Population genomic approaches to understanding the genetics and evolution of social insects

Brock A. Harpur

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, 2017

Abstract

Eusocial animals are comprised of distinct and specialized individuals that carry out specific tasks within colonies. Despite decades of research on eusocial insects (e.g. bees, wasps, termites, and ants), we lack knowledge on the genetics underlying social traits, and how the genomes of eusocial insects evolved over relevant timescales. I pioneered the use of next generation sequencing of populations of eusocial insects – population genomics – to characterize genomic regions that influence fitness and to study the genetics of two social traits. I first identified which genes have evidence of adaptive evolution within two genera: the primitively eusocial bumble bees and the highly eusocial honey bees (Chapters 2 and 3). Using a comparative approach, I found clear differences in which genes contribute to fitness within each lineage and in the caste-specific contributions to fitness at these two stages of social evolution. I then discovered the genes underpinning social immunity (Chapter 4) and colony defense (Chapters 5 and 6) in honey bees. I uncovered strong support that variation in social immunity arises through differential regulation of highly conserved neuronal developmental genes and that these genes have historical patterns of adaptive evolution. After creating a large genomic data set for a highly defensive honey bee population (Chapter 5), I discovered that variation in colony defense is underpinned by differences in inheritance of ancestral alleles. Finally, using population genomic data on honey bees, I tested the utility of a single nucleotide polymorphism assay to study the ancestry of Canadian honey bees, creating a powerful tool for securing the importation of honey bees into Canada (Chapter 7). My research highlights the importance of genomic data for understanding the genetics and evolution of social traits.

Acknowledgements

I would like to first thank the Zayed lab members, current and past, who have been there to listen to my ideas, to help with analysis, and to hit the town with occasionally. I would specifically like to thank Nadia, Alivia, and Katie. I wish you three the best; it's been a pleasure to share an office with you. Special thanks go to Clement. Clement welcomed me to Toronto and has been a constant source of inspiration both in and out of the lab. I would probably still be trying to load data into R if it weren't for his guidance. Clement, I have enjoyed talking about science, the markets, and science-fiction for the entirety of my degree. I do hope to continue it after—the next beer is on me. My visit to Brazil would not have been possible without my good friend Samir. It will always be a pleasure to talk to him about bees. Please, come back to Canada, Samir! At York, there are too many to thank for their support. I do I hope don't miss any: Ken Davey for his constant encouragement and inspiration, Laurence Packer for ensuring I knew about the other 19,999 + bees, Joel Shore and Bridget Stutchbury for their support in all forms through this process, Sheila Colla and Scott MacIvor for their advice along the way, Chris and Paul for being great friends, and Tamara Kelly for teaching me how to teach. Of course, I have to thank Amro. He has been the most encouraging supervisor, allowing me to explore whatever I want in the lab and helping me grow into a better academic (and person). My family has always supported me in this journey and I thank them that. Sorry I had to move across the country to take up this career. Love you all, thank you. I gained a wonderful family during my time at York. To the Buckleys and Watsons I say thank you for welcoming me with open arms. I'm sure the honey helped, so I promise to keep it flowing! Finally, I thank my partner Katey. Thank you for always supporting me, thank you for letting me work weekends, thank you for everything. It takes a special person to support a workaholic, and I love you for that and more.

Table of Contents

List of Tables

List of Figures

Chapter 1:

Sociobiology in the 'Omics Era: An Overview of Chapters to Follow

Brock A. Harpur

What more powerful form of study of mankind could there be than to read our own instruction book?

— Francis S. Collins, Address to the public following the completion of the first survey sequence of the human genome

Genome sequencing—the conversion of chemical genetic information into computationally-readable information—has revolutionized biology. As Francis Collins states above, sequencing provides a look into the instructions of the forms of life. These instructions come in the form of gene sequences and the organization of the genome as a whole. Variation in each of these can provide valuable information into the evolutionary history of a lineage. It was in the late 1990's that the scientific community sequenced *Haemophilus influenza* (Fleischmann et al. 1995), *Saccharomyces cerevisiae* (Goffeau et al. 1996), and *Caenorhabditis elegans* (Consortium 1998) and we got the first glimpses into how genomes varied among model species. Sequencing efforts continued and by the release of the first draft of the human genome in 2001 we were well into the 'Omics Era with dozens of genome sequences released in the early 2000's (Lander et al. 2001, Venter et al. 2001). Interest in sequencing, particularly from medical fields, drove innovations and cost reductions for faster and less expensive sequencing technologies (Mardis 2011). The exponential decline in costs has yielded an exponential increase in the number of new genomes (Mardis 2011). This also caused a shift in our attention from the sequencing of individual genomes within a species to sequencing many, allowing us to explore how variants within a genome contribute to phenotypic diversity.

Population Genomics and Eusociality

Here is where evolutionary biology has reaped a great benefit from newer sequencing technologies and modern techniques for analyzing data. With a population genomics approach—sequencing multiple individuals within a given population (or species)—one is able to form connections from genotype to fitness and from genotype to phenotype, connections at the core of Evolutionary Biology (Barrett and Hoekstra 2011). This approach has been successful in many model species; however, prior to the chapters I present here, no population genomic approaches had been used to explore adaptation and genetic variation within eusocial bees.

Our long fascination with eusocial species is a product of their unique life histories. A eusocial species is one with reproductive division of labour, corporative brood care, and overlapping generations (Michener 1974, Hölldobler and Wilson 1990). Within the Hymenoptera, reproductive individuals in a eusocial species (often called queens or gynes) are the primary egg-layers within a colony. Non-reproductive individuals (often called workers) perform all other aspects of colony upkeep; they are the brood care specialists, nest defenders, and foragers for a colony. Eusociality has evolved dozens of times independently within the Animalia (Bourke 2011). Darwin's interest in social insects was chiefly in understanding how the "neuters" (i.e. non-reproductive workers) evolved elaborate morphological and behavioural specializations when they do not reproduce. He called this problem his "special difficulty" and postulated that perhaps they were able to specialize by increasing the output of their queen (Darwin 1860, Herbers 2009). Darwin's idea was not formalized until the 1960's by W. D. Hamilton (Hamilton 1964a, b). It is surprising then that, there have been no population genomic analyses to ask where and how evolution has acted on a eusocial genome.

Population Genomics – Scanning for Evidence of Positive Selection

With a set of genomes, one can directly measure the levels of genetic diversity within a population (Nielsen 2005). Current levels of diversity in a population are the result of both neutral and non-neutral evolutionary processes (Kingman 2000). Neutral forces—genetic drift—influence diversity as a function of population size and demography and act across the genome as a whole. Non-neutral forces—such as positive selection—act on distinct loci that influence fitness (Nielsen 2005). Evidence of positive selection at a locus is therefore an indication that alleles at that locus contributed to fitness. Positive selection can be detected on a genome

because it leaves a distinct fingerprint surrounding selected loci (Nielsen 2005). Consider a new allele acted on by strong positive selection. Selection increases the frequency of this allele within a population over time, fixing it in the population (assuming that the population size is sufficiently large and that the stochastic influence of genetic drift on allele frequencies is sufficiently small). Sites linked to the selected allele will similarly be "dragged" to fixation reducing surrounding levels of polymorphism (Braverman et al. 1995). Selection can therefore be detected as regions of low polymorphism with high linkage disequilibrium. Further, selection will influence levels of population differentiation. For example, assuming populations experience different selective pressures over space, positive selection at a locus in one population will cause a change in frequencies of alleles at the locus between populations, leading to a high level of genetic differentiation (Nielsen 2005). Therefore, identifying loci with relatively high levels of genetic differentiation between populations can be a highly effective method to detect loci experiencing positive selection (Nielsen 2005, Nielsen et al. 2007, Zayed and Whitfield 2008, Harpur and Zayed 2013, Harpur et al. 2014b).

When acting over longer evolutionary times (i.e. between species) positive selection will act to fix non-synonymous mutations (e.g. mutations that change the amino-acid sequence) at a rate faster than the fixation of synonymous mutations (e.g. mutations that do not change the amino acid sequence) by random genetic drift. The McDonald-Krietman (MK) test makes use of this phenomenon and can highlight genes under selection (Mcdonald and Kreitman 1991). The MK test compares the ratio of synonymous and nonsynonymous polymorphisms within a species to fixed synonymous and nonsynonymous mutations between species. An excess of fixed vs polymorphic nonsynonymous mutations relative to synonymous mutations is indicative of positive selection rapidly fixing beneficial mutations. Adaptive change in allele frequency can be detected using the power of population genomics and the theoretical underpinnings of population genetics.

Which Genes Contribute to Fitness in Eusocial Lineages (Chapters 2 and 3)?

Using a population genomics approach and scanning for evidence of positive selection across two bee genera, I explored where and how evolution acts on genes within eusocial lineages. Hypotheses explaining the evolution of eusociality rest upon two key assumptions: 1)

mutations affecting the phenotype of sterile workers evolve by positive selection if the resulting traits benefit fertile kin, and 2) worker traits provide the primary mechanism through which social insects adapt to their environment (Linksvayer and Wade 2009). Despite the common view that positive selection drives phenotypic evolution of workers, we know very little about the prevalence of positive selection acting on the genomes of eusocial insects. It is also unclear if worker traits disproportionately contribute to fitness in all eusocial insects. Across Animalia, there is a spectrum of sociality with honey bees representing an extreme tail (Michener 1974, Hölldobler and Wilson 1990, 2009).

In Chapters 2 and 3, I identified which genes are positively selected across the genomes of *Apis* (Harpur et al. 2014b) and *Bombus* using genome sequence of 5 species and 65 individual genomes. I found little evidence that positive selection acts on a common set of genes or gene functions. Further, worker-expressed genes experienced differing degrees of positive selection within each lineage. Finally, I found that genes associated with honey bee worker were highly enriched for adaptive protein and *cis*-regulatory evolution. In addition to providing a unique insight into the process of adaptive evolution in social bees, Chapters 2 and 3 provide genomic datasets that have allowed researchers to study how specific genes and gene groups influence fitness in bees (Jasper et al. 2015, Kapheim et al. 2015, Kent and Zayed 2015, Rehan and Toth 2015, Vojvodic et al. 2015).

Population Genomics of Social Traits (Chapters 4, 5 and 6)

If a trait has a genetic basis it is possible to identify the genomic region(s) responsible and make a connection between genotype and phenotype. This practice has been of interest to geneticists for at least a century (Provine 1971), and is of growing interest to evolutionary biology, particularly since the Modern Synthesis (Barrett and Hoekstra 2011). Classically, connecting phenotype to genotype involved controlled crosses to associate the expression of a phenotype with the presence (or absence) of a marker genotype (Provine 1971, Lander et al. 2001). Population genomics allows for the association of genomic variants to phenotypic variation by correlating genotypes across the genome to phenotypic variation among (or within) populations (Csanadi et al. 2001, Colosimo et al. 2004, Visscher et al. 2012, Laine et al. 2014). When carefully applied, these methods can be useful tools to uncover associated genetic variants.

For example, there are associated variants for milk yield in cattle (Grisart et al. 2002), exploratory behaviour and armor plating in sticklebacks (Colosimo et al. 2004, Laine et al. 2014), muscle growth in pigs (Van Laere et al. 2003), and protein and oil production in soy (Csanadi et al. 2001).

For the next three chapters of my thesis, I used a combination of association mapping and population genomics to study the genetics and evolution of two social traits in honey bees: social immunity (Chapter 4) and colony defense (Chapters 5 and 6). The immune system of animals evolves rapidly (Weinstock et al. 2006, Sackton et al. 2007, Harpur and Zayed 2013). After nearly a century of investigation on model organisms, we have a deep understanding of which genes are expressed during an individual's immune response, how those genes interact with pathogens, and how those genes evolve through time (Sackton et al. 2007, Riddell et al. 2014). In contrast, we know less about the genetics and evolution of behavioural and social immunity. Social organisms can combat pathogens through individual innate immune responses or through social behaviours that limit transmission within groups — called social immunity (Cremer et al. 2007, Cremer and Sixt 2009).

Honey bees express an effective form of social immunity know as hygienic behaviour. Hygiene is the ability of nurse bees to detect and remove infected larvae from the comb. It is effective behaviour for limiting the spread of bacterial and fungal diseases that would otherwise kill a colony (Spivak and Gilliam 1998b, a, Spivak and Reuter 2001). Further, hygiene is highly heritable and known to be influenced by at least six regions within the genome (Oxley et al. 2010). It therefore provides a novel system to identify the genes underpinning social immunity and quantify how those genes have evolved through time.

I used high-resolution genomic comparisons of 40 colonies with varying levels of hygienic expression to find loci associated with hygienic behaviour. I confirmed that most of the loci mapped to previously identified quantitative trait loci (QTLs). However, my approach significantly narrowed these regions, allowing me to associate individual genes with hygienic expression. I found that genes associated with hygienic behaviour are involved in neuronal development and sensory perception; a finding that is in-line with previous mechanistic hypotheses for the trait. Finally, I found signs of adaptive evolution on the identified 'hygienic'

loci within the honey bee genus, supporting adaptive hypotheses for the evolution of social immunity in social insects.

I next explored which genes underpin colony defence. Honey bees provide a unique system to explore nest defence because there is variation in nest defence between subspecies; colonies of the least defensive subspecies will not sting a passerby while those of the most defensive can sting hundreds of times a minute (Winston 1987, 1992, Guzman-Novoa et al. 2002, Breed et al. 2004). I focussed on the Africanized honey bee—or the Killer honey bee as it has been called in media. The Africanized honey bee is among the most defensive honey bee populations, but colonies vary in defensive response (Winston 1987, 1992, Guzman-Novoa et al. 2002, Breed et al. 2004). We quantified defense response for 116 colonies in Brazil, and performed pooled-sequencing on 30 of the most phenotypically divergent samples. This yielded the largest available SNP data set for Africanized honey bees (Chapter 5; Kadri et al. 2016), a data set that will enable high-resolution studies of the population dynamics, evolution, and genetics of this successful biological invader. In addition to facilitating the development of SNP-based tools for identifying Africanized honey bee (Chapter 5; Kadri et al. 2016). Because the samples we sequenced varied for defense response I was able to scan the genomes of more- and less-defensive colonies and identify regions that had large differences in allele frequency between the two groups. I uncovered a set of genes associated with defense response within the Africanized honey bee, genes that overlapped with previously-reported sets of QTLs.

Mobilizing Genomic Data for Industry (Chapter 7)

The scientific advances described above demonstrate the value of population genomics approaches, but can genomic data be mobilized for use in applied beekeeping? Human genomic data have been used to create personalized medicine and identify mutations associated with inherited conditions (Hamburg and Collins 2010). Agricultural genomics is poised to allow breeders to select more effectively for desired traits (Hiendleder et al. 2005, Tieman et al. 2017).

Using the genomic resources I generated in Chapter 2 (Harpur et al. 2014b), I created a tool for beekeepers to explore genetic diversity within their own colonies and across Canada (Harpur et al. 2015). I used a citizen science approach that engaged a diverse group of beekeepers across the country. Beekeepers around the country sent a total of 855 worker honey

bees that I genotyped at 91 ancestrally-informative single nucleotide polymorphisms (SNPs). With this data set, I found low levels of genetic differentiation within Canada and small but significant differences in ancestry among provinces. Honey bee populations in Northern and Western Canada were more closely related to subspecies from Southern and Mediterranean Europe. I attributed this pattern to differences in importation practices within Canada. Finally, I was able to accurately discriminate between Africanized bees and Canadian bees using the ancestrally-informative SNPs, supporting the use of SNPs for accurately detecting Africanized honey bees and providing valuable insights into the genetic structure of Canadian bees, all while engaging beekeepers in the scientific process.

Chapter 2:

Population genomics of the honey bee reveals strong signatures of positive selection on worker traits

Brock A. Harpur, Clement F. Kent, Daria Molodtsova, Jonathan M. D. Lebon, Abdulaziz S. Alqarni, Ayman A. Owayss, and Amro Zayed[1]

Introduction

Eusocial behavior evolved multiple times in insects and is characterized in part by extreme asymmetries in the reproductive potential of individuals (Wilson and Holldobler 2005). This asymmetry is most pronounced in advanced eusocial insects with their fertile queen and sterile worker castes. Darwin first recognized that natural selection cannot directly optimize worker phenotypes because workers are usually sterile (Darwin 1860). Hamilton developed kin selection theory to describe the conditions that allow natural selection to indirectly optimize worker phenotypes if such phenotypes benefit their fertile kin (Hamilton 1964a, b). It is commonly believed that worker traits such as sib-care, foraging, and colony defense play important roles in allowing colonies to adapt to their environment (Wilson 1985, Sagili et al. 2011, Wray et al. 2011). However, despite the central role of kin-selection and inclusive fitness theory in the field of Sociobiology (Abbot et al. 2011, Strassmann et al. 2011), we lack knowledge on the pattern and prevalence of positive selection acting on the genomes of eusocial insects.

Population genomic studies provide unprecedented opportunities to detect signatures of selection on DNA sequences over different timescales (Begun et al. 2007). There are several tests of selection that can be applied to genome-wide datasets. The McDonald-Kreitman test is arguably the best method for detecting selection on protein coding sequences because of its robustness to changes in a species' demography, which often confounds other tests of selection

[1] This published manuscript has been reprinted with the permission of its co-authors and publisher from the original manuscript: Harpur BA*, et al.* (2014) Population genomics of the honey bee reveals strong signatures of positive selection on worker traits. *P Natl Acad Sci USA* 111(7):2614-2619.

(Begun et al. 2007). A recent Bayesian implementation of this classic test utilizes genome-wide estimates of polymorphism and divergence to improve statistical power (Eilertson et al. 2012). Outlier tests of selection are also less sensitive to population demography, which affect all loci within a genome; loci under selection thereby appear as outliers in the empirical distribution of genome-wide data (Akey et al. 2002, Nielsen 2005, Qanbari et al. 2012). In spatially structured populations, outlier tests of genetic differentiation are especially useful in identifying loci underlying local adaptation (Begun et al. 2007, Hohenlohe et al. 2010).

The honey bee, *Apis mellifera,* provides an ideal system for applying population genomics to understand the evolutionary forces shaping eusocial insect genomes. The honey bee is arguably the most well-known social insect at the level of behavior, physiology, and genetics, and there are many rich datasets that detail caste-specific transcriptomic and proteomic phenotypes (Chandrasekaran et al. 2011, Chan et al. 2013). The bee genome is relatively small (236 Mb) and lacks many repetitive elements (Weinstock et al. 2006) making assembly via short-read sequencing highly feasible. Finally, the honey bee's genetically and phenotypically distinct population groups in Africa, Asia, and Europe (Whitfield et al. 2006a) provide an opportunity to examine how the honey bee genome adaptively diverged in response to the different selective pressures experienced across its large and diverse native range (Zayed and Whitfield 2008, Chavez-Galarza et al. 2013).

To this end, we undertook a comprehensive population genomic study of the honey bee by sequencing the genomes of 40 individual bees from different geographic regions, including a closely related species. Our goals were to first identify genomic regions with signs of positive selection and then examine the degree to which genes associated with worker traits contribute to adaptive evolution. Our study provides unparalleled insights on the genes and traits underlying adaptation in social insects.

Results

Genomic diversity in *Apis mellifera*. We sequenced the diploid genomes of *A. mellifera* workers sampled from the four genetically distinct honey bee lineages (Whitfield et al. 2006a)in Africa (N=11 workers), Asia (N=10), East Europe (N=9) and West/Northern Europe (N=9) at an average coverage of 38X. We also sequenced a single *A. cerana* worker as an outgroup. We

conducted preliminary Sanger sequencing of several randomly chosen exons to ensure that our collected specimens were not admixed (Harpur et al. 2012). We discovered 12,041,303 single nucleotide polymorphisms (SNPs) in the 39 sequenced *A. mellifera* genomes, many of which were validated using independent datasets. We used the discovered SNPs to confirm the population structure of the sampled bees. As expected, the 39 *A. mellifera* workers were assigned to four distinct populations and our sampled bees had very low levels of admixture (Fig. 2.S1). Given that human management increases admixture levels in honey bees, the non-admixed bees studied herein provide the best approximation of the four *A. mellifera* evolutionary lineages prior to human management (Harpur et al. 2012).

Signatures of positive selection over intermediate timescales. We used a Bayesian implementation of the McDonald-Kreitman (MK) test (Eilertson et al. 2012) to estimate the strength and direction of selection on 12,303 genes since divergence between *A. mellifera* and *A. cerana* approximately 5 to 25 MYA (Arias and Sheppard 2005, Kotthoff et al. 2013). The MK test requires polymorphism data from at least one species (i.e. *A. mellifera*) and divergence data from at least one outgroup sequence (i.e. *A. cerana*) (Bustamante et al. 2005, Begun et al. 2007, Hartl and Clark 2007), and the Bayesian implementation of the MK test allows for the estimation of the population size-scaled selection coefficient γ on replacement mutations (Eilertson et al. 2012). Although the MK test is very robust to changes in population demography (Eilertson et al. 2012), we conservatively implemented this test using the polymorphism data from African bees only, which represent a large stable population that is minimally impacted by human management (Kent et al. 2011, Harpur et al. 2012). We found that most genes in the bee genome (ca. 90%) have γ between -1 and 1 (Fig. 2.1A). 0.9% of genes have $\gamma < -1$ consistent with strong purifying selection, while 9.3% of genes have $\gamma > 1$ consistent with strong positive selection.

Signatures of positive selection over short timescales. Positive selection facilitating local adaptation creates loci with outlier levels of genetic differentiation (F_{ST}) relative to the rest of the genome (Akey et al. 2002, Nielsen 2005). We used outlier levels of F_{ST} to identify loci that have likely experienced geographically restricted positive selection since divergence of *A. mellifera*'s four evolutionary lineages approximately 1 MYA to 11,000 years ago (Arias and Sheppard 2005,

Kotthoff et al. 2013). We used two approaches to detect genomic windows (≥5 kb) and SNPs with outlier levels of F_{ST} (Fig. 2.1B) in the six pairwise population comparisons between the four bee lineages. The two approaches were highly concordant: Outlier SNPs were significantly enriched within outlier windows (Fisher's exact test; $p<2.2x10^{-16}$ for all pairwise comparisons) and, on average, 55.5% of SNPs within outlier windows were themselves outlier SNPs. We detected an average of 5,715 outlier windows with extreme levels of genetic differentiation in the six pairwise population comparisons. Outlier SNPs contained alleles that were either nearly or completely fixed in pairwise population comparisons (F_{ST} ranged from 0.89 to 1). We found that SNPs with outlier F_{ST} in *A. mellifera* occur mostly in putative *cis*-regulatory regions: 18.5% of SNPs found 500 bp upstream of genes are outliers relative to 12.3% in exons, 8.5% in introns and 8.6% in intergenic regions. However, there is still a considerable amount of positive selection acting on protein sequences: 11% of nonsynonymous mutations were outlier SNPs and outliers SNPs were enriched for nonsynonymous SNPs (Fisher-Exact test, $p<2.2x10^{-15}$).

Biological significance of loci underlying positive selection. We used Gene Ontology (GO) tools (Huang et al. 2009) to investigate the possible function of adaptively evolving loci. Genes associated with G-protein coupled receptors (GPCRs) and GPCR-signaling were enriched among adaptively evolving protein and regulatory loci over intermediate and short timescales. GPCRs translate sensory inputs into cellular responses and are thus crucial for tuning an organism's physiology and behavior in response to the environment; this is particularly intriguing given the degree to which pheromones within a colony affect the biology of the different honey bee castes. We also found many annotation clusters enriched among adaptively evolving loci, including genes associated with adult behavior, cognition, nervous system development, metabolism, and steroid hormones (see Online Supplemental Materials (Harpur et al. 2014b)).

Selection on taxonomically restricted genes. The gene content of genomes is dynamic over evolutionary time, and genomes contain both 'old' genes and 'new' genes. Old genes originated in an evolutionary-distant common ancestor and orthologous copies are found across many distant taxa, while new genes originated recently and are found only in specific taxonomic groups. Taxonomically-restricted genes (TRGs) have been the subject of recent attention as they

are predicted to be drivers of phenotypic evolution (Chen et al. 2013). The genomes of social insects harbor many TRGs, which are hypothesized to play an important role in the elaboration of sociality (Simola et al. 2013). TRGs in ants (Feldmeyer et al. 2013), bees (Johnson and Tsutsui 2011) and wasps (Ferreira et al. 2013) tend to show, on average, worker-biased expression, which suggests that they play an important role in the evolution of worker phenotypes. We used the hierarchical catalogue of orthologs in OrthoDB v.6 (Waterhouse et al. 2013) to classify honey bee genes to four mutually exclusive groups: *Apis*-restricted, Apoidea-restricted, and Hymenoptera-restricted genes, as well as genes found in honey bees and at least one other insect order . We then asked if TRGs exhibit differences in adaptive protein evolution over intermediate timescales.

We found a significantly higher proportion of *Apis*-restricted, Apoidea-restricted, and Hymenoptera-restricted genes with signs of strong positive selection ($\gamma > 1$) relative to genes found in other insects; 20.4% of *Apis* genes (N = 88), 21.8% of Apoidea genes (N = 215) and 15% of Hymenoptera genes (N = 1,321) have $\gamma > 1$ relative to 9% of genes found in other insects (N = 8,686; Chi^2 $p < 0.0003$ for all tests comparing *Apis,* Apoidea and Hymenoptera genes relative to genes found in other insects). Further. Apoidea-restricted genes have a significantly higher proportion of genes with $\gamma > 1$ relative to Hymenoptera-restricted genes (Chi^2, $p = 0.025$). We also found that among *A. mellifera* genes with signs of positive selection ($\gamma > 0$), those found in all insects had the lowest average γ, those found in the Hymenoptera had intermediate average γ, and those found in the Apoidea had the highest average γ. *Apis*- and Apoidea-specific genes did not differ with respect to γ, but the differences between these two groups (i.e. *Apis* + Apoidea) and Hymenoptera- and Insect genes were highly significant (Wilcoxon test, $p<10^{-10}$). Average γ for Apoidea was more than three times higher than γ for genes found in all insects (Fig. 2.2A). We also observed differences in the prevalence of negative selection ($\gamma < 0$) among TRGs, with *Apis*-specific genes having significantly stronger purifying selection relative to Hymenoptera-specific genes (Fig. 2.2B; Wilcoxon test, $p<0.01$).

Adaptive evolution of queen-biased and worker-biased proteins. We investigated the degree to which worker and queen phenotypes contribute to colony fitness by examining if proteins with caste-biased expression show differences in the prevalence of positive selection. We used a list

of caste-biased proteins from the Honey Bee's Protein Atlas (Chan et al. 2013), which provides quantitative proteomic data for 26 tissues assayed in queens and workers. Most honey bee proteins are expressed in both queen and worker tissues. We obtained γ estimates for 90 and 79 proteins that were identified as significantly worker- or queen-biased based on average whole body expression (Chan et al. 2013); these proteins show consistent caste-biased expression in *most* of the 26 tissues in the Honey Bee Protein Atlas. Although few in number, caste-biased proteins provide an objective way to identify sets of genes that are relevant to caste-biased phenotypes. We make the reasonable assumption that the evolution of worker-biased proteins is mostly shaped by forces acting on worker phenotypes, and that the evolution of queen-biased proteins is mostly shaped by forces acting on queen phenotypes. We found that worker-biased proteins had a significantly higher γ relative to queen-biased proteins (Workers: Avg. γ=0.77; Queens: Avg. γ=0.42; Wilcoxon test; $p< 0.0016$). Proteins that were not differentially expressed between queens and workers (N=1,095) are expected to have the greatest levels of pleiotropy and constraint (Fisher 1930), and indeed they have significantly lower γ relative to worker-biased and queen-biased proteins (Wilcoxon test; $p< 0.013$) (Fig. 2.3A). We also found that worker-biased proteins were enriched for signatures of local adaptation. When benchmarked against non-differentially expressed proteins, we found that worker-biased proteins showed a greater enrichment of nonsynonymous outlier SNPs in more tissues and over a larger number of pairwise lineage comparisons relative to queen-biased proteins (Fisher's exact test: p=0.0005).

Worker traits and colony fitness. We investigated if genes that are *a priori* known to influence worker phenotypes showed signatures of positive selection:

Worker brain gene expression and behavior: There is strong evidence that shifts in brain gene expression mediate shifts in behavior in workers (Whitfield et al. 2006b, Chandrasekaran et al. 2011, Zayed and Robinson 2012). Given the considerable and possibly adaptive differences in worker behavior between the honey bee's four evolutionary lineages (Winston 1987), we predicted that differentially expressed genes (DEGs) associated with worker behavior would be enriched for signs of adaptive divergence. We queried 27 microarray experiments from the BeeSpace project that assayed the brain transcriptomes of nearly 1,000 workers across several natural or experimentally-induced behavioral states (reviewed by Chandrasekaran et al. 2011,

Zayed and Robinson 2012). We found that DEGs associated with 23 out of 27 behavioral states have regulatory regions with significantly more outlier SNPs than expected by chance in at least one pairwise population comparison after correcting for multiple tests (FDR, α=0.001; $P<2.2x10^{-4}$; Fig. 2.S3). 18 out of 27 behavioral states were enriched for coding sequences with significantly more nonsynonymous outlier SNPs than expected by chance (FDR, α=0.001; $P<2x10^{-6}$; Fig. 2.S3). The enrichment of outlier loci in DEGs across most BeeSpace experiments indicates that genes associated with worker behavior are enriched for signatures of positive selection underlying local adaptation.

Worker division of labor: Worker honey bees undergo an age-related division of labor that allows them to transition from in-hive tasks to foraging and colony defence over time. This division of labor is regulated through an unusual interaction between the egg yolk protein Vitellogenin (Vg), JH and JH-signalling, and insulin-like/TOR signalling (Sullivan et al. 2000, Amdam and Omholt 2003, Nelson et al. 2007, Ament et al. 2008, Ament et al. 2010, Wang et al. 2010, Ament et al. 2012) (Fig. 2.3B). The mutually repressive relationship between Vg and JH is unique to worker honey bees prompting researchers to hypothesize that these conserved genes and signalling pathways were co-opted via natural selection to regulate worker division of labor in *Apis* (Amdam and Omholt 2003, Amdam et al. 2004). Vg was previously shown to be under positive selection based on analysis of several exons (Kent et al. 2011) and our complete analysis shows its selection coefficient to be even higher than previously reported (γ = 4.97 vs. 1.88). Vg in turn regulates the central insulin/Tor growth pathway (Corona et al. 2007) and both of the bee's insulin receptors and the Depdc5 gene – part of a complex which sensitizes Tor signalling to cellular amino acid levels (Bar-Peled et al. 2013) – are under positive selection. Juvenile hormone acid methyltransferase (*Jhamt*) and juvenile hormone esterase (*Jhe*) are the proximal biosynthetic and catabolic enzymes for juvenile hormone (Jindra et al. 2013), and Met/Gce2 is the key cofactor in juvenile hormone receptor complexes (Li et al. 2011, Bernardo and Dubrovsky 2012); all of these are under significant and strong positive selection. We also investigated if *foraging* (Ben-Shahar et al. 2002) and *malvolio* (Ben-Shahar et al. 2004) – both implicated in worker division of labor – experience positive selection. We had previously estimated that *foraging* experiences nearly-neutral evolution based on analysis of 4 exons (Kent et al. 2011), but our complete analysis herein indicated that *foraging* experiences positive

selection ($\gamma = 0.99$). The gene *malvolio* on the other hand appears to be constrained ($\gamma = -0.33$). Given their causal involvement in regulating worker division of labour, signatures of selection on the above-mentioned genes (Fig. 2.3B) supports the hypothesis that worker division of labor has major influence on colony fitness.

Major Royal Jelly Proteins: Workers have specialized hypopharyngeal glands that are used to synthesize royal jelly for feeding nestmates (Winston 1987). The honey bee genome contains several genes that encode Major Royal Jelly Protein (Mrjp), and most of these genes are highly expressed in the hypopharyngeal glands of workers (Drapeau et al. 2006). The eight Mrjp genes studied herein had significantly higher gamma relative to other genes (Wilcoxon test, $p=0.0015$) and 3 out of 8 genes had $\gamma > 2$ (binomial $p = 0.00003$) indicating high levels of positive selection. This list included Mrjp1 (*royalactin*) which is essential for inducing queen-worker differentiation (Kamakura 2011). We also detected significant signs of positive selection on Mrjp4 and Mrjp7, which are known to be expressed only in workers and not in any other caste or developmental stage (Drapeau et al. 2006).

Discussion

The honey bee is a model eusocial organism and our analyses provide novel insights on the process of adaptive evolution in social insects. We found strong evidence of positive selection acting on protein coding sequences in the honey bee. The highest levels of selection were observed in genes that were taxonomically restricted to bees, while Hymenoptera-specific genes had intermediate levels of selection. The fact that Apoidea-specific genes had similar selection coefficients relative to *Apis*-specific genes suggests that adaptive evolution in the social honey bee is partially fueled by novel genes that were found in solitary ancestors. Although there is evidence that sociality evolved by co-opting conserved genetic toolkits (Toth and Robinson 2007), our results, along with others (Simola et al. 2013), suggest that taxonomically restricted genes play an important and disproportionately large role in the adaptive evolution of social insects. Additionally, we uncovered a substantial amount of adaptive regulatory sequence evolution when contrasting differences in allele frequency between the four honey bee lineages studied herein. Our results, along with recent findings of rapid evolution of transcription factor

binding sites in social insects (Simola et al. 2013), suggests that *cis*-regulatory changes play an important role in the evolution of insect societies.

The fitness of a colony is determined by the traits of fertile members who monopolize reproduction and by the traits of sterile workers who build and maintain the colony, feed the queen and the brood, collect food and resin, maintain temperature homeostasis, and sacrificially defend the colony against intruders (Winston 1987). It is often thought that worker behavior and phenotypic plasticity provide the primary mechanism that allows insect colonies to adapt to their environment (Wilson 1985, Sagili et al. 2011, Wray et al. 2011), and our population genomic data support this view. We showed that proteins with worker-biased expression have significantly higher selection coefficients relative to queen-biased proteins. We also showed that genes with known effects on worker division of labor and genes associated with nursing brood to be under strong positive selection in honey bees. Further, we showed that genes associated with worker behavior and behavioral plasticity, based on extensive studies of brain gene expression, were enriched for signatures of adaptive *cis*-regulatory and protein evolution.

It was previously shown that genes with worker-biased brain expression have lower rates of protein evolution relative to queen-biased genes based on analysis of *Apis* and *Nasonia vitripennis* alignments (Hunt et al. 2010); a result that is apparently inconsistent with our finding of higher rates of adaptive evolution of worker-biased proteins. However, our study used a more comprehensive database of caste-biased proteins (i.e. proteomic differences assayed in 26 tissues versus transcriptomic differences assayed in one tissue), included TRGs that we have shown to experience higher rates of natural selection (i.e. *Apis*-*Nasonia* alignments would have excluded *Apis*- and Apoidea-specific genes), and directly quantified adaptive evolution (Eilertson et al. 2012) (i.e. general measures of protein evolution (Hunt et al. 2010) are affected by both adaptive, neutral, and non-adaptive causes (Harpur and Zayed 2013)). Our population genomics study strongly indicates that worker transcriptomic and proteomic phenotypes are enriched for signatures of positive selection.

Workers honey bees are effectively sterile but they can produce haploid sons in queenless colonies. Given the rarity of worker reproduction under queenright conditions (Visscher 1989), the lower number of drones produced by queenless colonies relative to queenright colonies (Page and Erickson 1988), and lack of evidence showing that worker-laid drones have similar fitness as

queen-laid drones, it is reasonable to assume that indirect kin-selection is mostly responsible for the adaptive evolution of worker traits. Recent theory suggests that, all other factors being equal, indirect selection on workers will be effectively weaker than direct selection on queens (Linksvayer and Wade 2009), especially when queens are polyandrous as in *A. mellifera* (Hall and Goodisman 2012). However, our work shows that indirect selection does not necessarily impede adaptive evolution of the worker caste, possibly because mutations in worker-biased genes tend to – on average – have higher colony-level fitness effects.

Our study presents – to our knowledge – the first map of positive selection for a social insect. The field of genomics has greatly enriched research in sociobiology by providing knowledge on the molecular basis underlying caste differentiation and caste-specific phenotypes. Our population genomics approach allowed us to identify loci that affect fitness in honey bees – "the alleles that matter!" (Rockman 2012). We have shed some light on the biological and social relevance of such loci but more studies are needed to understand the molecular and phenotypic basis of adaptation in honey bees (Barrett and Hoekstra 2011). We believe that the rich genomic resources provided herein will be instrumental in developing and testing mechanistic and evolutionary-explicit models of how and why social behavior evolves.

Materials and Methods

Sequencing, alignment, and SNP calling. Genomic DNA was extracted from each bee using a DNeasy Blood & Tissue Kit from Qiagen, and sent for Illumina Hi-Seq sequencing (50bp reads) at Génome Québec Innovation Centre at McGill University. Each bee was sequenced in a single Hi-Seq lane. We developed the following bioinformatic pipeline: 1) FASTQ files were initially aligned to the *A. mellifera* genome assembly AMELv4.5 using the default parameters of BWA and alignments were then imported into SAMtools (Li et al. 2009) in BAM format. 2) We remapped each bee's sequence using Stampy (Lunter and Goodson 2011) at a substitution rate of 0.02 to better align divergent sequences. 3) We subsequently re-aligned sequences with GATK's RealignerTargetCreator followed by IndelRealigner to reduce any potential erroneous alignments close to indels (McKenna et al. 2010). We detected SNPs and created variant calling files (i.e. VCF) using mpileup (–Q 20 option), bcftools (mutation rate of 0.05), and varfilter (–d 3 –Q 15 –D64) (Li et al. 2009). We filtered out highly repetitive regions and recently duplicated genes from our analyses by first performing a blastn match of 50bp segments of the *A. mellifera*

genome back to the reference genome; we excluded any 50bp segment matching two or more locations with fewer than 3 mismatches and blastn E-value of 2E-20. An average of 3.2% of SNPs were masked with this protocol. We also excluded 6.47 Mb of sequence from unmapped scaffolds (scaffolds 17.2000 and above in AMELv4.5) because of low sequencing coverage in these small (mean 1,957bp) and gene-poor scaffolds. We aligned the *A. cerana* sequences to the reference *A. mellifera* genome using the same methods as above, except we set the Stampy divergence threshold to d=0.05. Overall, we were able to study genetic diversity in 227.6 Mb (~96%) of *A. mellifera*'s genome, and the *A. mellifera* workers had an average coverage depth of 38X. Five researchers manually examined over 100 kb of sequence to ensure the accuracy of our alignment and SNP calls.

Validation of SNPs. We used three datasets to validate the SNPs discovered herein (NGS SNPs). 1) Some of the bees analyzed by us were previously used to sequence several nuclear genes using Sanger technology (Kent et al. 2011, Harpur et al. 2012, Kent et al. 2012, Harpur and Zayed 2013). We compared 270 different Sanger sequences covering 169,791 bp to our NGS dataset: 97% of sequences had identical numbers of SNPs. 2) we compared 1,088,415 SNPs from the reference *A. mellifera* genome (Weinstock et al. 2006) to NGS SNPs: 88% of the SNPs were present in our dataset, either as SNPs (82.2%) or as indel polymorphisms (5.8%). 3) We also validated 85% of SNPs derived from sequencing Africanized honey bees (Weinstock et al. 2006). Given the large level of genetic diversity in honey bees, we do not expect to find a high (>95%) correspondence between NGS SNPs. The large level of validation reported herein, especially when comparing NGS and Sanger sequences derived from the same bees (97% validation), indicates that the vast majority of SNP calls are accurate.

Population structure. We utilized the program ADMIXTURE (Alexander et al. 2009) to estimate the population origin and admixture levels of the sequenced bees. We tested K=1-6 populations (100 times per K) assuming no prior knowledge of population origin. We randomly selected 25,000 SNPs separated by at least 5 KB from across the genome; singleton SNPs (i.e. derived allele present in a single bee) were excluded from this analysis. We repeated this analysis with three sets of 25,000 randomly chosen SNPs to test the robustness of ADMIXTURE results.

McDonald-Kreitman (MK) Analyses. We employed a Bayesian implementation of the MK test (Eilertson et al. 2012) to estimate the prevalence of selection acting on genes. We used perl scripts to determine if SNPs were nonsynonymous or synonymous using predictions from the bee's official gene set (OGSv3.2). Divergence data was based on fixed mutations between *A. cerana* and *A. mellifera* sequences. We restricted our MK analysis to genes with sequence coverage in all African bees. We employed the following measures to guard against spurious alignment of coding sequences in *A. mellifera* and non-coding sequences in *A. cerana*: 1) We used expression data derived from RNA sequencing of *Apis cerana* worker brains (Wang et al. 2012) to mask portions of *A. mellifera* exons that have no evidence of expression in *A. cerana*. 2) We checked all coding-sequence alignments for the presence of frame-shifting indels. When we discovered a frame-shifting indel in an exon, we excluded the downstream sequence of that exon. Genes with no SNPs were excluded from analyses.

Outlier SNPs and Windows. F_{ST} was estimated for all six pairwise comparisons involving the four sampled *A. mellifera* populations following Weir and Cockerham (Weir and Cockerham 1984) as implemented in GENEPOP v4.2 (Raymond and Rousset 1995). Weir and Cockerham's method provides accurate estimates of F_{ST} given uneven and/or small sample sizes (Willing et al. 2012). In each population comparison, SNPs with a minimum allele frequency < 0.025 and SNPs not meeting our masking criteria were excluded from analysis. We used two independent methods to identify loci and regions with outlier levels of F_{ST}. First, we classified any SNP in the top 5% of the empirical distribution of F_{ST} as an outlier. Across our dataset, outlier SNPs were significantly differentiated based on exact G-tests (Goudet et al. 1996) ($q << 10^{-8}$ after FDR correction). Second, we utilized a creeping window algorithm (Qanbari et al. 2012) that estimates mean F_{ST} for overlapping 5 kb windows containing at least 30 SNPs. Analyses were also performed with 7 and 10 kb windows and results remained consistent across the different window sizes. To avoid estimating F_{ST} across sequence gaps, windows with SNPs spaced greater than 5 kb apart were skipped (Qanbari et al. 2012). For the creeping window approach, outlier windows were statistically identified using simulation as follows: 1) we re-scanned the genome 10 million times and randomly sampled new F_{ST} values for every SNP in a given window (Qanbari et al. 2012), 2) windows were deemed outliers if observed average F_{ST} in a window was

above the 95th percentile of the empirical distribution of expected F_{ST}, following stringent FDR correction ($q<0.025$) (Storey and Tibshirani 2003). Within a range of overlapping windows, only the most significant window was considered an outlier. Because the two methods of detecting outlier loci were highly concordant (see text), we used the first method for most analyses because it allowed us to precisely determine genomic context (i.e. coding vs. noncoding) of outlier loci. All F_{ST}-based analyses were performed on each pairwise population comparison (n=6) and corrected for multiple testing using FDR (Benjamini and Hochberg 1995).

GO analysis. We used the program DAVID 6.7 (Huang et al. 2009) to examine if adaptively evolving loci are enriched for specific functional annotation clusters using default parameters. We first identified the *Drosophila* homologs of positively selected bee genes using blastp match (evalue threshold 1e-10). We were able to find fly homologs for 54.3% of genes in OGSv3.2.

Bee Protein Atlas. The Honey Bee Protein Atlas (Chan et al. 2013) provides protein expression data in 26 tissues in queens and workers for 1,728 proteins in OGSv3.2. We examined if significantly worker-biased proteins, averaged across the different tissues (Chan et al. 2013), have different γ relative to significantly queen-biased proteins using a Wilcoxon non-parametric test. We also counted the number of cases where worker-biased proteins were enriched for nonsynonymous outlier SNPs relative to all proteins found in a given tissue; this analysis was repeated for 26 tissues and for each of the 6 pairwise population comparisons (a total of 156 tests). We performed a similar analysis for queen-biased genes. After first ensuring that queen-biased and worker-biased proteins did not significantly differ in length, we compared the number of significant and non-significant (FDR; $\alpha<0.05$) tests of enrichment in worker-biased and queen-biased proteins using a Fisher's Exact test. The Bee Protein Atlas also provided a proteomic contrast of drones and workers. Worker-biased proteins had higher selection coefficients relative to drone-biased proteins but the number of drone-biased proteins was too small to warrant a statistical analysis.

BeeSpace Project. We obtained lists of DEGs in the brains of worker honey bees from 27 microarray experiments targeting several aspects of worker behavior associated with behavioral

maturation, foraging, and aggression (Chandrasekaran et al. 2011, reviewed by Zayed and Robinson 2012). We compared the number of outlier SNPs in putative *cis*-regulatory sequences (i.e. 500 bp upstream of start codon), and the number of outlier nonsynonymous SNPs in exons, in DEGs and non-DEGs for each of the 27 experiments. Across the experiments, DEGs were not significantly longer than non-DEGs, and thus enrichment of outlier SNPs in the exons of DEGs was not caused by differences in gene length.

Statistical Analyses and Power. All statistical analyses were performed in R (Team 2011). All comparisons were performed with non-parametric tests unless otherwise stated. FDR corrections were based on the methods of either Benjamini–Hochberg (α values reported) (Benjamini and Hochberg 1995) or Storey (q values reported) (Storey and Tibshirani 2003); the latter was used when the number of statistical tests was large. We employed appropriate samples sizes for estimating γ (20 haploid chromosomes from Africa) (Andolfatto 2008) and F_{ST} (18 to 20 haploid chromosomes per population) (Willing et al. 2012).

Figure Legends

Figure 2.1. Loci underpinning adaptive evolution in honey bees. Histograms of (A) the population-size scaled selection coefficient (γ) for 12,303 genes, and (B) pairwise genetic differentiation (F_{ST}) between African and West European honey bees for 3,392,632 SNPs. F_{ST} histograms for the other 5 pairwise comparisons are found in Figure 2.S2. Areas in red represent outlier loci with signatures of adaptive evolution.

Figure 2.2. (A) Taxonomically restricted genes have higher rates of adaptive evolution. For genes with signs of positive selection ($\gamma > 0$), γ is significantly higher in *Apis*-restricted and Apoidea-restricted genes, intermediate in Hymenoptera-restricted genes, and lowest for genes found in other insect orders. (B) For genes with signs of negative selection ($\gamma < 0$), *Apis*-restricted genes have the highest levels of negative selection. Error bars denote SEM.

Figure 2.3. Genes associated with worker phenotypes show signs of adaptive evolution in honey bees. (A) Worker-biased proteins have significantly higher selection coefficients relative to queen-biased proteins, and non-differentially expressed proteins (NDEG). (Error bars denote SEM; **=$p<0.01$) (B). Genes causally associated with worker division of labor have very high selection coefficients in the honey bee.

Supporting Figure Legends

Figure 2.S1. Population structure and ancestry for bees used in this study. We tested K=1-6 populations (100 times per K) assuming no prior knowledge of the population origin using 25,000 randomly selected SNPs from the bee genome (See Methods). The dataset best fit a model with four distinct populations (A, Y, M, and C; statistics from a single run: K=4, CV error =0.18078, LL= -891627). Each column represents the relative ancestry of each sampled bee to the 4 populations delineated by the Bayesian analysis. The sampled bees were very pure, and on average, each bee had a 99.39% ancestry to its inferred population.

Figure 2.S2. Histograms of pairwise genetic differentiation (F_{ST}) between all populations (a) C vs. M (b) C vs. Y (c) C vs. A (d) M vs. Y (e) A vs. Y. Areas in red indicate SNPs with high F_{ST} values (>95% of data).

Figure 2.S3. Enrichment of F_{ST} outlier SNPs in (a) 500bp regions upstream of genes and (b) nonsynonymous SNPs in exons for 27 experiments performed in the BeeSpace project. The 27 experiments were labeled according to the behavioral / genotypic contrast (see ref. 17 in paper). Afr = Africanized bees; Eur = North/South American bees, likely of European descent; C = East European bees; M = West/Northern European Bees.

Figure 2.1

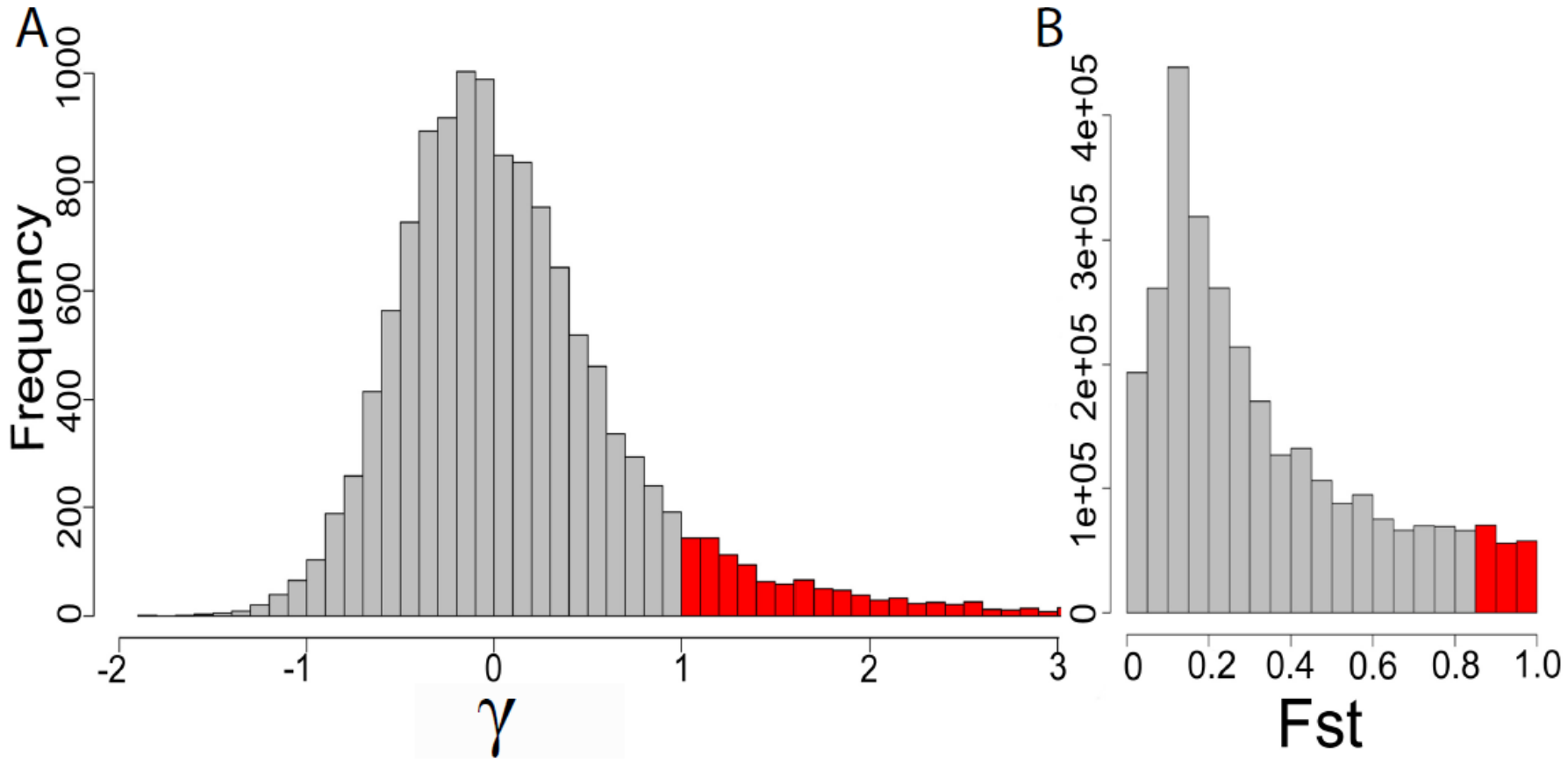

A
Frequency
0
200
400
600
800
1000
-2
-1
0
1
2
3
γ
B
0
1e+05
2e+05
3e+05
4e+05
0
0.2
0.4
0.6
0.8
1.0
Fst

Figure 2.2

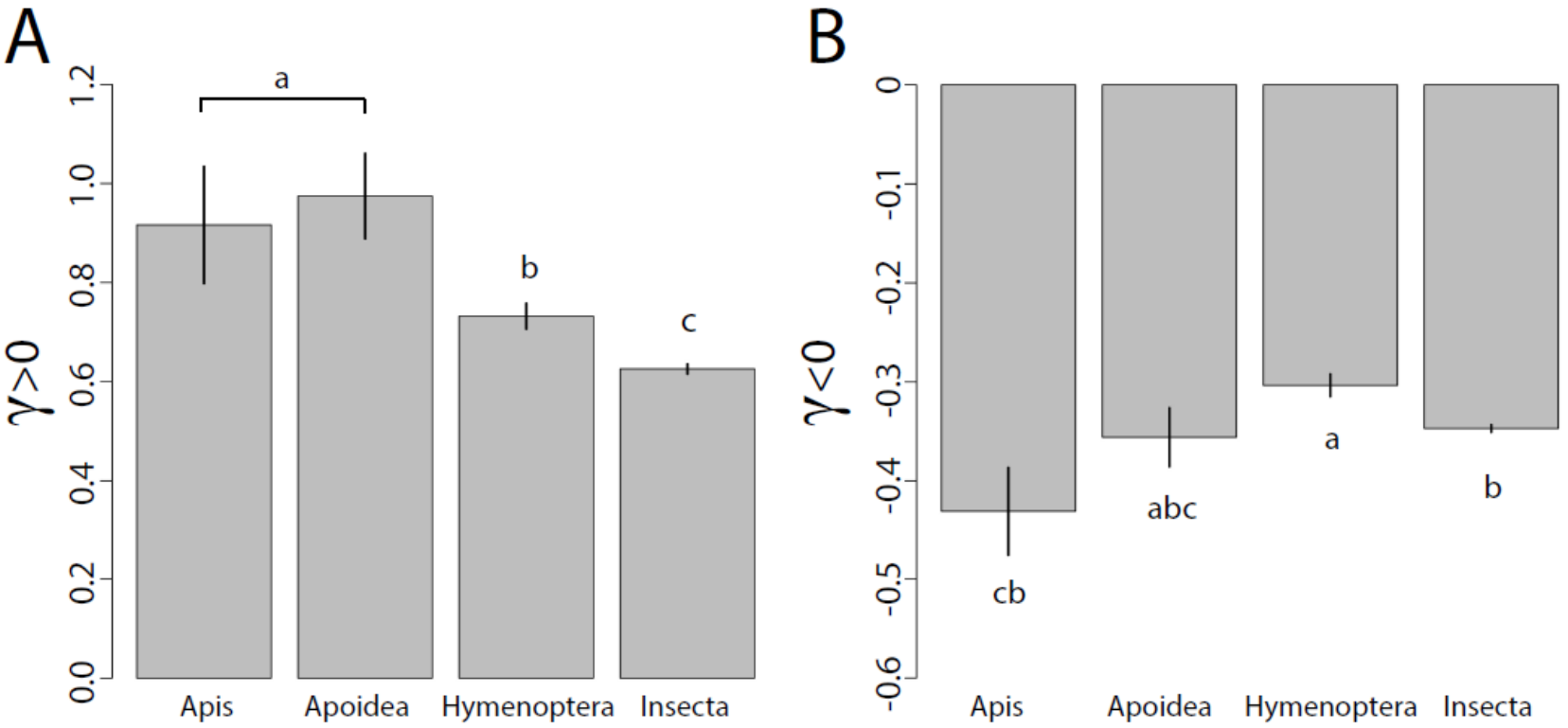

A
γ>0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
a
b
c
Apis
Apoidea
Hymenoptera
Insecta
B
γ<0
0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
cb
abc
a
b
Apis
Apoidea
Hymenoptera
Insecta

Figure 2.3

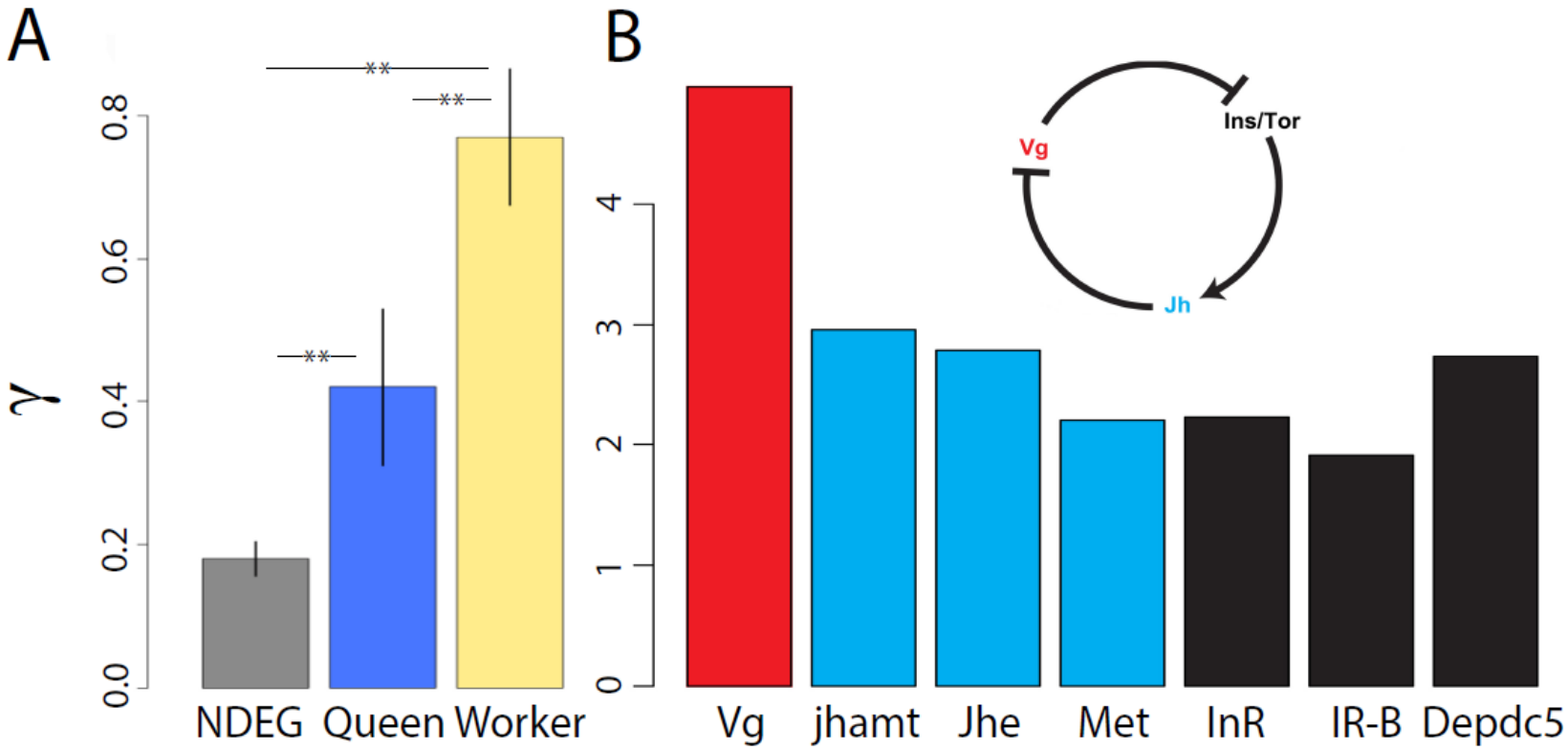

A
B
γ
0.0
0.2
0.4
0.6
0.8
**
**
**
NDEG
Queen
Worker
0
1
2
3
4
Ins/Tor
Vg
Jh
Vg
jhamt
Jhe
Met
InR
IR-B
Depdc5

Figure 2.S1

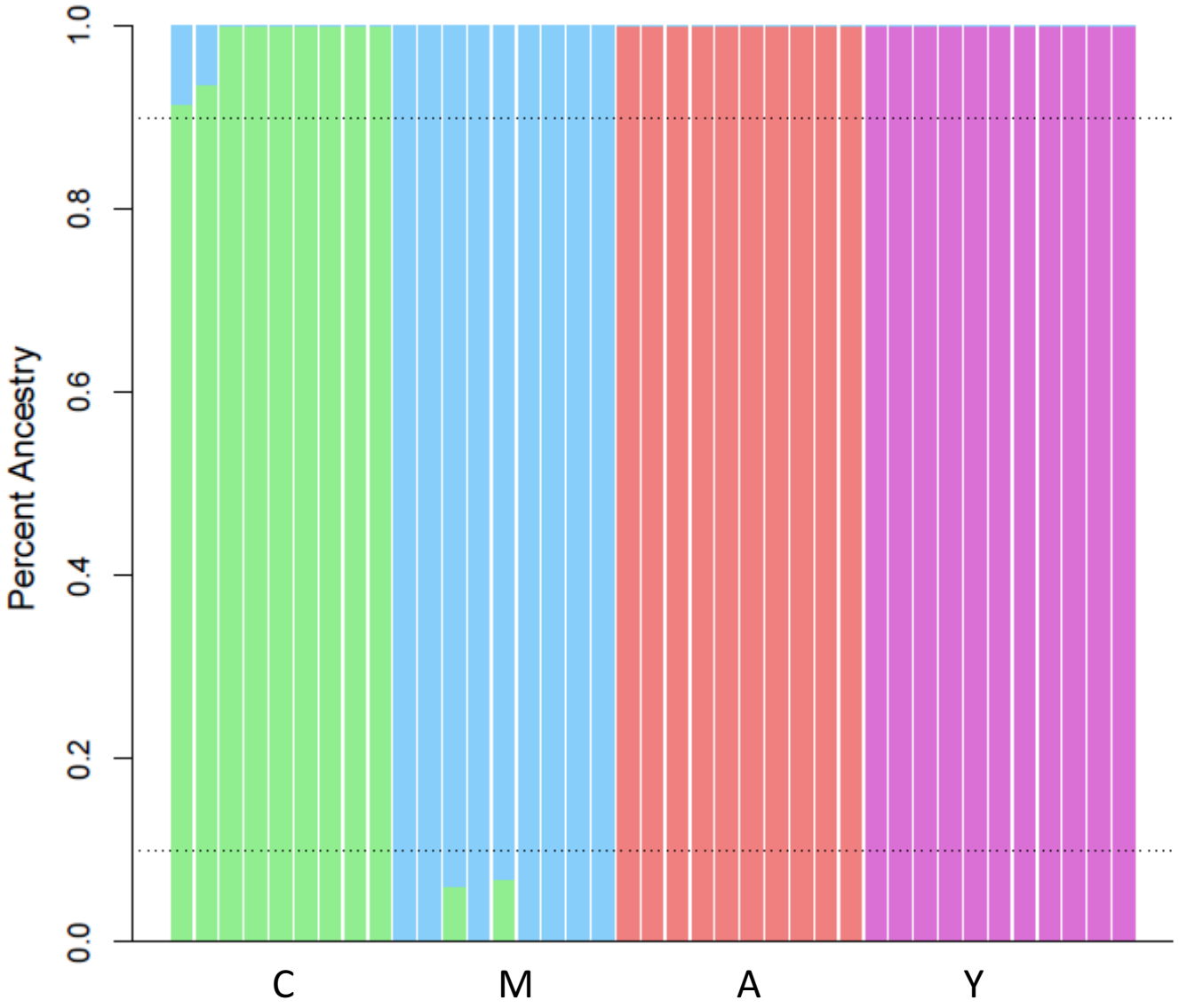


Figure 2.S2

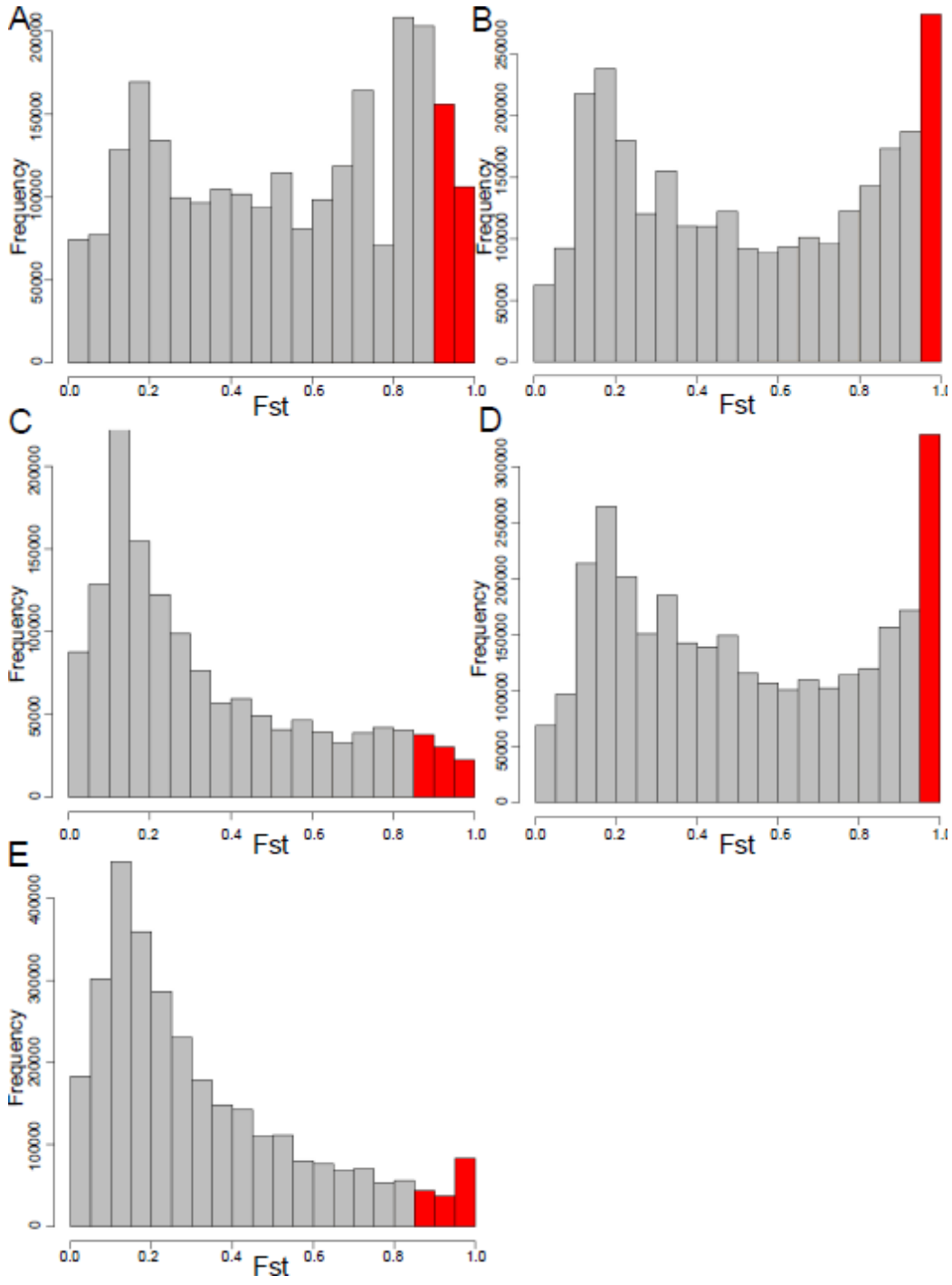
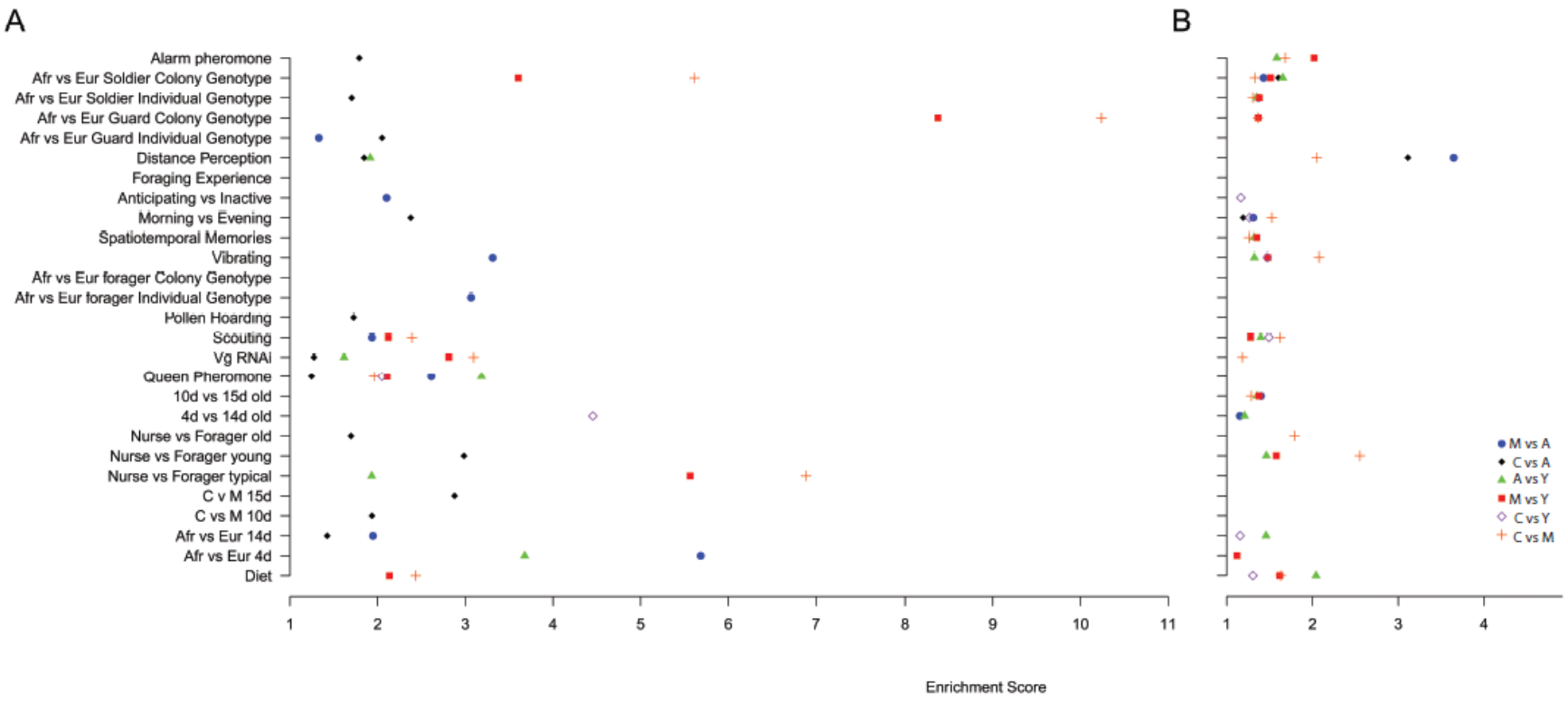


Figure 2.S3


A
B
Alarm pheromone
Afr vs Eur Soldier Colony Genotype
Afr vs Eur Soldier Individual Genotype
Afr vs Eur Guard Colony Genotype
Afr vs Eur Guard Individual Genotype
Distance Perception
Foraging Experience
Anticipating vs Inactive
Morning vs Evening
Spatiotemporal Memories
Vibrating
Afr vs Eur forager Colony Genotype
Afr vs Eur forager Individual Genotype
Pollen Hoarding
Scouting
Vg RNAi
Queen Pheromone
10d vs 15d old
4d vs 14d old
Nurse vs Forager old
Nurse vs Forager young
Nurse vs Forager typical
C v M 15d
C vs M 10d
Afr vs Eur 14d
Afr vs Eur 4d
Diet
1
2
3
4
5
6
7
8
9
10
11
Enrichment Score
M vs A
C vs A
A vs Y
M vs Y
C vs Y
C vs M

Chapter 3:

Contribution of queen and worker traits to adaptive evolution differs between bumble bees and honey bees

Brock A. Harpur, Alivia Dey, Jennifer R. Albert, Sani Patel, Heather M. Hines, Martin Hasselmann, Laurence Packer, Amro Zayed[2]

Introduction

Within a eusocial colony, labour is divided between the queens—responsible for most of the reproduction—and their workers—responsible for colony upkeep as the brood care specialists, nest defenders, and foragers (Wheeler 1910, Wilson 1985, Winston 1987, Hölldobler and Wilson 1990, Sagili et al. 2011, Wray et al. 2011). The separation and subsequent specialization of these roles is the result of Darwinian selection that has acted directly on mutations contributing to queen phenotypes and indirectly on mutations that influence worker traits (Wilson 1985, Sagili et al. 2011, Wray et al. 2011). We do not yet have an understanding of the relative role of queen or worker phenotypes to the fitness of eusocial lineages, a knowledge gap that has hindered out ability to understand the evolutionary processes responsible for caste divergence across different stages of social evolution and the resulting changes in social complexity.

Until recently, it was challenging to objectively compare the fitness effects of mutations influencing queen and worker traits. However, advances in population and functional genomics of social insects have allowed researchers to identify genes that are associated with worker and queen traits and quantify their relative importance to adaptive evolution in social lineages (Hasselmann et al. 2015, Kent and Zayed 2015). The first population genomic study of a social

[2] This submitted manuscript has been reprinted by permission from its co-authors: Harpur BA*, et al.* (2017). Contribution of queen and worker traits to adaptive evolution differs between bumble bees and honey bees. *Genome Biology and Evolution* [Submitted].

insect demonstrated that mutations in genes with worker-biased expression were, on average, significantly more beneficial relative to mutations in queen-biased genes in the eusocial honey bee, *Apis mellifera* (Harpur et al. 2014b). However, the relative importance of worker traits in other eusocial species is not so well understood and may be substantially different as a result of variation in social lifestyles among taxa. In the honey bees (*Apis* sp.), for example, colonies are perennial and contain thousands of individual workers that are morphologically distinct from their single queen (Michener 1974, Rehan and Toth 2015). In contrast, bumble bee colonies (*Bombus* sp.) are small (tens to hundreds of individuals), annual, and have a solitary workerless phase during the early stages of colony development (Michener 1974, Winston 1987, Rehan and Toth 2015). Although *Bombus* colonies have workers, they are absent for one of the most challenging life history stages, where foundresses (future queens) undertake the difficult task of beginning a colony and then performing all or a subset of the behavioural repertoire of workers to provision their first brood (Alford 1969, Crespi and Yanega 1995, Gadagkar 1997, Bourke 2011, Rehan and Toth 2015). In these annual eusocial societies, the success of a colony may be more influenced by traits expressed by foundresses and queens than those expressed by workers (Michener 1974).

The corbiculate bees are an ideal group to study the relative contribution of queen-acting and worker-acting mutations to fitness because of their considerable variation in social organization (Rehan and Toth 2015). Moreover, honey bees and bumble bees share a common social ancestor, and consequently, have been subject to the genomic impacts of social evolution for the same length of time (Romiguier et al. 2016). We carried out a comparative population genomics study of bumble bees and honey bees to identify and characterize genes with signatures of adaptive evolution in the two lineages, and compare the fitness effects of mutations influencing queen and worker phenotypes in the bumble bees relative to the perennially eusocial honey bees.

Results & Discussion

Adaptively evolving genes in *Bombus* and *Apis* are largely different

We used population genomic approaches (Hasselmann et al. 2015, Kent and Zayed 2015) to estimate the strength of selection acting on genes in *Bombus* and genes in *Apis* over the

approximately 5-25 MYA when sister species within each genus diverged (Arias and Sheppard 1996, Cameron et al. 2007, Hines 2008, Kotthoff et al. 2013). The bumble bee dataset comprised 21 newly sequenced genomes representing *B. impatiens, B. terrestris,* and *B. melanopygus* sequenced at high depth (16.5X; Materials and Methods). We compared this data set to a recently published *Apis* population genomic study that was sequenced and analysed using similar methods (Harpur et al. 2014b). We used a Bayesian implementation of the McDonald-Kreitman test (Eilertson et al. 2012) to estimate γ, the average selection coefficient of replacement mutations scaled by the effective population size, for 10,008 protein-coding genes in *Bombus* and 12,303 protein-coding genes in *Apis* (Materials and Methods; Fig. 3.1). We found that, on average, genes within both genera evolved neutrally and were not significantly different from $\gamma = 0$ (Fig. 3.1). However, we found evidence of strong positive selection acting on sets of genes within both lineages: 17.8% of *Bombus* genes and 9.3% of genes within *Apis genes* (Harpur et al. 2014b) had evidence of strong positive selection ($\gamma > 1$; Fig. 3.1) and both groups had ~2% of genes with $\gamma > 2$.

The positively selected genes in honey bees and bumble bees are largely unique to each lineage. There was a weak correlation between the selection coefficient (γ) between all genes in *Bombus* and their putative orthologous in *Apis* (Pearson Correlation, $r = 0.17$, $P < 2.2x10^{-16}$). However, this likely reflects shared patterns of evolutionary constraint on protein-coding sequences in both groups. As the selection coefficient increases in either lineage, the correlation coefficient between γ in *Apis* and *Bombus* rapidly becomes negative and non-significant (Fig. 3.2).This is consistent with a recent phylogenomic study that found little evidence for common patterns of accelerated amino acid evolution across independently derived social lineages (Kapheim et al. 2015). Given that honey bees and bumble bees share a common social ancestor; our finding that genes with high levels of positive selection were largely unique to each genus indicates that adaptive divergence involves very different genes even in closely related social lineages.

Adaptively evolving gene functions in *Bombus* and *Apis*

While the genes acted on by strong positive selection within *Bombus* are largely not the same as those acted on by strong positive selection within *Apis*, there may be overlap in the

biological, molecular, or cellular functions of these genera-specific positively-selected genes, if eusocial societies face similar selective pressures. We explored this hypothesis by identifying the Gene Ontology (GO) terms associated with genes underlying adaptive evolution in both genera (Huang et al. 2009) (Materials and Methods). Similar to our gene-specific analysis, we found little overlap between the functions of genes underlying adaptive evolution in bumble bees and honey bees. Within *Bombus*, the most significantly enriched GO terms (Hypergeometric Test; Bonferroni $P < 0.001$) were mitochondrion (GO:0005739), oxidative reduction (GO:0055114), mitochondrial organization (GO:0007005), NADH dehydrogenase activity (GO:0003954), and mitochondrial ATP synthesis coupled electron transport (GO:0042775). As we removed all mitochondrial genome scaffolds from our analyses (Materials and Methods), these genes represent adaptively evolving nuclear genes that are involved in mitochondrial function. In contrast, positively selected genes within *Apis* were enriched for 41 significant GO terms that often related to behaviour, including sensory perception (GO:0007600 sensory perception of smell (GO:0007608), co-factor binding (GO:0048037), channel-activity (GO:0015267), and cognition (GO:0050890) (see online supplement (Harpur et al. 2014b)). We found very little overlap between GO terms acted on by selection within *Apis* and *Bombus*: only two terms were enriched within each of the Biological and Molecular Functions between genera. If we examine only genes with very strong evidence of positive selection ($\gamma > 2$), there is no overlap in the GO terms acted on by selection within *Apis* and *Bombus*. Overall, this indicates that much of the adaptive protein evolution is lineage-specific and a common social origin does not drive adaptive selection.

Adaptive evolution of queen-biased vs. worker-biased genes in *Apis* and *Bombus*.

We found little overlap in the genes experiencing strong positive selection in both genera, suggesting that the traits underlying adaptive evolution differ between the two. We directly tested this hypothesis and found a shift in the relative importance of genes with an expression bias towards adult workers between these two eusocial genera. We had previously reported, and here replicated, that worker-biased genes in honey bees have higher selection coefficients relative to queen-biased and non-differentially expressed genes in *Apis* and have a higher proportion of genes acted on by positive selection (Harpur et al. 2014b) (Fig. 3.3). Using a

similar data set for *Bombus* that examined differential brain gene expression across each life history stage and caste (Woodard et al. 2014), we made the same comparison as in *Apis*. In contrast to previous results, we found that the genes expressed in *Bombus* female reproductives (queens and foundresses) had significantly higher selection coefficients and a higher proportion of genes acted on by positive selection than genes expressed in the non-reproductive workers and those that are not differentially expressed (Fig. 3.3).

This shift in the strength of selection on workers versus reproductives may reflect fundamental differences in the life histories of these two lineages. For example, the solitary founding phase in *Bombus* imposes a strong selective filter on reproductive individuals (Free and Butler 1959, Goulson 2010). At this life history stage, the foundress is solely responsible for the success of a future colony's output and there are strong metabolic demands to produce eggs, forage, and maintain the colony (Free and Butler 1959, Goulson 2010). We predicted that genes expressed at this life history stage may be those contributing to the patterns of positive selection we have detected on the *Bombus* genome. To test this prediction, we analysed a recent transcriptomic dataset from the fat bodies of virgin or mated female reproductives, diapausing female reproductives, and egg-laying foundresses for signatures of selection (Amsalem et al. 2015). We found that genes differentially-expressed by foundresses had a significantly higher proportion of genes acted on by strong positive selection (21.8%) relative to genes highly expressed during any other life history stage in this study (13.2% of genes across all other stages; Fisher Exact tests; $P<0.0001$). This analysis suggests that genes expressed by foundresses early in the colony cycle are the major source of adaptive evolution in bumble bees

Conclusions

Our analyses provide unique insights into the factors that influence adaptive caste divergence in social organisms. The stark differences in the relative importance of queen and worker traits to adaptive evolution of bumble bees and honey bees is particularly intriguing because it suggests that the evolution of eusociality *per se* does not necessary lead to conditions that render worker phenotypes to be of primary importance for the fitness of eusocial lineages. Honey bee workers are present during the entire life cycle and thus their traits can continuously influence the fitness of a colony. However, bumble bee workers are present only after colony

founding and, according to our results; their overall contribution to fitness is smaller relative to queens, perhaps as a result of the strong selective pressure on queens during the solitary founding stage.

Solitary nest founding is a common feature of most primitively eusocial insects. Our results suggest this life history trait leads to faster rates of evolution in traits expressed by reproductives relative to workers. The finding that strong selection acts on different genes within the genomes of *Apis* and *Bombus* and that those genes under strongest selection in each group act on traits relevant to different castes is of considerable sociobiological importance. It suggests that the loss of queen totipotency causes a dramatic change in the architecture of selection pressures upon the social insect genome. Switches from eusociality to solitary behaviour have occurred many times but there seem to have been few switches from swarm to independent colony founding among social insects (Packer 1997, Noll 2002, Cronin et al. 2013) . Our results suggest that divergent selection regimes may have made the latter transition much less likely.

Materials and Methods

Sampling, Sequencing, Alignment, and SNP Calling

We sampled haploid males from populations of the bumble bees *Bombus impatiens* (Toronto, Canada; N=10 and *B. melanopygus* (Oregon, United States; N=3), both in subgenus *Pyrobombus*, and *B. terrestris* (subgenus *Bombus* s.s.) (Norwich, United Kingdom; N=8). Each bee sample was paired-end sequenced (150 bp) with Illumina Hi-Seq Sequencing at either Génome Québec Innovation Centre's sequencing facility or the Penn State Huck Institutes of the Life Sciences Genome Core Facility to an average read-depth at each SNP of 16.5X. All reads were aligned to the *B. impatiens* genome assembly v 2.0 and annotated with Official Gene Set v 2.0 (Sadd et al. 2015) using the default parameters of BWA v 7.5 and SAMtools v 1.19 (Li and Durbin 2010). Because sequences were diverse and divergent relative to the reference genome, we remapped each bee's sequence using STAMPY v 1.0 (Lunter and Goodson 2011) at a substitution rate of 0.02. We subsequently re-aligned with GATK v 3.1 RealignerTargetCreator followed by IndelRealigner to reduce any potential erroneous alignments close to indels (DePristo et al. 2011). VCF files were created using GATK UnifiedGenotyper for both Single Nucleotide Polymorphisms (SNPs) and indels using --ploidy 1. We used three filters to reduce

the chance of making erroneous genotype or variant calls. First, we removed all SNPs within 10 bp of an indel using GATK's --maskExtension command. Second, we removed all SNPs in areas of outlier depth using a 1.5 x Inter Quartile Range cutoff. Third, we broadly removed all SNPs within repetitive or potentially paralogous areas of the genome. To perform this filter, we performed a BLAST of 150bp sequences across the *B. impatiens* genome and excluded any SNP within an area with multiple BLAST best-matches (E-value cut-off of 1^{-6}). Finally, we removed any SNP that could potentially be misgenotyped due to its local sequence complexity. We performed this filter by allowing GATK to call SNP genotypes for 3 randomly selected *B. impatiens* samples using the --ploidy 2 option. Because all of our samples were haploid, any site called as heterozygotic is erroneous. We compiled a list of such sites and removed all SNP calls within 5 bp from all samples we sequenced.

We identified all SNPs within protein coding genes and identified if those SNPs were non-synonymous or synonymous using SNPEFF v 3.6 (Cingolani et al. 2012) and excluded all genes lacking start codons, lacking stop codons, or containing premature stop codons (N = 1018 genes excluded). Because OGS v.2 contains isoforms of each gene, we included either the longest isoform or, in the case of isoforms being the same size, we randomly selected a single isoform for our analyses.

Relatedness and Population Structure

Because we sampled individuals of the species *B. impatiens* and *B. terrestris* each within the same municipality we tested whether samples within each species was a sibling or close relative. We used the program RELATEDNESS 4.2 (Queller and Goodnight 1989) to determine the average relatedness of individuals within each species using the genotypes at 127 randomly selected SNPs with MAF > 0.1 over 10 runs, each selecting a new set of 127 random SNP genotypes. No two individuals within any of these two *Bombus* species had significant evidence of being closely related (relatedness not significantly different from 0; $P > 0.25$ for all comparisons). To ensure each sample was indeed a member of its designated species and to ensure there was no evidence of population structure within species, we used the program ADMIXTURE v 1.22 (Alexander et al. 2009). Within each species we estimated K, the number of groups within a dataset, by randomly selecting 10% of SNPs with MAF > 0.1 and estimating

K= 1 to N - 2 where N is the number of samples for a given species. We tested each value of K 5 times with different sets of randomly selected SNPs. We used the cross-validation (CV) method to determine the optimal value for K. We used the same method above, but for SNPs shared across all species to ensure our sampling represented three distinct bumble bee lineages. There was no evidence of population structure within any species using ADMIXTURE (K=1 with all species individually).

Analysis of Positive Selection

We estimated the strength of selection within the *Bombus* genus for 10048 genes using a Bayesian implementation of the McDonald-Kreitman test (Eilertson et al. 2012). After identifying synonymous and non-synonymous mutations (above), we classified mutations as being fixed or polymorphic within species pairs within the genus (e.g. within and between *Bombus impatiens.* and B. *terrestris*). We ran Bayesian implementation of SNIPRE for 15000 iterations after 100000 burnin steps. After a Bayesian-equivalent False Discovery Rate correction, SNIPRE outputs estimates of the scaled selection coefficient, γ ($2N_e s$) and its 95% confidence interval. We also performed the same analyses above using *B. impatiens* and *B. melanopygus* to derive counts of fixed and polymorphic SNPs within exons. Our estimates of selection were highly correlated between the two potential outgroups ($t_{10006} = 85$; $r = 0.65$, $p < 2.2 \times 10^{-6}$). Finally, to validate our estimates of γ further, we calculated α , a measure of the proportion of nonsynonymous mutations fixed by selection (Eyre-Walker 2006) to ensure our results were consistent across methodologies. We found that γ and α correlated significantly (Figure 3.S1; $r = 0.80$, $P < 2.2 \times 10^{-16}$) and that high γ genes tended to also have $\alpha >> 0$.

Caste-Biased Genes and Genes Expressed During Diapause

We followed the same procedure used by our group previously to identify differential expression among castes of honey bees (Harpur et al. 2014b) using the Honey Bee Protein Atlas (Chan et al. 2013). The Atlas provides a list of proteins which were found to be differentially expressed consistently across 26 tissues of queens and workers. We identified *Bombus* genes with caste-biased expression patterns by using results kindly provided to us from previous micro-array analyses that examined brain gene expression of *B. terrestris* within queens, workers,

foundresses, and gynes (Woodard et al. 2014). By comparing gene expression among castes, Woodard et al. (Woodard et al. 2014) were able identify differentially expressed genes between reproductive- and non-reproductive castes and brood-caring versus non-brood-caring castes. We defined a gene as having caste-biased expression if it had been found to be significantly differentially-expressed in one caste relative to all other castes (Woodard et al. 2014). We classified genes as functioning in reproduction or brood care by making use of an ANOVA model performed previously that classified genes as being over-or under-expressed in castes performing either function (Woodard et al. 2014). We were able to calculate γ on 5643 genes within this dataset and found that 24.5% of these genes have significant evidence of having caste-bias expression patterns. To explore which reproductive-expressed genes were acted on by selection within *Bombus*, we analysed another transcriptomic dataset from the fat bodies of virgin or mated female reproductives, diapausing female reproductives, and egg-laying foundresses (Amsalem et al. 2015 see Table S3a). We compared γ of genes expressed highly between foundresses and all other comparisons to find if genes expressed in foundresses were enriched for positive selection relative to all other queen-specific life history stages.

GO Analysis

To identify functional relevance of genes with evidence of positive selection, we used Gene Ontology (GO) analysis as executed by DAVID v 6.7. We followed the same procedure used by our group previously (Harpur et al. 2014b) in order to compare GO terms between the current study and work examining selection within the genus *Apis*. For both the *Apis* and *Bombus* data sets, we identified putative fly orthologues using a BLASTN best match (E-value 1^{-6}). We used default parameters but output only MF_FAT, CC_FAT, and BP_FAT and output all significant GO results and KEGG pathways, following correction for False Discovery Rate (Bonferoni < 0.05).

Statistical analyses and Data Accession

All analyses and pipelines can be found on the author's GitHub (https://github.com/harpur/Bombus), including all supplemental data used in this study. We performed analyses using all values of γ, as well as for $\gamma > 1$, which we termed "high gamma".

Where appropriate, we used parametric models for all statistical tests, unless otherwise stated. All sequence data have been deposited with NCBI's Short-Read Archive (BioProject PRNJA347806).

Figure Legends

Figure 3.1 Distribution of the selection coefficient of replacement mutations scaled by the effective population size ($\gamma = 2N_e s$) for protein-coding genes within *Bombus* (top) and *Apis* (bottom).

Figure 3.2 Genes acted on by strong positive selection within *Apis* are not the same as those acted on by strong positive selection within *Bombus*: as the selection co-efficient increases in both species, genes with strong signs of positive selection in one tend to be neutrally evolving in the other. Red line is x = y line. Insert shows correlation coefficient and its significance as γ is increased in both species.

Figure 3.3 In *Apis* (right) genes associated with worker phenotypes show signs of adaptive evolution relative to genes expressed in queens ($F_{2,1688} = 11.97$; $P = 0.0000007$; Tukey $P < 0.01$ for all comparisons); however in *Bombus* (left), this pattern in reversed and genes expressed in female reproductive castes show signs of positive selection greater than those expressed in workers ($F_{2,5640} = 10.7$; $P = 0.00002$; Tukey $P < 0.03$ for all comparison). Error bars denote SEM; NDEG = Non-Differentially Expressed Gene. Percentages within bars are percent genes with high gamma ($\gamma > 1$).

Figure 3.S1. Relationship between the selection coefficient γ and of the proportion of nonsynonymous mutations fixed by selection (α). These values correlated significantly ($r = 0.80$, $P < 2.2 \times 10^{-16}$) and that high γ genes tended to also have $\alpha >> 0$.

Figure 3.1

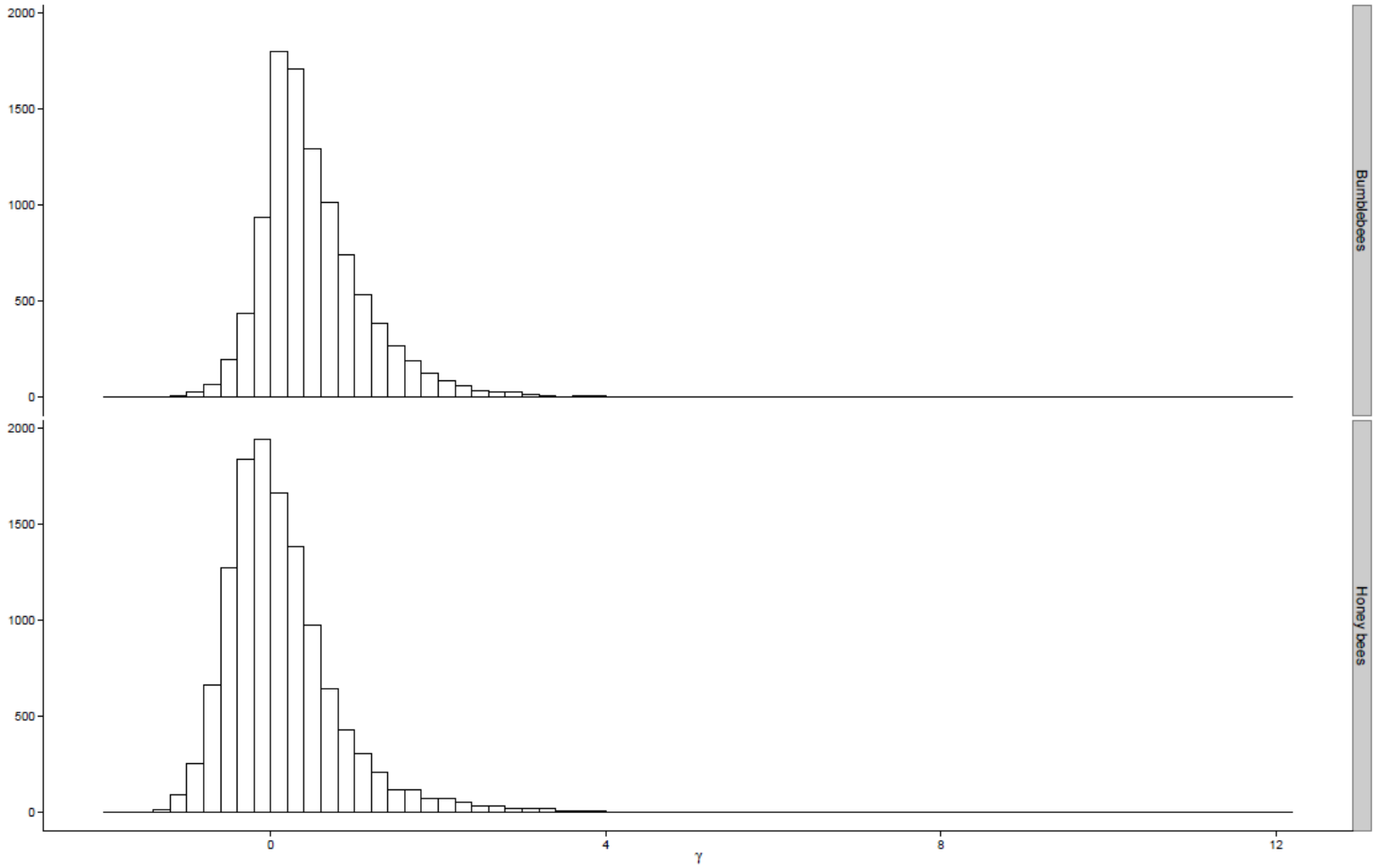

2000
1500
1000
500
0
Bumblebees
2000
1500
1000
500
0
Honey bees
0
4
8
12
γ

Figure 3.2

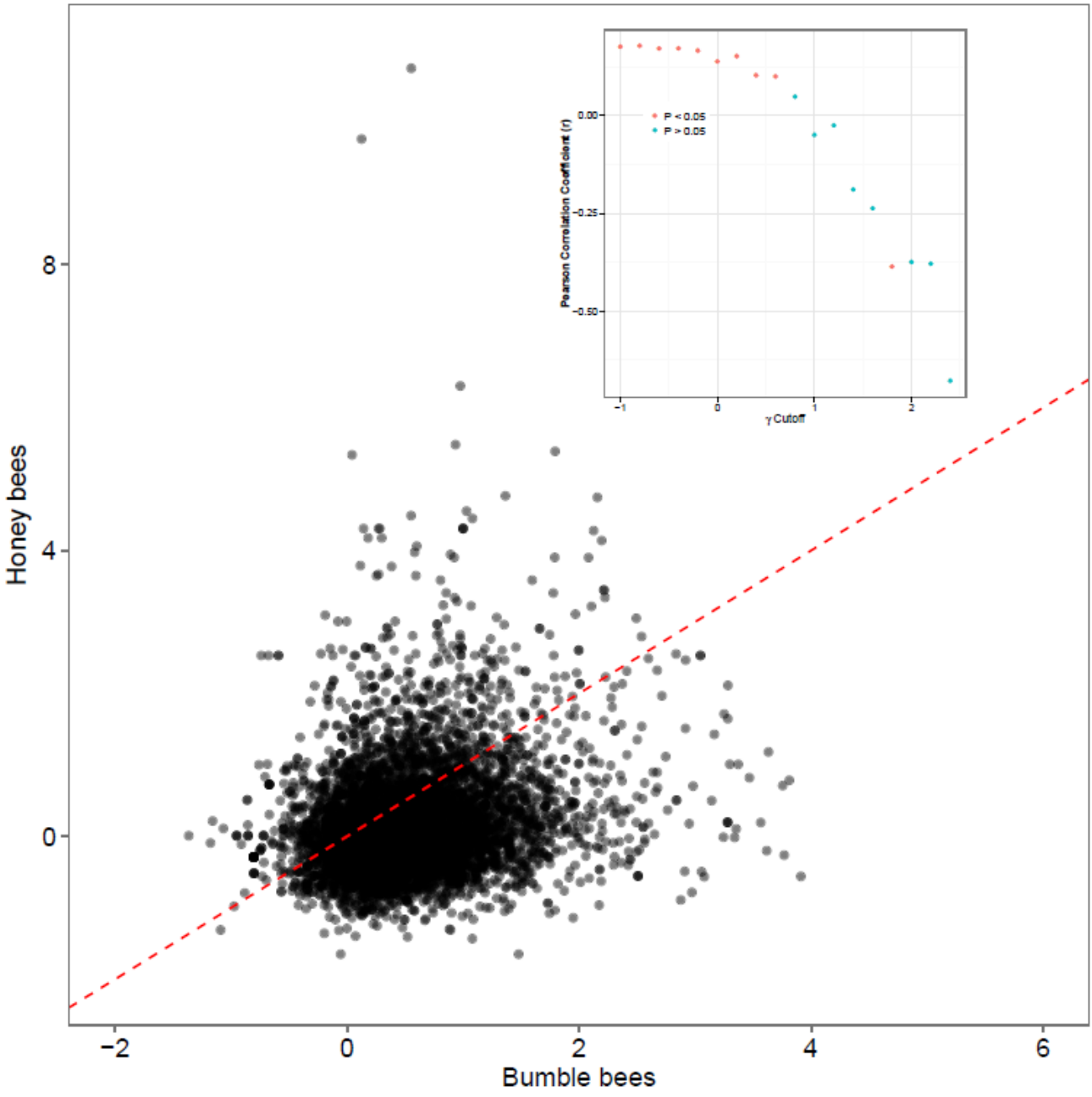

Honey bees
8
4
0
−2
0
2
4
6
Bumble bees
Pearson Correlation Coefficient (r)
0.00
−0.25
−0.50
P < 0.05
P > 0.05
−1
0
1
2
γ Cutoff

Figure 3.3

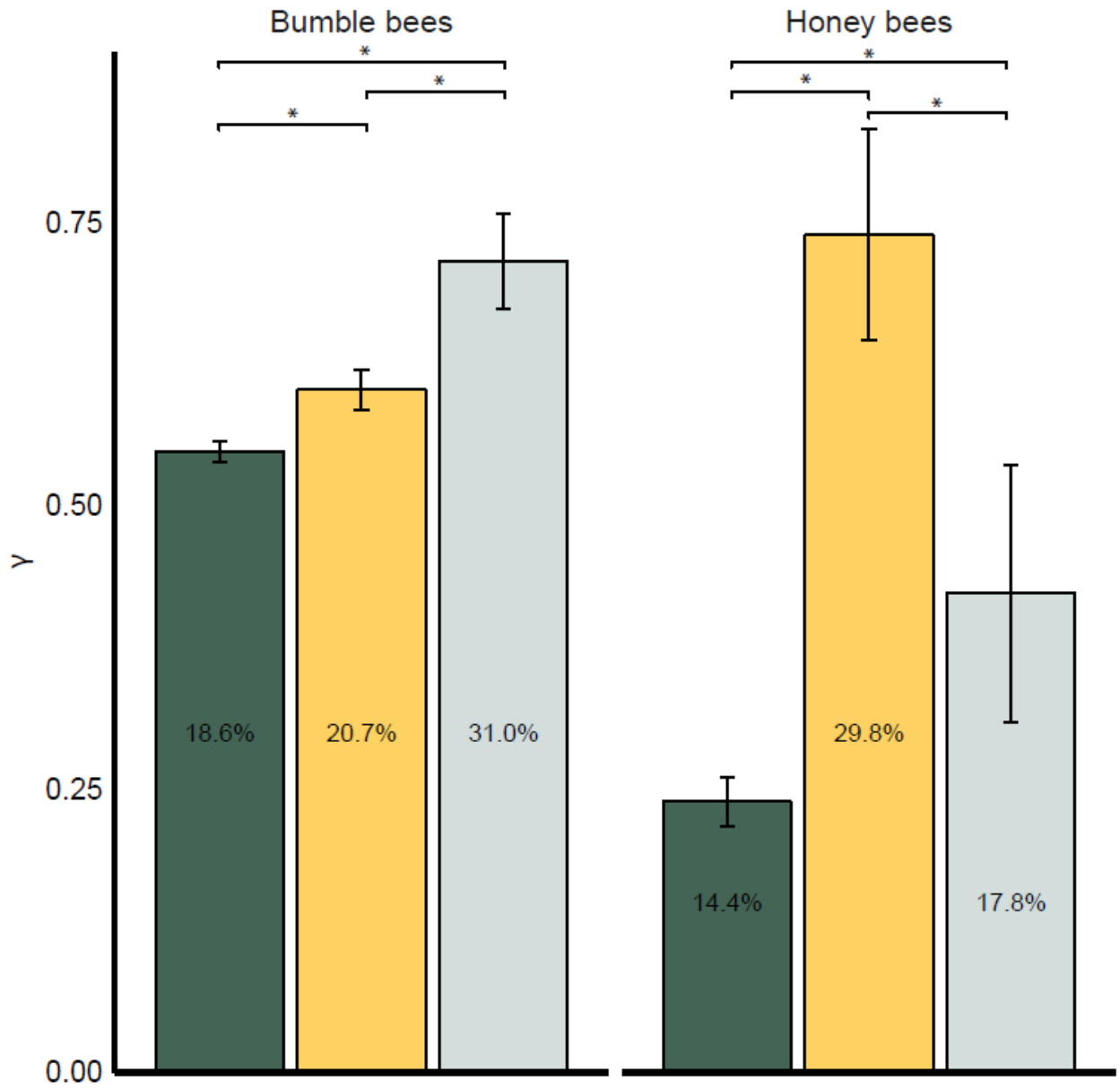
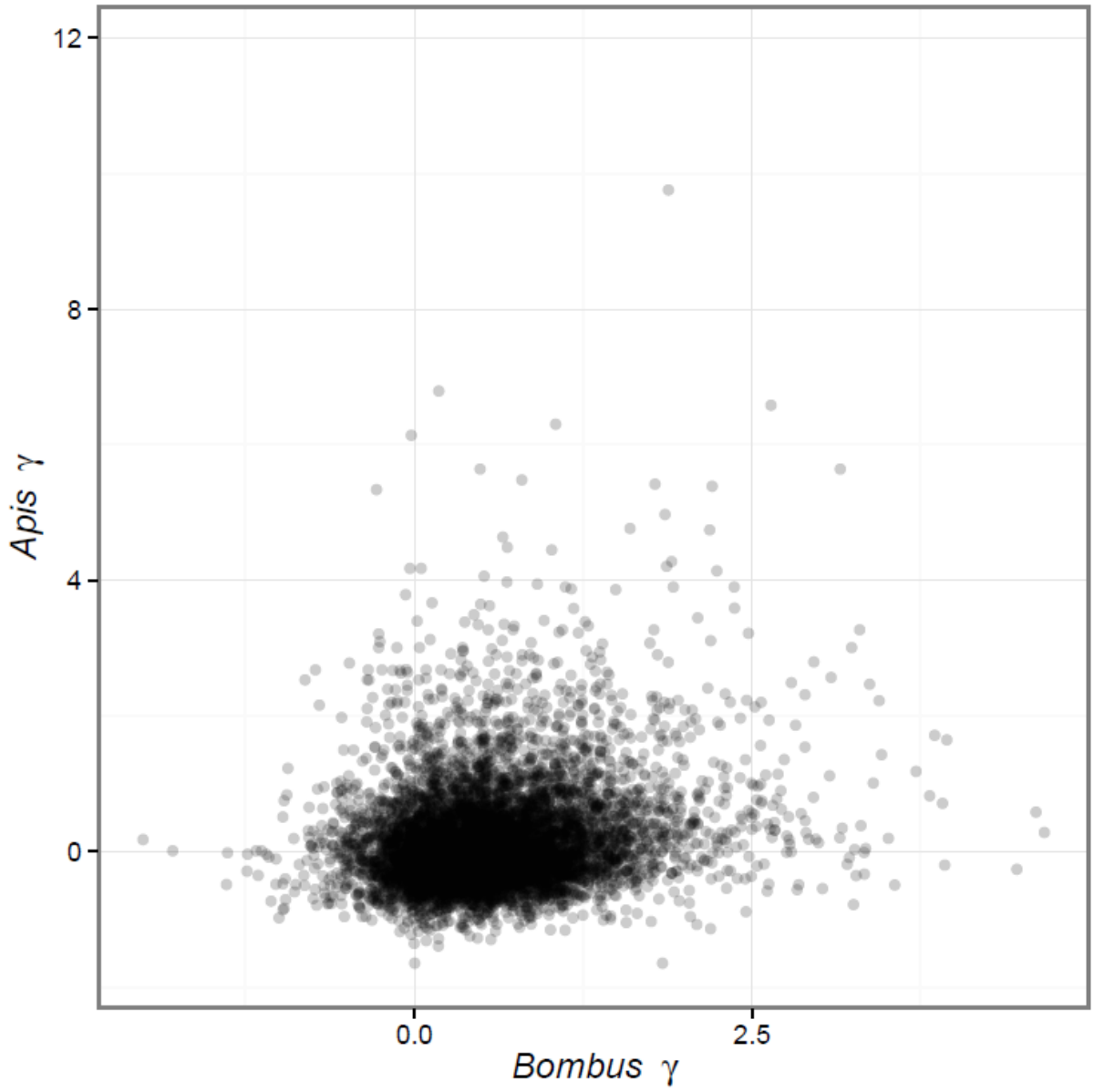

Bumble bees
Honey bees
*
*
*
*
*
*
0.75
0.50
0.25
0.00
y
18.6%
20.7%
31.0%
14.4%
29.8%
17.8%

Figure 3.S1



Chapter 4:

It's good to be clean: integrative genomics reveals adaptive evolution of the honey bee's (*Apis mellifera*) social immune system

Brock A. Harpur, M. Marta Guarna, Elizabeth Huxter, Heather Higo, Kyung-Mee Moon, Shelley E. Hoover, Abdullah Ibrahim, Andony P. Melathopoulos, Suresh Desai, Robert W. Currie, Stephen F. Pernal, Leonard J. Foster, and Amro Zayed[3]

Introduction

Living at high densities with close relatives increases the risk of epizootic outbreaks, yet these are the exact conditions in which social insects successfully live (Schmid-Hempel 1994, Zasloff 2002, Lawniczak et al. 2007, Nunn et al. 2015). Their success is due in part to their ability to mitigate the risk of pathogenic outbreak through two forms of immunity. The first is the innate immune system (Evans et al. 2006) that is comprised of well-characterized sets of genes that are conserved across social and solitary taxa. This system is activated by a set of generally acting recognition proteins that detect pathogens and, through downstream signalling pathways, elicit the expression of proteins that eliminate or reduce the pathogenic threat. We have a deep understanding of the genetics and evolution of the innate immune system in social insects, in part because the genes underpinning innate immunity are taxonomically ancient and are largely conserved across insects (Evans et al. 2006, Harpur and Zayed 2013, Barribeau et al. 2015).

The second form of immunity is the social immune system, an evolutionarily derived system of prophylactic or curative altruistic responses against pathogens (Cremer et al. 2007). The responses that are elicited by social species include secretions that act to limit bacterial and

[3] This submitted manuscript has been reprinted with the permission of its co-authors from the original submitted manuscript: Harpur BA*, et al.* (2017) It's good to be clean: integrative genomics reveals adaptive evolution of the honey bee's (*Apis mellifera*) social immune system. [Submitted]

fungal growth (Poulsen et al. 2003), self- or social-exclusion from all or part of the colony (Heinze and Walter 2010, Lecocq et al. 2016), removal or cannibalism of infected or deceased workers (Sun and Zhou 2013), grooming (Rosengaus et al. 1998), and/or the removal of dead or infected larvae (Figure 4.1A; Rothenbuhler 1964b, a). These responses are very effective at eliminating the risk of epizootics. For example, in honey bees (*Apis*), some workers are able to detect and remove infected brood before they become infective – a trait referred to as hygienic behaviour (Figure 4.1A). Field trials have shown that hygienic behaviour eliminates the risk of developing clinical symptoms of Chalkbrood disease and reduces the risk of developing symptoms of American foul brood disease by 61% (Spivak and Reuter 2001). Because of its evolutionary novelty in social insects, we do not yet know the genetic mechanisms underpinning social immunity, hampering our efforts to understand how social immunity evolves, and the possible existence of genetic or evolutionary trade-offs between innate and social immunity (e.g. Sackton et al. 2007, Harpur and Zayed 2013, Barribeau et al. 2015).

Here, we take an integrative genomic approach to study the genetics and evolution of loci associated with social immunity in honey bees. Hygienic behaviour provides an ideal model for this study. It shows substantial phenotypic variation within and among honey bees (Spivak and Gilliam 1998a, Woyke et al. 2004, Woyke et al. 2012, Uzunov et al. 2014), has been the target of several bee breeding programs around the world (e.g. Spivak and Reuter 2001, Buchler et al. 2010, Pernal et al. 2012, Guarna et al. 2015), and has a foundation of decades of research demonstrating that hygiene is highly heritable and variation can be explained by several broad (totalling ~12Mb) Quantitative Trait Loci (QTLs) (Rothenbuhler 1964b, a, Lapidge et al. 2002, Oxley et al. 2010, Harpur et al. 2014b). In this study, we created two artificially selected populations that highly express hygienic behaviour and, making use of high-depth full-genome sequencing, identified loci contributing to the variation in the expression of hygiene. We then integrated multiple independent genomic data sets (Rothenbuhler 1964b, a, Lapidge et al. 2002, Oxley et al. 2010, Harpur et al. 2014b) to quantify patterns of natural and artificial selection at loci associated with hygienic behaviour in honey bees.

Results and Discussion

Sampling and Genome Sequencing

After three generations of artificial selection our two selected populations expressed hygienic behaviour significantly more (mean = 92% of dead brood and cappings completely removed 24h post freezing) than the baseline population (68%; ANOVA; $F_{1,37} = 47.6$; $P < 0.00001$; Figure 4.1B). We sampled a total of 125 haploid drone larvae (about 3 per colony) from 41 colonies from each of the three populations. The queen genotypes of each colony were inferred given the genomes of their haploid drone sons, each sequenced to the same average mean site depth (mean = 34.3X; ANOVA $F_{1,37} = 2.15$; $P = 0.15$; See Materials and Methods). Following alignment and Single Nucleotide Polymorphisms (SNP) calling, we were able to identify 2,340,950 SNPs segregating in the colonies sampled.

Selection Mapping

Strong selective events are expected to: i) increase the degree of differentiation between selected and unselected populations at causal loci influencing the selected trait (Nijhout and Paulsen 1997), ii) increase differentiation at loci that are nearby causal mutations due to hitchhiking effects—called selective sweeps (Nielsen 2005), and iii) cause a shift in the allele frequency spectrum away from neutral expectations at and nearby causal loci in selected populations (Nielsen 2005). We used these expectations to identify regions of the genome that are associated with hygienic behaviour. To that end, we made use of three independent tests for selection. The first was the haplotype-based outlier approach hapFLK (Qanbari et al. 2012, Fariello et al. 2013) applied on the selected populations using the baseline population as an outgroup. The hapFLK statistic is a measure of haplotype frequency differentiation scaled by relatedness between populations: a high hapFLK value is indicative of positive selection (i.e. artificial selection in our study) (Qanbari et al. 2012, Fariello et al. 2013). We also estimated the shift in the allele frequency spectrum within the selected populations using Tajima's D (Tajima 1989) —lower Tajima's D relative to the genomic average is indicative of positive artificial selection. As a final metric, we estimated the integrated haplotype score (iHS) (Voight et al. 2006) for selected populations at each SNP within the genome. This statistic detects evidence of recent positive selection at a locus by comparing levels of linkage disequilibrium around alleles. These three statistics were combined into a single composite statistic (CSS) (Randhawa et al. 2014) that allowed us to find regions of the genome with robust signatures of artificial selection. This approach yielded 132 candidate regions across the genome that had significant evidence of

positive selection within the selected populations (Figure 4.2). Combined, these regions account for at least 1,255 Kb and 10,140 SNPs across the genome.

Overlap with previous QTL studies

Genomic regions associated with hygienic behaviour as revealed by our genomic contrasts often overlapped with, or were near to, previous QTLs for the trait (Figure 4.2; Lapidge et al. 2002, Oxley et al. 2010, Spotter et al. 2012, Tsuruda et al. 2012). Our regions fell directly inside the most informative QTL identified to date: *hyg2* on chromosome 5 (Figure 4.2) that accounted for 13% of the phenotypic variation in the expression of hygienic behaviour in an independent study (Oxley et al. 2010). Our regions also overlapped with two QTLs on chromosomes 1 and 9 that explain 3.9% and 6%, respectively, of the phenotypic variance of *Varroa*-Specific Hygiene (VSH)—a form of hygienic behaviour specific to brood parasitized by *Varroa* mites (Harbo and Harris 1999, 2005, Tsuruda et al. 2012). Two selected regions on chromosomes 10 and 9 also overlapped with QTLs that explain 7% of the variation in brood removal and 7% of the variation in brood uncapping behaviour, respectively (Oxley et al. 2010). Finally, we confirmed loci on chromosomes 3 and 6 that were found to be associated with hygienic behaviour from a low resolution genome association study (Figure 4.2) (Spotter et al. 2012, Spotter et al. 2016). The overlap between our work and previous genetic studies of hygienic behaviour strongly supports our approach for identifying loci underpinning hygienic behaviour in honey bees. However, our approach has much higher resolution: hygienic-associated regions span 1,255 Kb in our study, relative to a total of 12 MB previously implicated in hygienic or associated behaviours from QTL studies.

Candidate regions explain variation in hygienic behaviour in the baseline population.

Overlap with known QTLs provides strong evidence that the regions we identified are associated with hygiene. However, we were able to provide additional support for using a targeted haplotype association approach (Purcell et al. 2007). We asked if 'hygienic' loci inferred from our population genomic contrasts contained SNPs or haplotypes that actually explained phenotypic variation in hygiene in our baseline population (See Methods). We found 1443 haplotypes (2058 SNPs) within 99 of the 132 candidate loci inferred from population

genomic contrasts that were significantly associated ($P<0.05$) with differences in hygienic behaviour in the baseline population. We then asked if the haplotypes that are statistically associated with hygienic behaviour within the 99 candidate regions in the baseline population had, on average, a higher frequency in the artificially selected populations, relative to haplotypes within the same 99 candidate regions that were not phenotypically associated with the trait. As expected, hygiene-associated haplotypes in the baseline population had significantly higher frequency in the artificially selected populations, relative to non-associated haplotypes (Wilcoxon Paired Test; $P < 0.01$). This confirms that the hygienic behaviour of the artificially selected populations was increased through selection on standing genetic variation within the baseline population.

For the proceeding functional and evolutionary analysis, we only included the 99 genomic regions (977 Kb) that had significant evidence of selection in our genomic contrast between selected and baseline populations, and contained haplotypes that were significantly associated with hygienic behaviour in our baseline population.

Candidate Genes for Hygienic Behaviour

By integrating both selection and association mapping, we have narrowed the candidate loci underpinning variation in hygiene from the approximately 12 MB of bee's genome previously implicated in QTL studies to approximately 977 Kb, representing an order of magnitude improvement in mapping resolution. We next identified genes associated with hygiene by extracting those with significant evidence of differentiation in and around the 99 windows ($-\log_{10}$(HapFLK P) > 2.5). In doing so, we have also narrowed the putative candidate genes to a set of 73 protein-coding genes (49 of which are within QTL regions) that provide a proximate, functional hypothesis for how genetic variation impacts the expression of hygienic behaviour.

Variation in hygienic behaviour is the result of variance in the response threshold of nurse bees to "dead-brood" signals (Masterman et al. 2001) potentially caused by over-active octopaminergenic neurons in the antennal lobes or mushroom bodies of the brain (Spivak et al. 2003). Dead-brood signals are detected at olfactory chemo-sensory neurons of the antennae which are then transmitted to the antennal lobes and processed by the mushroom bodies.

Hygienic bees are more receptive to these signals as a result of structural variation in the brain and have distinct patterns of gene and protein expression in brain and antennal regions (Parker et al. 2012, Boutin et al. 2015, Guarna et al. 2015). Our data suggest that differences in the expression of hygienic behaviour between bees is likely the result of differences in developmental trajectory during adult behavioural maturation or larval development, as we discuss below.

After classifying these 73 candidate genes associated with hygienic behaviour based on their phylogenetic origins (Harpur et al. 2014b, Jasper et al. 2015), we found that 85-98.7% of them are shared among Hymenopterans, Insects, or Arthropods, respectively. Using enrichment analysis based on Gene Ontology, we found that candidate genes were enriched for terms associated with neuronal development and early axon guidance (GO0048812:neuron projection morphogenesis; GO0032502:developmental process; GO Analysis; $P < 0.05$). By comparing to a developmental time-course transcriptomic study of honey bees, we found that 22 of the 73 candidate genes are known to be expressed in diploid honey bee eggs between 0 and 24hrs post-laying (Pires et al. 2016). The most highly significant SNPs ($-\log_{10}$(HapFLK P) > 2.5) within the 73 genes were predominately found within introns (94% of all of SNPs), a pattern that suggests the genes underpinning hygiene are differentially regulated. The above observations paint an interesting portrait of the mechanisms underlying hygienic behaviour in honey bees; hygienic behaviour appears to be orchestrated by taxonomically ancient genes that influence brain and neuronal development.

Examining the most significantly differentiated genes and those within or near to previous QTLs, we recapitulate the broader results reported above. The significant CSS peak on chromosome 6 contains three genes (*abscam*, *goosecoid*, and *tropomysin-2-like*), all of which are critical to early neuronal development (Hahn and Jäckle 1996, Li and Gao 2003, Funada et al. 2007, Posnien et al. 2011). The most significantly differentiated of the candidates is *abscam* (GB45774) an ortholog of the *Drosophila* gene *dscam2*. *Abscam* is one of the few honey bee genes that has been functionally characterized and is known to play a role in axon guidance (Funada et al. 2007). Isoforms of *abscam* are expressed during early development within the lamina, medulla, and lobula of the optic lobes, the glomeruli of the antennal lobes, the central body, and the mushroom bodies where expression promotes neural outgrowth, particularly of

olfactory neural axons (Funada et al. 2007). It is the many isoforms of *abscam* that are involved in neuronal outgrowth and patterning, isoforms created by including or excluding immuno-globin domains through alternative splicing (Funada et al. 2007). The most significantly differentiated of the SNPs within this gene are intronic and are within or flank splice-site recognition regions surrounding immuno-globin domains (Funada et al. 2007).

The highest peaks at chromosomes 11 and 9 contain the ortholog to the *Drosophila* gene *dyschronic* (Chromosome 11; GB45054) and *Insulin-like receptor* (Chr. 9; GB53353). *Dyschronic* is expressed during development and encodes several splice forms whose expression can affect axon guidance, overall neuroanatomy and locomotion (Jepson et al. 2012). In adult *Drosophila*, *dyschronic* protein is expressed in the mushroom bodies, ellipsoid body and antennal lobes where it interacts with Big Potassium (BK) channels and regulates neuronal excitability (Jepson et al. 2012). Variants of *dyschronic*, may act to alter the response thresholds of hygienic bees through its association with BK channels. BK channels are known to limit the action potential duration (Bean 2007) and their interaction with *dyschronic* can change the shape of response thresholds (Jepson et al. 2014). Highly differentiated mutations within *dyschronic* include one mutation within a splice site region and two nonsynonymous variants. *Insulin-like receptor* on chromosome 9 shares similar functions with *abscam* and *dyschronic*: it is involved in neuronal pruning and axon guidance (Song et al. 2003, Wong et al. 2013). The CSS highest peak, on Chromosome 5, contains GB44550 (similar to Drosophila *sidestep*), again known to be involved in axon guidance during development (Sink et al. 2001).

The genes we identified likely have neuronal developmental effects but could still be expressed later in life and expressed differentially within hygienic bees. To test this, we incorporated data from previous expression studies that highlighted 96 differentially expressed genes and 9 differentially expressed proteins (Parker et al. 2012, Boutin et al. 2015, Guarna et al. 2015) in hygienic nurse bees. We found no overlap between candidate genes in our study and differentially expressed genes or proteins in hygienic versus unhygienic adult workers. However, 4 differentially expressed genes and a single protein (GB43112) were within 100 Kb of our selected windows. Two of these genes (GB54226, *myosin 20;* and GB54295, *Syn1*) were within QTL regions. This indicates that differentially expressed genes in adults are likely downstream in the regulatory pathways harbouring causal mutations for hygienic behaviour.

Evidence of Positive Selection on Social Immune Loci

Social immunity is argued to be effective at reducing the risk of infection to such an extent that it relaxes constraint on the innate immune system (Evans et al. 2006, Cotter and Kilner 2010, Harpur et al. 2014a, Lopez-Uribe et al. 2016). If the genes underpinning social immunity contribute to fitness in social lineages, we would expect those genes to be acted on by natural selection. To date, no such study has explored the evolution of social immunity because the underlying genes were not known. Here, we examined patterns of adaptive evolution at our candidate genes relative to the rest of the honey bee's protein-coding genome over the past ~5 to 25 MYA (Arias and Sheppard 2005, Kotthoff et al. 2013). We achieved this by directly estimating selection coefficients at hygienic loci and comparing them to other genes in the genome using a variant of the MK test applied to sequence data from *A. mellifera* and its sister species *A. cerana*.

We found that 13.6% of the 73 hygienic genes had evidence of strong positive selection and that these genes had significantly higher selection coefficients than all other similar sized sets of genes in the genome (permutation test N = 10000; P = 0.005). If we restrict our analysis to only the 49 candidate genes within QTL regions, we again find that hygiene candidates are more highly enriched for evidence of selection with 23.2% of those genes having evidence of selection (P = 0.01). As a set, the hygienic candidates had higher selection coefficients than 90% of all honey bee genes sets in the Gene Ontology Biological Process 4 database (Huang et al. 2009), with levels of selection similar to the biological processes of regulation of neurotransmitter levels (GO:0001505), learning or memory (GO:0007611), and detection of external stimulus (GO:0009581). Our analysis strongly supports the hypothesis that social immunity is important for fitness in honey bees and that this fitness benefit is likely to have occurred throughout the history of the genus *Apis* and not strictly a result of beekeeping as our estimates of selection were derived from the African honey bee genome; a population of honey bees that is not typically used in commercial beekeeping.

C-lineage alleles are associated with hygienic behaviour in managed bees.

Comparisons within and across multiple studies suggest that subspecies of the honey bee's C-lineage (e.g. *A. m. ligustica* or *A. m. carnica*) are more hygienic than subspecies of the M-lineage (e.g. *A. m. mellifera*) in Europe (Flores et al. 2001, Perez-Sato et al. 2009, Bak et al. 2010, Balhareth et al. 2012, Uzunov et al. 2014, Gerula et al. 2015). Managed North America honey bees are highly admixed, originating from both the C- and M-lineage bees of Europe (Harpur et al. 2015). If the differences in hygienic behaviour between the C- and M-lineages are genetically influenced, then we could expect to find a higher frequency of C-lineage alleles in managed North American populations that have been artificially selected for hygienic behaviour.

In our artificially selected populations, we found that hygienic loci have significantly more C-lineage ancestry (median 87% C) relative to the baseline population (79% C) and relative to the genome as a whole (Figure 4.3A; Wilcoxon Test, $P < 0.0001$). We found this same pattern of differential admixture at hygienic loci within an independent population of Canadian honey bees – colonies from the province of Ontario that have been subjected to artificial selection for hygienic behaviour for more than a decade (Harpur et al. 2012, Harpur et al. 2015) (Figure 4.3B). Within the candidate genes above, at the most extreme, SNPs within *Insulin-like receptor* (chr 9; GB53353) are almost entirely fixed for C-lineage variants within selected and North American hygienic populations (median 95% C in selected and 91% within North America) but not within the baseline population (53% C).

Conclusions

We used an integrative genomic approach to identify regions of the honey bee genome associated with hygienic behaviour and to study the molecular function and evolutionary trajectory of these regions. We show that genes associated with hygienic behaviour are highly conserved and enriched for regulatory mutations that likely act to influence brain and neuronal development of worker bees. Over the course of the honey bee's evolutionary history we found that genes associated with hygienic behaviour have evolved through positive selection. Our study provides a clear link between hygienic behaviour and fitness of honey bee colonies. The strong conservation of genes associated with hygienic behaviour support an Evo-Devo hypothesis (Toth and Robinson 2007) for the origin of social immunity, and our study suggests a 'hygienic behaviour tool-kit'. The evolution of hygiene within the genus *Apis* and more broadly across

other hygienic corbiculate bees may be underpinned by differential developmental canalization resulting from variation in the regulation of an underlying, common, set of genes.

Materials and Methods

Beekeeping and breeding

Honey bee sampling, field testing, and breeding was performed at four locations in Western Canada: selective breeding for hygienic behaviour was conducted near Grand Forks, BC while unselected colonies were maintained at the Research Farm of Agriculture and Agri-Food Canada in Beaverlodge, AB and at the University of Manitoba in Winnipeg and propagated near Abbotsford, BC (Guarna et al. 2015). Colonies were assessed for hygienic behaviour using the freeze-killed brood method (Spivak and Gilliam 1998b, a), where the proportion of sealed cells that nurse bees fully uncap and remove dead pupae from is counted at 24 h using two separate tests performed one week apart on each colony. From a baseline population of 600 colonies, two selected populations were maintained for three generations and selectively bred for either high hygienic behaviour or a combination of hygienic behaviour and expression of protein markers associated with hygiene (Guarna et al. 2015). For the first two generations, selected colonies were crossed using instrumental insemination in which selected virgins where crossed with pooled semen collected from drones from 8-12 breeder colonies per site. Virgin queens from the third generation of selection were naturally closed mated, with mating apiaries located in an isolated mountain valley near Grand Forks and Christina Lake, Canada, respectively, where there were no other known feral or domestic sources honey bees. We also sampled 8 diploid adult workers from a random set of colonies within Ontario, Canada. We included these samples to look for evidence of non-random introgression at hygienic loci.

Genome Alignment and SNP Calling

The McGill University and Génome Québec Innovation Centre sequenced high molecular weight DNA from a total of 125 haploid male honey bees (drones) using Illumina HiSeq 2500 Rapid with 150 bp paired-ended reads to a mean depth of 33.07 reads. Drones were collected as larvae from 41 colonies from each of the control and selected lines with an average of 3.1 drones collected per colony. All samples were aligned, processed, and had SNPs called following a similar pipeline used previously by our group (Harpur et al. 2014b; and

https://github.com/harpur/HygSel). Raw reads were trimmed of leading and tailing sequence with Trimmomatic v0.32, aligned to the honey bee reference genome (AMEL v4.5) using NextGenMapaligner v 0.4.12 (Sedlazeck et al. 2013), and removed of duplicate reads with Picard v1.8. For each colony, we created Variant Call Files (VCF) with GATK v 3.5 first by re-aligning around indels with RealignerTargetCreator followed by IndelRealigner to reduce any potential erroneous alignments (McKenna et al. 2010) then using UnifiedGenotyper (with options -stand_call_conf 60.0 -stand_emit_conf 40.0 –dcov 200 --min_base_quality_score 20) to call SNPs and then indels. We hard-filtered SNPs using VariantFiltration (QD < 5.0, FS > 40.0, MQ < 25.0, DP < 100.0) and excluded sequence from all unmapped scaffolds (AMEL v4.5; Groups 17 or Groups Un) because of low sequencing coverage in these small and gene-sparse scaffolds. Several genomic features can result in sequence data falsely calling SNPs and inflating local diversity (McKenna et al. 2010, Hodgkinson and Eyre-Walker 2011, Leffler et al. 2013). To account for these problems, we applied three additional filters to our dataset prior to scanning for selection. First, we removed all SNPs within 10 bp of an indel using GATK's VariantFiltration. Second, we eliminated 1.5xIQR outliers for depth within any alignment. Third, we aligned all drones individually to the honey bee reference genome; however, when calling SNPs with GATK (as above) we allowed the calls to be made as diploid with the expectation that heterozygotic calls would indicate areas of low complexity that may lead to subsequent sequencing error (Wallberg et al. 2014). We excluded any SNP within 5bp of these low-complexity sites. This alignment procedure was followed for each drone as well as for pooled alignments of drones from the same colony. The later allowed us to infer the queen's genotype for each colony, the data set we proceeded with for all analyses. SNPs were identified as non-synonymous or synonymous using SNPEff v3.6 (Cingolani et al. 2012).

Identifying Positively Selected Loci

Artificial positive selection shifts the allele frequency spectrum around selected loci by driving causal mutations and those linked to them to fixation (Nijhout and Paulsen 1997, Nielsen 2005). Alleles that are associated with a given trait will be among the first to fix and be detectable by differences in allele frequency between populations (Nijhout and Paulsen 1997, Akey et al. 2002, Nielsen 2005, De Kovel 2006). By sequencing the genomes of selected and

unselected lines, we were able to look for these differences in allele-frequency between lines using scans of pairwise F_{ST} (Weir and Cockerham 1984) with the understanding that regions of high F_{ST} relative to the rest of the genome are likely to be those acted on by selection (Akey et al. 2002). We used hapFLK analysis (Bonhomme et al. 2010, Fariello et al. 2013) to identify local haplotype clusters acted on by positive selection. We first ran hapFLK on each of the 16 chromosomes individually across all populations to create pairwise Reynolds' distances between populations. Using this kinship matrix, we used 20 haplotype clusters and scanned across each chromosome for 20 expectation maximization (EM) iterations with hapFLK using our baseline population as the outgroup. We estimated significance using chi-squared density and we corrected for False Discovery Rate by using Storey's Method (Storey and Tibshirani 2003) and taking only $P < 0.000001$ ($Q < 0.01$). We estimated the integrated haplotype score (iHS) (Voight et al. 2006) using the R package rehh (Gautier et al. 2016). We estimated the shift in the allele frequency spectrum within selected populations using Tajima's D (Tajima 1989) within 1 Kb windows as estimated through VCFTOOLs v1.11 (Danecek et al. 2011). We compiled each of these three measures of selection into a single statistic, the single composite statistic (CSS) (Randhawa et al. 2014). We scanned each chromosome using a running median of 101 SNPs and extracted all regions with a ($-\log_{10}$(CSS P) > 1.3). Any region that was within 5 Kb of any other significant region was pooled. For these methods, and all other methods requiring phased data, we phased all queen genotypes together for each chromosome individually using SHAPEIT v2.2 (O'Connell et al. 2014) with the additional options --rho 0.39 --window 0.5.

Comparisons to Previous Hygienic Behaviour Associations: QTLs, Association Maps, and Differentially Expressed Genes

Broad Quantitative Trait Loci (QTLs) have been previous identified for hygienic behaviour (Lapidge et al. 2002, Oxley et al. 2010). We tested to see if SNPs and genes acted on by selection in our analysis localized to these broader regions. We re-mapped QTL regions based on microsatellites by using BLASTN to identify the homologous regions within the most recent release of the honey bee genome (Oxley et al. 2010). We also tested if associated genes within our analysis could be found in previous reports of differentially expressed brain genes (Boutin et

al. 2015) and proteins (Guarna et al. 2015) in hygienic honey bees. To quantify selection acting since the split of *A. mellifera* from its sister species *A. cerana* we used previous estimates of the selection coefficient ($\gamma = 2N_e s$) on most genes within the honey bee genome (Harpur et al. 2014b); a selection coefficient greater than one is more indicative of positive selection driving the fixation of beneficial alleles.

Phenotype Association Analysis

We targeted our association analyses to quantify the relationship between haplotypes and the quantitative expression of hygiene within the 132 regions acted on by selection. Haplotype analysis was performed within the baseline population only for a moving 3 SNP window using PLINK v 1.07 (–hap) (Purcell et al. 2007, Chang et al. 2015). We extracted all 1443 haplotypes that were significantly associated ($P < 0.05$) with hygienic behaviour. We then estimated the frequency of all 3-SNP haplotypes within the selected populations (--hap-freq) and compared the frequency of haplotypes across the genome to those within the 99 regions and those that were within the 99 regions and associated with hygiene in control populations.

Admixture analyses

We scanned the genome for evidence of differential admixture between selected and baseline populations and within North American populations using ELAI v 1.0 (Guan 2014). For each chromosome, we estimated local ancestry using the recommended default parameters of ELAI and assuming 200 generations since the initial admixture of source populations. Each run included both selected and baseline populations together as well as an independent set of 6 diploid bees from Ontario.

GO Analyses

We used DAVID (Huang et al. 2009) to identify if our gene set was enriched for Gene Ontology (GO) terms. All tests we performed using *Drosophila* homologs identified with BLASTP match (E-value threshold 1e-10) and because of our limited gene list, we accepted any GO term with $P < 0.1$.

Statistical Analyses

All statistical analysis was performed with R v3.30 (R Core Team 2010). All scripts and workflows are available either as Supplemental Material or on GitHub (https://github.com/harpur/HygSel). Statistical tests are reported within text and we performed parametric tests where data permitted such analysis, otherwise we report non-parametric results.

Figure 4.1. A. Result of Freeze-Killed Brood (FKB) Assay for two colonies showing left panel = low uncapping and removal rates after 24 hrs and right panel = high uncapping and removal after 24 hrs. The FKB assay is performed by freezing a designated section of capped honey bee brood (see left image) with liquid nitrogen. Once thawed, the frozen section is placed back inside the colony for 24 hrs when the section is removed once more and the number of uncapped and removed cells is counted. Hygienic performance is the percentage of cells uncapped and/or removed divided by the number initially frozen. **B.** Violin-Boxplot of hygienic response for the 41 colonies included in this study with selected populations pooled. Boxplots inside violin plots show the median (center line), first and third quartiles (box top and bottom) and the minimum and maximum. Violin plots show the probability density of the data at each data point.

Figure 4.2. Selection map highlighting regions associated with hygienic behaviour. Each plot presents the significance of the Composite Selection Statistic (CSS) for a single chromosome. Horizontal, dotted line represents significance cut-off. Red boxes are regions (+/- 1 Mb on either side) that both have significant evidence of positive selection and have evidence of having haplotypes that associated with hygiene within baseline populations. Top-level horizontal bars are QTL regions for hygienic behaviour (Oxley et al. 2010, Tsuruda et al. 2012) and below those QTLs for hygiene-associated behaviours of uncapping and brood removal (Oxley et al. 2010). Dots are the location of SNPs tentatively associated with hygiene (Spotter et al. 2012, Spotter et al. 2016).

Figure 4.3. A Proportion of C-lineage ancestry at hygienic loci within selected populations compared with baseline populations. Y-axis represents the proportion of C-lineage ancestry in selected populations minus that of the baseline population; increasing values are indicative of more C-lineage ancestry in the selected populations **B.** This is a pattern that we also found within highly hygienic North American populations not included within our artificially selected populations. ("***" indicates $P<0.01$).

Figure 4.1

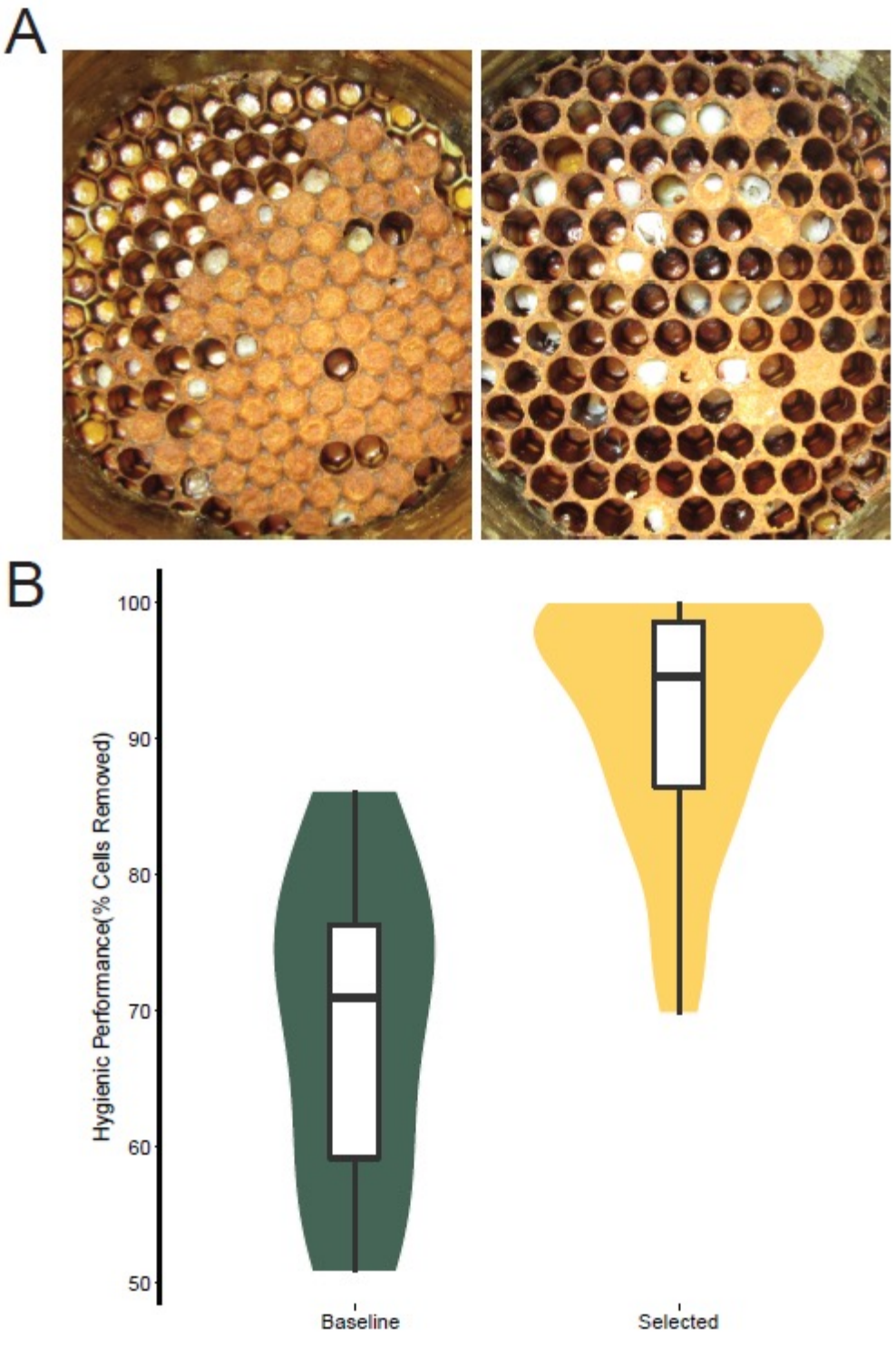

A
B
Hygienic Performance(% Cells Removed)
100
90
80
70
60
50
Baseline
Selected

Figure 4.2

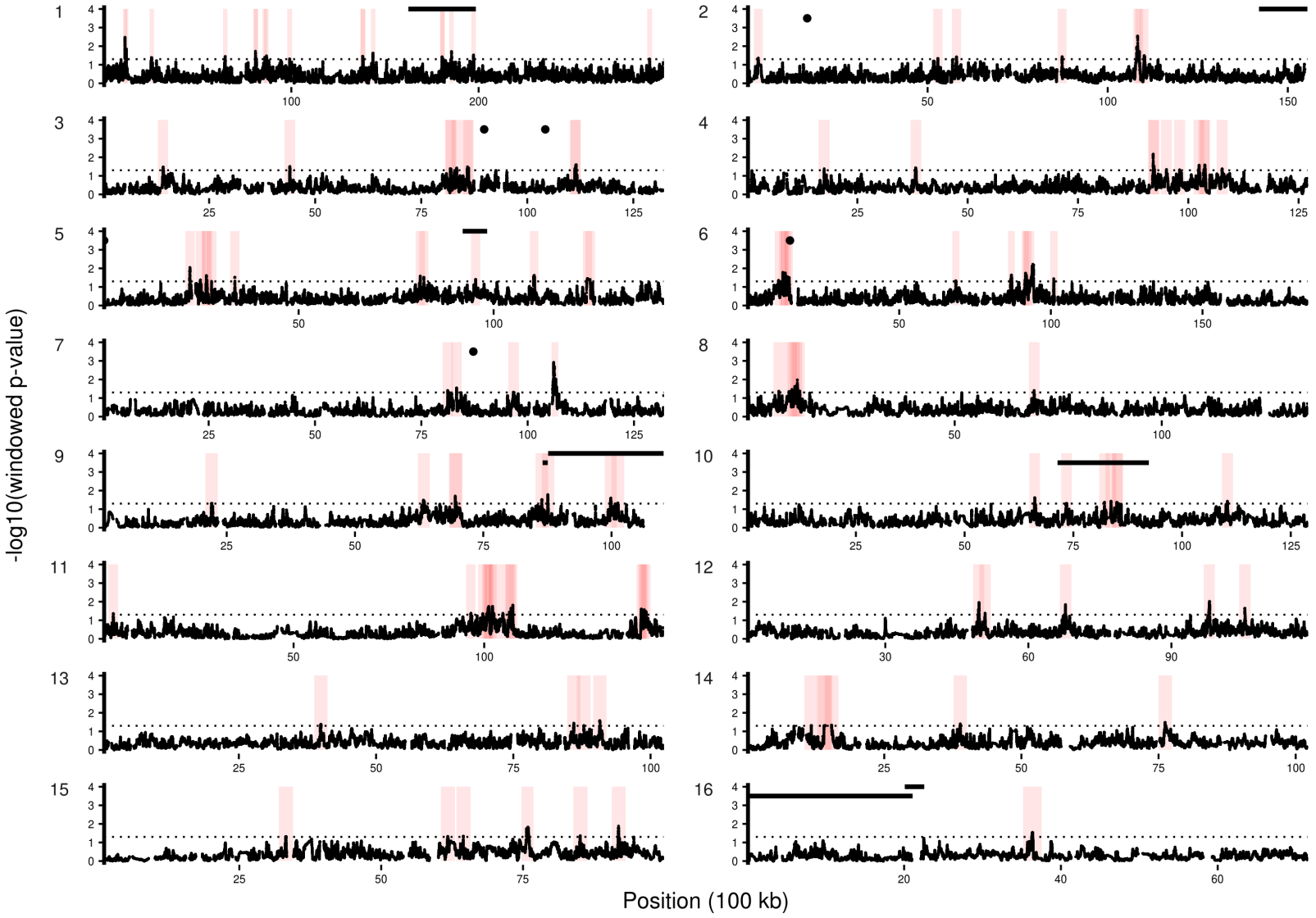
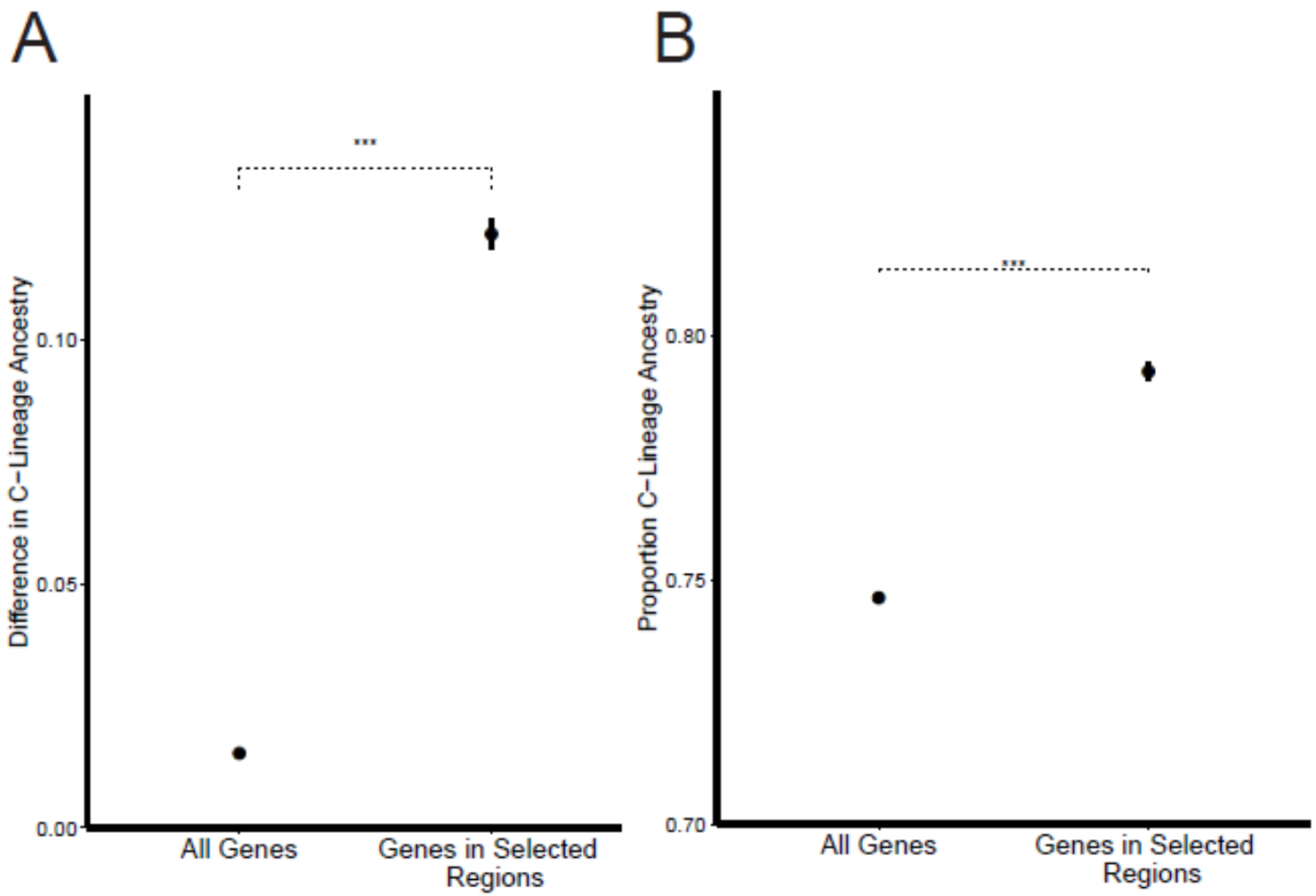

-log10(windowed p-value)
Position (100 kb)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16

Figure 4.3


A
B
***
***
Difference in C−Lineage Ancestry
Proportion C−Lineage Ancestry
0.00
0.05
0.10
0.70
0.75
0.80
All Genes
Genes in Selected Regions
All Genes
Genes in Selected Regions

Chapter 5:

A variant reference data set for the Africanized honeybee, *Apis mellifera*

Samir M. Kadri, Brock A. Harpur, Ricardo O. Orsi, Amro Zayed[4]

Background & Summary

The Western honey bee (*Apis mellifera*) was introduced to North and South America from Old World populations in the early 18th century (Tarpy et al. 2015). In its native range, the honey bee is divided geographically and genetically into five ancestral lineages – the M and C lineages of Europe, the A lineage of Africa, and the Y and O lineages of Asia (Ruttner 1988, Garnery et al. 1992, Arias and Sheppard 1996, Whitfield et al. 2006a, Harpur et al. 2014b) – that encompass approximately 22 subspecies (Ruttner 1988).

European settlers, in the early 18th century, introduced subspecies of the M lineage (*A. m. mellifera* and *A. m. iberica*) into North America (Sheppard 1989a, b). By the 20th century, C lineage (*A. m. ligustica* and *A. m. carnica*) and some O lineage (*A. m. caucasia*) subspecies were introduced (Sheppard 1989a, b). It was during this century that Brazilian beekeepers first imported honey bees, chiefly *A. m. mellifera* and *A. m. carnica*, followed by *A. m. ligustica* and *A. m. caucasia* (Crane 1999). These subspecies were used exclusively in Brazil until 1956 when *A. m. scutellata* was imported from Africa for breeding and genetics research. Several mated *A. m. scutellata* queens arrived from South Africa and one from Tanzania (Nogueira-Neto 1964) to breeding stations in Rio Claro, São Paulo, Brazil. The intention of the breeding program was to cross *A. m. scutellata* with commercial stock to serve as a base population in selection programs

[4] This published manuscript has been reprinted with permission from its co-authors and publisher from the original manuscript: Kadri SM, Harpur BA, Orsi RO, & Zayed A (2016) A variant reference data set for the Africanized honeybee, *Apis mellifera*. *Nature Scientific Data* 3:160097.

(Kerr 1957). Famously, queens and drones escaped and hybridized with the existing population (Kerr 1957, Winston 1987). These hybrids became one of the most astounding insect invaders in recent history: feral populations of the "Africanized" bees retained the highly defensive trait of their African ancestors and are now the most common honey bee found from Central South America (Brazil and Northern Argentina) to Mexico and the southern United States (Winston 1992).

The sequencing of the honey bee genome in 2006 (Weinstock et al. 2006) was a landmark for the field of sociogenomics and has created valuable resources for the beekeeping industry(Harpur et al. 2012, Chapman et al. 2015b, Harpur et al. 2015, Munoz et al. 2015), but because this genome was derived from a typical admixed North American honey bee (Weinstock et al. 2006) it provides little information about the underlying genetic variation present in Africanized populations.

Here, we present the pooled genomes of 360 AHBs from Brazil along with a reference SNP database for this population. Genomic resources for AHBs will be beneficial for both pure and applied research questions. First, there is a growing need to quickly and reliably detect Africanized colonies to secure international trade in honey bees (Chapman et al. 2015b, Harpur et al. 2015). Second, Africanized bees are highly defensive and they are commonly used for studying the genetics of nest defence (Chandrasekaran et al. 2015). Finally, Africanized bees are highly invasive: within the last 60 years they have become the most common genotype across much of the Southern United States (Winston 1992) and have evidence of adaptive introgression during their invasion (Zayed and Whitfield 2008).

Methods

Sampling and Sample Information

We collected 12 diploid worker bees from the brood frames of each of 30 Africanized honey bee colonies from four apiaries all located at Iaras city, São Paulo, Brazil (Table 5.1), located over 200 km from Rio Claro. These colonies were obtained from natural swarms within the State of São Paulo where the Africanized honey bee invasion began. We performed a single DNA extraction for each colony by pooling ¼ of a thorax from each of the 12 workers (Ferretti

et al. 2013). We used the Mag-Bind® Blood DNA kit (Omega Biotek Store) with the manufacturer´s recommended protocol to extract an average of 4.71 ± 1.42 µg of high-quality DNA from each of the colonies.

DNA samples (one from each colony) were submitted to The Centre for Applied Genomics (Toronto, ON) for library preparation and high throughput sequencing. In brief, DNA was quantified by Qubit HS assay and 200 ng of DNA was used as input material for the TruSeq Nano DNA Sample Preparation protocol (Illumina, Inc.) following Illumina's recommendation. DNA was sheared to 550-bp on average using a Covaris S2 system (Duty cycle: 10%; Intensity 2; Burst per second: 200; Treatment time: 44 seconds; Mode: Frequency sweeping). The sheared DNA was end-repaired and the 3' ends were adenylated prior to ligation of the TruSeq adapters. The library was enriched by PCR using different indexed adapters to allow for multiplex sequencing using the following conditions: 95^{o}C for 3 minutes followed by 8 cycles of 98^{o}C for 20 seconds, 60^{o}C for 15 seconds and 72^{o}C for 30 seconds, and finally an extension step at 72^{o}C for 5 minutes.

Final TruSeq Nano DNA genomic libraries were validated on a Bioanalyzer 2100 DNA High Sensitivity chip (Agilent Technologies) for size and by qPCR using the Kapa Library Quantification Illumina/ABI Prism Kit protocol (KAPA Biosystems) for quantities. Ten libraries were pooled in equimolar quantities and sequenced on a HiSeq 2500 platform on a high throughput flowcell with the Illumina TruSeq V4 sequencing chemistry following Illumina's recommended protocol to generate paired-end reads of 150-bases in length.

Genome Alignment and Variant Calling

Each colony's sequenced reads (N = 30 colonies) were trimmed of Illumina Adaptors using Trimmomatic v0.32 then aligned to the most recent version of the honey bee genome AMEL_4.5 (Munoz-Torres et al. 2011) using BWA aligner v0.7.5(Li and Durbin 2010). Paired alignments were then merged with SAMTOOLS v 0.1.19 and re-aligned using STAMPY v1.0.21(Lunter and Goodson 2011) with divergence (--d) set at 0.02 . We marked and removed duplicate reads with PICARD v 1.141 and re-aligned around indels using GATK IndelRealigner v 3.1 (DePristo et al. 2011) (Figure 5.1; Data Citation 1).

To identify all variants found within our samples, we used two independent variant callers (Figure 5.1): VARSCAN(Koboldt et al. 2009) v2.3.7 and GATK UnifiedGenotyper (Data Citations 2 and 3) We used GATK set to --ploidy 2 to identify only the location of variants and were unconcerned with specific genotype calls. We called all variant sites in GATK and VARSCAN using default parameters. We removed indels and all SNPs within 10 bp of indels, removed all unmapped scaffolds (Scaffolds 17.XXX or GroupUn) and mitochondrial sequence (Scaffold 18.1), and removed SNPs of low quality (Q < 25) or in areas of low genomic complexity, thus reducing the potential for calling erroneous SNPs due to paralogous sequence or misaligned reads (Harpur et al. 2014b) (Figure 5.1). We retained all SNPs that were identified using both variant callers and that passed our conservative filtering procedures, above. Because our data consist of pooled sequence for 30 colonies, we report the allele frequency at each site as called by VARSCAN in Variant Call Format (Data Citation 3).

SNP Validation: Population Differentiation and Admixture

To quantify differentiation among contemporary Africanized populations in Brazil and ancestral honey bee populations, we used POPOOLATION2 v1.201 (Kofler et al. 2011).We created a single input file containing our Africanized bee samples pooled into a single alignment as well as population-pooled alignments from ancestral honey bee populations from Africa (A lineage, N= 11) and Europe (M lineage, N = 9; C lineage N = 9). The latter sequence data were obtained from a recent honey bee population genomics study performed by our group(Harpur et al. 2014b) and represents ancestral populations from which Africanized populations are derived. We generated a single MPILEUP file and extracted from it the 3,606,720 SNPs identified above in our Africanized honey bee samples. We estimated pairwise population differentiation on all sites with --min-count 6 --min-coverage 100 --max-coverage 800 --min-covered-fraction 0.8.

Code availability

We have not used any custom code and relied on previously available, validated, software packages; however, we have left our general pipeline available for re-use at the author's GitHub (https://github.com/harpur/afz/blob/master/AHBPipeline.sh)

Data Records

We have curated a set of 3,606,720 SNPs identified in 360 Africanized honey bees across 30 colonies (Data Citations 2 and 3). The data consist of SNPs called across the most recent honey bee reference genome (Amel_4.5(Munoz-Torres et al. 2011)) in Variant Call File format on placed scaffolds. Because we utilized a pooled-sequencing method, all variant sites include the frequency of each alternate allele call for each colony. All sequence data are also available in BAM format (Data Citation 1; Table 5.1) allowing subsequent researchers to use updated SNP calling and genotype software when available.

Technical Validation

To validate that our samples are indeed Africanized and to confirm our SNP calls, we compared our current SNPs to those of a previous honey bee population genomics study that sequenced and analysed honey bee samples using similar methods as described herein (Harpur et al. 2014b). Africanized bees are known to be derived from three of the major honey bee population groups: A, M, and C, (Whitfield et al. 2006a, Harpur et al. 2015). We found that 99.8% of the 3,606,720 SNPs found in AHBs, were also found within one or more of these ancestral populations. Africanized populations are expected to have higher A lineage ancestry relative to C and M lineage ancestry. Using a regression model (Chiang et al. 2010), we demonstrated that allele frequencies within Africanized bees are more correlated with A lineage allele frequencies (GLM; $r = 0.529$, $p < 2.2x10^{-16}$) relative to both M lineage allele frequency ($r = 0.102$, $p < 2.2x10^{-16}$) and C lineage allele frequency ($r = -0.08$, $p < 2.2x10^{-16}$; Figure 5.2), as we would expect from an Africanized population. As well, we find that AHB and A lineage are more similar genetically (Fst = 0.02) than AHB vs M-lineage (0.04) and AHB vs C-lineage (0.05).

Data Citations

1 *NCBI SRA* BioProject PRJNA324081 (2016).

2 *NCBI dbSNP* Batch ID 1062539 http://www.ncbi.nlm.nih.gov/projects/SNP/snp_viewBatch.cgi?sbid=1062539 (2016).

3 Harpur, B.A., Kadri, S. M., Orsi, R. O. & Zayed, A. *Figshare* https://figshare.com/s/d411a20130d4d4be2863 (2016).

Figure Legends

Figure 5.1. Overview of alignment and SNP calling pipeline

Figure 5.2.Correlation of allele frequencies between the Brazilian Africanized honey bee population and A) A-lineage bees B) C-lineage bees and C) M-lineage bees. Red line shows results of linear model fit (GLM, $_{F3,78904} = 15470$, $p < 2.2x10^{-16}$).

Tables

Table 5.1: Sample Sequencing and Accession information

Sample ID	Accession No.	Average Sequencing Depth
HDB139	SAMN05194651	29.00
HDB179	SAMN05194655	29.01
HDB199	SAMN05194661	11.67
HDB303	SAMN05194664	12.53
HDB175	SAMN05194654	21.95
HDB302	SAMN05194663	23.25
HDB187	SAMN05194657	21.09
HDB191	SAMN05194659	33.12
HDB189	SAMN05194658	25.23
HDB288	SAMN05194662	13.74
HDB195	SAMN05194660	20.51
HDB148	SAMN05194652	25.31
HDB150	SAMN05194653	24.98
HDB183	SAMN05194656	6.82
HDB30-S	SAMN05194665	16.55
LDB6	SAMN05194666	18.04
LDB127	SAMN05194667	18.68
LDB136	SAMN05194668	18.81
LDB29	SAMN05194669	17.73
LDB162	SAMN05194670	23.04
LDB9	SAMN05194671	21.99
LDB153	SAMN05194672	19.96
LDB8	SAMN05194673	15.48
LDB23-S	SAMN05194674	30.04
LDB181	SAMN05194675	25.04
LDB35-S	SAMN05194676	18.42
LDB40-S	SAMN05194677	21.15
LDB196	SAMN05194678	17.18
LDB5	SAMN05194679	12.11
LDB221	SAMN05194680	13.64

Figure 5.1

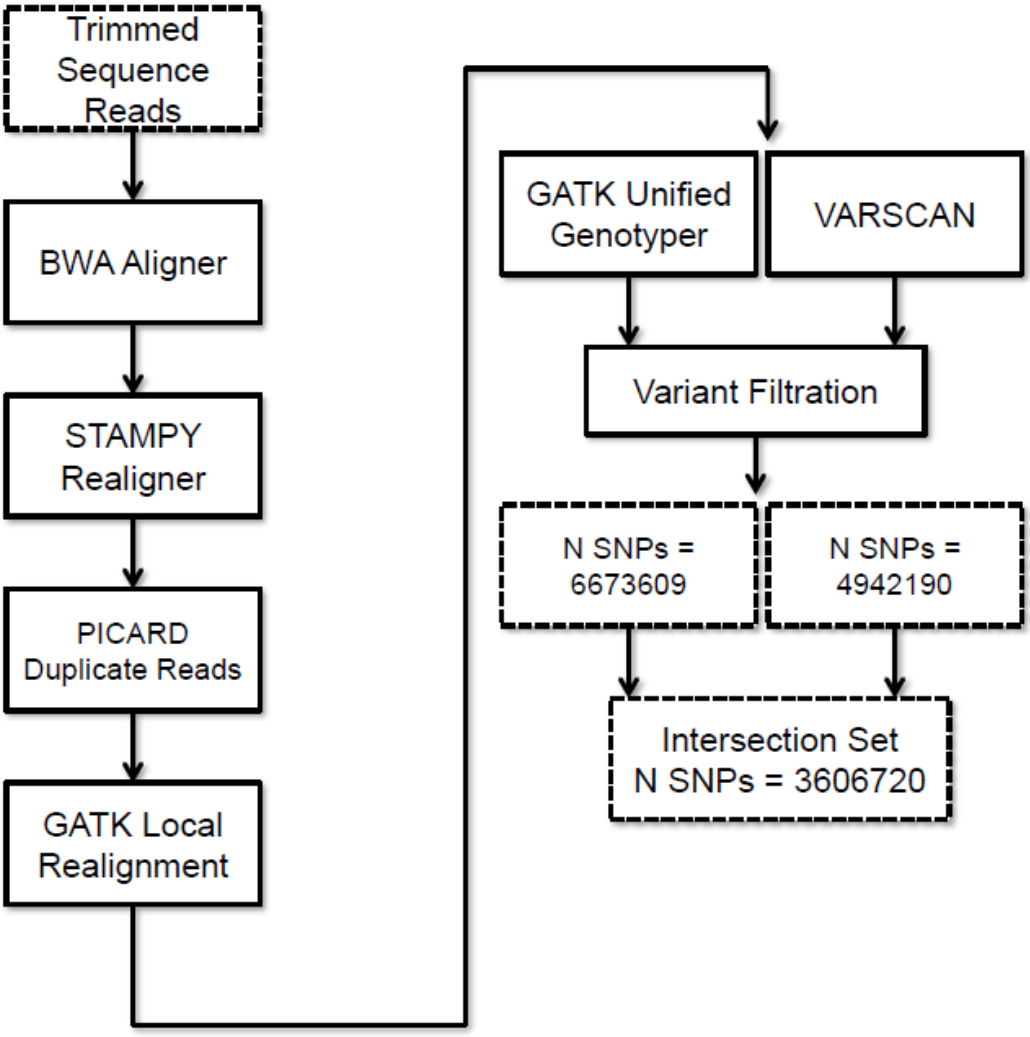
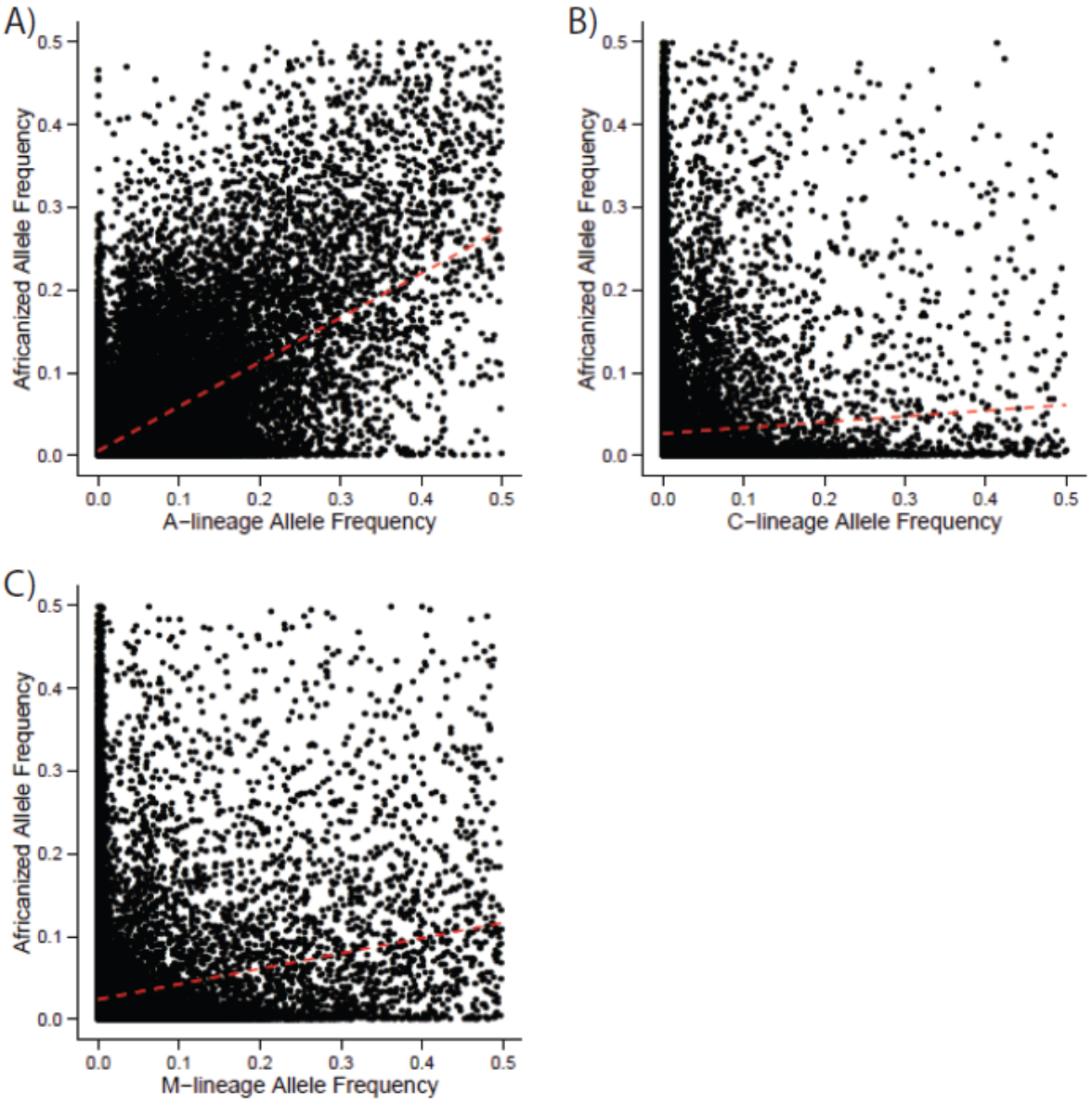

Trimmed Sequence Reads
BWA Aligner
STAMPY Realigner
PICARD Duplicate Reads
GATK Local Realignment
GATK Unified Genotyper
VARSCAN
Variant Filtration
N SNPs = 6673609
N SNPs = 4942190
Intersection Set
N SNPs = 3606720

Figure 5.2


A)
Africanized Allele Frequency
A−lineage Allele Frequency
B)
Africanized Allele Frequency
C−lineage Allele Frequency
C)
Africanized Allele Frequency
M−lineage Allele Frequency
0.0
0.1
0.2
0.3
0.4
0.5

Chapter 6:

Defence response in Africanized honey bees (*Apis mellifera* L.) is underpinned by complex patterns of admixture

Brock A. Harpur, Samir M. Kadri, Ricardo O. Orsi, Charles W. Whitfield, Amro Zayed

Introduction

Among the most successful and most publicized invasive insects to date is the Africanized or "Killer" honey bee (*Apis mellifera* L.). The Africanized honey bee, the most common honey bee from South America to the Southern United States, originated from a Brazilian research facility in 1956. At the time, Brazil's commercial stock was an admixed population that originated from at least two European sources: honey bee subspecies of the M-lineage (e.g. *A. m. mellifera and A.m. iberiensis*) and C-lineage (e.g. *A. m. ligustica* and *A. m. carnica*) (Kerr 1957, Sheppard 1989a, b, Crane 1999). Researchers hoped to produce a docile, subtropical-adapted honey bee by crossing commercial populations with South African honey bee subspecies (*A.m. scutellata*; A-lineage) (Kerr 1957, Nogueira-Neto 1964, Winston 1987, 1992). Unfortunately, the resulting managed-African hybrids were less desirable than hoped. They swarmed often, they absconded more frequently, and they were highly defensive.

Following their escape 61 years ago, the Africanized honey bee (AHB) has expanded across Brazil, north into Mexico and into the Southern United States. Within Brazil, their spread and establishment has purged almost all of the C-lineage progenitors and today AHB populations are characterized by a genome consisting of mostly A-lineage alleles (~80%) with the remainder from the M-lineage (Clarke et al. 2002, Whitfield et al. 2006a, Zayed and Whitfield 2008, Chapman et al. 2015b). As well, they have maintained their high defensiveness across their range, with the exception of a gentle AHB population in Puerto Rico (Rivera-Marchand et al. 2012). Their high defence response has been of major concern for apicultural industries and the

public (Winston 1992, Schneider et al. 2004) particularly as AHB is the most common honey bee from Northern Argentina to the Southern United States (Winston 1992).

Decades of research has investigated the genetic underpinnings of defence response in honey bees. The first quantitative genetic analyses used crosses between highly defensive AHBs and less defensive European honey bees (of mixed European background), a comparison used to this day to identify the underlying genetics of defence response in honey bees (Hunt et al. 1998, Hunt et al. 1999, Breed et al. 2004, Alaux et al. 2009, Chandrasekaran et al. 2011, Li-Byarlay et al. 2014, Chandrasekaran et al. 2015, Gibson et al. 2015). Hunt's foundational work demonstrated that components of the honey bee's defence response and the production alarm pheromone are all high heritability and influenced by at least 15 broad Quantitative Trait Loci (QTLs) (Hunt et al. 1998, Hunt et al. 1999). Those initial studies, and more recent crosses also revealed parent-of-origin effects for defence response: more defensive colonies originated from European-derived queens mated to African drones than the inverse cross (Breed et al. 2004, Gibson et al. 2015), and potential epistasis acting among QTLs (Breed et al. 2004). The parent-of-origin effects are likely the result of differentially-expressed gene clusters within two of the major-effect QTL regions known as *sting-1* on chromosome 3 and *sting-2* on chromosome 12 (Gibson et al. 2015). Finally, gene expression studies using similar comparisons have revealed that core metabolic genes are associated with defence response: bees exposed to alarm pheromone and bees more likely to express defense response down-regulate genes associated with oxidative phosphorylation in the brain while up-regulating those of the glycolytic pathway (Chandrasekaran et al. 2011, Li-Byarlay et al. 2014, Chandrasekaran et al. 2015, Rittschof et al. 2015b). These studies are invaluable for our understanding of the genetics of defense response; however, none have explored how defense response varies neither within AHB nor within the ancestral populations from which New World bees originated.

AHB provides a unique opportunity to explore how introgression of long-separated populations can influence phenotypic evolution in an invasive population (Rius and Darling 2014). In the case of AHB, a single introgression event resulted in a hybridized population with clear phenotypic differences and higher fitness than local populations (Winston 1992). There is evidence that admixture contributes to phenotypic diversity within AHB. A previous study found

evidence of a non-random pattern of admixture in AHB suggesting that some combination of European and African alleles were adaptive in the invasive AHB population, but the study employed a small number of markers and was unable to determine specific traits underlying adaptive admixture (Zayed and Whitfield 2008).

By using a recently curated population genomics data set for AHB (Kadri et al. 2016) and pairing that with phenotypic data, we explored which genes are associated with defence response within AHB and how differential admixture within these genes contributes to variation in defence response. We demonstrate that defence response is not simply the product of more A-lineage ancestry but rather an interaction between M- and A-lineage ancestral alleles in the genome.

Materials and Methods

Defence Response Assay and Sampling

We quantified the defence response of 116 Africanized honey bee colonies from four apiaries within São Paulo State (Figure 6.1). Two weeks prior to testing we standardized colonies within Langstroth boxes to consist of seven brood frames and three nectar frames. Each colony was assayed for defence using the Black Suede Ball test (Stort 1975). The test is performed by swinging a suede ball in front of the colony entrance causing bees to react and sting it. After one minute, the ball is removed and the stings remaining within the ball are counted. We repeated this test three times over three days and averaged the number of stings across the tests. From each colony, we then collected at least 15 workers from the brood chamber in 95% ethanol and stored them at -80C until gDNA extraction.

Genome Sequencing, Alignment, and SNP Calling

From the sample of 116 colonies, we extracted high-quality genomic DNA from the fifteen most and fifteen least defensive colonies (top and bottom ~10% of the data; Figure 6.1). Each colony's genomic DNA was extracted as a pool of 12 worker legs. Each of these pools was then sequenced with Illumina Hi-Seq 2500. The resulting data set and detailed bioinformatics

methods are available as an open-access resource (Kadri et al. 2016). In brief, we aligned the reads for each colony individually to the most recent version of the honey bee genome AMEL_v4.5 (Elsik et al. 2014) using BWA v0.7.5 (Li and Durbin 2010) and STAMPY v1.0.21(Lunter and Goodson 2011) and jointly called SNPs using VARSCAN v2.3.7 (Koboldt et al. 2009) and GATK UnifiedGenotyper (DePristo et al. 2011). All alignment and SNP calling was performed jointly on the high-defence and low-defence cohorts.

Differentiated Sites

Sites associated with defence response are expected to have significant differences in allele frequency between the high- and low-defense cohorts. To identify such sites, we calculated the pairwise fixation index (Fst) between the two cohorts and the difference in allele frequency at each site using POPOOLATION 2 (Kofler et al. 2011). We estimated local Fst within a region of 301 SNPs using a running median, and deemed any region of the genome as highly divergent if it had an average Fst greater than 99.95% of Fst values (median Fst > 0.0063) across the genome, was within 15 Kb of any other site with similarly high Fst, and contained at least 3 outlier Fst SNPs (Fst > 95% of all Fst values). To determine if the difference in allele frequency between cohorts was significant, we utilized a permutation test that randomly sampled 100000 sets of SNPs from across the genome to determine the expected distribution of allele frequency difference for similar sized sets.

Estimating Local Ancestry

To estimate levels of local ancestry (introgression of A- or M-lineage alleles across the genomes of each AHB sample), we made use of ANCESTRY_HMM (Corbett-Detig and Nielsen 2017). This method estimates local ancestry within samples of arbitrary ploidy through the use of a hidden Markov Model. We first extracted ancestrally informative markers from the A- and M-lineages that were at least 5Kb apart (Harpur et al. 2014b). We estimated the recombination rate between markers in the ancestral populations using a recent recombination map (Liu et al. 2015). For each informative site, we extracted read depth from AHB samples and performed runs of ANCESTRY_HMM with a range of Ne from 60 to 10000. We repeated this analysis for C-lineage ancestry to confirm that C-lineage alleles have largely been purged from AHBs.

For each cohort, we estimated the mean level of M-lineage ancestry at each site, above. Between cohorts we estimated the difference in M-lineage ancestry between the two cohorts as Dm = mean(M-lineage ancestry in high-defence cohort) – mean(M-lineage ancestry in low-defence cohort); where $Dm < 0$ there is more M-lineage ancestry at a site in the low-defence cohort (i.e. more A-lineage ancestry in highly defensive cohorts). To estimate significant deviations in the level of admixture between high- and low-defensive cohorts, we used a permutation protocol. From the entire dataset, we randomly sampled 2 sets of 15 colonies without replacement. For each site in the genome of this sample, we estimated the mean M-lineage ancestry and Dm. After 10000 permutations, a significantly differentiated site was then any site that has a two-sided probability < 0.001. This corresponded to any site with $|Dm| \geq 0.08$, or an 8% difference in ancestral allele frequency (henceforth "ancestry") between the two cohorts.

Re-Mapping QTL regions and Genes with Reported Defence-Response Association

We remapped all six previously reported defence response QTL regions that had been reported with microsatellite markers (Hunt et al. 1998, Hunt et al. 1999, Guzman-Novoa et al. 2002, Lobo et al. 2003, Gibson et al. 2015). We used the microsatellite sequence and extracted BLASTN matches against the honey bee genome (E-value< 1e-5). For all analysis involving QTLs, we added an additional 50Kb to either side of the QTL site to reflect uncertainty in the exact position of the causal mutations underlying each QTL (Lynch and Walsh 1998).

Statistics and Gene Ontology Analyses

We performed hypergeometric tests with DAVID 6.8 (Huang et al. 2009) to identify if our gene set was enriched for Gene Ontology (GO) and KEGG pathway terms using *Apis mellifera* gene calls against a background of all genes in the honeybee genome. We exported any result with $P < 0.05$. All tests were performed in R v 3.3.2 (R Core Team 2010) and were parametric unless otherwise stated.

Results

Identifying Defence Response-Associated Sites

Assuming defense response is heritable (Hunt et al. 1998, Hunt et al. 1999), mutations that influence defensiveness should exhibit differences in allele frequency (i.e. genetic differentiation) between the high and low defensive cohorts. Conversely, as these samples were obtained from a contiguous population, regions of the genome that are not associated with defence response should have relatively low levels of genetic differentiation. As expected, the high and low defensive cohorts exhibited virtually no genetic differentiation at most SNPs (Average Fst between the two cohorts was 0.0064 +/-0.0086 SD). However, in the most extreme instances (greater than the 95% quantile), we find mutations with Fst > 0.041, a 9% difference in allele frequency, and as high as Fst = 0.16, a 38.3% difference in allele frequency between the two populations.

After scanning the genome, we identified 63 genomic loci containing 285 genes with relatively high levels of genetic differentiation between the high- and low-defensive cohorts (Figure 6.2; Table 6.S1; hereafter called defence associated loci). These regions had, on average a difference in allele frequency of 7.1% between the two cohorts. SNPs with the highest Fst within each of the 63 loci (Tables 6.S2) have, on average, a difference in allele frequency of 22% between the high and low cohorts. This difference is significantly higher than any other similar sized set of SNPs chosen at random from the AHB genome (Permutation Test N = 100000; mean difference = 0.052; $P < 0.001$).

We re-mapped 6 loci that had been identified with microsatellite markers (Hunt et al. 1998, Hunt et al. 1999, Guzman-Novoa et al. 2002, Lobo et al. 2003, Gibson et al. 2015). Our defence associated loci overlapped with or were within 50Kb of at least 2 of the previously-reported QTL: the QTL on chromosome 12 influencing the production of the primary alarm pheromone component, and on chromosome 3 that associates with defense response (Figure 6.2). Taken together, these data provide evidence that the regions of the genome we have identified quantitatively contribute to defence response in AHB.

Defence Associated Regions can be differential admixed

If non-random patterns of admixture between highly- and less-defensive cohorts contribute to quantitative differences in defence response, we would expect to find different ancestral alleles segregating between the two cohorts. To test for this, we made use of a recent

method developed to estimate ancestral proportions in pooled sequencing data (Corbett-Detig and Nielsen 2017). Our admixture mapping procedure confirmed that, on average, AHBs have 86% of their genome originating from the A-lineage and the remainder almost entirely originating from the M-lineage (Weinstock et al. 2006). We found little evidence of C-lineage ancestry remaining within AHBs, supporting previous findings (Clarke et al. 2002, Whitfield et al. 2006a, Zayed and Whitfield 2008). Across the genome, 1.5% of alleles originate from the C-lineage on average. Because we estimated the level of C-lineage ancestry for each sampled colony, we are able to compare ancestral allele frequencies between the two cohorts at each site across the genome—estimating differential ancestry in a given region. We found no significant evidence of a genome-wide difference in C-lineage ancestry between the two cohorts across all sites ($t = -0.68$; $P = 0.49$). If we repeat this analysis within only defence-associated regions of the genome, we again find no significant evidence of different levels of C-lineage ancestry between the two cohorts ($t = 1.85$; $P = 0.064$) and at most a 3% difference in ancestral allele frequency between the two cohorts in these regions.

When we compared the level of M-lineage ancestry between the two cohorts, we found that less-defensive colonies (mean M = 14.0%) had a slightly, but significantly higher levels of M-lineage ancestry genome-wide when compared to highly-defensive colonies (mean M = 13.8%; $t = 3.37$; $P = 0.0008$). This slight level of differential admixture between the two cohorts is likely driven by distinct regions of the genome that are differentially admixed. To investigate further, we quantified if the 63 defence-associated regions overlapped with regions of the genome that had large differences in ancestral allele frequencies between the two cohorts. We found significant differences in both M- and A-lineage allele frequencies between the high- and low-defense cohorts ($|Dm| > 0.08$) within the 62 defense associated regions (Figure 6.2). Only a single associated region on chromosome 9 had higher A-lineage ancestry within more defensive colonies while all others had higher M-lineage ancestry in more defensive colonies (Figure 6.2). This was also true for defence associated loci within or nearby previously mapped QTLs. The locus on chromosome 3 was enriched for M-lineage ancestry in the highest defensive cohort. This indicates that the variation in defensive behaviour within AHB is the result of differences in ancestry from both ancestral lineages and not simply a product of A-lineage alleles.

Defence-Associated Genes Have Metabolic Function

We next determined if the set of 285 candidate genes have previously been found to be expressed or regulated differentially in defensive honey bees. Several recent studies identified genes and gene networks that are differentially expressed in guards, soldiers, or bees exposed to alarm pheromone (Alaux et al. 2009, Chandrasekaran et al. 2011, Rittschof et al. 2014, Gibson et al. 2015). We compared our set of 285 genes to each of these sets and found no significant enrichment between genes differentially expressed by guards, by soldiers, by more defensive individual honeybees, nor in the brains of bees exposed to alarm pheromone (Fisher Exact test, $P > 0.1$ for all comparisons) (Alaux et al. 2009, Rittschof et al. 2014, Chandrasekaran et al. 2015).

Although we found no evidence of significant overlap between previous gene expression studies and our own, we did find evidence of common functional categories underpinning defence response in AHBs. Previous works discovered that aggressive bees shift their brain metabolic activity from oxidative phosphorylation to glycolysis (Alaux et al. 2009, Li-Byarlay et al. 2014, Chandrasekaran et al. 2015).We found that variation in defence response is likely a result of differential admixture at genes within or associated with these pathways. We found a slight but significant enrichment of our candidate genes within three metabolic pathways: "Ribosome biogenesis" (KEGG PATHWAY; ame03008), "Galactose metabolism", and (KEGG PATHWAY; ame00052), and "Starch and sucrose metabolism" (ame00500; Hypergeometric Test; $P<0.05$; Table 6.1). Among these genes was a major regulator of glycolysis: *hexokinase2* (GB47079; Figure 6.2, Figure 6.3, Table 6.1, and Table 6.S2). This gene contained 3 SNPs with Fst greater than 0.001 % of all values in the genome between high and low-defence cohorts—an average of 21% difference in allele frequency. All of highly differentiated SNPs fell within introns (Figure 6.3). We also find this gene to be highly differentiated between A- and M-lineage populations (mean Fst = 0.25) relative to the rest of the genome (genomic mean = 0.21; Figure 6.3) (Harpur et al. 2014b).

An Imprinted Gene Cluster is Associated with Defence Response

More aggressive colonies are obtained from crosses of European-derived queens mated to African drones than the inverse cross. This effect is likely driven by imprinting at least two gene clusters, one on chromosome 3 and the other on chromosome 12 (Gibson et al. 2015). Both of these clusters overlap with previously-reported defensive response QTLs (Figure 6.2; Figure 6.4), but the cluster on chromosome 3 contained a region of high Fst between highly-defensive and less-defensive cohorts and more defensive colonies had significantly more M-lineage ancestry (mean M = 0.17) relative to less-defensive colonies (M = 0.11; ANOVA; $P < 2.2e\text{-}16$; Figure 6.4) and up to a 10% difference within the most significant windows within that region.

Discussion

Since at least the 19th century, bee breeders have been intentionally admixing long-separated and highly differentiated subspecies of honey bee (Langstroth 1865, Ruttner 1988, Whitfield et al. 2006a, Harpur et al. 2014b, Byatt et al. 2016). In doing so, they have created semi-naturalized experimental genetic populations within which we can explore how admixture contributes to phenotypic variation and ultimately evolutionary change. Defence response in AHB provides an exciting avenue to explore these questions as it has a clear genetic basis and is often cited as an adaptation contributing to the success of their expansion (Fletcher 1978). Our population genomic analysis of highly- and less-defensive AHB colonies allowed us to improve our knowledge of the genes associated with defensiveness and to shed light on the phenotypic consequences of a complex admixture event.

Our genomic contrasts allowed us to identify 63 loci within the bee genome with substantial levels of genetic differentiation between the most defensive and least defensive AHB colonies. While the number of loci is certainly large considering the few QTLs previously discovered (Hunt et al. 1998, Hunt et al. 1999, Guzman-Novoa et al. 2002, Lobo et al. 2003, Gibson et al. 2015), several lines of evidence suggest that our list of candidate loci contains true positives: 1) many of our candidates fall within or near to previously identified QTLs, 2) our candidates have molecular functions that are consistent with their involved in defensive behaviour based on functional genomic studies of defensive bees and 3) some of our candidates overlapped with regions of the bee genome that that exhibit imprinting in association with

defensive behaviour. Moreover, we applied stringent field and bioinformatics methods to reduce environmental noise and false positives. It is important to note that previous QTL experiments utilized 'far' crosses between AHBs and European bees (Hunt et al. 1998, Hunt et al. 1999), and QTLs tend to underestimate the number of loci, while over estimating their effect size (i.e. the Beavis effect) (Xu 2003). Our study was designed to test for the effect of standing levels of genetic diversity on aggression in the focal AHB population in Brazil. Finally, genomic analysis on the fruit fly *Drosophila melanogaster* identified mutations in more than 50 genes that influence aggression (Edwards et al. 2009) – a similar but likely less complex trait relative to defensiveness in honey bees.

We found non-random patterns of admixture at several defence associated loci, suggesting that a combination of African and West-European alleles play a role in defensive behaviour in AHBs. This may seem at odds with common wisdom that AHBs are defensive strictly because of *A. m. scutellata* alleles. However, our finding that defence response is underpinned by a mosaic pattern of ancestry fits well with the current understanding of both introgression and the genetics of defence response (aggression) in other species, as we will discuss below. We propose two potential hypotheses for future investigation. First, there may be distinctly different ancestral alleles acting on different aspects of defense response within AHB. Both the A- and M-lineages are noted as being defensive (Fletcher 1978, Pinto et al. 2014), and may vary in the behaviours ultimately leading to sting release (Breed et al. 2004). Selection acting on defense response within AHB may have acted on sets of "defense alleles" from each of the two ancestral lineages. Second, ancestral alleles may act epistatically. An interaction between M- and A-lineage alleles between loci may cause higher levels of defense response than A- or M-alleles alone. Selection acting on defense response would then fix alleles of both ancestries. In *Drosophila*, aggression is underpinned by epistatic interactions among at least 50 genes (Zwarts et al. 2011). The same has been observed in honey bees where there is evidence of alleles from AHBs and mixed European populations act epistatically at QTL loci (Breed et al. 2004, Gibson et al. 2015). If these interactions are manifested between M- and A-lineage alleles we may observe a mosaic pattern of ancestry.

Perhaps the clearest example of how M-lineage alleles contribute to defence response in AHBs can be found on chromosome 3. This region was highly differentiated between high and low defensive colonies, was significantly differentially admixed between the two cohorts, and overlapped with a previous QTL for defence response (Figure 6.2; Figure 6.4). Previous research suggests that this locus contributes to variation in defence response within AHB through parent-specific gene expression. When an admixed European queen (C- and M-lineage) is crossed with an Africanized drone, the genes expressed in this cluster have the European queen's genotype (Gibson et al. 2015) and the colonies are significantly more defensive than the inverse cross (Breed et al. 2004, Gibson et al. 2015). We found that the most defensive AHB colonies had significantly more M-lineage ancestry at this gene cluster than less-defensive colonies. Our data suggests that this may be due to this locus being more M-lineage-like in queens that have mated to drones with A-lineage-like alleles at this locus.

When examining our candidate genes as a set, we found that the most significant classification was metabolic function. As others have shown, defence response is correlated with a shift in brain metabolism to aerobic glycolysis (Alaux et al. 2009, Chandrasekaran et al. 2015). Highly defensive bees (A-lineage or AHB) and bees exposed to alarm pheromone have elevated whole-body metabolic rates, a measure of oxidative phosphorylation, and higher rates of glycolysis in the brain (Southwick et al. 1990, Alaux et al. 2009, Chandrasekaran et al. 2015, Rittschof et al. 2015b). Events that lead to defence response in honey bees seem to cause a hypoxia-like brain state (Rittschof and Robinson 2013). During hypoxia, glycolysis is initiated in part by *hexokinase2* (Semenza 2007, Wolf et al. 2011). At the level of the neuron this shift has been suggested to result in changes in excitability (Juge et al. 2010, Li-Byarlay et al. 2014, Chandrasekaran et al. 2015, Valdebenito et al. 2016).

Future studies exploring variation in defence response, or any phenotype in New World honey bees, should consider the important role of introgression to phenotypic diversity (Rius and Darling 2014). New World honey bees originate from at least three ancestral lineages (Whitfield et al. 2006a, Harpur et al. 2015). If the interaction of ancestral alleles contributes to phenotypic diversity in New World populations, this implies that the same alleles may not be influencing the phenotype in ancestral populations in the same way. The pooled-sequencing approach we used

here is particularly useful for identifying genomic variation underpinning colony-level traits such as nest defence. As a super-organism, honey bee colonies are composed of thousands of individuals of up to 20 different patrilines. The interactions between individuals within a colony can have drastic influences on colony-level phenotypes such as aggression (Rittschof et al. 2015a). By creating a pooled "colony genome" we could use within-colony allele-frequencies to look at the ultimate expression of the phenotype across the colony as a whole. This procedure should be very useful in future iterations of association mapping within social insects.

Acknowledgements

This study was funded by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada and an Early Researcher Award from the Ontario Ministry of Research and Innovation (to AZ). SMK was supported by a scholarship from Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP Process 2014/10150-2). BAH would like to thank Joshua Gibson, Greg Hunt, Ernesto Guzman, Gard Otis, and Arian Avalos for their informative discussions on the biology and genetics of defence response in honey bees.

Figure Legends

Figure 6.1. Histogram of average sting response for 116 colonies in Brazil. Each colony was tested for their sting response using the black suede ball assay on 3 different days. A colony's sting response is the number of stings left within a black suede ball in a minute, averaged across the three trial days. We sequenced the genomes of 15 of the most and least aggressive colonies (fewer than 39 stings or greater than 87 stings in a minute).

Figure 6.2. Average Fixation index (F_{ST}) between highly- and less-defensive AHB colonies in Brazil. Red boxes are regions (+/- 1 Mb on either side) that have significant evidence of differentiation between the two cohorts. Grey boxes are QTL regions for defence response (Hunt et al. 1998, Hunt et al. 1999). Black bars show areas which are both highly differentiated and have evidence of enriched M-lineage ancestry within less-defensive colonies (on top of plot) or within highly-defensive colonies (below plot). Red bars show the locations of imprinted gene clusters underpinning defence response. Point on Chromosome 15 marks the start of *hexokinase2* (GB47079).

Figure 6.3. Fixation index (F_{ST}) between (top) highly- and less-defensive AHB colonies in Brazil and (bottom) A-lineage and M-lineage populations within *hexokinase2* (GB47079). Coding direction indicated with arrow, gene region defined by line, grey boxes delineate exons. Dotted line indicates the 95% quantile.

Figure 6.4. Average frequency (+/- standard error) of M-lineage alleles within highly-defensive (circles) and less-defensive (triangles) AHB colonies along an imprinted gene cluster on chromosome 3 associated with defense response. Red box is a region of significantly high Fst between the two cohorts. Black boxes along x-axis are protein-coding gene sequences.

Tables

Table 6.1: Gene Ontology (GO) and KEGG Pathway (ame) terms enriched within genes associated with defence response in Africanized honey bees.

	Term	P	GeneID
ame03008	Ribosome biogenesis in eukaryotes	0.013	GB44445, GB54677, GB52153, GB47469, GB47420
ame00500	Starch and sucrose metabolism	0.022	GB54661, GB47079, GB53384
ame00052	Galactose metabolism	0.030	GB54661, GB47079, GB53384
GO:0005975	Carbohydrate metabolic process	0.068	GB53312, GB54661, GB47079, GB44978, GB53384, GB49439
GO:0035556	Intracellular signal transduction	0.087	GB43729, GB49120, GB53311, GB45036, GB54498, GB49505

Figure 6.1

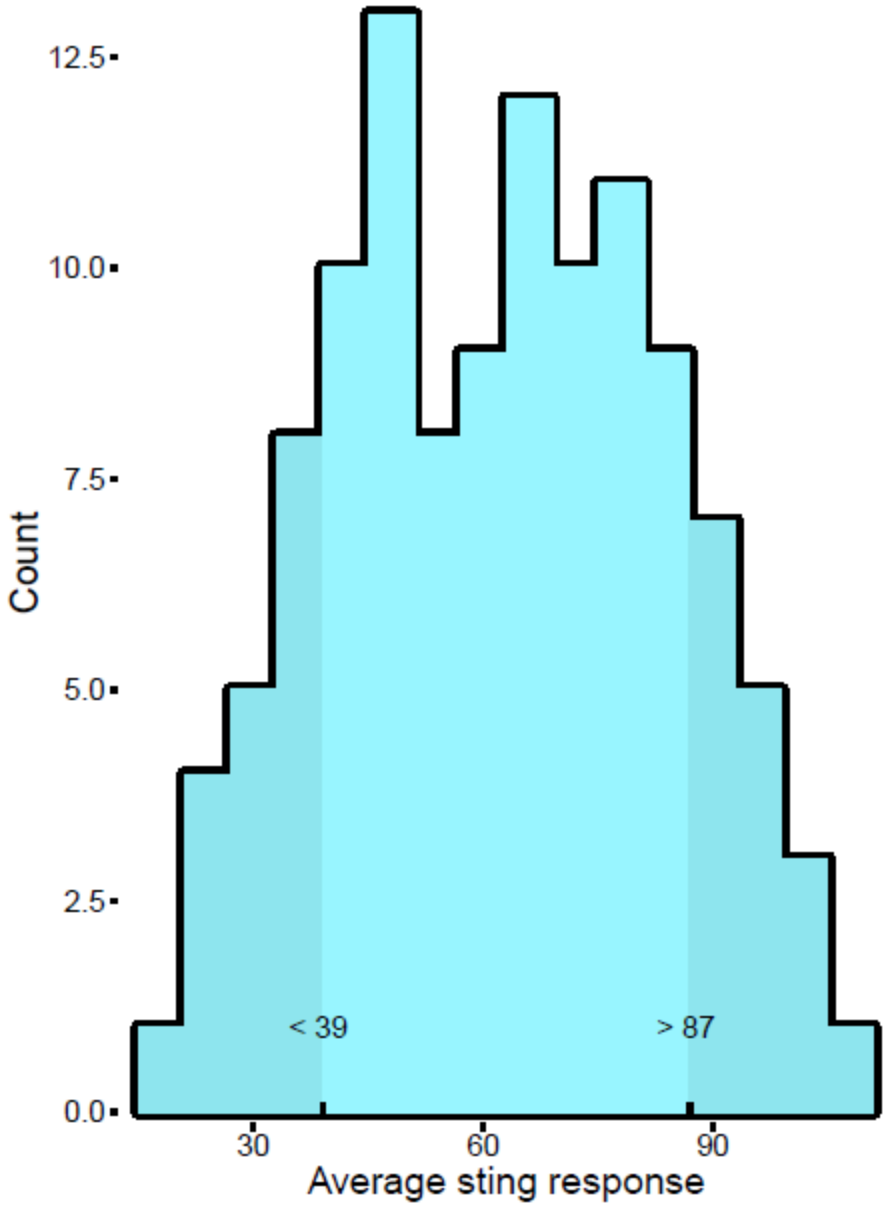


Figure 6.2

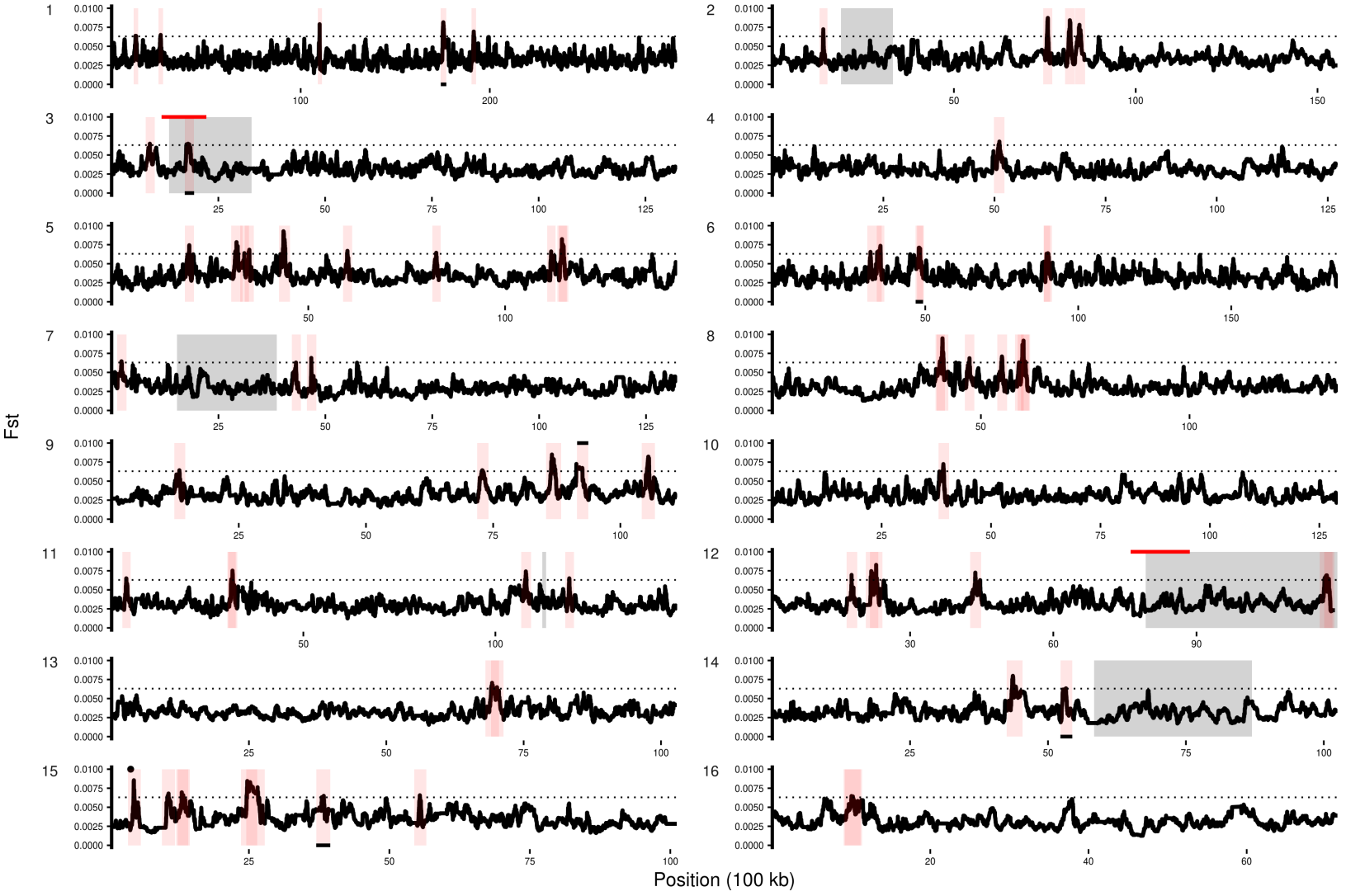


Figure 6.3

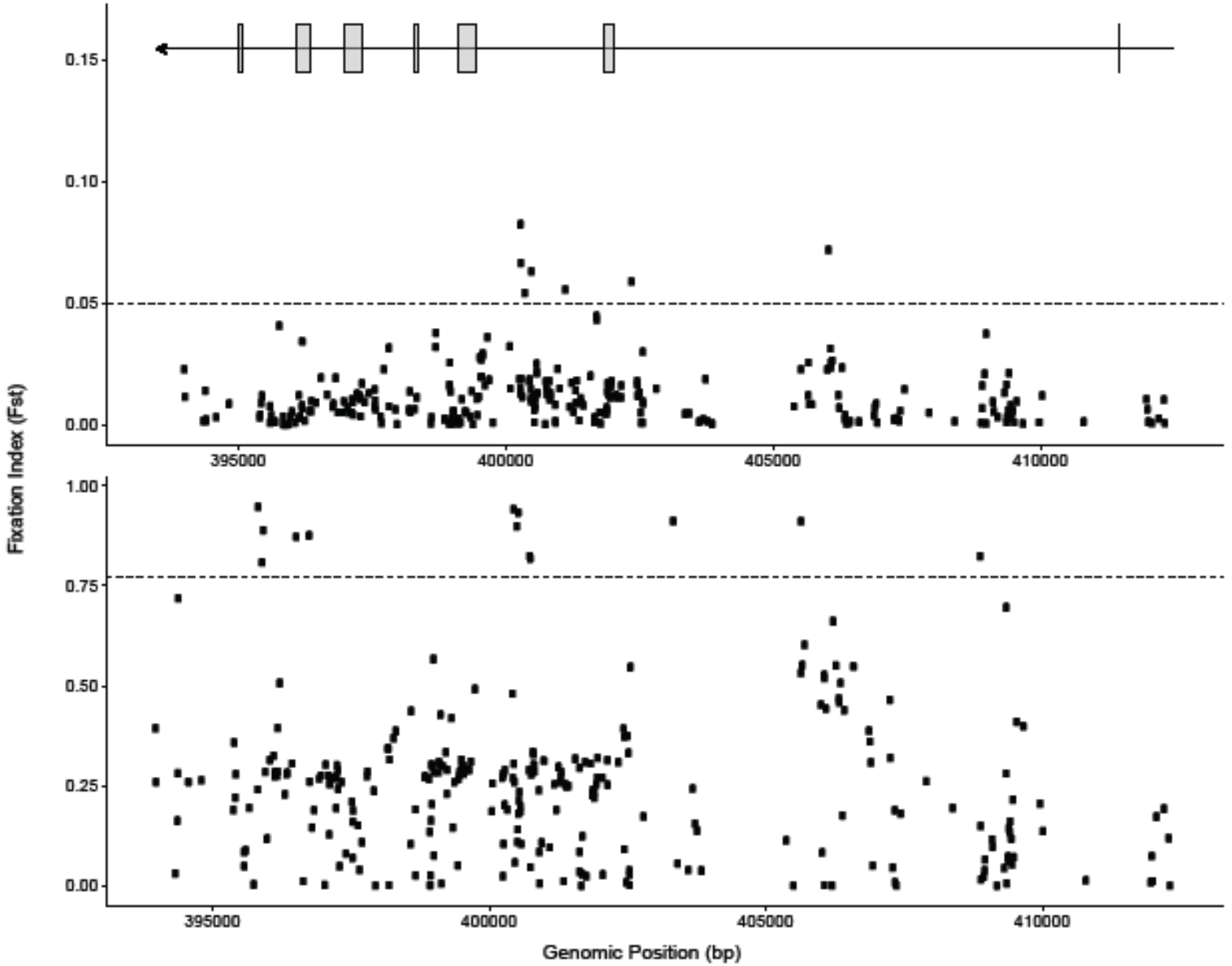
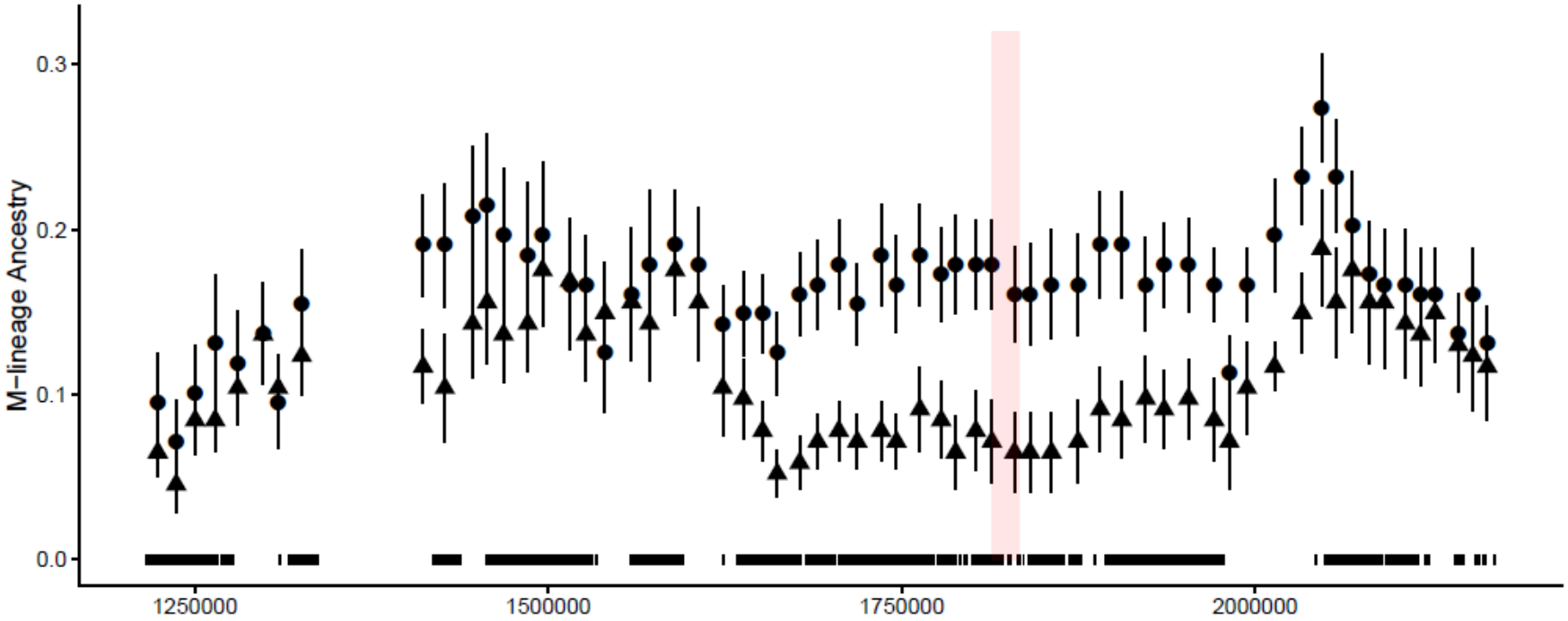

0.15
0.10
0.05
0.00
395000
400000
405000
410000
1.00
0.75
0.50
0.25
0.00
395000
400000
405000
410000
Fixation Index (Fst)
Genomic Position (bp)

Figure 6.4.


M−lineage Ancestry
0.3
0.2
0.1
0.0
1250000
1500000
1750000
2000000

Supplemental Tables

Table 6.S1: Genes found within highly differentiated windows between highly- and less-defensive Africanized Honey bees

Chromosome	Gene ID
1	GB55311
1	GB55338
1	GB47425
1	GB47424
1	GB47466
1	GB47423
1	GB47422
1	GB47467
1	GB47468
1	GB47420
1	GB47469
1	GB46546
1	GB46547
1	GB46548
1	GB42188
1	GB42144
1	GB42189
1	GB42143
1	GB42190
1	GB42191
1	GB42142
2	GB46296
2	GB46315
3	GB49081
3	GB49080
3	GB49118
3	GB49119
3	GB49079
3	GB49120
3	GB49121
3	GB49122
3	GB49123
4	GB43799
4	GB43805
4	GB43798

5	GB44661
5	GB44453
5	GB44452
5	GB44451
5	GB44662
5	GB44450
5	GB44449
5	GB44663
5	GB44448
5	GB44447
5	GB44479
5	GB44478
5	GB44638
5	GB44639
5	GB44477
5	GB46771
5	GB40915
5	GB40914
5	GB40913
5	GB40912
5	GB51391
5	GB51390
5	GB51389
5	GB47173
5	GB47170
5	GB47169
5	GB47174
5	GB55036
5	GB55037
5	GB55038
5	GB55039
5	GB55040
5	GB55041
5	GB55042
5	GB55027
5	GB44663
5	GB44447
5	GB44446
5	GB44664
5	GB44665

5	GB44445
5	GB44666
5	GB44667
5	GB44668
5	GB52279
5	GB52320
6	GB41506
6	GB41507
6	GB52156
6	GB52155
6	GB52154
6	GB52208
6	GB52209
6	GB52153
6	GB52210
6	GB52152
6	GB52211
6	GB52151
6	GB52212
6	GB52150
6	GB52149
6	GB52148
6	GB48496
6	GB48497
6	GB54460
6	GB54459
6	GB54487
7	GB46160
7	GB49248
7	GB49225
7	GB49249
7	GB49224
7	GB49223
7	GB43263
8	GB54493
8	GB54496
8	GB54495
8	GB54557
8	GB54493
8	GB54554

8	GB54555
8	GB54498
8	GB54556
8	GB54497
8	GB54496
8	GB40565
8	GB40338
8	GB40337
9	GB53316
9	GB53315
9	GB53426
9	GB53314
9	GB53313
9	GB53312
9	GB53427
9	GB53428
9	GB53311
9	GB53310
9	GB53309
9	GB53308
9	GB53307
9	GB53306
9	GB53429
9	GB53380
9	GB53381
9	GB53379
9	GB53382
9	GB53384
9	GB53378
9	GB53377
9	GB42581
9	GB42580
9	GB42899
9	GB42900
9	GB42901
9	GB42902
9	GB42903
9	GB42867
9	GB42868
9	GB42869

9	GB42870
9	GB42615
9	GB42871
9	GB42614
9	GB42872
9	GB42873
9	GB42874
9	GB42613
9	GB42875
9	GB42612
9	GB41403
9	GB41405
9	GB41402
10	GB49511
11	GB44981
11	GB44980
11	GB45188
11	GB44978
11	GB45189
11	GB44977
11	GB44976
11	GB45036
11	GB45035
11	GB55241
11	GB55242
11	GB55139
11	GB55243
11	GB55244
11	GB55245
11	GB55246
11	GB55247
11	GB55138
11	GB46621
11	GB46639
11	GB46620
11	GB46640
11	GB46641
11	GB46619
11	GB46642
11	GB46643

11	GB46644
11	GB46618
12	GB51967
12	GB52113
12	GB52114
12	GB52115
12	GB51968
12	GB52116
12	GB52117
12	GB52118
12	GB51967
12	GB52551
12	GB52552
12	GB55087
12	GB55084
12	GB55086
12	GB55085
12	GB51778
13	GB47776
13	GB47780
13	GB47779
13	GB47778
14	GB43711
14	GB43710
14	GB43709
14	GB43729
14	GB43574
14	GB43573
14	GB43572
14	GB43571
14	GB43570
14	GB43621
14	GB43622
14	GB43569
14	GB43568
14	GB43623
15	GB47935
15	GB47933
15	GB49440
15	GB49439

15	GB49438
15	GB49499
15	GB49500
15	GB49501
15	GB49437
15	GB49502
15	GB49503
15	GB49436
15	GB49504
15	GB49435
15	GB49434
15	GB49505
15	GB54668
15	GB54667
15	GB54671
15	GB54666
15	GB54672
15	GB54673
15	GB54674
15	GB54665
15	GB54675
15	GB54664
15	GB54663
15	GB54676
15	GB54662
15	GB54677
15	GB54661
15	GB54678
15	GB54660
15	GB54679
15	GB54680
15	GB46188
15	GB46187
15	GB46211
15	GB46198
15	GB46201
15	GB46202
15	GB46197
15	GB46203
15	GB46204

15	GB46196
15	GB46205
15	GB46194
15	GB46206
15	GB47083
15	GB47082
15	GB47081
15	GB47079
15	GB47078
15	GB47077

Table 6.S2: Characterization of most highly differentiated sites windows between highly- and less-defensive Africanized Honey bees

SNPID	Fst (High vs Low)	GeneID	Predicted Effect
10.16_38407	0.16077394	GB49511	INTRON
2.15_396583	0.12843079		INTERGENIC
12.8_78241	0.12334849		INTERGENIC
9.12_1301622	0.10884138	GB53427	SYNONYMOUS_CODING
5.8_404237	0.10360521		INTERGENIC
11.18_1909997	0.10328877		INTERGENIC
2.5_242395	0.10045361	GB46296	INTRON
13.10_635941	0.09804213		INTERGENIC
5.2_1391417	0.09367025	GB52279	INTRON
8.7_1816314	0.09299351	GB54493	INTRON
5.9_1203759	0.09210684	GB46771	INTRON
5.5_440198	0.09203307		INTERGENIC
4.8_437137	0.0873302	GB43799	INTRON
12.8_147244	0.08545198		INTERGENIC
6.14_692402	0.08499561	GB52150	INTRAGENIC
15.2_400292	0.08232654	GB47079	INTRON
15.13_395783	0.07984855		INTERGENIC
12.17_2687525	0.07917939		INTERGENIC
5.7_21158	0.07905504	GB51391	INTRON
15.16_62877	0.07821379		INTERGENIC
2.15_148621	0.07578157		INTERGENIC
15.10_53736	0.07535187	GB54673	INTRON
1.29_373738	0.07481621	GB47424	INTRON
7.12_395617	0.0746672	GB49225	SYNONYMOUS_CODING
6.12_288878	0.0743986	GB43054	INTRON
12.17_2739411	0.07382251	GB51967	INTRON
8.6_1805754	0.07101142	GB40565	INTRON
5.14_2186361	0.06938425	GB44663	INTRON
13.10_722146	0.06889874		INTERGENIC
9.10_2884693	0.06877883	GB42874	INTRON
9.10_4225587	0.0679407		INTERGENIC
1.32_623904	0.06758861		INTERGENIC
14.9_1214730	0.06722159	GB43621	INTRON
8.7_535107	0.0671162		INTERGENIC
15.5_318288	0.0670483	GB46187	INTRON
9.4_122086	0.0669054	GB41405	INTRON
1.18_133156	0.06623341		INTERGENIC

6.11_90997	0.06614486	GB48497	INTRON
5.6_82087	0.06455548		INTERGENIC
2.15_1026602	0.06383476	GB46593	INTRON
6.10_407683	0.06376112		INTERGENIC
6.20_1149077	0.06375698	GB41507	INTRON
12.6_45714	0.06053765		INTERGENIC
8.7_1878406	0.06032831	GB54493	INTRON
11.6_1381974	0.05913715	GB55241	INTRON
12.11_362681	0.05879709		INTERGENIC
15.5_54916	0.05867725		INTERGENIC
3.4_424438	0.05722289	GB49119	INTRON
1.1_1282030	0.05712338	GB42142	INTRON
8.7_1306552	0.05675113		INTERGENIC
5.14_1922702	0.05665595		INTERGENIC
5.14_2214156	0.05545655	GB44663	INTRON
16.2_1090434	0.05452807		INTERGENIC
1.3_570639	0.05375935	GB50376	INTRON
5.12_350443	0.05327347	GB48655	INTRON
9.12_5570	0.04923012	GB53380	INTRON
3.3_444030	0.0491785		INTERGENIC
7.2_28470	0.04841591		INTERGENIC
11.18_3019741	0.04690689	GB44980	INTRON
11.1_394928	0.04258921	GB46640	INTRON
7.14_34894	0.04110493		INTERGENIC
8.7_1725008	0.03518126	GB54496	INTRON
14.10_400249	0.03277042		INTERGENIC

Chapter 7:

Assessing Patterns of Admixture and Ancestry in Canadian Honey Bees

Brock A. Harpur, Nadine C. Chapman, Lior Krimus, Philip Maciukiewicz, Vijay Sandhu, Keshna Sood, Julianne Lim, Thomas E. Rinderer, Michael H. Allsopp, Benjamin P. Oldroyd, and Amro Zayed[5]

Introduction

The Western honey bee, *Apis mellifera* L., is native to the Old World where it has five major evolutionary lineages: the A lineage of Africa, the C and M lineages of Europe, the O lineage of Asia, and the Y lineage of North Eastern Africa and parts of the Middle East (Ruttner 1988, Franck et al. 2001, Whitfield et al. 2006a, Alqarni et al. 2011). These lineages are delineated geographically, morphologically, and genetically and they include approximately 24 subspecies (Ruttner 1988, Garnery et al. 1992, Garnery et al. 1993, Arias and Sheppard 1996, Franck et al. 2000, Palmer et al. 2000, Whitfield et al. 2006a, Wallberg et al. 2014). The current honey bee populations of North America are the result of centuries of importation, chiefly from the two European lineages (C and M). Canada's honey bee population originated from European settlers who introduced colonies from the M lineage (e.g. *A. m. mellifera*) (Root 1985, Seeley 1985, Cornuet 1986), followed by the C lineage (e.g. *A. m. ligustica* and *A. m. carnica*), and with minor introductions from the O lineage (e.g. *A. m. caucasica*) (Sheppard 1989a, b). Each new introduction of a lineage or subspecies into Canada was usually an effort by beekeepers to introduce "favourable" traits. Historically, C lineage bees have been favored for their high honey production and docility (Langstroth and Dadant 1889), but beekeepers often introduced variation from other regions of the world. For example, in Canada Beekeepers experimented with

[5] This published manuscript has been reprinted by permissions from its co-authors and publisher from the original manuscript: Harpur BA*, et al.* (2015) Assessing patterns of admixture and ancestry in Canadian honey bees. *Insect Soc* 62(4):479-489.

introductions from the A lineage (e.g. *A. m. intermissa* and *A. m. lamarckii*) (Root 1985, Sheppard 1989a, b, Pinto et al. 2007). Intentional admixture such as this dates back to at least Brother Adam's work in the United Kingdom to breed the 'ideal honey bee for beekeeping' (Root 1985). Brother Adam's own "Buckfast Bee" is a mix of several subspecies from each lineage and is still bred by a small number of Canadian beekeepers today.

Remarkably, there has been no large-scale investigation of the genetic ancestry of Canadian honey bees, despite a history older than the country itself (Crane 1999). We undertook a study on the genetics of Canadian honey bees using a citizen science approach to characterize their genetic ancestry and to study how geography and management practices influences their genetics. We also used the population genetics dataset to test the hypothesis that beekeepers in Northern Canada maintain honey bees more related to the Northern European subspecies *A. m. mellifera* (M lineage). A similar pattern has been noted in Australia, where colonies in colder regions of Tasmania maintain higher proportions of *A. m. mellifera* ancestry relative to colonies in warmer regions (Oldroyd et al. 1995), suggesting that *A. m. mellifera* is locally adapted for colder climes than the more Mediterranean C lineage (Ruttner 1988, Le Conte and Navajas 2008), and we would therefore expect that M lineage bees would perform better in Northern Canada.

Finally, we tested the utility of a SNP assay for discriminating between Canadian honey bees and Africanized honey bees from the United States and Brazil. Africanized honey bees can be highly aggressive and are continuously distributed from South America to the Southern United States (Rinderer et al. 1991, Sheppard et al. 1991, Collet et al. 2006, Szalanski and Magnus 2010). Africanized honey bees are the result of an introduction of the African lineage subspecies *A. m. scutellata* into Brazil in 1956 (Kerr 1967). Controlled crosses of Brazilian commercial honey bees with imported *A. m. scutellata* were performed with the hope that the resulting hybrid colonies would be better suited for Brazilian beekeepers. Unfortunately, the resulting "Africanized" colonies are often highly defensive (Collins et al. 1982, Breed et al. 2004, but see: Galindo-Cardona et al. 2013), swarm frequently, and typically abscond in response to adverse conditions (Winston 1992). Current tests for detecting Africanized honey bees with mtDNA and wing morphometrics are not reliable: they may miss cases of paternal

Africanization (Sheppard and Smith 2000) and are unable to detect low to medium levels of Africanization (Guzman-Novoa et al. 1994). Canadian beekeepers import hundreds of thousands of queens from the USA annually and the chance of accidental importation of Africanized honey bees is rated as moderate to high by the Canadian Food Inspection Agency. We had previously shown that an ancestry informative SNP panel was very successful at identifying Africanized bees in commercial honey bee populations from the United States and Australia (Chapman et al. 2015b), and we wanted to examine if the same panel will be suitable for use in Canada.

Methods

Citizen Science Project and Population sampling

From July 2013 to June 2014, we asked beekeepers across Canada to voluntarily take part in a genotyping study. Solicitations were made through social media, our personal websites, telephone, and announcements at beekeeping meetings. Beekeepers indicated their willingness to join our study by filling in an online form. This information was used to send each beekeeper a pamphlet containing sampling instructions, a small survey, sampling tubes, and a return envelope (Online Supplemental Files 1 and 2). Beekeepers were instructed to sample two workers (diploid females) per colony, from up to six colonies in their operation. We asked beekeepers to identify the location of their colonies, the number of colonies they manage, and the location of their queen breeder. In total, 145 beekeepers across 9 provinces and 1 territory submitted a total of 857 sampling tubes (Figure 7.1A; Online Supplemental Datasets 1, 2)

DNA Extraction and SNP genotyping

We extracted DNA from a single diploid worker from each sampling tube returned to us (N = 857 individual bees). High molecular weight DNA was extracted with phenol-chloroform isoamyl alcohol (25:24:1) from half of a bee's thorax. Each sample was then purified using EMD Multiscreen Millipore purification (Merck) and genotyped using the Sequenom MassARRAY MALDI-TOF (Agena) system in four multiplexes at Génome Québec Innovation Centre.

The SNP panel was created to differentiate between each of the three major lineages thought to be most abundant in North American honey bee populations: C, M, and A (Chapman et al. 2015b, Chapman et al. 2015a). Briefly, SNPs were randomly chosen from a set of more than 20 000 with high genetic differentiation as measured by pairwise F_{st} (Weir and Cockerham 1984) between each population group from a previous full-genome re-sequencing study (Harpur et al. 2014b) and conditioned on being at least 5,000 bp apart. We included an additional 19 SNPs from a previous study that also showed high genetic differentiation between European and African honey bees (Table S3 in Whitfield et al. 2006a). The final panel of 144 SNPs was chosen based on its ability to be multiplexed in an inexpensive SNP genotyping platform using the Sequenom Assay Design Suite (v1.0 Sequenom, CA, USA). In the final panel, all SNPs were separated by at least 45,945 bp (average 1,734,863 bp) and were effectively unlinked because of the honey bee's very high recombination rate (19 cM/Mb; Beye et al. 2006).

Population Admixture Analyses

To estimate each sample's ancestry, we used STRUCTURE v2.3.4 (Pritchard et al. 2000) using all polymorphisms with minimum allele frequency >0.05 and only for markers with a call rate >0.66 (N = 91 markers of 144 on the panel met these criteria). Similarly, we only included samples that could be successfully genotyped at 66% of all markers (N = 855 samples). We evaluated population structure using a burn-in phase of 50 000 iterations followed by 100 000 Markov-Chain Monte Carlo iterations with admixture assumed and uncorrelated allele frequencies. We included in each STRUCTURE run a set of 29 reference bees, known to be of pure descent from each of the three major lineages (African: A, Western and Northern Europe: M, and Eastern and Southern Europe: C). The reference bees were used in three previous population genetic analyses performed by our group (Harpur et al. 2012, Harpur and Zayed 2013, Harpur et al. 2014b); their genotype at each of the 91 SNPs was extracted from their full genome sequences. To reduce the influence of the large query population compared against a smaller reference population and to increase processing speed by parallelizing runs, we divided the dataset into 10 smaller datasets consisting of the reference population and 85 randomly selected samples. No *a priori* information was provided regarding population identity or

location. We performed 10 replicates for each of K = 1 to 4 populations. We used Structure Harvester v 0.6.94 (Earl and Vonholdt 2012) to estimate the most appropriate fit of K and to implement Evanno's method for estimating ΔK (Evanno et al. 2005). For each sample, we then identified the genomic contribution of each ancestral lineage (e.g. 70% C, 20% M, and 10% A) and the level of admixture (1 - maximum ancestry; e.g. if a bee is 70% C, 20% M, and 10% A, then admixture = 1-0.7). Finally, we used GENEPOP v4.0.11(Raymond and Rousset 1995) to report F_{st} statistics among provinces and countries and tested if pairwise F_{st} was significant using Arlequin 3.5.12 with a False Discovery Rate <0.05 (Excoffier and Lischer 2010) .

Accuracy of a SNP panel as a diagnostic test for Canada

To investigate the utility of our a SNP panel to detect Africanized honey bees among Canadian imports, we studied how well a selection of SNPs on our panel is able to accurately classify Africanized honey bees as African, and Canadian honey bees as European. We replicated a two stage procedure used previously to identify a cut-off at which the proportion of African ancestry is indicative of an Africanized honey bee (Chapman et al. 2015b, Chapman et al. 2015a).

We first estimated the True Positive rate of the genotyped SNPs by identifying at what minimum proportion of African ancestry (5%-60% in 5% increments) would bees known to be Africanized (true Africanized bees) correctly be identified as such. True Africanized samples were obtained from populations in Brazil (N = 55) and the United States (N = 86) (Chapman et al. 2015b, Chapman et al. 2015a). We also included *A. m. capensis* clonal lineage (N = 3), *A. m. capensis* (N = 104) and *Scutellata*-*Capensis* hybrids (N = 17), and 128 *A. m. scutellata* as samples that should be correctly identified as African. At each cut-off we determined the proportion of African/Africanized samples correctly identified as African/Africanized.

We then estimated the False Positive rate by repeating the above analysis with commercial true non-Africanized honey bees to identify at what maximum proportion of African ancestry true non-Africanized bees would be incorrectly classified as African. Our true non-Africanized samples were represented by the reference C and M populations (see above) as well as commercial and feral Australian (N = 93) and commercial populations from Canada (N = 10; imported into Australia) and the United States (N = 55). All reference samples were previously

genotyped in an Australian study at 95/144 of the SNPs available on the panel (those having minimum allele frequency > 0.05 and call rate > 0.66) (Chapman et al. 2015b, Chapman et al. 2015a). Of the 95 SNPs in this previous study, 81 were also used within the Canadian samples obtained from beekeepers (81/91 markers from N = 855 bees). Therefore, for all between-country analyses, including identifying cut-offs (above), we took only the genotypes of our Canadian samples at these 81 sites common between the two studies.

Statistical Analyses

All statistical analyses were performed in R v3.2.0 (R Core Team 2010). For geographic analyses, we binned statistics into 0.5˚ latitudinal and longitudinal bins. We identified trends across provinces both individually and as groups. We grouped Prairie Provinces (Alberta, Saskatchewan and Manitoba) and compared admixture levels among bees from Western Provinces and Territories (Yukon and British Columbia), Ontario and Quebec, and the Maritimes (Newfoundland, New Brunswick, and Nova Scotia). When performing multiple family-wise statistical tests, we corrected for False-Discovery Rate using the Benjamini-Hochberg method (Benjamini and Hochberg 1995) at $\alpha = 0.05$. Our datasets are available as supplemental tables on GitHub (https://github.com/harpur/CanadAdmix)

Results

Sampling Overview

We sent a total of 1633 individual sampling tubes across Canada and received back 857, a 52.4% return rate. Most samples came from British Columbia (N = 243) and Ontario (N = 199; Figure 7.1A; Online Supplemental Dataset 2). From each returned sampling tube, we genotyped a single (diploid) worker honey bee. Only two workers could not be successfully genotyped at 66% of all markers, so all population genetic and ancestry analyses were performed on 855 Canadian samples. Beekeepers could self-report the origins of their colonies. Of those who did self-report we found that most of the samples were bred in Canada (N = 665). Samples of workers from queens bred outside of Canada originated from the United States (N = 71), New Zealand (N = 27) or Denmark (N = 2). The beekeepers that responded to our study managed between 1 and 10500 colonies (mean = 400 ± 58.9 SE), indicating that we successfully solicited

interest from both hobbyists with a few colonies and commercial beekeepers with hundreds to thousands. We asked beekeepers to self-identify the subspecies or race of their bees. We received this information for 574 out of 855 colonies sampled and genotyped in this study. The largest proportions of beekeepers (30.2%) identified their bees as "Italian" or "Mixed" (13.7%) (Online Supplemental Dataset 1).

Admixture of Canadian honey bees

Analyses using STRUCTURE significantly supported models with K = 3 ancestral populations (A, M and C) both with the lowest average Ln[P(D)] = -1436.21 method, and by using Evanno's method to calculate ΔK (Figure 7.1B). Canadian bees were not classified as a distinct population, but instead a mix of the three ancestral lineages (Figure 7.1B). Canadian colonies had, on average, a large proportion of their ancestry originating from the C group (mean 74.2%), with the remainder consisting of M group (19.6%) and A group (6.2%; Figure 7.1B-C). As a result, differences in admixture between Canadian honey bee populations were driven by the level of M and A ancestry: where increasing M and/or A ancestry lead to increased admixture (Spearman Rank Correlation; Rho > 0.51; $P < 2.2x10^{-16}$). We found small but significant differences in the level of admixture between provinces (Figure 7.S2; ANOVA; $F_{9,845} = 6.167$; $P = 1.9x10^{-8}$). These differences tended to be between Prairie Provinces (Alberta, Saskatchewan and Manitoba) and others (Figure 7.S2). We confirmed this trend by pooling the bees from the Prairie Provinces and comparing their admixture to bees from Western Provinces and Territories (Yukon and British Columbia), Ontario and Quebec, and the Maritimes (Newfoundland, New Brunswick, and Nova Scotia). From this comparison, we found populations in the Prairie Provinces had lower levels of admixture when compared to populations in each of Canada's other major geographic regions (ANOVA; $F_{3,851} = 8.424$; $P = 1.6x10^{-5}$; Tukey's HSD $P < 0.0035$; Figure 7.S2).

Provinces also differed in their patterns of ancestry. We found small but significant differences in the average mean proportion (per sample) of C (Figure 7.1C; ANOVA; $F_{9,845} = 6.167$; $P = 1.9x10^{-8}$) and M ancestry (ANOVA; $F_{9,845} = 5.36$; $P = 3.7x10^{-7}$) among provinces, but did not detect differences in the level of A ancestry (P = 0.091). High C (low M) ancestry is more common in the Prairie Provinces than in the Western Provinces, Quebec and Ontario, and

Maritime Provinces, which had significantly lower C ancestry (Figure 7.S3; ANOVA; $F_{3,851}$ = 8.424; P = $1.6x10^{-5}$) and a trend towards higher levels of M ancestry (ANOVA; $F_{3,851}$ = 2.81; P = 0.0382; Tukey's HSD P > 0.052). Although we found these minor differences in the overall level of admixture and ancestry, Canadian provinces have very low levels of differentiation at the loci examined (Mean F_{st} = 0.0078; Table 7.1).

We found no significant evidence that samples from any self-identified subspecies or group have more A-lineage ancestry than any others; however, Buckfast bees (A = 8.1%; N = 33) tended to have higher levels of A ancestry than non-Buckfast bees (A = 6.0%; one-tailed t test; P = 0.06) and the sample with the highest proportion of A ancestry within Canada (A = 30.1%) is of Buckfast origin.

Distributions of honey bee lineages across Canada – Local Adaptation or Management Practices?

We predicted that Northern Canada may favour genotypes derived from honey bee subspecies accustomed to similar environments, such as the M group subspecies (Ruttner 1988, Le Conte and Navajas 2008). To test this hypothesis, we investigated associations between ancestry (C, M, or A) and geographic location (Figure 7.2). Following corrections for False Discovery Rate (Benjamini and Hochberg 1995), we found a significant negative correlation between M ancestry and latitude (P = 0.008; r = -0.48) and a positive relationship between C ancestry and latitude (P = 0.0046; r = -0.51) indicating that colonies in Northern Canada tended to have higher proportions of C lineage (Figure 7.2). There was no trend for A ancestry (P = 0.45). In addition, colonies from Northern Canada tended to be less admixed: there is a negative correlation between admixture and latitude (P = 0.0066; r = -0.14).We found that there is a significant positive correlation between M ancestry and longitude (P = 0.014; r = 0.29), a negative relation between C ancestry and longitude (P = 0.0006; r = -0.4), and a trend for more A ancestry in Eastern Canada (P = 0.052; r = 0.23; Figure 7.2).

It may be that the relationship between M lineage ancestry and geography are not a result of local adaptation but by differences in beekeeping practices. For example, small-scale beekeepers may prefer different subspecies of honey bee than commercial beekeepers. We found no relationship between the number of colonies managed by a beekeeper and the levels of C, M nor A ancestry of his/her samples (Spearman's Rank Correlation, $P > 0.38$ for all comparisons), nor level of admixture of his/her colonies and the number of colonies managed ($P = 0.46$).

Regional importation practices did seem to influence ancestry. We found significant regional differences in importation practices across Canada. Beekeepers at latitudes >50˚ reported purchasing more queens outside of Canada than beekeepers at lower latitudes (<50˚; 22% vs 14%; Fisher Exact test; $P = 0.039$). Western beekeepers (longitude < -100˚) also reported importing more queens than Eastern beekeepers (> -100˚; 17.1% vs 9.4%; $P = 0.018$). We found that imported colonies had significantly more C ancestry (ANOVA; $F_{1,763} = 18.21$; $P = 2.2x10^{-5}$) and significantly less M ($F_{1,763} = 5.096$; $P = 0.024$), and A ancestry ($F_{1,763} = 5.82$; $P = 0.0218$; Figure 7.3) relative to Canadian-bred and purchased bees.

Admixture on a Global Scale

We compared our dataset of Canadian honey bee ancestries those of commercial honey bee populations in Australia and non-Africanized populations in the United States that were genotyped using the same SNP panel (Chapman et al. 2015a). We found significant differences in the levels of admixture between countries (ANOVA; $F_{2,1000} = 33.1$; $P < 1.2x10^{-14}$), with Canadian samples (mean = 25%) having similar levels of admixture as Australian samples (Tukey's HSD; $P = 0.054$; mean=31%) and both having higher admixture than United States commercial samples (Tukey's HSD; $P < 0.00001$; mean = 23%). We found no differences in the level of African ancestry of commercial colonies between these countries ($P = 0.297$), but we did find significant differences in the degree of M ancestry ($F_{2,1002} = 96.95$; $P < 2.2x10^{-16}$) with significantly higher levels in commercial Australia (mean=30.5%) relative to both Canada (19.3%) and the United States (Tukey's HSD; $P < 0.0001$; 18.0%;). An inverse trend was found for C ancestry: Canada (74.2%) and the United States (76.6%) had more C ancestry than Australia (Tukey's HSD; P<0.0001; 64.1%). Even with these differences, we found no

significant evidence of differentiation between countries: average F_{st} between countries was 0.04 (Table 7.2).

Accuracy of SNP-based Africanized test in Canada

We previously characterized thresholds for identifying Africanized samples using this SNP panel for the use as a diagnostic assay in Australia (Chapman et al. 2015b, Chapman et al. 2015a). Africanized bees have higher levels of African ancestry (over 50%), compared to non-Africanized bees (less than 25%) allowing us to distinguish potentially Africanized samples using a predetermined threshold (Chapman et al. 2015b, Chapman et al. 2015a). We re-evaluated this cut-off in light of the ancestry of Canadian honey bee populations (Figure 7.4A). When we used a minimum cut-off threshold of 15% - 25% African ancestry, we obtained a True Positive rate of 1 and all true Africanized samples were correctly identified as such. When these same thresholds were applied to true non-Africanized commercial stocks we obtained a False Positive rate of 0.05 at a threshold of 15% (95% of true non-Africanized commercial stocks were classified as not African). At a more conservative threshold (25% A ancestry), we obtained a False Positive rate of 0 (100% of true non-Africanized commercial stocks were classified as not African; Figure 7.4B). Therefore, using the more conservative cut-off threshold of 25%, which has the maximum True Positive rate and minimized the False Positive rates, we found that 99.82% of the 855 Canadian honey bees genotyped herein could be classified as not African, as expected (Figure 7.1).

Discussion

Patterns of Admixture within Canada

Canada has no native populations of *A. mellifera*; resident populations are the result of centuries of importation from around the world, predominately from the C and M lineages of Europe (e.g. *A.m. ligustica* and *A. m. mellifera*) (Seeley 1985, Cornuet 1986, Sheppard 1989a, b, Pinto et al. 2007). We have demonstrated here that contemporary Canadian honey bees are largely derived of C lineage subspecies, very similar to populations in the United States (Seeley 1985, Sheppard 1988, 1989a, b, Pinto et al. 2007, Delaney et al. 2009) and Australia (Oxley and

Oldroyd 2009, Chapman et al. 2015b). This pattern is likely a result of both North American and Australian beekeepers favouring C lineage bees for their docility and honey production (Langstroth and Dadant 1889). Beekeepers have regularly imported and admixed local populations with *A. m. ligustica* (a practices once called "Italianizing") to introduce these favourable phenotypes (Jensen et al. 2005, Moritz et al. 2005). The large C lineage component of Canadian honey bees is likely a result of past importation preferences and the use of "Italianized" colonies that continues today.

Previous studies have discovered differences in ancestry between feral and commercial populations (Sheppard 1988, Schiff and Sheppard 1995, Chapman et al. 2008, Delaney et al. 2009, Chapman et al. 2015a), with feral bees having higher levels of M ancestry. This pattern is thought to be the result of beekeepers either favouring the use of C lineage bees or selection in feral populations favouring M ancestry (Pinto et al. 2005). We did not include feral populations in our Canada survey. However, we did find that a colony's location was correlated with its ancestry. North-western Canada had more C ancestry (less M) than South-eastern Canada. This is counter to expectation: northern colonies would be expected to be comprised of more northern-derived (i.e. M lineage) ancestry (Oldroyd et al. 1995). We attribute this pattern not to selective differences between parts of the country, but rather to beekeepers in North-western Canada self-reporting that they imported more colonies/queens from international sources that have higher C ancestry than colonies reported to be purchased within Canada.

Commercial populations of honey bee have been noted previously for their relatively low levels of differentiation within their introduced ranges (Delaney et al. 2009, Harpur et al. 2012, Chapman et al. 2015b, Chapman et al. 2015a). Three factors contribute to this pattern: high gene flow, similar importation histories, and the relatively young age of commercial populations. The Canadian samples included in this study were separated by as much as 4772 km and our international samples much further. Nonetheless, inter-population comparisons confirm that commercial colonies have very low levels of differentiation. Our data suggest that gene flow within Canada is very high; most beekeepers (86.9%) reported queens from breeders within Canada rather than rearing their own queens locally. Similarly, we found the lowest levels of differentiation between commercial US and Canadian populations (Table 7.2), two populations

that exchange honey bees frequently. Collectively, Canadian beekeepers import 150 000 to 200 000 queen bees from the United States each year (Tavares 2014). Commercial populations in Canada, the United States and Australia are also relatively young and originate from similar source populations. North America has only had resident populations of honey bees since the 17th century (Sheppard 1989a, b) and much like Australia and the United States (Hopkins 1886, Ruttner 1976, Sheppard 1989a, b, Oldroyd et al. 1992, Koulianos and Crozier 1996, 1997, Jolly 2004, Chapman et al. 2008, Oxley and Oldroyd 2009), the Canadian populations examined herein were likely first derived from the M lineage and later shifted to C lineage. Because North American and Australian populations are relatively young, drift has less time to alter allele frequencies, and potential differences are flooded by gene flow. Taken together, the young age of these populations, their similar importation histories and high gene flow have likely contributed to the current low levels of genetic differentiation.

Admixture in Global Commercial Populations

While introgression can be detrimental to the conservation of honey bees within their native ranges (De la Rua et al. 2009, Meixner et al. 2010, De la Rua et al. 2013, Pinto et al. 2014), it is actively sought after in regions without native *A. mellifera* populations, such as North America (Cobey et al. 2012, Sheppard 2012). It has been well documented that genetic diversity is important to the health of colonies (Tarpy 2003, Jones et al. 2004, Mattila and Seeley 2007), and beekeepers seek novel genotypes resistant to pests (Rinderer et al. 2010, Cobey et al. 2012, Sheppard 2012). Admixture has been shown to increase levels of genetic diversity in honey bees (Harpur et al. 2012, 2013) and beekeepers have been intentionally interbreeding subspecies of honey bee for at least a century in North America (e.g. Root 1985), often not targeted in a systematic way. Using tools such as the SNP panel herein (Chapman et al. 2015b), or similar approaches, it can be possible for regulators to target and manage the introduction of novel genetic stock to areas most in need or where it will be most beneficial. A corollary, is that these SNP panels can also be used to test for introgression of unwanted genetic stock such as C lineage in ancestral M lineage ranges (Pinto et al. 2014, Munoz et al. 2015) or Africanized bees in North America.

The utility for a SNP-based assay for monitoring Canadian imports

The current tests available to distinguish Africanized from non-Africanized colonies prior to importation can be unreliable. The SNP panel used in this study was designed to identify bees with African ancestry regardless of their maternal or paternal backgrounds, including Africanized honey bees. Using the frequency of SNPs indicative of African ancestry, we were able to demonstrate that Africanized honey bees can confidently be detected: We were able to detect true Africanized bees with 100% accuracy with a false-positive rate of less than 0.05. Additionally, the SNP panel (Chapman et al. 2015b, Chapman et al. 2015a) allows for the estimation of a bees's ancestry to each of the major honey bee lineages in Africa and Europe. Using this panel, we found low but pervasive levels of African ancestry in Canadian honey bees. Levels of African ancestry in Canadian bees ranged from 0.1 to33%, very similar to levels of African ancestry found in Australia (0.3-32.8%; Chapman et al. 2015b). To our knowledge, there have been no deliberate introductions of Africanized bees into Canada. We suggest that the low level of African ancestry in Canadian bees most likely resulted from early importations of A lineage subspecies other than *A. m. scutellata*. Most likely, Canadian beekeepers imported *A. m. intermissa* (Seeley 1985), *A. m. lamarkii* (Nielsen et al. 2000) , or other North African subspecies. Canadian beekeepers have imported bees admixed with other African lineages. Beekeepers in Ontario have maintained Buckfast bees (those developed by Brother Adam) since the 1960's (Otis 2015; Pers. Comm.). The first Buckfast bees brought into Canada came from the daughters of breeders of *A. m. saharensis* and *A. m. monticola* (Otis 2015; Pers. Comm.). Even with the few Buckfast bees represented in our dataset (N=33), we found a trend for Buckfast bees having higher A-lineage ancestry relative to all other subspecies or groups identified by beekeepers. Although we are unable to differentiate African subspecies with the current version of the SNP panel, the addition of informative alleles for each A-lineage subspecies, particularly *A. m. scutellata*, will enable us to better determine the origins of this pattern in the future.

Conclusions

Our data are the first in-depth assessment of the genetic structure of honey bees in Canada. Honey bees in this country, like most in the world, live predominantly in managed populations and management practices have significantly impacted genetic structure and admixture, as we have demonstrated here and elsewhere (Harpur et al. 2012, 2013). How these

management practices influence wild populations or contribute to the long-term success of managed populations still remain largely unanswered questions.

Tables

Table 7.1: Pairwise Fixation index (F_{st}) between each Canadian Province in this study. No comparisons were significant (FDR<0.05)

	British Columbia	Alberta	Saskatchewan	Manitoba	Ontario	Quebec	New Brunswick	Nova Scotia	Newfoundland
Alberta	0.0013								
Saskatchewan	0.0021	0.0016							
Manitoba	0.0035	0.0017	0.0026						
Ontario	0.0019	0.0019	0	0.0027					
Quebec	0.0020	0.0017	0.0042	0.0043	0.0026				
New Brunswick	0.0030	0.0047	0.0019	0.0089	0.0047	0.0043			
Nova Scotia	0.0046	0.0062	0.0104	0.0133	0.0052	0.0028	0.0077		
Newfoundland	0.0015	0.0020	0	0.0022	0.0001	0.0099	0.0045	0.0131	
Yukon Territory	0.0264	0.0253	0.0242	0.0257	0.0259	0.0248	0.0246	0.0209	0.0059

Table 7.2: Pairwise Fixation index (F_{st}) between Canadian, Australian and United States honey bee Colonies. No comparisons were significant (FDR<0.05)

	Canada	Australia
Australia	0.050	
United States	0.038	0.040

Figure legends

Figure 7.1: A) Map of sampling locations (red dots) and average proportion ancestry in each province with province code (Yellow, C; Black, M; Red, A). **B)** Ancestry of Canadian honey bees to major honey bee lineages. The first 29 solid bars are known reference samples of C, M, and A lineage bees. All bars following the white gap represent 855 Canadian honey bee samples. **C)** Proportion of ancestry derived from each major lineage within each sampled Canadian province. We found small but significant differences in the proportions of C ($P=1.9x10^{-8}$) and M ancestry ($P=3.7x10^{-7}$) among provinces, but did not detect differences in the level of A ancestry (P=0.091). High C (low M) ancestry is more common in the Prairie Provinces (Alberta, Saskatchewan and Manitoba) than in the Western Provinces and Territories (Yukon and British Columbia). Quebec and Ontario, and Maritime Provinces (Newfoundland, New Brunswick, and Nova Scotia), which had significantly lower C ancestry.

Figure 7.2: Relationships between latitude, longitude and percent ancestry (percentage C, M and A). Latitude negatively correlated with M ancestry and negatively correlated with C ancestry, but was not significantly correlated to A ancestry. Longitude positively correlated with M ancestry, negatively with C ancestry but was not significantly correlated to A ancestry.

Figure 7.3: Ancestry and country-of-origin of Canadian honey bee stocks. Canadian-bred colonies had significantly less C ancestry relative to internationally purchased colonies.

Figure 7.4: A) To identify an optimal true-positive rate, we estimated the proportion of African ancestry at which true-Africanized (N=393) bees collected from source populations in Africa, Brazil and the United States would be correctly identified as Africanized. **B)** To identify an optimal false-positive rate, we estimated the proportion of African ancestry at which all true-non-Africanized bees (N=187) would be correctly identified as not Africanized (i.e. not incorrectly identified as African) using 5% increments of A ancestry.

Figure 7.S1: Evanno's Method for the identification of K, following STRUCTURE analyses, showing optimal K=3 populations.

Figure 7.S2: Average admixture (1 - maximum ancestry; e.g. if 70% C, 20% M, and 10% A, then admixture = 1-0.7) of each Canadian Province represented in our study.

Figure 7.S3: Proportion of ancestry derived from each major lineage within each pooled Canadian province: Prairie Provinces (Alberta, Saskatchewan and Manitoba), Western Provinces and Territories (Yukon and British Columbia), Ontario and Quebec, and the Maritimes (Newfoundland, New Brunswick, and Nova Scotia). High C (low M) ancestry is more common in the Prairie Provinces than in the Western Provinces Quebec and Ontario, and Maritime Provinces, which had significantly lower C ancestry

Figure 7.1

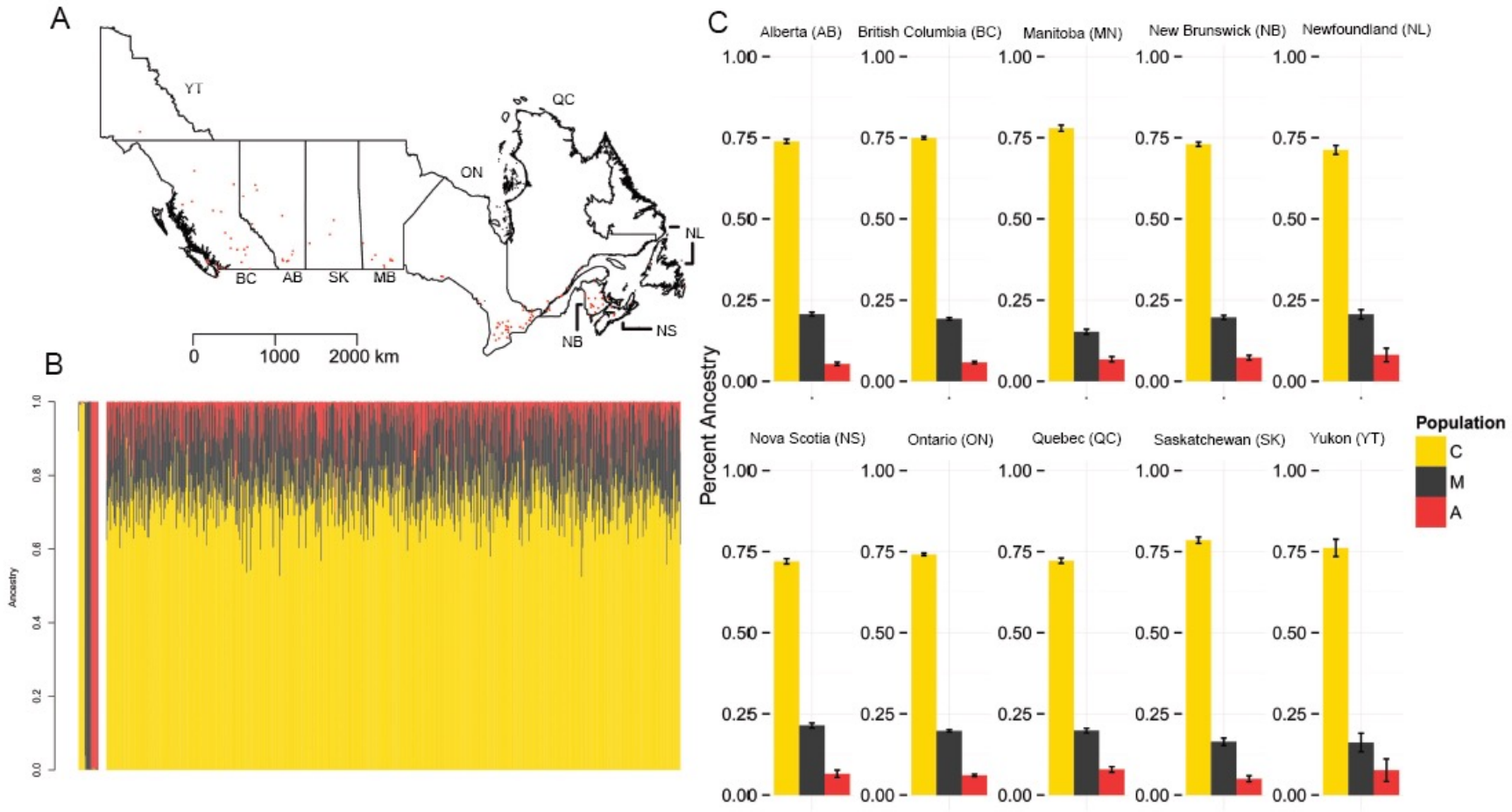

A
YT
QC
ON
NL
BC
AB
SK
MB
NB
NS
0
1000
2000 km
B
Ancestry
C
Alberta (AB)
British Columbia (BC)
Manitoba (MN)
New Brunswick (NB)
Newfoundland (NL)
Nova Scotia (NS)
Ontario (ON)
Quebec (QC)
Saskatchewan (SK)
Yukon (YT)
Percent Ancestry
Population
C
M
A

Figure 7.2

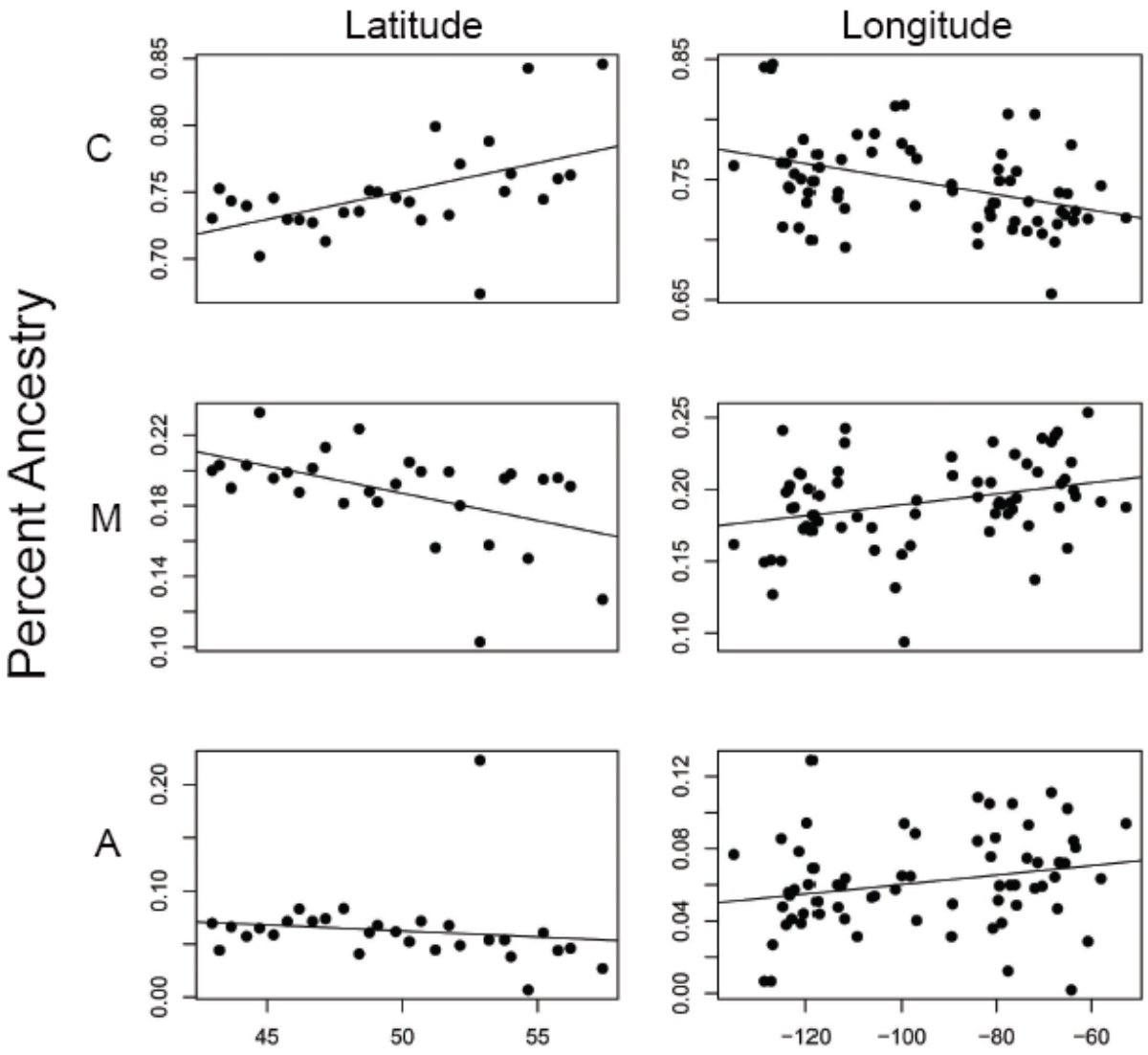

Latitude
Longitude
Percent Ancestry
C
M
A
0.70
0.75
0.80
0.85
0.65
0.10
0.14
0.18
0.22
0.15
0.20
0.25
0.00
0.04
0.08
0.12
45
50
55
−120
−100
−80
−60

Figure 7.3

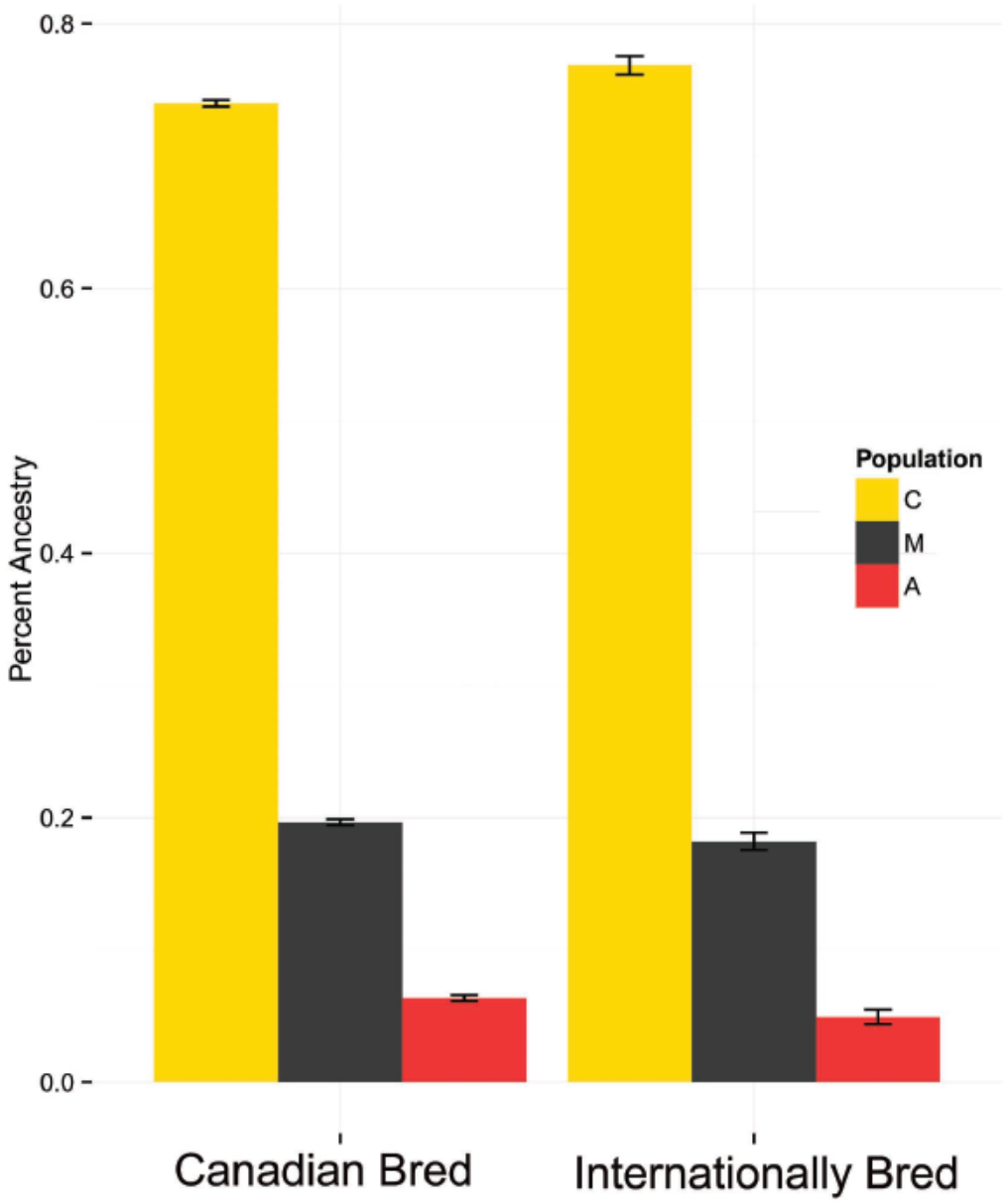

0.8
0.6
0.4
0.2
0.0
Percent Ancestry
Population
C
M
A
Canadian Bred
Internationally Bred

Figure 7.4

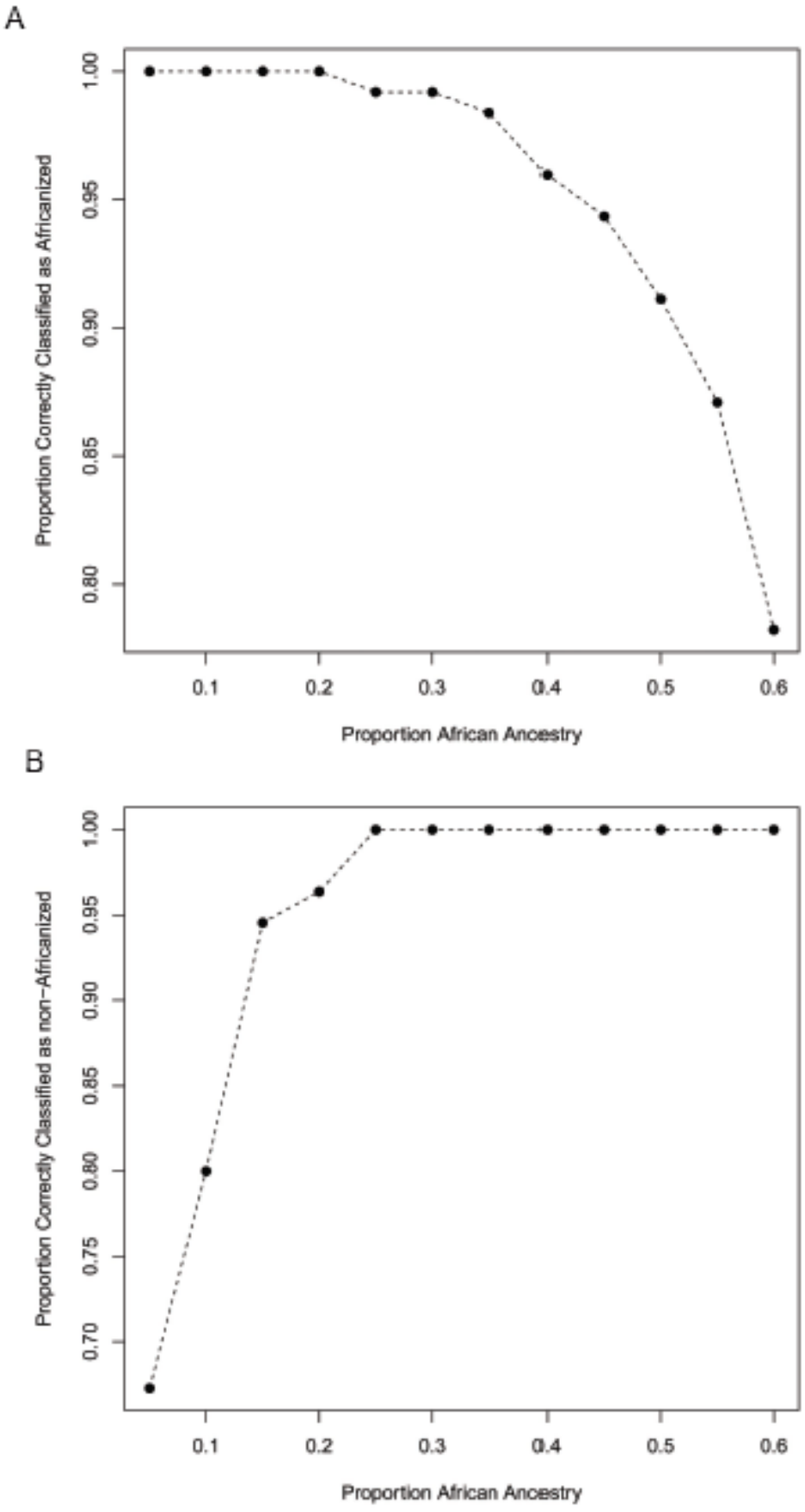

A
Proportion Correctly Classified as Africanized
1.00
0.95
0.90
0.85
0.80
0.1
0.2
0.3
0.4
0.5
0.6
Proportion African Ancestry
B
Proportion Correctly Classified as non−Africanized
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.1
0.2
0.3
0.4
0.5
0.6
Proportion African Ancestry

Figure 7.S1

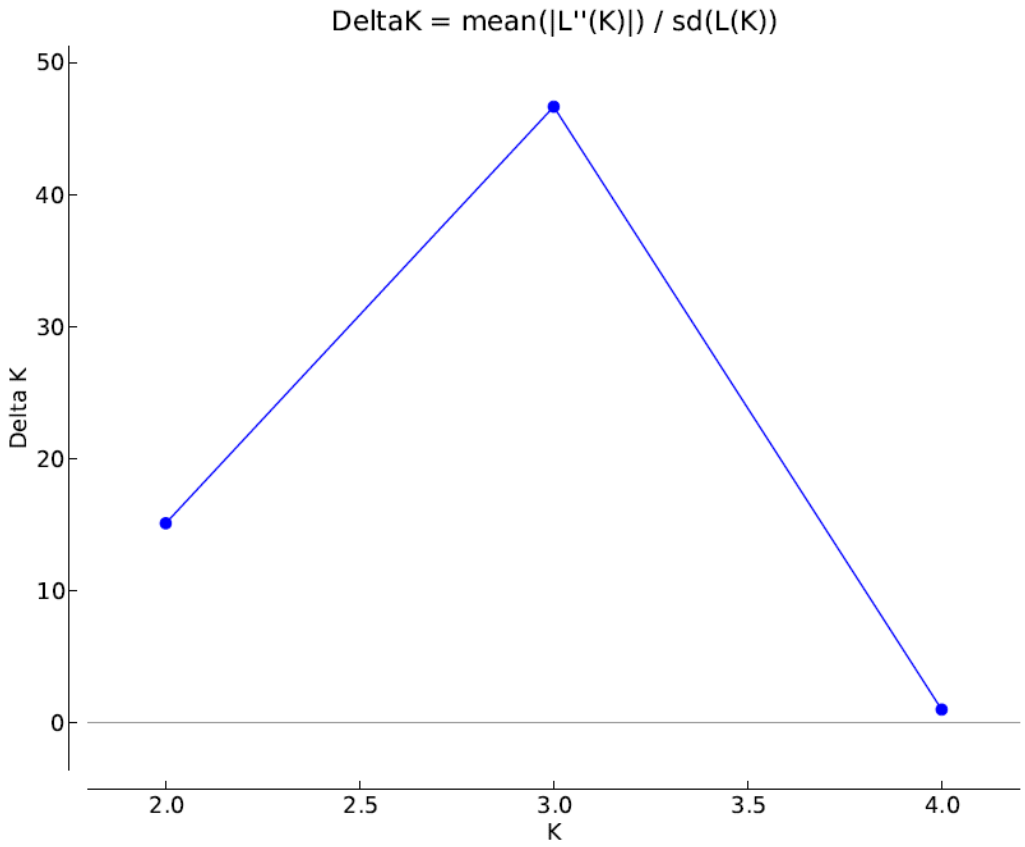


Figure 7.S2

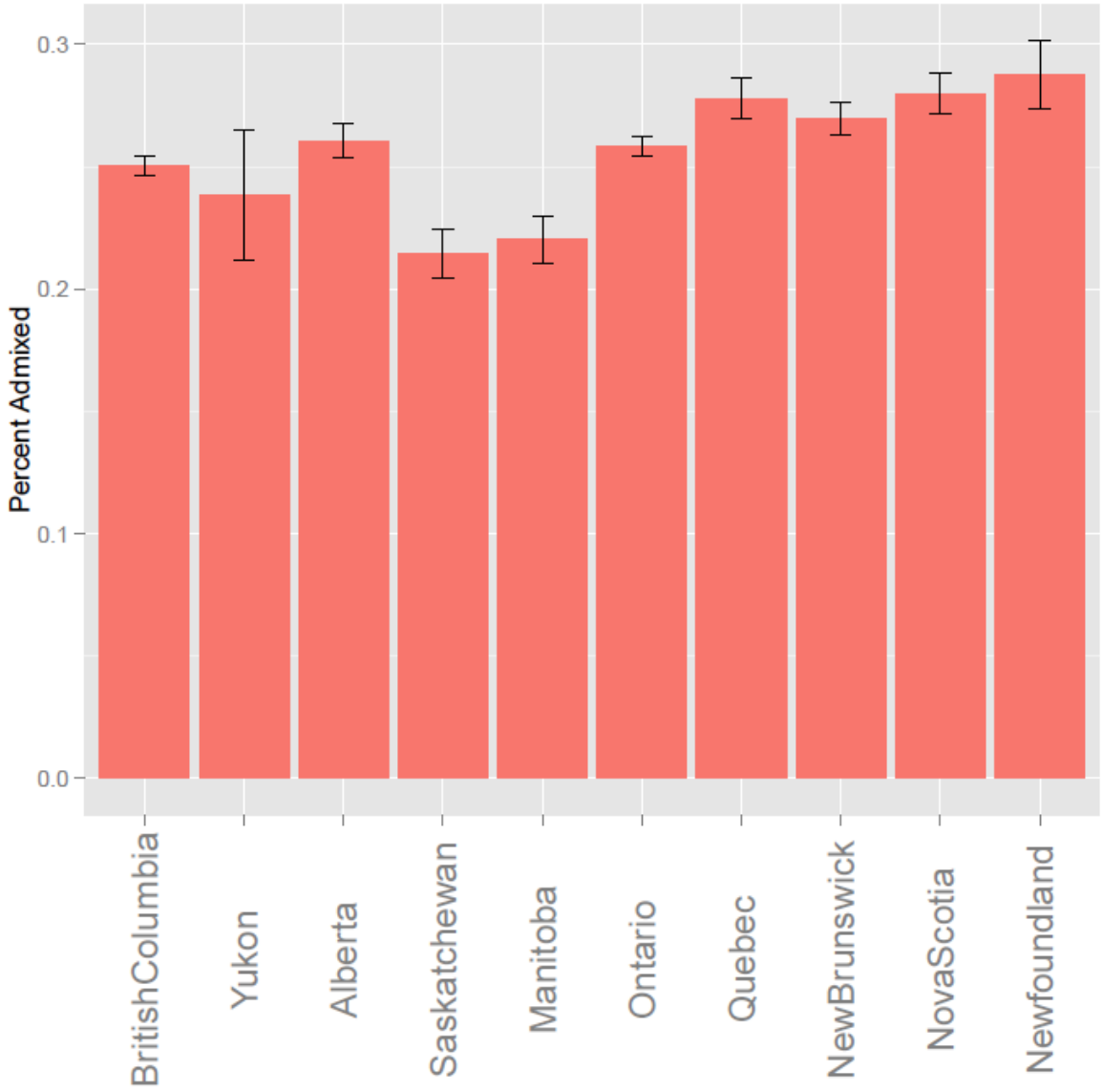
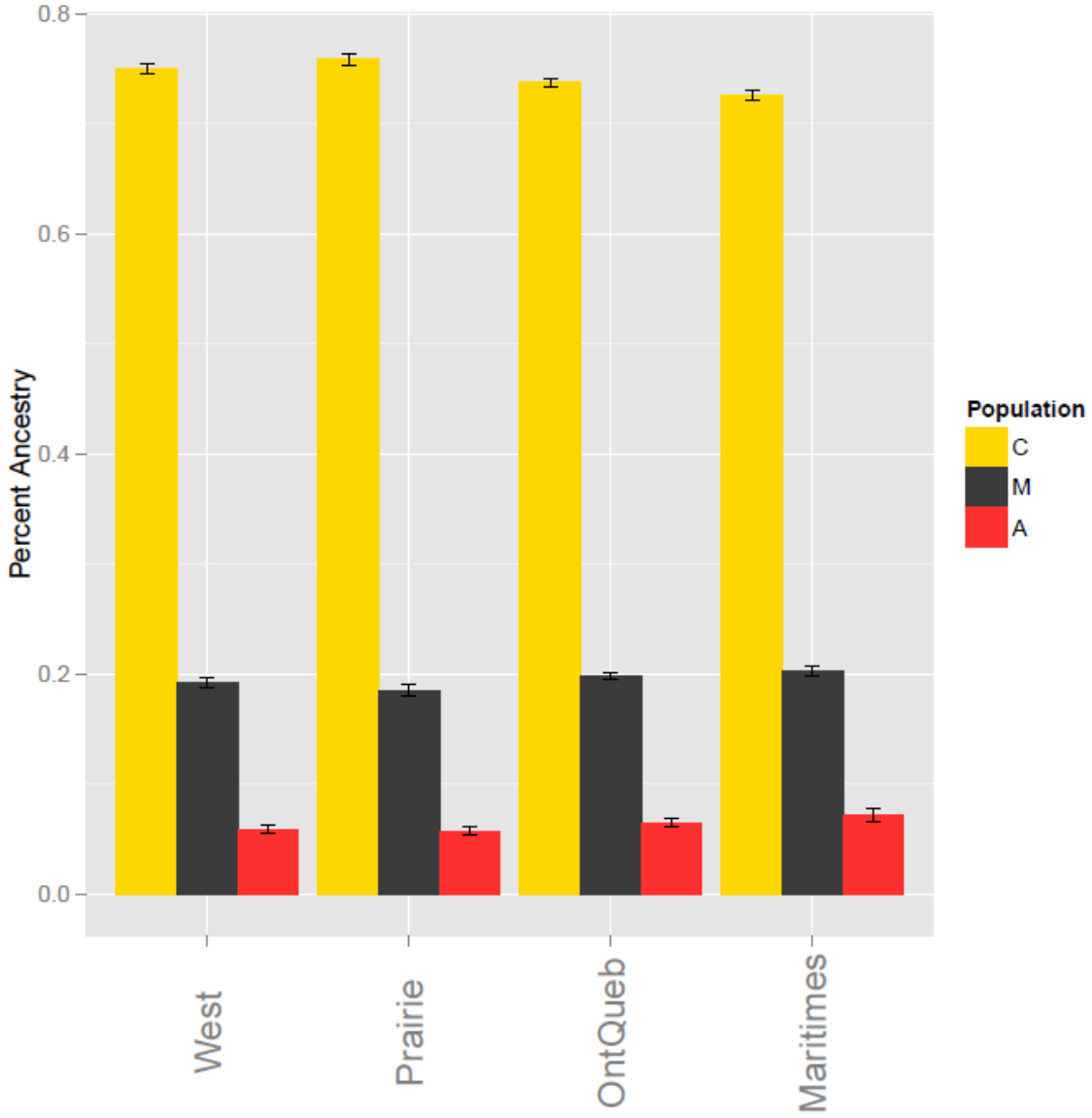

Percent Admixed
0.3
0.2
0.1
0.0
BritishColumbia
Yukon
Alberta
Saskatchewan
Manitoba
Ontario
Quebec
NewBrunswick
NovaScotia
Newfoundland

Figure 7.S3


0.8
0.6
0.4
0.2
0.0
Percent Ancestry
West
Prairie
OntQueb
Maritimes
Population
C
M
A

Bibliography

Abbot, P. and J. Abe and J. Alcock and S. Alizon and J. A. Alpedrinha, et al. 2011. Inclusive fitness theory and eusociality. Nature **471**:E1-4.

Akey, J. M., G. Zhang, K. Zhang, L. Jin, and M. D. Shriver. 2002. Interrogating a high-density SNP map for signatures of natural selection. Genome Research **12**:1805-1814.

Alaux, C., S. Sinha, L. Hasadsri, G. J. Hunt, E. Guzman-Novoa, et al. 2009. Honey bee aggression supports a link between gene regulation and behavioral evolution. Proceedings of the National Academy of Sciences of the United States of America **106**:15400-15405.

Alexander, D. H., J. Novembre, and K. Lange. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Research **19**:1655-1664.

Alford, D. 1969. A study of the hibernation of bumblebees (Hymenoptera: Bombidae) in southern England. The Journal of Animal Ecology:149-170.

Alqarni, A. S., M. A. Hannan, A. A. Owayss, and M. S. Engel. 2011. The indigenous honey bees of Saudi Arabia (Hymenoptera, Apidae, *Apis mellifera jemenitica* Ruttner): Their natural history and role in beekeeping. Zookeys:83-98.

Amdam, G. V., K. Norberg, M. K. Fondrk, and R. E. Page, Jr. 2004. Reproductive ground plan may mediate colony-level selection effects on individual foraging behavior in honey bees. Proceedings of the National Academy of Sciences of the United States of America **101**:11350-11355.

Amdam, G. V. and S. W. Omholt. 2003. The hive bee to forager transition in honeybee colonies: the double repressor hypothesis. Journal of Theoretical Biology **223**:451-464.

Ament, S. A., M. Corona, H. S. Pollock, and G. E. Robinson. 2008. Insulin signaling is involved in the regulation of worker division of labor in honey bee colonies. Proceedings of the National Academy of Sciences of the United States of America **105**:4226-4231.

Ament, S. A., Y. Wang, C. Chen, C. A. Blatti, F. Hong, et al. 2012. The transcription factor *ultraspiracle* influences honey bee social behavior and behavior-related gene expression. PLoS Genetics **8**:e1002596.

Ament, S. A., Y. Wang, and G. E. Robinson. 2010. Nutritional regulation of division of labor in honey bees: toward a systems biology perspective. Wiley Interdiscip. Rev. Syst. Biol. Med. **2**:566-576.

Amsalem, E., D. A. Galbraith, J. Cnaani, P. E. Teal, and C. M. Grozinger. 2015. Conservation and modification of genetic and physiological toolkits underpinning diapause in bumble bee queens. Molecular Ecology **24**:5596-5615.

Andolfatto, P. 2008. Controlling type-I error of the McDonald-Kreitman test in genomewide scans for selection on noncoding DNA. Genetics **180**:1767-1771.

Arias, M. C. and W. S. Sheppard. 1996. Molecular phylogenetics of honey bee subspecies (*Apis mellifera* L.) inferred from mitochondrial DNA sequence. Molecular Phylogenetics and Evolution **5**:557-566.

Arias, M. C. and W. S. Sheppard. 2005. Phylogenetic relationships of honey bees (Hymenoptera:Apinae:Apini) inferred from nuclear and mitochondrial DNA sequence data. Molecular Phylogenetics and Evolution **37**:25-35.

Bak, B., J. Wilde, and M. Siuda. 2010. Comparison of Hygienic Behaviour between Five Honey Bee Breeding Lines. Journal of Apicultural Science **54**:17-24.
Balhareth, H. M., A. S. Alqarni, and A. A. Owayss. 2012. Comparison of Hygienic and Grooming Behaviors of Indigenous and Exotic Honeybee (Apis mellifera) Races in Central Saudi Arabia. International Journal of Agriculture and Biology **14**:1005-1008.
Bar-Peled, L., L. Chantranupong, A. D. Cherniack, W. W. Chen, K. A. Ottina, et al. 2013. A tumor suppressor complex with GAP activity for the Rag GTPases that signal amino acid sufficiency to mTORC1. Science **340**:1100-1106.
Barrett, R. D. H. and H. E. Hoekstra. 2011. Molecular spandrels: tests of adaptation at the genetic level. Nature Reviews Genetics **12**:767-780.
Barribeau, S. M., B. M. Sadd, L. du Plessis, M. J. F. Brown, S. D. Buechel, et al. 2015. A depauperate immune repertoire precedes evolution of sociality in bees. Genome biology **16**:83.
Bean, B. P. 2007. The action potential in mammalian central neurons. Nature Reviews Neuroscience **8**:451-465.
Begun, D. J., A. K. Holloway, K. Stevens, L. W. Hillier, Y. P. Poh, et al. 2007. Population genomics: Whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. PLoS biology **5**:2534-2559.
Ben-Shahar, Y., N. L. Dudek, and G. E. Robinson. 2004. Phenotypic deconstruction reveals involvement of manganese transporter *malvolio* in honey bee division of labor. Journal of Experimental Biology **207**:3281-3288.
Ben-Shahar, Y., A. Robichon, M. B. Sokolowski, and G. E. Robinson. 2002. Influence of gene action across different time scales on behavior. Science **296**:741-744.
Benjamini, Y. and Y. Hochberg. 1995. Controlling the false discovery rate - a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B-Methodological **57**:289-300.
Bernardo, T. J. and E. B. Dubrovsky. 2012. The *Drosophila* juvenile hormone receptor candidates methoprene-tolerant (MET) and germ cell-expressed (GCE) utilize a conserved LIXXL motif to bind the FTZ-F1 nuclear receptor. The Journal of biological chemistry **287**:7821-7833.
Beye, M., I. Gattermeier, M. Hasselmann, T. Gempe, M. Schioett, et al. 2006. Exceptionally high levels of recombination across the honey bee genome. Genome Research **16**:1339-1344.
Bonhomme, M., C. Chevalet, B. Servin, S. Boitard, J. Abdallah, S. Blott, and M. SanCristobal. 2010. Detecting Selection in Population Trees: The Lewontin and Krakauer Test Extended. Genetics **186**:241-U406.
Bourke, A. F. G. 2011. Principles of social evolution. Oxford University Press, Oxford ; New York.
Boutin, S., M. Alburaki, P. L. Mercier, P. Giovenazzo, and N. Derome. 2015. Differential gene expression between hygienic and non-hygienic honeybee (*Apis mellifera* L.) hives. Bmc Genomics **16**:500.
Braverman, J. M., R. R. Hudson, N. L. Kaplan, C. H. Langley, and W. Stephan. 1995. The Hitchhiking Effect on the Site Frequency-Spectrum of DNA Polymorphisms. Genetics **140**:783-796.

Breed, M. D., E. Guzman-Novoa, and G. J. Hunt. 2004. Defensive behavior of honey bees: Organization, genetics, and comparisons with other bees. Annual Review of Entomology **49**:271-298.

Buchler, R., S. Berg, and Y. Le Conte. 2010. Breeding for resistance to *Varroa destructor* in Europe. Apidologie **41**:393-408.

Bustamante, C. D., A. Fledel-Alon, S. Williamson, R. Nielsen, M. T. Hubisz, et al. 2005. Natural selection on protein-coding genes in the human genome. Nature **437**:1153-1157.

Byatt, M. A., N. C. Chapman, T. Latty, and B. P. Oldroyd. 2016. The genetic consequences of the anthropogenic movement of social bees. Insectes Sociaux **63**:15-24.

Cameron, S. A., H. M. Hines, and P. H. Williams. 2007. A comprehensive phylogeny of the bumble bees (*Bombus*). Biological Journal of the Linnean Society **91**:161-188.

Chan, Q. W., M. Y. Chan, M. Logan, Y. Fang, H. Higo, and L. J. Foster. 2013. Honey bee protein atlas at organ-level resolution. Genome Research **23**:1951-1960.

Chandrasekaran, S., S. A. Ament, J. A. Eddy, S. L. Rodriguez-Zas, B. R. Schatz, N. D. Price, and G. E. Robinson. 2011. Behavior-specific changes in transcriptional modules lead to distinct and predictable neurogenomic states. Proceedings of the National Academy of Sciences of the United States of America **108**:18020-18025.

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

Chang, C. C., C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience **4**:7.

Chapman, N. C., B. A. Harpur, J. Lim, T. E. Rinderer, M. H. Allsopp, A. Zayed, and B. P. Oldroyd. 2015a. Hybrid origins of Australian honey bees (*Apis mellifera*). Apidologie **47**:26–34.

Chapman, N. C., B. A. Harpur, J. Lim, T. E. Rinderer, M. H. Allsopp, A. Zayed, and B. P. Oldroyd. 2015b. A SNP test to identify Africanized honeybees via proportion of "African" ancestry. Molecular Ecology Resources **15**:1346-1355.

Chapman, N. C., J. Lim, and B. P. Oldroyd. 2008. Population genetics of commercial and feral honey bees in Western Australia. Journal of Economic Entomology **101**:272-277.

Chavez-Galarza, J., D. Henriques, J. S. Johnston, J. C. Azevedo, J. C. Patton, I. Munoz, P. de La Rua, and M. A. Pinto. 2013. Signatures of selection in the Iberian honey bee (*Apis mellifera iberiensis*) revealed by a genome scan analysis of single nucleotide polymorphisms (SNPs). Molecular Ecology **22**:5890-5907.

Chen, S. D., B. H. Krinsky, and M. Y. Long. 2013. New genes as drivers of phenotypic evolution. Nature Reviews Genetics **14**:645-660.

Chiang, C. W. K., Z. K. Z. Gajdos, J. M. Korn, F. G. Kuruvilla, J. L. Butler, et al. 2010. Rapid Assessment of Genetic Ancestry in Populations of Unknown Origin by Genome-Wide Genotyping of Pooled Samples. PLoS Genetics **6**:e1000866.

Cingolani, P., A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Y. Lu, and D. M. Ruden. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w(1118); iso-2; iso-3. Fly **6**:80-92.

Clarke, K. E., T. E. Rinderer, P. Franck, J. G. Quezada-Euán, and B. P. Oldroyd. 2002. The Africanization of honeybees (*Apis mellifera* L.) of the Yucatan: a study of a massive hybridization event across time. Evolution **56**:1462-1474.

Cobey, S., W. S. Sheppard, and D. R. Tarpy. 2012. Status of breeding practices and genetic diversity in domestic US honey bees. Honey Bee Colony Health: Challenges and Sustainable Solutions. CRC, Boca Raton, FL:39-49.

Collet, T., K. M. Ferreira, M. C. Arias, A. E. E. Soares, and M. A. Del Lama. 2006. Genetic structure of Africanized honeybee populations (Apis mellifera L.) from Brazil and Uruguay viewed through mitochondrial DNA COI-COII patterns. Heredity **97**:329-335.

Collins, A. M., T. E. Rinderer, J. B. Harbo, and A. B. Bolten. 1982. Colony defense by Africanized and European honey bees. Science **218**:72-74.

Colosimo, P. F., C. L. Peichel, K. Nereng, B. K. Blackman, M. D. Shapiro, D. Schluter, and D. M. Kingsley. 2004. The genetic architecture of parallel armor plate reduction in threespine sticklebacks. PLoS biology **2**:635-641.

Consortium, C. e. S. 1998. Genome sequence of the nematode C-elegans: A platform for investigating biology. Science **282**:2012-2018.

Corbett-Detig, R. and R. Nielsen. 2017. A Hidden Markov Model Approach for Simultaneously Estimating Local Ancestry and Admixture Time Using Next Generation Sequence Data in Samples of Arbitrary Ploidy. PLoS Genet **13**:e1006529.

Cornuet, J.-M. 1986. Population Genetics. Pages 235-254 *in* T. E. Rinderer, editor. Bee Genetics and Breeding. Academic Press, Orlando, USA.

Corona, M., R. A. Velarde, S. Remolina, A. Moran-Lauter, Y. Wang, K. A. Hughes, and G. E. Robinson. 2007. Vitellogenin, juvenile hormone, insulin signaling, and queen honey bee longevity. Proceedings of the National Academy of Sciences of the United States of America **104**:7128-7133.

Cotter, S. C. and R. M. Kilner. 2010. Personal immunity versus social immunity. Behavioral Ecology **21**:663-668.

Crane, E. 1999. The world history of beekeeping and honey hunting. Routledge, New York.

Cremer, S., S. A. O. Armitage, and P. Schmid-Hempel. 2007. Social immunity. Current Biology **17**:R693-R702.

Cremer, S. and M. Sixt. 2009. Analogies in the evolution of individual and social immunity. Philosophical Transactions of the Royal Society B-Biological Sciences **364**:129-142.

Crespi, B. J. and D. Yanega. 1995. The Definition of Eusociality. Behavioral Ecology **6**:109-115.

Cronin, A. L., M. Molet, C. Doums, T. Monnin, and C. Peeters. 2013. Recurrent Evolution of Dependent Colony Foundation Across Eusocial Insects. Annual Review of Entomology, Vol 58 **58**:37-55.

Csanadi, G., J. Vollmann, G. Stift, and T. Lelley. 2001. Seed quality QTLs identified in a molecular map of early maturing soybean. Theoretical and Applied Genetics **103**:912-919.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, et al. 2011. The variant call format and VCFtools. Bioinformatics **27**:2156-2158.

Darwin, C. 1860. On the origin of species by means of natural selection, or The preservation of favoured races in the struggle for life. D. Appleton and company, New York,.

De Kovel, C. G. F. 2006. The power of allele frequency comparisons to detect the footprint of selection in natural and experimental situations. Genetics Selection Evolution **38**:3-23.

De la Rua, P., R. Jaffe, R. Dall'Olio, I. Munzos, and J. Serrana. 2009. Biodiversity, conservation and current threats to European honeybees. Apidologie **40**:263-284.

De la Rua, P., R. Jaffe, I. Munoz, J. Serrano, R. F. A. Moritz, and F. B. Kraus. 2013. Conserving genetic diversity in the honeybee: Comments on Harpur *et al*. (2012). Molecular Ecology **22**:3208-3210.

Delaney, D. A., M. D. Meixner, N. M. Schiff, and W. S. Sheppard. 2009. Genetic Characterization of Commercial Honey Bee (Hymenoptera: Apidae) Populations in the United States by Using Mitochondrial and Microsatellite Markers. Annals of the Entomological Society of America **102**:666-673.

DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics **43**:491-498.

Drapeau, M. D., S. Albert, R. Kucharski, C. Prusko, and R. Maleszka. 2006. Evolution of the Yellow/Major Royal Jelly Protein family and the emergence of social behavior in honey bees. Genome Research **16**:1385-1394.

Earl, D. A. and B. M. Vonholdt. 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genetics Resources **4**:359-361.

Edwards, A. C., L. Zwarts, A. Yamamoto, P. Callaerts, and T. F. Mackay. 2009. Mutations in many genes affect aggressive behavior in Drosophila melanogaster. BMC biology **7**:29.

Eilertson, K. E., J. G. Booth, and C. D. Bustamante. 2012. SnIPRE: Selection Inference Using a Poisson Random Effects Model. Plos Computational Biology **8**:e1002806.

Elsik, C. G., K. C. Worley, A. K. Bennett, M. Beye, F. Camara, et al. 2014. Finding the missing honey bee genes: lessons learned from a genome upgrade. Bmc Genomics **15**:86.

Evanno, G., S. Regnaut, and J. Goudet. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Molecular Ecology **14**:2611-2620.

Evans, J. D., K. Aronstein, Y. P. Chen, C. Hetru, J. L. Imler, et al. 2006. Immune pathways and defence mechanisms in honey bees Apis mellifera. Insect Molecular Biology **15**:645-656.

Excoffier, L. and H. E. L. Lischer. 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecology Resources **10**:564-567.

Eyre-Walker, A. 2006. The genomic rate of adaptive evolution. Trends in Ecology & Evolution **21**:569-575.

Fariello, M. I., S. Boitard, H. Naya, M. SanCristobal, and B. Servin. 2013. Detecting Signatures of Selection Through Haplotype Differentiation Among Hierarchically Structured Populations. Genetics **193**:929-U448.

Feldmeyer, B., D. Elsner, and S. Foitzik. 2013. Gene expression patterns associated with caste and reproductive status in ants: worker-specific genes are more derived than queen-specific ones. Molecular Ecology:doi: 10.1111/mec.12490.

Ferreira, P. G., S. Patalano, R. Chauhan, R. Ffrench-Constant, T. Gabaldon, R. Guigo, and S. Sumner. 2013. Transcriptome analyses of primitively eusocial wasps reveal novel insights into the evolution of sociality and the origin of alternative phenotypes. Genome biology **14**:R20.

Ferretti, L., S. E. Ramos-Onsins, and M. Perez-Enciso. 2013. Population genomics from pool sequencing. Molecular Ecology **22**:5561-5576.

Fisher, R. A. 1930. The Genetic Theory of Natural Selection. Dover, New York.
Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, et al. 1995. Whole-Genome Random Sequencing and Assembly of *Haemophilus influenzae* Rd. Science **269**:496-512.
Fletcher, D. J. 1978. The African bee, *Apis mellifera adansonii*, in Africa. Annual Review of Entomology **23**:151-171.
Flores, J. M., J. A. Ruiz, J. M. Ruz, F. Puerta, and M. Bustos. 2001. Hygienic behaviour of *Apis mellifera iberica* against brood cells artificially infested with varroa. Journal of Apicultural Research **40**:29-34.
Franck, P., L. Garnery, A. Loiseau, B. P. Oldroyd, H. R. Hepburn, M. Solignac, and J. M. Cornuet. 2001. Genetic diversity of the honeybee in Africa: microsatellite and mitochondrial data. Heredity **86**:420-430.
Franck, P., L. Garnery, M. Solignac, and J. M. Cornuet. 2000. Molecular confirmation of a fourth lineage in honeybees from the Near East. Apidologie **31**:167-180.
Free, J. B. and C. G. Butler. 1959. Bumblebees. Collins, London,.
Funada, M., H. Hara, H. Sasagawa, Y. Kitagawa, and T. Kadowaki. 2007. A honey bee Dscam family member, AbsCAM, is a brain-specific cell adhesion molecule with the neurite outgrowth activity which influences neuronal wiring during development. European Journal of Neuroscience **25**:168-180.
Gadagkar, R. 1997. The evolution of caste polymorphism in social insects: genetic release followed by diversifying evolution. Journal of Genetics **76**:167-179.
Galindo-Cardona, A., J. P. Acevedo-Gonzalez, B. Rivera-Marchand, and T. Giray. 2013. Genetic structure of the gentle Africanized honey bee population (gAHB) in Puerto Rico. Bmc Genetics **14**:65.
Garnery, L., J. M. Cornuet, and M. Solignac. 1992. Evolutionary history of the honey bee Apis mellifera inferred from mitochondrial DNA analysis. Molecular Ecology **1**:145-154.
Garnery, L., M. Solignac, G. Celebrano, and J. M. Cornuet. 1993. A simple test using restricted PCR amplified mitochondrial DNA to study the genetic structure of *Apis mellifera* L. Experientia **49**:1016-1021.
Gautier, M., A. Klassmann, and R. Vitalis. 2016. rehh 2.0: a reimplementation of the r package rehh to detect positive selection from haplotype structure. Molecular Ecology Resources **17**:78-90.
Gerula, D., P. Wegrzynowicz, B. Panasiuk, M. Bienkowska, and W. Skowronek. 2015. Hygienic Behaviour of Honeybee Colonies with Different Levels of Polyandry and Genotypic Composition. Journal of Apicultural Science **59**:107-113.
Gibson, J. D., M. E. Arechavaleta-Velasco, J. M. Tsuruda, and G. J. Hunt. 2015. Biased Allele Expression and Aggression in Hybrid Honeybees may be Influenced by Inappropriate Nuclear-Cytoplasmic Signaling. Frontiers in Genetics **6**.
Goffeau, A., B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, et al. 1996. Life with 6000 genes. Science **274**:546.
Goudet, J., M. Raymond, T. deMeeus, and F. Rousset. 1996. Testing differentiation in diploid populations. Genetics **144**:1933-1940.
Goulson, D. 2010. Bumblebees : behaviour, ecology, and conservation. 2nd edition. Oxford University Press, Oxford ; New York.

Grisart, B., W. Coppieters, F. Farnir, L. Karim, C. Ford, et al. 2002. Positional candidate cloning of a QTL in dairy cattle: Identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Research **12**:222-231.
Guan, Y. 2014. Detecting structure of haplotypes and local ancestry. Genetics **196**:625-642.
Guarna, M. M., A. P. Melathopoulos, E. Huxter, I. Iovinella, R. Parker, et al. 2015. A search for protein biomarkers links olfactory signal transduction to social immunity. Bmc Genomics **16**:63.
Guzman-Novoa, E., G. J. Hunt, J. L. Uribe, C. Smith, and M. E. Arechavaleta-Velasco. 2002. Confirmation of QTL effects and evidence of genetic dominance of honeybee defensive behavior: results of colony and individual behavioral assays. Behav Genet **32**:95-102.
Guzman-Novoa, E., R. E. Page, and M. K. Fondrk. 1994. Morphometric Techniques Do Not Detect Intermediate and Low-Levels of Africanization in Honey-Bee (Hymenoptera, Apidae) Colonies. Annals of the Entomological Society of America **87**:507-515.
Hahn, M. and H. Jäckle. 1996. Drosophila goosecoid participates in neural development but not in body axis formation. The EMBO journal **15**:3077.
Hall, D. W. and M. A. D. Goodisman. 2012. The effects of kin selection on rates of molecular evolution in social insects. Evolution **66**:2080-2093.
Hamburg, M. A. and F. S. Collins. 2010. The Path to Personalized Medicine. New England Journal of Medicine **363**:301-304.
Hamilton, W. D. 1964a. The genetical evolution of social behavior I. Journal of Theoretical Biology **7**:1-16.
Hamilton, W. D. 1964b. The genetical evolution of social behavior II. Journal of Theoretical Biology **7**:17-52.
Harbo, J. R. and J. W. Harris. 1999. Heritability in honey bees (Hymenoptera : Apidae) of characteristics associated with resistance to Varroa jacobsoni (Mesostigmata : Varroidae). Journal of Economic Entomology **92**:261-265.
Harbo, J. R. and J. W. Harris. 2005. Suppressed mite reproduction explained by the behaviour of adult bees. Journal of Apicultural Research **44**:21-23.
Harpur, B. A., N. C. Chapman, L. Krimus, P. Maciukiewicz, V. Sandhu, et al. 2015. Assessing patterns of admixture and ancestry in Canadian honey bees. Insectes Sociaux **62**:479-489.
Harpur, B. A., A. Chernyshova, A. Soltani, N. Tsvetkov, M. Mahjoorighasrodashti, Z. X. Xu, and A. Zayed. 2014a. No Genetic Tradeoffs between Hygienic Behaviour and Individual Innate Immunity in the Honey Bee, Apis mellifera. Plos One **9**: e104214.
Harpur, B. A., C. F. Kent, D. Molodtsova, J. M. D. Lebon, A. S. Alqarni, A. A. Owayss, and A. Zayed. 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., S. Minaei, C. F. Kent, and A. Zayed. 2012. Management increases genetic diversity of honey bees via admixture. Molecular Ecology **21**:4414-4421.
Harpur, B. A., S. Minaei, C. F. Kent, and A. Zayed. 2013. Admixture increases diversity in managed honey bees: Reply to De la Rua etal. (2013). Molecular Ecology **22**:3211-3215.
Harpur, B. A. and A. Zayed. 2013. Accelerated Evolution of Innate Immunity Proteins in Social Insects: Adaptive Evolution or Relaxed Constraint? Molecular Biology and Evolution **30**:1665-1674.

Hartl, D. L. and A. G. Clark. 2007. Principles of population genetics. 4th edition. Sinauer Associates, Sunderland, Mass.
Hasselmann, M., L. Ferretti, and A. Zayed. 2015. Beyond fruit-flies: population genomic advances in non-Drosophila arthropods. Briefings in Functional Genomics **14**:424-431.
Heinze, J. and B. Walter. 2010. Moribund Ants Leave Their Nests to Die in Social Isolation. Current Biology **20**:249-252.
Herbers, J. M. 2009. Darwin's 'one special difficulty': celebrating Darwin 200. Biology Letters **5**:214-217.
Hiendleder, S., S. Bauersachs, A. Boulesteix, H. Blum, G. J. Arnold, T. Frohlich, and E. Wolf. 2005. Functional genomics: tools for improving farm animal health and welfare. Revue Scientifique Et Technique-Office International Des Epizooties **24**:355-377.
Hines, H. M. 2008. Historical biogeography, divergence times, and diversification patterns of bumble bees (Hymenoptera : Apidae : Bombus). Systematic Biology **57**:58-75.
Hodgkinson, A. and A. Eyre-Walker. 2011. Variation in the mutation rate across mammalian genomes. Nature Reviews Genetics **12**:756-766.
Hohenlohe, P. A., P. C. Phillips, and W. A. Cresko. 2010. Using Population Genomics to Detect Selection in Natural Populations: Key Concepts and Methodological Considerations. International Journal of Plant Sciences **171**:1059-1071.
Hölldobler, B. and E. O. Wilson. 1990. The ants. Belknap Press of Harvard University Press, Cambridge, Mass.
Hölldobler, B. and E. O. Wilson. 2009. The superorganism : the beauty, elegance, and strangeness of insect societies. 1st edition. W.W. Norton, New York.
Hopkins, I. 1886. Illustrated Australasian Bee Manual and Complete Guide to Modern Bee Culture in the Southern Hemispere. 3rd edition. Issac Hopkins, Auckland, New Zealand.
Huang, D. W., B. T. Sherman, and R. A. Lempicki. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols **4**:44-57.
Hunt, B. G., S. Wyder, N. Elango, J. H. Werren, E. M. Zdobnov, S. V. Yi, and M. A. Goodisman. 2010. Sociality is linked to rates of protein evolution in a highly social insect. Molecular Biology and Evolution **27**:497-500.
Hunt, G. J., A. M. Collins, R. Rivera, R. E. Page, and E. Guzman-Novoa. 1999. Quantitative trait loci influencing honeybee alarm pheromone levels. Journal of Heredity **90**:585-589.
Hunt, G. J., E. Guzman-Novoa, M. K. Fondrk, and R. E. Page. 1998. Quantitative trait loci for honey bee stinging behavior and body size. Genetics **148**:1203-1213.
Jasper, W. C., T. A. Linksvayer, J. Atallah, D. Friedman, J. C. Chiu, and B. R. Johnson. 2015. Large-Scale Coding Sequence Change Underlies the Evolution of Postdevelopmental Novelty in Honey Bees. Molecular Biology and Evolution **32**:334-346.
Jensen, A. B., K. A. Palmer, J. J. Boomsma, and B. V. Pedersen. 2005. Varying degrees of *Apis mellifera ligustica* introgression in protected populations of the black honeybee, *Apis mellifera mellifera*, in northwest Europe. Molecular Ecology **14**:93-106.
Jepson, J. E. C., M. Shahidullah, A. Lamaze, D. Peterson, H. H. Pan, and K. Koh. 2012. dyschronic, a Drosophila Homolog of a Deaf-Blindness Gene, Regulates Circadian Output and Slowpoke Channels. PLoS Genetics **8**:549-563.
Jepson, J. E. C., M. Shahidullah, D. Liu, S. J. Le Marchand, S. Liu, M. N. Wu, I. B. Levitan, M. B. Dalva, and K. Koh. 2014. Regulation of synaptic development and function by the *Drosophila* PDZ protein Dyschronic. Development **141**:4548-4557.

Jindra, M., S. R. Palli, and L. M. Riddiford. 2013. The juvenile hormone signaling pathway in insect development. Annual Review of Entomology **58**:181-204.
Johnson, B. R. and N. D. Tsutsui. 2011. Taxonomically restricted genes are associated with the evolution of sociality in the honey bee. Bmc Genomics **12**:164.
Jolly, B. 2004. South Australia's early Ligurian beekeeping - and a lingering Kangaroo Island fable. Journal of the Historical Society of South Australia **32**:69-81.
Jones, J. C., M. R. Myerscough, S. Graham, and B. P. Oldroyd. 2004. Honey bee nest thermoregulation: Diversity promotes stability. Science **305**:402-404.
Juge, N., J. A. Gray, H. Omote, T. Miyaji, T. Inoue, et al. 2010. Metabolic Control of Vesicular Glutamate Transport and Release. Neuron **68**:99-112.
Kadri, S. M., B. A. Harpur, R. O. Orsi, and A. Zayed. 2016. A variant reference data set for the Africanized honeybee, *Apis mellifera*. Scientific Data **3**:160097.
Kamakura, M. 2011. Royalactin induces queen differentiation in honeybees. Nature **473**:478-483.
Kapheim, K. M., H. L. Pan, C. Li, S. L. Salzberg, D. Puiu, et al. 2015. Genomic signatures of evolutionary transitions from solitary to group living. Science **348**:1139-1143.
Kent, C. F., A. Issa, A. C. Bunting, and A. Zayed. 2011. Adaptive evolution of a key gene affecting queen and worker traits in the honey bee, *Apis mellifera*. Molecular Ecology **20**:5226-5235.
Kent, C. F., S. Minaei, B. A. Harpur, and A. Zayed. 2012. Recombination is associated with the evolution of genome structure and worker behavior in honey bees. Proceedings of the National Academy of Sciences of the United States of America **109**:18012-18017.
Kent, C. F. and A. Zayed. 2015. Chapter Nine-Population Genomic and Phylogenomic Insights into the Evolution of Physiology and Behaviour in Social Insects. Advances in Insect Physiology **48**:293-324.
Kerr, W. E. 1957. Introdução de abelhas africanas no Brasil. Brasil Apicola **3**:211-213.
Kerr, W. E. 1967. The history of the introduction of Africanized honey bees to Brazil. South African Bee Journal **39**:3-5.
Kingman, J. F. C. 2000. Origins of the coalescent: 1974-1982. Genetics **156**:1461-1463.
Koboldt, D. C., K. Chen, T. Wylie, D. E. Larson, M. D. McLellan, E. R. Mardis, G. M. Weinstock, R. K. Wilson, and L. Ding. 2009. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics **25**:2283-2285.
Kofler, R., R. V. Pandey, and C. Schlotterer. 2011. PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). Bioinformatics **27**:3435-3436.
Kotthoff, U., T. Wappler, and M. S. Engel. 2013. Greater past disparity and diversity hints at ancient migrations of European honey bee lineages into Africa and Asia. Journal of Biogeography **40**:1832-1838.
Koulianos, S. and R. Crozier. 1996. Mitochondrial DNA sequence data provides further evidence that the honeybees of Kangaroo Island, Australia are of hybrid origin. Apidologie **27**:165-174.
Koulianos, S. and R. Crozier. 1997. Mitochondrial sequence characterisation of Australian commercial and feral honeybee strains, *Apis mellifera* L. (Hymenoptera: Apidae), in the context of the species worldwide. Australian Journal of Entomology **36**:359-363.

Laine, V. N., G. Herczeg, T. Shikano, J. Vilkki, and J. Merila. 2014. QTL Analysis of Behavior in Nine-Spined Sticklebacks (*Pungitius pungitius*). Behav Genet **44**:77-88.
Lander, E. S. and I. H. G. S. Consortium and L. M. Linton and B. Birren and C. Nusbaum, et al. 2001. Initial sequencing and analysis of the human genome. Nature **409**:860-921.
Langstroth, L. and C. Dadant. 1889. Langstroth on the hive and honey bee. C. Dadant & son, United States.
Langstroth, L. L. 1865. A practical treatise on the hive and honey-bee. 3d edition. Lippincott, Philadelphia,.
Lapidge, K. L., B. P. Oldroyd, and M. Spivak. 2002. Seven suggestive quantitative trait loci influence hygienic behavior of honey bees. Naturwissenschaften **89**:565-568.
Lawniczak, M. K. N., A. I. Barnes, J. R. Linklater, J. M. Boone, S. Wigby, and T. Chapman. 2007. Mating and immunity in invertebrates. Trends in Ecology & Evolution **22**:48-55.
Le Conte, Y. and M. Navajas. 2008. Climate change: impact on honey bee populations and diseases. Revue Scientifique Et Technique-Office International Des Epizooties **27**:499-510.
Lecocq, A., A. B. Jensen, P. Kryger, and J. C. Nieh. 2016. Parasite infection accelerates age polyethism in young honey bees. Scientific Reports **6**.
Leffler, E. M., Z. Y. Gao, S. Pfeifer, L. Segurel, A. Auton, et al. 2013. Multiple Instances of Ancient Balancing Selection Shared Between Humans and Chimpanzees. Science **339**:1578-1582.
Li-Byarlay, H., C. C. Rittschof, J. H. Massey, B. R. Pittendrigh, and G. E. Robinson. 2014. Socially responsive effects of brain oxidative metabolism on aggression. Proceedings of the National Academy of Sciences of the United States of America **111**:12533-12537.
Li, H. and R. Durbin. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics **26**:589-595.
Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, et al. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**:2078-2079.
Li, M., E. A. Mead, and J. Zhu. 2011. Heterodimer of two bHLH-PAS proteins mediates juvenile hormone-induced gene expression. Proceedings of the National Academy of Sciences of the United States of America **108**:638-643.
Li, W. J. and F. B. Gao. 2003. Actin filament-stabilizing protein tropomyosin regulates the size of dendritic fields. Journal of Neuroscience **23**:6171-6175.
Linksvayer, T. A. and M. J. Wade. 2009. Genes with Social Effects Are Expected to Harbor More Sequence Variation within and between Species. Evolution **63**:1685-1696.
Liu, H. X., X. H. Zhang, J. Huang, J. Q. Chen, D. C. Tian, L. D. Hurst, and S. H. Yang. 2015. Causes and consequences of crossing-over evidenced via a high-resolution recombinational landscape of the honey bee. Genome biology **16**:15.
Lobo, N. F., L. Q. Ton, C. A. Hill, C. Emore, J. Romero-Severson, G. J. Hunt, and F. H. Collins. 2003. Genomic analysis in the sting-2 quantitative trait locus for defensive behavior in the honey bee, *Apis mellifera*. Genome Research **13**:2588-2593.
Lopez-Uribe, M. M., W. B. Sconiers, S. D. Frank, R. R. Dunn, and D. R. Tarpy. 2016. Reduced cellular immune response in social insect lineages. Biology Letters **12**:20150984.
Lunter, G. and M. Goodson. 2011. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Research **21**:936-939.

Lynch, M. and B. Walsh. 1998. Genetics and analysis of quantitative traits. Sinauer Sunderland, MA.
Mardis, E. R. 2011. A decade's perspective on DNA sequencing technology. Nature **470**:198-203.
Masterman, R., R. Ross, K. Mesce, and M. Spivak. 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-Sensory Neural and Behavioral Physiology **187**:441-452.
Mattila, H. R. and T. D. Seeley. 2007. Genetic diversity in honey bee colonies enhances productivity and fitness. Science **317**:362-364.
Mcdonald, J. H. and M. Kreitman. 1991. Adaptive Protein Evolution at the Adh Locus in Drosophila. Nature **351**:652-654.
McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research **20**:1297-1303.
Meixner, M. D., C. Costa, P. Kryger, F. Hatjina, M. Bouga, E. Ivanova, and R. Buchler. 2010. Conserving diversity and vitality for honey bee breeding. Journal of Apicultural Research **49**:85-92.
Michener, C. D. 1974. The social behavior of the bees; a comparative study. Belknap Press of Harvard University Press, Cambridge, Mass.,.
Moritz, R. F. A., S. Hartel, and P. Neumann. 2005. Global invasions of the western honeybee (*Apis mellifera*) and the consequences for biodiversity. Ecoscience **12**:289-301.
Munoz-Torres, M. C., J. T. Reese, C. P. Childers, A. K. Bennett, J. P. Sundaram, K. L. Childs, J. M. Anzola, N. Milshina, and C. G. Elsik. 2011. Hymenoptera Genome Database: integrated community resources for insect species of the order Hymenoptera. Nucleic Acids Research **39**:D658-D662.
Munoz, I., D. Henriques, J. S. Johnston, J. Chavez-Galarza, P. Kryger, and M. A. Pinto. 2015. Reduced SNP panels for genetic identification and introgression analysis in the dark honey bee (*Apis mellifera mellifera*). Plos One **10**:e0124365.
Nelson, C. M., K. E. Ihle, M. K. Fondrk, R. E. J. Page, and G. V. Amdam. 2007. The gene *vitellogenin* has multiple coordinating effects on social organization. PLoS biology **5**:e62.
Nielsen, D. I., P. R. Ebert, R. E. Page, G. J. Hunt, and E. Guzman-Novoa. 2000. Improved polymerase chain reaction-based mitochondrial genotype assay for identification of the africanized honey bee (Hymenoptera : Apidae). Annals of the Entomological Society of America **93**:1-6.
Nielsen, R. 2005. Molecular signatures of natural selection. Annual Review of Genetics **39**:197-218.
Nielsen, R., I. Hellmann, M. Hubisz, C. Bustamante, and A. G. Clark. 2007. Recent and ongoing selection in the human genome. Nature Reviews Genetics **8**:857-868.
Nijhout, H. F. and S. M. Paulsen. 1997. Developmental models and polygenic characters. American Naturalist **149**:394-405.
Nogueira-Neto, P. 1964. The spread of a fierce African bee in Brazil. Bee World **45**:119-121.
Noll, F. B. 2002. Behavioral phylogeny of corbiculate Apidae (Hymenoptera; Apinae), with special reference to social behavior. Cladistics **18**:137-153.

Nunn, C. L., F. Jordan, C. M. McCabe, J. L. Verdolin, and J. H. Fewell. 2015. Infectious disease and group size: more than just a numbers game. Philosophical Transactions of the Royal Society B-Biological Sciences **370**.

O'Connell, J., D. Gurdasani, O. Delaneau, N. Pirastu, S. Ulivi, et al. 2014. A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. PLoS Genetics **10**:e1004234.

Oldroyd, B. P., J. M. Cornuet, D. Rowe, T. E. Rinderer, and R. H. Crozier. 1995. Racial admixture of *Apis mellifera* in Tasmania, Australia - similarities and differences with natural hybrid zones in Europe. Heredity **74**:315-325.

Oldroyd, B. P., W. S. Sheppard, and J. A. Stelzer. 1992. Genetic characterization of the bees of Kangaroo Island, South Australia. Journal of Apicultural Research **31**:141-148.

Otis, G. W. 2015. Statement regarding use of A. m. sahariensis and A. m. monticola in Ontario's buckfast breeding program.

Oxley, P. R. and B. P. Oldroyd. 2009. Mitochondrial sequencing reveals five separate origins of 'Black' *Apis mellifera* (Hymenoptera: Apidae) in Eastern Australian commercial colonies. Journal of Economic Entomology **102**:480-484.

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

Packer, L. 1997. The relevance of phylogenetic systematics to biology: examples from medicine and behavioural ecology. Pages 11-29 *in* P. Grandcolas, editor. The Origin of Biodiversity in Insects: Phylogenetic Tests of Evolutionary Scenarios Museum National d'Histoire Naturelle.

Page, R. E. and E. H. Erickson. 1988. Reproduction by worker honey bees (*Apis mellifera* L.). Behavioral Ecology and Sociobiology **23**:117-126.

Palmer, M. R., D. R. Smith, and O. Kaftanoglu. 2000. Turkish honeybees: genetic variation and evidence for a fourth lineage of *Apis mellifera* mtDNA. Journal of Heredity **91**:42-46.

Parker, R., M. M. Guarna, A. P. Melathopoulos, K. M. Moon, R. White, E. Huxter, S. F. Pernal, and L. J. Foster. 2012. Correlation of proteome-wide changes with social immunity behaviors provides insight into resistance to the parasitic mite, Varroa destructor, in the honey bee (*Apis mellifera*). Genome biology **13**:R81.

Perez-Sato, J. A., N. Chaline, S. J. Martin, W. O. H. Hughes, and F. L. W. Ratnieks. 2009. Multi-level selection for hygienic behaviour in honeybees. Heredity **102**:609-615.

Pernal, S. F., A. Sewalem, and A. P. Melathopoulos. 2012. Breeding for hygienic behaviour in honeybees (*Apis mellifera*) using free-mated nucleus colonies. Apidologie **43**:403-416.

Pinto, M. A., D. Henriques, J. Chaves-Galarza, P. Kryger, L. Garnery, et al. 2014. Genetic integrity of the Dark European honey bee (*Apis mellifera mellifera*) from protected populations: a genome-wide assessment using SNPs and mtDNA sequence data. Journal of Apicultural Research **53**:269-278.

Pinto, M. A., W. L. Rubink, J. C. Patton, R. N. Coulson, and J. S. Johnston. 2005. Africanization in the United States: Replacement of feral European honeybees (*Apis mellifera* L.) by an African hybrid swarm. Genetics **170**:1653-1665.

Pinto, M. A., W. S. Sheppard, J. S. Johnston, W. L. Rubink, R. N. Coulson, N. M. Schiff, I. Kandemir, and J. C. Patton. 2007. Honey bees (Hymenoptera : Apidae) of African origin

exist in non-Africanized areas of the Southern United States: Evidence from mitochondrial DNA. Annals of the Entomological Society of America **100**:289-295.
Pires, C. V., F. C. D. Freitas, A. S. Cristino, P. K. Dearden, and Z. L. P. Simoes. 2016. Transcriptome Analysis of Honeybee (*Apis mellifera*) Haploid and Diploid Embryos Reveals Early Zygotic Transcription during Cleavage. Plos One **11**.
Posnien, N., N. D. B. Koniszewski, H. J. Hein, and G. Bucher. 2011. Candidate Gene Screen in the Red Flour Beetle *Tribolium* Reveals Six3 as Ancient Regulator of Anterior Median Head and Central Complex Development. PLoS Genetics **7**.
Poulsen, M., A. N. M. Bot, and J. J. Boomsma. 2003. The effect of metapleural gland secretion on the growth of a mutualistic bacterium on the cuticle of leaf-cutting ants. Naturwissenschaften **90**:406-409.
Pritchard, J. K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. Genetics **155**:945-959.
Provine, W. B. 1971. The origins of theoretical population genetics. University of Chicago Press, Chicago,.
Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics **81**:559-575.
Qanbari, S., T. M. Strom, G. Haberer, S. Weigend, A. A. Gheyas, et al. 2012. A High Resolution Genome-Wide Scan for Significant Selective Sweeps: An Application to Pooled Sequence Data in Laying Chickens. Plos One **7**:e49525.
Queller, D. C. and K. F. Goodnight. 1989. Estimating Relatedness Using Genetic-Markers. Evolution **43**:258-275.
R Core Team. 2010. R: a language and environment for statistical computing. R Foundation for Statistical Computing Vienna Austria.
Randhawa, I. A. S., M. S. Khatkar, P. C. Thomson, and H. W. Raadsma. 2014. Composite selection signals can localize the trait specific genomic regions in multi-breed populations of cattle and sheep. Bmc Genetics **15**:34.
Raymond, M. and F. Rousset. 1995. Genepop (v. 1.2) - Population genetics software for exact tests and ecumenicism. Journal of Heredity **86**:248-249.
Rehan, S. M. and A. L. Toth. 2015. Climbing the social ladder: the molecular evolution of sociality. Trends in Ecology & Evolution **30**:426-433.
Riddell, C. E., J. D. L. Garces, S. Adams, S. M. Barribeau, D. Twell, and E. B. Mallon. 2014. Differential gene expression and alternative splicing in insect immune specificity. Bmc Genomics **15**:1031.
Rinderer, T. E., J. W. Harris, G. J. Hunt, and L. I. de Guzman. 2010. Breeding for resistance to Varroa destructor in North America. Apidologie **41**:409-424.
Rinderer, T. E., J. A. Stelzer, B. P. Oldroyd, S. M. Buco, and W. L. Rubink. 1991. Hybridization between European and Africanized Honey-Bees in the Neotropical Yucatan Peninsula. Science **253**:309-311.
Rittschof, C. C., S. A. Bukhari, L. G. Sloofman, J. M. Troy, D. Caetano-Anolles, et al. 2014. Neuromolecular responses to social challenge: common mechanisms across mouse, stickleback fish, and honey bee. Proc Natl Acad Sci U S A **111**:17929-17934.

Rittschof, C. C., C. B. Coombs, M. Frazier, C. M. Grozinger, and G. E. Robinson. 2015a. Early-life experience affects honey bee aggression and resilience to immune challenge. Sci Rep **5**:15572.
Rittschof, C. C., C. M. Grozinger, and G. E. Robinson. 2015b. The energetic basis of behavior: bridging behavioral ecology and neuroscience. Current Opinion in Behavioral Sciences **6**:19-27.
Rittschof, C. C. and G. E. Robinson. 2013. Manipulation of colony environment modulates honey bee aggression and brain gene expression. Genes Brain and Behavior **12**:802-811.
Rius, M. and J. A. Darling. 2014. How important is intraspecific genetic admixture to the success of colonising populations? Trends in Ecology & Evolution **29**:233-242.
Rivera-Marchand, B., D. Oskay, and T. Giray. 2012. Gentle Africanized bees on an oceanic island. Evolutionary Applications **5**:746-756.
Rockman, M. V. 2012. The QTN program and the alleles that matter for evolution: all that's gold does not glitter. Evolution **66**:1-17.
Romiguier, J., S. A. Cameron, S. H. Woodard, B. J. Fischman, L. Keller, and C. J. Praz. 2016. Phylogenomics Controlling for Base Compositional Bias Reveals a Single Origin of Eusociality in Corbiculate Bees. Molecular Biology and Evolution **33**:670-678.
Root, A. I. 1985. ABC and XYZ of bee culture. 41 edition. A I Root Co, United States.
Rosengaus, R. B., A. B. Maxmen, L. E. Coates, and J. F. A. Traniello. 1998. Disease resistance: a benefit of sociality in the dampwood termite Zootermopsis angusticollis (Isoptera : Termopsidae). Behavioral Ecology and Sociobiology **44**:125-134.
Rothenbuhler, W. C. 1964a. Behavior Genetics of Nest Cleaning in Honey Bees .4. Responses of F1 and Backcross Generations to Disease-Killed Brood. American Zoologist **4**:111-123.
Rothenbuhler, W. C. 1964b. Behaviour Genetics of Nest Cleaning in Honey Bees .I. Responses of 4 Inbred Lines to Disease-Killed Brood. Animal Behaviour **12**:578.
Ruttner, F. 1976. Isolated populations of honeybees in Australia. Journal of Apicultural Research **15**:97-104.
Ruttner, F. 1988. Biogeography and Taxonomy of Honeybees.
Sackton, T. B., B. P. Lazzaro, T. A. Schlenke, J. D. Evans, D. Hultmark, and A. G. Clark. 2007. Dynamic evolution of the innate immune system in *Drosophila*. Nature Genetics **39**:1461-1468.
Sadd, B. M. and S. M. Barribeau and G. Bloch and D. C. de Graaf and P. Dearden, et al. 2015. The genomes of two key bumblebee species with primitive eusocial organization. Genome biology **16**.
Sagili, R. R., T. Pankiw, and B. N. Metz. 2011. Division of Labor Associated with Brood Rearing in the Honey Bee: How Does It Translate to Colony Fitness? Plos One **6**.
Schiff, N. M. and W. S. Sheppard. 1995. Genetic-Analysis of Commercial Honey-Bees (Hymenoptera, Apidae) from the Southeastern United-States. Journal of Economic Entomology **88**:1216-1220.
Schmid-Hempel, P. 1994. Infection and Colony Variability in Social Insects. Philosophical Transactions of the Royal Society of London Series B-Biological Sciences **346**:313-321.
Schneider, S. S., G. D. Hoffman, and D. R. Smith. 2004. The African honey bee: Factors contributing to a successful biological invasion. Annual Review of Entomology **49**:351-376.

Sedlazeck, F. J., P. Rescheneder, and A. von Haeseler. 2013. NextGenMap: fast and accurate read mapping in highly polymorphic genomes. Bioinformatics **29**:2790-2791.
Seeley, T. D. 1985. Honeybee Ecology: A Study of Adaptation in Social Life. Princeton University Press, Princeton, USA.
Semenza, G. L. 2007. Oxygen-dependent regulation of mitochondrial respiration by hypoxia-inducible factor 1. Biochemical Journal **405**:1-9.
Sheppard, W. S. 1988. Comparative-Study of Enzyme Polymorphism in United-States and European Honey Bee (Hymenoptera, Apidae) Populations. Annals of the Entomological Society of America **81**:886-889.
Sheppard, W. S. 1989a. A history of the introduction of honey bee races into the United States. Part I. American Bee Journal **129**:617-619.
Sheppard, W. S. 1989b. A history of the introduction of honey bee races into the United States. Part II. American Bee Journal **129**:664-667.
Sheppard, W. S. 2012. Managed pollinator CAP coordinated agricultural project a national research and extension initiative to reverse pollinator decline honey bee genetic diversity and breeding-towards the reintroduction of European germ plasm. American Bee Journal **152**:155-158.
Sheppard, W. S. and D. R. Smith. 2000. Identification of African-derived bees in the Americas: A survey of methods. Annals of the Entomological Society of America **93**:159-176.
Sheppard, W. S., A. E. E. Soares, D. Dejong, and H. Shimanuki. 1991. Hybrid Status of Honey-Bee Populations near the Historic Origin of Africanization in Brazil. Apidologie **22**:643-652.
Simola, D. F., L. Wissler, G. Donahue, R. M. Waterhouse, M. Helmkampf, et al. 2013. Social insect genomes exhibit dramatic evolution in gene composition and regulation while preserving regulatory features linked to sociality. Genome Research **23**:1235-1247.
Sink, H., E. J. Rehm, L. Richstone, Y. M. Bulls, and C. S. Goodman. 2001. sidestep encodes a target-derived attractant essential for motor axon guidance in Drosophila. Cell **105**:57-67.
Song, J. B., L. L. Wu, Z. Chen, R. A. Kohanski, and L. Pick. 2003. Axons guided by insulin receptor in Drosophila visual system. Science **300**:502-505.
Southwick, E., D. Roubik, and J. Williams. 1990. Comparative energy balance in groups of Africanized and European honey bees: ecological implications. Comparative Biochemistry and Physiology Part A: Physiology **97**:1-7.
Spivak, M. and M. Gilliam. 1998a. Hygienic behaviour of honey bees and its application for control of brood diseases and varroa - Part II. Studies on hygienic behaviour since the Rothenbuhler era. Bee World **79**:169-186.
Spivak, M. and M. Gilliam. 1998b. Hygienic behaviour of honey bees and its application for control of brood diseases and varroa Part I. Hygienic behaviour and resistance to American foulbrood. Bee World **79**:124-134.
Spivak, M., R. Masterman, R. Ross, and K. A. Mesce. 2003. Hygienic behavior in the honey bee (Apis mellifera L.) and the modulatory role of octopamine. Journal of Neurobiology **55**:341-354.
Spivak, M. and G. S. Reuter. 2001. Resistance to American foulbrood disease by honey bee colonies *Apis mellifera* bred for hygienic behavior. Apidologie **32**:555-565.

Spotter, A., P. Gupta, M. Mayer, N. Reinsch, and K. Bienefeld. 2016. Genome-Wide Association Study of a *Varroa*-Specific Defense Behavior in Honeybees (*Apis mellifera*). Journal of Heredity **107**:220-227.
Spotter, A., P. Gupta, G. Nurnberg, N. Reinsch, and K. Bienefeld. 2012. Development of a 44K SNP assay focussing on the analysis of a varroa-specific defence behaviour in honey bees (Apis mellifera carnica). Molecular Ecology Resources **12**:323-332.
Storey, J. D. and R. Tibshirani. 2003. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A **100**:9440-9445.
Stort, A. C. 1975. Genetic study of the aggressiveness of two subspecies of Apis mellifera in Brazil. IV. Number of stings in the gloves of the observer. Behav Genet **5**:269-274.
Strassmann, J. E., R. E. Page, G. E. Robinson, and T. D. Seeley. 2011. Kin selection and eusociality. Nature **471**:E5-E6.
Sullivan, J. P., O. Jassim, S. E. Fahrbach, and G. E. Robinson. 2000. Juvenile hormone paces behavioral development in the adult worker honey bee. Hormones and Behavior **37**:1-14.
Sun, Q. and X. G. Zhou. 2013. Corpse Management in Social Insects. International Journal of Biological Sciences **9**:313-321.
Szalanski, A. L. and R. M. Magnus. 2010. Mitochondrial DNA characterization of Africanized honey bee (*Apis mellifera* L.) populations from the USA. Journal of Apicultural Research **49**:177-185.
Tajima, F. 1989. Statistical-Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. Genetics **123**:585-595.
Tarpy, D. R. 2003. Genetic diversity within honeybee colonies prevents severe infections and promotes colony growth. Proceedings of the Royal Society of London, Series B: Biological Sciences **270**:99-103.
Tarpy, D. R., D. A. Delaney, and T. D. Seeley. 2015. Mating frequencies of honey bee queens (*Apis mellifera* L.) in a population of feral colonies in the Northeastern United States. Plos One **10**:e0118734.
Tavares, A. 2014. Statistical Overview of the Canadian Honey Industry-2013. Pages 1-24 *in* M. A. a. I. Section, editor. Government of Canada, Canada.
Team, R. D. C. 2011. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Tieman, D., G. Zhu, M. F. R. Resende, T. Lin, C. Nguyen, et al. 2017. A chemical genetic roadmap to improved tomato flavor. Science **355**:391-394.
Toth, A. L. and G. E. Robinson. 2007. Evo-devo and the evolution of social behavior. Trends in Genetics **23**:334-341.
Tsuruda, J. M., J. W. Harris, L. Bourgeois, R. G. Danka, and G. J. Hunt. 2012. High-Resolution Linkage Analyses to Identify Genes That Influence Varroa Sensitive Hygiene Behavior in Honey Bees. Plos One **7**.
Uzunov, A., C. Costa, B. Panasiuk, M. Meixner, P. Kryger, et al. 2014. Swarming, defensive and hygienic behaviour in honey bee colonies of different genetic origin in a pan-European experiment. Journal of Apicultural Research **53**:248-260.
Valdebenito, R., I. Ruminot, P. Garrido-Gerter, I. Fernandez-Moncada, L. Forero-Quintero, K. Alegria, H. M. Becker, J. W. Deitmer, and L. F. Barros. 2016. Targeting of astrocytic glucose metabolism by beta-hydroxybutyrate. Journal of Cerebral Blood Flow and Metabolism **36**:1813-1822.

Van Laere, A. S., M. Nguyen, M. Braunschweig, C. Nezer, C. Collette, et al. 2003. A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. Nature **425**:832-836.
Venter, J. C. and M. D. Adams and E. W. Myers and P. W. Li and R. J. Mural, et al. 2001. The sequence of the human genome. Science **291**:1304-+.
Visscher, P. K. 1989. A quantitative study of worker reproduction in honey bee colonies. Behavioral Ecology and Sociobiology **25**:247-254.
Visscher, P. M., M. A. Brown, M. I. McCarthy, and J. Yang. 2012. Five Years of GWAS Discovery. American Journal of Human Genetics **90**:7-24.
Voight, B. F., S. Kudaravalli, X. Q. Wen, and J. K. Pritchard. 2006. A map of recent positive selection in the human genome. PLoS biology **4**:446-458.
Vojvodic, S., B. R. Johnson, B. A. Harpur, C. F. Kent, A. Zayed, K. E. Anderson, and T. A. Linksvayer. 2015. The transcriptomic and evolutionary signature of social interactions regulating honey bee caste development. Ecology and Evolution **5**:4795-4807.
Wallberg, A., F. Han, G. Wellhagen, B. Dahle, M. Kawata, et al. 2014. A worldwide survey of genome sequence variation provides insight into the evolutionary history of the honeybee Apis mellifera. Nature Genetics.
Wang, Y., N. S. Mutti, K. E. Ihle, A. Siegel, A. G. Dolezal, O. Kaftanoglu, and G. V. Amdam. 2010. Down-regulation of honey bee IRS gene biases behavior toward food rich in protein. PLoS Genetics **6**:e1000896.
Wang, Z. L., T. T. Liu, Z. Y. Huang, X. B. Wu, W. Y. Yan, and Z. J. Zeng. 2012. Transcriptome analysis of the Asian honey bee *Apis cerana cerana*. Plos One **7**:e47954.
Waterhouse, R. M., F. Tegenfeldt, J. Li, E. M. Zdobnov, and E. V. Kriventseva. 2013. OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. Nucleic Acids Research **41**:D358-D365.
Weinstock, G. M. and G. E. Robinson and R. A. Gibbs and K. C. Worley and J. D. Evans, et al. 2006. Insights into social insects from the genome of the honeybee Apis mellifera. Nature **443**:931-949.
Weir, B. S. and C. C. Cockerham. 1984. Estimating F-statistics for the analysis of population structure. Evolution **38**:1358-1370.
Wheeler, W. M. 1910. Ants; their structure, development and behavior. Columbia university press, New York,.
Whitfield, C. W., S. K. Behura, S. H. Berlocher, A. G. Clark, J. S. Johnston, et al. 2006a. Thrice out of Africa: ancient and recent expansions of the honey bee, Apis mellifera. Science **314**:642-645.
Whitfield, C. W., Y. Ben-Shahar, C. Brillet, I. Leoncini, D. Crauser, Y. LeConte, S. Rodriguez-Zas, and G. E. Robinson. 2006b. Genomic dissection of behavioral maturation in the honey bee. Proceedings of the National Academy of Sciences of the United States of America **103**:16068-16075.
Willing, E. M., C. Dreyer, and C. van Oosterhout. 2012. Estimates of genetic differentiation measured by Fst do not necessarily require large sample sizes when using many snp markers. Plos One **7**:e42649.
Wilson, E. O. 1985. The Sociogenesis of Insect Colonies. Science **228**:1489-1495.
Wilson, E. O. and B. Holldobler. 2005. Eusociality: Origin and consequences. Proceedings of the National Academy of Sciences of the United States of America **102**:13367-13371.

Winston, M. L. 1987. The biology of the honey bee. Harvard University Press, Cambridge, Mass.
Winston, M. L. 1992. Killer bees : the Africanized honey bee in the Americas. Harvard University Press, Cambridge, Mass.
Wolf, A., S. Agnihotri, J. Micallef, J. Mukherjee, N. Sabha, R. Cairns, C. Hawkins, and A. Guha. 2011. Hexokinase 2 is a key mediator of aerobic glycolysis and promotes tumor growth in human glioblastoma multiforme. Journal of Experimental Medicine **208**:313-326.
Wong, J. J. L., S. Li, E. K. H. Lim, Y. Wang, C. Wang, et al. 2013. A Cullin1-Based SCF E3 Ubiquitin Ligase Targets the InR/PI3K/TOR Pathway to Regulate Neuronal Pruning. PLoS biology **11**.
Woodard, S. H., G. M. Bloch, M. R. Band, and G. E. Robinson. 2014. Molecular heterochrony and the evolution of sociality in bumblebees (Bombus terrestris). Proceedings of the Royal Society B-Biological Sciences **281**.
Woyke, J., J. Wilde, and M. Phaincharoen. 2012. First evidence of hygienic behaviour in the dwarf honey bee Apis florea. Journal of Apicultural Research **51**:359-361.
Woyke, J., J. Wilde, and C. C. Reddy. 2004. Open-air-nesting honey bees Apis dorsata and Apis laboriosa differ form the cavity-nesting Apis mellifera and Apis cerana in brood hygiene behaviour. Journal of Invertebrate Pathology **86**:1-6.
Wray, M. K., H. R. Mattila, and T. D. Seeley. 2011. Collective personalities in honeybee colonies are linked to colony fitness. Animal Behaviour **81**:559-568.
Xu, S. Z. 2003. Theoretical basis of the Beavis effect. Genetics **165**:2259-2268.
Zasloff, M. 2002. Antimicrobial peptides of multicellular organisms. Nature **415**:389-395.
Zayed, A. and G. E. Robinson. 2012. Understanding the relationship between brain gene expression and social behavior: Lessons from the honey bee. Annual Review of Genetics **46**:591-615.
Zayed, A. and C. W. Whitfield. 2008. A genome-wide signature of positive selection in ancient and recent invasive expansions of the honey bee *Apis mellifera*. Proceedings of the National Academy of Sciences of the United States of America **105**:3421-3426.
Zwarts, L., M. M. Magwire, M. A. Carbone, M. Versteven, L. Herteleer, R. R. H. Anholt, P. Callaerts, and T. F. C. Mackay. 2011. Complex genetic architecture of *Drosophila* aggressive behavior. Proceedings of the National Academy of Sciences of the United States of America **108**:17070-17075.

Appendix A: Statement on contributions

Chapter 1:

This introductory chapter was written by BAH.

Chapter 2:

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

BAH extracted DNA, carried out the majority of the data analysis, and wrote the manuscript. Postdoctoral researcher CFK, with the help of undergrad student JMDL, developed the bioinformatics pipeline to assemble bee genomes, and carried some data analysis. Graduate student DM helped carry out quality control checks on the genomics data. Collaborators ASA and AAO from Saudi Arabia provided samples of pure honey bees from Yemen. AZ provided funding, guidance on the analysis, and helped edit the manuscript.

Chapter 3:

Harpur, B.A., Dey, A., Albert, J.R., Patel, S., Hines, H.M., Hasselmann, M., Packer, L. Zayed, A. Contribution of queen and worker traits to adaptive evolution differs between bumble bees and honey bees. To be submitted.

BAH carried out all of the data analysis, and wrote the manuscript. Prof. MH (Germany) provided samples of *Bombus terrestris* from Europe, Research Associate JA collected samples of *Bombus impatiens* from Ontario, Prof. MHM (USA) provided unpublished genome sequences for *Bombus melanopygus*. Lab manager AD extracted DNA, and AD and SP carried out preliminary bioinformatics assembly of the *B. terrestris* and *B. impatiens* genomes. LP and AZ funded the research. HMH, MH, LP and AZ helped edit the manuscript.

Chapter 4:

Harpur, B.A. Guarna, M. Huxter, E. Kyung-Mee, M., Hoover, S.E., Ibrahim, A. Melathopoulos, A. P., Desai, S., Currie, R.W., Pernal, S.F., Foster, L.J., and Zayed. A. It's good to be clean:

integrative genomics reveals adaptive evolution of the honey bee's (*Apis mellifera*) social immune system.

BAH extracted DNA, and carried out all of the data analysis, and wrote the manuscript. This project was made possible by the availability of bees from colonies that were selected for hygienic behaviour. The bees that BAH sequenced and analyzed were provided by the BeeIPM project which is composed of MG, EH, SEH, AI, AM, APD, RC, SP, and LF. The same team also provided BAH with phenotypic data for the sequenced colonies. AZ provided funding, and helped edit the manuscript.

Chapter 5:

Kadri, S.M*, Harpur, B.A.*, Orsi, R.O., and Zayed, A. A variant reference data set for the Africanized honeybee, *Apis mellifera*. * = joint first authors.

BAH's name appears second on the joint-first-author list. This project represents a close collaboration between BAH and SMK. SMK sampled and measured aggression in over 100 colonies of honey bees from Brazil. This work was funded by SMK's Brazilian supervisor, ROO. SMK then visited our lab and extracted DNA from 30 colonies, which we then sequenced. BAH carried out all of the data analysis, and co-wrote the manuscript with SMK. The paper was written by BAH, SMK, and AZ. AZ funded the genome sequencing and helped edit the manuscript.

Chapter 6:

Harpur, B.A. Kadri, S.M, Orsi, R.O., Whitfield, C.W., and Zayed, A. Defence response in Africanized honey bees (*Apis mellifera* L.) is underpinned by complex patterns of admixture.

BAH carried the majority of the data analysis, and wrote the manuscript. Visiting PhD student SMK sampled bees and extracted DNA. SMK's Brazilian supervisor ROO funded and supervised sampling in Brazil. Sabbatical visitor Prof. CWK (USA) collaborated with BAH to carry out ancestry mapping, with the resulting data being used in this chapter. AZ funded the genome sequencing and helped edit the manuscript.

Chapter 7:

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

BAH carried the data analysis and wrote the manuscript. LS, PM, VS, and KS are YorkU undergrads that helped extract DNA from a very large number of bee samples, in addition to organizing the citizen science program that facilitated sample collection. The SNP assay used for the study was designed by NCC. TER, MA, JL, and BPO provided access to samples and data. I funded the genome sequencing and helped edit the manuscript.

Sincerely,

Brock A. Harpur

Approved by,

Amro Zayed, PhD
Associate Professor of Biology
York Research Chair in Genomics