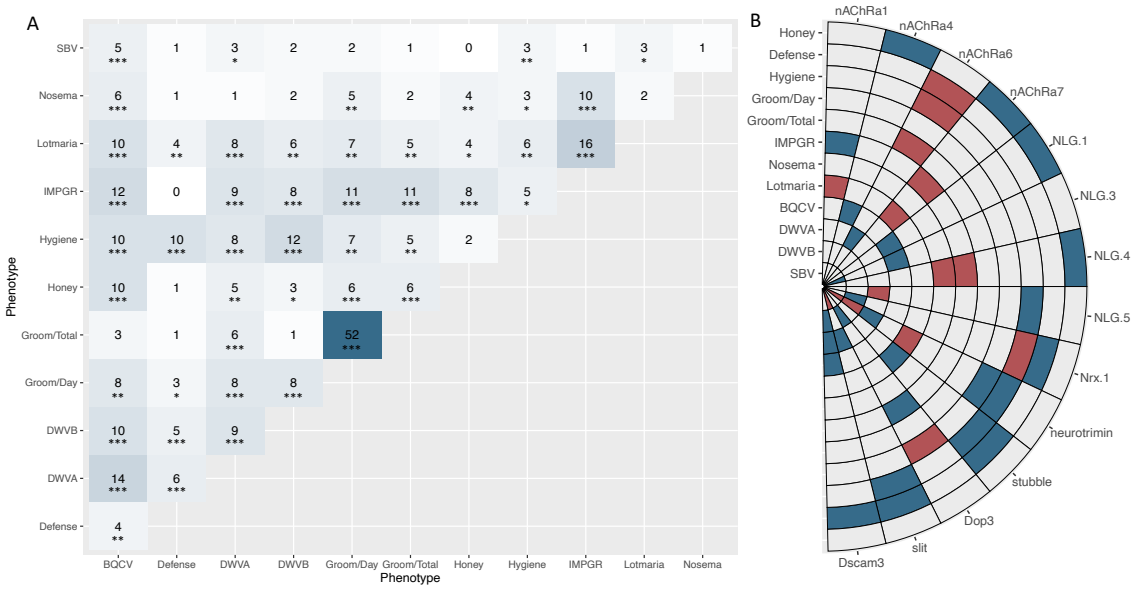


Figure 3.3. Evidence of pleiotropy among genes associated with colony-level phenotypes. **A)** Pairwise comparisons across phenotypic traits revealed an average of seven genes shared between trait pairs, indicating widespread pleiotropic effects. *FDR* < 0.05 (*), *FDR* < 0.01 (**), *FDR* < 0.001 (***) **B)** Visualization of pleiotropic effects for a subset of genes of particular interest. Blue indicates that the alternative allele 'improves' the trait from a beekeeping perspective (i.e. more honey, less pathogens), while red indicates the opposite. Effects were estimated from the beta values generated from the per-SNP linear mixed effect model. Aggregated effects were used in cases of multiple associated SNPs per gene per phenotype.

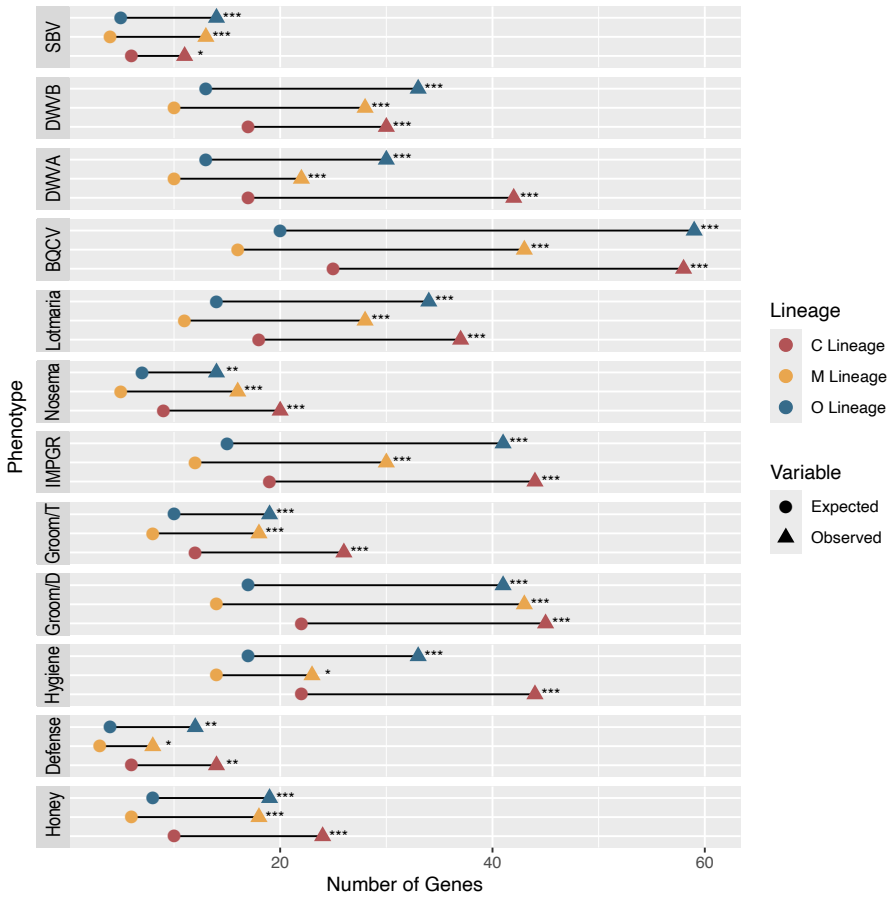


Figure 3.4: Genes associated with colony-level traits are enriched for signatures of selection. Genes containing candidate SNPs identified via GWAS exhibited significant enrichment for signatures of selection, as measured by extreme genetic differentiation (Chouvenc et al.) among the M (Western Europe), C (Eastern Europe) and O (Western Asia) lineage of honey bees, which comprise the genetic composition of Canadian honey bee colonies studied herein. Enrichment was assessed relative to genome-wide expectations, and statistical significance was determined using a Fisher's exact test with false discovery rate (FDR) correction. Asterisks indicate significance thresholds: $FDR < 0.05$ (*), $FDR < 0.01$ (**), $FDR < 0.001$ (***).

Data Tables:

Table 3.1: Estimated genomic architecture of honey bee traits. A Bayesian sparse linear mixed model (BSLMM) was used to estimate the proportion of phenotypic variance (PVE),

proportion of genetic variance sparse effect loci (PGE), and the estimated number of variants with sparse effects (gamma). Narrow sense heritability (h^2) was derived from the product of mean PVE and PGE.

Phenotype	Mean PVE (SD)	Mean PGE (SD)	h^2	Mean gamma (SD)
BQCV	0.99 (0.03)	0.56 (0.08)	0.55	253 (34)
Defensive Behaviour	0.70 (0.14)	0.28 (0.26)	0.20	88 (81)
DVWB	0.97 (0.01)	0.59 (0.13)	0.57	246 (14 et al.)
DWVA	0.81 (0.04)	0.41 (0.18)	0.33	222 (64)
Honey Production	0.96 (0.02)	0.49 (0.18)	0.47	156 (73)
Hygienic Behaviour	0.81 (0.05)	0.29 (0.17)	0.23	181 (88)
IMPGR	0.89 (0.06)	0.44 (0.31)	0.39	160 (99)
Lotmaria infection	0.71 (0.07)	0.41 (0.22)	0.29	175 (91)
Grooming per day	0.88 (0.09)	0.75 (0.18)	0.66	190 (72)
Grooming total	0.90 (0.09)	0.40 (0.26)	0.36	150 (89)
Nosema infection	0.49 (0.11)	0.35 (0.26)	0.17	120 (104)
SBV	0.19 (0.06)	0.69 (0.25)	0.13	25 (34)

Table 3.2: Power analysis to estimate the approximate sample size needed to detect a statistically significant association with the given r2 value, with statistical power of 0.8. Two sample sizes were calculated given an α threshold of 0.05 and $5x10^{-8}$, the later often consider then genomic threshold. The median R-squared value (r^2) was calculated for SNPs of interest, including SNPs with outlier associations (p-value lowest 1%) in the per SNP linear mixed model (LMM) and SNPs with a posterior inclusion probability (PIP) ≥0.01 in the BSLMM, to estimate the effects of individual loci on the trait and to determine model power.

Phenotype	Median r^2 of $P < 1\%$	Sample size for α = 0.05	Sample size for α = 5E8	Median r^2 of PIP > 0.01	Sample size for α = 0.05	Sample size for α = 5E8
BQCV	3.52E-03	2227	11237	2.10E-03	3736	18853

Defensive Behaviour	9.89E-03	795	4017	1.58E-02	485	2451
DVWB	8.41E-03	928	4684	9.15E-03	853	4304
DWVA	4.18E-03	1874	9457	7.92E-03	986	4977
Honey Production	4.39E-03	1778	8976	1.06E-02	742	3747
Hygienic Behaviour	8.34E-03	940	4747	5.37E-03	1448	7309
IMPGR	8.68E-03	900	4543	4.10E-03	1907	9626
Lotmaria infection	4.53E-03	1728	8722	5.78E-03	1354	6833
Grooming per day	1.12E-02	698	3526	9.12E-03	855	4316
Grooming total	1.41E-02	551	2785	1.27E-02	612	3092
Nosema infection	6.22E-03	1258	6353	1.21E-02	643	3250
SBV	1.05E-02	745	3763	1.61E-02	482	2435

Table 3.3: Performance metrics of models used to predict honey bee traits. The number of features indicates the number of informative SNPs used in the model. The SNPs were chosen based results from the per SNP linear mixed model test and the Bayesian sparse linear mixed and pruned based on distance and linkage metrics to reduced collinearity. Model coefficient of determination (R^2), Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the Pearson correlation (*r*) calculated between the predicted values and observed values.

Phenotype	# of features	R^2	MAE	RMSE	*r*
Hygienic Behaviour	102	0.584	10.967	14.242	0.67
Honey Productions	63	0.676	7.693	10.811	0.759
Defensive Behaviour	34	0.456	9.827	14.938	0.612
IMPGR	98	0.699	4.40E-02	6.48E-02	0.721
Nosema	48	0.406	1.22E+05	3.68E+05	0.505
Lotmaria	97	0.544	2.83E+06	7.97E+06	0.481
BQCV	142	0.187	4.99E+07	1.02E+08	0.151

SBV	31	0.283	4.40E+18	4.37E+19	0.286
DWVA	93	0.579	8.10E+08	3.57E+09	0.588
DWVB	89	0.692	1.95E+09	8.18E+09	0.625
Mite Drop Total	57	0.668	9.94E-03	1.53E-02	0.698
Mite Drop Per Day	118	0.831	2.81E-03	4.38E-03	0.782

Table 3.4: Pairwise comparisons between phenotypes of genes that overlapped in their association with traits. Significance of overlap was assessed with a Fisher's Exact test and adjusted for false discovery rate (FDR) using Benjamini-Hochberg correction.

Phenotype	Phenotype	Overlap	FisherOdd	P-value	FDR
Defense	Groom/Total	1	3.46585756	0.25752208	0.27469022
Defense	Honey	1	4.06298327	2.25E-01	2.48E-01
Defense	*Nosema*	1	5.42626579	1.74E-01	1.99E-01
Defense	SBV	1	7.55198517	1.29E-01	1.53E-01
DWVA	*Nosema*	1	1.81331746	0.42989792	0.43672169
DWVB	Groom/Total	1	1.17935067	0.57589471	0.57589471
Groom/Total	SBV	1	3.21148667	0.27436474	0.28785809
IMPGR	SBV	1	2.48007525	0.3388764	0.34980789
Nosema	SBV	1	5.02743457	0.18649043	0.20939276
DWVB	*Nosema*	2	3.79271555	0.10407797	0.12809596
DWVB	SBV	2	5.31703415	5.97E-02	7.79E-02
Groom/Day	SBV	2	4.07837453	9.32E-02	1.17E-01
Groom/Total	*Nosema*	2	4.74928804	0.07170685	0.09178477
Honey	Hygiene	2	2.16897877	0.24270697	0.26327536
Lotmaria	*Nosema*	2	3.62654363	0.1119128	0.13513998
BQCV	Groom/Total	3	2.35101717	1.44E-01	1.68E-01
Defense	Groom/Day	3	6.84563749	0.01188263	0.01728383
DWVA	SBV	3	8.10298059	7.58E-03	1.15E-02

DWVB	Honey	3	4.33781728	0.03628947	0.04838595
Hygiene	*Nosema*	3	4.47466965	0.03394439	0.04622215
Hygiene	SBV	3	6.3179855	0.01453578	0.0202237
Lotmaria	SBV	3	7.88793658	0.00813506	0.012108
BQCV	Defense	4	7.8704405	2.42E-03	4.08E-03
Defense	*Lotmaria*	4	11.8236825	5.68E-04	1.18E-03
Honey	*Nosema*	4	11.8835681	0.00053561	0.00118203
Honey	*Lotmaria*	4	5.65286745	0.00701764	0.01095436
BQCV	SBV	5	9.35831192	0.00034358	0.00080238
Defense	DWVB	5	16.071872	3.07E-05	9.82E-05
DWVA	Honey	5	7.42936437	0.00085852	0.00161604
Groom/Day	*Nosema*	5	7.84922316	6.96E-04	1.39E-03
Hygiene	Groom/Total	5	4.88159985	0.00488836	0.00782137
Hygiene	IMPGR	5	3.73034803	0.01387821	0.0197379
Lotmaria	Groom/Total	5	6.1161866	0.00193984	0.0034486
BQCV	*Nosema*	6	8.02933218	0.00019143	0.00049006
Defense	DWVA	6	19.6789673	1.75E-06	7.99E-06
DWVA	Groom/Total	6	7.70394618	0.00022405	0.0005515
DWVB	*Lotmaria*	6	5.92781293	0.00083181	0.0016132
Honey	Groom/Day	6	7.06950182	0.00035104	0.00080238
Honey	Groom/Total	6	11.752139	2.49E-05	8.40E-05
Hygiene	*Lotmaria*	6	4.50834833	0.00310896	0.00510188
Hygiene	Groom/Day	7	4.25129047	0.0020118	0.00347988
Lotmaria	Groom/Day	7	5.34648806	0.0005736	0.0011842
BQCV	Groom/Day	8	4.0837748	1.28E-03	2.34E-03
DWVA	Groom/Day	8	6.38816594	7.63E-05	2.04E-04
DWVA	Hygiene	8	6.38816594	7.63E-05	0.00020357
DWVA	Lotmaria	8	8.04771304	1.63E-05	5.81E-05

DWVB	Groom/Day	8	6.51209301	6.73E-05	0.00019585
DWVB	IMPGR	8	8.28039804	1.35E-05	5.07E-05
Honey	IMPGR	8	12.5144973	7.92E-07	4.61E-06
DWVA	DWVB	9	9.67757016	1.23E-06	6.57E-06
DWVA	IMPGR	9	9.31176423	1.65E-06	7.99E-06
BQCV	DWVB	10	6.95898853	5.31E-06	2.27E-05
BQCV	Honey	10	10.6236546	1.60E-07	1.05E-06
BQCV	Hygiene	10	5.24512915	5.12E-05	1.56E-04
BQCV	*Lotmaria*	10	6.63209413	7.86E-06	3.14E-05
Defense	Hygiene	10	29.7673587	2.58E-11	5.50E-10
IMPGR	*Nosema*	10	22.9517498	1.92E-10	3.06E-09
IMPGR	Groom/Day	11	9.06494215	1.65E-07	1.05E-06
IMPGR	Groom/Total	11	15.4441411	1.06E-09	1.13E-08
BQCV	IMPGR	12	8.29114856	1.18E-07	9.42E-07
DWVB	Hygiene	12	10.4652671	1.10E-08	1.01E-07
BQCV	DWVA	14	10.1902414	1.03E-09	1.13E-08
IMPGR	*Lotmaria*	16	18.4107325	2.17E-14	6.95E-13
Groom/Day	Groom/Total	52	170.676792	2.97E-79	1.90E-77

Table 3.5: Significant gene ontology (GO) terms for 204 pleiotropic genes, as determined with DAVID v 2024q4.

Term	PValue	Benjamini
GO:0045211~postsynaptic membrane	1.80E-07	1.33E-05
GO:0045202~synapse	6.58E-07	2.43E-05
GO:0007268~chemical synaptic transmission	1.23E-07	2.63E-05
GO:0007411~axon guidance	2.70E-06	2.90E-04
GO:0050804~modulation of chemical synaptic transmission	2.63E-05	0.00188776

GO:0030154~cell differentiation	8.85E-05	0.00475429
GO:0007158~neuron cell-cell adhesion	1.74E-04	0.00534624
GO:0097105~presynaptic membrane assembly	1.74E-04	0.00534624
GO:0097104~postsynaptic membrane assembly	1.74E-04	0.00534624
GO:0043005~neuron projection	4.06E-04	0.01000419
GO:0005886~plasma membrane	7.58E-04	0.01403112
GO:0005911~cell-cell junction	0.00116214	0.01680239
GO:0009986~cell surface	0.00136236	0.01680239
GO:0098609~cell-cell adhesion	9.72E-04	0.02612449
GO:0038023~signaling receptor activity	1.57E-04	0.02764413

Table 3.6: Enrichment statistics for phenotypes associated with outlier SNP genes across M, C, and O ancestral lineages. Observed values are the number of genes associated with outlier SNPs among each trait, while Expected values are the expected number of overall genes. OR: odds ratio of the Fisher Exact test, P-value: uncorrected p-value calculated from the Fisher Exact test. FDR correction was made using the Benjamini-Hochberg method calculated per trait, and per lineage. Percent change is the increase in the observed from expected value.

Trait	Lineage	Observed	Expected	OR	*P*-value	FDR *P*-value per trait	FDR p-value per lineage	Percent Change
BQCV	M Lineage (Europe)	43	16	3.72	1.9740E-10	1.97E-10	1.18E-09	174%
BQCV	C Lineage	58	25	3.44	5.2560E-11	7.88E-11	6.31E-10	130%
BQCV	O Lineage	59	20	4.70	3.5530E-16	1.07E-15	4.26E-15	193%
Defensive Behaviour	M Lineage (Europe)	8	3	2.79	1.8300E-02	1.83E-02	1.83E-02	167%

Defensive Behaviour	O Lineage	12	4	3.88	7.1400E-04	1.07E-03	7.79E-04	200%
Defensive Behaviour	C Lineage	14	6	3.91	4.0580E-04	1.07E-03	4.87E-04	133%
DWVA	M Lineage (Europe)	22	10	2.54	6.0450E-04	6.05E-04	7.25E-04	112%
DWVA	O Lineage	30	13	2.97	7.1730E-06	1.08E-05	1.43E-05	125%
DWVA	C Lineage	42	17	4.05	5.8340E-10	1.75E-09	3.50E-09	152%
DWVB	C Lineage	30	17	2.25	8.0780E-04	8.08E-04	8.81E-04	80%
DWVB	M Lineage (Europe)	28	10	3.59	4.2420E-07	6.36E-07	1.35E-06	170%
DWVB	O Lineage	33	13	3.46	1.8680E-07	5.60E-07	5.60E-07	148%
Groom/Day	C Lineage	45	22	2.81	4.2650E-07	4.27E-07	1.28E-06	106%
Groom/Day	O Lineage	41	17	3.12	6.4510E-08	9.68E-08	2.58E-07	134%
Groom/Day	M Lineage (Europe)	43	14	4.63	9.3220E-13	2.80E-12	1.12E-11	216%
Groom/Total	M Lineage (Europe)	18	8	2.95	3.1790E-04	0.0003179	4.24E-04	137%
Groom/Total	O Lineage	19	10	3.49	6.4410E-05	0.0001037	1.10E-04	95%
Groom/Total	C Lineage	26	12	2.96	6.9100E-05	0.0001037	9.55E-05	113%
Honey Production	O Lineage	19	8	3.08	3.4500E-04	3.45E-04	4.14E-04	138%
Honey Production	C Lineage	24	10	3.48	2.5670E-05	3.86E-05	4.40E-05	140%
Honey Production	M Lineage (Europe)	18	6	3.80	2.5710E-05	3.86E-05	4.11E-05	200%

Hygienic Behaviour	M Lineage (Europe)	23	14	1.88	1.2290E-02	1.23E-02	1.34E-02	64%
Hygienic Behaviour	O Lineage	33	17	2.28	1.9080E-04	2.86E-04	2.54E-04	94%
Hygienic Behaviour	C Lineage	44	22	2.70	1.5660E-06	4.71E-06	3.76E-06	100%
IMPGR	M Lineage (Europe)	30	12	3.39	4.5150E-07	4.52E-07	1.35E-06	159%
IMPGR	C Lineage	44	19	3.59	3.8380E-09	5.76E-09	1.54E-08	137%
IMPGR	O Lineage	41	15	4.17	1.2950E-10	3.89E-10	7.77E-10	176%
Lotmaria	C Lineage	37	18	2.89	2.7970E-06	2.80E-06	5.59E-06	110%
Lotmaria	M Lineage (Europe)	28	11	3.31	1.5410E-06	2.31E-06	3.70E-06	155%
Lotmaria	O Lineage	34	14	3.32	2.5880E-07	7.76E-07	6.21E-07	141%
Nosema	O Lineage	14	7	2.57	5.6410E-03	5.64E-03	5.64E-03	106%
Nosema	C Lineage	20	9	3.50	7.1590E-05	1.07E-04	9.55E-05	135%
Nosema	M Lineage (Europe)	16	5	4.22	2.7410E-05	4.11E-05	4.11E-05	202%
SBV	C Lineage	11	6	2.30	3.6600E-02	3.66E-02	3.66E-02	83%
SBV	O Lineage	14	5	4.59	7.3390E-05	1.10E-04	1.10E-04	192%
SBV	M Lineage (Europe)	13	4	5.32	2.6780E-05	8.03E-05	4.11E-05	248%

Supplementary Methods:

Experimental Design and Measurement of Phenotypes

IM colonies were established as single brood chambers split from parent colonies using four frames of predominantly unsealed brood, three frames of honey, three frames with empty comb, 1kg of bees, and a 1-year-old queen. IM colonies were managed by research teams to allow for more accurate standardization and management protocols. Colonies were prevented from swarming by providing honey supers and removing swarm cells. Supplemental feeding was provided in the spring (1:1 v/v sucrose syrup and 0.5 kg pollen patties containing 15% pollen; Global Patties Canada, Airdrie, AB, Canada) and fall (1:1 v/v sucrose syrup), and all colonies were overwintered indoors at approximate 5°C. Colonies were treated for Varroa mites or Nosema after phenotype measurements and before overwintering. SM colonies were managed by cooperating beekeepers and were established using either one-year-old queens or newly bred queens that had been laying for at least 50 days. Management practices were under the cooperating beekeeper's discretion but were standardized within an apiary. IM colonies followed the same protocol established in the prior year, except for overwintering techniques in British Columbia and Ontario where colonies were kept outdoors.

Hygienic behaviour

Hygienic behaviour was measured using queenright colonies, where comb with a solid patch of sealed worker brood at the pupal stage was collected. A 5cm diameter PVC pipe was pressed down to the midrib of the comb, using a twisting motion to seal the connection. Liquid nitrogen was then poured into the tube, to freeze-kill the brood. After the liquid nitrogen boiled off, the wax was allowed to thaw before removing the tube. The number of sealed brood cells within the 5 cm diameter was recorded, and the frame was returned to the colony. After 24 hrs, the proportion of pupae completely removed was counted.

Defensive behaviour

Defensive behaviour was measured using queenright colonies, where a black suede leather patch (7.6 cm x 7.6 cm) was swung across the brood chamber 5 cm from the frames. The swinging action was repeated over a two-minute period, after which the patch was sealed in a bag. The total number of stings in the patch was recorded.

Trypanosomatid (Lotmaria passim) parasite load

Samples of 60 honey bees per colony were homogenized in 60 mL of 70% ethanol. A 200 µL aliquot from the homogenized sample was centrifuged and pelleted. The 70% ethanol was aspirated with the remaining left to evaporate from the pellet, which was left to dry at room temperature. DNA was extracted using the NucleoSpin®Tissue kit following manufacturer instructions (Macherey-Nagel Gmbh & Co. KG, Düren, Germany). Quantification of L. passim infection levels were determined by real-time PCR using primers and SSoAdvanced™ Universal SYBR® Green Supermix (Bio-Rad Laboratories, Hercules, USA). Amplification assays were performed by triplicate employing 60ng of genomic DNA in a CFX384 Touch™ Real-Time Detection System (Bio-Rad Laboratories, Hercules, USA). RPS5 (Adnane et al.) and was chosen as reference gene. Standard curves were prepared from plasmids harboring the target amplicons with copy numbers diluted from 107 to 102. PCR conditions were 3 min at 98°C for initial denaturation/enzyme activation followed by 40 cycles of 10 seconds at 98°C and 20 seconds at 60°C. Specificity was checked by performing a melt-curve analysis 65-95°C increments of 0.5°C 2 seconds/step. Results were analyzed with the CFX Manager™ Software and exported to an Excel spreadsheet to calculate copy numbers per bee per colony.

Quantification of honey bee viruses

Samples of 60 honey bees per colony were homogenized in 12mL of GITC buffer (Jay D. Evans et al., 2013). An aliquot of 200µL was used to isolate total RNA using the NucleoSpin®RNA kit following manufacturer instructions (Macherey-Nagel Gmbh & Co. KG, Düren, Germany). cDNA was synthesized from 800ng of total RNA for 20 minutes at 46°C in a final volume of 20 µL using the iScript cDNA synthesis kit (Bio-Rad laboratories, Hercules, USA). cDNA was diluted with 60µL of molecular grade water to a total of 80µL from which 3µL were used for qPCR quantification. Quantification of BQCV, DWVA, DWVB, and SBV infection levels was determined by real-time PCR using primers and SSoAdvanced™ Universal SYBR® Green Supermix (Bio-Rad Laboratories, Hercules, USA). Amplification assays were performed by triplicate employing ~30ng of cDNA (Viruses) in a CFX384 Touch™ Real-Time Detection System (Bio-Rad Laboratories, Hercules, USA). RP49 (RNA) were chosen as a reference gene. Standard curves were prepared from plasmids harboring the target amplicons with copy numbers diluted from 107 to 102. PCR conditions were 3 min at 95°C for initial

denaturation/enzyme activation followed by 40 cycles of 10 seconds at 95°C and 30 seconds at 60°C (except IAPV, where annealing/extension was 45 seconds at 60°C). Specificity was checked by performing a melt-curve analysis 65-95°C increments of 0.5°C 2 seconds/step. Results were analyzed with the CFX Manager™ Software and exported to an Excel spreadsheet to calculate copy numbers per bee.

Grooming behaviour

To collect the data on mite damage, sticky boards made from plastic-coated freezer paper (Reynolds Kitchens®, Richmond, VA, USA) covered with petroleum jelly and labelled with hive identification numbers and insertion dates were placed on hardboard panels beneath the screened bottom boards of the colonies. This was done for two consecutive 72-hour periods in the fall, before miticide treatment. Afterward, the sticky boards were collected, combined, and used to assess grooming behavior. All mites were carefully removed from the sticky boards and preserved in 70% ethanol for analysis. For colonies with more than 200 mites on the board, a random subsample of 200 mites was selected, while colonies with fewer than five fallen mites were excluded from the analysis. Mites were examined under 32-50x magnification (Leica MS5, Leica Microsystems, Wetzlar, Germany), with both dorsal and ventral sides inspected for damage to their mouthparts, legs, and ventral shields (dented idiosomas were not considered damage). Injuries were categorized as mild (partial damage or parts of one leg missing), moderate (a whole leg or mouthparts missing), or severe (multiple legs or entire mouthparts missing).

Mite population growth rate

Varroa mite resistance was measured using the Instantaneous Mite Population Growth Rate (IMPGR). IMPGR was calculated using the formula $P2=P1\times e^{rn}$, where r represents the weekly growth rate, n is the duration in weeks (13–17 weeks), and e is the base of the natural logarithm (Harris et al., 2003). The initial mite population (P1) was determined at the start of the experiment from 500 bees using the alcohol wash method. The alcohol wash method was adapted from (De Jong, Roma, & Goncalves, 1982) where bee samples in ethanol were shaken at 130 rpm for five minutes using a horizontal shaker. The shaking process was repeated until no more mites were retrieved from the ethanol. The final mite population (P2) was assessed over six weeks after miticide treatment with amitraz strips (Apivar®; Veto-pharma, Palaiseau, France). Mite drop was monitored by placing sticky boards (Varroa Mite Sticky Boards, Bee Maid,

Spruce Grove, AB, Canada) under the screened hive bottom. The total number of mites collected on the sticky boards was recorded as P2 (Borba et al., 2022).

Supplementary Discussion:

Our study identified several loci associated with complex social phenotypes measured in honey bee colonies. To validate these associations, we constructed a multivariate generalized linear model with candidate SNPs and used cross-validation to generate a distribution of performance metrics (Henshall, 2013, Daetwyler et al., 2013). This approach evaluates the ability of the genetic variants identified in the GWAS to predict the trait of interest across diverse subsets of the dataset, offering insights into the reliability and consistency of these associations.

We observed variation in the prediction accuracy of the model in its ability to predict the measured phenotypes. On average, the candidate SNPs explained 55% of the phenotypic variance, with the model for mite drop per day accounting for 83% of the variance (Table 3.3). The correlation between the predicted phenotype values made by the model and the observed phenotype was highly variable ranging from 0.78 to 0.15, averaging 0.57 (Table 3.3). Overall, the models for predicting viral and pathogen loads performed below average, showing high prediction errors and low average R^2 (45%) and *r* (0.44) values (Table 3.3).

Though we do not have a completely independent population to validate how well the candidate SNPs can generalize across unrelated samples, these models suggest promising results about their association with the measured traits. This is especially evident by the high correlation between predicted and observed values, which are comparable to other agricultural plant and animal species (Budhlakoti et al., 2022; Haile et al., 2021; Lee et al., 2008; Sehrawat et al., 2023), where a correlation of 0.6 is considered a moderate level of success (McGaugh et al., 2021). However, predictions made within populations have been shown to perform better relative to between populations (Lee et al., 2008, Daetwyler et al., 2010, Haile et al., 2021), due to factors such as variation in phenotype, genetic relatedness, and subpopulation structure (Daetwyler et al., 2013, Henshall, 2013, McGaugh et al., 2021). Thus, it is unclear if these models will generalize outside of commercial honey bee colonies or even outside of colonies in Canada. Despite this limitation, this is a first step towards estimating complex honey bee traits, which is valuable for developing more productive and resilient managed bee populations through selective breeding initiatives.

Chapter four: Genetic architecture and regulatory networks underlying honey bee (*Apis mellifera*) innate immunity

Tanushree Tiwari, Alexandra Sébastien, Kathleen A Dogantzis, Rodney T Richardson, Clement F Kent, Stephen Rose, Harshilkumar Patel, Alivia Dey, Ida M Conflitti, Judith Trapp, Bradford Vinson, Nikolay Stoynov, Kyung-Mee Moon, Queenie W.T. Chan, Nonno Hasegawa, Renata B. Labuschagne, Shelley E Hoover, Rob W Currie, Pierre Giovenazzo, M Marta Guarna, Stephen F Pernal, Leonard J Foster, Amro Zayed

Summary:

Understanding the genetic basis of innate immune regulation is critical for improving disease resilience in honey bees, which are increasingly challenged by pathogen pressure and global population declines. In this study, we combined colony-level whole-genome sequencing of 50 pooled workers with quantitative proteomic measurements of immune proteins in more than 1300 colonies to conduct genome-wide association analyses and identify *cis* and *trans*-acting candidate SNPs associated with variation in immune protein abundance. We then examined relationships between immune protein variation and pathogen traits at both the phenotypic and genotypic levels, revealing shared genetic influences across multiple immune challenges. Despite the relatively limited number of innate immune proteins in honey bees, our results show that these proteins collectively mediate all major immune functions and respond to diverse infections and stressors. Together, these findings show honey bee immunity as a genetically structured, multistep process shaped by interactions among immune proteins, pathogens, and genetic variation at individual loci, providing a foundation for understanding immune resilience at the protein level.

Introduction:

Honey bees (*Apis mellifera*) are highly eusocial insects characterized by extreme social organization (Michener, 1974) and cooperative division of labor within colonies (Seeley & Visscher, 1985). Colony members engage in frequent physical interactions, including cooperative brood care, allogrooming, and nutrient exchange through trophallaxis (Farina, 1996), which are essential for colony cohesion and function. However, this high level of social contact also creates

favorable conditions for the transmission of infectious agents, including viruses, bacteria, fungi, protists, and parasitic *Varroa* mites (Evans & Schwarz, 2011; Schmid-Hempel, 1998).

To counteract this elevated disease risk, honey bees rely primarily on an innate immune system that provides broad-spectrum protection against diverse pathogens. This defense system includes physical barriers, chemical defenses, cellular immune responses, and humoral mechanisms (Wilson-Rich et al., 2008). Physical barriers such as the cuticle and the peritrophic membrane of the gut serve as the first line of defense, while chemical defenses including enzymes such as glucose oxidase, contribute antimicrobial activity that suppresses microbial growth (Evans & Spivak, 2010). When pathogens cross these barriers, immune activation is triggered through the recognition of pathogen associated molecular patterns (PAMPs) (Brutscher et al., 2015) by germline-encoded pattern recognition receptors (PRRs), initiating downstream immune signaling cascades (Brutscher et al., 2015). These cascades coordinate cellular immune responses mediated by hemocytes, including phagocytosis, nodulation, and encapsulation (Schmid-Hempel, 2005). Humoral responses further complement these defenses through the production of antimicrobial peptides (AMPs) and activation of the prophenoloxidase cascade (Söderhäll & Cerenius, 1998), leading to melanization of pathogens and wound sites (Chan et al., 2009).

In addition to these classical immune mechanisms, honey bees use RNA interference (RNAi) as an antiviral defense strategy (Brutscher et al., 2017). During viral replication, double-stranded RNA (dsRNA) is processed by Dicer into small interfering RNAs (siRNAs), which guide the RNA-induced silencing complex (RISC) to degrade viral RNA and limit infection (Brutscher & Flenniken, 2015; Niu et al., 2014). Immune responses are further regulated by conserved signaling pathways, including Toll, Imd, JAK/STAT, and JNK (Evans et al., 2006). The Toll pathway primarily responds to Gram-positive bacteria and fungi, while the Imd pathway is activated mainly by Gram-negative bacteria. The JAK/STAT and JNK pathways contribute to antiviral defense, immune regulation, and stress responses (Galbraith et al., 2015). Several immune-related genes within these pathways have been shown to influence colony health, including *Dicer* and *Argonaute-2* in antiviral defense, *Dorsal*, *Relish (Schlüns & Crozier, 2007)*, and *Defensin-1* in antibacterial and antifungal responses, and *Domeless* in broader immune regulation and seasonal adaptation. Together, these components form an interconnected

immune network that underpins honey bee resilience to pathogens. Despite this well-characterized immune architecture, the genetic mechanisms regulating variation in immune gene and protein expression remain poorly understood.

Protein abundance is not a direct representation of mRNA levels but instead reflects regulation across multiple interconnected molecular layers (Vogel & Marcotte, 2012). Broadly, protein levels are regulated by various processes at various levels. These processes include transcriptional control of mRNA synthesis, post-transcriptional regulation of mRNA and translation, and post-translational mechanisms responsible for protein stability, modification, and degradation. At the transcriptional level, transcription factors and chromatin state determine when and how much mRNA is produced. Post transcriptional processes, including mRNA stability, alternative processing, and translational efficiency, control how much protein is synthesized from available transcripts (Liu et al., 2016). Finally, post-translational regulation, such as protein modification and abundance, directly controls protein activity within the cell. In insect immune systems, rapid post-translational regulation is particularly important, enabling swift immune responses while limiting physiological trade-offs associated with prolonged activation (Lemaitre & Hoffmann, 2007).

Genetic variation influences these regulatory layers through both local (*cis*) and distal (*trans*) mechanisms. In honey bees, these variants strongly shape immune traits and colony-level phenotypes (Evans et al., 2006; Harpur & et al., 2019). While *cis*-acting variants typically affect expression locally, *trans*-acting variants often originate from distal regulatory networks and can form regulatory hotspots with coordinated effects on multiple genes or proteins (Grundberg et al., 2012; Rockman & Kruglyak, 2006) . By modulating everything from signaling pathways to protein turnover, these genetic drivers provide the essential mechanistic framework for interpreting protein quantitative trait loci (pQTLs) and their associated functional enrichments in the honey bees immune architecture.

In this study, we aim to identify genetic loci associated with variation in immune protein expression in honey bees. We employed protein quantitative trait loci (pQTL) mapping, analogous to expression QTL (eQTL) analysis (Nica & Dermitzakis, 2013), to link genetic variants to differences in protein abundance. Identified markers were classified as *cis* or *trans*-acting based on their genomic proximity to the encoded proteins. We further examined

associations between immune protein expression and pathogen related phenotypes, including four RNA viruses *(*BQCV, DWVA, DWVB, SBV*)* as well as *Nosema spp.*, *Lotmaria*, and *Varroa* mite growth rate, to assess the relevance of these regulatory loci to colony defense. Finally, we identified genetic loci that jointly influence immune protein expression and pathogen burden. This large-scale analysis includes more than 1,300 honey bee colonies sampled across five Canadian provinces and captures substantial genetic and phenotypic diversity. Together, these data provide a unique framework for investigating the genetic regulation of immune function in honey bees. The loci identified in this study represent strong candidates for future functional validation and may also inform breeding programs aimed at improving disease resistance and overall colony health.

Methods:

Sample collection

The honey bee samples were collected from five provinces (Alberta, British Columbia, Manitoba, Ontario and Quebec**)** in Canada in 2016 and 2017. A total of 926 honey bee colonies were established in May 2016 and 424 colonies in May 2017, for a total of 1,350 colonies across both years, see (Borba et al., 2022) for more information. Adult worker honey bees were collected from frames containing brood and immediately placed on dry ice in the field. Samples were stored and shipped on dry ice to Dr. Foster's lab at the University of British Columbia for proteomics analysis and to Dr. Zayed's lab at York University for whole genome sequencing.

Antennal tissue processing and protein extraction

A total of 1,322 (Supplementary Table 4.1) colonies were sampled, with each colony represented by three biological replicates (A, B, and C). Antennae from 20 individual bees were dissected per replicate (20 bees-1 pair of antennae per bee) and were stored in a -20°C freezer. Before processing, the samples were washed three times using 500 µL of ice-cold 1X Phosphate-buffered saline (PBS). Each wash was followed by vortexing and brief centrifugation, after which the cleaning solution was carefully removed. After the third wash, 250 µL of lysis buffer was added to the sample. Homogenization tubes were then supplemented with four 2.8 mm ceramic beads. The homogenization cycle (shaking 20 sec at 6.5 m/s followed by a brief spin) was repeated three times. The samples were then centrifuged 10 min at 8 000 rpm at 4°C. The

supernatant containing the proteins was collected into 1.5 mL tubes, carefully avoiding the antenna pieces. Antennal tissues were homogenized at 4°C and samples kept on ice in between to preserve protein integrity.

After this, ethanol/glycogen/sodium acetate precipitation was done. Approximately 200 µL of each sample was transferred to a clean 1.5 mL microcentrifuge tube. A total of 1,030 µL of ethanol precipitation mix was added, followed by immediate vortexing. Samples were left at room temperature for 2 hours and 30 min before centrifugation at 16,000 × g for 15 minutes. After discarding the supernatant, samples were spun again for 3 minutes to remove any residual ethanol. The resulting pellets were dried in a SpeedVac for 3 minutes at room temperature and stored at −20°C.

Frozen protein pellets were resuspended in 50 µL of solubilization buffer, sonicated in an ice-cold water bath for 5 minutes, vortexed and briefly centrifuged. The samples were incubated at room temperature for approximately 30 minutes and centrifuged at 16,000 × g for 10 minutes. The supernatant, representing the solubilized protein lysate, was collected and stored in 1.5 mL tubes.

For protein quantification, each lysate was diluted 1:5 by mixing 7 µL of the sample with 28 µL of ultrapure water in a 0.5 mL tube. After vortexing and a quick spin, 10 µL of the diluted sample was transferred into a flat-bottom 96-well microplate in triplicate, alongside a standard curve (0–2 µg/µL) prepared in duplicate. Bradford reagent (300 µL) was added to each well, and the plate was mixed gently using a thermomixer (600 rpm, 1 minute). After a 10-minute dark incubation at room temperature, absorbance was measured at 595 nm. Protein concentrations were normalized to 1 µg/µL using solubilization buffer, and samples were stored at −72°C.

Protein digestion was performed in 96-well plates, with each well containing 15 µL of lysate (15 µg of protein). lysine was added, and samples were agitated for 5 min at 400 rpm and incubated for 3 hours at room temperature. DTT (15uL of 0.02 ug/uL) were added to each sample, the plate was agitated for 5 min at 400 rpm and incubated for 30 min in the dark. Iodoacetamide (IAA, 15 uL of a 0.5 µg/µL solution) was then added, followed by incubation in the dark for 20 minutes followed by diluted trypsin. Digestion proceeded overnight at room temperature. The next day, samples were transferred to 0.5 mL tubes and stored at −20°C or immediately processed further.

Sample preparation and running Mass Spectrometry

Dimethyl labelling was performed in three parallel 96-well plates representing light, medium, and heavy isotopic conditions. Peptides were resuspended in 10 µL of 100 mM TEAB (Triethylammonium bicarbonate), vortexed, sonicated briefly, and transferred to new plates. Formaldehyde (CH_2O for light, CD_2O for medium, and CD_2O for heavy; 20 µL) was added, followed by 3 µL of sodium cyanoborohydride (ALD) for light/medium or sodium cyanoborodeuteride for heavy labeling. After gentle mixing (600 rpm, 1 minute), samples were incubated in the dark at room temperature for 2 hours. To stop the reaction, 10 µL of 3 M ammonium chloride was added, followed by a 10-minute incubation. Acidification was achieved by adding either 60 µL of 1% TFA. Following labelling, peptides were cleaned on C-18 STop And Go Extraction (STAGE) tips (Ishihama et al., 2002). Samples were acidified (pH <2.5) and pooled. The stage-tips were conditioned with methanol and Buffer A, samples were loaded, washed, and eluted twice with 100 µL of Buffer B. Eluates were dried and stored at −20°C until mass spectrometry (MS) analysis.

For MS analysis, peptides were diluted in 17.4 µL of 0.1% formic acid, achieving a final concentration of ~0.4 µg/µL. Samples were vortexed, sonicated and spun briefly. A 6 µL aliquot was loaded into a 96-well plate, sealed, and spun to remove bubbles. Injections were performed using Impact II Qtof (Bruker Daltonics) coupled to nano LC (Thermo Scientific) using the acquisition parameters. Remaining sample volumes were stored at −80°C for future use. The runs generated 3673 files for the 2016 and 1269 files for the 2017 samples.

Data analysis

Raw data was processed by MsFragger 17.0 with modifications to the default setting: fixed carbamidomethylation (+57.021464 Da) to cysteine and variable modifications of N-terminal acetylation (+42.0106Da), methionine oxidation (+15.9949), isotopic demethylation of N-termini and lysine side chain (+28.0313 Da for light, +32.056408 Da for medium, +36.07567 Da for heavy). The data was searched against a protein FASTA library containing Apis_OGS3.2 (Elsik et al., 2014) and 507 sequences of bee viruses and their known variants, then compiled by MsFragger to include common contaminants and reverse decoy sequences. To normalize the proteomic dataset, raw peptide intensities from light, medium, and heavy (L, M, H) channels for each data file were log10 transformed. For each sample, the median of all log-transformed

peptide intensities was calculated, generating a list of individual medians. From this list, the overall median of medians and the standard deviation were computed. The two standard deviations from the median were used as an outlier threshold. Samples with file medians deviating by more than two standard deviations from the overall median were considered outliers and excluded. As there were differences in 2016 and 2017 sample numbers, the 2017 dataset was aligned to the 2016 baseline. A correction factor was computed for each remaining file by dividing its median by the 2016 median of medians. Each peptide intensity was then normalized by dividing it by this correction factor, resulting in median-centred peptide intensities relative to the 2016 dataset. Finally, protein-level intensities were obtained by averaging the normalized intensities of all constituent peptides per protein within each sample.

Identification of Immune Proteins (IPs)

A literature review was conducted on published literature queried with keywords related to innate immunity, immune-related proteins, and pathogen responses in honey bees. By surveying the literature, we generated a curated list of proteins associated with immune function (Supplementary Table 4.2). Immune proteins identified through this were subsequently searched in proteomic datasets from 2016 and 2017, generated by mass spectrometry. This led to the identification of immune-related proteins, which were included in this study and used for downstream analyses.

Genotype data

DNA extraction and sequencing

Adult worker honey bees from 1,350 colonies were used for whole genome sequencing. The front left legs of 50 bees from each colony were pooled and ground in a mortar and pestle for DNA extraction. To extract DNA, honey bee legs were first flash-frozen by immersing the tubes in liquid nitrogen. The frozen tissues were then finely ground with a pestle to maximize the yield of pooled tissue. For tissue lysis, 350 µL of Tissue Lysis Buffer and 20 µL of Proteinase K (20 mg/ml) were added to the ground material, followed by overnight incubation at 50 °C. DNA was subsequently extracted using the Mag-Bind Blood & Tissue DNA HDQ 96 Kit (Omega Bio-tek Inc., USA), which is optimized for the KingFisher Flex Purification System (Thermo Fisher Scientific Inc., USA). The extraction yielded approximately 75 µL of DNA eluent per pool of 50

bee legs. DNA concentration was measured with a NanoDrop™ 2000 Spectrophotometer (Thermo Fisher Scientific Inc., USA), and quality was verified using 0.8% agarose gel electrophoresis. The extracted pooled DNA was then sent to the Genome Québec Innovation Centre's Next Generation Sequencing facility at McGill University. Sequencing was carried out on the Illumina HiSeqX platform, producing 150 bp paired-end reads at an average genome-wide depth of ~110× coverage. Pooling DNA from 50 individuals provides a sufficiently large sample size to capture the full range of genetic variants contributed by multiple drone fathers(Tarpy, 2000; Tarpy & Nielsen, 2002), thereby enabling estimation of colony allele frequencies.

Sequence alignment and SNP discovery

Raw sequencing data were trimmed with Trimmomatic (Bolger et al., 2014b) to remove low-quality bases with quality scores below 20 and Illumina adapters, using a 25 bp sliding window and retaining reads at least 50 bp in length. The trimmed reads were aligned to the Apis mellifera reference genome (HAv3.1) (Wallberg et al., 2019) using NextGenMap v0.4.12 (Sedlazeck et al., 2013a). The genome-aligned BAM files were indexed and sorted using SAMtools v1.3.1 (Li et al.) and duplicate reads were marked using Picard v2.1.0 (https://broadinstitute.github.io/picard/). Base quality scores across BAM files were recalibrated with GATK v4.0.11.0 BaseRecalibrator (Van der Auwera et al., 2013a), incorporating previously identified variants as a training dataset. SNP and indel detection in the pooled genomes was performed with Lofreq v2.1.2 (Wilm et al., 2012), which uses mapped reads to estimate allele frequencies for reference and alternative alleles.

Variant filtration

Individual sample VCF files were filtered to remove variants with strand bias (SB) > 50, depth > 500 or < 10, and SNP quality < 90. SNPs within 5 bp of indels, from unplaced scaffolds, multiallelic sites, and loci with ≥90% missing genotypes were also discarded. Only SNPs located within genes or 5 kb upstream or downstream of a gene, based on the HAv3.1 genome gene file (gff3) (Wallberg et al., 2019), were retained, resulting in 1,742,882 loci. These were used for downstream analysis.

Identification of pQTL loci

To map genetic loci underlying immune protein abundance, we performed protein quantitative trait locus (pQTL) analysis (Melzer et al., 2008), treating protein abundance as a

quantitative trait and genome-wide SNPs as markers. Two approaches were used first, a linear mixed model (LMM) with GridLMM (Runcie & Crawford, 2019) and second a Bayesian Sparse Linear Mixed Model (BSLMM) with GEMMA (Zhou & Stephens, 2014). The LMM accounts for population structure, while BSLMM integrates sparse regression with polygenic effects, enabling detection of both small and large effect loci. For each protein, two independent BSLMM runs were conducted with 20 million iterations (10 million burn-in), sampling every 10,000 steps. In addition, BSLMM outputs the proportion of variance in protein abundance attributable to genetic effects (PVE) and the fraction of this variance due to large-effect loci (PGE). Their product (PVE × PGE) was used as an estimate of narrow-sense heritability (h^2) (Bresadola et al., 2019).

Filtering and identification of candidate SNP markers

To identify candidate SNPs strongly associated with protein expression, we integrated results from the linear mixed model and the Bayesian sparse regression. We selected SNPs with large effects using the following threshold: gamma ($\beta\gamma$) is greater than 1, posterior inclusion probability (PIP)(gamma; γ) $\geq$0.01(Comeault et al., 2014), and a significant p-value below the 1% quantile of the genome distribution. Additionally, we included SNPs with a posterior inclusion probability (PIP > 0.10), regardless of their p-value, as these SNPs were used in the model more than once.

Classification and functional annotation of SNPs

The candidate SNPs identified through pQTL analysis were classified as *cis*- or *trans*-regulatory based on their genomic location relative to the associated immune protein. *Cis*-acting variants were defined as SNPs located within the gene itself or 1 Mb upstream or downstream of gene boundaries (Grundberg et al., 2012), whereas *trans*-acting variants encompassed SNPs located elsewhere in the genome. Following this classification, SNPs were annotated using SnpEff v5.2a (Cingolani, 2012) to determine their location within protein-coding and non-coding regions, including exons, introns, and flanking sequences. To further understand biological functions, Gene Ontology (GO) enrichment analyses were conducted using DAVID v2025_1(Huang et al., 2009), with p-values adjusted for multiple testing using the Benjamini-Hochberg correction (Benjamini & Hochberg, 1995).

Comparative analysis between immune proteins and pathogen levels

The adult bees from the colonies were quantified for RNA viruses *(*Black Queen Cell Virus (BQCV), Deformed Wing Virus A and B and SacBrood virus*)*, and *Lotmaria passim* using RT-qPCR. All pathogens were quantified against endogenous housekeeping genes using quantitative real-time PCR assays based on published primers and protocols (Jay D Evans et al., 2013). *Nosema spp*. the spores were quantified by counting them under the electron microscope. The Instantaneous Mite Population Growth Rate (IMPGR) was calculated using the formula $P2=P1\times e^{rn}$, where r represents the weekly growth rate, n is the duration in weeks (13-17 weeks), and e is the base of the natural logarithm (Harris et al., 2003). The data were recorded in an Excel sheet and normalized using the R package BestNormalize (Peterson, 2021b).

To characterize innate immune responses, we performed Pearson correlation analyses between immune protein (IP) abundance and colony-level pathogen loads. All pathogen phenotypes were quantified from the same colonies in which protein abundances were measured, enabling direct colony-level association analyses. For each protein pathogen pair, Pearson correlation coefficients ® were calculated, and statistical significance was assessed using the Benjamini-Hochberg false discovery rate (FDR) correction to account for multiple testing across all comparisons. Associations were considered significant at FDR-adjusted $p < 0.05$.

Following phenotype-level correlations, we investigated shared genomic architecture between immune proteins and pathogens by examining overlap among associated genetic markers. For each significant trait, candidate SNPs were identified and a 200 bp window was applied around each marker to capture nearby linked variants. Overlap was defined as shared markers or linked markers falling within this window between immune proteins and pathogen traits. The 200 bp window size was selected based on previous evidence that linkage disequilibrium (LD) in *Apis mellifera* decays rapidly, typically within ~200 bp (Dogantzis, 2022). This window therefore captures variants likely to be in LD and representing the same local genomic signal. All analyses were implemented using custom Python scripts.

Results:

Overview

The proteins analyzed in this study were selected through a comprehensive literature review and span multiple functional categories relevant to honey bee immunity and stress responses (Table 4.1). We included proteins involved in stress tolerance and protein folding, such as the molecular chaperones Hsc70-4, Hsp90, and Hsp83, along with their co-chaperone activator Aha1, which together support cellular stability under infection and environmental stress (Henriques et al., 2021). To capture immune recognition and signaling processes, we analyzed β-1,3-glucan recognition and binding proteins, peptidoglycan recognition proteins (PGRP-S2 and PGRP-S3), toll-interacting protein, CD109 antigen, phenoloxidase, and several antimicrobial peptides, all of which play established roles in pathogen detection and immune activation in honey bees (Doublet et al., 2017b; Evans et al., 2006). We also measured antioxidant and metabolic enzymes that contribute to immune defense by limiting oxidative damage and supporting cellular metabolism, including catalase, glutathione peroxidases (Gtpx1 and PHGPx), superoxide dismutase (Sod1), and phospholipase A2 (Evans et al., 2006; Henriques et al., 2021). In addition, proteins involved in intracellular transport and structural organization were included, such as Ran GTPase, actin-related proteins, and the NPC cholesterol transporter, reflecting the importance of cellular trafficking and cytoskeletal dynamics during immune responses. Finally, the protein set encompassed serine proteases and protease inhibitors involved in immune cascades, including chymotrypsin and antitrypsin, as well as key regulators of signal transduction such as JNK and STAT5B, which modulate downstream immune and stress-responsive pathways (Doublet et al., 2017b; Henriques et al., 2021). The details of the proteins are provided in Supplementary Table 4.2.

Using this biologically informed protein set, we conducted a pQTL analysis on 66 immune proteins quantified by mass spectrometry in 1,322 colonies sampled across five provinces. The dataset comprised 922 colonies collected in 2016, and 400 colonies collected in 2017, with three biological replicates per sample in each year. pQTL mapping identified 2,559 unique candidate loci distributed across 16 linkage groups. Although 66 proteins were initially targeted, two lacked significant genetic associations and were excluded from downstream

analyses, leaving 64 immune proteins for investigation of the genetic architecture underlying innate immune variation.

Across these 64 proteins, we detected substantial genetic contributions to phenotypic variation. Estimates of narrow-sense heritability (h^2) averaged 0.152 and ranged from 0.005 to 0.485 (Supplementary Figure 4.1), indicating heterogeneous but often appreciable genetic influence on immune protein abundance. Together, these results establish a population-scale framework for dissecting *cis*- and *trans*-regulatory mechanisms, regulatory hot spots, and pathogen-associated genetic effects described in the following sections.

Identification and functional classification of candidate loci

Cis-acting loci

Cis-acting variants were defined as SNPs located within the gene itself or within 1 Mb upstream or downstream of gene boundaries, distinguishing *cis* from *trans* effects based on physical proximity. From 2,559 candidate loci across 16 linkage groups, we identified 20 *cis*-acting loci (Figure 4.1) associated with 15 immune-related proteins. Most loci were separated by large physical distances, well beyond the short-range LD observed in honey bees (Dogantzis, 2022), resulting in 19 unique *cis* loci representing distinct regulatory regions.

Cis-regulatory loci influencing immune protein abundance were unevenly distributed across the genome, with the strongest clustering observed for a subset of immune proteins on linkage groups LG15 and LG7. On LG15, five unique *cis* loci formed a composite regulatory cluster spanning 3,302.8 kb (mean inter-locus distance: 825.7 kb), including multiple loci associated with GB50522 (FLO11-like) and GB50259 (FMR1), as well as a single locus linked to GB50538 (STAT5B). On LG7, the immune protein GB40977 (staphylococcal nuclease domain-containing protein) was associated with three *cis* loci distributed across 386.3 kb (mean inter-locus distance: 193.2 kb), suggesting localized regulatory complexity within this linkage group. Other linkage groups exhibited more spatially dispersed *cis*-regulatory loci. On LG5, *cis* loci associated with GB42082 and GB45735 were separated by 5,502.8 kb and exhibited no linkage disequilibrium, while on LG10, loci linked to GB46360 and GB48343 were separated by 4,149.9 kb. The remaining seven immune proteins (GB46479, GB53013, GB45417, GB47938, GB42685, GB43229, and GB48204) were each associated with a single unique *cis*-regulatory locus.

To characterize the genomic features underlying these *cis*-regulatory loci, SNPs associated with immune protein abundance were annotated using SnpEff (Cingolani, 2012). This annotation yielded 27 distinct functional assignments, reflecting instances in which individual SNPs mapped to multiple genes or transcripts. Most annotated SNPs were located within protein-coding genes, while the remainder mapped to long non-coding RNAs. Within protein-coding genes, most SNPs were positioned in intronic regions, with only a single synonymous variant identified; the remaining variants were located upstream or downstream of annotated genes (Figure 4.2). This enrichment of intronic and proximal regulatory variants is consistent with previous studies in honey bees and other insects, which have shown that *cis*-regulatory variation affecting immune and social traits frequently resides outside of protein-coding exons (Evans et al., 2006).

Trans-acting loci

Trans-acting loci were detected for all 64 immune proteins, underscoring their widespread role in shaping immune protein abundance. In total, 2,539 *trans*-acting SNPs (Figure 4.1) were identified and annotated using SnpEff (Cingolani, 2012), yielding 3,713 functional annotations. These variants mapped predominantly to protein-coding genes (3,071), with additional contributions from long non-coding RNAs (594), microRNAs (29), transfer RNA and one snoRNA. Consistent with *cis*-acting loci, *trans*-acting variants were largely non-coding, with most annotations located in introns (2,363) or upstream/downstream regulatory regions (1,081) (Figure 4.2). Only a small subset was predicted to directly alter protein sequence or splicing (42), indicating that *trans*-regulatory effects are primarily mediated through regulatory rather than structural changes.

Gene Ontology enrichment and functional annotation clustering of *trans*-associated genes revealed a coherent functional profile dominated by membrane-associated signaling and cellular communication. The strongest enrichment was observed for plasma membrane and cell periphery components (GO:0005886; GO:0071944), accompanied by biological process terms related to cell surface receptor signaling (GO:0007166) and signal transduction (GO:0007165). Molecular function categories were enriched for molecular transducer activity (GO:0060089), encompassing receptor-mediated and enzyme-linked signaling pathways. In addition, moderate enrichment for transcriptional regulation-related terms, including DNA-binding transcription

factor activity (GO:0003700), regulation of transcription (GO:0006355), and regulation of gene expression (GO:0010468). Collectively, these results indicate that *trans*-acting variation influences immune protein expression primarily through membrane-initiated signaling cascades, with transcription factors functioning as downstream effectors, consistent with trans-regulatory architectures described in honey bees and other insects.

Regulatory hot spots for immune protein expression

Within this broader *trans*-regulatory landscape, we identified 171 candidate loci that function as regulatory hot spots, collectively influencing the abundance of multiple immune-related proteins. Most of these loci were *trans*-acting, with only three showing combined *cis*- and *trans*-regulatory effects, indicating that pleiotropic regulation is largely driven by distal control. Two prominent hot spots are located on chromosome 7 (CM009937.2:11,394,917) and chromosome 12 (CM009942.2:7,082,895), each affects 13 immune proteins (Figure 4.3). The chromosome 7 locus preferentially regulated proteins involved in immune signaling, chromatin-associated processes, and melanization, whereas the chromosome 12 locus primarily targeted proteins linked to neural development and cell-cell interaction. Several proteins, including Hsp90, its ATPase activator, RNA helicase Bel, Vps60, CD109, and Tudor-SN, were shared between these loci, suggesting convergence on core cellular stress-response and signaling pathways

Functional annotation of genes associated with regulatory hot spots revealed the same signaling-centered architecture observed for *trans*-acting loci overall, but in a more concentrated form. Hot spot associated genes were strongly enriched for plasma membrane localization (GO:0005886), cell surface receptor signaling (GO:0007166), and signal transduction (GO:0007165), reinforcing the idea that regulatory hot spots act as central nodes controlling immune protein expression.

Understanding the interactions between immune proteins (IPs) and pathogens

After characterizing the functional and genomic context of candidate loci that influence immune protein expression, we next examined correlations between immune protein expression and pathogen phenotypes (Figure 4.4). Negative correlations were observed in cases where higher protein abundance was significantly associated with reduced pathogen levels, particularly for kinases (GB55483), heat shock proteins (GB45495, GB40976), and adaptor molecules

(GB43303, GB52655). In contrast, positive correlations were observed for proteins such as phospholipase (GB48228), phospholipid-binding proteins (GB48634), and neuroglian (GB44650), in which higher abundance was associated with increased pathogen levels. One kinase (GB65012) showed no detectable correlation, while several proteins displayed mixed patterns across pathogens.

The strongest associations (Figure 4.4), identified using Pearson correlation coefficients (R) and statistical significance determined after Benjamini-Hochberg false discovery rate (FDR) correction (adjusted $p < 0.05$), were observed with Black Queen Cell Virus (BQCV), involving 38 immune proteins that exhibited both positive and negative relationships. Instantaneous Mite Growth Rate (IMPGR) showed significant correlations with 29 proteins, and Lotmaria with 26 proteins. These patterns underscore the multilayered and context-dependent nature of innate immune responses; wherein individual proteins may exert divergent effects depending on the specific pathogen pressure.

Candidate loci linking immune proteins and pathogens

To investigate whether the observed phenotypic correlations between immune protein expression and pathogen abundance had shared genetic control, we performed a SNP-based window analysis using a 200 bp physical distance threshold. In this approach, we identify genomic regions that may simultaneously influence both trait categories, either through pleiotropic effects or closely linked regulatory variants. The total of 56 candidate SNPs influencing both immune protein expression and pathogen abundance were identified in two trait categories. This shared genetic architecture was highlighted by three immune proteins, Catalase (CAT/GB41427), Ran (RAN/GB43229), and Corkscrew (CSW/GB47575), which play vital roles in antioxidant defense, antiviral phagocytosis, and signal transduction cascades, respectively. Furthermore, one protein specific to *Nosema spp.* abundance was identified, VPS37B (GB53228), associated with intracellular trafficking and autophagy. Collectively, these findings suggest that honey bee resilience is modulated not only by classical immune signaling but also by supportive genes that maintain cellular health and structural integrity during infection (Figure 4.5).

Discussion:

This study provides a comprehensive view of the genetic architecture underlying immune regulation in the honey bee by integrating genomic variation, immune protein expression, and pathogen associated phenotypes. By identifying both *cis* and *trans*-acting loci, our results demonstrate that a combination of local genetic effects and extensive long-range regulatory interactions shapes immune regulation (Albert & Kruglyak, 2015) in honey bees. The predominance of *trans*-acting loci highlights the highly interconnected nature of immune regulatory networks, consistent with the coordinated responses required to manage multiple pathogens within densely populated colonies.

Two genomic regions emerged as prominent regulatory hubs, located on chromosome 7 (CM009937.2:11,394,917) and chromosome 12 (CM009942.2:7,082,895), each influencing the expression of a broad spectrum of immune-related proteins. These proteins include signaling molecules, stress-response factors, and regulators of RNA metabolism, suggesting that these loci coordinate multiple cellular processes relevant to immune defense. Genes within these regions include transcriptional regulators, chromatin-associated proteins, and signaling adaptors, indicating that variation at these loci may affect immune function indirectly by modulating downstream regulatory pathways. The chromosome 7 locus aligns with previous quantitative trait loci associated with hygienic behavior and reduced *Varroa destructor* reproduction, reinforcing its importance in honey bee social immunity and disease resistance (Le Conte et al., 2010; Mondet et al., 2020; Spötter et al., 2016). In contrast, the chromosome 12 locus has not been highlighted in earlier studies, identifying it as a novel candidate region for immune regulation.

Functional enrichment analysis of genes associated with these regulatory hotspots revealed a predominance of processes related to plasma membrane localization (GO:0005886), cell surface receptor signaling (GO:0007166), and signal transduction (GO:0007165)**.** Plasma membrane localization refers to the positioning of proteins at the cell membrane, which is critical for sensing external signals and facilitating interactions between immune cells and pathogens. Cell surface receptor signaling encompasses the recognition of extracellular cues through membrane-bound receptors, initiating intracellular signaling cascades that regulate downstream immune responses. Signal transduction represents the process by which these signals are

propagated through intracellular networks, ultimately influencing transcriptional, post-transcriptional, and post-translational regulation of protein abundance. Together, these processes conceptually link upstream perception at the cell surface to the modulation of protein expression, highlighting how variation in signaling and receptor-associated genes can indirectly shape immune protein levels.

The predominance of membrane associated, and receptor linked functions suggests that genetic variation at these loci primarily affects how immune cells perceive, integrate, and transmit external or internal cues, rather than directly modifying immune effector proteins. Changes at this upstream level are expected to influence protein abundance indirectly, through modulation of transcription, RNA processing, translation, and protein turnover (Sun et al., 2018). Consistent with this view, the enrichment of transcriptional and RNA regulatory functions indicates that signaling alarms are propagated through multiple regulatory layers, enabling stepwise and context-dependent control of immune protein expression.

Patterns emerging from protein pathogen associations are broadly consistent with immune defense strategies described in honey bees and other insects. The tendency for pathogens to be negatively associated with proteins involved in signaling, stress response, and protein homeostasis suggests that these pathways contribute to constraining infection and mitigating pathogen-induced damage. Conversely, associations with lipid-related or regulatory proteins may reflect endogenous immune regulation, including the downregulation of negative regulators required for immune effector activation, rather than solely pathogen driven metabolic reprogramming or energetic trade-offs. Such regulatory dynamics are well documented in insect immune signaling pathways (Evans et al., 2006). The involvement of cell-adhesion components further points to a role for immune cell coordination and trafficking in shaping pathogen dynamics. The particularly strong associations observed with *Black Queen Cell Virus* are in line with its documented capacity to elicit pronounced immune and stress responses in honey bees (Doublet et al., 2015). Collectively, these patterns reinforce the view of honey bee immunity as a dynamic and integrated system, in which signaling, cellular stress management, and pathogen-mediated physiological modulation jointly influence disease outcomes.

Integrating pathogen associated SNPs with immune protein pQTLs highlights genomic regions where regulatory variation links protein expression directly to pathogen abundance.

Proteins mapping to these shared regions are central to immune defense, including oxidative stress mitigation, intracellular trafficking, signal transduction, and RNA regulation. Together, these functions point to coordinated mechanisms through which honey bees limit pathogen proliferation, maintain cellular homeostasis, and sustain immune cell activity under infection pressure. The involvement of pathways related to gut integrity, autophagy, and stress signaling is particularly consistent with known host responses to viral and microsporidian pathogens, including *Nosema* and *Black Queen Cell Virus,* and mirrors immune strategies reported across insect systems.

Taken together, our results support a model in which honey bee immunity is governed by an integrated regulatory network that couples signaling, stress adaptation, metabolic balance, and cellular maintenance. By leveraging population scale pQTL analyses, this study provides high genomic resolution loci and reveals a layered architecture composed of both localized regulatory effects and extensive *trans*-acting networks. While these associations provide a framework for understanding genetic contributions to immune variation, future functional validation will be essential to establish causality and to determine how these regulatory hubs shape colony-level resilience to pathogen challenge.

Conclusion:

This study is one of the first to identify candidate SNPs based on the interaction between immune proteins and pathogens phenotypes and genotypes. It utilizes advanced methodologies, including whole-genome sequencing, mass spectrometry, and RT-qPCR, to generate the data used for analysis. Although honey bees possess a limited number of innate immune and immune-related proteins, these proteins carry out all major immune functions, hence proven once again. The data suggest that individual proteins can be activated by various infections and stressors. This research lays the foundation for understanding the complexity of honey bee immune responses as a multistep process involving diverse proteins, pathogens, and mite interactions at individual SNP marker loci. We did identify the candidate loci on the same linkage groups as previous QTLs, but we cannot map them as different gene sets and assemblies have been used.

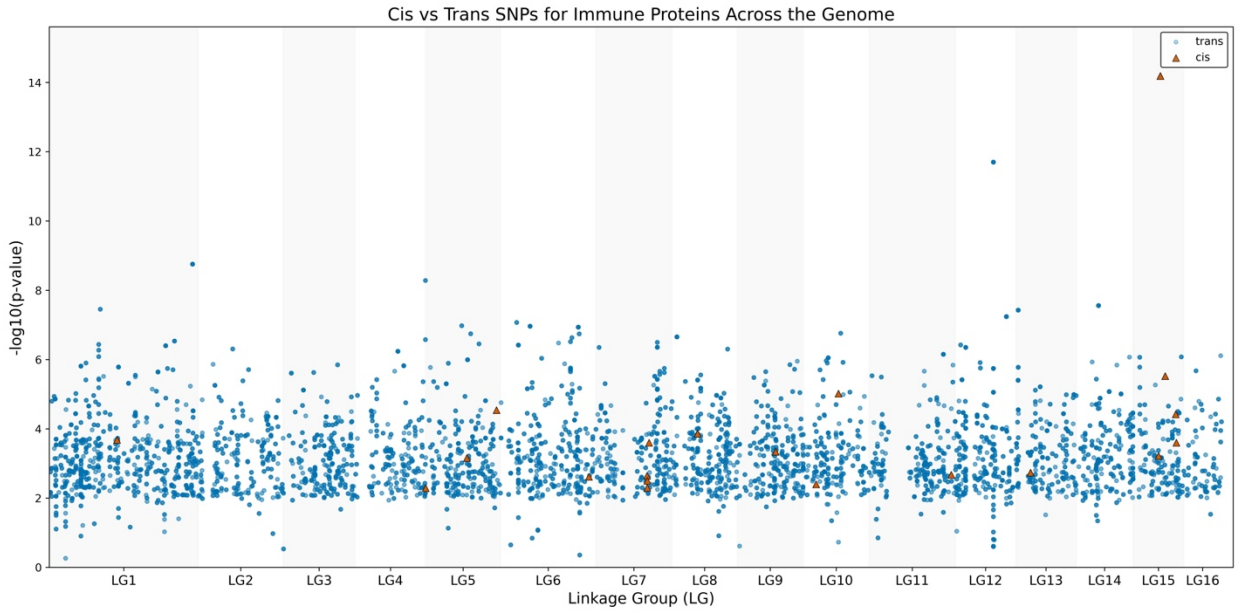
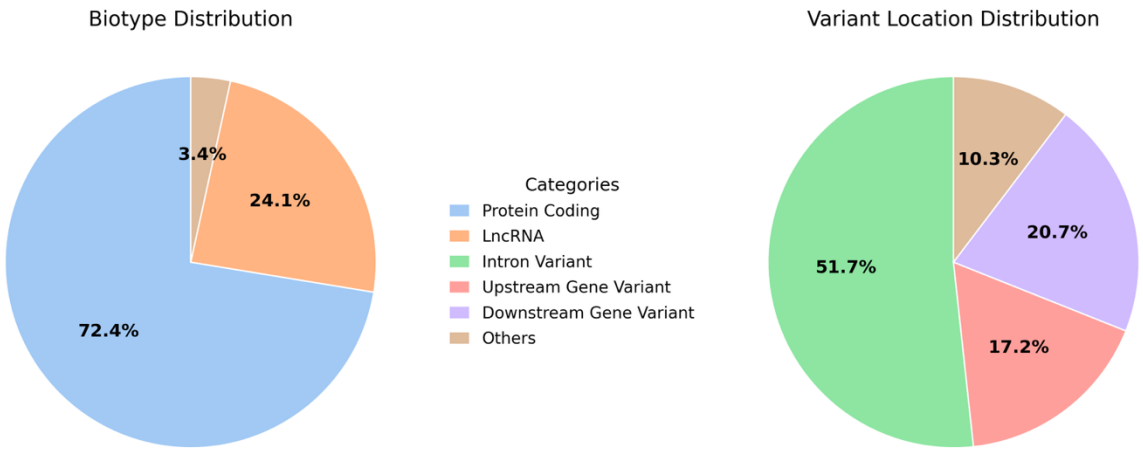


Figure 4.1. Global regulatory landscape of innate immunity in honey bees. Manhattan plot illustrating the genomic distribution of 2,532 *trans*-acting (blue circles) and 20 *cis*-acting (orange triangles) pQTLs across 16 linkage groups. The broad architectural signal suggests that the immune response is not limited to specific genomic regions but is a highly integrated system characterized by complex, distal genetic control.

a. *Cis* SNPs.



b. *Trans* SNPs

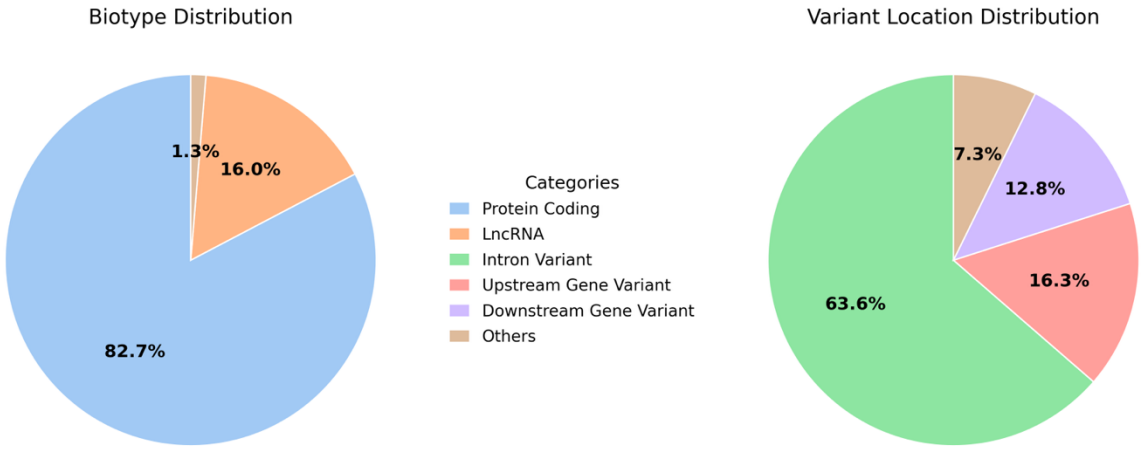


Figure 4.2. Pie charts showing the distribution of *cis* (a) and *trans* (b) acting loci based on gene type and predicted variant effects from snpEff. The gene type chart distinguishes loci mapping to protein-coding genes, lncRNAs, miRNAs, tRNAs, snoRNAs, and other categories. The variant location chart highlights the relative contributions of intronic, upstream, downstream, synonymous, UTR, missense, and splice region variants.

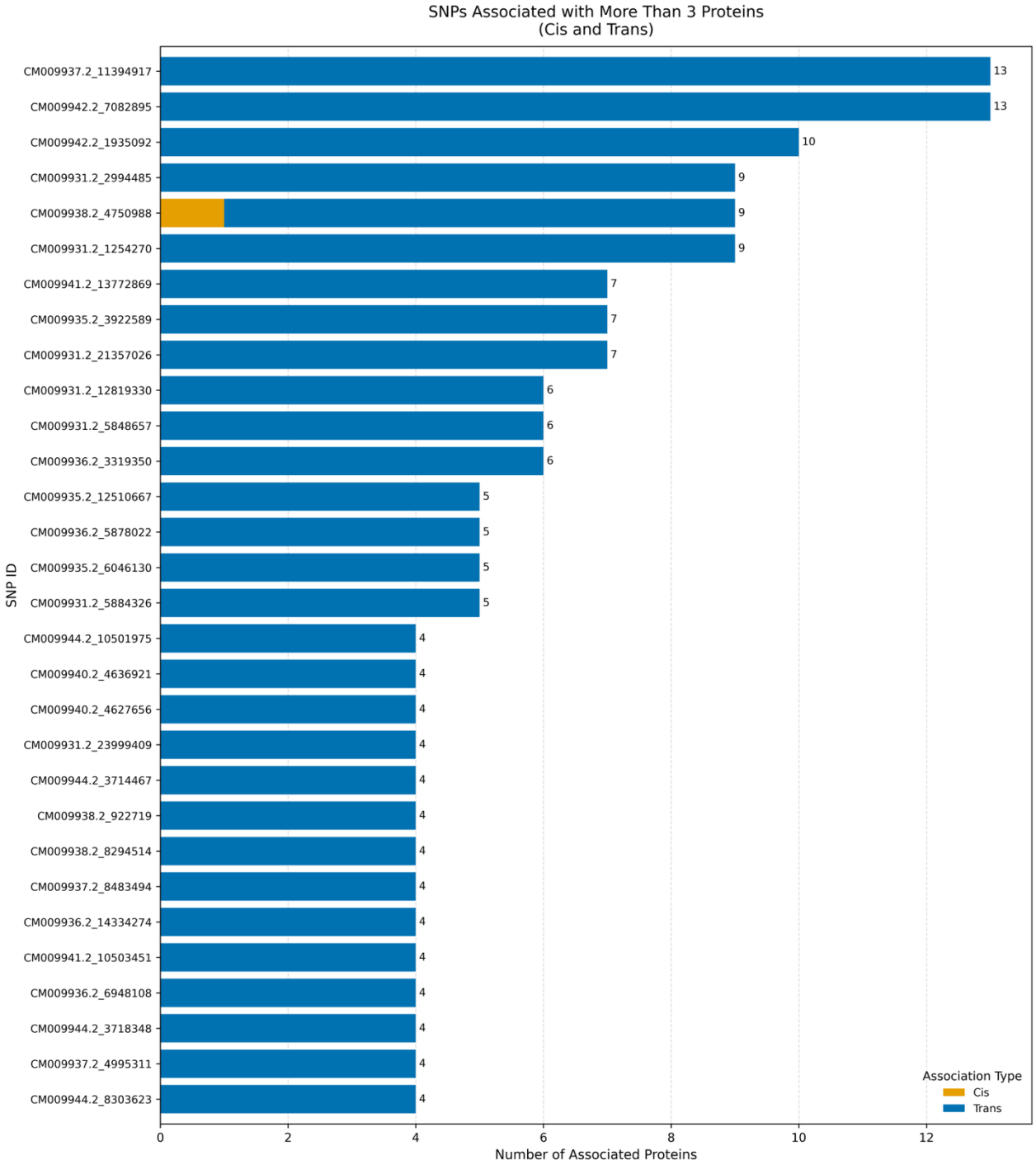
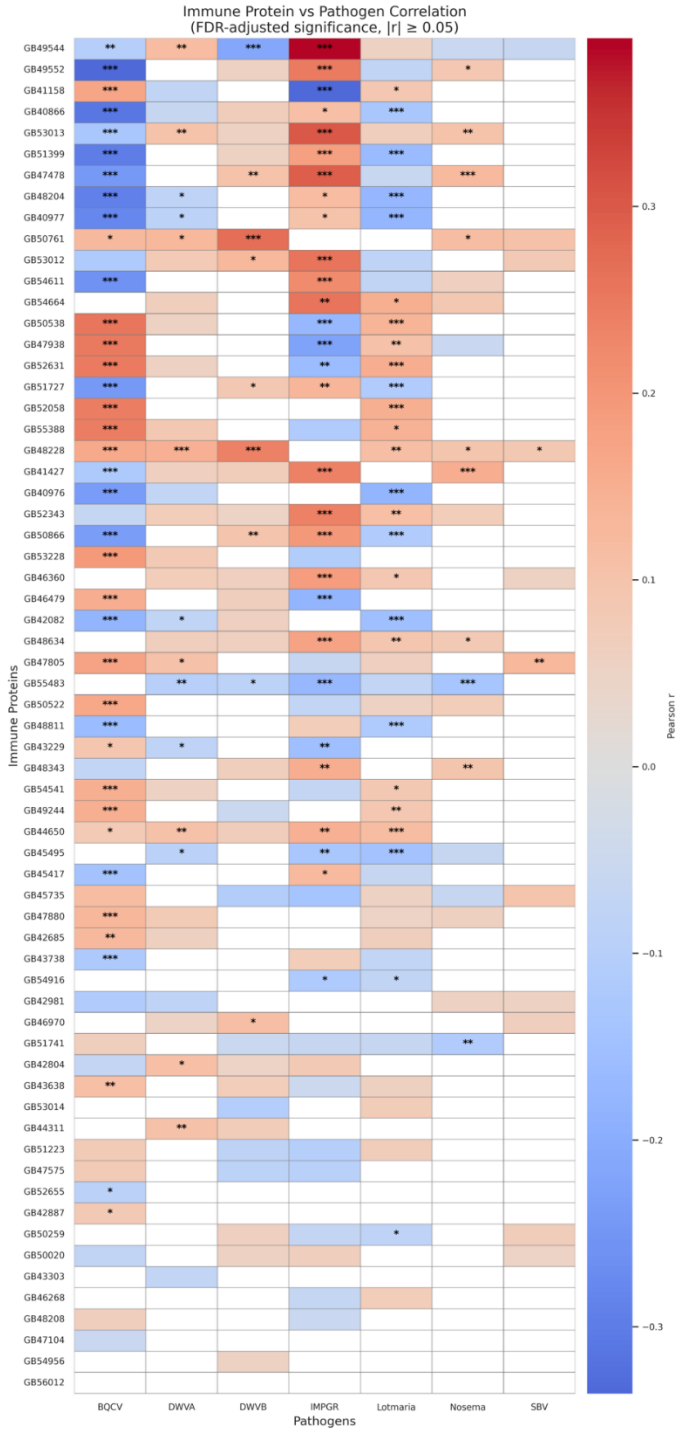


Figure 4.3: Identification of master regulatory hubs in the innate immune system. Horizontal stacked bar plot showing candidate SNPs associated with the abundance of more than three immune proteins. The striking predominance of trans-acting (blue) over cis-acting loci (orange) suggests that the former act as high-level regulatory hubs.



.

Figure 4.4. Heatmap of Pearson correlation coefficients (R) between pathogen phenotypes and immune protein (IP) abundances across colonies. Correlation values range from -1 to +1,

with orange indicating positive associations and blue indicating negative associations. Star denote significant associations (*FDR < 0.05, **FDR < 0.01, ***FDR < 0.001).

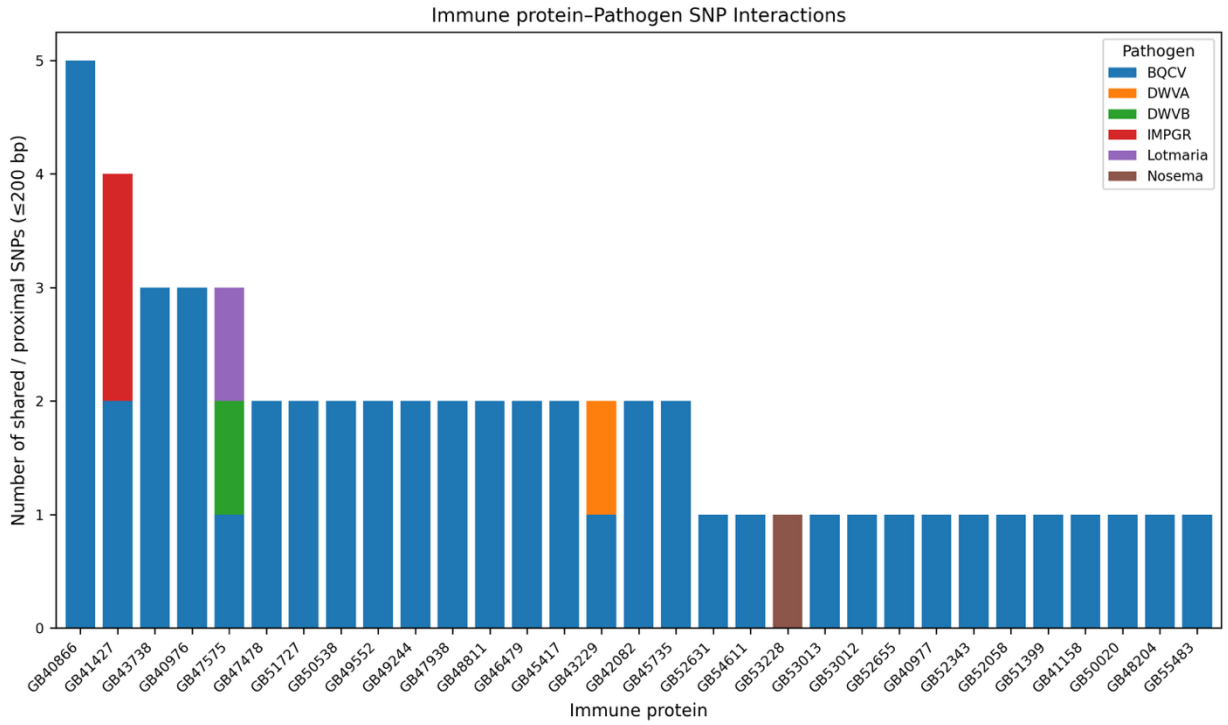
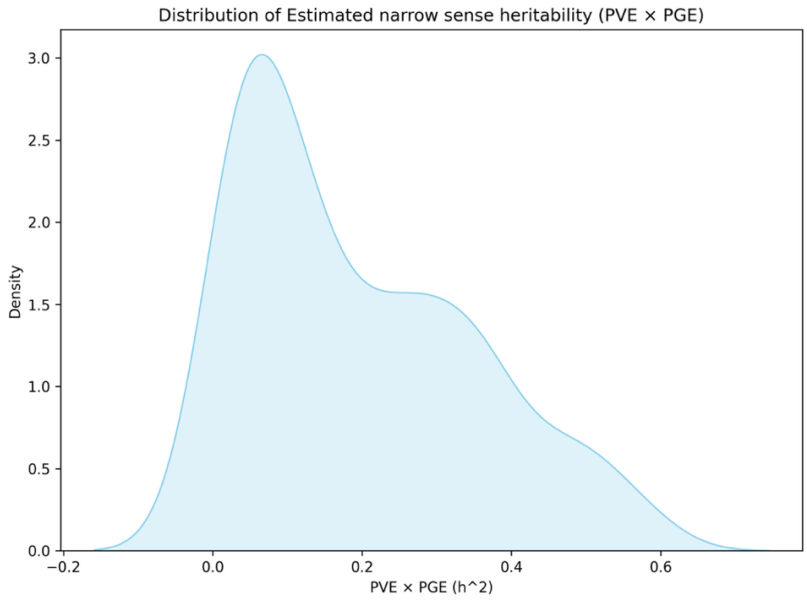


Figure 4.5. Genomic Co-localization of Immune Protein abundance and Pathogen abundance within colonies. Stacked bar plot illustrating the number of candidate SNPs that influence both pathogen loads and immune protein abundance within a 200 bp window. These loci are excellent candidates for understanding aspects of innate immunity that concurrently influence common pathogens found in honey bee colony.

Table 4.1: Classification of Immune Proteins (IPs) into different classes based on pathways and biological functions.

Pathways/ Mechanisms	IPs Count
Autophagy/Phagosome	6
Endocytosis	3
FoxO signaling pathway	2
Jak/STAT pathway	4
Lysosome	2
MAPK signaling pathway	3
Spliceosome	2
Toll and Imd signaling pathway	7

Others	37



Supplementary Figure 4.1: **Smooth kernel density plot showing the mean distribution of narrow-sense heritability estimates (h^2) of 64 immune proteins**.

Supplementary Table 4.1: Province and year-wise honey bee colonies used to measure Immune Proteins (IPs) abundance

Province/Year	**2016**	**2017**
Alberta (AB)	231	130
British Columbia (BC)	204	76
Manitoba (MB)	177	79
Ontario (ON)	155	42
Quebec (QB)	155	73

Supplementary Table 4.2: List of 64 IPs from literature review and mass spectrometry data. This table shows the name of Immune proteins, their NCBI ID, gene name and number of significant SNPs retained after GWAS filtration.

Protein ID	Gene ID	Gene Name
GB40866	Hsc70-4	heat shock protein cognate 4
GB40976	Hsp90	heat shock protein 90
GB40977	LOC552259	staphylococcal nuclease domain-containing protein 1
GB41158	LOC552034	toll-interacting protein
GB41427	Cat	catalase
GB42082	LOC552536	adipocyte plasma membrane-associated protein
GB42685	B-gluc1	beta-1,3-glucan recognition protein 1
GB42804	LOC412763	cardiomyopathy-associated protein 5
GB42887	LOC551922	NPC intracellular cholesterol transporter 2 homolog a
GB42981	LOC725832	beta-1,3-glucan-binding protein 1
GB43229	LOC410280	GTP-binding nuclear protein Ran,
GB43303	LOC551779	histone H2A.V
GB43638	LOC550857	protein enhancer of sevenless 2B
GB43738	PPO	phenoloxidase subunit A3
GB44311	Arp1	actin related protein 1
GB44650	LOC411829	neuroglian,
GB45417	LOC726883	CD109 antigen
GB45495	LOC411700	heat shock protein 83
GB45735	LOC413012	EH domain-containing protein 3
GB46268	LOC409410	plasminogen activator inhibitor 1 RNA-binding protein,
GB46360	LOC409849	immunoglobulin-like and fibronectin type III domain containing 12,
GB46479	LOC409093	flocculation protein FLO11,

GB46970	LOC412536	leukocyte elastase inhibitor
GB47104	LOC409663	lysozyme
GB47478	Gtpx1	glutathione peroxidase-like 1
GB47575	LOC412529	tyrosine-protein phosphatase corkscrew,
GB47805	Pgrp-s2	peptidoglycan recognition protein S2
GB47880	Sod1	superoxide dismutase 1
GB47938	LOC412825	sushi, von Willebrand factor type A, EGF and pentraxin domain-containing protein 1,
GB48204	LOC413980	CD109 antigen,
GB48208	LOC552062	protein argonaute-2,
GB48228	Pla2	phospholipase A2
GB48343	LOC413995	sensory neuron membrane protein 1,
GB48634	LOC726269	probable phospholipid hydroperoxide glutathione peroxidase,
GB48811	LOC408276	ATP-dependent RNA helicase bel
GB49244	LOC408939	charged multivesicular body protein 5
GB49544	Vg	vitellogenin
GB49552	LOC550671	Serine protease 3 (venom serine protease Bi-VSP)
GB50020	LOC413742	signal transducer and activator of transcription 5B
GB50259	LOC410580	synaptic functional regulator FMR1,
GB50522	LOC726457	flocculation protein FLO11,
GB50538	LOC412774	response to oxidative stress
GB50761	LOC410894	chymotrypsin-1
GB50866	LOC409228	scavenger receptor class B member 1,
GB51223	LOC406142	hymenoptaecin
GB51399	LOC551807	neurexin-4,
GB51727	LOC551720	activator of 90 kDa heat shock protein ATPase homolog 1
GB51741	PGRP-S3	peptidoglycan recognition protein S3

GB52058	LOC410395	chromobox protein homolog 1
GB52343	LOC409091	neurotrimin
GB52631	LOC100577844	Spatzle (spz), (Ortholog - FBgn0003495)
GB52655	LOC551140	signal transducing adapter molecule 1
GB53012	LOC408770	protein turtle
GB53013	LOC551628	protein borderless,
GB53014	LOC410982	protein croquemort,
GB53228	LOC724720	vacuolar protein sorting-associated protein 37B
GB54541	LOC100577408	leukocyte elastase inhibitor,
GB54611	LOC413749	serine protease inhibitor 88Ea
GB54664	LOC552479	protein draper,
GB54916	LOC409886	protein Mo25
GB54956	LOC552315	ubiquitin-like-conjugating enzyme ATG3
GB55388	LOC552073	arylsulfatase B
GB55483	LOC409821	twitchin
GB56012	LOC409286	stress-activated protein kinase JNK,

Chapter five: Integrating the genetics of Innate and Social immunity in honey bees

Tanushree Tiwari, Kathleen A Dogantzis, Alexandra Sébastien, Rodney T Richardson, Clement F Kent, Stephen Rose, Harshilkumar Patel, Alivia Dey, Ida M Conflitti, Judith Trapp, Bradford Vinson, Nikolay Stoynov, Kyung-Mee Moon, Queenie W.T. Chan, Nonno Hasegawa, Renata B. Labuschagne, Shelley E Hoover, Rob W Currie, Pierre Giovenazzo, M Marta Guarna, Stephen F Pernal, Leonard J Foster, Amro Zayed

Abstract:

Eusocial insects possess a dual-layered immune system composed of individual (innate) and social (behavioral) defenses, yet the genetic architecture coordinating these distinct mechanisms remains poorly understood. We quantified the genetic basis of social immunity, innate immunity, and pathogen load in 1,350 *Apis mellifera* colonies across Canada. Across all traits, we identified 1,797 candidate genes, with notable overlap among immune and pathogen categories, indicating shared genetic architecture underlying immune function and pathogen dynamics. From these, 41 core genes were shared across all three trait classes. We further identified 14 pleiotropic genes exhibiting antagonistic effects between social and innate immunity. Notably, a prominent subset of core genes encodes proteins involved in neural signaling and behavioral regulation, including Dop3, nAChRa6, FoxP, Nrx-1, and Antp. Together, these findings suggest a shared molecular framework linking sensory perception and signaling pathways to the coordination of innate and social immune responses in honey bees.

Introduction:

Honey bees (*Apis mellifera*) are one of the most economically and ecologically important pollinators worldwide (Klein et al., 2007; Potts et al., 2010). As managed pollinators, they enhance crop yields and quality, supporting global food security and agricultural sustainability (Aizen & Harder, 2009). Honey bees are eusocial, where the fitness of colonies depends on coordinated interactions among thousands of individual worker bees (Seeley, 1997). While this eusocial organization provides numerous advantages (e.g. cooperative brood care, division of labor, etc.), it simultaneously increases exposure to pathogens through high population density and frequent social interactions, ripe conditions for the spread of pathogens and diseases (Naug

& Camazine, 2002; Schmid-Hempel, 1998). This social structure creates intense pressure on immune mechanisms that must operate effectively across both individual and collective levels. This is important as colonies face a complex and large number of interacting stressors, including the ectoparasitic mite *Varroa destructor* and various RNA viruses, which frequently act together to amplify disease impacts (Doublet et al., 2015; Rosenkranz et al., 2010).

At the individual level, honey bees have a functional innate immune system that utilizes cellular defenses (Evans et al., 2006) and conserved signaling pathways such as Toll, Imd, JAK/STAT, JNK, and RNA interference (Brutscher & Flenniken, 2015). Comparative genomic analyses have revealed that *A. mellifera* possesses a significantly reduced number of innate immune genes (Barribeau et al., 2015), particularly in pathogen recognition and antimicrobial families (Evans et al., 2006). While the innate immune gene repertoire of honey bees may be small, it still appears effective in combating pathogens. For example, innate responses show substantial heritable variation in the regulation of antimicrobial peptides (AMPs) (Evans & Spivak, 2010) such as *defensin-1*, *abaecin*, and *hymenoptaecin*. Moreover, these genetic differences in the expression of AMPs are directly associated with colony-level pathogen tolerance and survival (Doublet et al., 2017a).

Social immunity comprises a suite of collective behaviors that reduce pathogen transmission and infection risk at the colony level (Cremer et al., 2007). One of the best-characterized forms is hygienic behaviour, in which workers detect and remove diseased or dead brood from capped cells, thereby interrupting pathogen replication and spread (Lapidge et al., 2002; Oxley et al., 2010; Spivak & Reuter, 1998). A specialized subtype of hygienic behaviour is Varroa-sensitive hygiene (VSH)(Harbo & Harris, 2005), where workers selectively identify and remove brood infested with *Varroa destructor*. By targeting reproducing mites within brood cells, VSH disrupts mite reproduction and reduces viral amplification within the colony. Adult workers also engage in grooming behaviours, including auto-grooming and allogrooming (Cini et al., 2020). Through these behaviors, bees can physically remove *Varroa* mites from nestmates, directly lowering parasite load and limiting pathogen transmission between individuals. Colonies may also express social fever**,** a thermoregulatory response in which workers elevate brood nest temperature to inhibit pathogen development. In addition, honey bees collect and deposit antimicrobial plant resins (propolis) within the hive (Simone-Finstrom & Spivak, 2012). This

"propolis envelope" reduces microbial growth and pathogen pressure that can modulate individual immune gene expression by lowering baseline immune activation. Like innate immunity, social immune traits are heritable and respond to selection. For example, breeding for VSH has been shown to result in lower *Varroa destructor* infestations and associated viral loads (Danka et al., 2013). The genetic basis of social immunity is primarily polygenic and often associated with genes involved in neural signaling and sensory perception (Chapter 3)(Harpur & et al., 2019; Harpur et al., 2019; Oxley et al., 2010). This indicates that social defenses emerge from regulatory and neural networks that link sensory perception of pathogens to task allocation and collective decision-making.

The honey bee's immune system is composed of two important layers: individual innate cascades and colony-level social behaviours (i.e., social immunity). Early genomic analyses of *Apis mellifera* revealed a small set of innate immune genes compared to *Drosophila melanogaster* (Evans et al., 2006). This finding led some researchers to hypothesize that that collective social behaviors in honey bees might offset the need for robust individual innate immune defenses. This in theory is expected to lead to relaxed molecular evolution on innate immune genes, leading to a small complement of innate immune genes (Evans et al., 2006). Several honey bee innate immune genes do show population genetic evidence of relaxed selection (Harpur & Zayed, 2013), which provides some support to this hypothesis. However, recent comparative studies suggest that immune gene numbers reflect phylogenetic inertia and evolutionary history, rather than eusociality *per se* (Barribeau et al., 2015; Mikhailova et al., 2024). For example, solitary fig wasps possess a streamlined immune system similar to the honey bee (Hou et al., 2021). Further evidence from a cross-order analysis of 27 insect species confirms that immune systems are shaped by long-term niche-specific adaptation (Shelke et al., 2025). While core signaling pathways remain highly conserved, key immune families such as PGRP-SA, Superoxide dismutase, Peroxidase, Cecropin A1 & C and Thioester-containing protein undergo frequent expansion and contraction across insect lineages. This dynamic gene turnover suggests that the honey bee's reduced immune repertoire reflects lineage-specific adaptation and functional specialization rather than simply being a by-product of eusociality (Shelke et al., 2025).

In addition to having a functional complement of innate immune genes(Cremer et al., 2007), eusocial insects have several colony-level behaviours that provide resilience against pathogens. For example Ants employ behaviors such as allogrooming and antimicrobial secretions to reduce disease transmission within colonies (Konrad et al., 2018), while termites utilize highly structured social immune strategies including nest hygiene and the removal of infected nestmates (Liu et al., 2019). Similarly, honey bees rely on collective behavioural defences such as grooming, dead brood removal, resin collection, nest sanitation and thermoregulation operating at the colony level, where coordinated worker responses function to limit pathogen spread and maintain colony health, reinforcing the concept of the colony as a superorganism (Evans & Spivak, 2010; Simone-Finstrom, 2017). Together, these findings support a revised framework in which social insect immunity is defined not by immune gene number alone, but by an integrated architecture combining conserved innate immune pathways with independently evolved social defense strategies. Within this framework, variation in immune gene repertoires reflects evolutionary history and ecological specialization, while social immunity represents a parallel adaptive layer enhancing colony-level resilience.

Here, we test the hypothesis that these layers of defense are not independent but rather share a partially overlapping genetic architecture. Specifically, we propose that genomic variation underlying innate immune function is linked to variation in social immune behaviors and pathogen load, such that shared genes simultaneously influence innate immunity and colony-level disease regulation. By identifying and studying these pleiotropic genes, we aim to quantify the degree of genetic interdependence within the honey bee immune system. This research represents one of the first holistic studies integrating innate immunity, social immunity, and pathogen-associated traits within a unified honey bee genomic framework.

Methods:

Experimental Design

Colonies were sampled over two years (2016-2018) across five Canadian provinces: British Columbia, Alberta, Manitoba, Ontario, and Quebec. In the first year, more than 900 colonies were established following 400 in the second year. Detailed colony management protocols are described in Borba et al. (Borba et al., 2022).

Phenotypic Measurements

Innate Immunity (II)

A total of 1,322 colonies were sampled, each represented by three biological replicates. Antennae from 20 worker bees per replicate were dissected and stored at -20°C. Samples were washed in ice-cold PBS, homogenized on ice, and proteins were extracted and precipitated using ethanol/glycogen. Protein pellets were resuspended, quantified and normalized to 1 µg/µL. From the assay proteins (15 µg per sample) were reduced with dithiothreitol, alkylated with iodoacetamide, and digested sequentially with LysC and trypsin. Peptides were isotopically labelled using dimethyl labelling (light, medium, heavy), pooled, desalted using C18 stage tips, and analyzed by LC Mass Spectrometry. A total of 3,673 MS files were generated for 2016 samples and 1,269 for 2017 samples. Raw peptide intensities were log10-transformed and median-centred. Outlier samples (>2 SD from the global median) were excluded. The 2017 dataset was aligned to the 2016 baseline using a median correction factor. Protein abundances were calculated as the mean normalized intensity of constituent peptides. A curated list of honey bee innate immune proteins was compiled from published literature. Immune-related proteins detected in antennal proteomes were retained for downstream analyses. Immune and Immune-related proteins are considered an Innate immunity (II) class for downstream analysis. Detailed protocols are described in Chapter 4.

Social Immunity (SI)

Hygienic behavior was assessed using the liquid nitrogen freeze-kill assay (M Marta Guarna et al., 2017). A 5 cm diameter section of sealed brood was freeze-killed and returned to the colony. The proportion of brood removed after 24 h was recorded. Colonies were assessed twice per season, and mean values were used for analysis. Phenotypes were normalized using ordered quantile normalization (Peterson, 2021b). Grooming behavior was quantified by assessing damage to Varroa mites collected on sticky boards over two consecutive 72-h periods before miticide treatment. Mites were examined under a microscope and classified as damaged or undamaged following established criteria. Grooming was expressed as the proportion of damaged mites and damage rate per day (Groom/Day). Colonies with fewer than five mites were excluded. Phenotypes were normalized using ordered quantile normalization (Peterson, 2021b). Detailed protocols are described in Chapter 3.

Pathogens

Varroa Mite Population Growth Rate was quantified using the instantaneous mite population growth rate (IMPGR), calculated following (Harris et al., 2003). Initial mite loads were estimated by alcohol wash (De Jong, Roma, & Gonçalves, 1982), and final populations were assessed using sticky board counts after amitraz treatment (Borba et al., 2022). Phenotypes were normalized using ordered quantile normalization (Peterson, 2021b). Nosema spp spore loads were quantified from homogenized worker bees using hemocytometer counts under phase-contrast microscopy. Mean spores per bee were calculated for each colony. Phenotypes were normalized using log-based transformation (Peterson, 2021b). Lotmaria passim load was measured by extracting DNA from homogenized worker bees, and L. passim infection levels were quantified by real-time PCR using SYBR Green chemistry (Jay D Evans et al., 2013). Absolute copy numbers were estimated using plasmid-based standard curves. Phenotypes were normalized using ordered quantile normalization (Peterson, 2021b). Viral Load was quantified using qPCR. RNA was extracted from pooled worker bees, reverse transcribed to cDNA, and viral loads of *BQCV,* DWVA*,* DWVB*,* and SBV were determined following established methods (Jay D Evans et al., 2013). Viral copy numbers were normalized per bee. Phenotypes were normalized using asinh(x) or ordered quantile normalization, depending on virus (Peterson, 2021b). Detailed protocols are described in Chapter 3.

Genotype Data

DNA was extracted from pooled legs of 50 worker bees per colony using a magnetic bead-based protocol optimized for high-throughput extraction. Libraries were sequenced on the Illumina HiSeqX platform, generating 150 bp paired-end reads at ~110× coverage. Reads were trimmed using Trimmomatic (Bolger et al., 2014b) and aligned to the Apis mellifera HAv3.1 reference genome (Wallberg et al., 2019) using NextGenMap (Sedlazeck et al., 2013a). Duplicate reads were marked using Picard, base quality scores recalibrated using GATK (Van der Auwera et al., 2013a), and SNPs identified using LoFreq (Wilm et al., 2012). Variants were filtered to remove sites with strand bias, low or excessive depth, low quality, proximity to indels, multiallelic loci, unplaced scaffolds, and loci with ≥90% missing data. The total number of 1,742,882 loci were retained after filtration. Detailed methods are described in Chapter 3.

Genome wide association analyses were conducted between all the trait categories and SNPs as predictors. Linear mixed models (LMM) were implemented using GridLMM (Runcie & Crawford, 2019), and Bayesian sparse linear mixed models (BSLMM) were implemented using GEMMA (Zhou & Stephens, 2014). Two independent BSLMM runs were conducted per protein (20 million iterations; 10 million burn-in). Identification of Candidate SNPs was done by integrating results from both LMM and BSLMM approaches. SNPs were retained if they showed large sparse effects ($\beta\gamma > 1$), posterior inclusion probability (PIP ≥ 0.01) (Comeault et al., 2014), and strong genome-wide association. SNPs with high confidence (PIP ≥ 0.10) were retained regardless of p-value. Detailed methods are described in Chapter 3.

Functional Annotation

Candidate SNPs from all the traits were annotated using SnpEff v5.2a (Cingolani, 2012). Following SnpEff annotation, genes associated with each candidate SNP were extracted for each phenotype based on genomic overlap. These gene sets were subsequently used for network construction and downstream comparative analyses across innate immunity, social immunity, and pathogen traits.

Analysis of Gene Overlap Across Phenotypic Traits

We used Venn diagrams to assess genomic similarity across phenotypic categories by comparing genes associated with annotated SNPs across innate immunity, social immunity, and pathogen-related traits. Using custom Python scripts, we identified intersections among phenotype-specific gene sets to distinguish unique genes from those shared across categories. These comparisons allowed us to characterize patterns of gene exclusivity and overlap among the three primary phenotypic groups.

To formally test whether innate immunity (II) and social immunity (SI) genes differed in their likelihood of overlapping with pathogen-associated genes, we conducted a contingency analysis using Fisher's exact test (Fisher, 1931). A 2×2 contingency table was constructed comparing the number of genes overlapping with pathogen traits to those not overlapping, separately for II and SI gene sets. Overlapping genes included both genes uniquely shared with pathogen traits and genes shared across all three categories. Fisher's exact test was implemented in Python using the SciPy statistical package to account for differences in category size and to evaluate whether the proportion of pathogen-associated genes differed significantly between

immune classes. Odds ratios were calculated to quantify the direction and magnitude of the association.

Gene-Based Network Analysis

Gene networks were constructed for pathogen effect size based on shared genes across immunity and pathogen traits. We analyzed three overlap categories: genes shared between innate immunity (II) and pathogen traits (II + Pathogen), genes shared between social immunity (SI) and pathogen (SI + Pathogen) and genes shared among innate immunity, social immunity, and pathogen simultaneously (II + SI + Pathogen). Networks were generated using custom Python scripts. Edge weights were derived from β coefficients estimated using gridLMM (Runcie & Crawford, 2019), with edge thickness scaled by the magnitude of effect size (β values) of pathogens and edge color representing effect direction. When multiple SNPs were associated with the same gene, the average β value was used to obtain a gene-level effect estimate. Network connectivity metrics reflecting pathogen associated interaction strength were calculated for each gene class. Differences among the three gene classes were assessed using analysis of variance (Cuevas et al., 2004; Tyanova et al.), followed by Tukey's honestly significant difference (HSD) test for post hoc pairwise comparisons (Tukey, 1949).

To evaluate potential trade-offs between innate immunity (II) and social immunity (SI), β values for II and SI associated SNPs were extracted for core genes (the genes present in both the Immune classes and pathogens) identified from gene-based Venn diagram analyses. When multiple SNPs were associated with a given gene, β values were averaged to obtain gene-level effect estimates for each immune category. These gene-level β values were visualized using a heatmap, enabling comparison of effect magnitude and direction across II and SI for shared genes. This visualization was used to explore concordant and contrasting effect patterns, providing insight into potential antagonistic or trade-off relationships between immune categories.

Results:

Gene Set Composition and Overlap Across Immune Categories

Based on the GWAS analyses, we compiled a comprehensive dataset of genes associated with candidate SNP loci underlying variation in pathogen load, innate immunity (II), and social immunity (SI) across 1,350 honey bee colonies. Across all traits, we identified 1,797 candidate

genes, comprising both unique and overlapping gene sets among II, SI, and pathogen-related traits (Figure 5.1). Among these, 193 genes were shared between pathogen traits and innate immunity, 13 genes were shared between pathogen traits and social immunity, and 41 genes were shared across all three trait categories. These shared genes represent candidates that jointly influence immune function and pathogen levels within colonies.

Although a greater absolute number of innate immunity genes overlapped with pathogen-associated traits compared to social immunity genes (234 vs. 54 genes), this pattern largely reflected the larger size of the innate immunity gene set. After accounting for differences in category size, Fisher's exact test revealed a significant association between immune category and pathogen overlap (odds ratio = 0.67, p = 0.024). Specifically, genes associated with social immunity were more likely (27.7%) to overlap with genes associated with variation in pathogen loads, relative to those associated with innate immunity (20.4%).

Gene interaction analysis

We visualized the network of 247 genes shared between immune traits and pathogen-associated traits across colonies (Figure 5.2). The network was dominated by genes with weak-to-moderate effect sizes, with a smaller subset showing larger effects. Here, β values represent effect sizes estimated from GWAS of pathogen-associated traits.

Among annotated genes with comparatively large effect sizes, several functional categories were represented. Within the innate immunity (II) category, LOC410006 (Rap GTPase-activating protein radish), a regulator of small GTPase signaling, showed strong associations with Sacbrood virus *(*SBV*)*. LOC410788 (*NMDA-type glutamate receptor-like*), involved in glutamatergic signaling, exhibited large effect sizes across viral load traits. Within the social immunity (SI) category, LOC551259 (*titin*), a structural muscle protein, was among the genes with the largest effect sizes for viral traits. LOC408960 (*slit guidance ligand*), identified in both II and SI categories and functionally linked to cell guidance and signaling pathways, also displayed consistently large effect sizes across viral traits. In addition to annotated genes, uncharacterized genes are also represented among the strong associations. LOC100577283, shared between II and SI categories, showed pronounced effects for Black Queen Cell Virus (BQCV), while the SI-associated locus LOC113218862 exhibited large effect sizes for Sacbrood virus *(*SBV*)*.

To determine whether genes associated with innate immunity (II), social immunity (SI), or both (II+SI) differed in their genetic effects on pathogen-level phenotypes, we compared the mean absolute effect sizes (β) across immune classes using a one-way ANOVA. No significant effect of immune class was detected (ANOVA: $F_{2,323} = 0.044$, $P = 0.957$). Post hoc Tukey HSD comparisons confirmed the absence of pairwise differences among groups (all adjusted $P = 0.90$). Mean absolute effect sizes were highly similar across categories (II: 0.60 ± 0.72; II+SI: 0.57 ± 0.84; SI: 0.59 ± 0.55). Substantial within-group variance was observed in all three immune classes, indicating broad dispersion of effect sizes irrespective of immune classification (Figure 5.3).

Pleiotropic genes and genetic tradeoffs in immune function

A lot of pleiotropic genes (Figure 5.4) are annotated for molecular function terms associated with neural signaling and behavioral regulation, including *Dop3* (dopamine receptor), *nAChRa6* (nicotinic acetylcholine receptor 6), *FoxP, Nrx-1,* and *Antp*. Beyond neural pathways, the pleiotropic gene set also includes genes involved in structural organization, signaling cascades, and transcriptional regulation. By comparing the directionality of the effect sizes between II and SI traits, we can distinguish putative cases of evolutionary trade-offs versus potentially coordinated innate and social responses. We found that 14 genes exhibited potentially antagonistic pleiotropy, where the allele associated with an increase in one immune class (e.g., an increase in the expression of an II protein) corresponded to a decrease in the other (a reduction in an SI behaviour). However, the majority (27 genes), including *NLG-5, nAChRa6,* and *Dop3*, showed effects in the same direction.

Discussion:

Network analysis of the 247 shared genes revealed that most genes had weak-to-moderate effect sizes, but a smaller set of genes, including neural, signaling, and structural genes, had large effects on multiple pathogen traits. However, we didn't observe a significant difference in the effect sizes of innate immune genes, social immune genes, or genes that affect both types of immunity on pathogen levels within colonies. This suggests that both social immunity and innate immunity play important roles in limiting the spread of pests and pathogens within colonies.

We discovered 41 pleiotropic genes that influence multiple innate immune traits. These genes exhibited two contrasting patterns: one group of genes exhibited putative antagonistic relationships where mutations in genes were associated with a numerical increase in innate immune protein expression and a decrease in the levels of social immunity (i.e. lower hygienic behavior or grooming), or vice versa. The other group of genes had consistent effect sizes for both types of immunity (i.e. higher levels of hygienic behavior and greater levels of expression of immune proteins). Genes with opposite phenotypic effects on social and innate immunity potentially reflect antagonistic pleiotropy, although we note, that while a reduction in social immunity has a clear effect on pathogens, understanding the direction of changes in innate immune proteins is not as clear. For example, one would expect the expression level of an innate immune effector protein to be directly related to its effect on pathogens (Lemaitre & Hoffmann, 2007). However, some signalling molecules associated with immunity can work in a complex and often context-dependent manner. Some genes act as activators while others act as repressors of innate immune systems (Aggarwal & Silverman, 2008; Boutros et al., 2002). In some cases, increased expression of regulatory genes may not enhance immune function but instead suppress components of immune signaling. For example, transcription factors within the Toll pathway, such as Dorsal, can differentially regulate downstream immune genes depending on regulatory context and interacting cofactors (De Gregorio et al., 2001). Consequently, higher expression of certain immune regulators could reduce the expression of antimicrobial effectors rather than enhance them. As such, it is difficult to determine whether pleiotropic genes exhibiting opposite effects on innate immune protein expression and social immune behaviours truly represent genetic antagonism or instead reflect complex regulatory relationships within immune signaling networks. Similarly, pleiotropic genes with consistent effects on both innate immunity and social immunity may not necessarily represent an enhancement of both traits. So, while our study does not conclusively point to genetic trade-offs in immune function, the relatively large number of genes that affect both innate immunity, social immunity and pathogen levels do create the conditions where potential genetic tradeoffs can exist.

An interesting feature of pleiotropic immune genes was their involvement in molecular functions associated with neural signaling, synaptic organization, and transcriptional regulation genes. This pattern suggests that core sensory perception and signal transduction pathways may act upstream of both innate immune activation and socially mediated defense behaviors. In

insects, pathogen detection relies on conserved signaling cascades (e.g., Toll, Imd, and JAK/STAT pathways) that integrate environmental cues and coordinate systemic immune responses (Evans et al., 2006; Lemaitre & Hoffmann, 2007). At the same time, social immunity behaviors such as hygienic brood removal and grooming depend on sensory discrimination, neuromodulator signaling, and rapid behavioral decision-making (Oxley et al., 2010; Perry & Barron, 2013; Spivak & Reuter, 1998). The repeated appearance of neuromodulators (e.g., dopaminergic and cholinergic receptors), synaptic regulators, and transcriptional control genes influencing both innate and social immunity suggests that these pathways may represent shared regulatory nodes linking pathogen detection to both physiological and behavioral outputs. This pattern is consistent with the hypothesis that components of ancestral immune detection and signaling networks have been co-opted during eusocial evolution to regulate derived social defenses (Cremer et al., 2007; Robinson et al., 2005; Whitfield et al., 2006). Eusocial insects frequently repurpose conserved molecular pathways to generate colony-level phenotypes (Robinson et al., 2008; Amy L Toth & Gene E Robinson, 2007). Under this framework, variation in upstream sensory and signaling genes could simultaneously modulate immune responsiveness at the individual level and behavioral responsiveness at the colony level.

In addition to well-characterized neural and signaling regulators, several of the strong associations involved unannotated genes with no clear orthologs in *Drosophila melanogaster*. These genes exhibited large effects on innate immunity, social immunity, and in some cases both immune domains. Honey bee functional annotation still relies heavily on orthologs to *Drosophila*, yet comparative analyses have shown that eusocial insects possess lineage-specific genes that are not represented in genetic model organisms such as *Drosophila* (Johnson & Tsutsui, 2011). Such taxonomically restricted (i.e. evolutionarily novel) genes have been repeatedly implicated in the emergence of derived social traits in eusocial insects (Amy L Toth & Gene E Robinson, 2007). Previous genomic comparisons suggested that honey bees possess fewer innate immune genes relative to dipterans (Consortium, 2006), but subsequent work has emphasized that immune complexity cannot be inferred solely from orthologs to model organisms (Evans et al., 2006) Our results are consistent with the possibility that some components of honey bee's innate immunity, particularly those interfacing with behavioral defenses are mediated by genes that remain uncharacterized because they are lineage-specific or rapidly evolving.

Conclusion:

Our study represents the first effort to study the genetic architecture of both social and innate immunity and how they influence pathogen loads within eusocial colonies.
Both innate immunity and social immunity had some independent compartments, but there was a notable set of pleiotropic genes that influenced both types of immune systems. The genetic effects of innate immune proteins on pathogen levels were not significantly different from those of social immune genes, indicating that social immunity is not necessarily a more powerful arm of immunity in honey bees. We discovered several key neuronal genes that influenced both innate and social immunity, a finding that suggest that there is a core upstream genetic machinery involved in sensory perception and signalling that is coordinating the actions of both innate and social immunity. Given that social immunity is derived, it is likely that the core sensory and signaling molecules used to trigger innate immune cascades have been co-opted to facilitate social immunity in insects. Our study sheds new insights on the genetics of immunity in social insects and provides several practical knowledge for enhancing the health of honey bees via selective breeding.

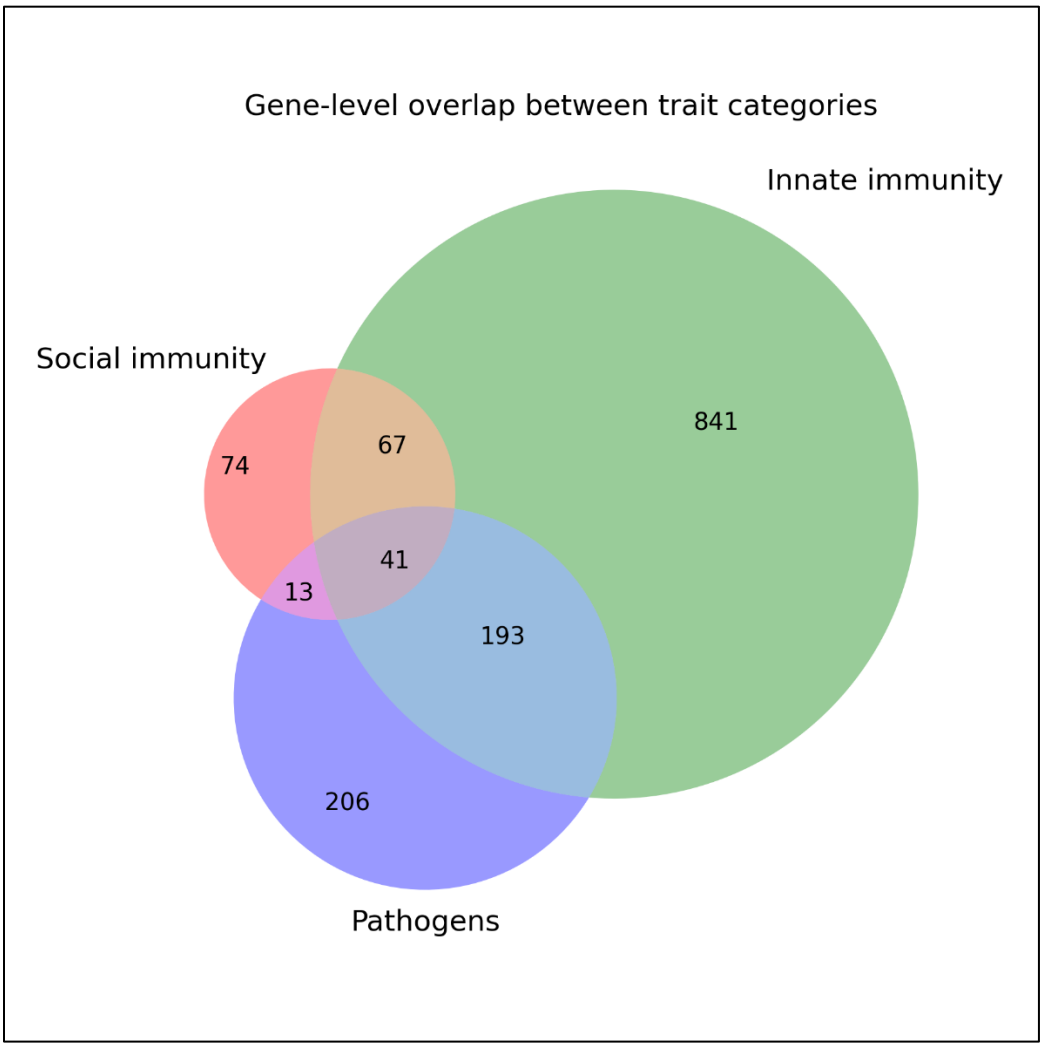


Figure 5.1. Shared genetic architecture of the honey bee's immune system and pathosphere. Venn diagram illustrating the gene-level overlap between innate immunity, social immunity, and pathogen loads. The intersection of all three categories reveals a core set of 41 pleiotropic genes that appear to influence innate immunity, social immunity and the abundance of pathogens within honey bee colonies.

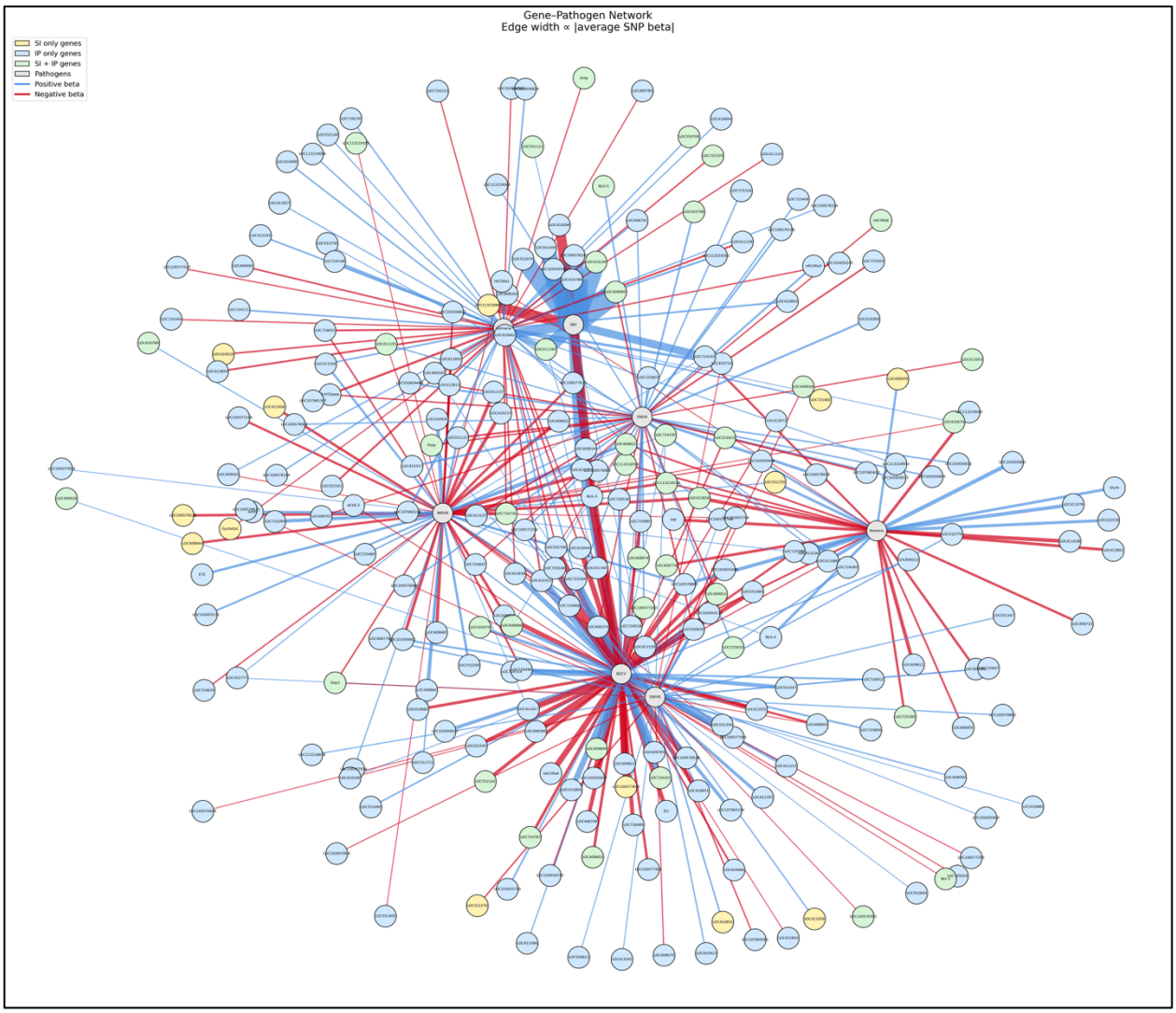


Figure 5.2. The interconnected web of innate and social immunity and the honey bee's pathosphere. Network analysis revealing the extensive genetic cross-talk between innate immunity (II), social immunity (SI), and pathogen load. The high density of interconnected nodes and shared pleiotropic hubs demonstrates that honey bee combat pathogens via a relatively integrated immune system that involves both social and innate immune pathways. Nodes represent genes and pathogens, with node colors indicating immune category. Edges represent SNP-based associations, with edge width proportional to the average effect size (β value) across

SNPs mapped to each gene and edge color indicating effect direction (red, negative; blue, positive).

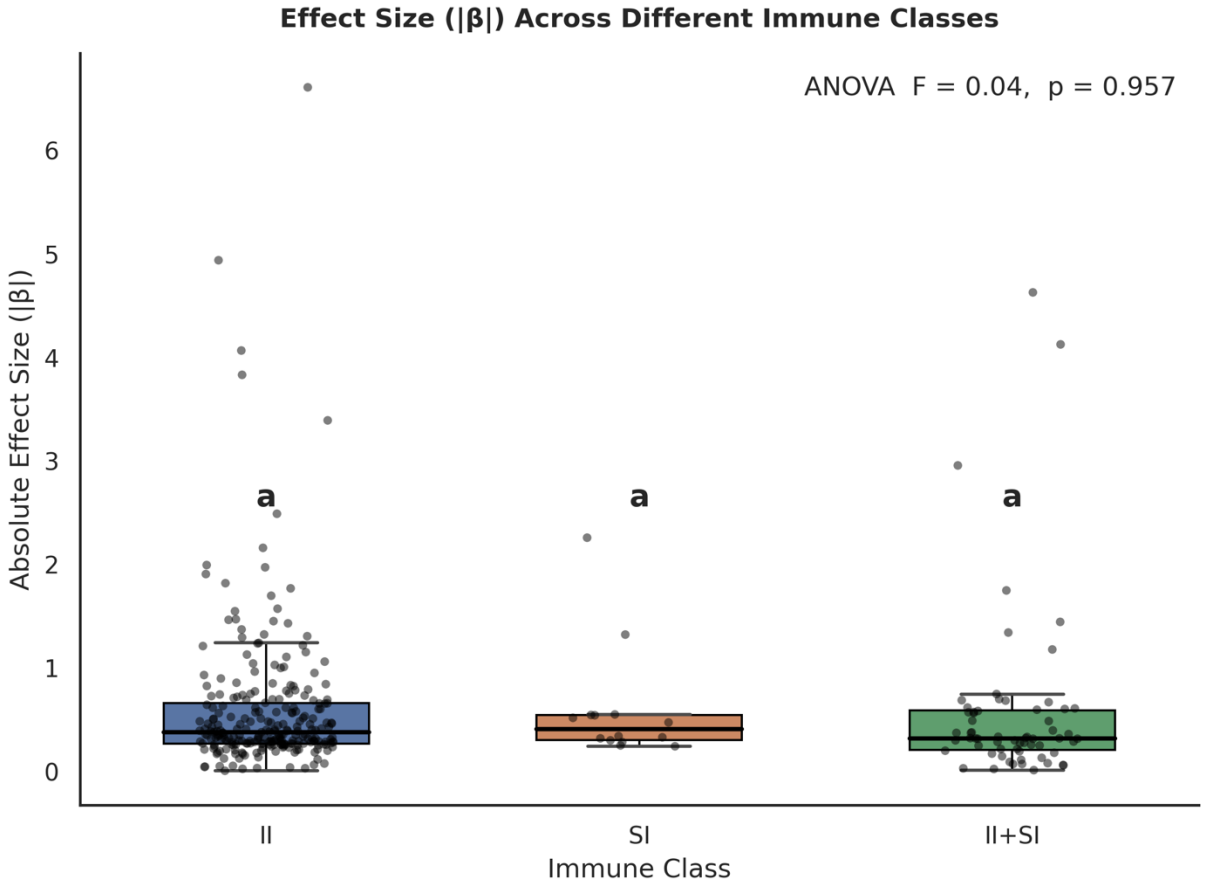


Figure 5.3: Homogeneity of genetic effect sizes across defense classes. Analysis of absolute effect sizes (β) for SNPs associated with innate immunity (II), social immunity (SI), and shared pleiotropic gene (II+SI). The lack of a significant difference in effect size (ANOVA, F=0.04, p=0.957) highlights a homogenous genetic architecture underlying honey bee health. This demonstrates that no single defense class predominates over the others.

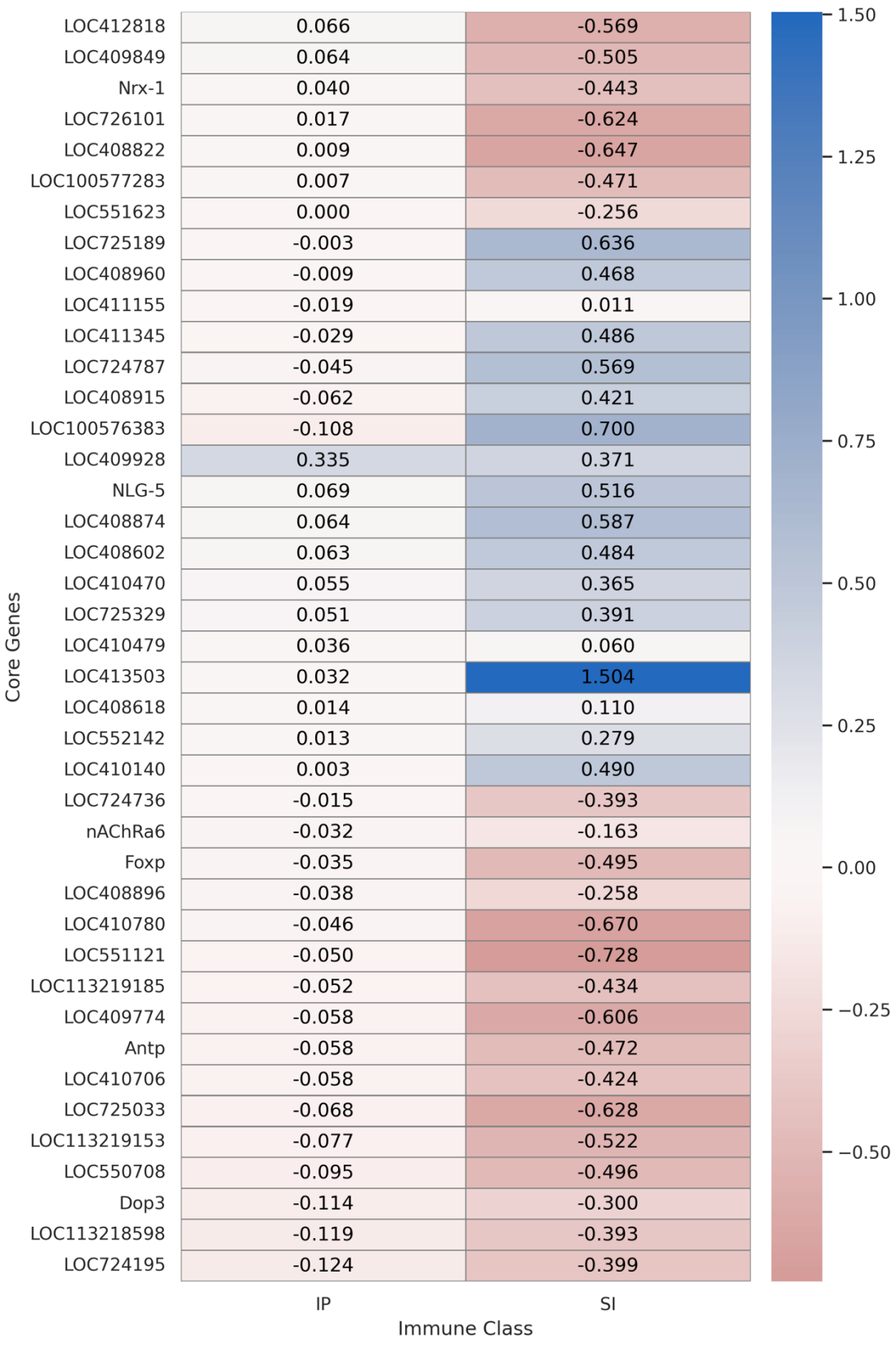

Effect size (β value) of 41 core genes in SI and IP
Core Genes
LOC412818 0.066 -0.569
LOC409849 0.064 -0.505
Nrx-1 0.040 -0.443
LOC726101 0.017 -0.624
LOC408822 0.009 -0.647
LOC100577283 0.007 -0.471
LOC551623 0.000 -0.256
LOC725189 -0.003 0.636
LOC408960 -0.009 0.468
LOC411155 -0.019 0.011
LOC411345 -0.029 0.486
LOC724787 -0.045 0.569
LOC408915 -0.062 0.421
LOC100576383 -0.108 0.700
LOC409928 0.335 0.371
NLG-5 0.069 0.516
LOC408874 0.064 0.587
LOC408602 0.063 0.484
LOC410470 0.055 0.365
LOC725329 0.051 0.391
LOC410479 0.036 0.060
LOC413503 0.032 1.504
LOC408618 0.014 0.110
LOC552142 0.013 0.279
LOC410140 0.003 0.490
LOC724736 -0.015 -0.393
nAChRa6 -0.032 -0.163
Foxp -0.035 -0.495
LOC408896 -0.038 -0.258
LOC410780 -0.046 -0.670
LOC551121 -0.050 -0.728
LOC113219185 -0.052 -0.434
LOC409774 -0.058 -0.606
Antp -0.058 -0.472
LOC410706 -0.058 -0.424
LOC725033 -0.068 -0.628
LOC113219153 -0.077 -0.522
LOC550708 -0.095 -0.496
Dop3 -0.114 -0.300
LOC113218598 -0.119 -0.393
LOC724195 -0.124 -0.399
IP
SI
Immune Class
1.50
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50

Figure 5.4. Core immune system genes in honey bees. Heatmap of mean (β) effect sizes showing the directional coupling between innate immunity (II) and social immunity (SI) across 41 core pleiotropic genes. Genes showing antagonistic effects (opposite (β) directions) can possibly be involved in genetic trade-offs between innate and social immunity.

Supplementary Table 5.1. List of 41 core pleiotropic genes and their associated pathogen, innate immunity (II), and social immunity (SI) traits.

Gene ID	Gene Name	Pathogen-Traits	II – Immune Proteins Traits	SI-Traits
LOC409849	neural cell adhesion molecule fasciclin 2(Fas2)	BQCV	GB45417	Hygiene
LOC726101	uncharacterized LOC726101	BQCV	GB50866, GB40866, GB48343, GB43229, GB46479, GB41427, GB50538	Hygiene
LOC551623	nucleolysin TIAR	BQCV	GB50866, GB40866, GB49552, GB48343, GB43229, GB46479, GB53012, GB41158, GB48204, GB47478	GroomingPerDay
LOC724787	protein RUFY3	BQCV	GB49544, GB48811	Hygiene
LOC408915	protein couch potato	BQCV	GB51741, GB41158, GB43738	Hygiene
LOC408602	MKL/myocardin-like protein 2	BQCV	GB51399, GB48228, GB42685	GroomingPerDay
LOC410479	defective proboscis extension response 1(dpr1)	BQCV	GB50522, GB40866, GB46479, GB40976, GB51727, GB45735, GB41427, GB50538	Hygiene
LOC552142	latrophilin Cirl	BQCV	GB50522, GB43229, GB41158, GB48811, GB50538	Hygiene
LOC100577283	uncharacterized LOC100577283	BQCV, DWVA	GB40866, GB45495, GB48811, GB48204, GB41427, GB52343, GB47478	Hygiene
Foxp	forkhead box transcription factor P, Foxp	BQCV, DWVA	GB49544	GroomingPerDay

NLG-5	neuroligin 5	DWVA	GB40866, GB45495, GB47478	Hygiene
LOC408874	uncharacterized LOC408874	DWVA	GB40866, GB48343, GB53228, GB40976, GB41158, GB42082, GB53014, GB40977, GB44650	GroomingPerDay
LOC413503	TWiK family of potassium channels protein 18-like	DWVA	GB50866, GB40866, GB40977	GroomingPerDay
nAChRa6	nicotinic acetylcholine receptor alpha6	DWVA	GB48343, GB50259, GB40976, GB45417, GB50538, GB40977	Hygiene
LOC410706	RhoGTPase activating protein(cv-c)	DWVA	GB50866, GB40866, GB54916, GB50761, GB51399, GB41427, GB52343	Hygiene
LOC725033	uncharacterized LOC725033	DWVA	GB48343, GB50761	GroomingPerDay
LOC408960	slit guidance ligand(sli)	DWVA, SBV	GB50522, GB50866, GB40866, GB50020, GB46479, GB40976, GB51399, GB48204, GB47938, GB42082, GB41427, GB47478, GB42685, GB48634, GB40977	Hygiene
Nrx-1	neurexin 1(Nrx-1)	DWVB	GB40866, GB43738, GB46360	Hygiene
LOC100576383	alpha-Mannosidase class I a(alpha-Man-Ia)	DWVB	GB50866, GB43229, GB47575, GB50761, GB48811, GB41427, GB53013, GB51223, GB40977	GroomingPerDay
LOC409928	Suppressor of Triplolethal(Su(Tpl))	DWVB	GB51223	Hygiene
LOC408618	furin-like protease 2(Fur2)	DWVB	GB50866, GB43229, GB46479, GB48204, GB42685, GB50538, GB44650	GroomingPerDay
LOC551121	dipeptidase 1(LOC551121)	DWVB	GB43738	Hygiene

LOC410780	Na[+]/H[+] hydrogen antiporter 1(Nha1)	IMPGR	GB50866	GroomingPerDay
LOC113219185	uncharacterized LOC113219185	IMPGR	GB49244, GB47575, GB46479, GB40976, GB53012, GB41158, GB43738, GB48204, GB42082, GB45417, GB41427	Hygiene
LOC411155	Teneurin-a transmembrane protein (Ten-a)	IMPGR, DWVA	GB50522, GB54916, GB50259, GB47575, GB40976, GB45735, GB48204, GB47478	GroomingPerDay
Dop3	dopamine D2-like receptor (Dop3)	IMPGR, DWVB	GB52655	GroomingPerDay
LOC408896	Semaphorin 2a(Sema2a)	IMPGR, BQCV, DWVA	GB49552, GB53228, GB52655, GB40976, GB51727, GB41158, GB42082, GB47478	GroomingPerDay
LOC725329	misexpression suppressor of ras 6(MESR6)	*Lotmaria*	GB50866, GB40866, GB47575, GB41158, GB43738, GB45417, GB41427	Hygiene
LOC406077	homeotic protein antennapedia(Antp)	*Lotmaria*	GB50522, GB47478	Hygiene
LOC550708	E3 ubiquitin-protein ligase mind bomb 1(mib1)	*Lotmaria*	GB40866, GB48228, GB41427, GB42685, GB50538	GroomingPerDay
LOC411345	neurotrimin	*Lotmaria,* DWVB, SBV	GB47880, GB50866, GB40866, GB47575, GB40976, GB47938, GB40977	Hygiene
LOC410470	polypeptide N-acetylgalactosaminyltransferase 9(Pgant9)	*Lotmaria,* IMPGR	GB50522, GB50866, GB40866, GB52631, GB52655, GB50538	Hygiene
LOC724736	semaphorin-1A	*Lotmaria,* IMPGR, DWVB	GB48343, GB46479, GB40976, GB49544, GB41427, GB50538	GroomingPerDay
LOC724195	irregular chiasm C-roughest protein teiresias(tei)	*Lotmaria,* IMPGR, BQCV,	GB50761, GB47938	Hygiene

		DWVB		
LOC725189	protein bric-a-brac 2	*Nosema*	GB41158, GB55483	Hygiene
LOC409774	Proteoglycan Cow	*Nosema*	GB41158, GB41427	GroomingPerDay
LOC113218598	uncharacterized LOC113218598	*Nosema*	GB50522, GB47104, GB41427	Hygiene
LOC412818	glutamate ionotropic receptor NMDA type subunit 2(Nmdar2)	*Nosema,* IMPGR	GB40866, GB43229, GB53228, GB45735, GB52343	GroomingPerDay
LOC113219153	uncharacterized LOC113219153	*Nosema,* IMPGR	GB46970, GB43229, GB55483	GroomingPerDay
LOC408822	uncharacterized LOC408822	*Nosema, Lotmaria,* IMPGR	GB50866, GB40866, GB43229, GB51399, GB41158, GB49544, GB43738, GB42082, GB45417, GB52343	GroomingPerDay
LOC410140	gamma-aminobutyric acid type B receptor subunit 2(GABA-B-R2)	SBV	GB50522, GB41427, GB47478, GB44650	GroomingPerDay

Chapter six: Genome-wide distribution and functional implications of short InDels in honey bees (*Apis mellifera).*

Tanushree Tiwari, Clement F Kent, Kathleen A Dogantzis, Rodney T Richardson, Stephen Rose, Harshilkumar Patel, Alivia Dey, Ida M Conflitti, Renata B. Labuschagne, Shelley E Hoover, Rob W Currie, Pierre Giovenazzo, M Marta Guarna, Stephen F Pernal, Leonard J Foster, Amro Zayed

Abstract:

Honey bees (*Apis mellifera*) are critical pollinators that provide indispensable pollination services to a large number of crops. Studies on the population genomics of honey bees, which have proven very useful for understanding the genetics of complex traits that can be used to inform breeding, have so far focused only on single nucleotide polymorphisms (SNPs). Insertions and deletions are the second most common source of genetic variation, which are important for fitness and quantitative traits in number of other organisms. Here, we characterize short (1-5 bp) InDel variation in honey bee colonies using population-scale whole genome sequencing of pooled workers from 1350 colonies sequenced at very high depth from across Canada. We identified over 125,000 short InDels, the majority of which were in non-coding regions, with intronic and intergenic variants predominating. We then asked how InDel mutations influence the genetic architecture of five heritable colony-level traits, which have been previously analyzed using SNP markers. Relative to SNPs, fewer InDels were associated with the traits in absolute number; however, the proportion of suggestive associations relative to the total number of variants tested was higher for InDels. Despite this, their estimated genetic effect sizes were generally smaller than SNPs. Together, these results provide new insight into the population genetics of short InDels in honey bees and highlight their complementary role in shaping complex colony-level traits central to colony fitness.

Introduction:

Genetic variation arises through multiple types of mutations, including single-nucleotide polymorphisms (SNPs), insertions and deletions (InDels), copy-number variants (CNVs), and larger structural rearrangements. These mutational processes provide the raw material for evolutionary change, with the fate of individual variants shaped by selection, drift, migration, and demographic history (Fisher, 1931; Kimura, 1985; Lynch, 2010) SNPs have dominated

population genomic and association studies largely because of their abundance and relative ease of detection. However, growing evidence suggests that other forms of variation, including InDels, contribute to functional and phenotypic diversity.

Insertions and deletions are defined as the gain or loss of short stretches of DNA, typically ranging from one to several base pairs. InDels represent one of the most abundant forms of genetic variation in eukaryotic genomes, second only to SNPs in frequency (Ajawatanawong & Baldauf; Mills et al., 2006). Despite their prevalence, InDels have been comparatively understudied due to challenges related to their accurate detection and genotyping from short-read sequencing data (Muzzey et al., 2015). Advances in next-generation sequencing methods, reduced costs, larger sample sizes, and advanced variant-calling algorithms have now made it possible to reliably characterize short InDels at the population scale, enabling renewed investigation into their evolutionary and functional roles.

Short InDels can have substantial phenotypic effects by altering protein coding sequences or modifying regulatory elements that control transcription, splicing, and chromatin accessibility (Lin et al., 2017). In coding regions, InDels can introduce frameshifts or alter protein length and are therefore often subject to strong purifying selection (Rockah-Shmuel et al., 2013) In contrast, InDels located in noncoding regions, particularly within promoters, enhancers, untranslated regions and introns, can be neutral or modulate gene expression without functional disruption, required for adaptive regulatory evolution (Montgomery et al., 2013). Consistent with this, population studies in humans have shown that short InDels contribute to variation in gene expression, disease susceptibility, and complex traits, often with effect sizes comparable to or larger than those of SNPs (Montgomery et al., 2013).

Evidence for the adaptive importance of InDels is not limited to humans, they are particularly abundant in plants, where both coding and regulatory InDels have been repeatedly implicated in ecologically and agronomically important traits. In maize and rice, short InDels contribute to variation in flowering time, plant body, and domestication related phenotypes, in some cases representing the causal variants underlying major trait differences (Kou et al., 2022; Zhou et al., 2024). Regulatory InDels affecting transcription factor binding or gene expression have also been shown to play key roles in adaptive divergence and crop domestication, highlighting the importance of small insertion-deletion mutations in shaping complex traits

(Meyer & Purugganan, 2013; Studer et al., 2011). These plant studies underscore the broader evolutionary relevance of InDels and suggest that their contribution to phenotypic variation may be systematically underestimated when analyses focus primarily on SNPs.

In insects, insertion and deletion (InDel) polymorphisms represent an important source of genetic variation and can influence development, fitness, and adaptive evolution. For example, in *Drosophila melanogaster*, InDels alter gene structure and exhibit strong signatures of selection in both coding and regulatory regions (Presgraves, 2006). Despite their functional potential to induce frameshifts or modify regulatory architecture, most population genomic and genome-wide association studies (GWAS) focus predominantly on single-nucleotide polymorphisms (SNPs) (Tiwari & Zayed, 2021). As a result, the contribution of InDels to complex adaptive traits remains comparatively understudied, especially for traits shaped by structural variation rather than single-base changes.

The western honey bee (*Apis mellifera*) is an important pollinator, but managed colonies are currently at risk from exposure to a plethora of biotic and abiotic stressors (French et al., 2024). Efforts to mitigate global colony declines have relied heavily on population and quantitative genomics to identify SNP markers of resilience and colony performance (Dogantzis & Zayed, 2019; Grozinger & Zayed, 2020a; Harpur & Zayed, 2013; Tiwari & Zayed, 2021). SNP-based population genomic studies in honey bees have helped elucidate the species biogeographical history and highlighted mutations with putative patterns of adaptive evolution (Chen et al., 2016; Dogantzis, 2021). SNP-based GWAS have been valuable for studying the genetic architecture of complex traits in bees (Chapter 3), including aggression (A. Avalos et al., 2020), hygienic behaviour (Harpur et al., 2019), and thelytokous parthenogenesis (Yagound et al., 2020). Unlike SNPs, InDels are not well characterized despite their potential to restructure genes and regulatory elements. Addressing this gap is therefore critical for achieving a more complete view of honey bee genomics.

In this study, we characterized short insertion and deletion (InDel) variants in *Apis mellifera* colonies by examining their types and genomic distribution and by evaluating how these mutations contribute to complex colony-level phenotypes. To do so, we used whole-genome pooled sequencing of 1350 colonies studied across Canada. Using this extensive dataset, which also contains measurements of a large number of colony level traits, we investigated the

relationship between genetic variation and colony-phenotype by performing genome-wide association analyses (GWAS) for five heritable colony-level traits. We previously presented a SNP-based GWAS of these traits (Chapter 3); here, we quantified the contribution of InDel variation to genetic architecture and assessed how patterns of association differ from those identified using SNPs. Together, our results provide new insights into the population genetics of short InDels in honey bees and clarify their relative importance for colony-level traits central to colony fitness.

Methods:

Dataset Description

Insertions and deletions (InDels) were detected in 1,350 honey bee (Apis mellifera) colonies sampled in 2016 and 2017 across five provinces in Canada (Borba et al., 2022). These colonies represent a broad geographic distribution and provide a population level resource for examining structural genetic variation associated with colony level phenotypes.

DNA Extraction and Whole-Genome Sequencing

For each colony, the front left legs of 50 worker bees were pooled and homogenized using a mortar and pestle before DNA extraction. Pooling workers enables estimation of allele frequencies at the colony level and captures genetic variation arising from the queen's polyandrous mating system, in which queens typically mate with 15-20 drones (Tarpy, 2000). Sequencing 50 workers per colony provides sufficient power to detect alleles contributed by multiple patrilines and to accurately estimate within-colony allele frequencies (Inbar et al., 2020). Extracted pooled DNA was sequenced at the Génome Québec Innovation Centre (McGill University) using the Illumina HiSeq X platform, generating 150 bp paired-end reads. Sequencing achieved an average genome-wide coverage of approximately 110× per colony, enabling reliable detection of InDels and robust allele frequency estimation from pooled samples.

Read Processing, Alignment, and InDel Calling

Raw sequencing reads were processed using an in-house bioinformatics pipeline to ensure high-quality genotype calls for each colony. Initial quality assessment was performed using FastQC (Brown et al., 2017). Reads that passed quality thresholds were trimmed using

Trimmomatic (Bolger et al., 2014b), which removes adapter sequences and discards reads shorter than 50 bp. This step ensured retention of high-quality reads of sufficient length for reliable InDel detection.

Trimmed reads were aligned to the Apis mellifera reference genome (HAv3.1(Wallberg et al., 2019)) using NextGenMap v0.4.12 (Sedlazeck et al., 2013a). Resulting BAM files were sorted and indexed with SAMtools v1.3.1 (Li et al., 2009), and duplicate reads that have potential PCR artifacts were marked using Picard v2.1.0 (https://broadinstitute.github.io/picard/). Base quality score recalibration was performed using GATK v4.0.11.0 BaseRecalibrator (Van der Auwera et al., 2013a), incorporating previously identified variants as known sites. InDel calling was conducted on cleaned, sorted BAM files using LoFreq v2.1.2 (Wilm et al., 2012), a variant caller optimized for detecting low-frequency variants in pooled sequencing data. LoFreq estimates reference and alternate allele frequencies from mapped reads and outputs variant call format (VCF) files containing genomic positions, quality scores, read depth information, and additional variant-level metrics.

Variant Filtering and Dataset Construction and Validation

Initial filtering of InDel VCF files was performed at the individual colony level to retain high-confidence variants, based on sequencing quality, read depth, and allele frequency. Variants exhibiting strand bias (SB > 50), read depth < 10 or > 500, or variant quality scores < 90 were excluded. In addition, loci with alternative allele frequencies < 0.20 or > 0.90 were removed to eliminate rare or near-fixed variants with limited power for downstream analyses. Following this initial quality and frequency filtering, individual colony-level VCF files were merged into a single dataset. A second round of filtering was applied to the merged VCF to remove variants located on unplaced scaffolds. Only InDels present in at least 80% of sampled colonies were retained for downstream analyses. Custom Python scripts were used to quantify unique insertions and deletions and characterize their length distributions. To assess the robustness of the final variant set, InDels detected in our study were compared with previously reported honey bee InDels deposited in the European Variation Archive (EVA) (https://www.ebi.ac.uk/eva/).

Genomic distribution of variant classes relative to centromeres

InDel density was quantified across chromosomes to characterize its genomic distribution and to test whether InDels are randomly distributed or show structured patterns in relation to

centromere position. Centromere coordinates were obtained from chromosome-level *Apis mellifera* genome assemblies in which centromeres were mapped using cytogenetic evidence, centromere-associated AvaI repeats, and regions of suppressed recombination (Wallberg et al., 2019). Each chromosome was divided into non-overlapping 500 kb windows, and InDels were counted within each window to calculate variant density as the number of variants per base pair.

Using the same window framework, high-confidence SNPs from the same dataset were then included to evaluate whether different classes of genetic variation show similar or distinct distribution patterns relative to centromeres. To test the relationship between variant density and centromere proximity while accounting for chromosome-specific differences, linear mixed-effects models were fitted separately for SNPs and InDels, with distance from the centromere included as a fixed effect and chromosome as a random intercept in the model (Equation 1).

$$Equation\ 1: variant\_density \sim \text{dist\_from\_centromere} + (1 \mid \text{chromosome})$$

The significance of the fixed effect was assessed using the z-statistic and associated p-value, enabling comparison of distance-dependent patterns between SNPs and InDels and testing whether these variant types differ significantly in their distribution relative to centromere proximity.

To further assess whether SNP and InDel densities exhibit similar genomic distribution patterns, we calculated Pearson's correlation coefficient across matched 500 kb windows. All analyses were performed in Python (version 3.5).

Functional Annotation of Variants

All InDels were functionally annotated using SnpEff (Cingolani, 2012) to assign genomic feature context. For each InDel, SnpEff was used to classify the variant according to the genomic feature it overlaps, including coding sequences, introns, untranslated regions (UTRs), and intergenic regions, as well as by functional category, such as frameshift or non-frame insertion or deletion. These annotations were used to compare the distribution of insertions and deletions across genomic feature classes.

To test whether insertions and deletions differ in their genomic distribution, contingency tables of variant counts across annotation categories were constructed and evaluated using chi-

square tests. High-confidence SNPs from the same dataset were then annotated using the same SnpEff framework, and analogous chi-square tests were performed to compare the distribution of SNPs and InDels across genomic feature classes. This approach allowed assessment of whether different classes of genetic variation show significantly different patterns of functional annotation within the genome.

Phenotypic Measurements [Refer to chapter 3 for more detailed methods]

Phenotypic data were collected from the same *Apis mellifera* colonies used for InDel discovery and genomic analyses, enabling direct integration of genetic and phenotypic information. Colony-level traits were selected based on prior evidence of heritability and were chosen to represent distinct functional classes of colony performance, including economic productivity, social immunity, and pathogen susceptibility. Economic productivity was measured as honey production, social immunity was assessed through hygienic behavior and grooming behavior per day, and pathogen susceptibility was quantified as viral load. This selection allowed genetic variation to be examined across multiple, biologically relevant dimensions of colony-level phenotypes.

Honey production was estimated as the change in total colony weight over two weeks during peak honey flow (Szabo & Lefkovitch, 1989) using a digital platform or hanging scale (Global Industrial Canada, Toronto, ON, Canada). Hygienic behavior was assessed using the liquid nitrogen freeze-killed brood assay following established protocols (M Marta Guarna et al., 2017; Spivak & Reuter, 1998), with each colony evaluated twice, 3–10 days apart. Hygienic behavior was quantified as the proportion of freeze-killed brood removed within 24 hours and averaged across the two assessments; colonies with a mean removal rate ≥ 0.95 were classified as hygienic. Grooming behavior was measured as the proportion of fallen *Varroa destructor* mites showing physical damage. Sticky boards made from plastic-coated freezer paper coated with petroleum jelly were placed beneath screened bottom boards for two consecutive 72-hour periods in the fall, prior to miticide treatment. Collected boards were combined, and mites were removed and preserved in 70% ethanol. Colonies with more than 200 mites were subsampled to 200 mites, while colonies with fewer than five mites were excluded. Mites were examined under 32–50× magnification for damage to mouthparts, legs, or ventral shields, and the proportion of grooming per day ("Grooming per day," Groom/D) was calculated by dividing the total

proportion of damaged mites by the number of sampling days. Pathogen susceptibility was assessed by quantifying viral loads of Black Queen Cell Virus (BQCV) and Deformed Wing Virus type B (DWVB) from pooled worker samples using quantitative PCR with established primers and standardized protocols (Jay D Evans et al., 2013). Viral copy numbers were calculated from standard curves and used as quantitative phenotypes. For all colony-level traits, phenotypes were normalized using ordered quantile normalization implemented in the bestNormalize function (Peterson, 2021b) before downstream analyses.

Genome-wide association analyses [Refer to chapter 3 for more detailed methods]

Genome-wide association analyses were conducted for honey production, hygienic behaviour, grooming per day and two RNA viruses (BQCV, DWVB) using two complementary mixed-model approaches. In both analyses, InDels were treated as single-locus variants and tested individually across the genome. First, locus specific associations between phenotypes and InDel alternative allele frequencies were tested using a linear mixed model (LMM) implemented in GridLMM (Runcie & Crawford, 2019). This approach is well suited for pooled sequencing data, as it directly models allele frequencies at each locus. Population structure among colonies was accounted for by including a genetic relatedness matrix (GRM) as a random effect, constructed from 83,196 unlinked markers obtained during the imputation analysis. Additional shared environmental and management effects were controlled by including apiary and sampling year as a nested random effect. Model fitting was performed using restricted maximum likelihood, and statistical significance was assessed using Wald tests.

To further characterize the genetic architecture of each trait, we applied a Bayesian sparse linear mixed model (BSLMM) implemented in GEMMA (Zhou et al., 2013). This framework allows simultaneous estimation of polygenic background effects and identification of loci with large effects. For this analysis, InDel allele frequencies were converted to consensus genotypes using frequency-based thresholds and formatted for GEMMA input. A centered GRM, derived from the same set of unlinked markers, was used to account for relatedness among colonies. Two independent BSLMM runs were performed for each trait using long Markov chain Monte Carlo chains to ensure convergence.

Candidate InDels associated with phenotypic variation were identified by integrating results from both the LMM and BSLMM analyses. Variants were prioritized if they showed

evidence of large effects in the Bayesian model, including high posterior inclusion probabilities (PIP)and extreme sparse effect sizes relative to the genome-wide distribution. Additional loci with strong Bayesian support were retained even in the absence of strong single-locus significance, allowing detection of variants contributing to trait architecture beyond strict p-value thresholds.

In addition, from BSLMM analyses, we estimated key parameters describing trait architecture. The proportion of phenotypic variance explained by all InDels (PVE) and the proportion attributable to major-effect loci (PGE) were extracted from posterior distributions. Narrow-sense heritability (*h2*) was estimated as the product of these two parameters, providing an estimate of the phenotypic variance explained by large-effect InDels. For comparison with SNP-derived heritability, we performed a sign test across the same five traits to statistically assess whether InDel-based h^2 values were consistently lower than SNP-based h^2.

Comparison of SNP and InDel association signals

To assess whether InDels provide independent signals beyond SNP based associations, we compared the physical genomic locations of SNPs and InDels associated with the same phenotype. Overlap and proximity between variant classes were evaluated to determine whether InDel associations co-localized with SNP signals or identified distinct genomic regions. This comparison allowed us to assess whether InDels primarily validate SNP-based findings or reveal additional, potentially independent contributors to phenotypic variation.

Extraction and functional annotation of candidate genes

The genes associated with candidate InDels were identified using the snpEff annotation output**.** Gene names were extracted from the annotation file, and for each gene, functional information was retrieved from the NCBI Gene database**.** Extracted information was compiled into a summary table, providing a systematic overview of the potential roles of candidate genes in the traits of interest.

Results:

InDel Discovery and Distribution

We discovered a total of 125,190 short InDels that passed our filtering criteria (see Methods) across the honey bees 16 chromosomes. Comparison with the European Variation

Archive (EVA) showed that approximately 70% of our InDels overlapped with previously reported variants, while the remaining ~30% appear novel or population-specific, likely reflecting unique variation in Canadian colonies. Classification of InDels revealed that insertions were slightly more frequent than deletions, a pattern observed across most chromosomes except for LG2, LG9, LG14, and LG15, where deletions predominated (Supplementary Table 6.1). Analysis of variant lengths showed a strong bias toward single-base pair changes, with the majority of InDels being shorter than 5 bp, and a rapid decline in frequency with increasing size (Figure 6.1). These results are consistent with the detection limits of short-read Illumina sequencing (2*150 bp), which can reliably capture short InDels but are unable to resolve longer InDels.

Genome-wide InDel distribution

Genome-wide analysis revealed that InDels are unevenly distributed, with regions of lowest InDel density closely overlapping known centromere positions, suggesting that InDel accumulation is constrained in these low-recombination regions (Figure 6.2a). A notable exception was observed on Chromosome 1, where centromere windows did not coincide with the low-density InDel regions, and more than one low-density window was present, indicating that this chromosome may have unique structural or assembly features that need further investigation. It is important to note that technical limitations related to short-read next-generation sequencing (NGS) data may also contribute to these patterns. The centromeric regions are rich in repetitive DNA and low-complexity sequences, including low-GC regions surrounding AvaI repeat clusters, which are difficult to assemble and align, potentially leading to underrepresentation of variants in these regions (Treangen & Salzberg, 2012).

We used linear mixed-effects models to examine the relationship between variant density and centromere proximity for both InDels and SNPs. Both InDels and SNPs showed a significant positive association with distance from the centromere, indicating that variant density increases further from centromeres. For InDels, the effect of distance was $z = 4.72$ ($p < 0.001$), and for SNPs, $z = 4.35$ ($p < 0.001$), with both effects in the same positive direction. These results demonstrate that, despite differences in variant type, InDels and SNPs exhibit broadly similar distribution patterns relative to centromere proximity across all 16 chromosomes analyzed (Windows analyzed: InDels = 449, SNPs = 450).

To further assess whether SNP and InDel densities exhibit coordinated genomic patterns beyond their shared association with centromere proximity, we studied the correlation between SNP and InDel densities across the same 500 kb windows. Genome-wide analysis revealed a significant positive correlation between SNP and InDel density (Pearson $r = 0.586$, $p = 8.23 \times 10^{-43}$) (Figure 6.2b). Chromosome-specific analyses showed consistently positive correlations across all 16 chromosomes ($r = 0.45$-0.73, all $p < 0.05$). These results indicate that genomic windows with higher SNP density also tend to show higher InDel density.

Functional Annotation

To characterize the potential biological impact of the identified InDels we performed functional annotation on all the InDels using SnpEff. The majority of InDels fell within intronic regions, accounting for 59.3% of insertions and 58.7% of deletions, followed by upstream, intergenic and downstream regions, of insertions and deletions of the respective totals (Figure 6.3a).

Only a small fraction of variants (0.3% in insertions and deletions) mapped to exonic sequences. In total, 210 insertions and 171 deletions were identified in exons. Among insertions, 102 (48.6%) resulted in frameshift mutations (FS), whereas among deletions, 55 (32.2%) were frameshift. Non-frameshift (NFS) events were correspondingly more frequent among deletions (116 of 171, 67.8%) than insertions (108 of 210, 51.4%). Insertions also contributed slightly more predicted high-impact variants compared to deletions, with 103 (49.0%) insertions and 55 (32.2%) deletions classified as high-impact. High-impact variants were defined as those predicted to disrupt protein coding, including frameshift mutations, stop-gained, and start-lost events.

Following SnpEff annotation, we used chi-square tests to assess whether insertions and deletions differed in their genomic distributions and biotype. Insertions and deletions showed largely similar distributions across biotypes (protein-coding, lncRNA, and others), with no significant differences detected ($\chi^2 = 7.51$, $p = 0.276$; Cramer's $V = 0.01$). However, their distribution across genomic location categories differed significantly ($\chi^2 = 252.81$, $p < 0.001$; Cramer's $V = 0.04$), although the effect size was small. Insertions occurred slightly more frequently than expected in intergenic regions, whereas deletions were relatively more abundant within genic features, including introns and upstream/downstream of gene. A similar pattern of

subtle but consistent differences was observed when comparing frameshift (FS) and non-frameshift (NFS) variants ($\chi^2 = 12.83$, $p = 0.00034$; Cramer's V = 0.01). Insertions were modestly overrepresented among frameshift variants relative to deletions (observed: 102 insertions vs 55 deletions), whereas deletions were slightly more frequent among non-frameshift variants. Although the effect sizes were small, these results indicate consistent directional differences in how insertions and deletions are distributed across genomic features and functional consequence categories.

Comparisons between SNPs and InDels revealed significant differences in functional annotation categories ($\chi^2 = 221{,}366.43$, $p < 0.001$; Cramer's V = 0.34; Fig. 3b). Relative to expectation, InDels occurred less frequently in introns, upstream and downstream of gene but more frequently in intergenic regions. In contrast, SNPs were observed more frequently than expected in introns and upstream of genes, and less frequently in intergenic regions. Biotype distributions also differed significantly, although the effect size was small ($\chi^2 = 16.54$, $p = 2.56 \times 10^{-4}$; Cramer's V = 0.003), with InDels occurring more frequently than expected in *LncRNA* regions. Together, these results indicate broadly similar genomic distributions between variant classes, with modest but statistically robust differences in specific annotation and biotype categories (Figure 6.3b).

InDel association patterns relative to SNP-based GWAS

We performed genome-wide association analyses of InDels for five heritable colony-level traits: Hygiene, Honey production, Black Queen Cell Virus (BQCV)*, Deformed Wing Virus B (*DWVB*)*, and Grooming (quantified as a per day measurement). These traits were measured in the same colonies from which InDel and SNP variants were identified. Across these traits, 92 candidates InDels were retained after stringent filtering, representing 0.073% of the 125,190 variants tested. None reached genome-wide significance, and most associations were observed at suggestive thresholds (Supplemental Tables 6.2–6.6). For comparison, SNP-based GWAS conducted on the same dataset (Chapter 3) identified 514 suggestive SNPs out of 1,742,884 (0.029%) of tested markers. A chi-square test of independence confirmed that the proportion of suggestive associations differed significantly between variant classes (χ^2 (1) = 68.37, $p = 1.36 \times 10^{-16}$). Although SNPs yielded a greater absolute number of loci, InDels showed a significantly higher proportion of suggestive associations relative to the number tested, indicating proportional enrichment of InDels in the present dataset.

We then estimated narrow-sense heritability (h^2) for the five traits using InDel markers, and the resulting values were low across all phenotypes (4–13%), indicating a limited contribution of InDels to phenotypic variance (Table 6.1). For each trait, InDel-based h^2 estimates were lower than those obtained from SNP-based analyses of the same dataset (23% to 66%) (Supplementary Table 6.7). A sign test confirmed that this consistent directional pattern was unlikely to occur by chance alone (binomial test, $p = 0.031$), indicating that SNP markers captured significantly greater additive genetic variance than InDels across all traits (Figure 6.4).

Limited Overlap Between InDel Associations and SNP GWAS Signals

To assess whether candidate InDels might reflect the same underlying genetic signals as SNP-based GWAS hits, we examined their physical proximity to significant SNPs identified for the same traits using the same dataset and analysis pipeline (Chapter 3). Only a small number of candidate InDels were located near (200–2,000 bp) significant SNPs. For these specific cases, we cannot determine if the causal mutations driving the phenotypic associations are SNPs or InDels because of linkage disequilibrium. In contrast, most candidate InDels were located at substantially greater distances, ranging from hundreds of kilobases to several megabases from the nearest significant SNP, a distance that should theoretically not be associated with linkage disequilibrium given the honey bee's high recombination rate of 19-20 cM/Mb (Beye et al., 2006; DeLory et al., 2020; Kent et al., 2012). The limited spatial overlap between candidate InDels and SNPs suggests that InDels and SNPs are capturing the effects of different causal loci associated with colony-level phenotypes.

Functional annotation of trait-associated InDels

The functional annotation of InDels associated with social immunity traits, hygienic behavior and grooming (mites dropped per day) were predominantly located near genes involved in neuronal signaling, cytoskeletal organization, intracellular transport, and transcriptional regulation. These functional categories align with the sensory processing, motor coordination, and behaviors required for colony level defense mechanisms (Supplementary Table 6.2 and Supplementary Table 6.3). In contrast, InDels associated with viral traits (BQCV and DWVB) mapped to genes implicated in signal transduction, membrane transport, proteolysis, and transcriptional control. Many candidates are involved in cellular regulation and stress-response pathways, suggesting that genetic variation underlying viral traits may reflect differences in host

tolerance or modulation of infection outcomes rather than direct antiviral resistance (Supplementary Table 6.4 and Supplementary Table 6.5) Finally, InDels associated with honey production formed a smaller, distinct set and mapped to genes involved in intracellular signaling and metabolic regulation, consistent with honey yield representing a performance-related colony trait (Supplementary Table 6.6).

Discussion:

Our analysis of 125,190 InDels provides a genome-wide overview of short structural variation in the honey bee genome. Consistent with patterns reported across insect and other eukaryotic genomes, the majority of detected InDels were short (<5 bp) and predominantly located in non-coding regions, including introns and intergenic regions, with only 0.3% mapping to exons. In total, 210 insertions and 171 deletions were identified in coding regions. Notably, nearly half of exonic insertions (48.6%) and one-third of exonic deletions (32.2%) resulted in frameshift mutations, and a similar proportion were classified as high-impact variants predicted to disrupt protein coding (including frameshift, stop-gained, and start-lost events). The fact that coding InDels are rare overall, yet frequently predicted to have disruptive consequences when present, is consistent with strong purifying selection limiting the persistence of protein-altering mutations in populations (Keightley & Lynch, 2003; Montgomery et al., 2013). The relatively low abundance of such variants in our dataset is consistent with this expectation. Finally, it is important to note that the paired-end 150 bp short-read sequencing strategy used here is biased toward the detection of short INDELs; larger or more complex structural variants, particularly in repetitive or low-complexity regions, are likely underrepresented (Treangen & Salzberg, 2012).

The uneven distribution of InDels across the *Apis mellifera* genome, with a pronounced reduction in the number of InDels in centromeric regions, aligns with observations from other insects in which variation is reduced near centromeres and increases toward chromosome arms. For example, in *Drosophila melanogaster*, reduced nucleotide diversity and structural variation near centromeres have been linked to suppressed recombination and linked selection (Begun & Aquadro, 1992). Similar centromere-associated reductions in variation have also been reported in mosquitoes and ants (Fontaine et al., 2017; Nouhaud, 2018), indicating that centromeric constraints on genetic diversity are a common genomic feature across diverse taxa. The overlap between low InDel density windows and independently defined centromere positions in honey

bee supports this pattern of constrained variation in low-recombination regions. Although most chromosomes exhibited this pattern, Chromosome 1 showed multiple low-density InDel windows that did not align precisely with annotated centromere locations. This deviation may reflect a unique structural organization or residual uncertainty in centromere annotation for this chromosome, requiring further investigation. We should also note that technical limitations of short-read sequencing also influence the interpretation of these patterns. Centromeric regions are enriched for repetitive and low-complexity DNA, including AT-rich AvaI repeat clusters, which are difficult to assemble and align using short reads, often leading to underrepresentation of variants in these regions (Treangen & Salzberg, 2012). Therefore, the apparent low density of INDELs and SNPs near centromeres likely reflects a combination of biological constraint and methodological bias.

Additionally, quantitative analysis using linear mixed-effects models reinforced the association between variant density and centromere proximity. Both InDels and SNPs showed significant positive relationships with distance from the centromere, indicating that overall sequence variation increases with centromere distance across the genome. Genome-wide, SNP and InDel densities were positively correlated ($r = 0.586$, $p = 8.23 \times 10^{-43}$), with chromosome-level correlations ranging from 0.45 to 0.73 (Supplementary Table 6.1). This indicates that genomic regions enriched for SNPs also tend to harbor elevated InDel variation. Both variant types showed higher density in regions distal to centromeres, highlighting a shared genomic distribution pattern. The consistency of this trend across variant classes suggests that the same factors that broadly influence the mutation rate for SNPs also influence the mutation rate for InDels.

Functional annotation revealed clearer contrasts between InDels and SNPs than their overall genomic distributions would suggest. Although both variant types were widely distributed across the genome, InDels were relatively enriched in intergenic and lncRNA regions, whereas SNPs showed stronger enrichment in intronic and upstream regulatory regions. These subtle yet statistically robust differences may be driven by differences in purifying selection acting on Indels and SNPs. Exonic InDels were rare (~0.3%), yet their predicted functional impact was disproportionately severe. Nearly half of exonic insertions and one-third of deletions resulted in frameshift mutations, and insertions contributed a higher proportion of high-impact

variants overall. Because frameshift mutations alter the downstream reading frame, they are far more likely than SNPs to produce truncated or nonfunctional proteins. In contrast, most SNPs typically result in synonymous or missense substitutions with comparatively localized effects. Together, these findings emphasize a qualitative difference between variant classes: SNPs are more abundant and broadly tolerated, whereas InDels though less frequent in coding regions, are more likely to exert substantial functional consequences when present. This contrast supports the hypothesis that differential selective pressures shape their genomic distribution and functional impact.

While we found that InDels can contribute to phenotypic variation in honey bees, their effects were often lower than that of SNPs. Heritability estimates based on BSLMM for InDels were significantly lower than those derived from SNPs across five colony-level traits. Although the absolute number of candidate loci identified by InDels was smaller than SNPs, the proportion of suggestive associations relative to the total number of variants tested was slightly higher for InDels (0.073% vs 0.029%), suggesting that InDels potentially because of their more disruptive nature are more likely to affect quantitative traits on a per mutation basis. However, InDels are far fewer than SNPs, which is perhaps why they generally explain less of the phenotypic variance relative to SNPs; if the majority of genetic mutations are SNPs, then it would be expected that the majority of genetic contributions to a phenotypic trait are driven by SNPs.

Functional annotation of genes associated with candidate InDels revealed convergent biological themes across trait categories. Behavioral and productivity-related traits, including honey yield, were enriched for genes involved in neurosensory function and signal integration, such as *hwt*, consistent with the central role of neural processing and task allocation in shaping colony-level performance (Whitfield et al., 2006). Social immunity traits, including hygienic behavior and grooming, were associated with genes involved in neuronal signaling, intracellular transport, and sensory perception, reinforcing the idea that collective defense behaviors are governed primarily by regulatory and neural mechanisms rather than immune effectors alone (Wilson-Rich et al., 2008). Disease-related traits highlighted trade-offs between immunity and broader physiological processes, with candidate genes such as *Ubx* and *osp* linking viral responses to developmental and reproductive pathways, consistent with pleiotropic limitations on immune evolution (Schwenke et al., 2016). Although individual InDels generally exhibited

modest effect sizes, this pattern is consistent with the polygenic architecture of complex traits and underscores the cumulative contribution of many small-effect regulatory variants.

Conclusion:

Overall, our study represents one of the first large scale genome-wide analysis of InDel variation in *Apis mellifera* and demonstrates that structural variation is a significant and complementary contributor to colony-level phenotypes. The results provide a framework for understanding how regulatory structural changes shape honey bee health, social immunity, behavior, and productivity**.** While SNP-based GWAS have been instrumental in identifying trait-associated loci, InDels capture independent genetic sources for phenotypic variation, emphasizing the importance of studying complex traits in honey bees.

Table 6.1: The table presents the number of significant candidate InDels and narrow-sense heritability (*h2*). The candidates were identified after stringent filtering, along with the mean Proportion of Variance Explained (PVE) by the BSLMM model and mean Proportion of Genetic Effect (PGE) attributable to sparse-effect InDels.

Phenotype	**# of Significant InDels**	**Mean PVE**	**Mean PGE**	h^2
GroomingPerDay	13	0.97	0.10	0.097
DWVB	16	0.99	0.05	0.049
BQCV	27	0.99	0.13	0.128
Honey	3	0.99	0.04	0.039
Hygiene	33	0.86	0.09	0.077

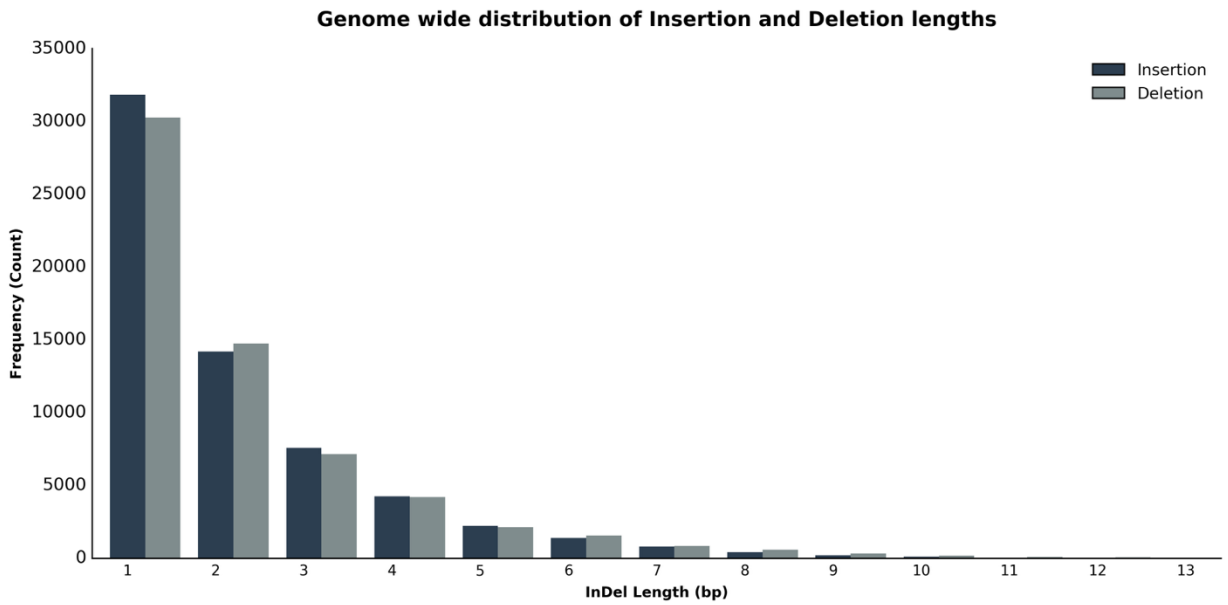


Figure 6.1. Genome-wide distribution of InDel lengths across the honey bee's 16 chromosomes. A bias toward short variants is evident across all linkage groups, with single base-pair (1 bp) InDels constituting the most abundant class. The vast majority of InDels fall within the 1-5 bp range. Insertions and Deletions are distinguished by color, with insertions (blue) and deletions (grey).

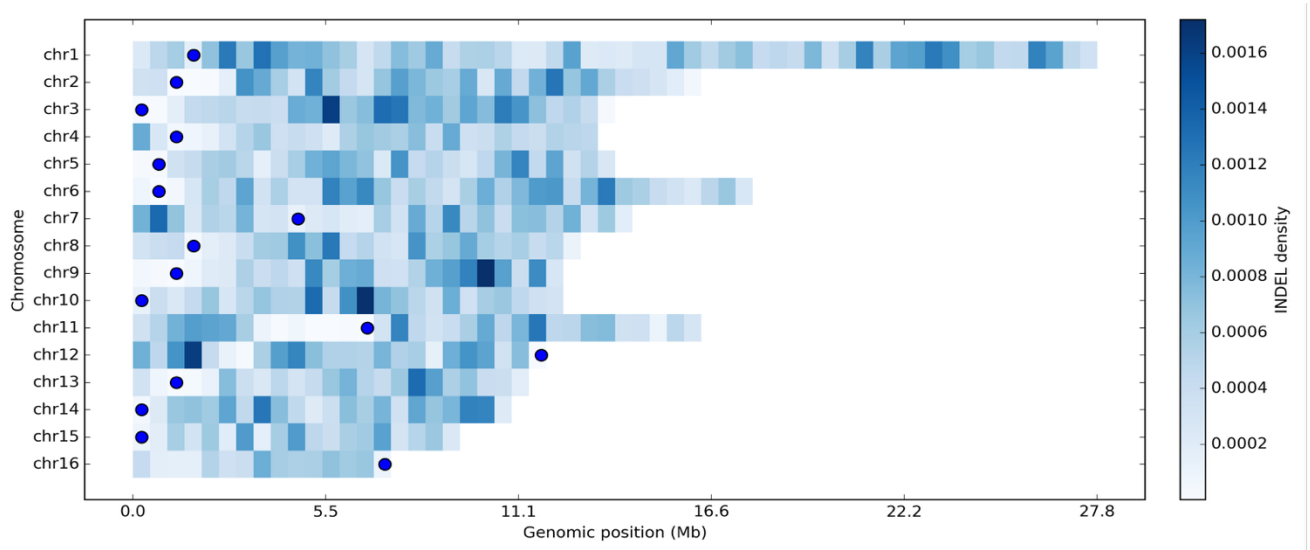


Figure 6.2a. The density of InDels across the honey bee genome. InDel density was calculated using a 500 kb non-overlapping sliding window. The color scale represents the number of variants per window, ranging from low (light blue) to high (dark blue) density. A consistent

reduction in variant density is observed in probable centromeric regions (small circles) across most chromosomes, mirroring known patterns in the honey bee genome.

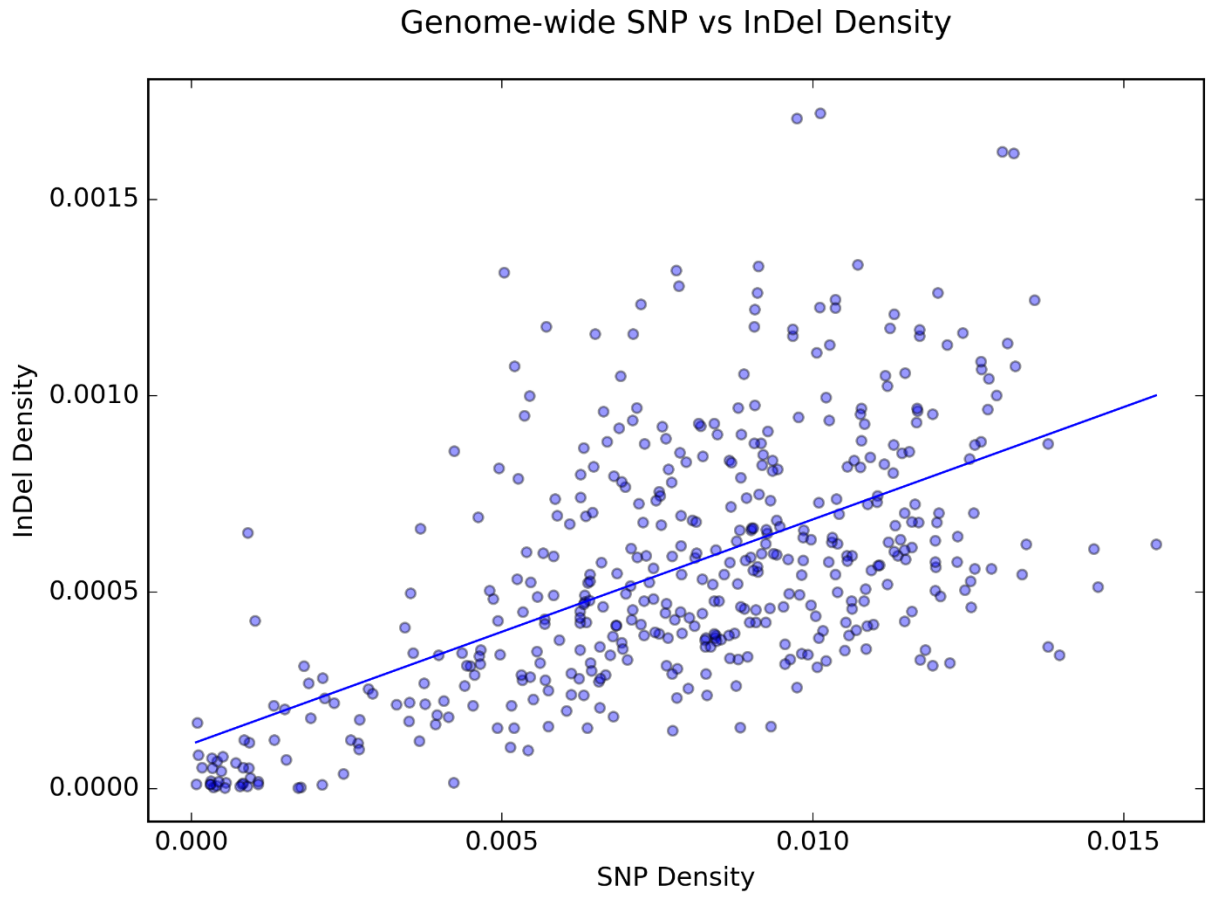


Figure 6.2b. SNP and InDel density are highly correlated across the honey bee genome. Scatter plot illustrating the genome-wide correlation between SNP and InDel densities within 500kb genomic windows. The highly significant positive correlation ($r = 0.586$, $p = 8.23 \times 10^{-43}$) demonstrates suggests that the evolutionary processes that generate and maintain diversity of SNPs are relatively similar to those generating and maintain InDels.

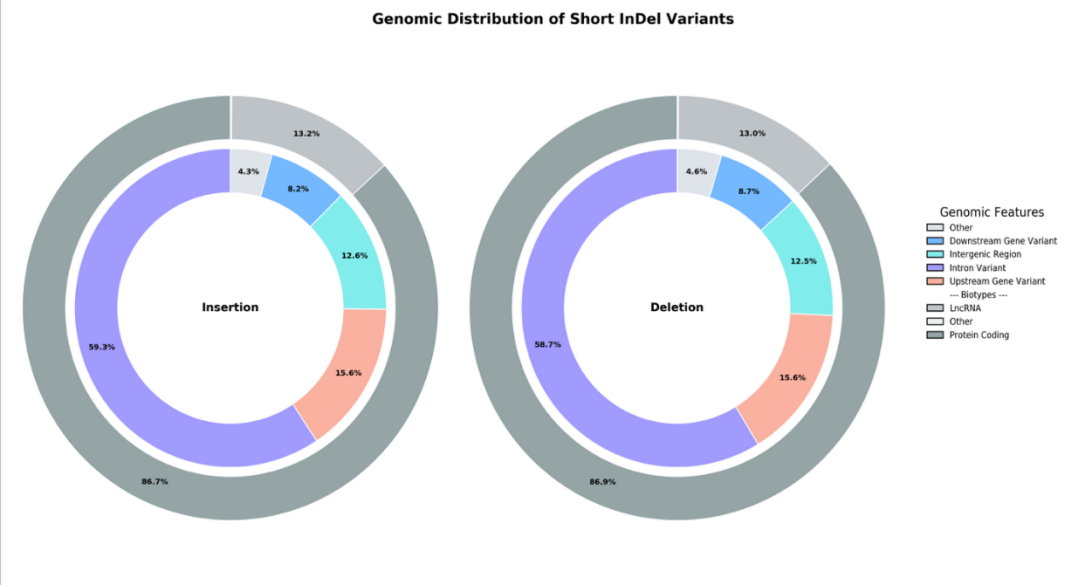


Figure 6.3a. Distribution of insertion and deletion variants across genomic regions and biotypes. Double-ring donut plots showing the proportional distribution of Insertion (left) and Deletion (right) variants. The inner ring represents genomic location, while the outer ring represents biotype classifications. Categories contributing less than 5% of the total variants were grouped into Other.

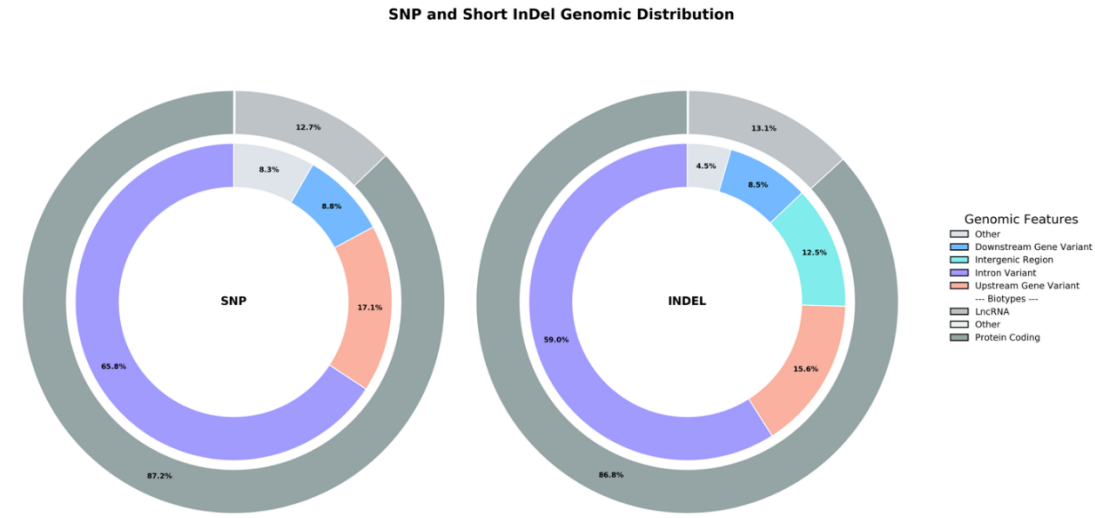


Figure 6.3b. Distribution of SNP and InDel variants across genomic regions and transcript biotypes. Double-ring donut plots showing the proportional distribution of SNP (left) and InDel

(right) variants. The inner ring represents genomic location, while the outer ring represents transcript biotype classifications. Categories contributing less than 5% of the total variants were grouped into Other.

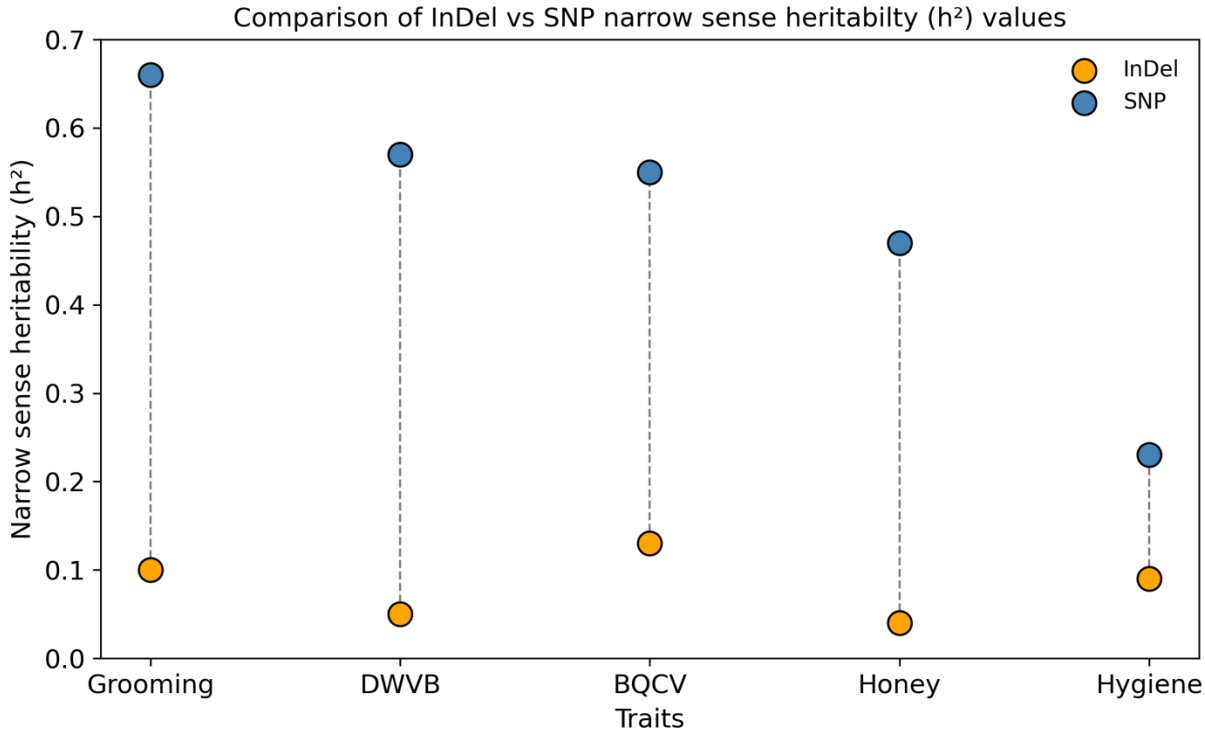


Figure 6.4. Total genetic effects associated with InDel loci are often much smaller than those associated with SNP loci. Comparison of total heritability (h^2) explained by InDels and SNPs across several honey bee traits. Each trait is represented on the x-axis, with heritability (h^2) on the y-axis. Orange and blue dots correspond to InDel and SNP-derived heritability estimates, respectively. Lower heritability estimates for InDel loci is likely attributed to the small number of InDel mutations relative to SNPs.

Supplementary Data:

Supplemental Table 6.1. Distribution of Insertions and Deletions across Honey Bee Linkage Groups (LG).

Chromosome Number	Linkage Group	Insertion	Deletion	Total	Dominant Variant Type
CM009931.2	1	8874	8588	17462	Insertion

CM009932.2	2	4476	4492	8968	Deletion
CM009933.2	3	4651	4568	9219	Insertion
CM009934.2	4	3100	3070	6170	Insertion
CM009935.2	5	3737	3624	7361	Insertion
CM009936.2	6	5139	5097	10236	Insertion
CM009937.2	7	3960	3823	7783	Insertion
CM009938.2	8	3484	3379	6863	Insertion
CM009939.2	9	3607	3648	7255	Deletion
CM009940.2	10	3827	3611	7438	Insertion
CM009941.2	11	4027	3969	7996	Insertion
CM009942.2	12	3907	3906	7813	Insertion
CM009943.2	13	2706	2669	5375	Insertion
CM009944.2	14	3475	3581	7056	Deletion
CM009945.2	15	2374	2386	4760	Deletion
CM009946.2	16	1718	1717	3435	Insertion

Supplemental Table 6.2. InDels associated with hygienic behavior in honey bees. The table lists genomic positions and linkage groups of InDels significantly associated with hygienic behavior, along with effect size estimates from Bayesian Sparse Linear Mixed Models (BSLMM) and GridLMM, corresponding GridLMM p-values, predicted genomic context, and annotated genes based on the *Apis mellifera* reference genome.

Chr and position of Hygiene InDels	**Linkage Group**	**SparseEffect-BSLMM**	**Effect Size-BSLMM**	**Effect Size-GridLMM**	**pvalue-GridLMM**	**Indel Location**	**NCBI Gene ID**	**Gene Name**
CM009932.2_14863134	LG2	0.021	0.003	-0.505	0.001	intron_variant	LOC726461	SH3 and multiple ankyrin repeat domains prosap(Prosap)
CM009932.2_7521907	LG2	0.100	0.017	-0.541	0.001	downstream_gene_variant	LOC102654085	uncharacterized LOC102654085(LOC102654085)

CM009932.2_8447109	LG2	0.011	0.002	0.495	0.006	intergenic_region	LOC113218598-LOC408702	ncharacterized LOC113218598-SAGA-associated factor 29(Sgf29)

CM009933.2_31 08577	LG 3	0.015	- 0.00 2	0.49 1	0.00 2	intron_variant	LOC408759	protein FAM102B(LOC408759)
CM009933.2_31 12194	LG 3	0.033	- 0.00 5	0.55 5	0.00 0	upstream_gene_variant	LOC107964202	uncharacterized LOC107964202(LOC107964202)
CM009933.2_59 32595	LG 3	0.013	0.00 2	- 0.47 1	0.00 3	intron_variant	LOC410951	RNA binding protein fox-1 homolog 2(LOC410951)
CM009933.2_74 36714	LG 3	0.075	- 0.01 3	0.43 4	0.00 4	intron_variant	LOC725792	MDS1 and EVI1 complex locus protein EVI1-B-like (LOC725792)
CM009935.2_11 090632	LG 5	0.264	0.04 8	- 0.76 9	0.00 0	intron_variant	LOC726101	uncharacterized LOC726101(LOC726101)
CM009935.2_52 33261	LG 5	0.015	- 0.00 2	0.42 2	0.00 6	intron_variant	LOC413021	proteoglycan-like sulfated glycoprotein papilin(Ppn)
CM009936.2_12 800433	LG 6	0.026	- 0.00 4	0.52 5	0.00 1	intron_variant	LOC411155	Teneurin-a transmembrane protein (Ten-a)
CM009936.2_13 121570	LG 6	0.012	- 0.00 2	0.52 5	0.00 1	downstream_gene_variant	LOC100578311	cytidine/uridine monophosphate kinase Dak1(Dak1)
CM009936.2_74 47982	LG 6	0.038	0.00 5	- 0.55 8	0.00 8	intron_variant	LOC100577517	protein unc13 homolog(unc-13)
CM009937.2_10 534830	LG 7	0.290	0.05 3	- 0.32 7	0.03 1	intron_variant	LOC551898	Tetraspanin 26A(Tsp26A)
CM009937.2_14 002381	LG 7	0.025	0.00 4	- 0.50 8	0.00 2	3_prime_UTR_variant	LOC408898	bestrophin family protein (Best2)
CM009938.2_12 562250	LG 8	0.016	0.00 3	- 0.52 7	0.00 3	downstream_gene_variant	LOC413264	yellow b(LOC413264)
CM009938.2_57 40433	LG 8	0.028	0.00 4	- 0.42 9	0.00 5	intron_variant	LOC725870	chaoptin (LOC725870)
CM009940.2_25 13290	LG 10	0.016	0.00 2	- 0.47 7	0.00 2	intergenic_region	LOC100576301-LOC102656656	uncharacterized LOC100576301- Serine-threonine kinase, Hsap SBK1
CM009940.2_32 56974	LG 10	0.014	- 0.00 2	0.58 6	0.00 0	downstream_gene_variant	LOC413270	kelch-like protein 10(LOC413270)

CM009941.2_11074586	LG11	0.012	0.002	-0.471	0.002	upstream_gene_variant	LOC413837	kinesin family member unc-104(unc-104)
CM009941.2_1317992	LG11	0.014	0.002	-0.556	0.003	upstream_gene_variant	LOC408337	bitesize, btsz
CM009941.2_13880605	LG11	0.023	-0.003	0.588	0.004	intron_variant	LOC552619	homeobox protein onecut (LOC552619)
CM009941.2_9369168	LG11	0.076	0.012	-0.519	0.002	5_prime_UTR_variant	LOC412865	collagen XV/XVIII-type protein multiplexin (Mp)
CM009942.2_3669773	LG12	0.013	-0.002	-0.597	0.000	intron_variant	LOC100578102	ankyrin repeat domain 50(LOC100578102)
CM009942.2_6504511	LG12	0.020	0.003	-0.540	0.000	intron_variant	LOC410362	DBH like monooxygenase olf413(olf413)
CM009943.2_5247833	LG13	0.021	0.003	-0.718	0.002	upstream_gene_variant	LOC725354	Dpr group, synapse organization
CM009943.2_6261051	LG13	0.024	-0.003	0.513	0.005	intron_variant	LOC408430	Calcium channel protein beta subunit (Ca-beta)
CM009943.2_8064193	LG13	0.013	-0.002	0.468	0.007	intergenic_region	LOC724914 - LOC100576460	bHLH protein deadpan(dpn)-EF-hand calcium-binding domain-containing protein 1
CM009943.2_8458553	LG13	0.022	0.004	0.661	0.000	intron_variant	LOC102655449	uncharacterized LOC102655449(LOC102655449)
CM009944.2_8058026	LG14	0.017	-0.002	0.676	0.004	3_prime_UTR_variant	LOC408463	uncharacterized LOC408463(LOC408463)
CM009945.2_3962717	LG15	0.011	-0.002	-0.672	0.000	intron_variant	LOC113219262	uncharacterized LOC113219262(LOC113219262)
CM009945.2_3962720	LG15	0.013	-0.002	-0.598	0.000	intron_variant	LOC113219262	uncharacterized LOC113219262(LOC113219262)
CM009945.2_4211605	LG15	0.011	0.002	-0.418	0.007	intron_variant	LOC724847	Ig-like domain-containing protein
CM009945.2_5539410	LG15	0.014	0.002	0.537	0.002	intron_variant	LOC551957	hdc homolog, cell cycle regulator(heca)

Supplemental Table 6.3. InDels associated with Grooming (mites drop per day) in honey bees. The table lists genomic positions and linkage groups of InDels significantly associated with hygienic behavior, along with effect size estimates from Bayesian Sparse Linear Mixed Models

(BSLMM) and GridLMM, corresponding GridLMM p-values, predicted genomic context, and annotated genes based on the *Apis mellifera* reference genome.

Chr and position of miteDropPerDay InDels	Linkage Group	SparseEffect-BSLMM	EffectSize-BSLMM	EffectSize-GridLMM	pvalue-GridLMM	Indel Location	NCBI Gene ID	Gene Name
CM009931.2_4727271	LG1	0.021	-0.005	0.922	0.001	upstream_gene_variant	LOC412511	Kap-alpha3, karyopherin alpha3
CM009932.2_9928221	LG2	0.014	-0.003	0.753	0.001	intron_variant	LOC100578436	uncharacterized LOC100578436
CM009933.2_11930747	LG3	0.012	-0.003	0.981	0.008	upstream_gene_variant	LOC102654417	uncharacterized LOC102654417
CM009934.2_11731007	LG4	0.011	-0.002	0.613	0.009	intron_variant	LOC408787	Tmtc2, Transmembrane O-mannosyltransferase targeting cadherins 2
CM009935.2_5859421	LG5	0.034	0.010	-0.593	0.004	5_prime_UTR_variant	LOC408830	Obsc, Obscurin
CM009937.2_969387	LG7	0.015	0.003	-0.737	0.001	intron_variant	LOC411185	Fanconi anemia group J protein
CM009938.2_11554415	LG8	0.013	-0.002	0.720	0.002	upstream_gene_variant	LOC551016	midasin
CM009938.2_1421734	LG8	0.039	-0.012	0.589	0.006	intron_variant	LOC725220	HGTX, HGTX homeodomain transcription factor
CM009938.2_5047864	LG8	0.083	-0.029	0.772	0.000	3_prime_UTR_variant	LOC408960	sli, slit guidance ligand
CM009939.2_5350028	LG9	0.040	0.010	-0.624	0.006	intron_variant	LOC409024	tmod, tropomodulin
CM009939.2_8416000	LG9	0.013	0.003	-0.882	0.011	intron_variant	LOC107964950	Sf3b6, splicing factor 3B subunit 6
CM009944.2_2910507	LG14	0.154	-0.043	0.752	0.002	upstream_gene_variant	LOC413213	tubulin polyglutamylase TTLL4
CM009945.2_7321886	LG15	0.017	0.004	-0.719	0.000	upstream_gene_variant	LOC100577656	uncharacterized LOC100577656

Supplemental Table 6.4. InDels associated with Black Queen Cell Virus (BQCV) in honey bees. The table lists genomic positions and linkage groups of INDELs significantly associated with hygienic behavior, along with effect size estimates from Bayesian Sparse Linear Mixed

Models (BSLMM) and GridLMM, corresponding GridLMM p-values, predicted genomic context, and annotated genes based on the *Apis mellifera* reference genome.

Chr and position of BQCV InDels	Linkage Group	SparseEffect-BSLMM	Effect Size-BSLMM	Effect Size-GridLMM	pvalue-GridLMM	Indel Location	NCBI Gene ID	Gene Name
CM009931.2_10378480	LG1	0.042	-0.025	1.477	0.005	intron_variant	LOC725189	protein bric-a-brac 2(LOC725189)
CM009931.2_5070772	LG1	0.126	-0.076	0.343	0.610	upstream_gene_variant	LOC408614	Secretogranin_V domain-containing protein 7B2(7B2)
CM009931.2_6685377	LG1	0.036	-0.018	1.765	0.002	intron_variant	LOC100577254	KCNQ potassium channel (KCNQ)
CM009933.2_10027693	LG3	0.029	0.017	-1.692	0.003	intron_variant	LOC724885	Glucose transporter 4 enhancer factor (Glut4EF)
CM009933.2_7526572	LG3	0.046	0.025	-1.563	0.004	intergenic_region	LOC725792-LOC102655392	MDS1 and EVI1 complex locus protein EVI1-B-like, transcription factor hamlet-like
CM009934.2_8678374	LG4	0.025	0.014	-1.549	0.002	intergenic_region	CPR19-LOC725327	cuticular protein 19-nuclear receptor subfamily 2 group E member tailless(tll)
CM009935.2_12260666	LG5	0.034	0.021	-1.511	0.004	downstream_gene_variant	LOC102656631	uncharacterized LOC102656631(LOC102656631)
CM009936.2_13915378	LG6	0.017	0.008	-1.741	0.003	intron_variant	LOC102656661	uncharacterized LOC102656661(LOC102656661)
CM009937.2_10975935	LG7	0.025	0.012	-1.435	0.008	upstream_gene_variant	LOC100576293	zinc finger protein 182(LOC100576293)
CM009937.2_893322	LG7	0.032	0.019	-1.799	0.000	intron_variant	LOC408953	peroxidase(LOC408953)
CM009938.2_4939853	LG8	0.027	-0.015	-4.451	0.006	intron_variant	LOC408960	slit guidance ligand(sli)
CM009938.2_5824435	LG8	0.157	-0.102	1.261	0.014	intergenic_region	LOC725870-LOC411260	chaoptin-RAS oncogene family member Rab27
CM009938.2_6109946	LG8	0.198	-0.126	2.473	0.002	intron_variant	LOC411264	aryl hydrocarbon receptor spineless(ss)

CM009939.2_7568576	LG9	0.026	-0.013	1.756	0.007	upstream_gene_variant	LOC725026	retinol dehydrogenase 10-A(LOC725026)
CM009939.2_7958251	LG9	0.025	-0.012	1.396	0.009	intron_variant	LOC411336	shavenoid
CM009940.2_10418431	LG10	0.026	0.014	-1.390	0.006	intergenic_region	LOC100576772-LOC726017	uncharacterized LOC100576772-lachesin
CM009940.2_4442063	LG10	0.029	0.016	-1.749	0.000	downstream_gene_variant	LOC100577407	uncharacterized LOC100577407(LOC100577407)
CM009940.2_7230743	LG10	0.030	-0.015	1.693	0.004	intron_variant	LOC724333	Syntrophin-like 1(Syn1)
CM009942.2_8317237	LG12	0.126	0.075	-0.554	0.379	downstream_gene_variant	LOC408381	sodium- and chloride-dependent GABA transporter 1(LOC408381)
CM009943.2_10807872	LG13	0.013	0.008	-2.043	0.000	upstream_gene_variant	LOC409715	sodium potassium chloride cotransporter(NKCC)
CM009943.2_9690546	LG13	0.021	0.010	-1.370	0.006	intergenic_region	LOC113219210-LOC113219183	RNA-binding protein 25-like-uncharacterized LOC113219183
CM009944.2_2201045	LG14	0.020	-0.010	2.111	0.003	intron_variant	LOC100577945	furin-like protease 1(LOC100577945)
CM009944.2_2551019	LG14	0.021	-0.012	1.874	0.003	intron_variant	LOC102655131	zinc finger protein 300-like(LOC102655131)
CM009945.2_4118314	LG15	0.042	-0.026	1.515	0.002	intron_variant	LOC113219262	uncharacterized LOC113219262(LOC113219262)
CM009945.2_609023	LG15	0.244	-0.165	0.685	0.332	intron_variant	LOC102653951	ras-related and estrogen-regulated growth inhibitor-like (LOC102653951)
CM009946.2_3693996	LG16	0.060	0.034	-1.449	0.006	intron_variant	Ubx	ultrabithorax(Ubx)
CM009946.2_5187302	LG16	0.116	-0.070	1.617	0.003	upstream_gene_variant	LOC100578793	uncharacterized LOC100578793(LOC100578793)

Supplemental Table 6.5. InDels associated with Deformed Wing Virus B (DWVB) in honey bees. The table lists genomic positions and linkage groups of INDELs significantly associated with hygienic behavior, along with effect size estimates from Bayesian Sparse Linear Mixed

Models (BSLMM) and GridLMM, corresponding GridLMM p-values, predicted genomic context, and annotated genes based on the *Apis mellifera* reference genome.

Chr and position of DWVB InDels	Linkage Group	SparseEffect-BSLMM	EffectSize-BSLMM	EffectSize-GridLMM	pvalue-GridLMM	Indel Location	NCBI Gene ID	Gene Name
CM009931.2_4295509	LG1	0.020	0.002	-0.337	0.000	intron_variant	LOC408618	Fur2, furin-like protease 2
CM009931.2_5097013	LG1	0.015	0.002	-0.287	0.008	upstream_gene_variant	LOC410751	phospholipase A1 member A
CM009932.2_11413712	LG2	0.029	-0.004	0.335	0.000	intron_variant	LOC410866	LIM domain only protein 3
CM009932.2_11641164	LG2	0.024	-0.003	0.269	0.005	intron_variant	LOC408688	Down syndrome cell adhesion molecule-like protein Dscam2
CM009933.2_5152643	LG3	0.027	-0.003	0.324	0.001	intron_variant	LOC410956	nmo,serine/threonine-protein kinase nemo
CM009933.2_5529447	LG3	0.018	-0.003	-0.308	0.002	intron_variant	osp	myosin phosphatase Rho interacting protein outspread
CM009934.2_3895414	LG4	0.013	-0.001	0.260	0.008	intron_variant	LOC727092	protein nubbin
CM009936.2_2156299	LG6	0.017	-0.002	0.249	0.006	intron_variant	LOC413071	rdgA, retinal degeneration A
CM009938.2_9308117	LG8	0.109	0.018	-0.173	0.060	intron_variant	LOC413568	probable G-protein coupled receptor 158
CM009940.2_698890	LG10	0.011	-0.002	-0.421	0.001	intron_variant	LOC413242	N-deacetylase and N-sulfotransferase sfl
CM009941.2_11177258	LG11	0.041	0.006	-0.252	0.008	downstream_gene_variant	LOC726150	transcription factor Sox-21-B
CM009942.2_10023489	LG12	0.014	0.001	-0.275	0.005	5_prime_UTR_variant	LOC413223	rictor, rapamycin-insensitive companion of Tor

CM009944.2_7993320	LG14	0.013	-0.001	0.284	0.002	intron_variant	LOC726697	coiled-coil domain-containing protein 137
CM009945.2_3225564	LG15	0.137	-0.018	0.199	0.070	intron_variant	LOC552099	homeotic protein empty spiracles
CM009945.2_6509849	LG15	0.017	0.002	-0.343	0.001	intron_variant	LOC410609	disintegrin and metalloproteinase domain-containing protein 10-like
CM009946.2_1583679	LG16	0.029	0.003	-0.291	0.003	intron_variant	LOC413633	sff,BRSK family serine/threonine-protein kinase sugar-free frosting

Supplemental Table 6.6. InDels associated with Honey Production in honey bees.

The table lists genomic positions and linkage groups of INDELs significantly associated with hygienic behavior, along with effect size estimates from Bayesian Sparse Linear Mixed Models (BSLMM) and GridLMM, corresponding GridLMM p-values, predicted genomic context, and annotated genes based on the *Apis mellifera* reference genome.

Chr and position of Honey InDels	**Linkage Group**	**SparseEffect-BSLMM**	**EffectSize-BSLMM**	**EffectSize-GridLMM**	**pvalue-GridLMM**	**Indel Location**	**NCBI Gene ID**	**Gene Name**
CM009936.2_5537497	LG6	0.0155	-0.0026903	0.41868227	0.00219441	intron_variant	LOC550870	hwt, SH2 domain-containing adapter heavyweight
CM009937.2_12982291	LG7	0.0105	-0.0015553	0.51976474	0.00030406	intron_variant	LOC100577159	protein PFC0760c
CM009943.2_970077	LG13	0.01825	-0.0026783	0.41314345	0.00347321	upstream_gene_variant	LOC102653778	uncharacterized LOC102653778

Supplemental Table 6.7: Narrow-sense heritability of SNPs. SNP-based estimates of the proportion of **phenotypic** variance explained (PVE), proportion of genetic variance attributable to

sparse-effect loci (PGE), and narrow-sense heritability *(h^2).* The was derived as the product of the mean PVE and PGE

Phenotype	**Mean PVE**	**Mean PGE**	***h^2***
BQCV	0.99	0.56	0.55
DVWB	0.97	0.59	0.57
Honey Production	0.96	0.49	0.47
Hygienic Behaviour	0.81	0.29	0.23
Grooming per day	0.88	0.75	0.66

Chapter seven: Conclusion and future directions

Tanushree Tiwari

This dissertation represents a major advancement in honey bee genomics and constitutes one of the largest datasets assembled to investigate the genetic architecture of colony-level traits in social insects. Before my research, most studies examining the genetics of colony-level phenotypes were limited in scope, typically focusing on a single behavioural or health-related trait, relatively small samples sizes, or employed lower-resolution experimental designs such as quantitative trait locus (QTL) mapping (Table 7.1). Although these foundational studies were important for advancing the field and providing qualitative knowledge about the genetics of colony level traits, they generally did not lead to gene-specific knowledge, which precluded more detailed inquiries into the genetics and evolution of colony level traits.

Among colony-level traits, social immunity received considerable attention because of its direct importance to colony survival. Hygienic behaviour in honey bees, in which worker bees detect and remove diseased or parasitized brood, became one of the earliest and most intensively studied examples of genetically influenced social immunity. Early QTL mapping studies by Lapidge et al. (2002), involving 90 colonies derived from hygienic and non-hygienic breeding lines, identified several genomic regions associated with hygienic behavior using microsatellite linkage mapping. Building upon this work, Oxley et al. (2010) performed genome-wide linkage analyses in colonies selected for *Varroa*-sensitive hygiene and identified multiple loci associated with brood uncapping and removal behaviour. These studies collectively demonstrated that hygienic behaviour is highly polygenic and influenced by multiple genomic regions distributed across the honey bee genome.

As genomic technologies advanced, subsequent studies improved marker resolution and expanded the molecular understanding of social immunity. Harpur et al. (2014, 2019) applied whole-genome sequencing approaches to compare hygienic and non-hygienic colonies and identified candidate genes associated with olfactory signalling, neuronal communication, sensory perception, and behavioural responsiveness. Similarly, Guarna et al. (2015) integrated behavioural phenotyping with proteomic analyses to identify protein biomarkers predictive of hygienic behaviour and Varroa resistance, thereby connecting colony-level behavioural

phenotypes with underlying molecular pathways. Together, these studies shifted the interpretation of social immunity away from single-gene resistance models toward a more complex framework involving coordinated behavioural, neurological, physiological, and immune-related mechanisms.

Genome-wide association and population genomic studies investigating aggression and defensive behaviour further reinforced the importance of neural and sensory pathways in shaping colony-level phenotypes. Avalos et al. analyzed approximately 470 honey bee colonies representing Africanized and European-derived populations from Puerto Rico and the Americas using genome-wide SNP analyses combined with colony-level behavioural assays. Their study identified loci associated with aggression thresholds, alarm response, sensory perception, and neural signalling. Several candidate genes identified were involved in neurotransmission and neuronal development, further emphasizing the neurogenetic basis of colony defence traits. Complementary transcriptomic studies similarly demonstrated that aggressive colonies exhibit differential expression of genes associated with synaptic signalling, neurodevelopment, and metabolic regulation.

Research outside of honey bees further demonstrated the value of genome-wide approaches for understanding colony-level traits in eusocial insects. In fire ants (*Solenopsis invicta*), Wang et al. (2013) identified a large social chromosome associated with colony social organization, specifically whether colonies exhibit monogyne or polygyne structure. Genome-wide population studies in bumble bees similarly identified loci associated with immunity, environmental adaptation, thermoregulation, and behavioural plasticity (Francisco et al., 2015; Kent et al., 2018). Although large colony-based GWAS are less common in bumble bees than in honey bees, Francisco et al. (2015) demonstrated that social and ecological traits are also governed by highly polygenic genomic architectures involving neural, immune, and stress-response pathways. Comparable genomic analyses in ants (Wang et al., 2013) and termites (Terrapon et al., 2014) further identified genes and regulatory regions associated with caste differentiation, reproductive division of labour, and chemical communication. Collectively, these studies across eusocial insects demonstrated that colony-level traits emerge through complex and interconnected genomic networks involving behaviour, neurobiology, immunity, development, and environmental adaptation rather than through isolated major-effect genes.

However, despite these important advances, the field remained fragmented because most studies focused on individual traits or specific pathogen systems independently. Traits such as hygienic behaviour, aggression, productivity, pathogen resistance, and immune responses were typically investigated separately, meaning that each study captured only one component of the broader biological system underlying colony health. Consequently, although previous studies collectively provided important insights into the genetics of social insects, they often represented only individual pieces of a much larger and interconnected genomic puzzle. As a result, the extent to which behavioural, immune, neurological, and pathogen-related traits share overlapping genetic architecture and coordinated molecular networks remained largely unresolved.

My dissertation addresses these limitations through the development of a large-scale, integrative systems-genomics framework for honey bee health. Specifically, this work involved sampling 1,350 colonies over two years to simultaneously quantify behavioural, productivity, innate immune, and pathogen-related traits. Unlike previous studies that examined individual traits or pathogens independently, this dissertation integrates analyses of multiple viruses, microsporidians, and trypanosomatids alongside behavioural and immune phenotypes within a single genomic framework. By applying GWAS to high-resolution whole-genome sequencing (WGS) data across this cohort, the results demonstrate that colony-level phenotypes are not isolated but instead emerge through highly interconnected, polygenic, and pleiotropic networks. Furthermore, integrating genomic variation with colony phenotypes and immune protein abundance establishes a systems-level framework in which behaviour, immunity, productivity, and pathogen resistance are treated as biologically interconnected components of overall colony health, evident from results in Chapter 3.

A major advancement of this dissertation is the extension of genomic association analyses into functional molecular phenotypes through protein quantitative trait locus (pQTL) analysis. To the best of our knowledge, this represents the first large-scale application of immune protein abundance as a quantitative trait in honey bees. Whereas previous GWAS studies primarily focused on DNA-level associations, this work directly links genomic variation with variation in immune protein abundance, enabling the identification of both *cis*- and *trans*-regulatory loci controlling innate immune responses. As demonstrated in Chapter 4, several genomic regions emerged as regulatory hubs influencing multiple immune-related proteins

simultaneously, suggesting the presence of coordinated regulatory networks underlying honey bee immunity. By directly connecting genomic variants with functional molecular outcomes, these analyses help bridge the gap between genotype and biological mechanism.

Beyond SNP-based analyses, this dissertation also expands the genomic framework by demonstrating that insertions and deletions (indels) are important and contribute to the understanding of association across behavioural, immune, and pathogen-related traits. Previous honey bee GWAS studies relied almost exclusively on SNP markers despite increasing evidence from other organisms that indels can substantially influence gene regulation and protein function. To the best of our knowledge, this dissertation represents the first large-scale investigation of indels associated with colony-level traits in honey bees. As shown in Chapter 6, many associated indels were identified within regulatory regions, untranslated regions, and candidate immune- and neurobiology-related loci, suggesting potential impacts on transcriptional regulation, neural signalling, protein structure, and downstream molecular pathways. By integrating both SNPs and indels within a unified analytical framework, this work provides a substantially more comprehensive representation of functional genomic variation underlying colony health.

Together, this dissertation reframes honey bee genomics from a collection of largely independent trait-mapping studies into an integrated systems-level model of colony health. By combining large-scale colony phenotyping, whole-genome sequencing, pQTL analysis, pathogen profiling, and structural variant discovery across one of the largest social insect datasets assembled to date, this work substantially advances our understanding of the genetic architecture underlying colony-level traits. More importantly, the results demonstrate that behavioural, immune, neurological, and pathogen-related phenotypes are governed through overlapping genetic and molecular networks rather than independent biological pathways. Ultimately, this dissertation provides a high-resolution framework for future functional validation, genomic selection, and selective breeding strategies aimed at improving honey bee health, disease resistance, environmental resilience, and productivity, while also advancing the broader understanding of social insect biology at the superorganismal level.

Table 7.1. From single-trait QTL/GWAS studies to multi-layer systems genomics of honey bee health.

This table contrasts landmark genetic studies of honey bee behavioral, productivity, and disease-related traits with the integrated framework developed in this dissertation. While previous studies primarily identified trait-specific QTLs and candidate genes, the present work expands this view by resolving polygenic, pleiotropic, and multi-omic (SNP, pQTL, and indel) architectures underlying colony health.

Trait/category	Peer-reviewed studies	Sample / approach	Traits studied	Major findings (QTL/GWAS outcome)	Key candidate genes / mechanisms	Current findings from the Dissertation
Hygienic behavior	Lapidge et al. 2002; Oxley et al. 2010; Harpur et al. 2019	Linkage mapping + early GWAS (~100–300 colonies)	Hygienic removal, disease resistance	QTLs on Chr 2, 6, 9, 13; later refined SNP-level signals	neurexin-1 (AmNrx-1), Dop3, GABA pathway genes	Trait is highly polygenic and pleiotropic, linked to neuronal and synaptic regulation; shared signals across multiple social immunity traits
Varroa Sensitive Hygiene (VSH)	Tsuruda et al. 2012; Spötter et al. 2012, 2016; Mondet et al. 2015, 2021	Targeted QTL + GWAS (~100–200 colonies)	Uncapping, recapping, mite removal	Major QTL on Chr 9; additional suggestive loci on Chr 1	Dop3, norpA, sensory odor response genes	NA
Grooming behavior	Guzmán-Novoa et al. 2012; Boutin et al. 2015	Behavioral QTL + colony assays	Grooming intensity, mite drop	Moderate-effect loci associated with grooming response variability	CYP9Q3, neurexin-1	Grooming is part of a distributed behavioral immunity network, strongly pleiotropic with hygiene
Aggression / stinging	Hunt et al. 1998; Arechavaleta-Velasco et al. 2003, 2012; Avalos et al. 2017, 2020	QTL + population genomics	Stinging, guarding, aggression thresholds	Sting-1 (Chr 14), Sting-2 (Chr 11), plus soft sweeps in natural populations	serotonin receptor (AmHtr2), dpr4, GABA pathways	Aggression is continuously distributed and polygenic, tightly linked to neural regulation and colony-level allele frequencies
Honey production / foraging	Hunt et al. 1995; Page et al. 2000	QTL mapping (pollen hoarding model)	Honey/pollen storage, foraging behavior	pln1, pln2, pln3 QTL system defining foraging syndrome	Amfor (foraging gene), vitellogenin	Honey production is part of a behavioral network, genetically correlated with social immunity and other traits
Varroa / mite resistance	Arechavaleta-Velasco et al. 2012; Mondet et al. 2020	QTL + behavioral phenotyping	Mite drop, mite reproduction suppression	groom-1 locus (Chr 5), MNR-linked traits	Ataxin, sensory/olfactory genes	Resistance emerges from multi-trait network rather than single loci and single trait
Viral / pathogen resistance	Huang et al. 2012; Khongphinitbunjong et al. 2015; Schwarz & Evans 2013	Linkage + transcriptomics	DWV, Nosema, gut pathogens	Broad immune regions (Chr 7, 12); AMP activation patterns	defensin, hymenoptaecin, PPO, ABC transporters	Pathogen resistance is controlled by coordinated immune-protein networks (SI and II)
Nosema / gut pathogens	Huang et al. 2012; Osterman et al. 2024	QTL + infection assays	Spore load, survival	Resistance loci on Chr 7 & 12	Abcb1, immune effector genes	Pathogen resistance is controlled by coordinated immune-protein networks (SI and II)

References:

Adnane, M., Kelly, P., Chapwanya, A., Meade, K. G., & O'Farrelly, C. (2018). Improved detection of biomarkers in cervico-vaginal mucus (CVM) from postpartum cattle. *BMC veterinary research*, *14*(1), 1-6.

Aggarwal, K., & Silverman, N. (2008). Positive and negative regulation of the Drosophila immune response. *BMB reports*, *41*(4), 267-277.

Aizen, M. A., & Harder, L. D. (2009). The global stock of domesticated honey bees is growing slower than agricultural demand for pollination. *Current biology*, *19*(11), 915-918.

Ajawatanawong, P., & Baldauf, S. L. (2013). Evolution of protein indels in plants, animals and fungi. *BMC Evolutionary Biology*, *13*(1), 140.

Albert, F. W., & Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, *16*(4), 197-212.

Allsopp, M., & Crewe, R. (1993). The cape honey bee as a Trojan horse rather than the hordes of Genghis Khan. *American bee journal (USA)*.

Arechavaleta-Velasco, M. E., Alcala-Escamilla, K., Robles-Rios, C., Tsuruda, J. M., & Hunt, G. J. (2012). Fine-scale linkage mapping reveals a small set of candidate genes influencing honey bee grooming behavior in response to Varroa mites. *PLoS One*, *7*(11), e47269.

Aumeier, P. (2001). Bioassay for grooming effectiveness towards Varroa destructor mites in Africanized and Carniolan honey bees. *Apidologie*, *32*(1), 81-90.

Avalos, A., Fang, M., Pan, H., Ramirez Lluch, A., Lipka, A. E., Zhao, S. D., Giray, T., Robinson, G. E., Zhang, G., & Hudson, M. E. (2020). Genomic regions influencing aggressive behavior in honey bees are defined by colony allele frequencies. *Proceedings of the National Academy of Sciences*, *117*(29), 17135-17141.

Avalos, A., Fang, M., Pan, H., Ramirez Lluch, A., Lipka, A. E., Zhao, S. D., Giray, T., Robinson, G. E., Zhang, G., & Hudson, M. E. (2020). Genomic regions influencing aggressive behavior in honey bees are defined by colony allele frequencies. *Proc Natl Acad Sci U S A*, *117*(29), 17135-17141.

Barribeau, S. M., Sadd, B. M., Du Plessis, L., Brown, M. J., Buechel, S. D., Cappelle, K., Carolan, J. C., Christiaens, O., Colgan, T. J., & Erler, S. (2015). A depauperate immune repertoire precedes evolution of sociality in bees. *Genome biology*, *16*(1), 83.

Batra, S. (1966). Nests and social behavior of halictine bees of India (Hymenoptera: Halictidae). *Indian Journal of Entomology*, *28*(3), 375-393.

Beggs, K. T., Hamilton, I. S., Kurshan, P. T., Mustard, J. A., & Mercer, A. R. (2005). Characterization of a D2-like dopamine receptor (AmDOP3) in honey bee, Apis mellifera. *Insect biochemistry and molecular biology*, *35*(8), 873-882.

Begun, D. J., & Aquadro, C. F. (1992). Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. *Nature*, *356*(6369), 519-520.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289-300.

Beye, M. (2004). The dice of fate: the csd gene and how its allelic composition regulates sexual development in the honey bee, Apis mellifera. *Bioessays*, *26*(10), 1131-1139.

Beye, M., Gattermeier, I., Hasselmann, M., Gempe, T., Schioett, M., Baines, J. F., Schlipalius, D., Mougel, F., Emore, C., & Rueppell, O. (2006). Exceptionally high levels of recombination across the honey bee genome. *Genome research*, *16*(11), 1339-1344.

Beye, M., Hasselmann, M., Fondrk, M. K., Page Jr, R. E., & Omholt, S. W. (2003). The gene csd is the primary signal for sexual development in the honeybee and encodes an SR-type protein. *Cell*, *114*(4), 419-429.

Bienefeld, K., & Pirchner, F. (1990). Heritabilities for several colony traits in the honeybee (Apis mellifera carnica). *Apidologie*, *21*(3), 175-183.

Biswas, S., Reinhard, J., Oakeshott, J., Russell, R., Srinivasan, M. V., & Claudianos, C. (2010). Sensory regulation of neuroligins and neurexin I in the honeybee brain. *PLoS One*, *5*(2), e9133.

Biswas, S., Russell, R. J., Jackson, C. J., Vidovic, M., Ganeshina, O., Oakeshott, J. G., & Claudianos, C. (2008). Bridging the synaptic gap: neuroligins and neurexin I in *Apis mellifera*. *PloS one*, *3*(10), e3542.

Bolger, A. M., Lohse, M., & Usadel, B. (2014a). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, btu170.

Bolger, A. M., Lohse, M., & Usadel, B. (2014b). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114-2120.

Borba, R. S., Hoover, S. E., Currie, R. W., Giovenazzo, P., Guarna, M. M., Foster, L. J., Zayed, A., & Pernal, S. F. (2022). Phenomic analysis of the honey bee pathogen-web and its dynamics on colony productivity, health and social immunity behaviors. *PloS one*, *17*(1), e0263273.

Boutros, M., Agaisse, H., & Perrimon, N. (2002). Sequential activation of signaling pathways during innate immune responses in Drosophila. *Developmental cell*, *3*(5), 711-722.

Brandt, A., Gorenflo, A., Siede, R., Meixner, M., & Büchler, R. (2016). The neonicotinoids thiacloprid, imidacloprid, and clothianidin affect the immunocompetence of honey bees (Apis mellifera L.). *Journal of Insect Physiology*, *86*, 40-47.

Breed, M. D., Diaz, P. H., & Lucero, K. D. (2004). Olfactory information processing in honeybee, Apis mellifera, nestmate recognition. *Animal behaviour*, *68*(4), 921-928.

Breed, M. D., Robinson, G. E., & Page, R. E. (1990). Division of labor during honey bee colony defense. *Behavioral Ecology and Sociobiology*, *27*(6), 395-401.

Bresadola, L., Caseys, C., Castiglione, S., Buerkle, C. A., Wegmann, D., & Lexer, C. (2019). Admixture mapping in interspecific Populus hybrids identifies classes of genomic architectures for phytochemical, morphological and growth traits. *New Phytologist*, *223*(4), 2076-2089.

Brown, J., Pirrung, M., & McCue, L. A. (2017). FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. *Bioinformatics*, *33*(19), 3137-3139.

Bruckner, S., Steinhauer, N., Rennich, K., Aurell, S., Caron, D., Ellis, J., Fauvel, A., Kulhanek, K., Nelson, K., & Rangel, J. (2018). Honey bee colony losses 2017–2018: preliminary results. *Bee Informed Partnership*.

Bruckner, S. S., N; Rennich, K; Aurell, SD; Caron, DM; Ellis, JD; Fauvel, AM;, Kulhanek, K. N., KC; Rangel, J; Rose, R; Sagili, R; Slater, GP; Snyder, R; Thoms,, CA; Wilkes, J. W., ME; vanEngelsdorp, D; Williams, GR; for the Bee Informed, preliminary, P. U. S. h. b. c.

l. , r., & 2019)., p.-r. A. M. (2019). United States honey bee colony losses 2017-2018: preliminary results.

Brutscher, L. M., Daughenbaugh, K. F., & Flenniken, M. L. (2015). Antiviral defense mechanisms in honey bees. *Current Opinion in Insect Science*, *10*, 71-82.

Brutscher, L. M., Daughenbaugh, K. F., & Flenniken, M. L. (2017). Virus and dsRNA-triggered transcriptional responses reveal key components of honey bee antiviral defense. *Scientific reports*, *7*(1), 1-15.

Brutscher, L. M., & Flenniken, M. L. (2015). RNAi and Antiviral Defense in the Honey Bee. *J Immunol Res*, *2015*, 941897.

Budhlakoti, N., Kushwaha, A. K., Rai, A., Chaturvedi, K., Kumar, A., Pradhan, A. K., Kumar, U., Kumar, R. R., Juliana, P., & Mishra, D. (2022). Genomic selection: A tool for accelerating the efficiency of molecular breeding for development of climate-resilient crops. *Frontiers in genetics*, *13*, 832153.

Bulus, M. (2023). Pwrss: Statistical power and sample size calculation tools. *R Package Version 0.3*, *1*.

Chan, Q. W. T., Melathopoulos, A. P., Pernal, S. F., & Foster, L. J. (2009). The innate immune and systemic response in honey bees to a bacterial pathogen, Paenibacillus larvae. *BMC genomics*, *10*(1), 387.

Chapman, N. C., Bourgeois, A. L., Beaman, L. D., Lim, J., Harpur, B. A., Zayed, A., Allsopp, M. H., Rinderer, T. E., & Oldroyd, B. P. (2017). An abbreviated SNP panel for ancestry assignment of honeybees (Apis mellifera). *Apidologie*, *48*(6), 776-783.

Chapman, N. C., Harpur, B. A., Lim, J., Rinderer, T. E., Allsopp, M. H., Zayed, A., & Oldroyd, B. P. (2015). A SNP test to identify Africanized honeybees via proportion of ‘African’ancestry. *Molecular Ecology Resources*, *15*(6), 1346-1355.

Chapman, N. C., Harpur, B. A., Lim, J., Rinderer, T. E., Allsopp, M. H., Zayed, A., & Oldroyd, B. P. (2016). Hybrid origins of Australian honeybees (Apis mellifera). *Apidologie*, *47*(1), 26-34.

Chavan, S. S., Pavlov, V. A., & Tracey, K. J. (2017). Mechanisms and therapeutic relevance of neuro-immune communication. *Immunity*, *46*(6), 927-942.

Chaves, J. A., Cooper, E. A., Hendry, A. P., Podos, J., De León, L. F., Raeymaekers, J. A., MacMillan, W. O., & Uy, J. A. C. (2016). Genomic variation at the tips of the adaptive radiation of Darwin's finches. *Molecular ecology*, *25*(21), 5282-5295.

Chen, C., Liu, Z., Pan, Q., Chen, X., Wang, H., Guo, H., Liu, S., Lu, H., Tian, S., & Li, R. (2016). Genomic analyses reveal demographic history and temperate adaptation of the newly discovered honey bee subspecies Apis mellifera sinisxinyuan n. ssp. *Molecular biology and evolution*, *33*(5), 1337-1348.

Chen, X., & Shi, W. (2020). Genome-wide characterization of coding and non-coding RNAs in the ovary of honeybee workers and queens. *Apidologie*, *51*(5), 777-792.

Chouvenc, T., Efstathion, C. A., Elliott, M. L., & Su, N.-Y. (2013). Extended disease resistance emerging from the faecal nest of a subterranean termite. Proc. R. Soc. B,

Christen, V., & Fent, K. (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.

Cingolani, P. (2012). Variant annotation and functional prediction: SnpEff. In *Variant calling: methods and protocols* (pp. 289-314). Springer.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*, *6*(2), 80-92.

Cini, A., Bordoni, A., Cappa, F., Petrocelli, I., Pitzalis, M., Iovinella, I., Dani, F. R., Turillazzi, S., & Cervo, R. (2020). Increased immunocompetence and network centrality of allogroomer workers suggest a link between individual and social immunity in honeybees. *Scientific reports*, *10*(1), 8928.

Claing, G., Ferland, J., Kempers, M., Kozak, P., Labuschagne, R., Marchant, P., Melanson, M., Menzies, C., Micholson, D., & Morfin, N. STATEMENT ON HONEY BEE WINTERING LOSSES AND DISEASES MANAGEMENT IN CANADA FOR 2025.

Claudianos, C., Ranson, H., Johnson, R., Biswas, S., Schuler, M., Berenbaum, M., Feyereisen, R., & Oakeshott, J. G. (2006). A deficit of detoxification enzymes: pesticide sensitivity and environmental response in the honeybee. *Insect molecular biology*, *15*(5), 615-636.

Cobey, S. W., Tarpy, D. R., & Woyke, J. (2013). Standard methods for instrumental insemination of Apis mellifera queens. *Journal of Apicultural Research*, *52*(4), 1-18.

Comeault, A. A., Soria-Carrasco, V., Gompert, Z., Farkas, T. E., Buerkle, C. A., Parchman, T. L., & Nosil, P. (2014). Genome-wide association mapping of phenotypic traits subject to a range of intensities of natural selection in Timema cristinae. *The American Naturalist*, *183*(5), 711-727.

Conlon, B. H., Frey, E., Rosenkranz, P., Locke, B., Moritz, R. F., & Routtu, J. (2018). The role of epistatic interactions underpinning resistance to parasitic Varroa mites in haploid honey bee (Apis mellifera) drones. *Journal of evolutionary biology*, *31*(6), 801-809.

Consortium, H. G. S. (2006). Insights into social insects from the genome of the honeybee Apis mellifera. *Nature*, *443*(7114), 931.

Cornman, R. S., Boncristiani, H., Dainat, B., Chen, Y., Weaver, D., & Evans, J. D. (2013). Population-genomic variation within RNA viruses of the Western honey bee, Apis mellifera, inferred from deep sequencing. *BMC genomics*, *14*(1), 154.

Craig, A. M., & Kang, Y. (2007). Neurexin–neuroligin signaling in synapse development. *Current opinion in neurobiology*, *17*(1), 43-52.

Cremer, S., Armitage, S. A., & Schmid-Hempel, P. (2007). Social immunity. *Current biology*, *17*(16), R693-R702.

Cuevas, A., Febrero, M., & Fraiman, R. (2004). An anova test for functional data. *Computational statistics & data analysis*, *47*(1), 111-122.

Danka, R. G., Harris, J. W., Villa, J. D., & Dodds, G. E. (2013). Varying congruence of hygienic responses to Varroa destructor and freeze-killed brood among different types of honeybees. *Apidologie*, *44*(4), 447-457.

De Gregorio, E., Spellman, P. T., Rubin, G. M., & Lemaitre, B. (2001). Genome-wide analysis of the Drosophila immune response by using oligonucleotide microarrays. *Proceedings of the National Academy of Sciences*, *98*(22), 12590-12595.

De Jong, D., Roma, D. D. A., & Goncalves, L. (1982). A comparative analysis of shaking solutions for the detection of Varroa jacobsoni on adult honeybees. *Apidologie*, *13*(3), 297-306.

De Jong, D., Roma, D. D. A., & Gonçalves, L. (1982). A comparative analysis of shaking solutions for the detection of Varroa jacobsoni on adult honeybees. *Apidologie*, *13*(3), 297-306.

Decanini, L. I., Collins, A. M., & Evans, J. D. (2007). Variation and heritability in immune gene expression by diseased honeybees. *Journal of Heredity*, *98*(3), 195-201.

DeLory, T., Funderburk, K., Miller, K., Zuluaga-Smith, W., McPherson, S., Pirk, C. W., Costa, C., Weinstein-Teixeira, É., Dahle, B., & Rueppell, O. (2020). Local variation in recombination rates of the honey bee (Apis mellifera) genome among samples from six disparate populations. *Insectes Sociaux*, *67*(1), 127-138.

Di Prisco, G., Cavaliere, V., Annoscia, D., Varricchio, P., Caprio, E., Nazzi, F., Gargiulo, G., & Pennacchio, F. (2013). Neonicotinoid clothianidin adversely affects insect immunity and promotes replication of a viral pathogen in honey bees. *Proceedings of the National Academy of Sciences*, *110*(46), 18466-18471.

Dogantzis, K. A. (2022). Understanding the evolutionary origin and ancestral composition of honey bee (Apis mellifera) populations.

Dogantzis, K. A., Tiwari, T., Conflitti, I. M., Dey, A., Patch, H. M., Muli, E. M., Garnery, L., Whitfield, C. W., Stolle, E., Alqarni, A. S., Allsopp, M. H., & Zayed, A. (2021). Thrice out of Asia and the adaptive radiation of the western honey bee. *Science Advances*, *7*(49), eabj2151.

Dogantzis, K. A., & Zayed, A. (2019). Recent advances in population and quantitative genomics of honey bees. *Current Opinion in Insect Science*, *31*, 93-98.

Doublet, V., Labarussias, M., de Miranda, J. R., Moritz, R. F., & Paxton, R. J. (2015). Bees under stress: sublethal doses of a neonicotinoid pesticide and pathogens interact to elevate honey bee mortality across the life cycle. *Environmental microbiology*, *17*(4), 969-983.

Doublet, V., Poeschl, Y., Gogol-Döring, A., Alaux, C., Annoscia, D., Aurori, C., Barribeau, S. M., Bedoya-Reina, O. C., Brown, M. J., & Bull, J. C. (2017a). Unity in defence: honeybee workers exhibit conserved molecular responses to diverse pathogens. *BMC genomics*, *18*(1), 207.

Doublet, V., Poeschl, Y., Gogol-Döring, A., Alaux, C., Annoscia, D., Aurori, C., Barribeau, S. M., Bedoya-Reina, O. C., Brown, M. J., & Bull, J. C. (2017b). Unity in defence: honeybee workers exhibit conserved molecular responses to diverse pathogens. *BMC genomics*, *18*(1), 1-17.

Elsik, C. G., Worley, K. C., Bennett, A. K., Beye, M., Camara, F., Childers, C. P., de Graaf, D. C., Debyser, G., Deng, J., & Devreese, B. (2014). Finding the missing honey bee genes: lessons learned from a genome upgrade. *BMC genomics*, *15*(1), 86.

Evans, J., Aronstein, K., Chen, Y. P., Hetru, C., Imler, J. L., Jiang, H., Kanost, M., Thompson, G., Zou, Z., & Hultmark, D. (2006). Immune pathways and defence mechanisms in honey bees Apis mellifera. *Insect molecular biology*, *15*(5), 645-656.

Evans, J. D., & Schwarz, R. S. (2011). Bees brought to their knees: microbes affecting honey bee health. *Trends in microbiology*, *19*(12), 614-620.

Evans, J. D., Schwarz, R. S., Chen, Y. P., Budge, G., Cornman, R. S., De la Rua, P., de Miranda, J. R., Foret, S., Foster, L., & Gauthier, L. (2013). Standard methods for molecular research in Apis mellifera. *Journal of Apicultural Research*, *52*(4), 1-54.

Evans, J. D., Schwarz, R. S., Chen, Y. P., Budge, G., Cornman, R. S., De la Rua, P., de Miranda, J. R., Foret, S., Foster, L., Gauthier, L., Genersch, E., Gisder, S., Jarosch, A., Kucharski, R., Lopez, D., Lun, C. M., Moritz, R. F. A., Maleszka, R., Muñoz, I., & Pinto, M. A. (2013). Standard methods for molecular research in Apis mellifera. *Journal of Apicultural Research*, *52*(4), 1-54.

Evans, J. D., & Spivak, M. (2010). Socialized medicine: individual and communal disease barriers in honey bees. *Journal of invertebrate pathology*, *103*, S62-S72.

Eynard, S. E., Mondet, F., Basso, B., Bouchez, O., Le Conte, Y., Dainat, B., Decourtye, A., Genestout, L., Guichard, M., Guillaume, F., Labarthe, E., Locke, B., Mahla, R., de Miranda, J., Neuditschko, M., Phocas, F., Canale-Tabet, K., Vignal, A., & Servin, B. (2025). Sequence-Based Multi Ancestry Association Study Reveals the Polygenic Architecture of Varroa destructor Resistance in the Honeybee Apis mellifera. *Mol Ecol*, *34*(3), e17637.

Farina, W. M. (1996). Food-exchange by foragers in the hive–a means of communication among honey bees? *Behavioral Ecology and Sociobiology*, *38*(1), 59-64.

Ferland, J., Kempers, M., Kennedy, K., Kozak, P., Lafrenière, R., Maund, C., & Wilson, G. (2020). Canadian Association of Professional Apiculturists statement on honey bee wintering losses in Canada. *Retrieved on*, *22*.

Fisher, R. A. (1931). XVII.—The distribution of gene ratios for rare mutations. *Proceedings of the Royal Society of Edinburgh*, *50*, 204-219.

Fontaine, A., Filipović, I., Fansiri, T., Hoffmann, A. A., Cheng, C., Kirkpatrick, M., Rašić, G., & Lambrechts, L. (2017). Extensive genetic differentiation between homomorphic sex chromosomes in the mosquito vector, Aedes aegypti. *Genome biology and evolution*, *9*(9), 2322-2335.

Francisco, L., Marxer, M., Schmid-Hempel, R., du Plessis, L., Antonio, D. S., Bourke, A. F., Mallon, E. B., Asher, C., de Graaf, D. C., & Otte, M. (2015). The genomes of two key bumblebee species with primitive eusocial organization.

French, S. K., Pepinelli, M., Conflitti, I. M., Jamieson, A., Higo, H., Common, J., Walsh, E. M., Bixby, M., Guarna, M. M., & Pernal, S. F. (2024). Honey bee stressor networks are complex and dependent on crop and region. *Current biology*, *34*(9), 1893-1903. e1893.

Frisch, K. v. (1993). *The dance language and orientation of bees*. Harvard University Press.

Fuller, Z. L., Niño, E. L., Patch, H. M., Bedoya-Reina, O. C., Baumgarten, T., Muli, E., Mumoki, F., Ratan, A., McGraw, J., & Frazier, M. (2015). Genome-wide analysis of signatures of selection in populations of African honey bees (Apis mellifera) using new web-based tools. *BMC genomics*, *16*, 1-18.

Galbraith, D. A., Yang, X., Nino, E. L., Yi, S., & Grozinger, C. (2015). Parallel epigenomic and transcriptomic responses to viral infection in honey bees (Apis mellifera). *PLoS pathogens*, *11*(3), e1004713.

Giordani, G., Cattabriga, G., Becchimanzi, A., Di Lelio, I., De Leva, G., Gigliotti, S., Pennacchio, F., Gargiulo, G., & Cavaliere, V. (2023). Role of neuronal and non-neuronal acetylcholine signaling in Drosophila humoral immunity. *Insect biochemistry and molecular biology*, *153*, 103899.

Giral, H., Landmesser, U., & Kratzer, A. (2018). Into the wild: GWAS exploration of non-coding RNAs. *Frontiers in cardiovascular medicine*, *5*, 181.

Goulson, D., Nicholls, E., Botías, C., & Rotheray, E. L. (2015a). Bee declines driven by combined stress from parasites, pesticides, and lack of flowers. *Science*, *347*(6229), 1255957.

Goulson, D., Nicholls, E., Botías, C., & Rotheray, E. L. (2015b). Bee declines driven by combined stress from parasites, pesticides, and lack of flowers. *Science*, *347*(6229).

Grosso, J. P., Barneto, J. A., Velarde, R. A., Pagano, E. A., Zavala, J. A., & Farina, W. M. (2018). An early sensitive period induces long-lasting plasticity in the honeybee nervous system. *Frontiers in Behavioral Neuroscience*, *12*, 11.

Grozinger, C. M., & Zayed, A. (2020a). Improving bee health through genomics. *Nature Reviews Genetics*, 1-15.

Grozinger, C. M., & Zayed, A. (2020b). Improving bee health through genomics. *Nature Reviews Genetics*, *21*(5), 277-291.

Grundberg, E., Small, K. S., Hedman, Å. K., Nica, A. C., Buil, A., Keildson, S., Bell, J. T., Yang, T.-P., Meduri, E., & Barrett, A. (2012). Mapping cis-and trans-regulatory effects across multiple tissues in twins. *Nature genetics*, *44*(10), 1084-1089.

Guarna, M. M., Hoover, S. E., Huxter, E., Higo, H., Moon, K.-M., Domanski, D., Bixby, M. E., Melathopoulos, A. P., Ibrahim, A., & Peirson, M. (2017). Peptide biomarkers used for the selective breeding of a complex polygenic trait in honey bees. *Scientific reports*, *7*(1), 1-10.

Guarna, M. M., Hoover, S. E., Huxter, E., Higo, H., Moon, K.-M., Domanski, D., Bixby, M. E. F., Melathopoulos, A. P., Ibrahim, A., Peirson, M., Desai, S., Micholson, D., White, R., Borchers, C. H., Currie, R. W., Pernal, S. F., & Foster, L. J. (2017). Peptide biomarkers used for the selective breeding of a complex polygenic trait in honey bees. *Scientific reports*, *7*(1), 8381.

Guarna, M. M., Melathopoulos, A. P., Huxter, E., Iovinella, I., Parker, R., Stoynov, N., Tam, A., Moon, K.-M., Chan, Q. W. T., Pelosi, P., White, R., Pernal, S. F., & Foster, L. J. (2015). A search for protein biomarkers links olfactory signal transduction to social immunity. *BMC Genomics*, *16*(1), 63.

Guichard, M., Dainat, B., Eynard, S., Vignal, A., Servin, B., Consortium, B., & Neuditschko, M. (2021). Identification of quantitative trait loci associated with calmness and gentleness in honey bees using whole-genome sequences. *Animal genetics*, *52*(4), 472-481.

Guichard, M., Dainat, B., Eynard, S., Vignal, A., Servin, B., Consortium, B., & Neuditschko, M. (2022). Two quantitative trait loci are associated with recapping of Varroa destructor-infested brood cells in Apis mellifera mellifera. *Animal genetics*, *53*(1), 156-160.

Haile, T. A., Walkowiak, S., N'Diaye, A., Clarke, J. M., Hucl, P. J., Cuthbert, R. D., Knox, R. E., & Pozniak, C. J. (2021). Genomic prediction of agronomic traits in wheat using different models and cross-validation designs. *Theoretical and Applied Genetics*, *134*, 381-398.

Hamiduzzaman, M. M., Emsen, B., Hunt, G. J., Subramanyam, S., Williams, C. E., Tsuruda, J. M., & Guzman-Novoa, E. (2017). Differential gene expression associated with honey bee grooming behavior in response to Varroa mites. *Behavior Genetics*, *47*, 335-344.

Hamilton, W. D. D. (1964). The genetical evolution of social behaviour. I. *Journal of Theoretical Biology*, *7*, 1-16.

Harbo, J. R., & Harris, J. W. (2005). Suppressed mite reproduction explained by the behaviour of adult bees. *Journal of Apicultural Research*, *44*(1), 21-23.

Harpur, B., Chapman, N., Krimus, L., Maciukiewicz, P., Sandhu, V., Sood, K., Lim, J., Rinderer, T., Allsopp, M., & Oldroyd, B. (2015). Assessing patterns of admixture and ancestry in Canadian honey bees. *Insectes Sociaux*, *62*(4), 479-489.

Harpur, B. A., & et al. (2019). Integrative genomics reveals the genetics and evolution of the honey bee's social immune system. *Genome biology and evolution*, *11*(3), 937-948.

Harpur, B. A., Guarna, M. M., Huxter, E., Higo, H., Moon, K.-M., Hoover, S. E., Ibrahim, A., Melathopoulos, A. P., Desai, S., & Currie, R. W. (2019). Integrative genomics reveals the genetics and evolution of the honey bee's social immune system. *Genome biology and evolution*, *11*(3), 937-948.

Harpur, B. A., Kadri, S. M., Orsi, R. O., Whitfield, C. W., & Zayed, A. (2020). Defense Response in Brazilian Honey Bees (*Apis mellifera scutellata* × spp.) Is Underpinned by Complex Patterns of Admixture. *Genome biology and evolution*, *12*(8), 1367-1377.

Harpur, B. A., Kent, C. F., Molodtsova, D., Lebon, J. M., Alqarni, A. S., Owayss, A. A., & Zayed, A. (2014a). Population genomics of the honey bee reveals strong signatures of positive selection on worker traits. *Proceedings of the National Academy of Sciences*, *111*(7), 2614-2619.

Harpur, B. A., Kent, C. F., Molodtsova, D., Lebon, J. M. D., Alqarni, A. S., Owayss, A. a., & Zayed, A. (2014b). Population genomics of the honey bee reveals strong signatures of positive selection on worker traits. *Proceedings of the National Academy of Sciences of the United States of America*, *111*, 2614-2619.

Harpur, B. A., Minaei, S., Kent, C. F., & Zayed, A. (2012). Management increases genetic diversity of honey bees via admixture. *Molecular ecology*, *21*(18), 4414-4421.

Harpur, B. A., Minaei, S., Kent, C. F., & Zayed, A. (2013). Admixture increases diversity in managed honey bees: reply to De la Rúa et al.(2013). *Molecular ecology*, *22*(12), 3211-3215.

Harpur, B. A., & Zayed, A. (2013). Accelerated evolution of innate immunity proteins in social insects: adaptive evolution or relaxed constraint? *Molecular biology and evolution*, *30*(7), 1665-1674.

Harris, J., Harbo, J., Villa, J., & Danka, R. (2003). Variable population growth of Varroa destructor (Mesostigmata: Varroidae) in colonies of honey bees (Hymenoptera: Apidae) during a 10-year period. *Environmental entomology*, *32*(6), 1305-1312.

Hayes, B. (2013). Overview of statistical methods for genome-wide association studies (GWAS). In *Genome-wide association studies and genomic prediction* (pp. 149-169). Springer.

Hemmling, C., Koeniger, N., & Ruttner, F. (1979). Quantitative Bestimmung der 9-oxodecensäure im Lebenszyklus der Kapbiene (Apis mellifera capensis Escholtz). *Apidologie*, *10*(3), 227-240.

Henriques, D., Lopes, A. R., Chejanovsky, N., Dalmon, A., Higes, M., Jabal-Uriel, C., Le Conte, Y., Reyes-Carreno, M., Soroker, V., & Martin-Hernandez, R. (2021). A SNP assay for assessing diversity in immune genes in the honey bee (Apis mellifera L.). *Scientific reports*, *11*(1), 15317.

Henriques, D., Parejo, M., Vignal, A., Wragg, D., Wallberg, A., Webster, M. T., & Pinto, M. A. (2018). Developing reduced SNP assays from whole-genome sequence data to estimate introgression in an organism with complex genetic patterns, the Iberian honeybee (Apis mellifera iberiensis). *Evolutionary Applications*, *11*(8), 1270-1282.

Hölldobler, B., & Wilson, E. O. (2009). *The superorganism: the beauty, elegance, and strangeness of insect societies*. WW Norton & Company.

Hou, H.-X., Huang, D.-W., Xin, Z.-Z., & Xiao, J.-H. (2021). Genome-wide analysis of gene families of pattern recognition receptors in fig wasps (Hymenoptera, Chalcidoidea). *Genes*, *12*(12), 1952.

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols*, *4*(1), 44-57.

Humann, F. C., Tiberio, G. J., & Hartfelder, K. (2013). Sequence and expression characteristics of long noncoding RNAs in honey bee caste development–potential novel regulators for transgressive ovary size. *PloS one*, *8*(10), e78915.

Hunt, G. (2007). Flight and fight: a comparative view of the neurophysiology and genetics of honey bee defensive behavior. *Journal of insect physiology*, *53*(5), 399-410.

Hunt, G. J., Amdam, G. V., Schlipalius, D., Emore, C., Sardesai, N., Williams, C. E., Rueppell, O., Guzmán-Novoa, E., Arechavaleta-Velasco, M., & Chandra, S. (2007). Behavioral genomics of honeybee foraging and nest defense. *Naturwissenschaften*, *94*(4), 247-267.

Inbar, S., Cohen, P., Yahav, T., & Privman, E. (2020). Comparative study of population genomic approaches for mapping colony-level traits. *PLoS computational biology*, *16*(3), e1007653.

Ishihama, Y., Rappsilber, J., Andersen, J. S., & Mann, M. (2002). Microcolumns with self-assembled particle frits for proteomics. *Journal of chromatography A*, *979*(1-2), 233-239.

Jernelöv, A. (2017). African “Killer Bees” in the Americas. In *The Long-Term Fate of Invasive Species* (pp. 251-260). Springer.

Jo, B.-S., & Choi, S. S. (2015). Introns: the functional benefits of introns in genomes. *Genomics & informatics*, *13*(4), 112.

Johnson, B. R., & Tsutsui, N. D. (2011). Taxonomically restricted genes are associated with the evolution of sociality in the honey bee. *BMC genomics*, *12*(1), 164.

Keightley, P. D., & Lynch, M. (2003). Toward a realistic model of mutations affecting fitness. *Evolution*, *57*(3), 683-685.

Kent, C. F., Dey, A., Patel, H., Tsvetkov, N., Tiwari, T., MacPhail, V. J., Gobeil, Y., Harpur, B. A., Gurtowski, J., & Schatz, M. C. (2018). Conservation genomics of the declining North American bumblebee Bombus terricola reveals inbreeding and selection on immune genes. *Frontiers in genetics*, *9*, 316.

Kent, C. F., Minaei, S., Harpur, B. A., & Zayed, A. (2012). Recombination is associated with the evolution of genome structure and worker behavior in honey bees. *Proceedings of the National Academy of Sciences*, *109*(44), 18012-18017.

Kimura, M. (1985). The role of compensatory neutral mutations in molecular evolution. *Journal of Genetics*, *64*(1), 7-19.

Klein, A.-M., Vaissière, B. E., Cane, J. H., Steffan-Dewenter, I., Cunningham, S. A., Kremen, C., & Tscharntke, T. (2007). Importance of pollinators in changing landscapes for world crops. *Proceedings of the royal society B: biological sciences*, *274*(1608), 303-313.

Knight, D., Xie, W., & Boulianne, G. L. (2011). Neurexins and neuroligins: recent insights from invertebrates. *Molecular neurobiology*, *44*, 426-440.

Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K., & Mardis, E. R. (2013). The next-generation sequencing revolution and its impact on genomics. *Cell*, *155*(1), 27-38.

Kono, Y., & Kohn, J. R. (2015). Range and frequency of Africanized honey bees in California (USA). *PLoS One*, *10*(9), e0137407.

Konrad, M., Pull, C. D., Metzler, S., Seif, K., Naderlinger, E., Grasse, A. V., & Cremer, S. (2018). Ants avoid superinfections by performing risk-adjusted sanitary care. *Proceedings of the National Academy of Sciences*, *115*(11), 2782-2787.

Korte, A., & Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. *Plant methods*, *9*(1), 29.

Kou, K., Yang, H., Li, H., Fang, C., Chen, L., Yue, L., Nan, H., Kong, L., Li, X., & Wang, F. (2022). A functionally divergent SOC1 homolog improves soybean yield and latitudinal adaptation. *Current biology*, *32*(8), 1728-1742. e1726.

Kunieda, T., Fujiyuki, T., Kucharski, R., Foret, S., Ament, S., Toth, A., Ohashi, K., Takeuchi, H., Kamikouchi, A., & Kage, E. (2006). Carbohydrate metabolism genes and pathways in insects: insights from the honey bee genome. *Insect molecular biology*, *15*(5), 563-576.

Lapidge, K. L., Oldroyd, B. P., & Spivak, M. (2002). Seven suggestive quantitative trait loci influence hygienic behavior of honey bees. *Naturwissenschaften*, *89*(12), 565-568.

Le Conte, Y., Ellis, M., & Ritter, W. (2010). Varroa mites and honey bee health: can Varroa explain part of the colony losses? *Apidologie*, *41*(3), 353-363.

Lee, S. H., Van Der Werf, J. H., Hayes, B. J., Goddard, M. E., & Visscher, P. M. (2008). Predicting unobserved phenotypes for complex traits from whole-genome SNP data. *PLoS genetics*, *4*(10), e1000231.

Lemaitre, B., & Hoffmann, J. (2007). The host defense of Drosophila melanogaster. *Annu. Rev. Immunol.*, *25*(1), 697-743.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987-2993.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. 692 (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078-2693.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & Subgroup, G. P. D. P. (2009). The sequence alignment/map format and samtools. *Bioinformatics*, *25*(16), 2078-2079.

Lin, M., Whitmire, S., Chen, J., Farrel, A., Shi, X., & Guo, J.-t. (2017). Effects of short indels on protein structure and function in human genomes. *Scientific reports*, *7*(1), 9313.

Lin, W., McBroome, J., Rehman, M., & Johnson, B. R. (2018). Africanized bees extend their distribution in California. *PLoS One*, *13*(1), e0190604.

Linksvayer, T. A. (2006). Direct, maternal, and sibsocial genetic effects on individual and colony traits in an ant. *Evolution*, *60*(12), 2552-2561.

Linksvayer, T. A., & Wade, M. J. (2009). Genes with social effects are expected to harbor more sequence variation within and between species. *Evolution*, *63*, 1685-1696.

Liu, L., Zhao, X.-Y., Tang, Q.-B., Lei, C.-L., & Huang, Q.-Y. (2019). The mechanisms of social immunity against fungal infections in eusocial insects. *Toxins*, *11*(5), 244.

Liu, Y., Beyer, A., & Aebersold, R. (2016). On the dependency of cellular protein levels on mRNA abundance. *Cell*, *165*(3), 535-550.

Lynch, M. (2010). Evolution of the mutation rate. *Trends in Genetics*, *26*(8), 345-352.

Masterman, R., Ross, R., Mesce, K., & Spivak, M. (2001). Olfactory and behavioral response thresholds to odors of diseased brood differ between hygienic and non-hygienic honey bees (Apis mellifera L.). *Journal of Comparative Physiology A*, *187*, 441-452.

Mattila, H. R., & Seeley, T. D. (2007). Genetic diversity in honey bee colonies enhances productivity and fitness. *Science*, *317*(5836), 362-364.

McDonald, J. H., & Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, *351*(6328), 652-654.

McGaugh, S. E., Lorenz, A. J., & Flagel, L. E. (2021). The utility of genomic prediction models in evolutionary genetics. *Proceedings of the Royal Society B*, *288*(1956), 20210693.

Melzer, D., Perry, J. R., Hernandez, D., Corsi, A.-M., Stevens, K., Rafferty, I., Lauretani, F., Murray, A., Gibbs, J. R., & Paolisso, G. (2008). A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS genetics*, *4*(5), e1000072.

Meyer, R. S., & Purugganan, M. D. (2013). Evolution of crop species: genetics of domestication and diversification. *Nature Reviews Genetics*, *14*(12), 840-852.

Michener, C. (1974). *The Social Behavior of the Bees: A Comparative Study* (Vol. 404). Harvard University Press (Belknap Press).

Mikhailova, A. A., Rinke, S., & Harrison, M. C. (2024). Genomic signatures of eusocial evolution in insects. *Current Opinion in Insect Science*, *61*, 101136.

Mills, R. E., Luttig, C. T., Larkins, C. E., Beauchamp, A., Tsui, C., Pittard, W. S., & Devine, S. E. (2006). An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome research*, *16*(9), 1182-1190.

Mondet, F., Beaurepaire, A., McAfee, A., Locke, B., Alaux, C., Blanchard, S., Danka, B., & Le Conte, Y. (2020). Honey bee survival mechanisms against the parasite Varroa destructor: a systematic review of phenotypic and genomic research efforts. *International journal for parasitology*, *50*(6-7), 433-447.

Moniruzzaman, M., Khatun, R., & Mintoo, A. (2014). Application of marker assisted selection for livestock improvement in Bangladesh. *Bangladesh Veterinarian*, *31*(1), 1-11.

Montgomery, S. B., Goode, D. L., Kvikstad, E., Albers, C. A., Zhang, Z. D., Mu, X. J., Ananda, G., Howie, B., Karczewski, K. J., & Smith, K. S. (2013). The origin, evolution, and functional impact of short insertion–deletion variants identified in 179 human genomes. *Genome research*, *23*(5), 749-761.

Muñoz, I., Henriques, D., Jara, L., Johnston, J. S., Chávez-Galarza, J., De La Rúa, P., & Pinto, M. A. (2017). SNP s selected by information content outperform randomly selected microsatellite loci for delineating genetic identification and introgression in the endangered dark European honeybee (Apis mellifera mellifera). *Molecular Ecology Resources*, *17*(4), 783-795.

Muzzey, D., Evans, E. A., & Lieber, C. (2015). Understanding the basics of NGS: from mechanism to variant calling. *Current genetic medicine reports*, *3*(4), 158-165.

Naug, D., & Camazine, S. (2002). The role of colony organization on pathogen transmission in social insects. *Journal of Theoretical Biology*, *215*(4), 427-439.

Neumann, P., & Carreck, N. L. (2010). Honey bee colony losses. *Journal of Apicultural Research*, *49*(1), 1-6.

Nica, A. C., & Dermitzakis, E. T. (2013). Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *368*(1620), 20120362.

Nielsen, R. (2005). Molecular signatures of natural selection. *Annu. Rev. Genet.*, *39*, 197-218.

Niu, J., Meeus, I., Cappelle, K., Piot, N., & Smagghe, G. (2014). The immune response of the small interfering RNA pathway in the defense against bee viruses. *Current Opinion in Insect Science*, *6*, 22-27.

Nouhaud, P. (2018). Long-read based assembly and annotation of a Drosophila simulans genome. *bioRxiv*, 425710.

Osborne, K., Robichon, A., Burgess, E., Butland, S., Shaw, R., Coulthard, A., Pereira, H., Greenspan, R., & Sokolowski, M. (1997). Natural behavior polymorphism due to a cGMP-dependent protein kinase of Drosophila. *Science*, *277*(5327), 834-836.

Oxley, P. R., Spivak, M., & Oldroyd, B. P. (2010). Six quantitative trait loci influence task thresholds for hygienic behaviour in honeybees (Apis mellifera). *Molecular ecology*, *19*(7), 1452-1461.

Parejo, M., Wragg, D., Gauthier, L., Vignal, A., Neumann, P., & Neuditschko, M. (2016). Using whole-genome sequence information to foster conservation efforts for the European Dark Honey Bee, Apis mellifera mellifera. *Frontiers in Ecology and Evolution*, *4*, 140.

Parejo, M., Wragg, D., Henriques, D., Charrière, J.-D., & Estonba, A. (2020). Digging into the genomic past of Swiss honey bees by whole-genome sequencing museum specimens. *Genome biology and evolution*, *12*(12), 2535-2551.

Park, O., Pellet, F., & Paddock, F. (1937). Disease resistance and American foulbrood. *Am. Bee J*, *77*(1), 20-25.

Perry, C. J., & Barron, A. B. (2013). Honey bees selectively avoid difficult choices. *Proceedings of the National Academy of Sciences*, *110*(47), 19155-19159.

Peterson, R. A. (2017). Package 'bestNormalize'.

Peterson, R. A. (2021a). Finding Optimal Normalizing Transformations via best Normalize. *R Journal*, *13*(1).

Peterson, R. A. (2021b). Finding optimal normalizing transformations via bestNormalize.

Picard toolkit. (2019). https://broadinstitute.github.io/picard/

Potts, S. G., Biesmeijer, J. C., Kremen, C., Neumann, P., Schweiger, O., & Kunin, W. E. (2010). Global pollinator declines: trends, impacts and drivers. *Trends in ecology & evolution*, *25*(6), 345-353.

Presgraves, D. C. (2006). Intron length evolution in Drosophila. *Molecular biology and evolution*, *23*(11), 2203-2213.

R Core Team. (2013). R: A language and environment for statistical computing. *201*.

Richardson, R. T., Conflitti, I. M., Labuschagne, R. S., Hoover, S. E., Currie, R. W., Giovenazzo, P., Guarna, M. M., Pernal, S. F., Foster, L. J., & Zayed, A. (2023). Land use changes associated with declining honey bee health across temperate North America. *Environmental Research Letters*, *18*(6), 064042.

Rittschof, C. C., Coombs, C. B., Frazier, M., Grozinger, C. M., & Robinson, G. E. (2015). Early-life experience affects honey bee aggression and resilience to immune challenge. *Scientific reports*, *5*(1), 1-8.

Robinson, G. E., Fernald, R. D., & Clayton, D. F. (2008). Genes and social behavior. *Science*, *322*(5903), 896-900.

Robinson, G. E., Grozinger, C. M., & Whitfield, C. W. (2005). Sociogenomics: social life in molecular terms. *Nature Reviews Genetics*, *6*(4), 257-270.

Rockah-Shmuel, L., Tóth-Petróczy, Á., Sela, A., Wurtzel, O., Sorek, R., & Tawfik, D. S. (2013). Correlated occurrence and bypass of frame-shifting insertion-deletions (InDels) to give functional proteins. *PLoS genetics*, *9*(10), e1003882.

Rockman, M. V., & Kruglyak, L. (2006). Genetics of global gene expression. *Nature Reviews Genetics*, *7*(11), 862-872.

Rose, A. B. (2019). Introns as gene regulators: a brick on the accelerator. *Frontiers in genetics*, *9*, 672.

Rosenkranz, P., Aumeier, P., & Ziegelmann, B. (2010). Biology and control of Varroa destructor. *Journal of invertebrate pathology*, *103*, S96-S119.

Rothenbuhler, W. C. (1964). Behaviour genetics of nest cleaning in honey bees. I. Responses of four inbred lines to disease-killed brood. *Animal behaviour*, *12*(4), 578-583.

Runcie, D. E., & Crawford, L. (2019). Fast and flexible linear mixed models for genome-wide genetics. *PLoS genetics*, *15*(2), e1007978.

Russo, R. M., Pietronave, H., Conte, C. A., Liendo, M. C., Basilio, A., Lanzavecchia, S. B., & Scannapieco, A. C. (2024). Stimulus-specific gene expression profiles associated with grooming behavior and Varroa destructor resistance in honey bees. *Frontiers in Bee Science*, *2*, 1441317.

Sanchez-Bayo, F., & Goka, K. (2014). Pesticide residues and bees–a risk assessment. *PLoS One*, *9*(4), e94482.

Schlüns, H., & Crozier, R. (2007). Relish regulates expression of antimicrobial peptide genes in the honeybee, Apis mellifera, shown by RNA interference. *Insect molecular biology*, *16*(6), 753-759.

Schmid-Hempel, P. (1998). *Parasites in social insects*. Princeton University Press.

Schmid-Hempel, P. (2005). Evolutionary ecology of insect immune defenses. *Annu. Rev. Entomol.*, *50*(1), 529-551.

Schwenke, R. A., Lazzaro, B. P., & Wolfner, M. F. (2016). Reproduction–immunity trade-offs in insects. *Annual review of entomology*, *61*(1), 239-256.

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

Sedlazeck, F. J., Rescheneder, P., & Von Haeseler, A. (2013b). NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics*, btt468.

Seeley, T. D. (1997). Honey bee colonies are group-level adaptive units. *The American Naturalist*, *150*(S1), s22-S41.

Seeley, T. D., & Visscher, P. K. (1985). Survival of honeybees in cold climates: the critical timing of colony growth and reproduction. *Ecological Entomology*, *10*(1), 81-88.

Sehrawat, S., Najafian, K., & Jin, L. (2023). Predicting phenotypes from novel genomic markers using deep learning. *Bioinformatics Advances*, *3*(1), vbad028.

Shelke, T., Gupta, V., & Gupta, I. (2025). Unraveling Insect Immunity: A Cross-Order Comparative Genomic Analysis of Key Immune Proteins. RECOMB International Workshop on Comparative Genomics,

Sherman, B. T., Hao, M., Qiu, J., Jiao, X., Baseler, M. W., Lane, H. C., Imamichi, T., & Chang, W. (2022). DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic acids research*, *50*(W1), W216-W221.

Sih, A., Bell, A. M., & Kerby, J. L. (2004). Two stressors are far deadlier than one. *Trends in ecology & evolution*, *19*(6), 274-276.

Simone-Finstrom, M. (2017). Social immunity and the superorganism: Behavioral defenses protecting honey bee colonies from pathogens and parasites. *Bee World*, *94*(1), 21-29.

Simone-Finstrom, M. D., & Spivak, M. (2012). Increased resin collection after parasite challenge: a case of self-medication in honey bees? *PLoS One*, *7*(3), e34601.

Söderhäll, K., & Cerenius, L. (1998). Role of the prophenoloxidase-activating system in invertebrate immunity. *Current Opinion in Immunology*, *10*(1), 23-28.

Soller, M., & Bar-Cohen, R. (1967). Some observations on the heritability and genetic correlation between honey production and brood area in the honeybee. *Journal of Apicultural Research*, *6*(1), 37-43.

Spivak, M., & Danka, R. G. (2021). Perspectives on hygienic behavior in Apis mellifera and other social insects. *Apidologie*, *52*(1), 1-16.

Spivak, M., & Reuter, G. S. (1998). Performance of hygienic honey bee colonies in a commercial apiary. *Apidologie*, *29*(3), 291-302.

Spötter, A., Gupta, P., Mayer, M., Reinsch, N., & Bienefeld, K. (2016). Genome-wide association study of a Varroa-specific defense behavior in honeybees (Apis mellifera). *Journal of Heredity*, *107*(3), 220-227.

Stort, A., & Gonçalves, L. (1991). The 'African'Honey Bee.

Studer, A., Zhao, Q., Ross-Ibarra, J., & Doebley, J. (2011). Identification of a functional transposon insertion in the maize domestication gene tb1. *Nature genetics*, *43*(11), 1160-1163.

Sun, B. B., Maranville, J. C., Peters, J. E., Stacey, D., Staley, J. R., Blackshaw, J., Burgess, S., Jiang, T., Paige, E., & Surendran, P. (2018). Genomic atlas of the human plasma proteome. *Nature*, *558*(7708), 73-79.

Szabo, T., & Lefkovitch, L. (1989). Effect of brood production and population size on honey production of honeybee colonies in Alberta, Canada. *Apidologie*, *20*(2), 157-163.

Szabo, T. I., & Heikel, D. T. (1987). Patterns of honeybee colony gain in Alberta, Canada. *Journal of Apicultural Research*, *26*(1), 47-52.

The major evolutionary transitions., 374 227-232 (1995).

Tarpy, D. R. (2000). *Behavioral and evolutionary mechanisms of polyandry in honey bees (Apis mellifera)*. University of California, Davis.

Tarpy, D. R., & Nielsen, D. (2002). Sampling Error, Effective Paternity, and Estimating the Genetic Structure of Honey Bee Colonies (Hymenoptera: Apidae). *Annals of the Entomological Society of America*, *95*(4), 513-528.

Tarpy, D. R., & Page, J., Robert E. (2000). No behavioral control over mating frequency in queen honey bees (Apis mellifera L.): implications for the evolution of extreme polyandry. *The American Naturalist*, *155*(6), 820-827.

Terrapon, N., Li, C., Robertson, H. M., Ji, L., Meng, X., Booth, W., Chen, Z., Childers, C. P., Glastad, K. M., & Gokhale, K. (2014). Molecular traces of alternative social organization in a termite genome. *Nature communications*, *5*(1), 3636.

Tiwari, T., & Zayed, A. (2021). Practical Applications of Genomics in Managing Honey bee Health. *Veterinary Clinics: Food Animal Practice*, *37*(3), 535-543.

Toth, A. L., & Robinson, G. E. (2007). Evo-devo and the evolution of social behavior. *Trends in Genetics*, *23*(7), 334-341.

Toth, A. L., & Robinson, G. E. (2007). Evo-devo and the evolution of social behavior. *Trends in Genetics*, *23*, 334-341.

Traniello, I. M., Avalos, A., Gachomba, M. J., Gernat, T., Chen, Z., Cash-Ahmed, A. C., Hamilton, A. R., Cook, J. L., & Robinson, G. E. (2025). Genetic variation influences food-sharing sociability in honey bees. *PLoS biology*, *23*(9), e3003367.

Treangen, T. J., & Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, *13*(1), 36-46.

Tsuruda, J. M., Harris, J. W., Bourgeois, L., Danka, R. G., & Hunt, G. J. (2012). High-resolution linkage analyses to identify genes that influence Varroa sensitive hygiene behavior in honey bees. *PloS one*, *7*(11), e48276.

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics*, 99-114.

Tyanova, S., Temu, T., & Cox, J. (2016). The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nature protocols*, *11*(12), 2301-2319.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., & Thibault, J. (2013a). From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current protocols in bioinformatics*, 11.10. 11-11.10. 33.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., & Thibault, J. (2013b). From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current protocols in bioinformatics*, *43*(1), 11.10. 11-11.10. 33.

VanEngelsdorp, D., Evans, J. D., Saegerman, C., Mullin, C., Haubruge, E., Nguyen, B. K., Frazier, M., Frazier, J., Cox-Foster, D., & Chen, Y. (2009). Colony collapse disorder: a descriptive study. *PLoS One*, *4*(8), e6481.

Verma, S., & Ruttner, F. (1983). Cytological analysis of the thelytokous parthenogenesis in the Cape honeybee (Apis mellifera capensis Escholtz). *Apidologie*, *14*(1), 41-57.

Vitti, J. J., Grossman, S. R., & Sabeti, P. C. (2013). Detecting natural selection in genomic data. *Annual review of genetics*, *47*, 97-120.

Vogel, C., & Marcotte, E. M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics*, *13*(4), 227-232.

Von Frisch, K. (1967). The dance language and orientation of bees.

Wallberg, A., Bunikis, I., Pettersson, O. V., Mosbech, M.-B., Childers, A. K., Evans, J. D., Mikheyev, A. S., Robertson, H. M., Robinson, G. E., & Webster, M. T. (2019). A hybrid de novo genome assembly of the honeybee, *Apis mellifera*, with chromosome-length scaffolds. *BMC Genomics*, *20*(1), 275.

Wallberg, A., Glémin, S., & Webster, M. T. (2015). Extreme recombination frequencies shape genome variation and evolution in the honeybee, Apis mellifera. *PLoS Genet*, *11*(4), e1005189.

Walsh, J. T., Garonski, A., Jackan, C., & Linksvayer, T. A. (2022). The collective behavior of ant groups depends on group genotypic composition. *Journal of Heredity*, *113*(1), 102-108.

Wang, J., Wurm, Y., Nipitwattanaphon, M., Riba-Grognuz, O., Huang, Y.-C., Shoemaker, D., & Keller, L. (2013). A Y-like social chromosome causes alternative colony organization in fire ants. *Nature*, *493*(7434), 664-668.

Whitfield, C. W., Ben-Shahar, Y., Brillet, C., Leoncini, I., Crauser, D., LeConte, Y., Rodriguez-Zas, S., & Robinson, G. E. (2006). Genomic dissection of behavioral maturation in the honey bee. *Proceedings of the National Academy of Sciences*, *103*(44), 16068-16075.

Wilm, A., Aw, P. P. K., Bertrand, D., Yeo, G. H. T., Ong, S. H., Wong, C. H., Khor, C. C., Petric, R., Hibberd, M. L., & Nagarajan, N. (2012). LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic acids research*, *40*(22), 11189-11201.

Wilson, D. S., & Wilson, E. O. (2007). Rethinking the theoretical foundation of sociobiology. *The Quarterly review of biology*, *82*(4), 327-348.

Wilson, E. O., & Hölldobler, B. (2005). Eusociality: origin and consequences. *Proceedings of the National Academy of Sciences*, *102*(38), 13367-13371.

Wilson-Rich, N., Dres, S. T., & Starks, P. T. (2008). The ontogeny of immunity: development of innate immune strength in the honey bee (Apis mellifera). *J Insect Physiol*, *54*(10-11), 1392-1399.

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

Woyke, J. (1963). What happens to diploid drone larvae in a honeybee colony. *Journal of Apicultural Research*, *2*(2), 73-75.

Wragg, D., Marti-Marimon, M., Basso, B., Bidanel, J.-P., Labarthe, E., Bouchez, O., Le Conte, Y., & Vignal, A. (2016). Whole-genome resequencing of honeybee drones to detect genomic selection in a population managed for royal jelly. *Scientific reports*, *6*(1), 1-13.

Wu, Y., Zheng, Z., Visscher, P. M., & Yang, J. (2017). Quantifying the mapping precision of genome-wide association studies using whole-genome sequencing data. *Genome biology*, *18*(1), 1-10.

Yagound, B., Dogantzis, K. A., Zayed, A., Lim, J., Broekhuyse, P., Remnant, E. J., Beekman, M., Allsopp, M. H., Aamidor, S. E., & Dim, O. (2020). A single gene causes thelytokous parthenogenesis, the defining feature of the cape honeybee Apis mellifera capensis. *Current biology*, *30*(12), 2248-2259. e2246.

Zayed, A. (2004). Effective population size in Hymenoptera with complementary sex determination. *Heredity*, *93*(6), 627-630.

Zhang, B., Zhang, C., Zhang, J., Lu, S., Zhao, H., Jiang, Y., & Ma, W. (2024). Regulatory roles of long non-coding RNAs in short-term heat stress in adult worker bees. *BMC Genomics*, *25*(1), 506.

Zhou, X., Carbonetto, P., & Stephens, M. (2013). Polygenic modeling with Bayesian sparse linear mixed models. *PLoS genetics*, *9*(2), e1003264.

Zhou, X., Li, J., Chen, L., Guo, M., Liang, R., & Pan, Y. (2024). The genomic pattern of insertion/deletion variations during rice improvement. *BMC genomics*, *25*(1), 1263.

Zhou, X., & Stephens, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature methods*, *11*(4), 407-409.

Appendix A: Statement on contributions

Tanushree Tiwari (TT) conducted all data analyses and led the writing and interpretation of results and discussion for Chapters 1, 2, 4, 5, 6, and 7, with guidance and feedback from the supervisor. For Chapter 3, data analysis and writing were equally contributed by Tanushree Tiwari (TT) and Kathleen Dogantzis (KAD).

Chapter 1: Introduction

Authors: Tanushree Tiwari

The introduction chapter was written by TT.

Chapter 2: Practical applications of genomics in managing honey bee health

Authors: Tanushree Tiwari and Amro Zayed

TT and AZ conceptualized the review chapter. TT wrote the manuscript. AZ provided comments on the manuscript.

Chapter 3: The genetics and evolution of the superorganism

Authors: Kathleen A Dogantzis†, Tanushree Tiwari†, Rodney T Richardson, Clement F Kent, Stephen Rose, Harshilkumar Patel, Alivia Dey, Ida M Conflitti, Renata B. Labuschagne, Shelley E Hoover, Rob W Currie, Pierre Giovenazzo, M Marta Guarna, Stephen F Pernal, Leonard J Foster, Amro Zayed

† These authors are considered co-first authors

KAD, TT, and AZ wrote the manuscript. KAD, and TT performed formal analyses and data visualization. KAD, TT, RTR, CFK, SR, and HP established methodology and data curation. AD

and IMC performed project administration, resource preparation and sequencing. SEH, RWC, PG, MMG, SFP, LJF and AZ acquired funding. All authors provided comments on the manuscript.

Chapter 4: Genetic architecture and regulatory networks underlying honey bee (*Apis mellifera*) innate immunity

Authors: Tanushree Tiwari, Alexandra Sébastien, Kathleen A Dogantzis, Rodney T Richardson, Clement F Kent, Stephen Rose, Harshilkumar Patel, Alivia Dey, Ida M Conflitti, Judith Trapp, Bradford Vinson, Nikolay Stoynov, Kyung-Mee Moon, Queenie W.T. Chan, Nonno Hasegawa, Renata B. Labuschagne, Shelley E Hoover, Rob W Currie, Pierre Giovenazzo, M Marta Guarna, Stephen F Pernal, Leonard J Foster, Amro Zayed

TT and AZ conceptualized the chapter. TT prepared and wrote the manuscript. TT carried out formal analyses and data visualization. TT, KAD, RR, CFK, SR, HP finalized sequenced data processing. AD and IMC performed project administration, resource preparation and sequencing. AS, JT, BV, NS, KM performed the sample preparation and Mass spectrometry data generation. SEH, RWC, PG, MMG, SFP, LJF and AZ acquired funding. AS, KAD, KM, QWTC, LJF, and AZ provided comments on the manuscript.

Chapter 5: Integrating the Genetics of Innate and Social Immunity in Honey Bees (*Apis mellifera).*

Authors: Tanushree Tiwari, Kathleen A Dogantzis, Alexandra Sébastien, Rodney T Richardson, Clement F Kent, Stephen Rose, Harshilkumar Patel, Alivia Dey, Ida M Conflitti, Judith Trapp, Bradford Vinson, Nikolay Stoynov, Kyung-Mee Moon, Queenie W.T. Chan, Nonno Hasegawa, Renata B. Labuschagne, Shelley E Hoover, Rob W Currie, Pierre Giovenazzo, M Marta Guarna, Stephen F Pernal, Leonard J Foster, Amro Zayed

TT and AZ conceptualized the chapter. TT prepared and wrote the manuscript. TT carried out formal analyses and data visualization. TT, KAD, RR, CFK, SR, HP finalized sequenced data processing. AD and IMC performed project administration, resource preparation and sequencing. AS, JT, BV, NS, KM performed the sample preparation and Mass spectrometry data generation.

SEH, RWC, PG, MMG, SFP, LJF and AZ acquired funding. AZ provided comments on the manuscript.

Chapter 6: Genome-wide distribution and functional implications of short InDels in honey bees (*Apis mellifera).*

Authors: Tanushree Tiwari, Clement F Kent, Kathleen A Dogantzis, Rodney T Richardson, Stephen Rose, Harshilkumar Patel, Alivia Dey, Ida M Conflitti, Renata B. Labuschagne, Shelley E Hoover, Rob W Currie, Pierre Giovenazzo, M Marta Guarna, Stephen F Pernal, Leonard J Foster, Amro Zayed*

TT prepared and wrote the manuscript. TT carried out formal analyses and data visualization. TT, CFK, AZ conceptualized the idea of InDel analysis; TT, KAD, RR, CFK, SR, HP finalized sequenced data processing. AD and IMC performed project administration and resource preparation and sequencing. SEH, RWC, PG, MMG, SFP, LJF and AZ acquired funding. CFK, KAD and AZ provided comments on the manuscript.

Chapter 7: Chapter seven: Conclusion, limitations, and future directions

Authors: Tanushree Tiwari

The conclusion chapter was written by T.T.

Appendix B: Project Acknowledgements:

We thank all cooperating beekeepers who participated in this study and/or donated queens to the project. We wish to thank all technicians (Abdullah Ibrahim, Michael Peirson, Heather Higo, Lynae Ovinge, Jeff Kearns), graduate students (Derek Micholson) research assistants (Andrée Rousseau, Zoe Rempel), the Ontario Technical Transfer Team staff (Les Eccles, Daniel Borges) and summer staff for their valuable help with sample collection, sample processing and field and lab work. We would also like to thank the National Bee Diagnostic Centre in Beaverlodge, Alberta, for virus, *Nosema* spp. and *Lotmaria passim* diagnostics, and William Cuthbert for their illustrative help with Figure 3.1. This work was supported by a Large-Scale Applied Research Project Grant from Genome Canada, Genome British Columbia, Genome Alberta, Genome Prairie, Ontario Genomics and Genome Quebec through the “BeeOMICS” Project (227Bee) awarded to

LJF/AZ/SFP/SEH/MMG/RWC/PG. Additional cash and in-kind co-funding was provided by: the University of British Columbia; the British Columbia Ministry of Agriculture, Food and Fisheries; the British Columbia Honey Producers Association; Bruker Daltonics; the Alberta Crop Industry Development Fund; Manitoba Agriculture, Food and Rural Initiatives (MAFRI); York University; the Ontario Ministry of Agriculture, Food and Rural Affairs; the Ontario Ministry of Colleges and Universities; the Ontario Beekeepers Association; and the Centre de recherche en sciences animales de Deschambault. Evolutionary aspects of this research were funded by a discovery grant from the Natural Sciences and Engineering Research Council of Canada to AZ.

Sincerely,

Tanushree Tiwari

Approved by,

Amro Zayed, PhD

Professor of Biology