Understanding the evolutionary origin and ancestral composition of honey bee (*Apis mellifera*) populations.

Kathleen A. Dogantzis

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

November 2022

Abstract

The honey bee, *Apis mellifera*, is arguably the most important managed pollinator globally. Yet despite its economic and ecological importance, there are still several unknowns regarding the species ancestral origin and ancestral complexity. Understanding the genetic composition of native and managed honey bee colonies is imperative for resolving the species life history and elucidating how ancestry may inform management strategies. In this dissertation, I take a deep dive into the evolutionary origins of *Apis mellifera* and learn how ancestral complexity has shaped the composition of contemporary populations. In Chapter two, I settle a long-standing debate about the ancestral origins of the species. I find that *Apis mellifea* diverged out of Western Asia via at least three colonization routes, which resulted in the evolution of at least seven genetically distinct lineages. Interesting, I find that these lineages were able to adapt to their current distribution by repeated selection among a core set of genes. In Chapter three, I take a closer look at the genetic complexity of managed Canadian honey bees by estimating the ancestral composition of colonies using the genomic dataset from Chapter two. I find that patterns of ancestry differ between Canadian provinces, and that admixture correlates strongly with levels of genetic diversity. Interestingly, I find that genomic intervals with elevated levels of admixture segregate non-randomly in the genome and are associated with genes related to parasite and xenobiotic tolerance. Though admixture may bear advantages for managed colonies, admixture among honey bee is not always valued. In Chapter four and five I make use of the ancestral composition of invasive Africanized honey bees to develop assays to identify and track populations. This was achieved using machine learning models to choose the most informative single nucleotide polymorphisms (Chapter 4) and insertion-deletion (Chapter 5) markers that best delineate Africanized genetics from managed European colonies. My research addresses many gaps in our understanding of honey bee origins and ancestral complexity.

Acknowledgments

This has been a long journey, which at times has been a fun, grueling, and complex. First, I would like to thank my family, friends, and husband for all their support over the years. Your words of encouragement, support, and celebration have kept me going. I love you all and thank you for everything,

Thank you to the past and present members of the Zayed lab. It has been a pleasure getting to know you all through my MSc and PhD journey. I would like to thank Brock, Alivia, Dova, Ida, and Mateus for sharing an office and tolerating my office plant collection. I would like to thank Tanushree, Brock, Harsil, Stephen, and Clement for all their help with coding and server problems. Shout out to Tanushree and Arshad for the memories 'iykyk'. I would like to thank all members for whom I shared a lunch hour with every day. This was a highlight and made the long commute to York worth it. I have missed chatting with you all. Thank you all for the fantastic lab parties and celebrations, day trips, conference road trips, lab BBQs, beekeeping, and honey extractions. The lab has been an amazing source of support and I wish you all the best

I would like to thank Dr. Amro Zayed, my supervisor, for whom I have been working with over the last 7 years. Thank you for being an encouraging supervisor and giving me the support and push I needed to advance in my career. Thank you for introducing me to the exciting world of honey bee genetics. I have learnt so much over the years and I am sad to see this chapter of my life come to an end.

Finally, thank you to my present and past committee members and examination committee Dr. Stutchbury, Dr. Bucking, Dr. Shore, Dr. MacDonald, Dr. Clare, and Dr. Rueppell. Thank you to the many people at York who have contributed to a positive experience.

Table of Contents:

List of Figures:

List of tables:

Chapter one: Introduction

Overview:

The western honey bee, *Apis mellifera*, is the most commonly managed bee in the world. The species, which is native to Europe, Africa, and parts of Western Asia has diversified into several morphologically (Ruttner, 1988) and genetically (Chen et al., 2016, Cridland et al., 2017) distinct subspecies. *Apis mellifera* has been translocated worldwide for the apiculture industry, which boomed to 94 million honey bee hives worldwide in 2020; equating to a 13% increase over the last decade (FAOSTAT, 2022). Honey production is estimated to be worth 8.17 billion USD worldwide (Fortune Business Insights, 2022), and in natural habitats, the honey bee is the most frequent floral visitor (13% of visits worldwide) and is responsible for exclusively pollinating 5% of plant species (Hung et al., 2018). Unfortunately, colony losses are frequently reported, and declines have been attributed to several factors, including pathogens, habitat loss, and climate change (Neov et al., 2021). However, understanding the genetic composition of honey bee colonies and how ancestry may influence phenotypes is imperative for mitigating declines, informing selective breeding practices, and improving the success of the beekeeping industry.

Many unknowns related to the population genetics of *Apis mellifera* still exist. For instance, the evolutionary origin of the honey bee is still hotly debated, but this knowledge is important for identifying derived and ancestral genetic mutations. This knowledge can then be extended to tracing the evolution of derived phenotypes and elucidating the genetic underpinnings of economically important traits. Next, admixture between honey bee subspecies is a pervasive event that can have a pronounced effect on the genetic and phenotypic diversity of colonies through new combinations of mutations. As such, categorizing the ancestral background of managed colonies is imperative for informing breeding and management strategies. Finally, admixture among honey bee is not always valued. Thus, the ability to identify and track unwanted introgression, using ancestry informative mutations, is important for maintaining managed colonies

Where did the honey bee come from?

The genus *Apis* is composed of at least 12 extant honey bee species, of which all but one are endemic to Asia. The exception, *Apis mellifera*, is native to Europe, Africa, and Western Asia, where upwards of 30 different subspecies (Ilyasov et al., 2020) have been described based on morphological variation (Ruttner, 1988). Genetic classification of the subspecies has revealed at least five genetically distinct lineages known as the M lineage of Eurasia, the C lineage of Europe, then O and Y lineages of western Asia, and the A lineage of Africa (Cridland et al., 2017). For over four decades, the evolutionary origin of these distinct lineages has remained disputed. Resolving this debate will enhance our ability to identify derived and ancestral mutations. This information is relevant for tracing the evolution of adaptive traits and determining their association with ancestral lineages. Ultimately, this data can be used to determine how locally adapted subspecies contribute to the fitness and diversity of managed colonies.

Historically, this topic was approached using morphological and mitochondrial DNA variation (Ruttner et al., 1978, Garnery et al., 1992). But with advancements in genome sequencing, large nuclear SNP (single nucleotide polymorphism) datasets encompassing several different subspecies have significantly advanced this area of research. The first large population genomic study on *Apis mellifera* used 1136 SNPs and resolved four ancestral lineages (M, C, O, and A) (Whitfield et al., 2006). The study hypothesized, based on phylogenic reconstructions that placed the root of the tree in Africa, that the evolutionary origins of the species was in Africa, and colonization of its current distribution occurred via two to three independent routes (Whitfield et al., 2006) (Fig 1.1). However, reanalysis of this data, which excluded a controversial admixed population, did not support an African origin (Han et al., 2012). Following this work, Wallberg et al. (2014) published an extensive population genomics study with 140 honey bee genomes that reaffirmed the existence of four genetically distinct lineages (M, C, O and A). However, their phylogenetic reconstruction placed the root of the tree between clades of temperately (M, C and O), and tropically (A) located lineages, and proposed the likely origin to be in Asia with all other extant species (Wallberg et al., 2014) (Fig 1.1). Finally, Cridland et al. (2017) most recently combined genomic datasets (Wallberg et al., 2014), including the newly identified Y lineage (Harpur et al., 2014), to provide a comprehensive analysis of *Apis mellifera* evolution. This study confirmed the presence of five genetically

distinct lineages and proposed an evolutionary origin in North Africa or the Middle east, with the A and Y lineages representing the earliest branches based on phylogenic reconstructions (Fig 1.1). Though a lot of progress has been made, there is still considerable contention between the out-of-Africa and out-of-Asia hypotheses that have been proposed.

In Chapter 2 I set out to disentangle the out-of-Africa and out-of-Asia debate using an extensive population genomic dataset comprising 251 individual genomes across 18 putatively identified subspecies. I found that honey bee subspecies cluster into seven genetically distinct groups, including two newly identified lineages, the L lineage composed of *Apis mellifera lamarckii* from Egypt and the U lineage composed of *Apis mellifera unicolor* from Madagascar (Dogantzis et al., 2021). Phylogenetic reconstructions identified two distinct topologies that differed with regard to the placement of the Y lineage and resolved two clades defined by the separation of the M, C and O lineages, from the L, A and U lineages (Dogantzis et al., 2021). I applied a biogeographic reconstruction to both topologies that revealed, with high probability, that the most likely ancestral origin of the species is in Asia (Dogantzis et al., 2021).

Understanding the adaptive radiation of Apis mellifera:

Across the native range of *Apis mellifera*, the species occupies geographically and ecological diverse space. As a response to the different selective pressures within the native range, the species has diversified into morphologically, behaviourally, and genetically distinct subspecies and lineages. For example, some of the most contrasting behavioral differences exist between temperately and tropically adapted bees. Temperate subspecies tend to build larger colonies with more workers and store a substantial quantity of honey to ensure winter survival (Winston, 1992, Seeley, 1983, Schneider et al., 2004). In contrast, tropically located bees, such as those in Africa, establish smaller nests, store less honey, and focus more on pollen collection (Winston, 1992, Seeley, 1983, Schneider et al., 2004). Tropical subspecies also have a greater tendency to swarm due to faster growth rates, abscond from predator attacks or a lack of resources, and exhibit greater pathogen resistance (Winston, 1992, Seeley, 1983, Schneider et al., 2004). Considering the advancements in genome sequencing and the growth of honey bee genome databases, recent population genomic studies have begun to elucidate the molecular underpinning of honey bee adaptions, primarily by examining patterns of positive selection across the genome (Zayed and Whitfield, 2008, Wallberg et al., 2014, Chen et al., 2016).

Positive selection acts on mutations that increase fitness and is therefore consider the primary mechanism for adaptation. Luckily, selection leaves detectable signatures across the genome. When a new mutation, that provides fitness benefits, appears in a population, positive selection will increase the frequency of that mutation over time (Vitti et al., 2013). Because a species occupies diverse space with different selective pressures, positive selection will lead to changes in allele frequency at different loci between populations (Hoban et al., 2016). Thus, when distinct populations are compared, they should show considerable differentiation between selected sites, relative to non-selected sites. Measures of F_{ST} (fixation index) can be used to compare the variance in allele frequency to identify regions consist with signatures of positive selection (Hoban et al., 2016, Vitti et al., 2013). These loci typically show outlier, or extreme measures of F_{ST} differentiation. This method has been shown to be effective for comparing *Apis mellifera* lineages. For example, several studies have noted significant genetic differentiation in coding regions between lineages that potentially reflect the adaptative evolution and phenotypic difference between groups (Zayed and Whitfield, 2008, Wallberg et al., 2014, Chen et al., 2016). However, with the discovery of new *Apis mellifera* lineages (Chapter 2), positive selection has not been investigated across the entire species. This work will increase our understanding of the molecular changes underpinning the species adaptive radiation, and will allow researchers to trace the evolution of traits to lineages or subspecies or origin. This is important for the management of honey bee colonies and understanding how populations may respond to future climatic changes.

In Chapter two, I studied the adaptive radiation of honey bee lineages by identifying patterns of selection using outlier measures of F_{ST}. I found that outlier SNPs were enriched within functional regions of a gene (regulatory and protein coding), but were underrepresented among introns (Dogantzis et al., 2021). Genes associated with outlier SNPs overlapped significantly between lineages, and there were 145 genes that contained at least one outlier SNP across all honey bee lineages (Dogantzis et al., 2021). This finding suggests that a core set of genes may be key to the adaptive response of the species. Genes associated with outlier SNPs were enriched for functions related to development, morphogenesis, and behaviour, and overlapped with previous studies related to colony behaviours (Dogantzis et al., 2021). Finally, I found that worker-biased differentially expressed genes were enriched for genes associated with

outlier SNPs, suggesting the worker caste may have a disproportionate influence on the adaptive radiation of the species.

Tracking admixture in honey bee populations:

Admixture is a process by which genetic information from divergent evolutionary lineages mix as a result of interbreeding (Hamilton and Miller, 2016). Studies suggest that admixture is a common phenomenon, thus, the outcome from such events should be an important consideration for the evolution of a species. For example, admixture has been suggested to be a novel source of genetic variation facilitating genetic rescue and species adaptation (Frankham, 2015). On the other hand, it is often argued that admixture can reduce fitness through the introduction of maladapted genotypes and the breakdown of local adaptation, potentially resulting in population decline (Muhlfeld et al., 2009). Regardless of the outcome, the transfer of genetic information from one lineage into the genome of another lineage can provide important information about a population's demographic and evolutionary history.

When previously isolated divergent lineages undergo gene flow, the exchange of genetic information generates recognizable chromosomal segments called "haplotype blocks" or "ancestry blocks" composed of segments from the respective source populations (Winkler et al., 2010). For example, a single chromosome can be 'stitched' together from multiple long segments of DNA with alternating ancestry associations (Tang et al., 2006). The length and maintenance of each haplotype block is a function of the time since the initial admixture event, and the duration of gene flow between each lineage (Winkler et al., 2010). The recognizable genomic patterns that result from admixture offers an opportunity to track the genetic makeup of genomes via a method called admixture mapping (Darvasi and Shifman, 2005, Winkler et al., 2010, Shriner, 2013). Admixture can be assessed using 'global' admixture mapping, which seeks to estimate the overall ancestry proportions of individuals. Such methods were applied in Chapter 2 to individual honey bees across the native range of *Apis mellifera*. Alternatively, 'local' admixture mapping is focused on mapping the length and location of distinct ancestry blocks throughout the genome of a species (Fig. 1.2). Such data can be used to associate loci with a particular ancestry to inform on ancestral composition and to isolate the effects of ancestry on phenotypes of interest (McKeigue, 2005). In Chapter 3, I apply local ancestry

mapping to Canadian honey bee populations to assess if ancestry has an effect on fitness of commercial colonies.

Identifying patterns of admixture in managed Canadian colonies.

Apiculture continues to be a growing industry but there are still major concerns about managed honey bee colony health. Understanding the ancestral complexity of colonies can provide valuable insight into the diversity of managed bees and how ancestry may influence colony phenotypes. Recent genetic assessments of managed colonies using ancestry informative markers has revealed populations to be highly admixed, primary composed of C and M ancestry, with some contribution from the A-lineage (Harpur et al., 2015). However, these estimates of ancestry proportions were conducted with data from three ancestral lineages, though we know from Chapter 2 that at least seven genetically distinct lineages exist. As such, a reassessment of the ancestral composition of managed bees is needed to ensure we have a full understanding of their genetic complexity. Additionally, we do not know how ancestry is distributed across the genome of the species. Recent work with Africanized bee populations has shown that admixture among colonies is maintained in a non-random pattern and has been linked to behaviours such as reproduction, foraging (Nelson et al., 2017), and colony defense (Harpur et al., 2020). Given the economic and ecological importance of honey bees in North America, understanding the ancestral complexity of colonies can provide valuable insight into the diversity of managed bees and how ancestry may influence colony phenotypes. For example, if mutations from a recently introgressed region offers a fitness advantage, the introgressed segment will likely increase in frequency across the population (Fig. 1.2). We can then use this information to trace its ancestral origins to a lineage or subspecies. This has significant implications for the conservation of native subspecies and the management of commercial honey bees.

In Chapter 3 I analyzed global and local ancestry patterns among 1350 managed Canadian honey bee colonies. Global ancestry revealed that admixture is pervasive among Canadian honey bees, which differed significantly between provinces. Additionally, the proportion of admixture was significantly positively correlated with the level of genetic diversity found within colonies. Finally, I 'mapped' patterns of ancestry across the genome of colonies and found highly admixed regions were associated with genes linked to honey bee health, including parasite and xenobiotic tolerance.

Using genomics and admixture patterns to identify Africanized honey bees:

Africanized honey bees (AHBs) are an invasive population that have rapidly spread throughout South America and into the southern United States (Kono and Kohn, 2015, Rangel et al., 2016, Porrini et al., 2019). AHB populations are often regarded as undesirable for beekeeping due to their high colony aggression, tendency to swarm, and frequent absconding. There is continued concern in the apiculture industry that Africanized bee populations will continue to expand their range, either through accidental movement, or through climate change. Luckily it is known that AHB are genetically admixed, composed primary of A-lineage ancestry, which makes them genetically discernable from managed European (C and M lineage) colonies. Given that the genetically distinct lineages are highly differentiated (F_{ST}), it is possible to use ancestry informative markers to track and identify AHBs using their genetic composition. Current methods for identifying and tracking Africanized bees primarily rely on SNP assays, and while effective, were developed with data from only three ancestral lineages (Chapman et al., 2017, Chapman et al., 2015). As I demonstrated in Chapter 2, honey bees have seven genetically distinct linages, and thus the current methods for identifying AHBs may not correctly represent the genetic composition of managed or Africanized bee colonies. Additionally, current assays rely on ancestry proportion thresholds for classification, but these measures may be confounded by incomplete ancestral representation. Accurate and up-to-date detection assays are needed to track the movement of Africanized honey bees and prevent the spread of Africanized bee genetics to regions currently free of invasive populations. This is especially important for regions that breed and export queen bees, where contamination of stocks would render these sources unusable.

Here, in Chapter 4 and 5 I developed two improved genetic assays for identifying Africanized honey bees. These assays use genetic information from all seven lineages, which provides a comprehensive genetic background to base classification. Molecular markers for the assays were chosen using machine learning tools to improve on the informativeness of markers. In Chapter 4 I found that 80 SNP markers, when combined with machine learning classification, can effectively identify Africanized bees as well as populations that share close African ancestry. In chapter 5, I designed a lower cost PCR alternative to SNP genotyping using insertion deletion

markers. This novel source of variation is effective for quickly screening bees for evidence of Africanization and could easily be used in combination with other methods.

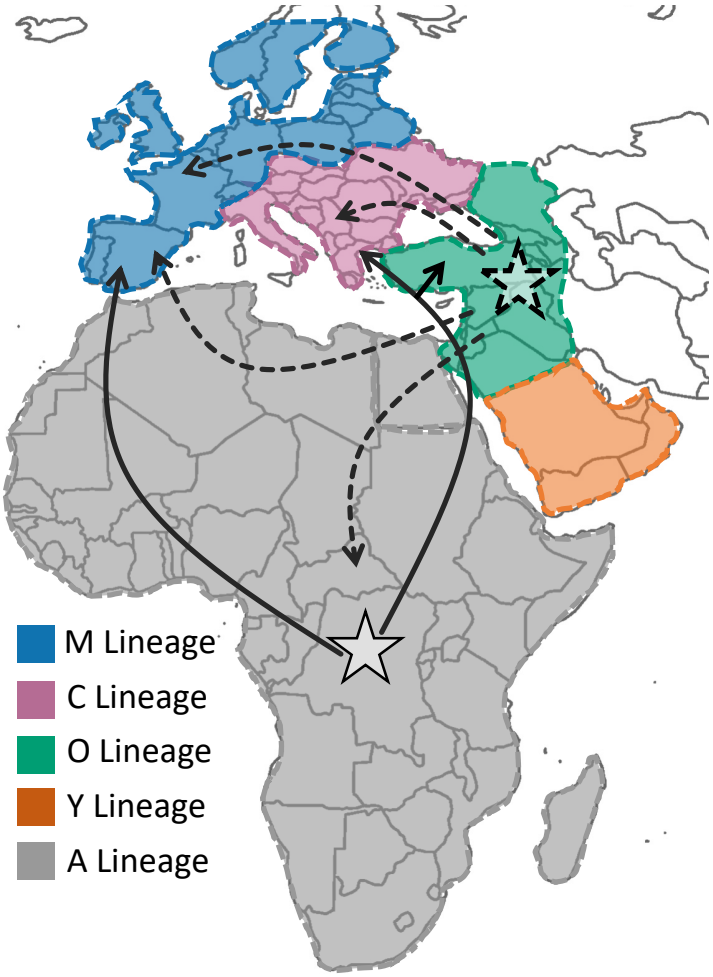

Fig. 1.1: Hypotheses regarding the evolution of *Apis mellifera*. There are at least five genetically distinct honey bee lineages distributed throughout Africa, Europe, and West Asia: The A lineage of Africa, the O and Y lineages in Western Asia, the M lineage of Eurasia, and the C lineage of Europe. There are two leading hypotheses about the colonization of the species. First, the out-of-Africa hypothesis (solid lines) posits that the species originated in Africa and colonized its current distribution via a western and eastern expansion route. The western expansion involved the M lineage, which colonized Eastern Europe via the Iberian Peninsula (Whitfield et al., 2006, Cridland et al., 2017), while the eastern route includes the diversification of the C and O lineage into Eastern Europe and Western Asia (Whitfield et al., 2006, Cridland et al., 2017). Second, the out-of-Asia hypothesis (Han et al., 2012, Wallberg et al., 2014) (dashed lines) suggests that the ancestral origin in likely is Asia and the species expanded to its current distribution via a combination of three expansion routes. The first route expanded into Africa (A lineage), while the second route expanded into Western Europe (C lineage) after dividing from

Western Asia (O lineage). The final expansion route involves the colonization of the M lineage, where it has been suggested the lineage could have colonized Eastern Europe via a northern expansion through Europe (Garnery et al., 1992), or an expansion from Africa via the Iberian Peninsula (Cridland et al., 2017, Ruttner et al., 1978). Settling the out-of-Africa and out-of-Asia debate has important implications for tracing the evolution of derived mutations and their association with adaptively significant phenotypes.

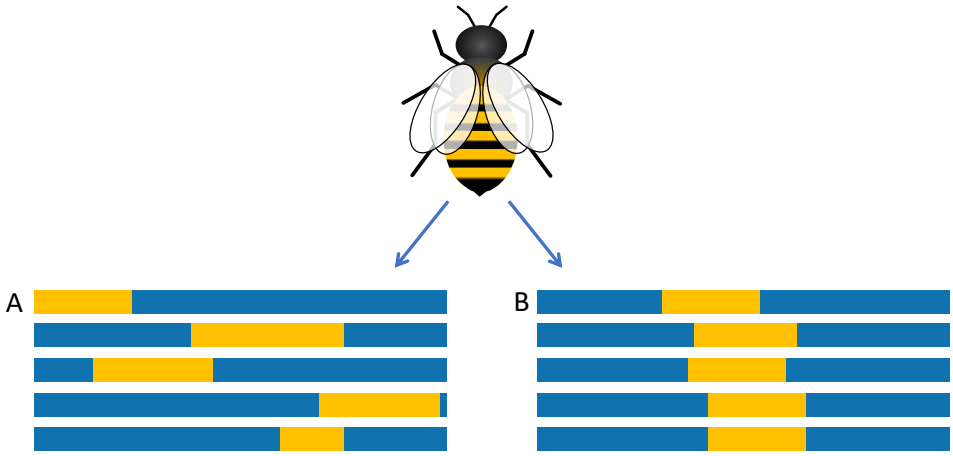

Fig. 1.2: Potential patterns of introgression in *Apis mellifera.* When admixture occurs between genetically distinct groups, the resulting offspring contain a mosaic of ancestry from the source populations (yellow introgressed into blue). Over time, if the introgressed regions have a neutral effect, ancestry will be 'randomly' distributed throughout the genome (A). Alternatively, introgressed regions that contain mutations that provide a fitness advantage will increase in frequency throughout the population (B). If regions of ancestral enrichment are detected among admixed populations, mutations can be traced back to the source subspecies or lineage (yellow), which can have important implications for conservation and breeding practices.

Chapter two: Thrice out of Asia and the adaptive radiation of the western honey bee

Kathleen A. Dogantzis, Tanushree Tiwari, Ida M. Conflitti, Alivia Dey, Harland M. Patch, Elliud M. Muli, Lionel Garnery, Charles W. Whitfield, Eckart Stolle, Abdulaziz S. Alqarni, Michael H Allsopp, Amro Zayed[1]

Summary:

The origin of the western honey bee *Apis mellifera* has been intensely debated. Addressing this knowledge gap is essential for understanding the evolution and genetics of one of the world's most important pollinators. By analyzing 251 genomes from 18 native subspecies, we found support for an Asian origin of honey bees with at least three expansions leading to African and European lineages. Remarkably, the adaptive radiation of honey bees involved selection on a few genomic 'hot-spots'. We discovered 145 genes with independent signatures of selection across all bee lineages, and these genes were highly associated with worker traits. Our results indicate that a core set of genes associated with worker and colony traits facilitated the adaptive radiation of honey bees across their vast distribution.

Introduction:

The genus *Apis* is composed of twelve extant species that form three distinct groups: giant honey bees, dwarf honey bees, and cavity nesting honey bees (Arias and Sheppard, 2005, Lo et al., 2010, Raffiudin and Crozier, 2007). All but one of the extant *Apis* species are endemic to Asia. The exception, *Apis mellifera*, is native to Europe, Africa, and Western Asia. Given the wide geographic spread of the species, *Apis mellifera* has diversified into several subspecies (Dogantzis and Zayed, 2019, Ruttner, 1988), of which there are approximately ten subspecies in Africa, nine in Asia, and potentially as many as 13 subspecies in Europe (Ilyasov et al., 2020). Each subspecies can be genetically and morphologically classified into at least five distinct evolutionary lineages; the M lineage of Eurasia, the C lineage of Europe, the O and Y lineages of Western Asia, and the A lineage of Africa (Dogantzis and Zayed, 2019, Ruttner, 1988). Though

[1]This published manuscript has been reprinted with the permission of its co-authors and publisher from the original manuscript: Dogantzis, Kathleen A., et al. "Thrice out of Asia and the adaptive radiation of the western honey bee." *Science Advances* 7.49 (2021): eabj2151.

it is reasonably accepted that the genus emerged in Asia, the ancestral origin and adaptive radiation of contemporary *Apis mellifera* lineages and subspecies remain unresolved.

Early fossil records from the Oligocene (34-23 MYA) place ancestral *Apis* within Europe, followed by a migration of the genus during the late Oligocene or during the Miocene (23-5.5 MYA) (Kotthoff et al., 2013, Ruttner, 1988). It has been hypothesized that ancestral *Apis* migrated from Europe into Asia, where it diversified into the three modern lineages of *Apis,* including *Apis mellifera* of the cavity nesting bees (Ruttner, 1988, Kotthoff et al., 2013). Alternatively, it has also been proposed that ancestral *Apis* remained widespread throughout Europe and Asia, where near the end of the Miocene, *Apis* colonized Africa via the Iberian Peninsula leading to the origin of *Apis mellifera*, while the remaining extent *Apis* species descended from ancestors in Asia (Kotthoff et al., 2013). These different hypotheses about the biogeography and diversification of the *Apis* genus are important for understanding the two competing hypotheses regarding the origin of *Apis mellifera* in Asia (Ruttner et al., 1978, Cridland et al., 2017, Wallberg et al., 2014, Garnery et al., 1992) or Africa (Whitfield et al., 2006, Wilson, 1971). The expansion from Asia is predicted to have occurred via two northwestern routes into Europe, one consisting of the M lineage and another consisting of the C and O lineages, and a colonization route extending into Africa (A lineage) (Garnery et al., 1992). However, it has also been proposed that the route into Africa could have acted as a western expansion of the M lineage into Europe (Ruttner et al., 1978). Comparatively, the species expansion from Africa is predicted to have occurred via two or three expansion routes including the colonization of the M lineage via the Iberian Peninsula, and then the C and O lineage through Northeast Africa and Western Asia (Whitfield et al., 2006).

Resolving the ancestral origin and evolutionary expansion of *Apis mellifera* will enhance our understanding and ability to identify derived and ancestral genetic mutations. This is especially relevant for tracing the evolution of novel phenotypes, and for discerning how locally adapted subspecies may contribute to the fitness and diversity of managed colonies (Harpur et al., 2012). Recent genomic studies of *Apis mellifera* have shown that with the addition of new subspecies and enhanced datasets (Cridland et al., 2017, Han et al., 2012) estimates of evolutionary origin can change. As such, the increased representation of samples from Africa and Western Asia – two historically under sampled regions (Dogantzis and Zayed, 2019,

Grozinger and Zayed, 2020) – may be the key to disentangling the out-of-Africa and out-of-Asia debate.

Here, we used an extensive population genomic dataset comprised of 251 individuals and 18 putatively identified subspecies from Europe (N=4), Africa (N=8), and Asia (N=6), to elucidate the evolutionary and adaptive origins of *Apis mellifera*. These samples were collected throughout the native distribution of *Apis mellifera*, with a concentrated effort on filling population and subspecies gaps within Africa and Western Asia. In this study, we aimed to evaluate the population structure of subspecies and determine their lineage classification, define evolutionary relationships using phylogenetic reconstruction, and use biogeography to estimate the most likely ancestral range of the species. Finally, we assessed patterns of selection among lineages to identify and categorize the genomic regions associated with the adaptative radiation of the species.

Methods:

Data processing

Methods for DNA extraction, genome alignment, and SNP detection are described in detail in the Supplemental Material. In brief, we generated a dataset of 251 individual *Apis mellifera* samples, of which 160 samples are newly sequenced, with the remaining samples downloaded from the Sequence Read Archive (SRA) in addition to 15 *Apis cerana* genomes (Chen et al., 2018) (Data S1). Sequence reads were trimmed of adapters and low quality bases uses Trimmomatic v0.36 (Bolger et al., 2014). Trimmed reads were aligned to the honey bee reference genome (Elsik et al., 2014) using NextGenMap aligner v0.4.12 (Sedlazeck et al., 2013) and duplicate reads were marked with Picard v2.1.0 (https://broadinstitute.github.io/picard/). SNPs were identified and filtered using GATK v3.7 (Poplin et al., 2017, Van der Auwera et al., 2013).

Population structure

The program ADMIXTURE v 1.3.0 (Alexander and Lange, 2011) was used to estimate ancestry proportions and population structure among the 251 *Apis mellifera* samples. To reduce the effects of uninformative and low frequency variants (Linck and Battey, 2019), 1M variants were selected among a pool of SNPs pruned for bi-allelic loci with a MAF >0.05. To account for

linkage disequilibrium (LD), SNPs were further pruned (38,493) for a minimum distance of 5,000 bp – a distance where LD typically decays to background level in the honey be genome (Wallberg et al., 2014). Both analyses were run with predicted K values 1-18 and used the 10X cross-validation procedure to estimate the optimal number of ancestral groups (K). A principal component analysis (PCA) was generated to examine the genetic relatedness and clustering patterns among *Apis mellifera* samples using the SNPrelate (Zheng et al., 2012) package in R (R Core Team, 2013) with all available SNP markers. Finally, we performed a hierarchal structure analysis using identity by state with the SNPrelate (Zheng et al., 2012) package in R (R Core Team, 2013) to qualitatively determine lineage assignment using the snpgdsCutTree function..

Phylogenetic reconstruction

We produced several phylogenetic trees using a SNP dataset pruned of ambiguous loci, as implemented by RaxML v8.2.12 (Stamatakis, 2014), and loci with low coverage (<0.8) in *Apis cerana*. Trees were constructed using three different datasets 1) SNPs located genome wide (2,126,091), 2) SNPs within coding regions (276,602), and 3) randomly selected SNPs located among intrageneric and intergenic regions (276,602). Neighbor-joining trees where constructed with all three SNP sets using allele-sharing distance with Adegenet (Jombart, 2008) and Ape (Paradis et al., 2004) in R (R Core Team, 2013). Confidence levels for bipartitions in the neighbor-joining tree were calculated using 100 bootstrap replicates as implemented in Ape (Paradis et al., 2004). Maximum-likelihood trees were constructed using SNP sets two and three using the program RaxML v8.2.12 (Stamatakis, 2014). Trees were constructed with the gamma model of rate heterogeneity (ASC_GTRGAMMA) with the Lewis ascertainment bias correction. A 100 rapid bootstrap analysis and search for the best-scoring tree was performed in a single program run. Finally, the program TreeMix v1.13 (Pickrell and Pritchard, 2012) was used to produced maximum-likelihood trees using dataset one. TreeMix infers population splits using genome-wide allele frequency data at the population level. The program assumes biallelic loci with no missing data, thus, missing genotypes were imputed using Beagle v5.0 (Browning and Browning, 2007) and only biallelic loci were retained (1,884,783 genomic SNPs). The analysis was performed with samples grouped into their respective lineages and previously determined subspecies grouping (Data S1; Supplementary Text).

Divergence time estimation

Divergence times were estimated with PAML 4.9 (Yang, 2007) using both resolved phylogenetic topologies. We used the putative coding regions of non-overlapping genes, with high (>0.9) sequence coverage among the outgroup (*Apis cerana*), concatenated into one supergene. For ease of phylogenetic reconstruction, we did not include the *A. m. monticola* cluster, which is clearly established within the A lineage. We also did not include *A. m. pomonella* or *A. m. syriaca* due to high levels of admixture. First, the substitution rate was estimated using BASEML in PAML 4.9 (Yang, 2007). We used the REV (GTR) model with the strict molecular clock and calibrated the divergence time between *Apis mellifera* and *Apis cerana* at 7.5MYA (Arias and Sheppard, 2005, Wallberg et al., 2014, Chen et al., 2016) using a time unit of 100MYA (@0.075). The calculated substitution rate per unit of time was used to calculate the rgene_gamma variable using the shape α and scale parameter β equations as per the PAML manual (Yang, 2007). Divergence times were estimated following the two-step approximate likelihood calculation with the MCMCtree package in PAML 4.9 (Yang, 2007). We used the REV (GTR) model with independent clock rates and root age was bound between >.06<.09 (Arias and Sheppard, 2005, Wallberg et al., 2014, Chen et al., 2016) using a time unit of 100MYA. The process was run for 10,000 samples, sampling every 10 iterations, after a burn-in of 50,000, for a total of 150,000 iterations.

Ancestral biogeography reconstruction

To infer the biogeographic history of *Apis mellifera*, we estimated the most probable model of geographic range expansion on the divergence time tree of both topologies using the R package BioGeoBEARS (Matzke, 2013a, Matzke, 2013b). BioGeoBEARS employs three different models of geographic range evolution: Extinction Cladogenesis (DEC) (Ree and Smith, 2008, Ree, 2005), a likelihood version of Dispersal-Vicariance Analysis (DIVA) (Ronquist, 1997), and a likelihood version of BayArea (BAYAREA) (Landis et al., 2013). Additionally, BioGeoBEARS can incorporate a jump dispersal or founder event speciation into the model, generating three additional models DIVA+J, DEC+J, and BAYAREA+J. We defined three biogeographic areas based on the current *Apis mellifera* distribution: Europe (E), Africa (F), and Asia (A). We tested all six biogeographic models provided by BioGeoBEARS and used the

Akaike Information Criterion (AIC) and the Log of the likelihood scores (LnL) to compare models and determine the best fit to the phylogeny.

Genetic diversity, genetic differentiation, and demography

We calculated several diversity and demographic statistics among lineage and subspecies groupings (Data S1). Nucleotide diversity (π) was calculated in 500bp sliding windows with a 250bp step size using VCFtools v0.1.17 (Danecek et al., 2011). Segregating sites (S) were calculated by counting the number of polymorphic loci, and singletons ($S_{\text{singletons}}$) were calculated by counting the number of sites with only one copy of an allele. To estimate theta (θ_w), we used the equation $\hat{\theta_w} = S/an$, where S is the number of segregating sites, and an is the harmonic number of n-1, where n is the number of chromosomes. To obtain the per-base-pair estimate of θ_w, $\hat{\theta_w}$ was divided by the total number loci that had sufficient coverage (≥0.8) across the entire genome. To estimate the effective population size (N_e) we used Watterson's theta estimator (Watterson, 1975) $\theta_w = 3Ne\mu$ (3 is used because *Apis mellifera* is haplodiploid), where μ is the mutation rate. N_e was calculated using two estimates of mutation rate 5.27×10^{-9} (Wallberg et al., 2014) and 3×10^{-9} (Liu et al., 2016). We calculated linkage disequilibrium (LD) as a measure of the squared correlation coefficient between variants (r^2). LD was measured within 5000bp windows using SNP variants that had ≥0.8 coverage and a MAF > 0.05 using VCFtools (Danecek et al., 2011). The pairwise F_{ST} matrix was calculated using Weir and Cockerham's weighted F_{ST} statistic with VCFtools v 0.1.17 (Danecek et al., 2011). Finally, the program ADMIXTOOLS (Patterson et al., 2012) was used to calculate outgroup f_3 statistics (Raghavan et al., 2015) which can be used to quantify the genetic distance between populations relative to an outgroup, *Apis cerana*, where higher values imply longer shared evolutionary time or greater share genetic drift.

Detecting and annotating loci under selection

We identified patterns of positive selection by means of outlier differentiation using pairwise measures of Weir and Cockerham's weighted F_{ST} statistic with VCFtools v 0.1.17 (Danecek et al., 2011). The genome-wide distribution of F_{ST} was measured between each pairwise lineage, and loci consistently within the top 0.95 quantile across each pairwise distribution were considered unique measures of genome outliers. This analysis was performed

on markers that had a MAF > 0.05 in at least one lineage and had ≥0.8 coverage across all lineages; we used 3,183,349 SNPs for this analysis. We exclude highly admixed samples (Data S1) and divided the European and Asian M lineage subspecies into separate populations as they are likely experiencing disparate selective pressures (Chen et al., 2016). We used the program SNPeff v 4.3t (Cingolani et al., 2012) to annotate SNPs at the gene and functional category level, including exons, introns and promoter regions, which were defined as the sequence 1000bp upstream of the start codon of a gene (Molodtsova et al., 2014) and excluded regions that overlapped with neighboring genes. Additionally, SNPeff v 4.3t was used to predict mutation effects on genes, such as amino acid changes. Finally, Gene Ontology (GO) enrichment was conducted with DAVID v 6.8 (Huang et al., 2009) using *Drosophila melanogaster* orthologs (Elsik et al., 2016). Gene ontology functional annotation clusters with an enrichment score ≥ 1.3, and gene ontology terms with $p < 0.05$ after Benjamini-Hochberg correction were of interest.

Resampling simulations

To ensure the overlap of outliers among genes was not due to chance, we used gene and SNP resampling to simulate the overlap of genes across lineages. Gene resampling was achieved by randomly selecting, from the background set of genes (12,916), the corresponding number of genes associated with outlier loci within each lineage (Data S2; Table S9). We then made pairwise comparisons between lineages to calculate the number of genes that overlapped between the randomly resampled lists. Simulations were repeated for 1000 iterations to generate a null distribution. We then carried out an additional analysis which takes into account gene size. Large genes may be expected to have more outlier SNPs per lineage, thus leading to greater overlap among lineages. In our dataset, we observed a total of 36,678 outlier SNPs in genes across the seven lineages studied (Table S9). We simulated the null distribution of overlap (i.e., genes with different outliers in more than one lineage) by randomly generating 36,678 unique outlier SNPs, corresponding to the same number of outliers per lineage as observed in our dataset (Table S9), across an equivalent coding genome as studied herein (i.e., same number of genes with identical sizes as predicted in the honey bee genome) (Data S2). In our simulations, the probability of observing an outlier locus within a gene scales linearly with the gene's size. After each simulation, we computed the average number of lineages with different outlier SNPs in the same gene, and the number of genes with unique outlier SNPs in all seven lineages. We ran this

simulation for 100 iterations, then compared the null distortions of these two parameters to our observed data. There were clear significant differences between the null distribution of the 100 iterations and our observed data that additional iterations were not necessary.

Differential gene expression

We identified caste differentially expressed genes between 96 hr-old queen and workers (Ashby et al., 2016). Reads were downloaded from the Sequence Read Archive (SRA) (BioProject PRJNA260604) and trimmed of adapters and low quantity bases (<20) using Trimmomatic v0.36 (Bolger et al., 2014). Trimmed reads were then aligned to the honey bee reference genome (Elsik et al., 2014) using multi-sample 2-pass mapping with STAR v2.7 (Dobin et al., 2013). Using the aligned RNA-seq data, a matrix of unnormalized read counts was constructed for annotated gene regions using featureCounts in Subread V 2.0.1 (Liao et al., 2014). Finally, DESeq2 (Love et al., 2014) in R (R Core Team, 2013) was used to identify differentially expressed genes. We used the matrix constructed with featureCounts as the countData and set the condition to caste phenotypes (queen or worker). We analyzed 10,000 genes and determined them to be differentially expressed between castes if they passed the following thresholds: fold change of ≥1.5, FDR <0.05 after applied standard error, and gene-level read counts ≥10 per individual in the upregulated caste. In addition, we used a protein atlas, which examined protein expression across 26 tissues in queen and worker honey bees (Chan et al., 2013). Harpur et al (Harpur et al., 2014) had previously generated mutually exclusive queen biased and worker biased proteins from this resource (1,582 proteins), which we used for our analysis.

Results:

Sequencing and variant detection

We curated a genomic dataset of 251 individual *Apis mellifera* samples representing 18 putative subspecies; of which 14 representative groups were retained (Supplementary Text; Data S1). The dataset is composed of several previously published samples (Harpur et al., 2014, Fuller et al., 2015, Haddad et al., 2015, Chen et al., 2016, Wallberg et al., 2017), and 160 newly sequences individuals that were collected across temporally and spatially diverse ranges to broaden the representation of populations. The average coverage depth for newly sequenced

samples was 66.1 ± 17.7x. After filtering raw variants, we retained a working dataset of 11.8 million SNPs.

Population structure and admixture

Using several population structures analyses, we investigated patterns of clustering and admixture among honey bee samples (Fig. 2.1B; Fig. S2.1-S2.3). The cross-validation of the ADMIXTURE analyses revealed the optimal number of genetic clusters to be eight (K=8). We confirmed the presence of previously identified honey bee evolutionary lineages in Africa (A lineage), Asia (Y and O lineage), Europe (C lineage), and Eurasia (M lineage) (Fig. 2.1A-B). Interestingly, two newly sequenced subspecies formed unique genetic clusters warranting classification as distinct lineages; *A. m. lamarckii* of Egypt (L lineage) and *A. m. unicolor* of Madagascar (U lineage) (Fig. 2.1A-B). At K=8, *A. m. intermissa* (North Africa), a highly admixture subspecies (27%), is identified as an independent genetic cluster (Fig. S2.1). However, this cluster is not consistent at other K values (Fig. S2.1) and likely does not represent a true lineage, but rather an artifact of high genetic admixture. As such, seven genetically distinct groups more accurately represent the number of biologically relevant lineages.

We detected additional patterns of admixture among subspecies, notably within *A. m syriaca*, which is composed primarily of O lineage ancestry (76.8%), and is admixed with the A (12.6%), Y (4.4%), and L (4.4%) lineages (Fig. S2.1). As noted in previous studies, *A. m. syriaca* is located within a contact zone between Africa (A and L) and Asia (O and Y) (Fig. 2.1A), which is the likely contributor to high levels of hybridization (Cridland et al., 2017, Wallberg et al., 2014). Introgression of the C lineage into the L and M lineages was detected (Fig. S2.1). These admixture patterns are not unexpected given the close geographic proximity of these lineages, and M and C lineage admixture has been documented extensively (Henriques et al., 2018a, Henriques et al., 2018b, Parejo et al., 2016, Muñoz et al., 2017, Muñoz et al., 2015, Pinto et al., 2014). Finally, we detect varying levels of admixture in samples from Kyrgyzstan (*A. m. pomonella*), with some samples displaying high levels of C lineage ancestry (Fig. S2.1) likely from imported European colonies used for commercial beekeeping.

Phylogenetic and biogeographic reconstruction

We constructed several phylogenetic trees using three different combinations of SNPs to determine the evolutionary relationships among *Apis mellifera* samples (Fig. 2.1C-D; Fig. S2.4-S2.7). Our analyses resolved two topologies that differed slightly with respect to the placement of the Y lineage (Fig. 2.1C-D). Subspecies are consistently clustered into previously recognized lineages and two definitive clades defined by the separation of the M, C, and O lineages from the L, A, and U lineages. Divergence dating based on nuclear coding sequences suggests that *Apis mellifera* lineages may have begun to diverge as early as c. 6 MYA (Fig. S2.8).

To predict the most likely ancestral range of the species and major clades, we applied a biogeographic reconstruction to both resolved topologies (Table S2.1). The ancestral range for the most recent common ancestor of the species was predicted to be in Asia with 64.5-71.4% probability, while probabilities for an African or European ancestral range were much lower (<6%) (Fig. 2.2; Fig. S2.9; Table S2.2). This finding complements a recent independent study that predicted the ancestral range for cavity nesting bees to be in Southeast Asia (Ji, 2021). The ancestral range of the most recent common ancestor of the M, C, and O clade was predicted to be in Asia with a 70% probability, while the ancestral range of the L, A, and U clade varied (70% Asia or 100% Africa) depending on the topology (Fig. 2.2; Fig. S9; Table S2.2). Microgeographic classification of subspecies likewise places the ancestral range of the species in Western Asia (Fig. S2.10). The use of an outgroup in biogeographic reconstructions is recommended to prevent the reconstruction of wide ancestral ranges (Lamm and Redelings, 2009). We used *A. cerana* as an outgroup for this analysis, but choosing a different cavity nesting bee would not have changed the biogeographic reconstruction, as the ancestor of all cavity nesting bees is predicted to be in Asia (Ji, 2021).

Contemporary patterns of diversity and demography

Recent demographic events, notably the last glacial period where temperate populations were constrained and the A lineage expanded to its population maxima (Wallberg et al., 2014), have likely shaped patterns of genetic diversity and effective population size among contemporary populations. For instance, genetic diversity is highest among the A lineage ($\boldsymbol{\pi}$ = 3.54E^{-03}, θ_w = 1.01E^{-02}), relative to European (average $\boldsymbol{\pi}$ =1.48E^{-03}, θ_w = 1.84E^{-03}), and Asian (average $\boldsymbol{\pi}$ =1.84E^{-03}, θ_w = 1.83E^{-03}) lineages (Table S2.3). Likewise, estimates of N_e were considerably larger for the A lineage (~640,000), relative to European or Asian lineages

(~116,000) (Table S2.3); as previously documented (Wallberg et al., 2014). Interestingly, the U lineage of Madagascar had relatively high levels of diversity ($\boldsymbol{\pi}$ =2.33E^{-03}, θ_w =1.70E^{-03}) and effective population size (~107,000) relative to European and some Asian subspecies, but considerably less than its parent lineage (A) of mainland Africa (Table S2.7). Additionally, we find that linkage disequilibrium (LD) is lowest and decays the quickest among A lineage samples, consistent with high estimates of N_e. Comparatively, LD was high among European and the U lineage, consistent with low N_e and historical population bottlenecks (Fig. S2.11-S2.12).

Measures of pairwise F_{ST} between lineages were high (0.528 ± 0.149) (Table S2.5), while estimates between subspecies within the same lineage were low (0.163 ± 0.073) (Table S2.6). Outgroup f_3 statistics, which are less sensitive to lineage specific drift (Harris and DeGiorgio, 2017), were used to quantify the genetic distance between lineages relative to an outgroup (*Apis cerana*). The analysis identified high f_3 between the O and C lineages, affirming a longer shared evolutionary history (Fig. S2.13). Pairwise f_3 values between the A, L, and U lineages were also high, suggesting a close evolutionary relationship between African lineages (Fig. S2.13). Finally, we observed high f_3 values between the M and C lineage (Fig. S2.13), despite having high genetic differentiation (F_{ST} = 0.66) (Table S2.5), suggesting a more recent common ancestor, but rapid divergence between the lineages. Overall, the relationships identified by f_3 statistics are congruent with the evolutionary relationships suggested by the phylogenetic tree and structure analyses.

Patterns of selection across the genome

We studied the adaptive radiation of honey bee lineages by identifying patterns of positive selection inferred from pairwise estimates of outlier genetic differentiation (F_{ST}) at SNP loci (Table S2.7; Data S2). Here, we focused our analyses on lineage-specific outliers defined as mutations that show extreme values of F_{ST} (highest 5%) in all pairwise comparisons involving a focal lineage. We excluded samples with high levels of recent admixture and separately analyzed European and Asian M lineage subspecies given they likely experience disparate selective pressures (Chen et al., 2016). Additionally, the A lineage was excluded from analyses due to a relatively low number of lineage-specific outlier markers (Table S2.7; Supplementary Text). While we and others (Harpur et al., 2014, Wallberg et al., 2014) have discovered a substantive number of outlier mutations when comparing the A group to any other honey bee lineage, there

were very few loci with outlier F_{ST} in all six pairwise comparisons involving the A lineage. This may be the result of demographic effects, shared evolutionary relationships, or local adaptation among the A lineage subspecies (Supplemental Text).

Outlier SNPs were enriched within protein coding (χ^2, $p < 6.43E^{-11}$) and putative promoter regions (χ^2, $p < 3.12E^{-02}$) of most lineages (Table S8). Comparatively, introns were deficient of outlier SNPs, which was significant among four lineages (χ^2, $p < 2.79E^{-02}$) (Table S8). Though each outlier SNP is distinct to a lineage, we discovered that their distribution was concentrated among a relatively small set of genomic hotspots as evident by a significant overlap of genes with outlier SNPs between pairwise lineage comparisons (781 ± 263 genes; $p < 3.57E^{-50}$) (Fig. 2.3; Table S2.9). In addition, 145 genes contained at least one outlier SNP across all honey bee lineages. We used gene and SNP resampling simulations to confirm that the overlap of genes and the distribution of outlier SNPs was not due to chance or gene length. Gene resampling indicated that the observed overlap between pairwise lineages was, on average, 140% greater relative to the simulations (Table S2.10). Likewise, randomly resampling outlier SNPs within genes, which corrects for the possibility that larger genes are more likely to have outlier loci because they tend to have more SNPs, indicated that the average number of lineages that overlapped across genes was significantly lower (29.6%, Mann-Whitney test, $p < 7.17E^{-48}$), relative to the simulations. This suggests that outliers in our dataset are concentrated in a smaller set of genes than expected by chance.

Genes associated with the adaptive radiation of Apis mellifera lineages

Loci underlying adaptive divergence were enriched for gene ontology (GO) terms (Data S3) related to morphogenesis and development of tissues and organs, including wing development, sensory organs, eye, muscle, and appendages, as well as development during the larvae and pupae stages. Notably, gene GB48653 was found to be under selection among all lineages and is orthologous to homothorax (FBgn0001235) in *D. melanogaster*, which is important for antennal development, appendage patterning, and cell division of the eye field (Corsetti and Azpiazu, 2013). We also found enrichment of GO terms related to neuron development, as well as receptor and signaling activity. There was also evidence for enrichment of genes related to learning and memory, as well as behaviour, including olfactory, aggression, and mating. Gene GB42603 (*NLG3*) was found to be under selection among all lineages, and it is

posited that changes in gene regulation may affect memory and learning tasks (Biswas et al., 2008). Additionally, among the 145 genes under selection across all lineages, there were several genes that have been found to be associated with colony behaviour traits including colony defense (Harpur et al., 2020), immunity (Mondet et al., 2020b, Lattorff et al., 2015, Zanni et al., 2017, Amiri et al., 2020), and the production of honey and royal jelly (Wragg et al., 2016) (Data S2). Intriguingly, we find several genes overlap among colony traits. For example, three genes (GB54493, GB51389, GB40915) overlapped between *Varroa* response and colony defense. Among genes associated with royal jelly production, two, GB52279 and GB43012, were found to overlap with colony defense and *Varroa* infection respectively. Finally, GB42671, which was associated with honey production was also associated with Varroa infection.

To further understand the phenotypic context of local adaptation of western honey bees, we evaluated the association between genes with outlier SNPs and queen and worker castes. We used published datasets to determine differences in the expression of genes in larvae (Ashby et al., 2016), and proteins in adults (Harpur et al., 2014, Chan et al., 2013) to define genes associated with queen and worker traits (i.e. queen-biased vs. worker-biased expression). We discovered, relative to expected values, that genes associated with local adaptation of honey bee lineages were significantly elevated in the worker caste (χ^2, $p < 7.16E^{-04}$), but often significantly underrepresented in the queen caste (χ^2, $p < 3.99E^{-07}$ larvae; N.S. among adults) (Fig. 2.4A). Likewise, the proportion of genes with outlier SNPs was significantly higher in worker biased genes, relative to queen biased genes (χ^2, $p < 3.11E^{-02}$) (Fig. 2.4B). Finally, genes (N=145) with independent signs of adaptive evolution across all lineages were overwhelmingly more likely to be worker-biased (N=64) than queen-biased (N=0) (Fisher's Exact Test, $p = 1.35E^{-13}$).

Discussion:

Deciphering the ancestral origin of contemporary *Apis mellifera* lineages is a major unsolved question with implications of understanding the evolution of this model eusocial species. There are currently two hypotheses that place the origin of *Apis mellifera* in either Africa (Whitfield et al., 2006, Wilson, 1971) or Asia (Ruttner et al., 1978, Cridland et al., 2017, Wallberg et al., 2014, Garnery et al., 1992). Our analysis supports the hypothesis of an Asian origin of *Apis mellifera*. *Apis mellifera* likely diverged from other cavity nesting bees in Southeast Asia (Ji, 2021), and colonized its current distribution from Western Asia. We find that

some of the phylogenetic reconstructions emphasize an ancestral divide between West Asian lineages (Y and O), which are resolved in separate clades. While phylogenies based on protein coding regions, resolve the Y lineage as the most basal branch. Both topologies indicate that the ancestor of contemporary *A. mellifera* lineages was most likely in Asia. These findings are more congruent with the hypothesis that all extent *Apis* species descended from a common ancestor in Asia, rather than *Apis mellifera* originating independently in Africa.

Once diverged from other cavity nesting bees, our biogeographic reconstruction provides several hypotheses for how *Apis mellifera* expanded to its current distribution. The M lineage, which forms a distinct evolutionary branch, likely colonized Europe via an independent Northern route. Though previous studies hypothesized that the M lineage expanded from Africa (Whitfield et al., 2006, Wilson, 1971, Cridland et al., 2017), which was supported by shared genetic similarity with the A lineage, these patterns are like the result of recent nuclear (Whitfield et al., 2006, Han et al., 2012, Chávez-Galarza et al., 2015) and mitochondrial (Chávez-Galarza et al., 2017, Chávez-Galarza et al., 2015, Cánovas et al., 2008, Pinto et al., 2013, Boardman et al., 2020a) introgression among geographically proximate populations. The C lineage colonized Southern Europe, which may have once been the southern limit of the M lineage, after splitting from a shared common ancestor with the O lineage in Western Asia. Finally, colonization of Africa potentially occurred via two dispersal events from Asia. The L lineage forms its own genetically distinct nuclear cluster and shares mitochondrial origins with some populations from desert Africa (El-Niweiri and Moritz, 2008, Hailu et al., 2020) and Western Asia (Franck et al., 2001), notably the Y lineage. In contrast, the A lineage, which comprises the remainder of Africa, possesses distinct nuclear and mitochondrial (Franck et al., 2001) variants, and is ancestral to the U lineage.

The adaptive radiation of *Apis mellifera* lineages is marked by repeated selection among several genomic hotpots. Notably, there is a significant overlap of genes with outlier loci among pairwise lineages, but also shared among all lineages. Repeated selection among genes has been shown to be common among taxa that descend from a common ancestor and are then exposed to similar environments (Conte et al., 2012). However, recent studies with *Apis cerana* have also uncovered patterns of gene reuse, which may be linked to radiation among diverse habitats (Ji et al., 2020). In our study, we find that genomic hot spots are prevalent among genes related to development, morphogenesis, and behaviour. This pattern of selection is consistent with the

extensive morphological and behavioural adaptations that have occurred among the species, especially between tropically and temperately adapted bees (Winston et al., 1983).

Finally, we find that genes with outlier loci are disproportionately related to the worker caste in the form of worker-biased genes and worker related phenotypes. Evidence for selection among the worker caste has been demonstrated previously (Harpur et al., 2014), and is hypothesized to be related to the eusocial nature of honey bee colonies (Harpur et al., 2017, Dogantzis et al., 2018). Honey bee colonies are composed of several thousand workers who contribute to important colony tasks such a brood rearing and resource provisioning. Though workers do not lay eggs, natural selection may indirectly select for worker phenotypes to optimize colony fitness (Harpur et al., 2014). This is relevant given the diverse colony adaptations that have arisen in response to environmental variables, including traits directly related to colony defense (Harpur et al., 2020), immunity (Mondet et al., 2020b, Lattorff et al., 2015, Zanni et al., 2017, Amiri et al., 2020), and the production of honey and royal jelly (Wragg et al., 2016). Interestingly, we also find signs of pleiotropy between worker phenotypes, indicating that not only is repeated selection among a common set of genes prevalent across *Apis mellifera* lineages, but the same genes are involved with increasing fitness among several different phenotypes.

In conclusion, we have presented compelling evidence that *Apis mellifera* emerged in Asia with the remainder of extant honey bees, but then expanded into its current distribution via Western Asia. This expansion event is marked by at least three independent colonization routes that gave rise to seven genetically distinct lineages. Modern populations of *Apis mellifera* maintain high genetic diversity, which has allowed the species to adapt to diverse ecological environments through repeated selection among a common set of genes. These genes are more often than not related to worker phenotypes, supporting that the worker caste is related to the success of the adaptative radiation of the species.

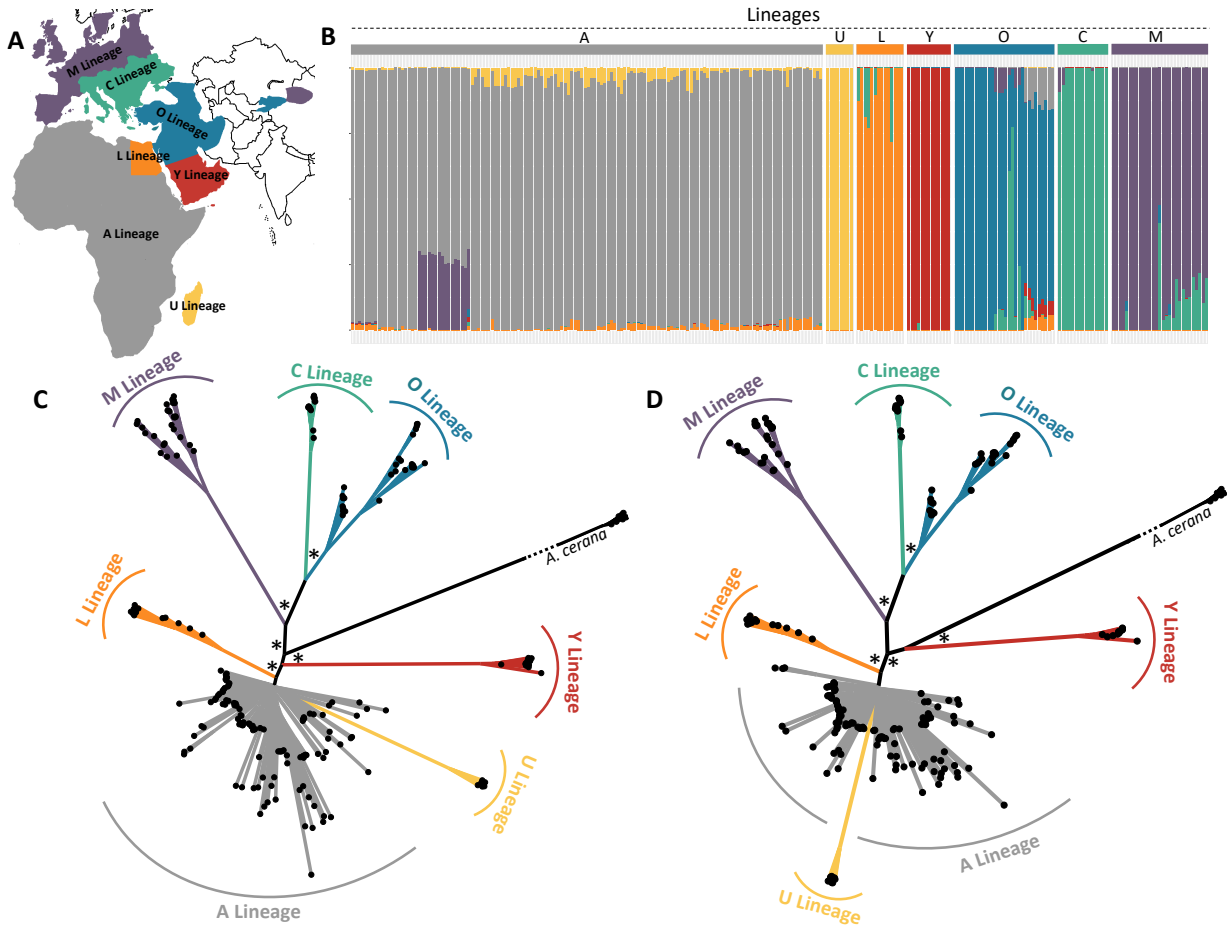

Fig. 2.1: Population structure and phylogenetic reconstruction of *Apis mellifera*. (**A**) Map of the native distribution of the seven genetically distinct lineages. (**B**) Patterns of ancestry and population structure identified with ADMIXTURE when K=7. Vertical bars represent individual bees and coloured segments represent the proportion of ancestry to the different clusters. (**C**) Evolutionary relationships among *Apis mellifera* samples reconstructed with a neighbor-joining tree using SNPs located genome wide. Asterisks represent node support of 100%. (**D**) Evolutionary relationships among *Apis mellifera* samples constructed with a neighbor-joining tree using SNPs located within protein coding regions. Asterisks represent node support of 100%. Node support and maximum likelihood phylogenetic trees can be found in the Supplementary Materials.

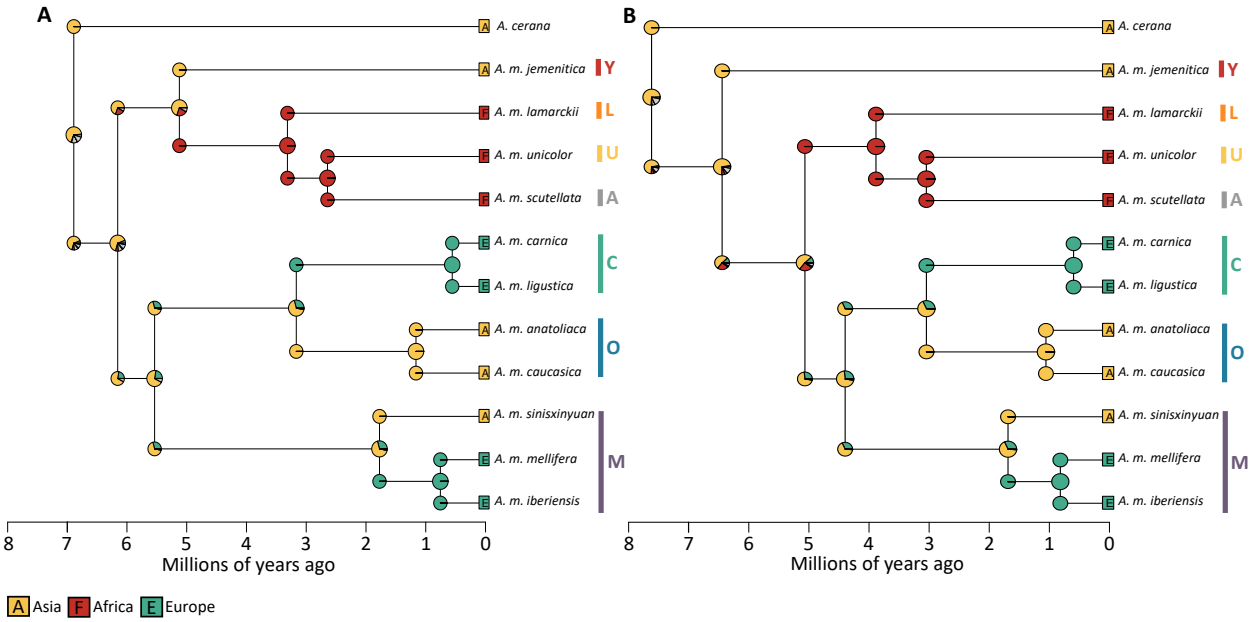

Fig. 2.2: Ancestral biogeographic range reconstruction of *Apis mellifera* using two resolved topologies. The current geographic range of subspecies is indicated at branch tips by letters A: Asia, F: Africa and E: Europe. Coloured bars to the right of the trees indicates the lineage association of the subspecies. Pie charts at nodes indicate the marginal maximum likelihood probabilities for the estimated ancestral range. The ancestral range is predicted to be in Asia with an estimated probability of 64-73%. Part (**A**) represents the topology reconstructed using SNPs located throughout the genome, while (**B**) represents the topology reconstructed with SNPs located in protein coding regions. Node probabilities and the biogeographic reconstruction of the *Apis* genus can be found in the Supplementary Materials.

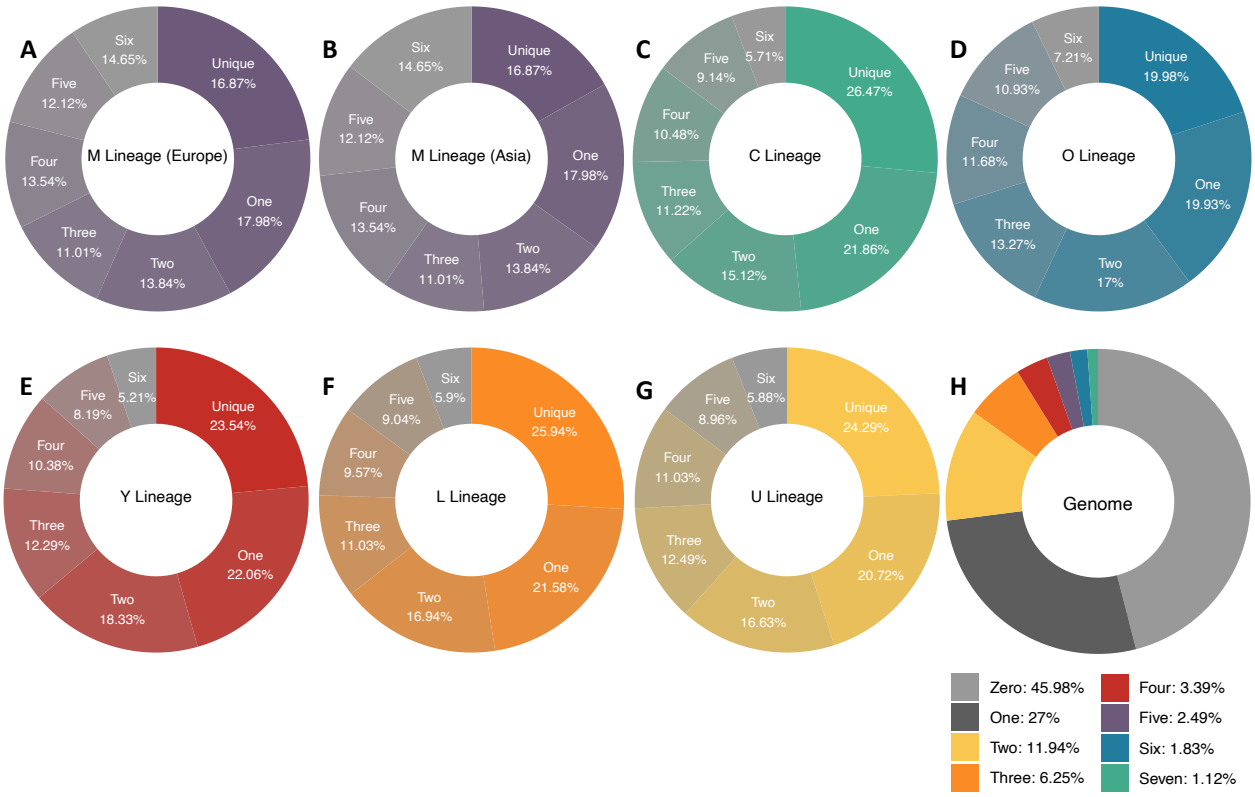

Fig. 2.3: Proportion of genes that overlap among lineages and across the *Apis mellifera* genome. Plots (**A** to **G)** illustrate the proportion of genes that are either unique to a lineage or share signs of selection among 1-6 other lineages. Plot (**H)** illustrates the proportion of genes across the genome that possess outlier SNPs among no lineages to all seven lineages.

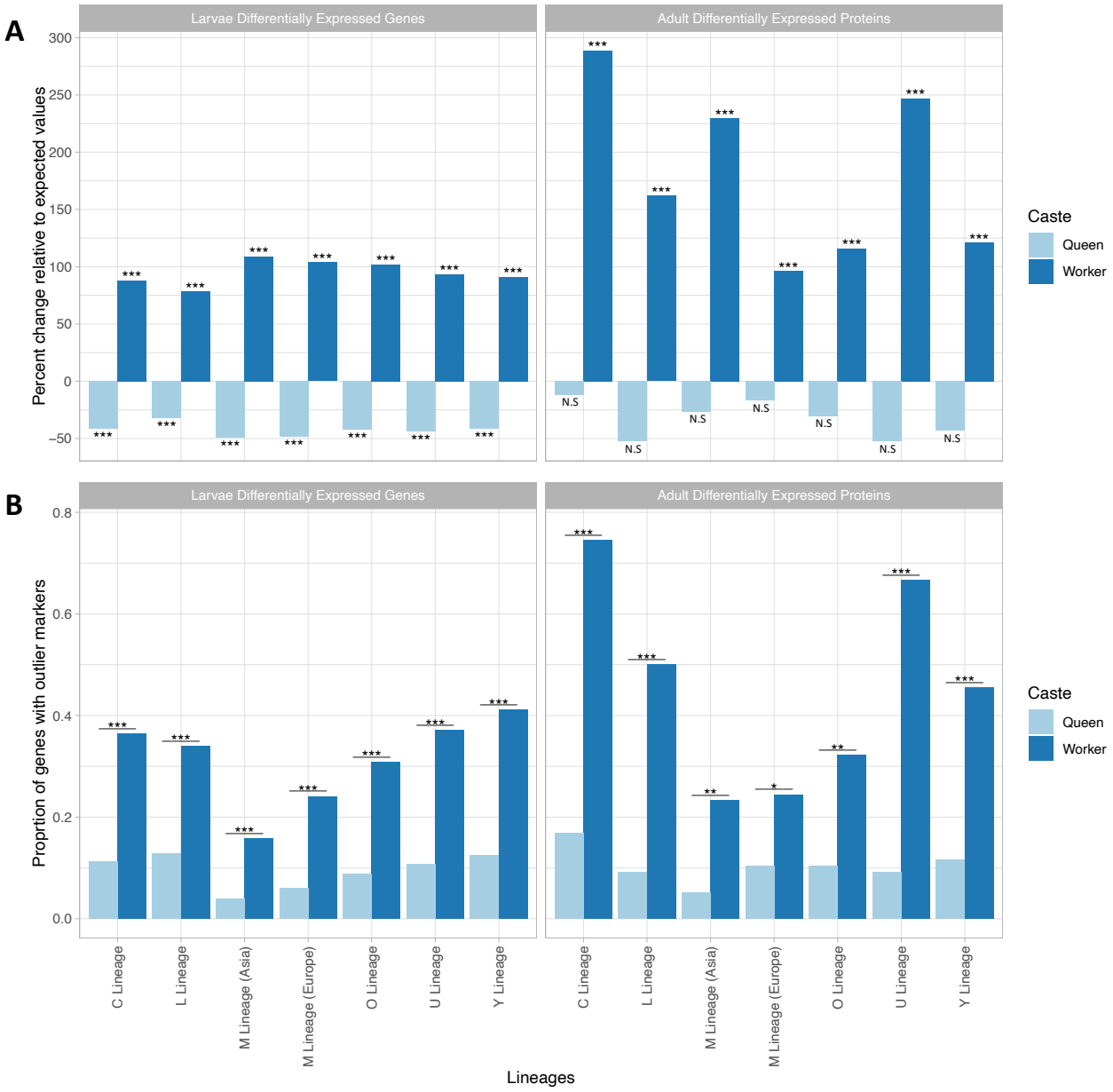

Fig. 2.4: The association of genes with outlier loci among the worker and queen caste. Figure (**A**) illustrations the percent change in the observed number of genes with outlier SNPs among queen and worker biased genes for larvae and adults, relative to expected values. For example, a negative change, suggests an underrepresentation of genes, a positive change represents an enrichment of genes, and no change suggests no difference from expected values. Asterisks represent the degree of significance of the change between observed and expected values (* < 0.05, ** < 0.01, *** < 0.001), while N.S is not statistically significant. Figure (**B**)

illustrates the proportion of genes with outlier SNPs among worker and queen caste biased genes for larvae and adults for each lineage. Asterisks represent the degree of significant difference between the proportions (* < 0.05, ** < 0.01, *** < 0.001).

Supplementary Methods:

Sample collection, DNA extraction, and sequencing

We composed a population genomic dataset of 251 individual *Apis mellifera* samples, of which 160 samples are newly sequenced, with the remaining samples downloaded from the Sequence Read Archive (SRA) (Data S1), including 15 *Apis cerana* genomes (Chen et al., 2018). New samples were collected across several years and wide geographic areas to ensure we were sampling one bee per colony. New samples were received as either whole bees, partial bees, or already extracted DNA. We extracted DNA from bee tissues using a Mag-Bind® Blood & Tissue DNA HDQ 96 Kit (Omega Bio-tek Inc., USA) optimised for the KingFisher™ Flex Purification System (Thermo Fisher Scientific Inc., USA). For tissue lysis, either half or whole bee heads or thoraces were flash frozen in liquid nitrogen and finely ground using a pestle. We then added 350µl Tissue Lysis Buffer, 20µl Proteinase K, and heated samples overnight at 50°C. After processing with the KingFisher System, samples were eluted using the Mag-Bind Kit Elution Solution (Thermo Fisher Scientific Inc., USA) to a final volume ranging from 70–85µl. DNA was quantified using NanoDrop™ 2000 Spectrophotometer (Thermo Fisher Scientific Inc., USA). DNA quality was assessed on 1.0% agarose gel using electrophoresis. Genome sequencing of new samples was carried out at The Centre for Applied Genomics (The Hospital for Sick Children, Ontario, Canada) using Illumina HiSeqX to generate 150bp pair-end reads.

Sequence alignment, variant discovery, and filtration

Sequencing reads of *Apis mellifera* and *Apis cerana* were trimmed of adapters and low quantity bases (<20) using Trimmomatic v0.36 (Bolger et al., 2014). Trimmed reads were retained for downstream assembly if >50bps and >35bps in length from 100-150bp or 50bp read length Illumina data respectively. Reads were then aligned to the *Apis mellifera* reference genome (Amel 4.5)(Elsik et al., 2014) using default parameters of NextGenMap aligner v0.4.12 (Sedlazeck et al., 2013). Midway through our analysis, a new assembly of the honey bee genome was released (Wallberg et al., 2019); we elected to continue to use Amel 4.5 to be able to utilize the large transcriptomic and functional genomic datasets generated using Amel 4.5. The average percentage of mapped reads across samples was 96.6%, and we are confident that there was little information loss. BAM files were sorted using SAMtools v1.3.1 (Li et al., 2009) and duplicate reads were marked using Picard v2.1.0 (https://broadinstitute.github.io/picard/). Base quality

scores were recalibrated using GATK v3.7 BaseRecalibrator (Van der Auwera et al., 2013) using previously identified variants as reference (Harpur et al., 2014, Harpur et al., 2019). Variants were identified by constructing intermediate gVCF files for each genome using HaplotypeCaller GATK v3.7 (Poplin et al., 2017, Van der Auwera et al., 2013), and then samples were aggregated using GenotypeGVCFs GATK v3.7. Variants were filtered using VariantRecalibrator GATK v3.7 using previously identified variants as reference (Harpur et al., 2014, Harpur et al., 2019) in addition to the following hard filter thresholds: MQ < 40.0, QD < 5.0, FS > 11.0, MQRankSum -2.0 < x > 2.0, and ReadPosRankSum -2.0 < x > 2.0. In addition, we excluded variants located within five base pairs of an indel and within five base pairs of areas with low complexity (Harpur et al., 2019), we excluded loci with greater than 20% missing data, and excluded variants from the unmapped scaffolds using GATK v3.7.

Supplementary Text:

Sample inclusion and population classification

The ADMIXTURE and PCA analyses revealed seven genetically distinct clusters. The hierarchical structure analysis grouped subspecies into their respective lineages, except four samples that were labeled as outliers. *A. m. intermissa* samples were grouped into their own cluster, mirroring results of the admixture analysis, likely the result of high admixture. The A lineage was divided into two clusters, one composed of putatively known *A. m. monticola* samples, and the second composed of remaining African subspecies and *A. m. scutellata*. Because there is little definition of subspecies groupings within the A lineage cluster, the two aforementioned A lineage groups were maintained for subspecies level analyses.

Based on results from the preceding structure analyses, all seven genetically distinct lineages, and the following subspecies classifications were retained: *A. m. mellifera* (M lineage), *A. m. iberiensis* (M lineage), *A. m. sinisxinyuan* (M lineage), *A. m. carnica* (C lineage), *A. m. ligustica* (C lineage), *A. m. anatoliaca* (O lineage), *A. m. caucasica* (O lineage), *A. m. syriaca* (O lineage), *A. m. pomonella* (O lineage), *A. m. jemenitica* also known as *A. m. yemenitica* (Y lineage), *A. m. lamarckii* (L lineage), *A. m. scutellata* (A lineage), *A. m. monticola* (A lineage), and *A. m. unicolor* (U lineage). The following samples were excluded from analyses: samples that ~50% of ancestry attribute to an unexpected lineage (n=3), *A. m. intermissa* (n=16), genetic outliers identified in the hierarchical structure analysis (n=4), and haploid samples (N=8) (Data S1). For

analyses that are lineage focused, we retained samples that had ≥90% assignment to one ancestral lineage to reduce the effects of admixture. At the subspecies level, ancestry exceptions were made for uniformly admixed subspecies; *A. m. syriaca*, *A. m. pomonella*, and *A. m. sinisxinyuan*. Additionally, we exclude individual samples that, while grouped with the expected lineage, did not clearly cluster with a putatively identified subspecies or population in the phylogenetic analyses, which made classification ambiguous (Data S1).

Identification of outlier loci

The A lineage has the least amount (2 SNPs) of outlier SNPs among all lineages. To account for the large sample size in the A lineage, we repeated the outlier detection analysis with ten representative samples but were only able to identify ~80 outlier SNPs. It is hypothesized that low outlier SNPs within the A lineage may be attributed to demographic effects resulting in substantially lower than expected F_{ST} values (Meirmans and Hedrick, 2011). Additionally, considering the longer evolutionary time between the A, L, and U lineages, outlier markers may be 'shared' among lineages and thus not represented as unique to the A lineage. Finally, it is also postulated that local adaption among the subspecies that make up the A lineage has disrupted or prevented congruent shifts in allele frequency that would result in high F_{ST} at the lineage level.

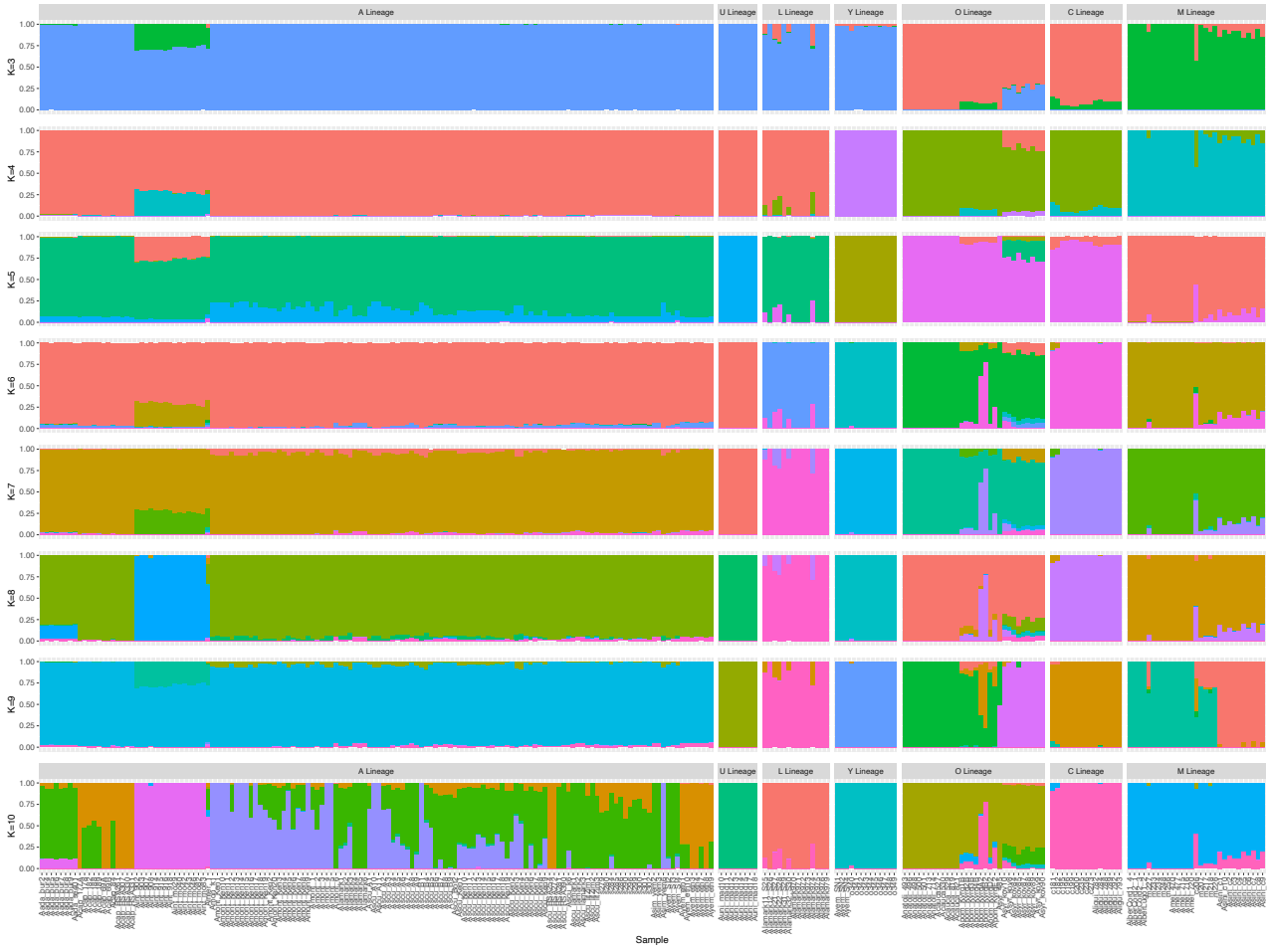

Fig. S2.1: Genetic clustering of *Apis mellifera* samples using nuclear SNPs. Patterns of ancestry for all *Apis mellifera* samples as estimated by the program ADMIXTURE using 1M loci that were selected among a pool of SNPs pruned for bi-allelic loci with a MAF >0.05. Vertical bars represent individual bees and coloured segments represent the proportion of ancestry estimated to K=3-10 genetic clusters. When the ADMIXTURE analysis is conducted with K predictive clusters 3-6, there is evidence for common ancestry between the O and C lineage, and between the A, L, U, and Y lineages. When K is increased above 7, we begin to see the division of subspecies into separate clusters.

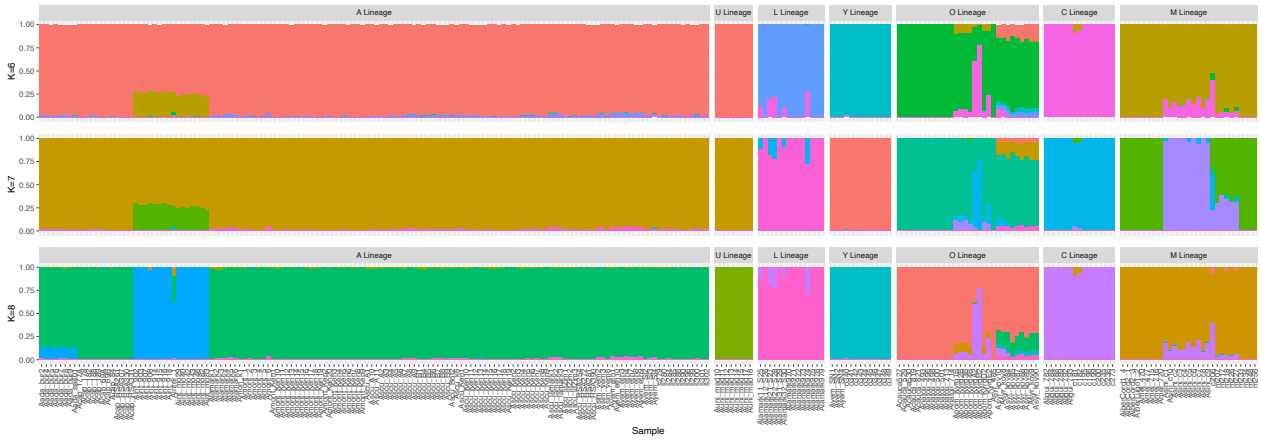

Fig. S2.2: Genetic clustering of *Apis mellifera* samples using unlinked nuclear SNPs.
Patterns of ancestry for all *Apis mellifera* samples as estimated by the program ADMIXTURE using unlinked SNPs (38,493). Vertical bars represent individual bees and coloured segments represent the proportion of ancestry estimated to K=6-8 genetic clusters.

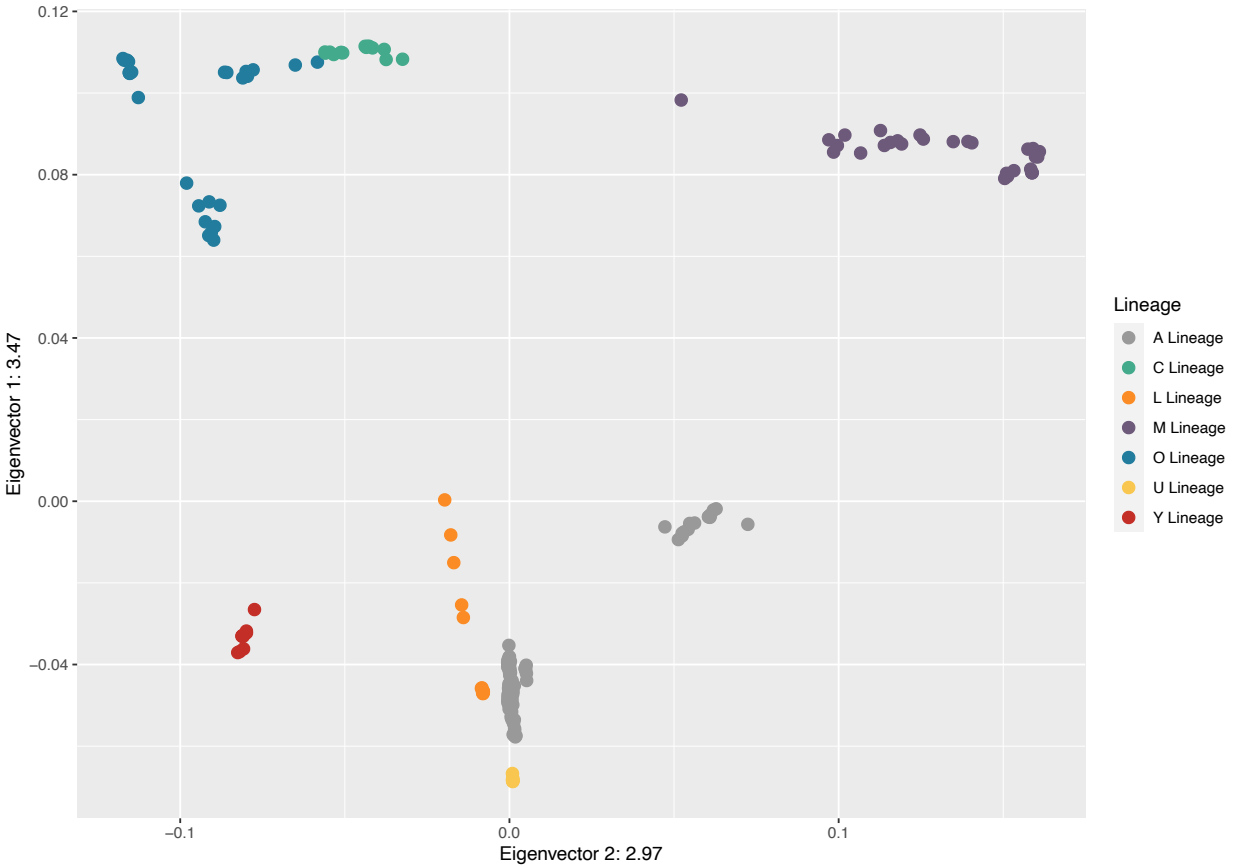

Fig. S2.3: Principal component analysis of diploid *Apis mellifera* samples using nuclear SNPs. The first two principal components grouped *Apis mellifera* samples into seven distinct clusters representative of the M, C, O, Y, L, A, and U lineages. The PCA analysis also clearly illustrates the effects of admixture on the genetic relatedness among lineages. For example, *A. m. intermissa*, which is highly admixed with M lineage ancestry, is situated between the A and M lineage clusters. Similarly, the putatively identified O lineage *A. m. pomonella* samples from Kyrgyzstan, which have high C lineage ancestry, are adjacent to the C lineage cluster. These relationships correlate with clustering patterns detected in the ADMIXTURE analysis.

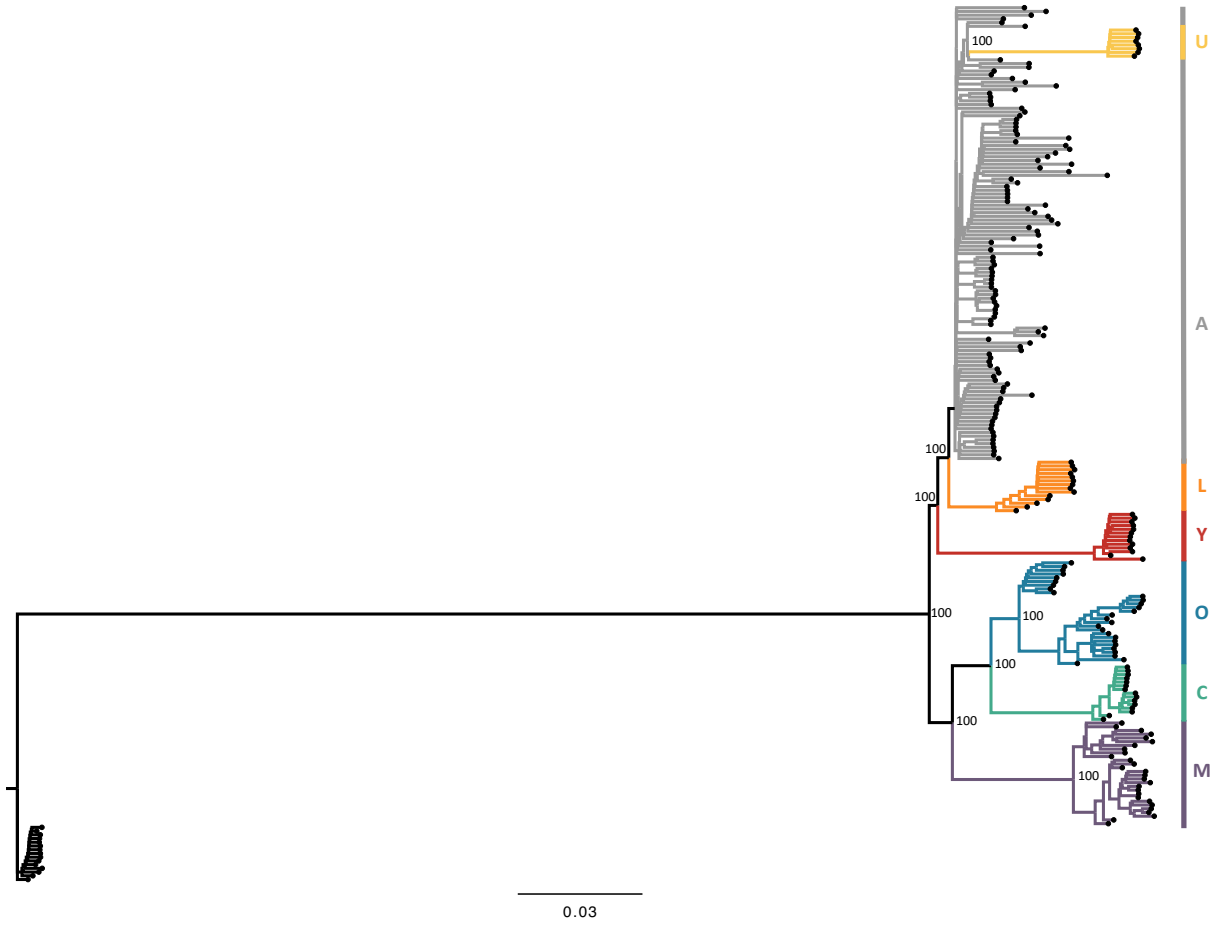

Fig. S2.4: Neighbor-joining phylogeny of *Apis mellifera* samples using SNPs located genome wide. The entire SNP dataset was pruned of ambiguous loci, as implemented by RAxML v8.2.12 (*63*), and loci with low coverage (<0.8) in *Apis cerana*. The neighbor-joining tree were constructed using SNPs located genome wide (2,126,091) using allele-sharing distance and was rooted with *Apis cerena*. Major nodes are labeled by bootstrap support.

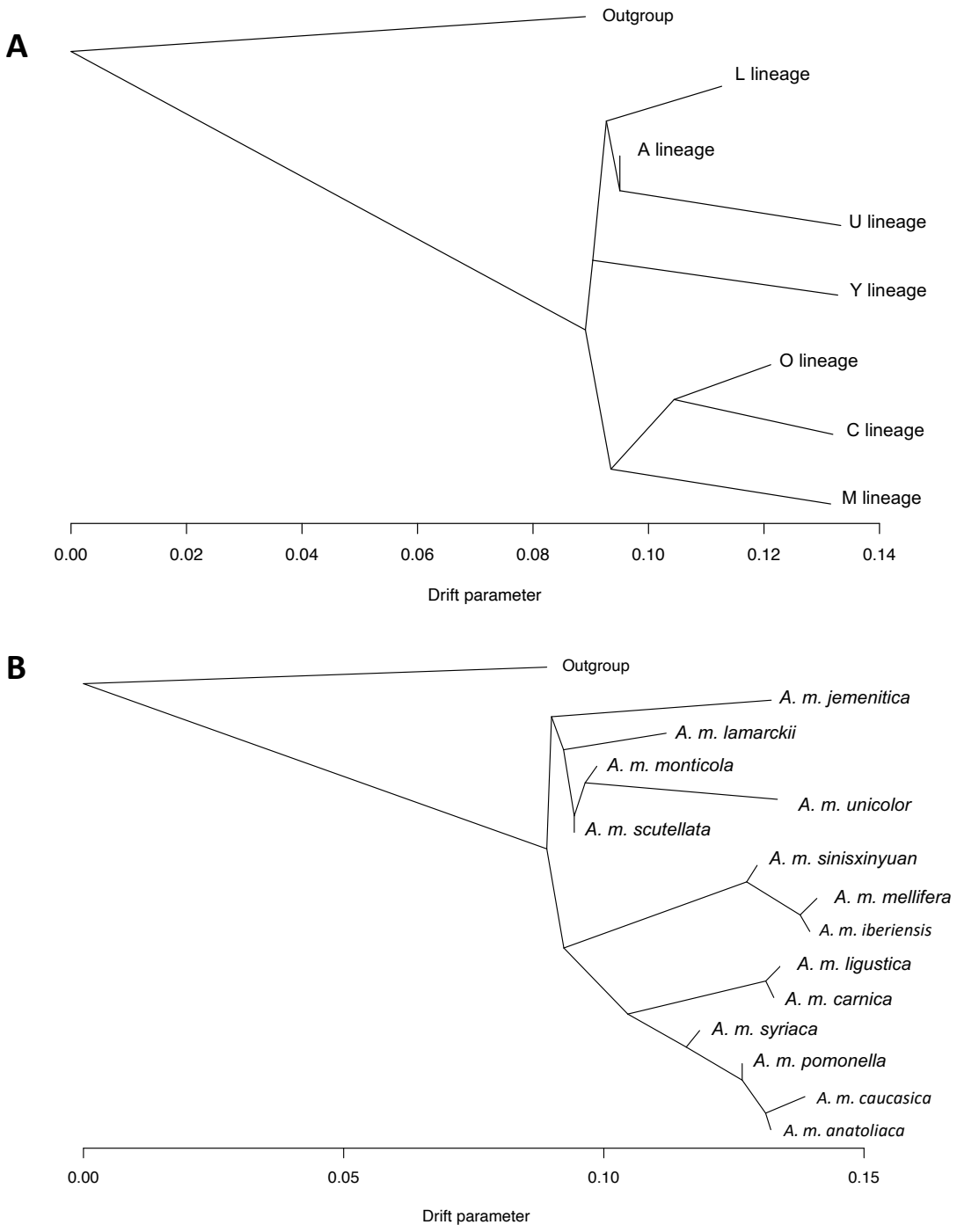

Fig. S2.5: Maximum-likelihood Phylogeny of *Apis mellifera* samples using SNPs located genome wide. The program TreeMix v1.13 (*66*) was used to produced maximum-likelihood trees using SNPs located genome wide that were further pruned for biallelic loci (1,884,783). The analysis was performed with SNPs formatted as allele frequencies with samples grouped into their respective lineages (**A**) and subspecies (**B**) grouping and was rooted using *Apis cerana*.

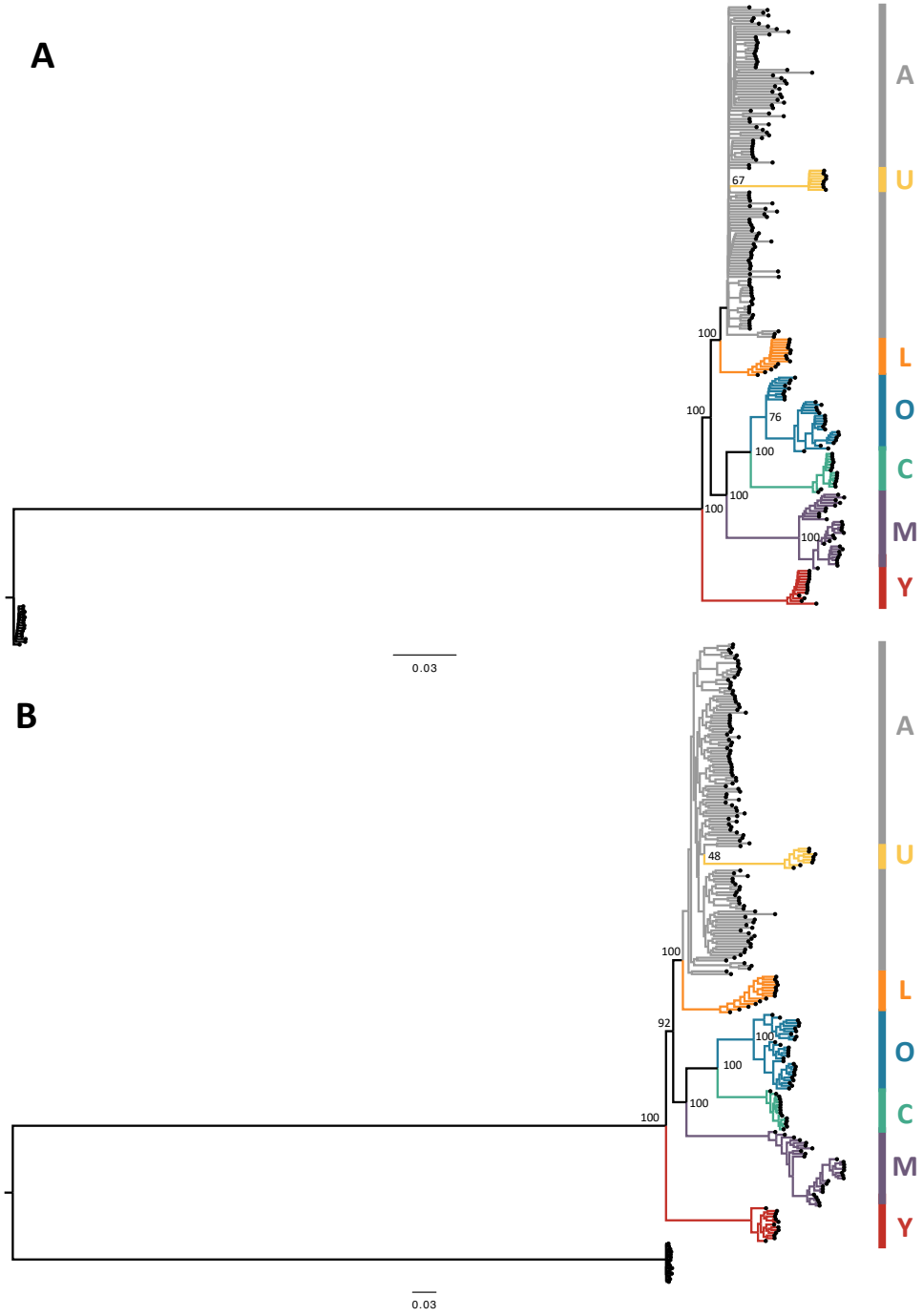

Fig. S2.6: Phylogeny of *Apis mellifera* using protein coding SNPs. The entire SNP dataset was pruned of ambiguous loci, as implemented by RAxML v8.2.12 (*63*), and loci with low coverage (<0.8) in *Apis cerana*. (**A**) A neighbor-joining tree constructed with SNPs located within protein coding regions (276,602) using allele-sharing distance and rooted with *Apis cerena*. Major nodes are labeled by bootstrap support. (**B**) A maximum-likelihood tree constructed with SNPs located

within protein coding regions (276,602) using the gamma model of rate heterogeneity (ASC_GTRGAMMA) with the Lewis ascertainment bias correction in RAxML v8.2.12 (*63*). The tree was rooted with *Apis cerena*. Major nodes are labeled by bootstrap support.

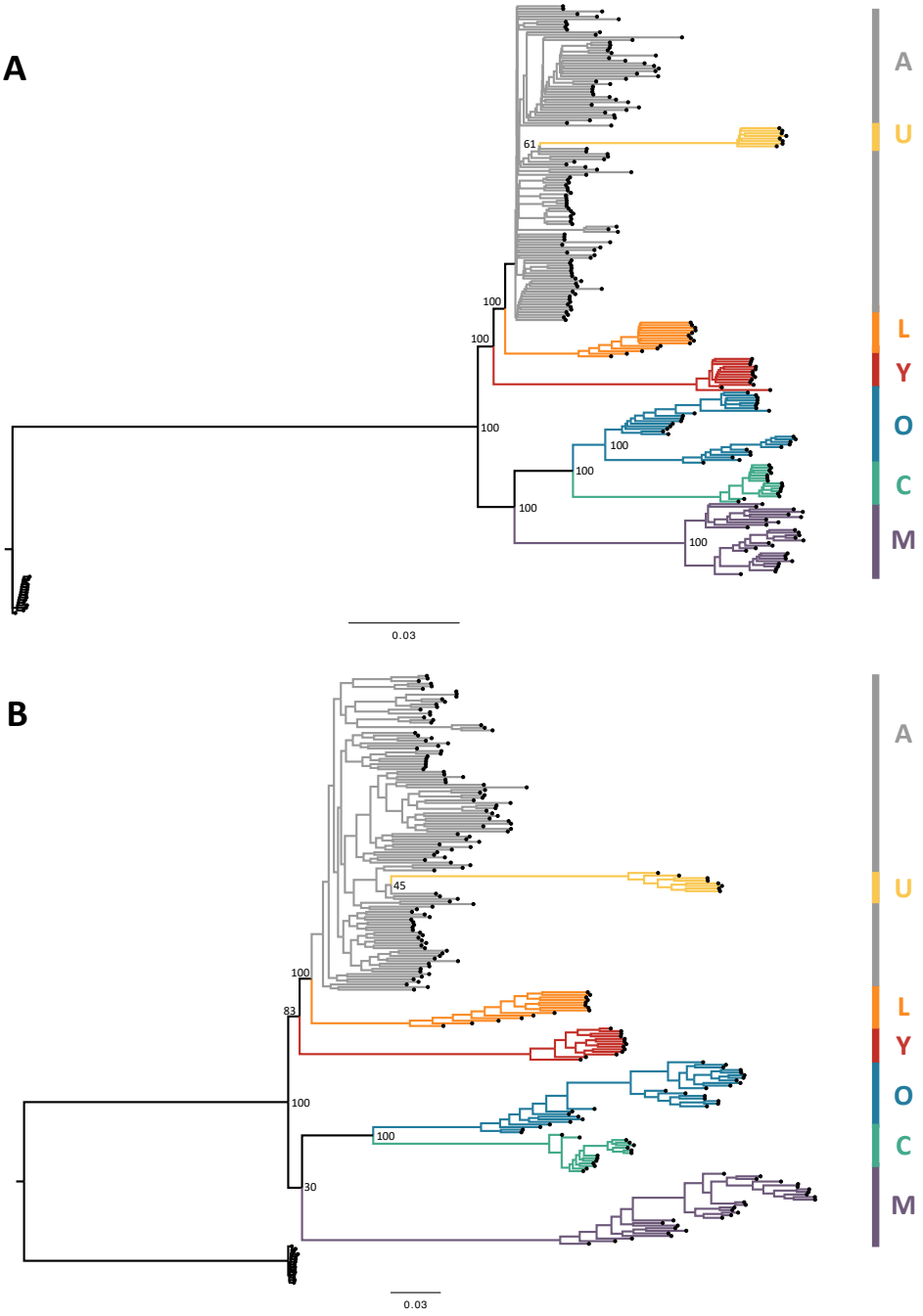

Fig. S2.7: Phylogeny of *Apis mellifera* using a subsample of SNPs. The entire SNP dataset was pruned of ambiguous loci, as implemented by RAxML v8.2.12 (*63*), and loci with low coverage (<0.8) in *Apis cerana*. (**A**) A neighbor-joining tree constructed with randomly selected SNPs located among intrageneric and intergenic regions (276,602) using allele-sharing distance and

rooted with *Apis cerena*. Major nodes are labeled by bootstrap support. (**B**) A maximum-likelihood tree constructed with randomly selected SNPs located among intrageneric and intergenic regions (276,602) using the gamma model of rate heterogeneity (ASC_GTRGAMMA) with the Lewis ascertainment bias correction in RAxML v8.2.12 (*63*). The tree was rooted with *Apis cerena*. Major nodes are labeled by bootstrap support.

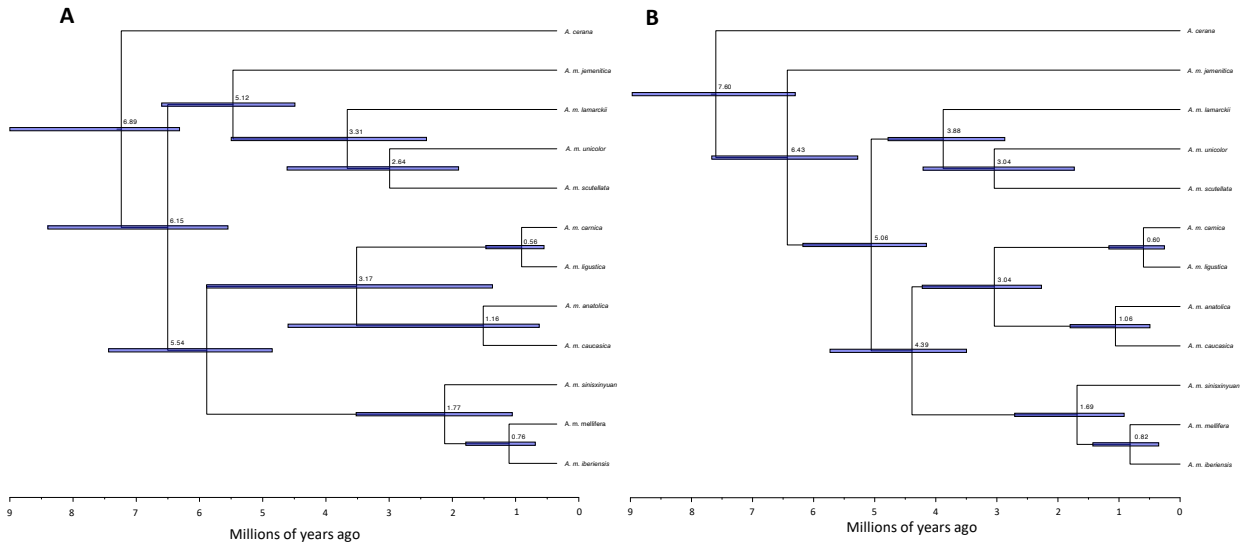

Fig S2.8: Divergence dating applied to both topologies resolved by the phylogenetic reconstruction of *Apis mellifera*. (**A**) Divergence time estimates applied to the phylogeny resolved using SNPs located genome wide (2,126,091). Estimated divergence times are labeled at the nodes, and purple bars at the nodes illustrate the 95% confidence interval of divergence times. (**B**) Divergence time estimates applied to the phylogeny resolved using SNPs located within coding regions (276,602). Estimated divergence times are labeled at the nodes, and purple bars at the nodes illustrate the 95% confidence interval of divergence times.

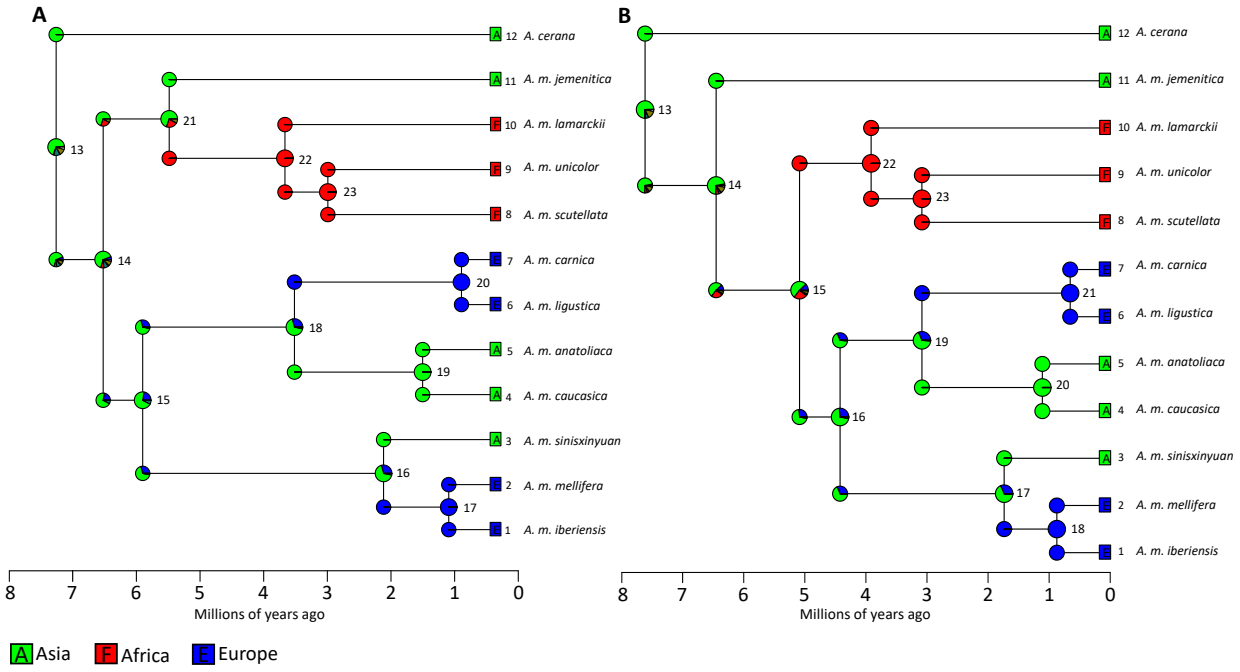

Fig. S2.9: Biogeographic range estimation applied to *Apis mellifera.* We applied a biogeographic range estimation to the topology (**A**) resolved using SNPs located genome wide (2,126,091) and to the topology (**B**) resolved using SNPs located within protein coding regions (276,602). We defined three biogeographic areas based on the current *Apis mellifera* distribution: Europe [E], Africa [F], and Asia [A] and tested all six biogeographic models provided by BioGeoBEARS (*69, 70*). We used the Akaike Information Criterion (AIC) and the Log of the likelihood scores (LnL) to compare models and determine the best fit to the phylogeny. The best fit model for each phylogeny (**A-B**) is highlighted in Table S1. Pie charts at nodes indicate the marginal maximum likelihood probabilities for the estimated ancestral range. Probabilities for each labeled node (1-23) in topologies (**A-B**) are outlined in Table S2.

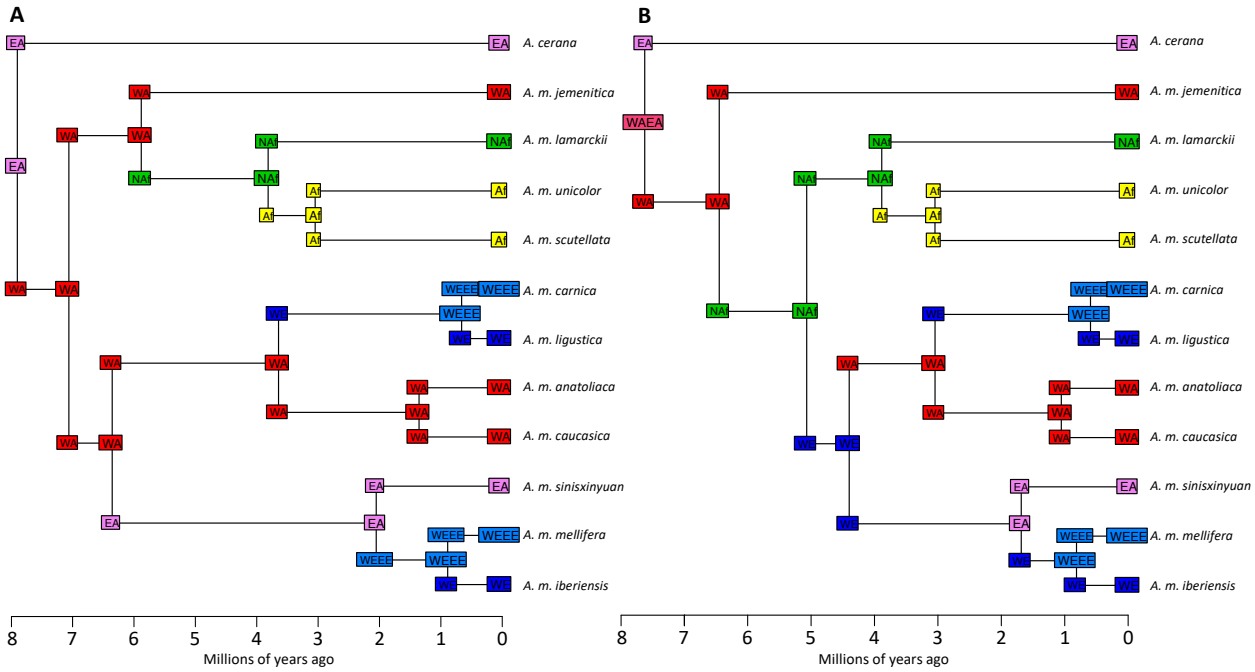

Fig. S2.10: Microregional biogeographic range estimation applied to *Apis mellifera.* We applied a biogeographic range estimation to the topology (**A**) resolved using SNPs located genome wide (2,126,091) and to the topology (**B**) resolved using SNPs located within protein coding regions (276,602). We defined six biogeographic areas based on the current *Apis mellifera* distribution: East Asia [EA], West Asia [WA], East Europe [EE], West Europe [WE], North Africa [NAf] and Africa [Af] and tested all six biogeographic models provided by BioGeoBEARS (*69, 70*). We restricted the analysis to only include states that had adjacent ranges, and limited the ranges occupied by the species to three (such as in the three-continent model; Fig S9). We used the Akaike Information Criterion (AIC) and the Log of the likelihood scores (LnL) to compare models and determine the best fit to the phylogeny, which was the DEC +J model for both topologies. The most likely ancestral regions are indicated at the nodes by letters corresponding to the defined biogeographic areas or combined states.

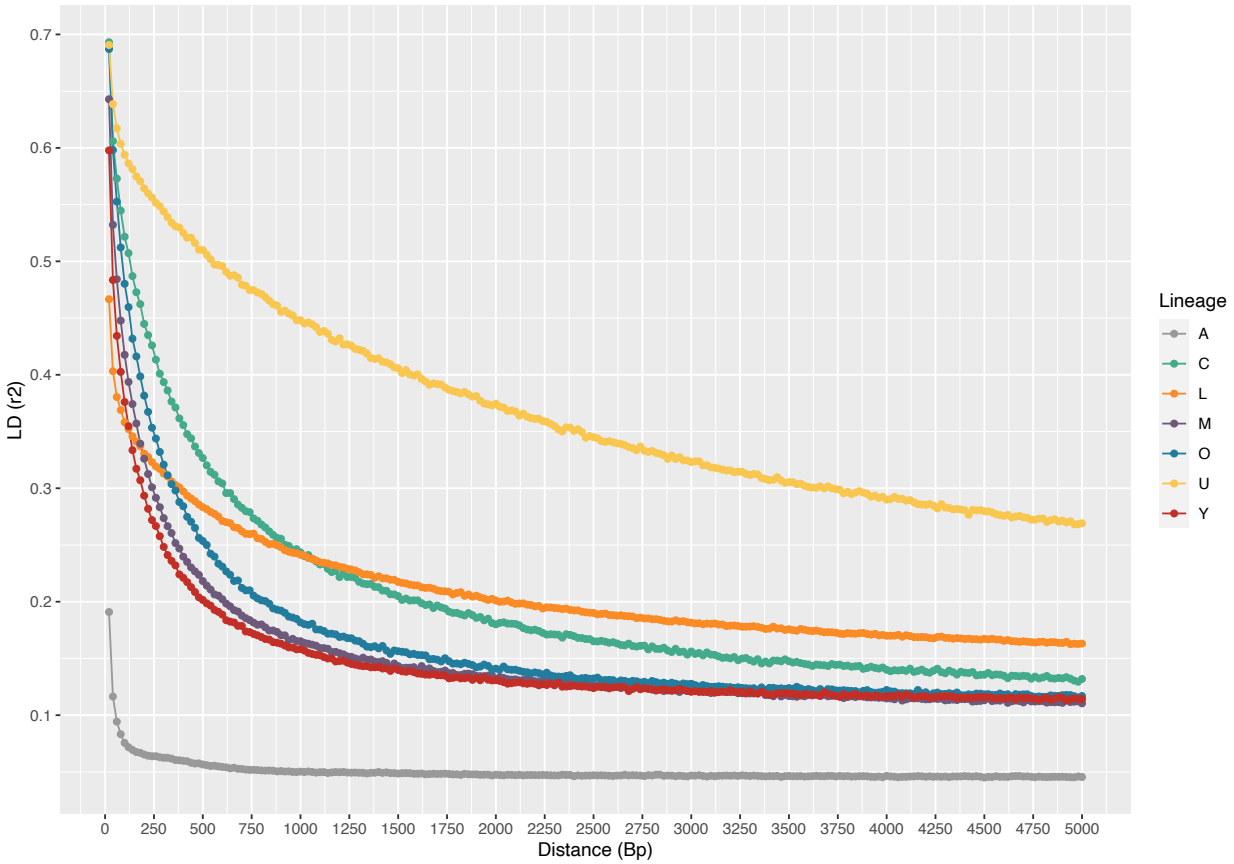

Fig. S2.11: Linkage disequilibrium decay among lineages. Linkage disequilibrium (LD) was measured for each lineage as the average squared correlation coefficient (r^2). LD was graphed as the average measure of r^2 per increasing intervals of 20bp up to a distance of 5000bp. LD decays by half at an average distance of 701bps among lineages, and decays the quickest among A lineage samples, reflective of a large effective population size.

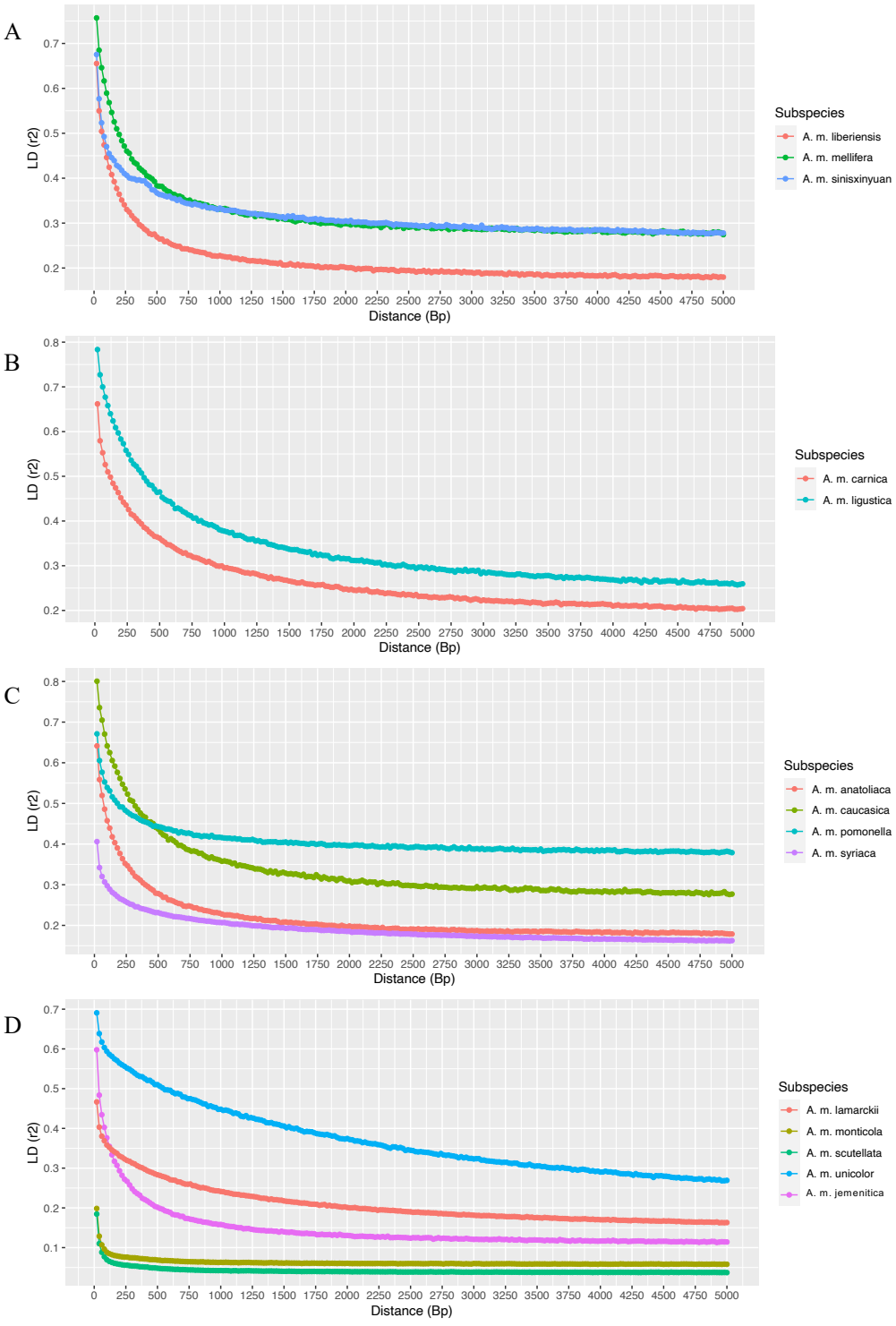

Fig. S2.12: Linkage disequilibrium decay among subspecies. Linkage disequilibrium (LD) was measured for each lineage as the average squared correlation coefficient (r^2). LD was graphed as the average measure of r^2 per increasing intervals of 20bp up to a distance of 5000bp. LD decays by half at an average distance of 716bp bps among lineages. (**A**) LD decay of M

lineage subspecies, (**B**) LD decay of C lineage subspecies, (**C**) LD decay of O lineage subspecies, (**D**) LD decay of A, L, Y and U lineage subspecies.

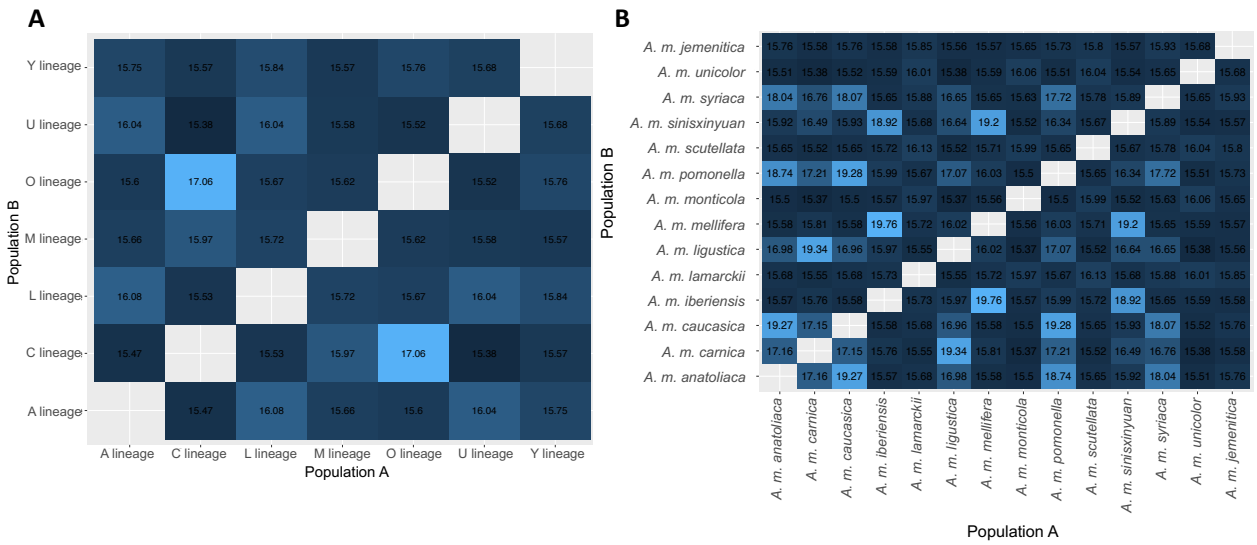

Fig. S2.13: Measures of shared genetic drift between lineages and subspecies. We calculated outgroup f_3 statistics, which is robust to lineage specific drift, to quantify the genetic distance between populations relative to an outgroup, *Apis cerana*. Higher f_3 values between lineage suggests greater shared genetic drift. (**A**) f_3 values estimates between lineages, (**B**) f_3 values estimates between subspecies.

Table S2.1: Models of ancestral range estimation applied to Fig. S2.9 (A-B). The best fit model to both topologies was DEC+J. LnL: Log-likelihood, k: number of parameters fitted to the model, d: rate of anagenetic range expansion, e: rate of anagenetic range contraction, j: weight of jump dispersal, AIC: Akaike Information Criterion.

Figure	Model	LnL	k	d	e	j	AIC
Fig.S2.9A	DEC	-15.18	2	2.83E+00	1.00E-12	0.00E+00	34.37
	DEC+J	**-9.81**	**3**	**1.00E-12**	**1.00E-12**	**1.16E-01**	**25.61**
	DIVA	-15.84	2	1.21E+00	4.76E+00	0.00E+00	35.67
	DIVA+J	-9.99	3	1.00E-12	4.06E-01	1.28E-01	25.99
	BAYAREA	-21.70	2	3.33E+00	5.00E+00	0.00E+00	47.39
	BAYAREA+J	-11.23	3	1.06E+00	2.26E-01	1.30E-01	28.46
Fig. S2.9B	DEC	-15.15	2	2.21E+00	1.00E-12	0.00E+00	34.29
	DEC+J	**-9.81**	**3**	**1.00E-12**	**1.00E-12**	**1.16E-01**	**25.63**
	DIVA	-15.48	2	2.47E+00	5.00E+00	0.00E+00	34.96
	DIVA+J	-9.85	3	1.00E-12	1.00E-12	1.22E-01	25.7
	BAYAREA	-39.54	2	9.58E-02	2.22E-01	0.00E+00	83.09
	BAYAREA+J	-10.09	3	1.00E-12	8.12E-08	1.07E-01	26.18

Table S2.2: Estimated probabilities of ancestral range for topologies in Fig. S2.9 (A-B). Node numbers refer to locations indicated on Fig. S10 (A-B). Each geographic location represents the three defined areas (Europe, Asia, Africa), or combined states.

Figure	Node	Europe	Asia	Africa	Europe + Asia	Europe + Africa	Asia + Africa	Europe + Asia + Africa
Fig. S2.9.A	1	1.000	0.000	0.000	0.000	0.000	0.000	0.000
	2	1.000	0.000	0.000	0.000	0.000	0.000	0.000
	3	0.000	1.000	0.000	0.000	0.000	0.000	0.000
	4	0.000	1.000	0.000	0.000	0.000	0.000	0.000
	5	0.000	1.000	0.000	0.000	0.000	0.000	0.000
	6	1.000	0.000	0.000	0.000	0.000	0.000	0.000
	7	1.000	0.000	0.000	0.000	0.000	0.000	0.000
	8	0.000	0.000	1.000	0.000	0.000	0.000	0.000
	9	0.000	0.000	1.000	0.000	0.000	0.000	0.000
	10	0.000	0.000	1.000	0.000	0.000	0.000	0.000
	11	0.000	1.000	0.000	0.000	0.000	0.000	0.000
	12	0.000	1.000	0.000	0.000	0.000	0.000	0.000
	13	0.015	0.675	0.015	0.121	0.004	0.124	0.046
	14	0.060	0.645	0.060	0.091	0.019	0.094	0.031
	15	0.223	0.700	0.000	0.077	0.000	0.000	0.000
	16	0.293	0.677	0.000	0.031	0.000	0.000	0.000
	17	1.000	0.000	0.000	0.000	0.000	0.000	0.000
	18	0.293	0.677	0.000	0.031	0.000	0.000	0.000
	19	0.000	1.000	0.000	0.000	0.000	0.000	0.000
	20	1.000	0.000	0.000	0.000	0.000	0.000	0.000
	21	0.000	0.697	0.222	0.000	0.000	0.080	0.000
	22	0.000	0.000	1.000	0.000	0.000	0.000	0.000
	23	0.000	0.000	1.000	0.000	0.000	0.000	0.000
Fig. S29.B.	1	1.000	0.000	0.000	0.000	0.000	0.000	0.000
	2	1.000	0.000	0.000	0.000	0.000	0.000	0.000
	3	0.000	1.000	0.000	0.000	0.000	0.000	0.000
	4	0.000	1.000	0.000	0.000	0.000	0.000	0.000
	5	0.000	1.000	0.000	0.000	0.000	0.000	0.000
	6	1.000	0.000	0.000	0.000	0.000	0.000	0.000
	7	1.000	0.000	0.000	0.000	0.000	0.000	0.000
	8	0.000	0.000	1.000	0.000	0.000	0.000	0.000

9	0.000	0.000	1.000	0.000	0.000	0.000	0.000
10	0.000	0.000	1.000	0.000	0.000	0.000	0.000
11	0.000	1.000	0.000	0.000	0.000	0.000	0.000
12	0.000	1.000	0.000	0.000	0.000	0.000	0.000
13	0.008	0.736	0.015	0.058	0.001	0.147	0.035
14	0.030	0.714	0.060	0.043	0.004	0.117	0.031
15	0.123	0.524	0.246	0.013	0.020	0.058	0.015
16	0.266	0.705	0.000	0.029	0.000	0.000	0.000
17	0.317	0.671	0.000	0.011	0.000	0.000	0.000
18	1.000	0.000	0.000	0.000	0.000	0.000	0.000
19	0.317	0.671	0.000	0.011	0.000	0.000	0.000
20	0.000	1.000	0.000	0.000	0.000	0.000	0.000
21	1.000	0.000	0.000	0.000	0.000	0.000	0.000
22	0.000	0.000	1.000	0.000	0.000	0.000	0.000
23	0.000	0.000	1.000	0.000	0.000	0.000	0.000

Table S2.3: Genetic variation of lineages. Genetic variation and measures of effective population size (N_e) for lineages using two different estimates of mutation rate (μ). Genetic variation measured by: π: nucleotide diversity, θ_w: Watterson's estimator, S: segregating sites, $S_{\text{singletons}}$: singletons.

Lineage	N	π	θ_w	S	$S_{\text{singletons}}$	N_e where $\mu = 5.27\text{x}10^{-9}$	N_e where $\mu = 3\text{x}10^{-9}$
M	17	0.00158	0.00183	1182374	413493	115620	203105
C	15	0.00138	0.00185	1158888	421866	116960	205459
O	13	0.00181	0.0017	1028854	281121	107802	189372
Y	13	0.00187	0.00196	1181707	331483	123819	217508
L	9	0.00277	0.00252	1369138	283530	159153	279579
A	111	0.00354	0.01012	9557653	3250486	640249	1124704
U	8	0.00233	0.0017	890294	140872	107277	188451

Table S2.4: Genetic variation of subspecies. Genetic variation and measures of effective population size (N_e) for subspecies using two different estimates of mutation rate (μ). Genetic variation measured by: π: nucleotide diversity, θ_w: Watterson's estimator, S: segregating sites, $S_{\text{singletons}}$: singletons.

Subspecies	N	π	θ_w	S	$S_{\text{singletons}}$	N_e where $\mu = 5.27\text{x}10^{-9}$	N_e where $\mu = 3\text{x}10^{-9}$
A. m. monticola	35	0.00349	0.00768	5840100	2409655	486049	853826
A. m. scutellata	76	0.00352	0.00934	8264280	3047949	590975	1038147
A. m. carnica	7	0.00148	0.00155	779953	384561	98070	172276
A. m. ligustica	6	0.00167	0.00135	645534	267854	85470	150143
A. m. lamarckii	9	0.00277	0.00252	1369138	283530	159153	279579
A. m. iberiensis	8	0.00172	0.00147	773919	231068	93255	163818
A. m. mellifera	5	0.00185	0.00138	618442	223847	87412	153554
A. m. sinisxinyuan	10	0.00177	0.00158	876114	185977	100243	176094
A. m. anatoliaca	7	0.00197	0.00173	872109	272238	109649	192617
A. m. caucasica	5	0.0018	0.00125	560203	195786	79179	139091
A. m. pomonella	5	0.00218	0.00211	944590	392631	133521	234552
A. m. syriaca	9	0.00254	0.00299	1628875	574300	189345	332617
A. m. unicolor	8	0.00233	0.0017	890294	140872	107277	188451
A. m. jemenitica	13	0.00187	0.00196	1181707	331483	123819	217508

Table S2.5: Measures of pairwise genetic differentiation calculated with weighted F_{ST} between lineages.

	M	C	O	Y	L	A	U
M	0.000						
C	0.664	0.000					
O	0.660	0.553	0.000				
Y	0.583	0.668	0.625	0.000			
L	0.657	0.609	0.557	0.533	0.000		
A	0.373	0.379	0.352	0.335	0.196	0.000	
U	0.655	0.686	0.635	0.614	0.481	0.274	0.000

Table S2.6: Measures of pairwise genetic differentiation calculated with weighted F_{ST} between subspecies.

	A. m. carnica	A. m. ligustica	A. m. monticola	A. m. scutellata	A. m. unicolor	A. m. caucasica	A. m. anatoliaca	A. m. pomonella	A. m. syriaca	A. m. mellifera	A. m. iberiensis	A. m. sinisxinyuan	A. m. lamarckii	A. m. jemenitica
A. m. carnica	0.000													
A. m. ligustica	0.129	0.000												
A. m. monticola	0.407	0.403	0.000											
A. m. scutellata	0.371	0.368	0.038	0.000										
A. m. unicolor	0.670	0.665	0.307	0.295	0.000									
A. m. caucasica	0.638	0.658	0.385	0.352	0.656	0.000								
A. m. anatoliaca	0.556	0.578	0.376	0.336	0.621	0.160	0.000							
A. m. pomonella	0.517	0.534	0.326	0.283	0.595	0.068	0.147	0.000						
A. m. syriaca	0.456	0.467	0.285	0.249	0.518	0.263	0.200	0.210	0.000					
A. m. mellifera	0.724	0.711	0.383	0.351	0.651	0.716	0.687	0.605	0.567	0.000				
A. m. iberiensis	0.701	0.690	0.404	0.367	0.647	0.719	0.675	0.624	0.569	0.108	0.000			
A. m. sinisxinyuan	0.594	0.581	0.382	0.345	0.613	0.644	0.604	0.539	0.510	0.207	0.260	0.000		
A. m. lamarckii	0.572	0.565	0.242	0.205	0.481	0.553	0.525	0.490	0.415	0.551	0.556	0.526	0.000	
A. m. jemenitica	0.650	0.650	0.379	0.343	0.614	0.641	0.609	0.589	0.530	0.654	0.666	0.631	0.533	0.000

Table S2.7: Summary of outlier SNPs among lineages. Number of outlier SNPs per lineage based on pairwise measures of F_{ST}. Number of genes associated with outlier SNPs for each lineage.

Lineage	Number of outlier SNPs	Genes
M (Europe)	6479	1538
M (Asia)	3885	990
C	10428	2539
O	7105	2012
Y	10854	2784
L	12148	2457
A	2	1
U	14434	2466

Table S2.8: Enrichment of outlier SNPs among genic and promoter regions. Chi-square (χ^2) test for the enrichment of outlier SNPs concentrated within three functional annotation categories (promoter, protein coding, and intronic) when compared to the expected genomic distribution. The percent change represents the percent increase or decrease of observed values compared to expected values. *P*-values were corrected for false discovery rate (FDR) using the Benjamini-Hochberg for each functional category.

Lineage	Annotation Region	Observed number of Outlier SNPs	Expected number of Outlier SNPs	χ^2	*P*-value	FDR	Percent Change
M (Europe)	Promoter	319	257	15.23	9.50E-05	**1.66E-04**	24.12
M (Asia)	Promoter	151	154	0.05	8.22E-01	8.22E-01	-1.95
C	Promoter	550	414	46.33	9.98E-12	**3.49E-11**	32.85
O	Promoter	319	282	4.91	2.67E-02	**3.12E-02**	13.12
Y	Promoter	544	431	30.74	2.96E-08	**6.91E-08**	26.22
L	Promoter	654	482	63.54	1.57E-15	**1.10E-14**	35.68
U	Promoter	649	573	10.39	1.26E-03	**1.77E-03**	13.26
M (Europe)	Protein Coding	606	468	43.34	4.59E-11	**6.43E-11**	29.49
M (Asia)	Protein Coding	302	281	1.64	2.01E-01	2.01E-01	7.47
C	Protein Coding	1087	754	158.69	2.18E-36	**1.53E-35**	44.16
O	Protein Coding	692	514	66.52	3.46E-16	**6.06E-16**	34.63
Y	Protein Coding	1008	785	68.43	1.32E-16	**3.08E-16**	28.41
L	Protein Coding	1176	878	108.87	1.74E-25	**6.09E-25**	33.94
U	Protein Coding	1097	1044	2.91	8.78E-02	1.02E-01	5.08
M (Europe)	Intron	2572	2668	5.81	1.59E-02	**2.79E-02**	-3.60
M (Asia)	Intron	1507	1600	9.05	2.62E-03	**6.12E-03**	-5.81
C	Intron	4125	4294	11.28	7.85E-04	**2.75E-03**	-3.94
O	Intron	2884	2926	0.99	3.21E-01	3.21E-01	-1.44
Y	Intron	4397	4469	1.97	1.60E-01	1.87E-01	-1.61
L	Intron	4816	5002	11.77	6.02E-04	**2.75E-03**	-3.72
U	Intron	5858	5944	2.08	1.49E-01	1.87E-01	-1.45

Table S2.9: Significance of overlap among genes under selection between lineages. A one-way Fisher's Exact test to evaluate whether the overlap of genes with outlier SNPs between lineages is greater than expected by chance. *P*-values were corrected for false discovery rate (FDR) using the Benjamini-Hochberg correction.

Population 1	Population 2	Observed Overlap	*P*-value	FDR
M lineage (Europe)	C lineage	640	1.61E-100	**1.69E-100**
M lineage (Asia)	C lineage	556	3.38E-157	**5.92E-157**
M lineage (Europe)	L lineage	660	1.90E-120	**2.34E-120**
M lineage (Asia)	L lineage	475	4.28E-104	**4.74E-104**
C lineage	L lineage	991	4.95E-174	**1.15E-173**
O lineage	L lineage	864	9.41E-165	**1.80E-164**
U lineage	L lineage	995	3.37E-172	**7.08E-172**
Y lineage	L lineage	1110	8.31E-194	**2.49E-193**
M lineage (Europe)	M lineage (Asia)	283	3.57E-50	**3.57E-50**
M lineage (Europe)	O lineage	610	1.23E-135	**1.84E-135**
M lineage (Asia)	O lineage	450	1.02E-121	**1.34E-121**
C lineage	O lineage	914	1.11E-201	**5.07E-201**
M lineage (Europe)	U lineage	701	1.35E-144	**2.18E-144**
M lineage (Asia)	U lineage	485	3.45E-110	**4.02E-110**
C lineage	U lineage	1036	1.21E-201	**5.07E-201**
O lineage	U lineage	915	3.25E-198	**1.14E-197**
M lineage (Europe)	Y lineage	816	2.13E-188	**5.60E-188**
M lineage (Asia)	Y lineage	536	2.64E-122	**3.69E-122**
C lineage	Y lineage	1135	1.46E-213	**1.02E-212**
O lineage	Y lineage	1043	1.95E-241	**2.05E-240**
U lineage	Y lineage	1186	3.45E-243	**7.24E-242**

Table S2.10: Simulations of the expected gene overlap between lineages relative to the observed overlap.

Population 1	Population 2	Expected Overlap (Min)	Expected Overlap (Mean)	Expected Overlap (Max)	Observed Overlap
M lineage (Asia)	C lineage	262	302	357	640
M lineage (Asia)	C lineage	160	195	231	556
M lineage (Europe)	L lineage	242	291	343	660
M lineage (Asia)	L lineage	155	189	224	475
C lineage	L lineage	429	482	546	991
O lineage	L lineage	319	388	438	864
U lineage	L lineage	415	469	546	995
Y lineage	L lineage	457	528	594	1110
M lineage (Europe)	M lineage (Asia)	89	117	149	283
M lineage (Europe)	O lineage	198	240	277	610
M lineage (Asia)	O lineage	120	154	188	450
C lineage	O lineage	347	296	444	914
M lineage (Europe)	U lineage	247	293	340	701
M lineage (Asia)	U lineage	158	189	223	485
C lineage	U lineage	430	484	536	1036
O lineage	U lineage	333	384	439	915
M lineage (Europe)	Y lineage	279	332	380	816
M lineage (Asia)	Y lineage	177	213	255	536
C lineage	Y lineage	489	547	604	1135
O lineage	Y lineage	372	433	508	1043
U lineage	Y lineage	478	531	602	1186

Data S1.
List of new and previously sequenced samples used in this study. File available online at: https://www.science.org/doi/full/10.1126/sciadv.abj2151.

Data S2
List of genes associated with at least one significant SNP among lineages. NA indicates genes that did not possess at least one outlier SNP for the corresponding lineage. This gene list also includes references to phenotypes associated with genes under selection among all lineages. File available online at: https://www.science.org/doi/full/10.1126/sciadv.abj2151.

Data S3.
List of significant gene ontology (GO) terms for each lineage with Benjamini-Hochberg corrected *p*-values. NA indicates the GO term was not present for the corresponding lineage. File available online at: https://www.science.org/doi/full/10.1126/sciadv.abj2151.

Chapter three: Patterns of admixture in Canadian honey bees are associated with genetic diversity and colony phenotypes.

Introduction:

The natural and anthropogenetic translocation of plants and animals has led to pervasive hybridization at the population (intraspecific) and species level (interspecific) (Mallet, 2005). Admixture between genetically distinct groups can lead to the exchange of novel mutations through geneflow. The introduction of new diversity may have benefits, including increased standing genetic variation and novel genotype associations, both of which can be acted on by natural selection (Hedrick, 2013). Alternatively, it has been argued that admixture can reduce population fitness through the breakdown of local adaptation and the introduction of maladapted or deleterious mutations (Kim et al., 2018, Muhlfeld et al., 2009). Regardless of the outcome, hybridization is considered an important evolutionary event that can facilitate population dispersal and adaptation (Seehausen, 2004), speciation (Abbott et al., 2013), and is recognized as an important consideration for conservation (vonHoldt et al., 2018, Hamilton and Miller, 2016).

When admixture occurs between genetically distinct populations or lineages it often results in the permanent and asymmetric infiltration of genes and alleles throughout the species genome (Winkler et al., 2010). This event generates a mosaic of "ancestry blocks" along the chromosome, the frequency and length of which is maintained as a function of time since admixture, the duration of geneflow (Winkler et al., 2010), recombination (Buerkle and Lexer, 2008), and selection (Abbott et al., 2016). One approach to detecting and understanding the architecture of admixture is to use global and local estimates of ancestry. Global ancestry seeks to identify the overall average ancestry proportions of focal individuals, while local ancestry is focused on mapping the distinct ancestral segments across the genome of focal individuals (Liu et al., 2013). Both tools offer a valuable way to gather insight about the ancestral complexity of populations, while local admixture provides a unique way to isolate the effects of ancestry on genes and phenotypes of interest.

The western honey bee, *Apis mellifera*, is a prime candidate to study admixture patterns as hybridization is a prominent occurrence. The species, which is native to Europe, Africa, and Western Asia, is composed of seven genetically distinct lineages: the M lineage of Eurasia, the C lineage of Europe, the O and Y lineage of Western Asia, and the A, L and U lineages in Africa

(Dogantzis et al., 2021). The honey bee has been introduced worldwide for apiculture, and in North America, historical introductions have included subspecies from serval different ancestral lineages (Carpenter and Harpur, 2021). Global ancestry analysis using ancestry informative markers (AIM) has shown that managed honey bees in North America are primarily composed of C-lineage ancestry (>70%), but also show signs of M and A lineage introgression (Chapman et al., 2015, Harpur et al., 2015). However, these estimates of ancestry proportions were conducted with data using only three ancestral lineages, while recent population genomic work has uncovered a least seven genetically distinct groups (Dogantzis et al., 2021). As such, a reassessment of the ancestral composition of managed bees is needed to ensure we have a full understanding of their genetic complexity. Additionally, we do not know how ancestry is distributed across the genome of north American commercial populations. Recent work using local ancestry on Africanized bee populations has revealed non-random introgression among admixed colonies, which has been linked to behaviours related to reproduction, foraging (Nelson et al., 2017), and colony defense (Harpur et al., 2020). Thus, elucidating the pattern and organization of admixture throughout the genome of managed colonies is imperative for understanding its effect on diversity and its association with economically relevant phenotypes.

Here, we analysed 1350 pooled genomes representative of colonies collected cross five Canadian provinces. Using global ancestry analysis, we categorized the ancestral composition of commercial Canadian honey bee colonies. This study is the first to use reference samples from all seven genetically distinct lineages and offers a comprehensive categorization of the ancestral composition of managed honey bees. Next, we examined the genetic diversity of the studied colonies. Within colony genetic diversity has been linked to higher colony productivity and survival (Tarpy et al., 2013). Previous work demonstrated that genetic admixture is an important factor influencing the diversity of managed honey bees (Harpur et al., 2012), thus we hypothesize that levels of genetic diversity should be positively correlated with admixture proportions. Finally, we inferred pattens of local ancestry among admixed colonies to determine if ancestry segregates non-randomly across the genome. In recent studies on Africanized bees, introgression from the minor donor lineage was maintained primarily on a single chromosome and encompassed genes related to colony phenotypes (Harpur et al., 2020, Nelson et al., 2017). Here, we hypothesize that genomic intervals with outlier levels of admixture will be focused in hotspots among the most functionally relevant parts of the honey bee genome (i.e. protein coding

regions), which may suggest that outlier levels of admixture are associated with some subset of phenotypes affecting fitness in commercial populations.

Methods:

Genome sequence processing and SNP detection

We used 1350 pooled genome sequences representative of commercial honey bee colonies sampled across five provinces in Canada: British Columbia (N=280), Alberta (N=373), Manitoba (N=267), Ontario (N=197), and Quebec (N=233). Details on sample collection have been described elsewhere (Borba et al., 2022). For each colony, the legs of 50 workers were pooled for DNA extraction and sequencing. DNA extraction followed a previously published protocol (Dogantzis et al., 2021). Genome sequencing was carried out at the Genome Quebec Innovation Centre using the Illumina HiSeqX 150bp paired-end sequencing platform at 90X coverage. The dataset was supplemented with previously sequenced reference samples from seven genetically distinct honey bee lineages; the A (N=16), L (N=9), and U (N=8) lineages in Africa, the Y (N=11) and O (N=12) lineages found in Western Asia, and the C (N=12) and M (N=13) lineages found in Europe. Information on sample collection, DNA extraction, and sequencing for these samples have been described elsewhere (Dogantzis et al., 2021). Sequence reads of pooled Canadian colonies and reference samples were trimmed of Illumina adapters and low-quality bases (<20) using Trimmomatic v0.36 (Bolger et al., 2014). Reads were aligned to the current *Apis mellifera* reference genome (HAv3.1) (Wallberg et al., 2019) using NextGenMap aligner v0.5.0 (Sedlazeck et al., 2013). BAM files were sorted using SAMtools v1.9 (Li et al., 2009) and marked for duplicates using MarkDuplicates in GATK v4.1.0.0 (Van der Auwera and O'Connor, 2020). Variant detection among the pooled genomes was conducted with Lofreq (Wilm et al., 2012), which uses mapped reads to determine allele frequencies for reference and alternative SNPs. Variants were removed if they were presented in less than 15% of colonies, and if the collective frequency of the minor allele fell below 0.2 of the interdecile range. Variants among reference samples were identified using HaplotypeCaller in GATK v4.1.0.0 (Poplin et al., 2017), and SNPs were retained for analyses if they were bi-allelic, had

sufficient coverage (> 0.8) and were previously identified among the Canadian population of honey bees. In total 1,509,553 SNPs were retained.

Consensus genotype for pooled samples

Variants from pooled sequenced data are typically identified as allele frequency estimates based on the aligned sequence reads. Ancestry mapping requires knowledge of genotypes, and as such, we had to estimate a consensus genotype for each colony using simulations. To convert the allele frequency calls of the pooled samples into a colony consensus genotype, we first created an algorithm to simulate segregation of alleles within a colony for a multiply mated honey bee queen. Queens were assigned a genotype of homozygous reference (0/0), heterozygous (0/1), or homozygous alternative (1/1), while drones genotypes were randomly set to *i*=0 or *j*=1. Simulations randomly chose one allele from the queen, and one allele from a pool of 20 drone alleles. Here, we assumed that queens are mated with up to 20 drones, which is within the range of average observed mating number 25 ± 13.11 and effective paternity 16.0 ± 9.28 for commercially bred queens (Delaney et al., 2011). The allele *i* was represented at a frequency of $p=k/20$, and *j* was represented at a frequency of $q=20-k/20$, where *k* was equal to 1-19. Simulations were run for 1000 replicates per gene pool, resulting in 19,000 genotype combinations per queen genotype. Allele frequencies where calculated based on the resulting genotypes per gene pool. We then cross referenced the allele frequencies of the pooled samples with the simulated dataset to determine the consensus genotype for the colony. In brief, the consensus genotypes at any SNP were assigned based on the following allele frequency thresholds: a consensus homozygous reference genotype was assumed when $q \leq 0.25$, a consensus heterozygous genotype was assumed when $0.25 < q < 0.75$, a consensus homozygous alternative genotype was assumed when $q \geq 0.75$. A full description of this process is included in the Supplementary Text.

Genetic diversity and differentiation

We calculated the genetic diversity of single colonies using the estimated allele frequencies of the reference and alternative allele, not using the consensus genotype.

Nucleotide diversity (π) was calculate in 1kb windows using the equation $(2*j*(n-j)/(n(n-1))$ (Begun et al., 2007) where j is the minor allele count, and n is the number of chromosomes. We assumed an equal representation of samples in the pooled genome (n=100). The pairwise F_{ST} matrix between provinces was calculated based on the average per-locus estimate of F_{ST}, using $F_{ST} = (H_t - H_s)/H_t$, where H_t is the total heterozygosity, and H_s is the average heterozygosity across subpopulations. The absolute divergence (d_{xy})(Nei and Li, 1979) was calculated in 5kb windows between ancestral lineages and Canadian colonies $d_{xy} = p_1(1-p_2) + p_2(1-p_1)$, where p_1 is the frequency of the alterative allele in group one, and p_2 is the allele frequency of the alternative allele in group two (Van Doren et al., 2017, Tavares et al., 2018).

Estimates of Global Ancestry

The program ADMIXTURE v 1.3.0 (Alexander and Lange, 2011) was used to estimate global ancestry and admixture proportions for all Canadian honey bee colonies using the consensus genotypes. The analysis was run in the supervised learning mode, which allows for a more accurate estimate of ancestry, using the 81 reference samples grouped into their known lineages.

Estimates of local ancestry

We estimated local ancestry across the genome of Canadian honey bee colonies using the C, M and O lineages as the reference source populations with the program Loter (Dias-Alves et al., 2018). Only the aforementioned lineages were used as they contribute, on average, to 96.4% of the total ancestry profile of Canadian honey bees, as found in this study. The variant dataset was pruned for ancestry informative markers (AIMs), based on the source populations, defined as loci that had F_{ST} values ≥0.50 consistently across pairwise lineage comparisons and was restricted to loci located on assembled chromosomes (1,373,668 SNPs). Genomes of the reference and Canadian samples were phased, and missing data was imputed using Beagle v5.0 (Browning et al., 2018). The program Loter was run with all three source populations with bagging enabled. Genomic regions enriched for an ancestral lineage were defined by >5kb of consecutive loci that had ancestral frequencies in the top 95th percentile.

Gene ontology and QTL mapping

Gene ontology (GO) enrichment was conducted with DAVID v6.8 (Huang et al., 2009) using *Drosophila melanogaster* orthologs. *Apis mellifera* orthologs to *D. melanogaster* were identified using reciprocal blastp matches and the HymenopteraMine v1.5 database (Elsik et al., 2016). *Drosophila* orthologs were found for 6773 genes, which were used as the background. Gene ontology was conducted with We only included GO functional annotation clusters with an enrichment score of ≥1.3 and GO terms with $P < 0.05$ after false-discovery rate in the results.

Broad quantitative trait loci (QTL) have been identified for *Varroa* tolerant traits. We used QTL segments, as approximated in Mondet, Beaurepaire et al. (2020), and mapped them to the honey bee genome AMEL 4.5 (Elsik et al., 2014). To find the equivalent sequences in the new genome assembly (HAv3.1) (Wallberg et al., 2019) we used NCBI's genome remapping service to remap intervals.

Results:

Sample overview

We sampled and pool-sequenced 50 individual honey bees from 1350 colonies across Canada. Colonies were sampled in five provinces, which included British Columbia (N=280), Alberta (N=373), Manitoba (N=267), Ontario (N=197), and Quebec (233) (Fig. 3.1). The dataset was supplemented with 81 genomes representative of the reference *Apis mellifera* lineages; the M and C lineages of Europe, the O and Y lineages of Western Asia, and the A, L and U lineages found in Africa. In total, there were >1.5 million SNP (single-nucleotide polymorphisms) identified among the Canadian colonies and reference ancestral lineages that were retained for analyses.

Categorizing global admixture in Canadian honey bee colonies

The structure analysis revealed that Canadian honey bee colonies have high levels of genetic admixture. Canadian honey bee colonies were primarily composed of ancestry from the C lineage (86.6% ± 5.3%), while the remainder of ancestry was admixed from the M (7.3% ±

3.1%), and O (5.5 ± 3.4%) lineages (Fig. 3.2A). Average ancestry from the remaining lineages was negligible (<0.2%). The levels of admixture, defined by the proportion of non-C lineage ancestry, were highest amongst honey bee colonies in Ontario (18% ± 3.2%), while levels of admixture were lowest in Manitoba (9.8% ± 6.7%) (Fig. 3.2A). Levels of mean admixture were significantly different between provinces ($F_{4,1344} = 99.23$, $p < 2e^{-16}$), except for Alberta and British Columbia (Tukey $p = 0.99$).

Most of the variance in admixture among honey bee colonies can be attributed to differences in the proportion of M and O lineage ancestry. Patterns of average M lineage ancestry were significantly different between provinces ($F_{4,1345} = 95.83$, $p < 2e^{-16}$) where ancestry proportions ranged between 0 and 23% among Canadian colonies. Pairwise comparisons between provinces were all significantly different (Tukey $p < 6.5e^{-4}$), except between British Columbia and Alberta (Tukey $p = 0.87$). Likewise, the average O lineage ancestry was significantly different between provinces ($F_{4,1345} = 46.27$, $p < 2e^{-16}$) and ancestry among colonies ranged between 0 and 37.2%. Pairwise comparisons between provinces were all significantly different (Tukey $p > 3.4e^{-3}$), except between British Columbia and Alberta (Tukey $p = 0.19$), Quebec and British Columbia (Tukey $p = 0.11$), and Quebec and Manitoba (Tukey $p = 0.85$). Despite the variability in admixture across Canadian colonies, there was low levels of genetic differentiation between provinces (mean $F_{ST} = 0.0053$) (Table S3.1).

Patterns of diversity among Canadian honey bee colonies

We assessed the genetic diversity of Canadian honey bee colonies by estimating nucleotide diversity (π). Nucleotide diversity within colonies was high and ranged between 0.00111 - 0.00204 with a mean of 0.00160 ± 0.00015 (Table 3.1). Averaged across colonies, π significantly differed between provinces ($F_{4,1344} = 153.5$, $p < 2e^{-16}$). Pairwise comparisons of π between provinces were all significantly different (Tukey $p < 1.38e^{-03}$) except between Alberta and British Colombia (Tukey $p = 0.84$) (Fig. 3.2B; Table 3.1). Measures of π were highest amongst colonies in Ontario 0.00169 ± 0.00012 and lowest among colonies in Manitoba 0.00145 ± 0.00014 (Table 3.1). Given the association between admixture and genetic diversity, we evaluated whether the proportion of admixture correlated with measures of genetic

diversity among Canadian honey bee colonies. We found a significant positive correlation between the level of non-C lineage introgression and levels of nucleotide diversity ($r_{(1348)} = 0.65$, $p < 2.2e^{-16}$) (Fig. 3.2C).

Finally, we compared measures of nucleotide diversity (π) between managed Canadian colonies and native *Apis mellifera* lineages. We found that average π among native *Apis mellifera* lineages, across the same loci, had markedly lower levels of diversity (0.00100 ± 0.00003), relative to the average π among Canadian honey bee colonies (0.00160 ± 0.00015) (Table 3.1). Measures of π were highest among the A (0.00144) and C-lineages (0.00124), and lowest among the Y-lineage (0.00058) (Table 3.1).

Identifying outlier regions of admixed ancestry

We estimated local ancestry among Canadian honey bee colonies using ancestry informative markers from the C, M and O source populations (see Methods). Only the aforementioned lineages were used as they contribute, on average, to 96.4% of the total ancestry profile. First, global ancestry was estimated for each colony to ensure ancestry proportions from Loter were consistent with global estimates from ADMIXTURE. We found that the difference in ancestry between programs was on average 9% for C-lineage ancestry, 6.3% for M-lineage ancestry, and 3.1% for O-lineage ancestry. While variation in ancestry proportion estimates is expected due to differences in software and the loci used, we tried to minimize variation by retaining samples that had less than a 10% difference among paired ancestry proportions. After filtering samples, we retained 876 colonies that had comparable estimates of global ancestry between Loter and ADMIXTURE for the C (77.0% vs 84.6% respectively), O (9.07% vs 6.54 % respectively), and M (14.0% vs 8.3% respectively) lineages.

Next, we investigated whether portions of the genome were enriched for ancestry from the minor donor lineages (M and O), which may indicate a genetic advantage over the dominant C-lineage. Enriched intervals were categorized as segments of the genome that had >5kbs of consecutive loci enriched for outlier proportions (highest 5%) of non-C-lineage ancestry (M and O combined). Since M and O ancestry are distributed as low frequencies across the genome, the top 5% corresponded to regions where >35% of the population ancestry was

attributed to the minor lineage. As such, we used a minimum threshold of 50%, to ensure we were targeting regions where at least half of the population retained the minor ancestry. In total, we identified 114 genomic intervals enriched for outlier proportions of non-C-lineage ancestry (M and O combined); hereby referred to as outlier intervals (Fig. 3.3). Among outlier intervals, the population wide level of non-C-lineage ancestry averaged 63.3%, and never went above 84.6%, indicating there were no loci fixed for the minor donor lineage.

We used measures of absolute divergence (d_{xy}) to support enrichment of the minor donor lineages among outlier intervals, relative to non-outlier regions. We expect outlier intervals to display low levels of d_{xy} between Canadian colonies and the M lineage and between Canadian colonies and the O lineage, indicating higher levels of geneflow. We found that average measures of d_{xy} between the M lineage and Canadian colonies was significantly reduced among outlier intervals ($d_{xy} = 1.48e^{-03}$) relative to the non-outlier regions ($d_{xy} = 3.66e^{-03}$) (Mann-Whitney U, $p < 2.2e^{-16}$). Additionally, we found that average measures of d_{xy} between the O lineage and Canadian colonies was significantly reduced among outlier intervals ($d_{xy} = 1.40e^{-03}$) relative to the non-outlier regions ($d_{xy} = 2.76e^{-03}$) (Mann-Whitney U, $p < 2.2e^{-16}$).

Categorization of outlier regions

Local admixture mapping identified 114 outlier segments enriched for non-C lineage ancestry (M and O lineages) (Fig. 3.3). The size of outlier intervals averaged 14.9kb in length, with the longest segment being >82kb. Interestingly, outlier intervals appear to be distributed nonrandomly throughout the genome, with 33 and 36 intervals residing on chromosome 4 and 7 respectively (60.5% of intervals). Among outlier intervals there were 74 unique genes that were associated with at least one ancestry informative marker (AIM). To gain insight into the functions associated with genes among outlier intervals, we performed a gene ontology (GO) enrichment analysis. We found no significant enrichment for outlier intervals after FDR correction. However, prior to FDR correction, we identified 20 significant GO terms that were primarily related to lipid and fatty acid biosynthesis or metabolism. There were several genes among outlier intervals found to overlap with loci of functional significance, as discovered by other researchers. For example, previous proteomic studies found that odorant binding protein

Obp15 (GB46224) (Hu et al., 2016) and Obp16 (GB46225) (Guarna et al., 2015), located on chromosome 15, are associated with response to *Varroa* infection (Mondet et al., 2020b) (Fig. 3.3). In addition, we found three genes among outlier intervals that were found to be differentially expressed in response to *Nosema* infection, GB46290 (DeGrandi-Hoffman et al., 2018), GB51236 (Li et al., 2019), and GB51580 (Badaoui et al., 2017). Interestingly, outlier intervals contained cytochrome P450 genes Cyp6be1 (GB46814), and Cyp6as3 (GB49887) (Fig. 3.3), which are associated with xenobiotic detoxification (Berenbaum and Johnson, 2015), and have been found to be differentially expressed in response to neonicotinoid exposure (Chen et al., 2021, Alptekin et al., 2016). Finally, outlier intervals contained genes related to caste differentiation, including GB47159, which has been associated with cell death regulation in worker honey bee ovaries (Ronai et al., 2016), and GB49642, which has been implicated in self-sacrifice behaviour among honey bee workers (Mullen and Thompson, 2015).

Overlap with previous Varroa QTL regions

Quantitative trait loci (QTL) mapping has identified several regions across the genome (Mondet et al., 2020a), particularly within chromosome 4, 7, 9 (Behrens et al., 2011) and 5 (Arechavaleta-Velasco et al., 2012), involved in *Varroa* tolerance. We found that genome segments with outlier admixed ancestry were concentrated among chromosome 4 and 7, and contained several genes differentially expressed with *Varroa* infection (see above). As such, we investigated if outlier segments overlapped with *Varroa* resistance QTLs (Mondet et al., 2020a). We found that outlier intervals significantly overlapped with QTL regions at a greater proportion (26%), relative to non-outlier intervals (13%) (χ^2, $p < 2.2e^{-16}$). This coverage was concentrated on chromosome 7 where all outlier segments in this region overlapped with a suspected QTL (Fig. 3.3). We did not find overlap anywhere else in the genome. Among these segments there were 12 genes of interest, which accounted for 16% of unique genes that were also associated with at least one ancestry informative marker.

Discussion:

The genomic patterns found in Canadian honey bee colonies are the result of a complicated ancestral history generated through human mediated introductions and breeding of different subspecies. Here, we find that contemporary populations of Canadian honey bees are highly admixed, composed primarily of C lineage ancestry with the remaining ancestry originating from the M and O lineages. Comparable levels of O lineage introgression have been previously detected in North American honey bee colonies (Whitfield et al., 2006, Wallberg et al., 2014) and likely confirms the introduction of subspecies from Western Asia (Sheppard, 1989). Contrary to previous estimates of ancestry in Canadian colonies we do not detect notable levels of A lineage ancestry (Harpur et al., 2015). The discrepancy in admixture may be due to the use of different reference lineages and SNP markers, which suggests that biases in the data from the absence of all potential source populations may lead to slight misclassifications in ancestry. Alternatively, by imputing the consensus genotypes, we could be losing low frequency alleles representative of A lineage ancestry. Despite this discrepancy, our analysis is still able to detect non-random patterns of admixture among the most common sources of genetic ancestry in Canadian colonies – the C, M and O groups.

We detected a statistically significant positive correlation between colony admixture and genetic diversity. It is widely postulated that admixture between disparate populations can increase diversity by increasing standing genetic variation. For example, admixture events during post-glacial expansions (Sakaguchi et al., 2011) and during species invasions (Kolbe et al., 2008) has been implicated in population success due to the accompanying increase in genetic diversity. The honey bee, *Apis mellifera*, is composed of seven distinct evolutionary lineages whereby genetic differences between lineages are large, but differences between subspecies within a lineage are relatively small (Dogantzis and Zayed, 2019, Dogantzis et al., 2021). The ancestral mosaic that comprises commercial honey bee colonies in Canada combines at least three lineages (C, M, O), and so it follows that we expect to detect greater genetic variation with increasing admixture proportions. We also find, when comparing the same variant sites among the progenitor lineages, average measures of nucleotide diversity among managed Canadian colonies were markedly higher relative to their source populations. This trend

parallels a previous finding that detected higher levels of admixture and genetic diversity in managed honey bees from North America and Europe, relative to their progenitor populations (Harpur et al., 2012, Harpur et al., 2013). Though our study shows a positive influence of admixture on genetic variation, we do not recommend the intentional admixture of pure honey bee subspecies due to the potential loss of genetic integrity and locally adapted gene complexes (Ellis et al., 2018, Espregueira Themudo et al., 2020).

Given that Canadian honey bees are a mosaic of ancestry from genetically distinct lineages, this presents a unique opportunity to study how ancestry is distributed across the genome of managed colonies. Using local admixture mapping, we identified several genomic segments enriched for non-C lineage ancestry. Interestingly, outlier segments overlapped with previous QTLs related to *Varroa* tolerance and contained genes associated with xenobiotic detoxification. Honey bees are often selectively bred for economically favourable traits, thus these finding may suggest that M and O lineage alleles offer an advantage over C ancestry at some loci, and that it is possible that the outlier levels of admixture we identify here is the result of artificial selection by commercial beekeepers and bee breeders.

Colony resistance to *Varroa destructor*, a parasitic mite that can lead to colony mortality, has been an intense focus of selective breeding practices. Several QTL's have been linked to *Varroa* resistance, and in this study, we find that 36% of outlier intervals overlapped with previously identified QTLs. This overlap was restricted to chromosome 7, which has been linked to *Varroa* tolerance by limiting mite reproduction (Behrens et al., 2011, Lattorff et al., 2015, Mondet et al., 2020b). Colony level phenotypes such as hygienic behaviour and *Varroa*-sensitive hygiene are also sought-after traits (van Alphen and Fernhout, 2020, Rinderer et al., 2010), and odorant binding proteins have been implicated as strong candidate genes underlying these phenotypes (Mondet et al., 2020a). In this study we found genes Obp15 and Obp16 among outlier regions. Obp15 has been associated with *Varroa* sensitive hygiene and is down regulated in antenna tissue of *Varroa* sensitive bees (Hu et al., 2016), while Obp16 has been linked to hygienic behaviour (Guarna et al., 2015). These genes, located on chromosome 15, are clustered among several other odorant binding proteins that may have experienced artificial selection in recent years.

Honey bee health can also be negatively impacted from exposure to different xenobiotic compounds (Berenbaum and Johnson, 2015). For example, sublethal levels of pesticides can dimmish colony performance by affecting immunity, cognition, behaviour, and reproduction (Chmiel et al., 2020). Honey bees have a suite of detoxification enzymes, like those encoded by cytochrome P450 (CYP450) genes, to mitigate the effects from xenobiotic exposure (Berenbaum and Johnson, 2015). In this study, we find gene Cyp6be1 and Cyp6as3 among outlier intervals. Gene Cyp6be1 has been previously shown to be upregulated in honey bees exposed to thiacloprid (Alptekin et al., 2016), while both aforementioned genes have been shown to be downregulated among worker honey bees exposed to imidacloprid (Chen et al., 2021). Additionally, gene Cyp6as3 is located proximally to Cyp6as5 on chromosome 13, which has shown activity against thiacloprid exposures (Manjon et al., 2018). Because honeybees are often kept near agricultural areas for pollination, it is possible that exposure to agrochemicals has increased the frequency of M or O alleles at some detoxification genes; assuming these alleles offer an enhanced ability to detoxify pesticides.

In conclusion, our study demonstrates that managed Canadian honey bee colonies maintain high levels of admixture. The variability in ancestry between colonies is influenced by changes in ancestry proportions among the C lineage of East Europe, the M lineage in Western Europe, and the O lineage located in Western Asia. As expected, we find that admixture levels correlate strongly with the genetic diversity of colonies. Finally, local ancestry analysis revealed that genomics segments with outlier levels of admixture were concentrated on chromosome 4 and 7. Genes among outlier regions had links to honey bee health, including parasite and xenobiotic tolerance, suggesting M or O alleles at some loci may confer an advantage for some honey bee traits. Follow-up studies are needed to better understand the functional significance of admixture on phenotypes associated with disease tolerance. For instance, selective breeding experiments with colonies who possess admixed ancestry among target outlier intervals could be examined for their association with phenotypes of interest. Such colonies could also be scanned for signatures of artificial selection to investigate the association of outlier intervals with selected loci. Such studies are needed to better understand the importance of admixture on adaption in honey bees.

Fig. 3.1: A map of approximate sampling locations across Canada. Sample were collected from British Columbia (N=280), Alberta (N=373), Manitoba (N=267), Ontario (N=197), and Quebec (233). Black dots represent the approximate sampling location.

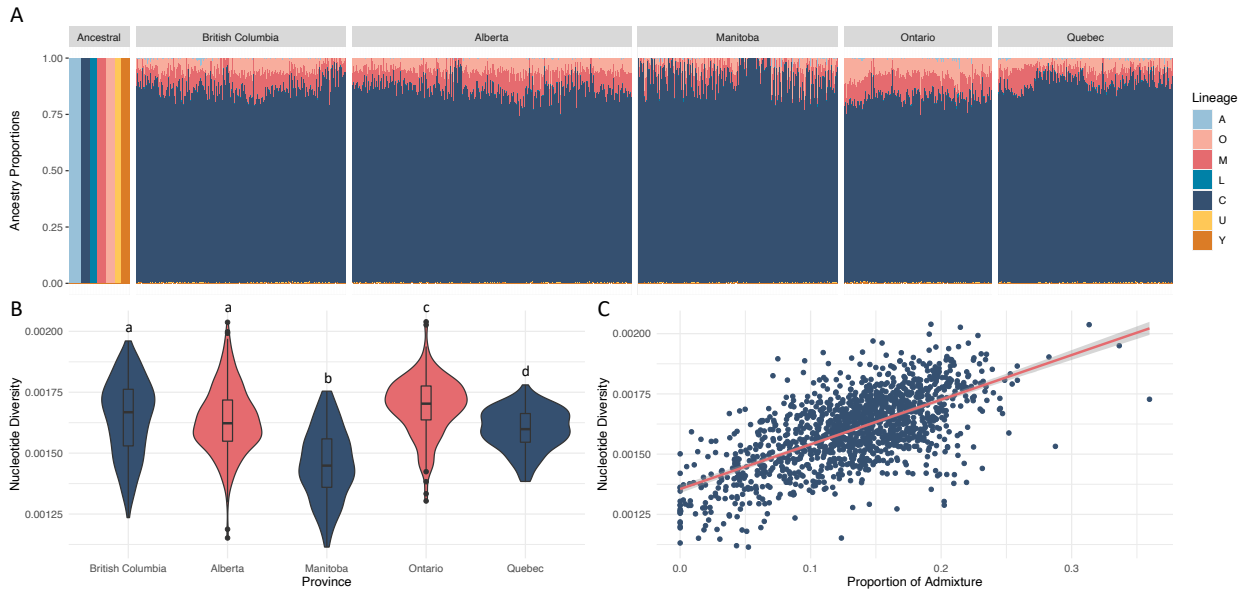

Fig. 3.2: Ancestry and diversity measures of Canadian honey bee colonies. A) Proportion of ancestry in each Canadian honey bee colony. Vertical bars represent individual colonies and coloured segments represent the proportion of ancestry attributed to different lineages. Samples under the ancestral heading represent the reference lineage bees: the A, L, and U lineages in Africa, the Y and O lineages found in Western Asia, and the C and M lineages found in Europe. The remaining samples present the 1350 colonies collected in their respective provinces. B) Distribution of nucleotide diversity (π) among provinces. All pairwise comparisons are significantly different (Tukey $p < 1.38e^{-3}$), except between Alberta and British Colombia (Tukey $p = 0.84$) ($F_{4,1344} = 153.5$, $p < 2e^{-16}$). C) There was a significant, positive correlation between the proportion of admixture and nucleotide diversity among Canadian colonies; $r = 0.65$, $p < 2.2e^{-16}$.

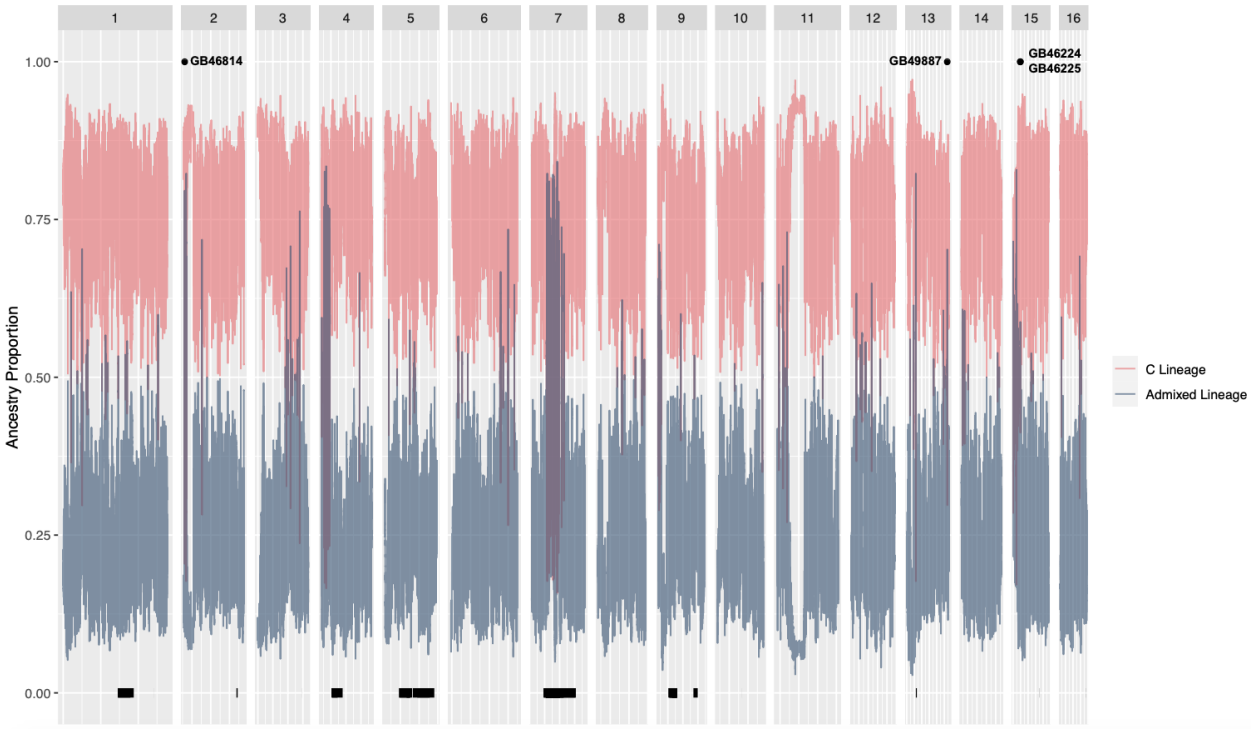

Fig. 3.3: Local ancestry mapping of Canadian honey bee colonies. Average proportion of ancestry of SNPs across managed Canadian honey bee colony genomes. SNPs have been pruned to every four loci to aid with graphing. Top red line indicates the proportion of ancestry contributed by the C lineage and the bottom blue line indicates the proportion of ancestry contributed by the M and O lineages combined (Admixed Lineages). Regions where admixed ancestry is high and C lineage ancestry is low is represented by the purple overlap. Black bars at the bottom of the figure represent the locations of QTLs related to *Varroa* tolerance, as approximated in Mondet et al. (2020a). Black dots at the top of the figure represent the approximate location of some genes of interest.

Table 3.1: Measures of nucleotide diversity (π). Values for Canadian colonies represent the average and standard deviation of nucleotide diversity for colonies.

Group	π
British Columbia	0.00164 ± 0.00016
Alberta	0.00163 ± 0.00013
Manitoba	0.00145 ± 0.00014
Ontario	0.00169 ± 0.00012
Quebec	0.00160 ± 0.00008
C lineage	0.00124
M lineage	0.00105
O lineage	0.00110
A lineage	0.00144
L lineage	0.00092
U lineage	0.00066
Y lineage	0.00058

Supplementary Text:

Estimating the consensus genotype

Pooled genome sequencing is an effective way to sample and sequence DNA from several individuals from a population. While this technique captures the diversity within a honey bee colony, it poses a challenge for population genomics as individual genotypes can't be detected. In our dataset, SNPs are identified as allele frequency estimates based on the aligned sequence reads. Ancestry mapping requires knowledge of genotypes, and as such, we had to estimate a consensus genotype for each colony using simulations (see methods, Fig. S3.1). Simulation revealed that when multiply-mated queens are homozygous for the reference allele at a locus, the consensus of a colony is homozygous reference (0,0). However, once that alternative allele frequency reaches 25% in the population, the consensus genotype of the colony shifts to heterozygous (0,1) (Fig. S3.1). The same pattern is mirrored among colonies where the queen has a homozygous alternative genotype at a locus. The consensus colony genotype is most likely to be homozygous alternative (1,1) up until the alternative allele frequency falls below 75%, at which point the consensus genotype shifts to heterozygous (0,1) (Fig. S3.1). Among colonies where the queen is heterozygous at a locus, the alternative allele frequency ranged between 25-75%, and the consensus genotype is always likely to be heterozygous (1,1) (Figure S1). Thus, loci where the alternative allele frequency was between 0 – 25% was assigned a homozygous reference consensus genotype. Loci that had an alternative allele frequency between < 25 – < 75%, were assigned a heterozygous consensus genotype. Finally, loci whose alternative allele frequencies was between 75 – 100% were assigned a heterozygous consensus genotype.

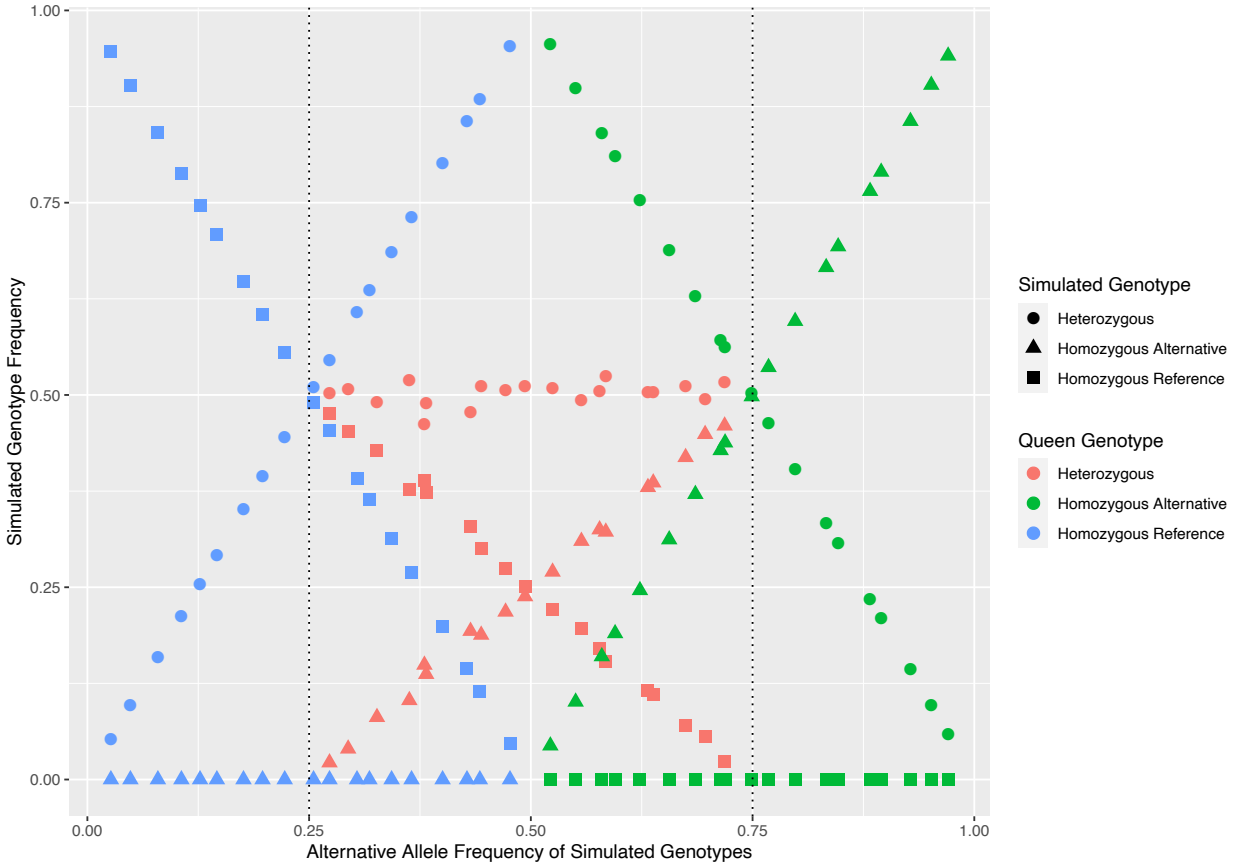

Fig. S3.1: Multiple mating simulations. Simulations of expected colony genotypes based on three possible queen genotypes and different drone allele frequencies based on a pool of 20 individuals. Queens were assigned a genotype of homozygous reference (0/0), heterozygous (0/1), or homozygous alternative (1/1), while drones genotypes were randomly selected from a pool of 20 where the alternative allele frequency was q=20-k/20, where k was equal to 1-19. For example, if the queen genotype was homozygous reference (blue) and could mate with a drone gene pool of 20 individuals with alternative allele frequency q=20-k/20, we could expect colony genotypes (simulated genotypes) and the alternative allele frequency to occur at the given frequencies (y- and x-axis respectively). The consensus genotype among colonies with a homozygous queen will be homozygous when the alterative allele frequency of the colony is below 25%. Above 25%, the consensus genotype is more likely to be heterozygous.

Table S3.1: Measures of pairwise genetic differentiation calculated as F_{ST} between provinces.

	British Columbia	Alberta	Manitoba	Ontario	Québec
British Columbia	-				
Alberta	0.0014	-			
Manitoba	0.0041	0.0041	-		
Ontario	0.0036	0.0040	0.0073	-	
Québec	0.0068	0.0073	0.0084	0.0063	-

Chapter four: Accurate detection of Africanized bees using a SNP-based diagnostic assay.

Introduction

Pollinating insects are critical to the ecological and economic stability of natural and agricultural systems, and it is estimated that they contribute billions of dollars annually in pollination services (Hanley et al., 2015, Gallai et al., 2009, Khalifa et al., 2021). The western honey bee (*Apis mellifera*) is universally identified for their prolific pollination abilities, and subsequently has been introduced globally for commercial use. However, declines of managed honey bee colonies have been extensively reported (Potts et al., 2010, Pettis and Delaplane, 2010, Smith et al., 2013, Gray et al., 2020), and in North America there is additional concern over the displacement of colonies by expanding hybrid Africanized bee populations (Huxel, 1999, Lin et al., 2018). Therefore, the ability to accurately identify and track the movement of Africanized honey bees is a high priority for the beekeeping industry.

The western honey bee (*Apis mellifera*) is native to Europe, Africa and parts of Asia and can be delineated into at least seven genetically distinct groups (Dogantzis et al., 2021). This includes the M-lineage of Eurasia, the C-lineage of Europe, the O and Y-lineages in western Asia, and the A, L and U-lineages of Africa (Dogantzis et al., 2021). The introduction of *Apis mellifera* to North America is suspected to have occurred prior to the end of the sixteenth century, and to parts of South America as early as the eighteenth century to as late as the twentieth century (Kent, 1988, Carpenter and Harpur, 2021). Such introductions began primarily with *A. m. mellifera* and *A. m. iberiensis* imported from western Europe (M-lineage) (Sheppard, 1989, Kent, 1988). Subsequent introductions were followed with *A. m. ligustica* and *A. m. carnica* from the C-lineage, and *A. m. caucasica* from Western Asia (O-lineage) (Kent, 1988, Sheppard, 1989). While apiculture quickly grew in North America, South American beekeepers found it difficult to establish hives and were dissatisfied with the low productivity of temperately adapted European ancestry bees (Kent, 1988). Consequently, in 1956, tropically adapted subspecies *A. m. scutellata* (A-lineage) was introduced to Brazil with the intention of interbreeding them with previously introduced populations to establish a tropically adapted honey bee strain (Kerr, 1967). Soon after the initial introduction, *A. m. scutellata* queens and

drones were unintentionally released into the natural environment and hybridized with local populations producing feral 'Africanized' bees (Winston, 1992).

Over the past 60 years hybrid populations of Africanized bees have rapidly spread across South America to Northern Argentina (Porrini et al., 2019) and through central America into the southwestern United States (Rangel et al., 2016, Kono and Kohn, 2015). Today, Africanized honey bees are the most abundant managed and feral honey bee strain across this region. The rapid and successful expansion of Africanized bees has been attributed to a combination of ecological and behavioural factors that contribute to higher fitness in Africanized bee populations relative to European strains (Schneider et al., 2004, Winston, 1992). For example, Africanized bees have retained many of the behavioural and physiological traits prominent among African (A-lineage) subspecies, including faster colony growth, and greater tendency to abscond and swarm (Schneider et al., 2004, Winston, 1992). Aggressive colony defense, a notorious trait among Africanized bees, is enhanced through hybridization with existing European strains (Harpur et al., 2020, Zayed and Whitfield, 2008). These traits, while adaptively advantageous in tropical habitats, make these populations unpopular for apiculture.

During their expansion in the Americas, Africanized honey bees are thought to have displaced or hybridized with previously abundant European colonies, radically changing the ancestral genetic composition of colonies (Pinto et al., 2005, Rangel et al., 2016). Established Africanized honey bee populations are comprised largely of African ancestry, which represents on average 75% of individual genetic composition (Chapman et al., 2015, Nelson et al., 2017, Zayed and Whitfield, 2008). The remaining ancestral proportions originate primarily from the M-lineage, with some contribution from the C-lineage (Chapman et al., 2015, Nelson et al., 2017). However, ancestral proportions are variable across the population distribution. This is especially prevalent within hybrid zones located at the Northern and Southern limits of expansion where proportion of African ancestry exhibits a cline between 5% to 77% (Calfee et al., 2020).

The rapid and dynamic invasion capabilities of Africanized bee populations demonstrate that this strain continues to pose a threat to regions currently free of Africanized honey bee genetics. To avoid the incorporation of undesirable behavioural phenotypes into commercial colonies, several countries, including Canada and Australia, have placed import restrictions from regions with known Africanized honey bees. Though there is evidence to suggest that

Africanized honey bees may have reached their range limit at temperate latitudes (Calfee et al., 2020, Porrini et al., 2019), changes in the environment could improve habitat suitability for Africanized bees, thus promoting the spread of invasive populations (Jarnevich et al., 2014, Stohlgren et al., 2014, Gill and Sangermano, 2016). Recent studies have already shown that Africanized bees are slowly expanding their distribution into regions that serve as queen breeding hubs for North America (Cridland, et al. 2018; Lin, et al. 2018). Without the ability to accurately detect and track the movement of Africanized bees, there remains a risk of accidently importing populations, especially from unknown recently colonized regions.

Traditional methods of identifying Africanized bee samples can be inaccurate and have the potential to misidentify samples. Errors are often a result of the variability in ancestry proportion of Africanized bees, which can confound conclusions based on morphology (Guzmán-Novoa et al., 1994) and maternally inherited mitochondrial DNA sequences (Sheppard and Smith, 2000). However, current diagnostic tools using single nucleotide polymorphisms (SNP) have demonstrated excellence in differentiating honey bee subspecies, for example (Henriques et al., 2018a, Henriques et al., 2018b, Parejo et al., 2016, Muñoz et al., 2017, Muñoz et al., 2015, Pinto et al., 2014). Recently, diagnostic assays using 95 and 37 SNP loci were developed to identify Africanized bees by estimating the proportion of African lineage ancestry (Chapman et al., 2017, Chapman et al., 2015). While these SNP assays are a significant improvement over the traditional methods for detecting AHB, they have some drawbacks: 1) The current 96 and 37 SNP diagnostic assays were developed using only three out of the seven ancestral lineages, and consequently, do not benefit from recent large scale genomic datasets on *Apis mellifera* subspecies (Dogantzis et al., 2021). 2) SNPs for these assays were not chosen based on an information criterion, and recent studies have shown that markers selected by information content outperform randomly selected SNPs (Muñoz et al., 2017). 3) The reliance on ancestry proportion thresholds for detecting Africanization can be confounded by the variance in ancestry levels among hybrid samples, especially within hybrid zones (Calfee et al., 2020).

Novel approaches, such as the use of machine learning, could greatly increase the efficacy of diagnostic assays. The use of machine learning in population genomic analyses is an emerging paradigm but has already demonstrated success in identifying selective sweeps, inferring demographic histories (Schrider and Kern, 2018), and has been used to discern *Apis mellifera* subspecies (Momeni et al., 2021) Notably, supervised machine learning algorithms,

which utilize prior knowledge to make predictions about new datapoints, are ideal for classification tasks. Here, we present an improved diagnostic assay designed to differentiate African ancestry and Africanized bees from European lineages and commercial honey bees, by employing two different supervised machine learning algorithms. The first aim of this study involved categorizing native honey bee samples into their respective ancestral lineages to determine the genetic composition of Africanized and managed honey bees. This initiative helped determine ancestry informative markers that may be effective in discriminating lineages. Next, we constructed a random forest classifier to subsample perspective loci and rank markers based on their informativeness for effectively differentiating Africanized and African lineage bees from European and commercial honey bees. A diagnostic assay was constructed based on 113 informative loci and was validated using 1263 honey bee samples collected from North America, South America, and Australia. Classification of samples based on genotyping results was estimated with a support vector machine classifier, which estimates the classification probability of a sample to a predetermined group. Overall, the diagnostic assay provides an accurate means for identifying individual honey bee samples and has the potential to provide accurate results with a reduced set of informative markers.

Methods:

Genome sequence processing and SNP detection

Our dataset consists of 243 previously sequenced honey bee genomes, in addition to newly sequence genomes representative of 16 hybrid Africanized bees, and six commercial North American honey bees (N=265). Sample preparation and genome sequencing of new hybrid Africanized bee samples follows a previously published protocol (Dogantzis et al., 2021). Similarly, sample preparation and genome sequencing of North American samples follows published protocols (Harpur et al., 2014). Sequence reads were trimmed of Illumina adapters and low quantity bases (<20) using Trimmomatic v0.36 (Bolger et al., 2014), and retained for downstream assembly if >50bps and >35bps in length from 100-150bp and 50bp Illumina sequencing data respectively. Reads were aligned to the *Apis mellifera* reference genome (Amel 4.5) (Elsik et al., 2014) using NextGenMap aligner v0.4.12 (Sedlazeck et al., 2013). BAM files were sorted using SAMtools v1.3.1 (Li et al., 2009) and reads were marked for duplicates using Picard v2.1.0 (https://broadinstitute.github.io/picard/). Base quality scores were recalibrated

using GATK v3.7 BaseRecalibrator (Van der Auwera et al., 2013) using previously identified variants as reference (Harpur et al., 2014, Harpur et al., 2019). SNPs were identified with GATK v3.7 (Poplin et al., 2017, Van der Auwera et al., 2013) HaplotypeCaller and filtered using VariantRecalibrator using previously identified variants as reference (Harpur et al., 2014, Harpur et al., 2019) in addition to the following hard filter thresholds: MQ < 40.0, QD < 5.0, FS > 11.0, MQRankSum -2.0 < x > 2.0, and ReadPosRankSum -2.0 < x > 2.0. Variants were also excluded if they were situated within five base pairs of an indel or area of low complexity (Harpur et al., 2019), or were located on unmapped scaffolds.

Sample processing for SNP Genotyping

To validate the diagnostic assay, 1263 samples of putatively known Africanized and non-Africanized commercial honey bees were collected from North America, South America, and Australia, including 838 previously analysed Canadian honey bee samples (Harpur et al., 2015). For newly collected samples, DNA extraction was performed using Mag-Bxind® Blood & Tissue DNA HDQ 96 Kit (Omega Bio-tek Inc., USA) optimised for KingFisher™ Flex Purification System (Thermo Fisher Scientific Inc., USA). For tissue lysis, either half or whole bee thoraces were flash frozen in liquid nitrogen and finely ground using a pestle. We then added 350µl Tissue Lysis Buffer, 20µl Proteinase K, and heated samples overnight at 55°C. After processing with the KingFisher System, samples were eluted in nuclease-free water (Thermo Fisher Scientific Inc., USA) to a final volume ranging from 50–80µl. DNA was quantified using NanoDrop™ 2000 Spectrophotometer (Thermo Fisher Scientific Inc., USA). DNA quality was assessed with 1.0% agarose gel electrophoresis. SNP genotyping was outsourced to Genome Quebec Innovation Centre (Quebec, Canada). This dataset was supplemented with an additional 29 reference Africanized honey bee genomes whose corresponding genotypes were extracted from published variant data (Kadri et al., 2016).

Population structure

ADMIXTURE v1.3.0 (Alexander and Lange, 2011) was used to estimate ancestry proportions and population structure of the 265 honey bee genomes. This analysis was performed with 1M randomly selected bi-allelic markers with a minor allele frequency of >0.10 among at least one of the predicted subspecies (see Dogantzis et al. (2021)). ADMIXTURE was run with

predicted K values 1-18 using the 10X cross validation procedure. A principal component analysis (PCA) was generated to examine the genetic relatedness among lineages in relation to Africanized honey bee and commercial North American populations. The PCA was constructed with the SNPRelate (Zheng et al., 2012) package in R v3.6.0 (R Core Team, 2013) using SNP markers with a minor allele frequency of >0.10 among at least one of the predicted subspecies, amounting to 3,488,846 markers.

SNP selection

To identify a set of informative genetic markers from which to construct the diagnostic assay, we calculated pairwise measures of F_{ST} (Weir and Cockerham, 1984) between genetically distinct honey bee lineages using VCFtools v0.1.17 (Danecek et al., 2011). SNPs of interest were bi-allelic and contained <5% missing data across all genome samples (N=392,389). Missing data for individuals was imputed using the consensus genotype representative of the lineage of origin. Final genotypes were coded as "0" representing homozygous reference genotypes, "1" representing heterozygous genotypes, and "2" representing homozygous alternative genotype calls. Using the reduced dataset, we used a random forest classification model (Breiman, 2001) to determine the importance of SNP markers for classifying samples as Africanized or African, relative to non-Africanized and non-African in origin. To train the random forest classifier, we divided the 265 honey bee genomes into a training group and a testing group, which contained 177 (66%) and 88 (33%) samples respectively. We ensured the testing and training group has an approximately equal proportion of samples from each lineage, Africanized population, and North American population. We used the GridSearchCV option as implemented in scikit-learn Python package (Pedregosa et al., 2011) to determine the optimal parameters of the random forest classify, including n_estimators, max_features, and max_depth. The random forest classifier, using optimal parameters, was run for 30 replicates using the reduced SNP marker dataset as implement by the scikit-learn Python package (Pedregosa et al., 2011). The feature importance for each replicate run was estimated for the top 100 markers using the feature_importances_ option as implement by the scikit-learn Python package (Pedregosa et al., 2011). Overall, there were 824 markers of interest that were present two or more times across replicates or were ranked among the top 75th percentile of the scored features. The testing samples were used as an

independent validation of the predictive accuracy of the selected markers in the SNP panel prior to final development.

SNP panel development and validation

We developed a three panel Agena iPLEX Gold SNP array estimated to hold approximately 120 SNP markers. To increase the design success of the panel, markers were submitted for inclusion on the array if they were free of secondary SNPs at least 16bp up or downstream of the target loci and were >5000bp apart from other informative loci to reduce linkage disequilibrium. Of the 824 top ranking SNPs, 249 markers fit the preceding requirements, and 113 markers were successfully designed for the panel. Panel design, production, and validation, in addition to oligo design were completed at the Genome Quebec Innovation centre (Quebec, Canada).

To validate the assay, we genotyped several putative Africanized bee samples from Brazil and the United States, in addition to several managed commercial and feral honey bees from the United States, Australia, and Canada (Harpur et al., 2015) (N=1263) (S1 Table 2). All samples were genotyped at 113 loci and were analysed using a linear support vector machine classification model (linear SVC). The model was initially trained using the training dataset (n=177), and then was tested on the unknown genotyped samples (N=1263) and the previously genome sequenced honey bees (n=117) (n=88 testing samples and n=26 Africanized honey bees). To run the model, genotypes were coded as "0" representing homozygous reference genotypes, "1" representing heterozygous genotypes, and "2" representing homozygous alternative genotype calls. The SVC was trained using a linear kernel with the GridSearchCV option to estimate optimal parameters for C, kernel, and gamma as implement by the scikit-learn Python package (Pedregosa et al., 2011). Classification probabilities of unknown samples to an Africanized or African origin relative to a non-Africanized and non-African origin was computed with the predict_proba option using the scikit-learn Python package (Pedregosa et al., 2011). Unfortunately, machine learning algorithms do not support data with missing values, as such, all missing genotypes where imputed conservatively as "2", representing the genotype associated with an African or Africanized origin.

Results:

Ancestry Estimation and Lineage assignment:

Our dataset comprises 243 previously sequenced honey bee genomes, which were used as refence lineages in our population structure analyses to estimate the ancestral proportions of Africanized and commercial honey bee samples. Genetic clustering results produced by the program ADMIXTURE, illustrates that when K=7, native honey bee samples clustered into previously identified lineages, including *A. m. unicolor* (U-lineage) from Madagascar and *A. m. lamarckii* (L-lineage) from Egypt (Dogantzis et al) (Fig. 4.1). When the structure analysis is conducted with K predictive values 3-6 (Fig. S4.1), we observe gradual separation of clusters with increasing K between the C and O lineages and the A, L, U, and Y lineages. These patterns are also reflected in the PCA analysis which depicts the proximate clustering of the A, L, U and Y lineages together, while the C and O lineage cluster closer together in space (Fig. S4.2). While most samples have a definitive lineage assignment (>80% to a single ancestry), the structure results highlight that several individual honey bees have low to moderate levels of admixture, likely a result of hybridization with geographically neighbouring lineages.

The structure analyses were also used to study the genetic composition of Africanized and North American honey bee samples (Fig. 4.1). Individuals from the Africanized bee population had a large portion of their ancestry originating from the A-lineage, representing on average 82.7% of the predicted ancestral proportion. The remaining genetic composition was comprised of M-lineage ancestry and C-lineage ancestry contributing on average 15.3% and 0.02% respectively. The PCA results also emphasize the extensive introgression of A-lineage ancestry into Africanized bees, depicted by the proximate clustering of these groups (Fig. S4.2). North American samples can also be classified as admixed, with an average of 73.1% of ancestry originating from the C-lineage, and M and O lineage contributing an average of 12.9% and 11.0% respectively. Similarly, the North American population clusters most closely with the C-lineage samples in the PCA analysis, reflective of shared ancestral origins (Fig. S4.2).

Given the clustering results from the ADMXITURE and PCA analyses, the A, U, Y and L lineage were determined to be more ‘Africanized’ or ‘African’ in their genetic composition, while the M, C and O lineages were determined to be closer to the genetic composition of “commercial” honey bees or “non- Africanized” and “non-African” honey bees.

Marker Selection and SNP Panel Design:

To select an informative set of diagnostic markers for the SNP panel we used a two-step selection process. First, we calculated pairwise measures of F_{ST} (Weir and Cockerham, 1984) between genetically distinct honey bee lineages. We specially focused on identifying loci that where highly differentiated ($F_{ST} > 0.8$) between the African (A) lineage and the remaining lineages as Africanized bees are comprised predominantly (>75%) of A-lineage ancestry. In total, we identified 392,389 SNP loci of interest. To narrow down the list of markers, we used a random forest classification model (Breiman, 2001) to determine the importance of SNP markers for classifying samples as Africanized or African, relative to non-Africanized and non-African in origin. A random forest is a supervised learning algorithm that operates by constructing multiple decision trees using a random subset of the given features (SNPs) and training points (samples) drawn with replacement (Breiman, 2001). The model can also evaluate the importance of the subsampled features for differentiating the target groups. After running the model over 30 replicates, we identified 824 markers of interest. Markers were then assessed for inclusion on the SNP panel per developer requirements, after which 113 markers were successfully designed for the panel.

Panel Validation and Probability Estimates:

To validate the diagnostic capabilities of the panel, we genotyped 1263 honey bee samples collected from South America, North America, and Australia. The final diagnostic SNP panel consisted of 113 informative markers, of which 87 could be successfully genotyped. We removed five loci due to a low call rate (< 70%), and an additional two markers due to monomorphic genotype calls (N=80). Based on the genotypes assigned by the SNP panel, we used a linear SVC to estimate the classification probability of samples to an Africanized or non-Africanized assignment. A linear SVC is a supervised learning model that classifies data by applying a best fit hyperplane to maximize the margin between groups (Cortes and Vapnik, 1995). The model was trained on the original 177 training samples and was then tested on the 1263 SNP-genotyped samples, in addition to 117 known samples whose genomes were previously sequenced. Metrics of model performance can be found in the supplementary material (Fig S4.3).

The commercial Canadian samples were assigned as non-Africanized with probabilities ranging from 98.5% to 89.9% (Table 4.1). The exception was one sample that had a predicted probability of 61.6% (not shown); likely due to 10 markers that failed genotyping and were conservatively imputed with an African genotype since SVC modeling cannot use missing data. When the SVC model is trained and tested without the missing loci, the classification probability of this Canadian sample to a non-Africanized assignment is 81.9%. This sample was removed from the proceeding analyses due to low genotyping coverage. The commercial and feral Australian honey bee samples were classified as non-Africanized with a 98.5% to 91.7% probability (Table 4.1). Honey bee samples from Brazil, presumed to be Africanized, were classified as Africanized with a 99.4% to 96.6% probability. We detected a wide variance in classification probability for feral populations in North America. For feral populations in Texas, the average probability to an Africanized classification was 82.9%, but ranged between 97.7% to 2.9%. Samples from feral populations in California had a higher average probability to a non-Africanized classification (58%), which ranged from 72.4% to 49.2% (Table 4.1).

Sample Classification:

Our classification model assigns individual bees to an Africanized or non-Africanized assignment using probability estimates. Designating a definitive classification to samples for monitoring can be done in two ways. First, samples can be classified based on how their probability estimates compare to known samples of an Africanized or non-Africanized group, and second, samples can be categorized based on an overall threshold, where samples above get placed into a non-Africanized group, while samples below get placed into an Africanized group. Here, we determine the accuracy of the diagnostic assay by measuring the false negative and false positive rate of each categorization approach. To estimate the false positive and false negative rate we used only reference samples, commercial samples, Africanized bees from Brazil, and feral samples from Australia as these bees show the closest measures of probability to reference samples. We used the feral Texas and California bees to apply the categorization methods, as these bees show the most variance in probability estimates and whose origins are not entirely known.

Using the first approach, if a strict probability threshold of 95% probability for a non-Africanized or Africanized classification is used, we identify a 2% false positive rate, and a 0%

false negative rate respectively (Fig. 4.2A). When a minimum threshold of 88% is used, the false positive and false negative rate is 0% for both groups. To provide some leeway in categorization, a threshold of 80% should provide enough confidence of classification to both groups while also taking into consideration genetic variance not captured in this study. Using the 80% threshold, 80% of feral honey bees from Texas were classified as Africanized, while 4% were classified as non-Africanized (Fig. 4.2A). All samples from California, and 14 samples from Texas (16%) did not fit into either an Africanized or non-Africanized classification and were considered unclassified. Given the uncertainty of an unclassified classification, these samples should conservatively be considered Africanized.

In the second categorization approach, we make the broad assumption that all samples are Africanized until proven otherwise. As such, setting a threshold for non-Africanized probability would classify all samples above the threshold as non-Africanized, while all samples below that threshold would be considered Africanized. If a strict threshold of 95% probability for a non-Africanized classification is used, there is a 2% false positive rate and 0% false-negative rate. When an 80% threshold for non-Africanized classification is used, there are no false-positives or false-negatives (Fig. 4.2B). When applying the 80% non-Africanized threshold to the feral bee samples, 3% of samples were classified as non-Africanized, while 97% of all feral were classified as Africanized (Fig. 4.2B; Fig. S4.3).

Reduction in Diagnostic Loci:

While the diagnostic panel genotypes at 80 informative markers, there may be circumstances where the number of loci available for classification is reduced. To assess how a reduction in diagnostic markers would affect the classification estimates of samples, a linear SVC was trained based on a random subset of 10-70 loci. The model was then tested on 117 known reference samples and 694 validation samples that were originally successfully genotyped across all 80 loci (n=91 Africanized samples, n=720 non-Africanized samples). Since we are using a random subset of diagnostic markers, this process was replicated five times across each subset to capture the variability in loci. When ≥30 diagnostic markers are used, the model estimated a classification probability greater than 80% for all Africanized samples, and a classification probability above 80% for all but two non-Africanized samples (Fig. S4.4). Using an 80% threshold, these results suggest a false negative rate of zero (n=2275) for Africanized

samples and a false positive rate of 0.01% (n=18000) for non-Africanized samples. Under extreme circumstances when only 10 loci are used, 13 Africanized samples are assigned a probability estimate below 80% (3% false negative rate, n=455), while only two non-African samples are assigned a probability estimate below 80% (0.06% false positive rate, n=3600) (Fig. S4.4).

Effect of Imputation on Classification:

The assisted machine learning models used in these analyses cannot use missing data as a feature, therefore, missing genotypes are coded conservatively as the representative genotype for Africanized or African ancestry ("2"). To assess the effects of imputation on the predicted probability estimates, 1 to 80 of the informative markers was randomly assigned a "2" genotype across 29 reference samples and 691 validation samples of known non-Africanized or non-African origin. Using a linear SVC trained with 177 training samples, the average predictive probability of an Africanized assignment was determined for each increase in the number of randomly assigned African ancestry genotypes. We found that when ≥43 markers are imputed, more than half of all samples are given a probability estimate of 80% or greater to an Africanized classification (59% false positive), and when ≥49 markers are imputed, all samples are classified as Africanized with >80% probability (100% false positive) (Fig. S4.5). On the lower end, we find that when 9 or few markers are imputed, all samples are classified as non-Africanized with a probability above 80% (Fig. S4.5). When 10-17 markers are imputed, we find that, on average, three samples fall below the 80% threshold for a non-Africanized classification. The probability of these samples ranged from 79.9% to 63.8%, averaging 74.5%, and resulted in an average false positive rate of 0.4% (Fig. S4.5). We detect a considerable change in false positive rate (23.3%) when ≥20 markers are imputed (Fig. S4.5). Thus, while sensitivity to imputation is sample specific, 9 or fewer imputations should not affect classification, and the imputation of <19 markers should not produce a substantial number of false positives (less then 3.3%).

Discussion:

As Africanized bees continue to expand their range, diagnostic tools that minimize classification uncertainty are needed to improve the detection and monitoring of these hybrid populations. Here, we developed a SNP diagnostic tool that when combined with assisted

machine learning classification can consistently and effectively differentiate known Africanized and commercial honey bees. The classification between samples is made with high probability estimates (>88%), resulting in a false positive and false negative rate of zero when a probability threshold of <88% is used. However, to remain conservative and take into consideration the genetic variability of samples not captured in this study, we recommend the use of a threshold of 80% for classification of unknown samples.

The accuracy of the diagnostic assay can be attributed to several factors considered during the design process. Previous studies have demonstrated that diagnostic assays constructed with SNPs chosen using an information criterion outperform those that employ randomly chosen markers (Muñoz et al., 2017). In this study, SNPs were evaluated on their discriminant power based on measures of F_{ST} (>0.8) and feature importance estimated with a random forest classifier. Previous studies have repeatedly shown that loci with high measures of F_{ST}, specifically fixed differences between target populations, are substantially advantageous for population discernment (Henriques et al., 2018b, Chapman et al., 2015, Willing et al., 2012, Muñoz et al., 2015). Markers that are highly differentiated are less likely to be lost to genetic drift and are more likely to reach fixation in a population. The random forest classifier added an extra measure of scrutiny to markers by measuring SNP informativeness – the ability of a SNP to categorize Africanized and African ancestry lineages from non-Africanized and non-African lineage bees.

We used a linear support vector machine classifier to predict the classification of individual honey bees based on probability estimates. This approach is advantageous over previous methods that rely on the proportion of A-lineage ancestry to identify Africanized bees. When testing samples of known commercial or Africanized origins, the model can classify samples into the correct group with >80% probability. However, we do find that in some cases, samples with unknown origins, such as those of feral bees from areas with known Africanized honey bee colonies, show a wide variance in probability estimates (97.7% - 2.9% Africanized assignment). The broad range of estimates suggest that this cohort contained Africanized honey bee samples, non-Africanized honey bee samples, and several samples that had intermediate levels of African introgression – likely the result of backcrossing with commercial strains. Although the model didn't definitively classify all samples to a single group (<2% overall, or 25% of feral bees), we can still categorize these bees as Africanized if they fall below the 80%

threshold required for a non-Africanized classification. Thus, we strongly recommend the use of the second classification system, which uses an overall threshold to categorize samples. Since our methodology integrates assisted machine learning models, over time additional reference samples of confirmed Africanized and non-Africanized origins, including those with variable levels of hybridization, can be added to improve classification. At present, we are limited to the training set produced in this study which does not currently reflect the true variance found across honey bees.

Issues with genotyping can complicate machine learning classification tasks through a reduction in data. To avoid the issue of missing data, we recommend removing loci if a significant portion of samples are affected. We demonstrate that reduced subsets of SNP markers retained the ability to classify samples with high predictive probabilities. Notably, when as low as 30 markers are used, all Africanized and all but two non-Africanized samples are assigned to the correct classification with a probability above 80%, equating to a 0% false negative and 0.01% false positive rate. The development of reduced SNP panels (Muñoz et al., 2015, Henriques et al., 2018a, Henriques et al., 2018b, Chapman et al., 2017) is a common goal since they provide cost savings and reduce the computational demand typically associated with large SNP panels or genome sequencing. Alternatively, if only a few samples are affected by missing genotypes, it is recommended to either omit the sample or conservatively impute genotypes representative of Africanized origins. Imputation on fewer then 20 markers does not adversely affect classification and retains probability estimates above 80%. Additionally, the results from the imputation analysis revealed a conservative bias towards positively identifying Africanized bees. For example, when ≥49 loci are represented by an African ancestry genotype, samples are assigned an Africanized classification ≥80% probability. This conservative bias presents a two-fold advantage by decreasing the chance of a false negatives and being sensitive enough to detect moderate levels of Africanization; greatly decreasing the chance for accidental misclassification and importation of Africanized bees.

In conclusion, we show that 80 markers when combined with machine learning classification can clearly and effectively classify samples by proving a non-ambiguous probability estimate to an Africanized or not Africanized assignment. This is advantageous over previous methods which relied on morphology or ancestry proportions, which can introduce some uncertainly, especially among moderately admixed populations. Furthermore, our results

suggest that a reduced SNP assay, using as few as 30 loci, retains accurate probability estimates. This is especially important for when genotyping may fail, or when cost saving measures are needed to be implemented.

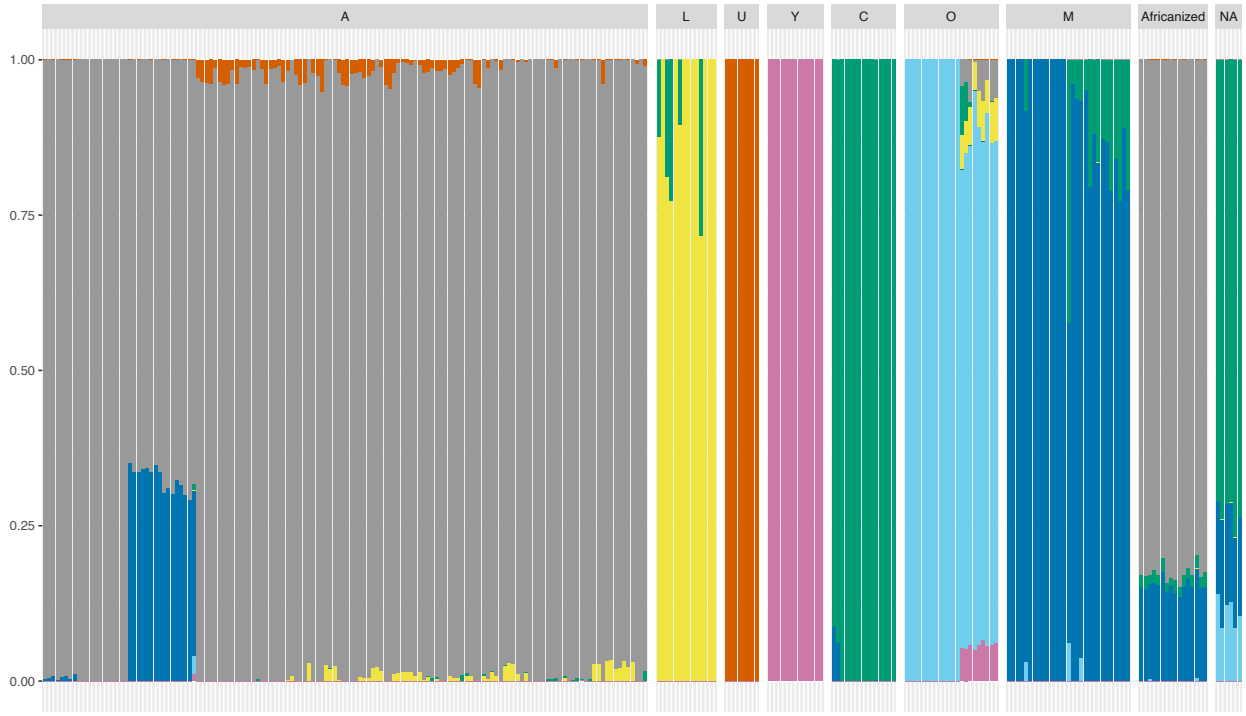

Fig. 4.1: Patterns of admixture and ancestry of ancestral lineages, Africanized honey bees, and North American honey bees. Patterns of ancestry and admixture for N=243 Native *Apis mellifera* samples grouped into their respective lineages, N=16 hybrid Africanized bee samples (Africanized), and N=6 commercial North American honey be samples (NA). Vertical bars represent individual bees and coloured segments represent the proportion of ancestry estimated to K=7 genetic clusters (lineages). Africanized bees possess an average A-lineage ancestry of 82.7%, and M-lineage of 15.3%, while North American honey bees are composed of an average 73.1% C-lineage, and an average of 12.9% and 11.0% to M and O lineage respectively.

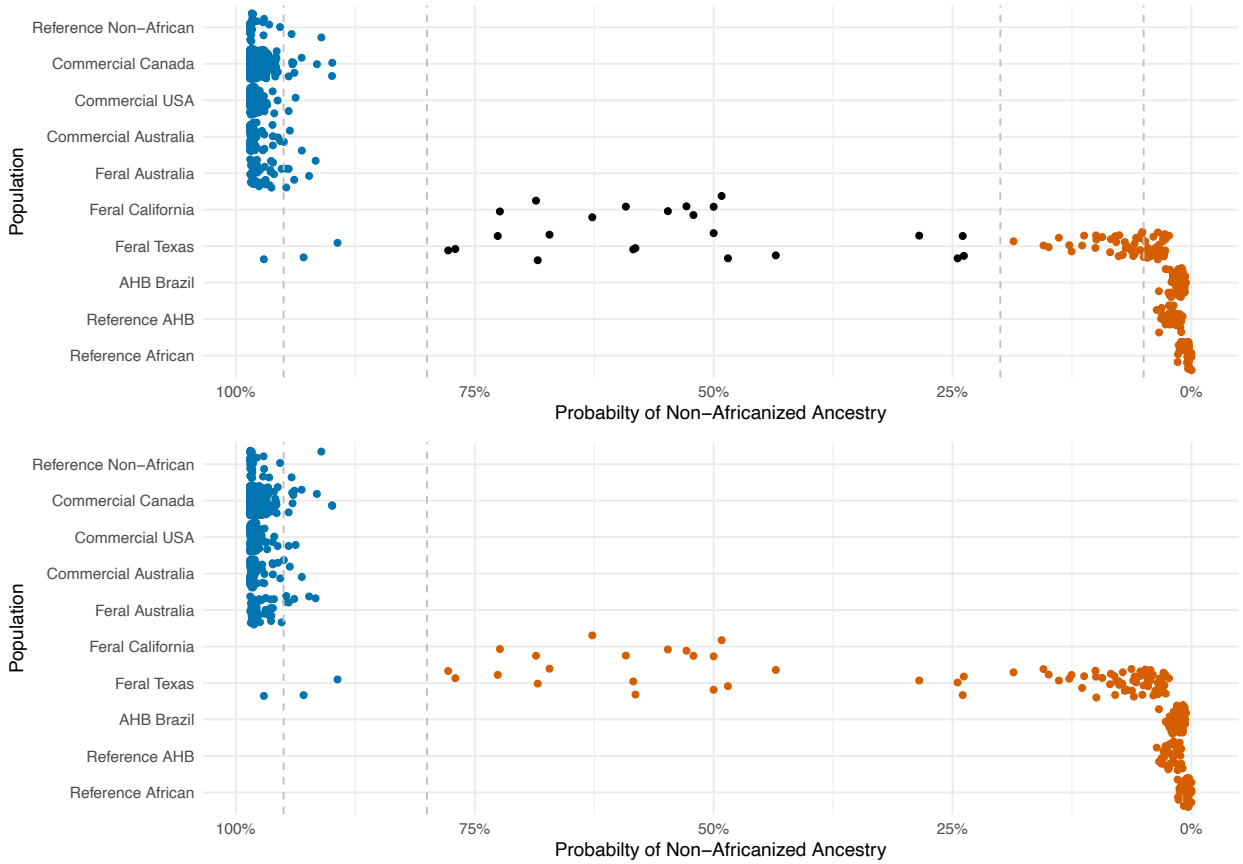

Fig. 4.2: Classification of validation samples using probability thresholds. A) When a threshold of 80% probability to non-Africanized assignment or an 80% (20% non-Africanized) probability to an Africanized assignment is used, there is a 0% false negative and false positive rate among known samples (Reference samples, Commercial Samples, Feral Australia, and Africanized honey bees Brazil). Among unknown samples (N=92; Feral California and Texas), 25% of samples are unclassified (black dots). Dashed lines indicate probability thresholds for a non-Africanized (95% and 80%) or Africanized (20% and 5% non-Africanized) classification. B) When an absolute threshold of 80% is used, all samples above this value are considered non-Africanized, while all samples below this threshold at considered Africanized. Using an 80% threshold produced a 0% false negative and false positive rate among known samples (Reference samples, Commercial Samples, Feral Australia, and Africanized honey bees Brazil). Among unknown samples (Feral California and Texas), 97% of samples are classified as Africanized. The dashed line indicated the 80% probability threshold for non-Africanized ancestry.

Table 4.1: Probability estimates to either a non-Africanized or Africanized classification. The table provides the upper probability estimate, the lower probability estimate, and the mean for each population group.

		Non-Africanized			Africanized		
Population	N	Upper Probability	Lower Probability	Mean	Upper Probability	Lower Probability	Mean
Commercial Canada	840	98.510%	89.907%	98.131%	10.093%	1.490%	1.869%
Reference Non-African	29	98.493%	91.055%	97.686%	8.945%	1.507%	2.314%
Commercial USA	115	98.493%	93.745%	98.027%	6.255%	1.507%	1.973%
Commercial Australia	88	98.493%	93.093%	97.954%	6.907%	1.507%	2.046%
Feral Australia	49	98.493%	91.651%	97.137%	8.349%	1.507%	2.863%
Feral Texas	83	97.065%	2.348%	17.143%	97.652%	2.935%	82.857%
Feral California	9	72.368%	49.165%	57.973%	50.835%	27.632%	42.027%
Reference AHB	33	3.632%	0.920%	2.175%	99.080%	96.368%	97.825%
Africanized Brazil	78	3.400%	0.563%	1.363%	99.437%	96.600%	98.637%
Reference African	55	1.442%	0.002%	0.449%	99.998%	98.558%	99.551%

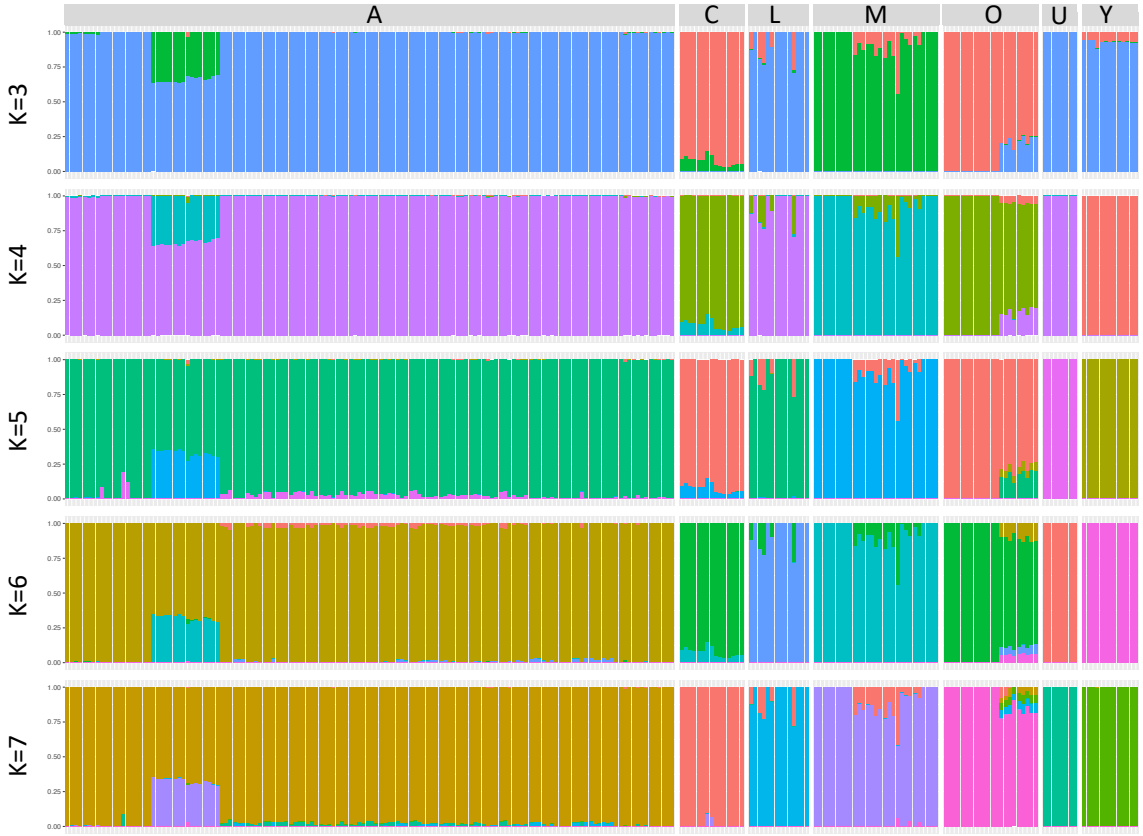

Fig. S4.1: ADMIXTURE results for native honey bee samples. Patterns of ancestry and admixture for all (N=243) native *Apis mellifera* samples, grouped into ancestral lineages, as estimated with the program AMDIXTURE. Vertical bars represent individual bees and coloured segments represent the proportion of ancestry estimated to K=3-7 genetic clusters.

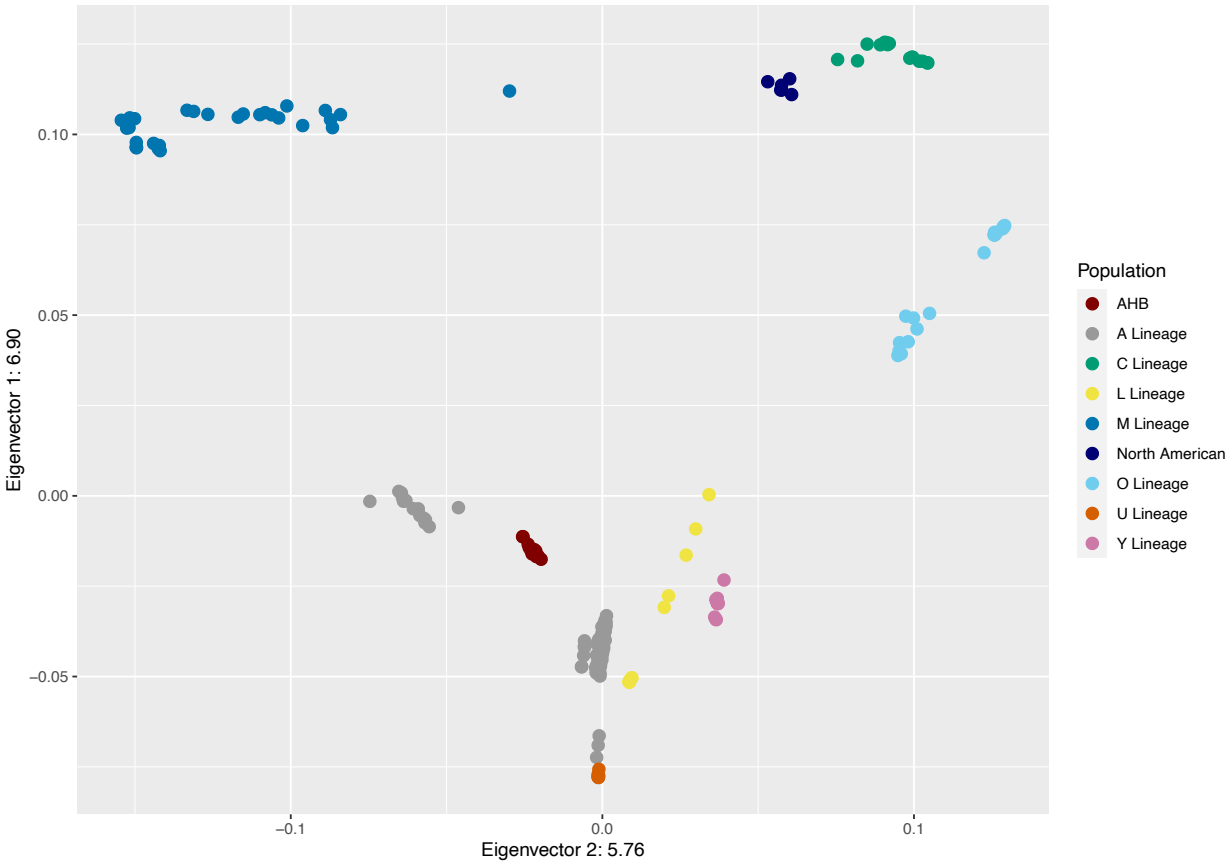

Fig. S4.2: Principal component analysis of honey bee samples. Samples cluster broadly into representative lineages and populations, with admixed samples clustering outside of lineage groups. The Africanized honey bee samples (AHB) cluster closely with A lineage samples, while North American commercial colony samples cluster closely with C lineage samples.

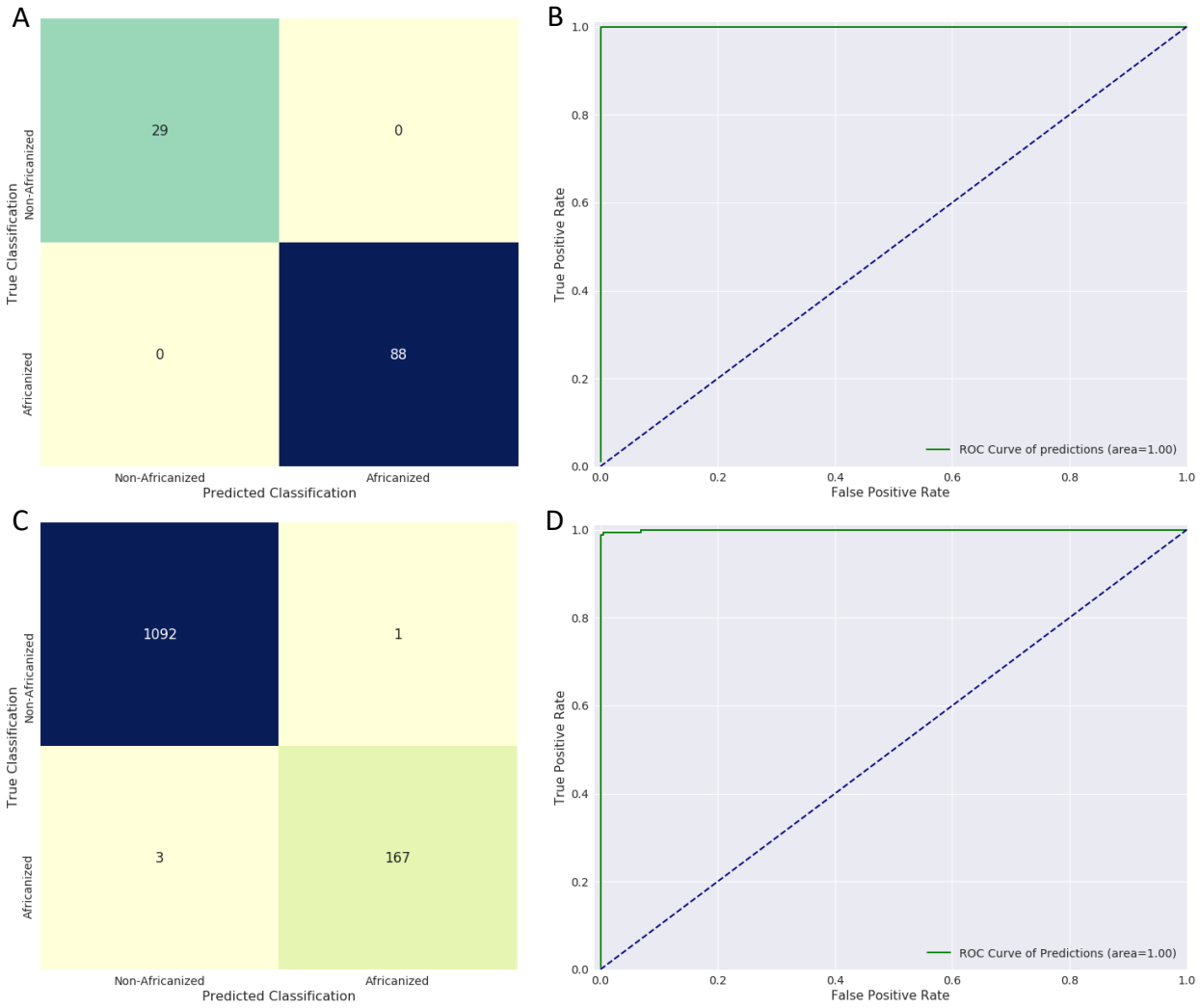

Fig. S4.3: Metrics of model performance of the SVC classifier. A) Figure A depicts a confusion matrix, which shows the predicted classification of samples by the trained model relative to their known classification. We used a classification threshold of 0.8 (80% probability) to a non-Africanized group ("0"), where samples above this threshold were labeled as non-Africanized, and all samples below this threshold were labelled as 'Africanized'. All reference Africanized honey bee samples (N=29) were correctly classified as Africanized, while all reference non-Africanized samples (N=88) were correctly classified as non-Africanized. The average probability for a non-Africanized classification ("0") was 0.977 (N=29), while the average probability for an Africanized classification ("1") was 0.989 (N=88) (0.010 for a non-Africanized classification). B) Figure B depicts the receiver operating characteristic (ROC) curve which illustrates the performance classification of the model at all classification thresholds. The dashed blue line is the performance of a random model, while the solid green line is the ROC

curve for the trained model tested on the reference (N=117) samples. C). Figure C depicts the confusion matrix for the 1263 samples used to validate the model, whose true classification was assumed based on collection location. Here, we assume that commercial honey bee samples from North America and all samples from Australia are likely of non-African origin, while Africanized honey bees form Brazil and feral honey bees from North America are likely of Africanized origin. Samples assigned a model classification >0.8 to a non-Africanized group ("0") were classified as 'non-Africanized', and all samples below this threshold were labelled as 'Africanized'. One non-Africanized sample was misclassified, while three Africanized samples were misclassified. The misclassified non-Africanized sample had a predicted probability of 0.616 but contained several imputed datapoints due to missing genotype calls. The three misclassified Africanized samples had a precited probability of >0.893 to a non-Africanized classification. These samples were collected from Texas where beekeeping with European colonies is prevalent, thus these samples are likely representative of commercial European colonies. Given we are making a priory assumption about the sample's true classification, the model correctly identified the mistake in our categorization. D) Figure D depicts the receiver operating characteristic (ROC) curve for the validation samples (N=1263). The dashed blue line is the performance of a random model, while the solid green line is the ROC curve of the model.

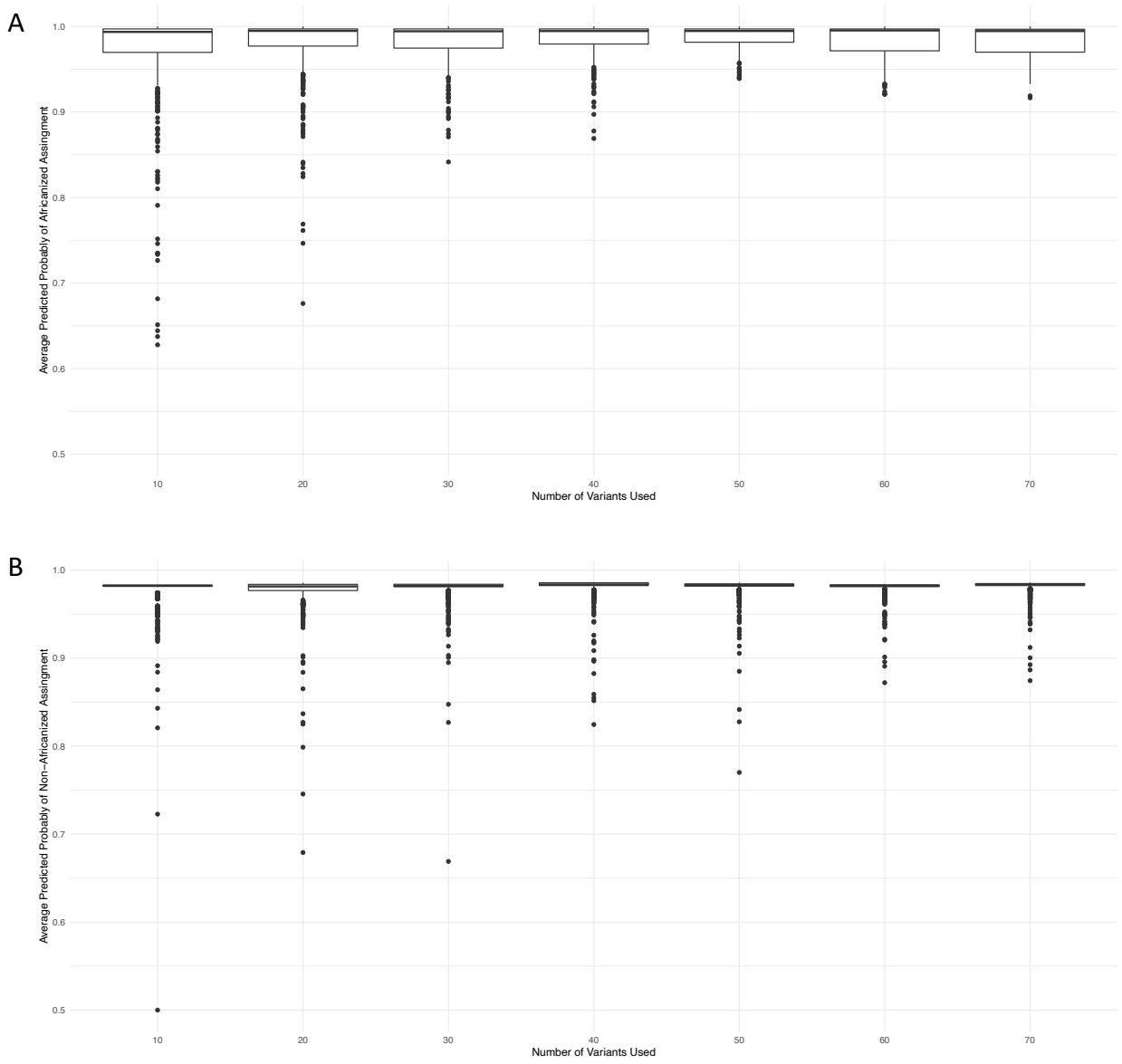

Fig. S4.4: Predicting the classification probability of samples using a reduced dataset. A linear SVC was trained with 177 training samples based on a random subset of 10-70 SNPs increasing by increments of 10. The trained model was tested over 5 replicates on 117 known reference samples and 694 validation samples that were originally successfully genotyped across all 80 loci (n=91 Africanized samples, n=720 non-Africanized samples). A) Figure A represents the predicted classification of Africanized honey bee samples (N=91) using a model trained with a random subset of SNPs over five replicates. B) Figure B represents the predicted classification of non-Africanized honey bee samples (N=720) using a model trained with a random subset of SNPs over five replicates.

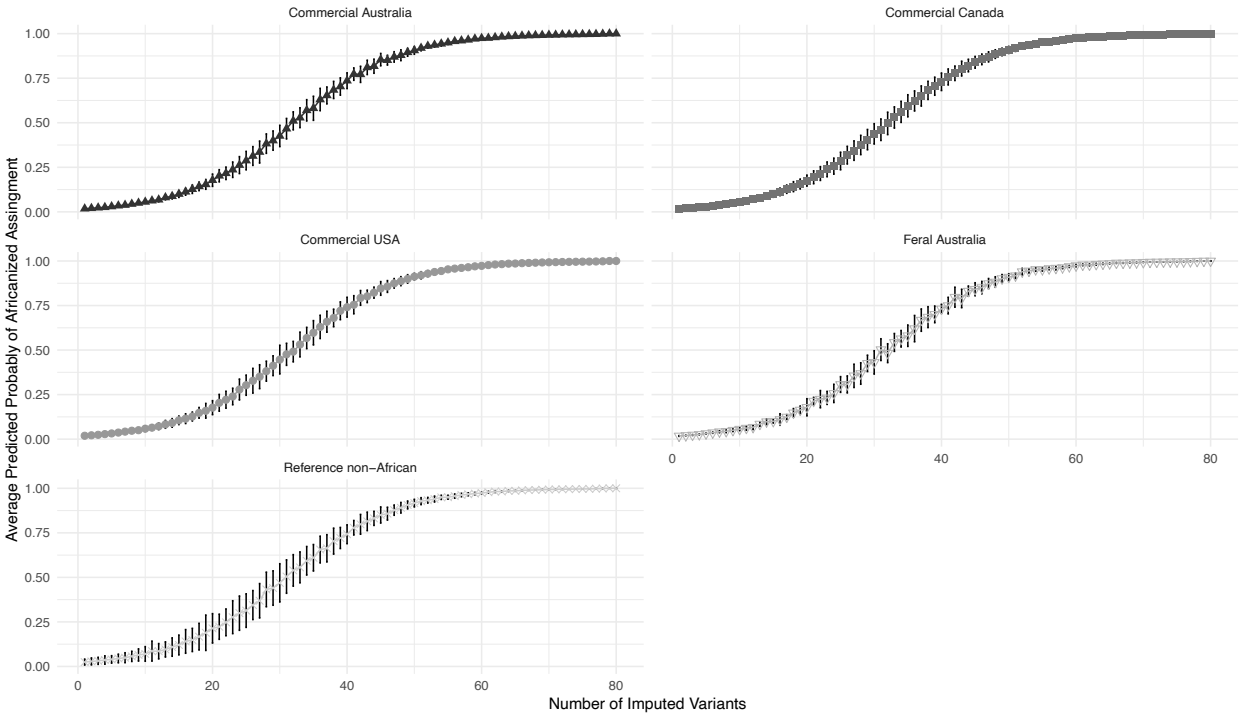

Fig. S4.5: Effects of imputation on probability estimates. To assess the effects of imputation on non-Africanized samples, 1-80 loci were randomly assigned an 'Africanized" genotype among 720 sample. A linear SVC model was then used to estimate the probability of assignment to an Africanized classification. Each graph depicts the curve of change in the predicted probability of an Africanized classification based on the number of imputed genotypes. Bars representant the standard deviation of estimates.

Chapter five: Developing a collection of insertion-deletion markers to identify Africanized honey bees

Introduction:

Africanized honey bees (AHB) are an invasive hybrid population with predominantly African ancestry (Chapman et al., 2017) that have rapidly spread across south America into the southern United States since their introduction to Brazil in 1956 (Calfee et al., 2020, Kerr, 1967). This strain of honey bee is often regarded as undesirable for beekeeping due to their high colony aggression and high tendency to swarm and abscond (Winston, 1992). Given the large-scale trade and movement of honey bee across the globe, several countries have placed strict import restrictions on regions with known Africanized honey bees. However, recent studies have shown that Africanized honey bee populations continue to expand their distribution (Lin et al., 2018), leaving the industry vulnerable to accidental importation. As such, effective tools for monitoring and identifying Africanized bees are crucially needed.

Several identification methods have been developed to detect Africanized bee populations relative to commercial honey bee populations, which are predominantly European in ancestry (Chapman et al., 2015, Harpur et al., 2015, Chapman et al., 2016). Early identification methods relied on morphological and mitochondrial differences, but both methods are unable to accurately detect low levels of Africanization (Guzmán-Novoa et al., 1994). Mitochondrial methods also cannot track paternal inheritance of African ancestry (Sheppard and Smith, 2000). Recent detection assays have instead focused primarily on nuclear single nucleotide polymorphism (SNP) markers, such as Chapter 4, which can easily and effectively delineate between African and European ancestral lineages (Chapman et al., 2015, Harpur et al., 2015, Chapman et al., 2016) and subspecies (Momeni et al., 2021). SNP markers however have some drawbacks, mainly the cost associated with genotyping individuals at a sufficient number of markers to ensure accurate identification; typically, 30 to 80 markers (Chapter 4). Identification using nuclear insertion-deletion (InDels) polymorphisms may alleviate the costs associated with SNP genotyping as InDels can be assayed using relatively inexpensive PCR methods.

Insertion-deletion polymorphisms are abundantly distributed across the genome and are characterized as short segments of DNA that are added or deleted to the DNA sequence (Väli et al., 2008). These variable loci can segregate at discernable frequencies between populations,

which make them an informative marker for population differentiation and ancestry identification (Pereira et al., 2012, Santos et al., 2015, Zaumsegel et al., 2013). Typically, individual samples can be easily and quickly genotyped with InDel markers using PCR amplification and post gel screening for presence and absence or amplicon size (Bashir and Hassan, 2016, Mitrečić et al., 2008). While insertion-deletion mutations have yet to be used for the identification of *Apis mellifera* samples, we believe that they can be an effective tool for detecting Africanized bees.

Here, we developed a PCR assay targeting InDel markers to differentiate between individual Africanized honey bees and commercial honey bees of European ancestry. Using 265 fully sequenced honey bee genomes from across *A. mellifera*'s native range, we identified several bi-allelic insertion-deletion mutations as putative targets for AHB classification. In total, we successfully designed nine primers that were validated using commercial honey bee samples from the United States and Australia, and Africanized honey bee samples from Brazil. Our assay provides an effective tool for quickly screening honey bee samples on their own, or in conjunction with existing detection methods as an added measure of control

Methods:

Genome processing

Our dataset consists of 243 previously sequenced honey bee genomes representative of seven genetically distinct lineages of native *Apis mellifera* (Dogantzis et al., 2021), in addition to newly sequence genomes representative of 16 hybrid Africanized bees, and six commercial North American honey bees. Sample preparation and genome sequencing of new hybrid Africanized bee samples follows a previously published protocol (Dogantzis et al., 2021). Similarly, sample preparation and genome sequencing of North American samples followed a previously published protocol (Harpur et al., 2014). Sequence reads were trimmed of Illumina adapters and low quality bases (<20) using Trimmomatic v0.36 (Bolger et al., 2014). Reads were aligned to the *Apis mellifera* reference genome (Elsik et al., 2014) using NextGenMap aligner v0.4.12 (Sedlazeck et al., 2013), and reads were marked for duplicates using Picard v2.1.0 (http://broadinstitute.github.io/picard/). Insertion-deletion mutations were identified using HaplotypeCaller and isolated from SNPs using the SelectVariants function in GATK v3.7 (Van der Auwera et al., 2013).

Indel marker selection

To determine a candidate set of informative InDel markers we identified bi-allelic loci with an allele frequency difference ≥0.8 between Africanized (n=16) and commercial North American (n=6) samples that also had <5% missing data across all 265 samples; 266,541 InDel markers were retained. Form the filtered dataset, missing genotypes for individuals was imputed using the consensus genotype representative of the lineage of origin as determined in (Dogantzis et al., 2021). Final genotypes were coded as "0" representing homozygous reference genotypes, "1" representing heterozygous genotypes, and "2" representing homozygous alternative genotype calls. Using the reduced dataset, we used a random forest classification model (Breiman, 2001) to determine the importance of InDel markers for classifying samples as Africanized or non-Africanized. To train the random forest classifier, we divided the 265 honey bee genomes into a training group and a testing group, which contained 177 (66%) and 88 (33%) samples respectively. The random forest classifier was trained using the training set (n=177) over ten replicates, while implementing the GridSearchCV option, to determine the optimal parameters of the model, including n_estimators, max_features, and max_depth, as implement by the scikit-learn Python package (Pedregosa et al., 2011). Feature importance for each replicate run was estimated for the top markers using the feature_importances_ option as implement by the scikit-learn Python package (Pedregosa et al., 2011). In total, we identified 105 InDel markers as candidates for primer design. These markers were highly ranked and free of secondary variants (including SNPs) within 20bps up or downstream of the target loci.

PCR Primer development

Primers to amplify the InDel mutations were designed using a 3' variable end approach (3' VPE), such that the 3' end of the reverse or forward primer contains the insertion variant (Mitrečić et al., 2008). During PCR amplification, samples with the insertion will yield a PCR product, while samples without the insertion will not yield a PCR product due to the sequence mismatch. All primers were designed using Primer3web version 4.1.0 (Untergasser et al., 2012) using the following parameters: primer length 15bp – 30bp, melting temperature 42°C – 65°C, and GC content 40 – 60%. Primer candidates were excluded if multiple unintended genome

targets were predicted. We successfully designed and optimized primers for nine InDel markers with amplicon sizes ranging between 150 bp – 700 bp (Table 5.1).

Sample collection and DNA extraction of validation samples

The InDel primers were validated on Africanized honey bees from Brazil (N=91), commercial honey bees from the USA (N=109), and commercial honey bees from Australia (N=66). DNA was extracted from samples using the Mag-Bind® Blood & Tissue DNA HDQ 96 Kit (Omega Bio-tek Inc., USA) optimised for the KingFisher™ Flex Purification System (Thermo Fisher Scientific Inc., USA). For tissue lysis, either half or whole bee thoraces were flash frozen in liquid nitrogen and finely ground using a pestle. We then added 350µl Tissue Lysis Buffer, 20µl Proteinase K, and heated samples overnight at 55°C. After processing with the KingFisher system, samples were eluted in nuclease-free water (Thermo Fisher Scientific Inc., USA) to a final volume ranging from 50–80µl. DNA was quantified using NanoDrop™ 2000 Spectrophotometer (Thermo Fisher Scientific Inc., USA) and DNA quality was assessed with 1.0% agarose gel electrophoresis.

PCR Protocol and genotyping of validation samples

Validation samples were screened for the presence or absence of the InDel variant using PCR amplification and gel electrophoresis. PCR amplification was performed in 12uL reactions with the following concentrations of reagents: 5ng/uL of DNA, 1X PCR buffer (ThermoFisher Scientific), 0.2 mg/mL bovine serum albumin (New Enlgland BioLabs), 0.2 mM dNTPs (ThermoFisher Scientific), 0.2µM forward and reverse primer (Integrated DNA Technologies), 0.05U/µl Taq Polymerase (ThermoFisher Scientific), and nuclease-free water. Concentrations of $MgCl_2$ were dependent on primer type (Table 5.1). The reaction was run with an initial denaturation at 95°C for 5 minutes, followed by 30 cycles of denaturation at 95°C for 30 seconds, annealing for 30 seconds, extension at 72°C for 30 seconds, and a final extension step at 72°C for 10 minutes. Annealing temperatures were dependent on primer type (Table 5.1). Each sample was also individually amplified using an internal control primer targeting the CYP9Q gene to validate the PCR reaction in absence of the indel. The CYP9Q primer sequences were obtain from (Tsvetkov et al., In review) and were amplified using the same conditions as the target InDel primer.

Finally, 10ul aliquots of PCR product was visualised by 2% agarose gel electrophoresis with 4ul of 10 mg/ml Ethidium Bromide. Genotypes were scored based on presence or absence of the insertion. In all but one instance, the amplification of the insertion indicates an African ancestry, the exception being primer AHB-INDEL-16, where an absence of amplification indicates African ancestry. Due to the binary nature of the genotype screening, samples that were heterozygous were scored based on presence of the insertion. Samples that failed to genotype at two or more loci were excluded from further analyses (N=10; N=3 Africanized honey bees and N=7 commercial honey bees from Australia).

Population structure analyses

To validate the discriminatory performance of the InDel markers in differentiating Africanized and non-Africanized samples, we evaluated the population structure and genetic clustering of samples. ADMIXTURE v1.3.0 (Alexander and Lange, 2011) was used to estimate the population assignment of genotyped samples (N=256) against known reference samples (N=265) whose genotypes were scored from genomic data. ADMIXTURE was run with predicted K values 1 – 8 using the 10X cross validation procedure using nine validated Indel markers. To examine patterns of genetic clustering among samples, we used a principal component analysis (PCA). The PCA was constructed using the SNPRelate (Zheng et al., 2012) package in R v3.6.0 (R Core Team, 2013) using genotypes from the 9 indel markers.

Accuracy of the indel markers

We determined the accuracy of the PCR assay by calculating the false positive and false negative rate based on the population assignments estimated with ADMIXTURE v1.3.0 (Alexander and Lange, 2011). This was achieved by determining the number of individuals that would be misclassified based on an African ancestry threshold ranging between 5-95% over 5% increments. To determine the false negative rate, we counted the number of misclassified Africanized or African individuals for each threshold and divided it by the total number of Africanized and African samples. The false positive rate was determined by counting the number of misclassified non-African or non-Africanized samples at each threshold and divided by the total number of non-African and non-Africanized samples.

Results:

We successfully designed nine InDel markers to differentiate between two classification groups: Africanized and non-Africanized samples. The PCR amplification had a high average success rate (>99%) across all loci, including primer AHB-INDEL-3 which had the higher proportion of failed amplifications (2.7%). Across individual honey bees, 241 (90.6%) samples were successfully genotyped across all nine loci, while 15 samples experienced failed amplification at one locus – indicated by failed amplification of the control locus. There were ten samples that failed amplification at two or more loci and were subsequently excluded from the analysis.

The population structure results produced with the program ADMIXTURE illustrates that when K=2, there are two ancestrally relevant clusters (Fig 5.1). The first cluster is comprised of samples that are reflective of African lineage ancestry, which includes Africanized and African ancestry samples, while the second cluster is reflective of European lineage ancestry and includes non-African and commercial honey bee samples. The average ancestral proportion estimated for each reference population differed significantly between groups suggesting these samples are an accurate reflection of their ancestral origins ($H_{(5)} = 454.83$, $p < 2e^{-16}$) (Fig. 5.2). On average, reference African and reference AHB samples had an African ancestry assignment of 0.977 ± 0.062 and 0.986 ± 0.056 respectively, which was significantly different relative to the average African ancestry assignment for reference non-African and reference commercial samples combined (Dunn $p < 1.82e^{-12}$ both comparisons) (Fig. 5.2A). Likewise, reference non-African and reference commercial honey bees had a combined European ancestry assignment of 0.963 ± 0.101, which was significantly different relative to the average European assignment for reference African and reference AHB samples (Dunn $p < 1.82e^{-12}$ both comparisons) (Fig. 5.2B).

Among the samples used to validate the InDel assay, Africanized bee samples from Brazil, had an average ancestral composition of 0.906 ± 0.10 African ancestry. We found no significant different in ancestry estimates between AHB from Brazil relative to reference AHB (Dunn p = 0.53), but there was a significant difference relative to reference African samples (Dunn p = 0.026) ($H_{(5)} = 454.83$, $p < 2e^{-16}$) (Fig. 5.2A). Validation samples of commercial honey bees from Australia and the USA had an average ancestral composition of 0.996 ± 0.022 and 0.976 ± 0.056 respectively, to European ancestry (Fig. 5.2B). We found no significant difference in European ancestry assignment between the commercial honey bee samples and the reference

non-African samples (Dunn $p > 0.89$ all comparisons), and as expected, we found a significant difference between the commercial samples and the Africanized honey bee from Brazil (Dunn $p < 2.09e^{-22}$) (Fig. 5.2B).

Structuring of the samples was further validated using a principal component analysis. The PCA delineates samples into two clusters that separate primarily along principal component one, which explains 81.72% of the variance among samples (Fig. S5.1). Like the ADMIXTURE results, the reference African and reference AHB samples cluster with the AHB from Brazil, revealing the similarity is genetic composition. Similar clustering can be observed among the reference non-African and commercial honey bee samples from Australia and the USA (Fig. S5.1). Principle component two explains a much smaller proportion of the variance (4.94%), and primarily separates samples among, rather than between, ancestrally relevant clusters.

To determine the accuracy of the PCR assay, we assessed the false positive and false negative rate associated with misclassifying samples. Misclassifications were based on samples being included or excluded from their target group based on different thresholds for African ancestry. Thresholds of African ancestry ranged between 5-95% and increased by 5% intervals. Samples of Africanized honey bees from Brazil, used to validate the assay, were labeled as false negatives if estimates for African ancestry fell below the target threshold. Likewise, commercial honey bee samples from Australia and the USA were labeled false positives if estimates for African ancestry fell above the target threshold. We found that when using an African ancestry threshold of 25% or higher, no commercial honey bee samples are misclassified as African, or in other words, no false positives are produced (Fig. 5.3). On average, only 1.7% of ancestry is contributed from the African (A) lineage, thus it is possible to use a lower threshold, where a small percentage of samples would be considered false positives. For example, when a threshold of 20% is used, only three samples (<2%) are misclassified. We begin to see a marked effect of the minimum threshold when <12% African ancestry is used. Here, approximately 13% of commercial samples would be misclassified. Likewise, when using a threshold of 55% or lower, no Africanized bee samples are misclassified (Fig. 5.3). When analyzing samples for Africanized honey bee classification, false negatives are of greater consequence and could result in the accidental movement of invasive strains. As such, a conservation threshold on the lower end of the range will account for variance of genetic admixture in Africanized bees and will minimize the chance of false negatives.

Discussion:

Current molecular based detection methods for identifying Africanized bees have primarily focused on using SNP markers. In contrast, insertion-deletion mutations (InDels) have been underutilized but provide an alternative means for capturing the genetic differentiation between honey bee populations. Here, we developed nine, ancestry informative InDel markers that can be used to effectively differentiate individual Africanized honey bees from commercial honey bees. When a threshold of 25% African ancestry is used to classify samples, all validation samples of Africanized honey bees from Brazil (N=88) and commercial honey bees from the United States (N=109) and Australia (N=59) are correctly categorized. A threshold of 25% African ancestry, while higher relative to ancestry based assays, developed with many more SNP markers (Chapman et al., 2017, Chapman et al., 2015), is still low enough to prevent false negatives as the lowest ancestry proportion for Africanized bees in this study was >55%. For beekeepers wishing to import colonies within the native distribution of *Apis mellifera*, there may be some populations that surpass the 25% threshold. For example some O lineage bees, mainly *A. m. syriaca*, located among a hybrid zone with other West Asian (Y) and African (A and L) lineages demonstrate higher levels of African ancestry relative to European and some other O lineage subspecies (Dogantzis et al., 2021). Nevertheless, we recommend the use of the 25% African ancestry threshold for use in detecting putative AHB among imports arising from North or South America.

Our assay provides some advantages over current molecular based detection methods. Relative to mitochondrial based tests, which rely on the detection of African haplotypes via maternal inheritance, the InDel assay can detect African ancestry alleles from either a paternal or maternal origin. This eliminates the chance of misidentifying offspring of African drones and European queens. Additionally, relative to short tandem repeats, InDel markers have a lower mutation rate, and in some cases, a smaller amplicon size. These factors make InDels suitable for the analysis of degraded or low quantities of DNA (Bashir and Hassan, 2016), and may allow for non-lethal screening of honey bee individuals (Châline et al., 2004). Finally, our PCR assay offers a cost effective and time efficient alternative to SNP genotyping as this assay uses fewer markers and can be performed with simple and standard molecular biology instruments (e.g., PCR machine, gel electrophoresis rig).

In conclusion, we have developed a novel PCR assay that uses the discriminative power of nine bi-allelic insertion-deletion markers to accurately and cost effectively differentiate individual Africanized and non-Africanized honey bees. When used alone, or in combination with previously developed molecular or morphology-based methods, this assay will provide additional screening measures for regulating honey bee imports and monitoring the movement of Africanizes honey bees. For example, this assay can be combined with the assay developed in Chapter 4 as additional screening or to provide different information for classification purposes. The assay in Chapter 4 was developed with SNP markers and provides a probability estimate to a classification group, while this PCR assay uses InDel markers and provides information on ancestry composition.

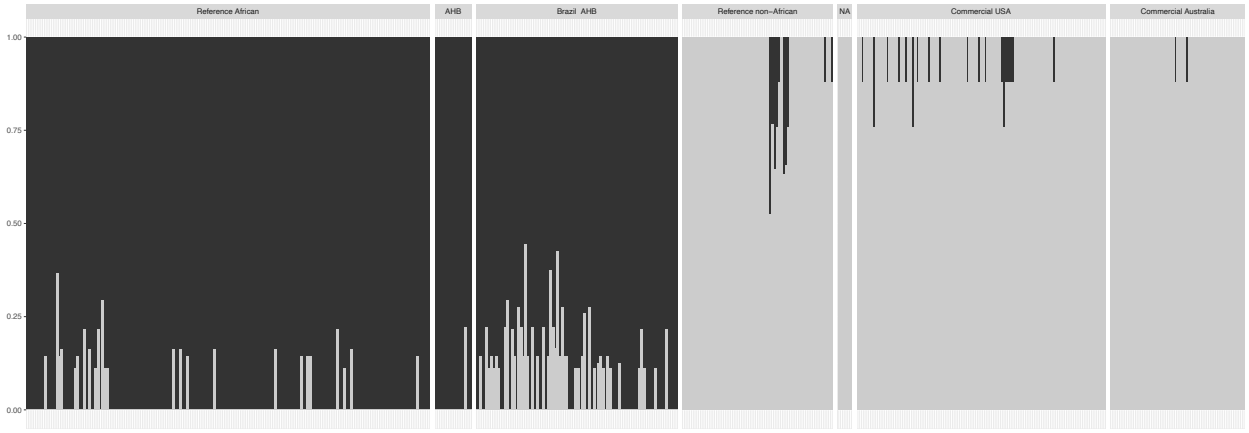

Fig. 5.1: Results of the ADMIXTURE analysis for K=2 clusters conducted with nine bi-allelic InDel markers. Vertical bars represent individual bees and dark grey segments represent the proportion of ancestry assigned to an African or Africanized origin, relative to the light grey segments, which represent non-African and non-Africanized origins. Africanized honey bee samples from Brazil, used to validate the assay, clearly cluster with the reference African and reference Africanized honey bee (AHB) samples. Comparatively, the commercial USA and Australia samples used to validate the assay clearly cluster with the reference non-African and reference commercial North American (NA) samples.

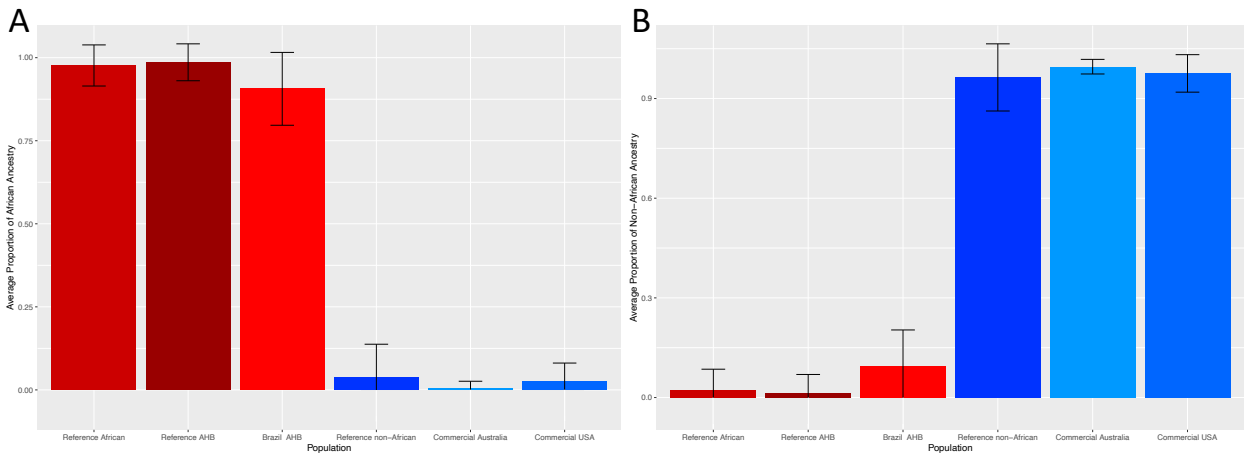

Fig. 5.2: Ancestry proportions of reference and validation samples. The average proportion of A) African and B) non-African (European) ancestry of reference and validation samples. In both figures the reference commercial North American samples are grouped with the reference non-African samples. Error bars represent the standard deviation.

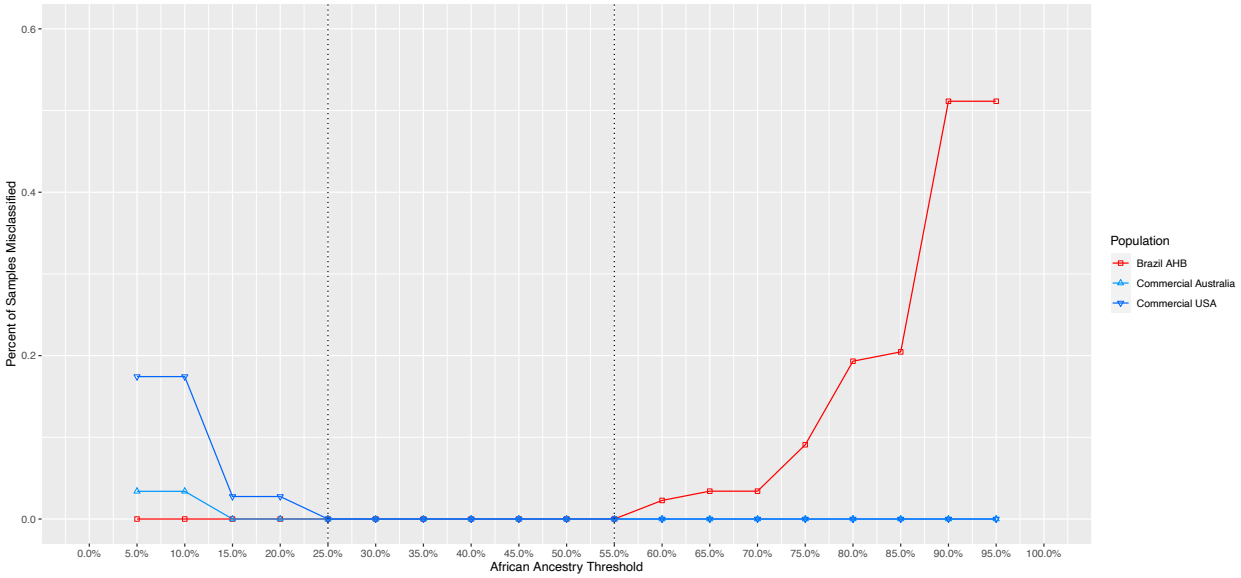

Fig. 5.3: The percentage of misclassified individuals among each validation population derived from the ADMIXTURE results. The false positive rate was defined as the proportion of commercial honey bee samples misclassified as Africanized given a threshold for African ancestry. Likewise, the false negative rate was defined as the proportion of Africanized honey bees misclassified as non-Africanized given a threshold for African ancestry. Threshold values were increase by 5% increments between 5%-95%.

Table 5.1: Primer sequences and amplification requirements for primers used to identify Africanized bee samples. Square brackets indicate the 3' variable end sequence of the insertion.

Primer Name	Sequence (5' - 3')	Product Size	Position	AHB Genotype	MgCl2 (mM)	Temperature
AHB-INDEL-18F	GCTAACAGTGAATCGAACCG	189	1.34:401139	Present	1.5	55°C
AHB-INDEL-18R	CGAGTAACTTTATTCCAGCTTA[TTT]	189	1.34:401139	Present	1.5	55°C
AHB-INDEL-16F	AAGCTCCACAGGAAGAGGAC	471	2.19:2971072	Absent	2	57°C
AHB-INDEL-16R	CGGTAGTTACACGGAAAGACC[CA]	471	2.19:2971072	Absent	2	57°C
AHB-INDEL-17F	GGGTACAAACGCGCATTCTA	162	5.12:180114	Present	1.5	55°C
AHB-INDEL-17R	TCCTTCACCTGACCGCC[G]	162	5.12:180114	Present	1.5	55°C
AHB-INDEL-10F	TGGAAGAAAAATCATATCAGCC	284	10.25:202579	Present	1.5	58°C
AHB-INDEL-10R	CATTGCAATAATAAAAAATTGCAA[AC]	284	10.25:202579	Present	1.5	58°C
AHB-INDEL-11F	TATGGGCACGCTATAATAGT[AAATG]	463	11.1:24300	Present	2	53°C
AHB-INDEL-11R	CGTTCTCGCGTTTATTCACA	463	11.1:24300	Present	2	53°C
AHB-INDEL-12F	TGCCTTTTACCAGTAGCTTGG	453	11.20:670989	Present	1.5	58°C
AHB-INDEL-12R	GGCCGGTATGCATGTGCT[ATG]	453	11.20:670989	Present	1.5	58°C
AHB-INDEL-5F	AGTTACAACATGAGTGGCCA	246	12.10:135321	Present	1.5	54°C
AHB-INDEL-5R	GATGGCAGAGATGTCGA[TGA]	246	12.10:135321	Present	1.5	54°C
AHB-INDEL-3F	AGTAGCGCAACAATGTTCAC	159	15.3:176276	Present	1.5	57°C
AHB-INDEL-3R	TATCGGCTTCCATCCAT[ATG]	159	15.3:176276	Present	1.5	57°C
AHB-INDEL-2F	AGTTAATGACACTCACAT[ACAC]	242	16.2:238025	Present	1.5	59°C
AHB-INDEL-2R	AGCGTGATATGTACATAACTTAA	242	16.2:238025	Present	1.5	59°C

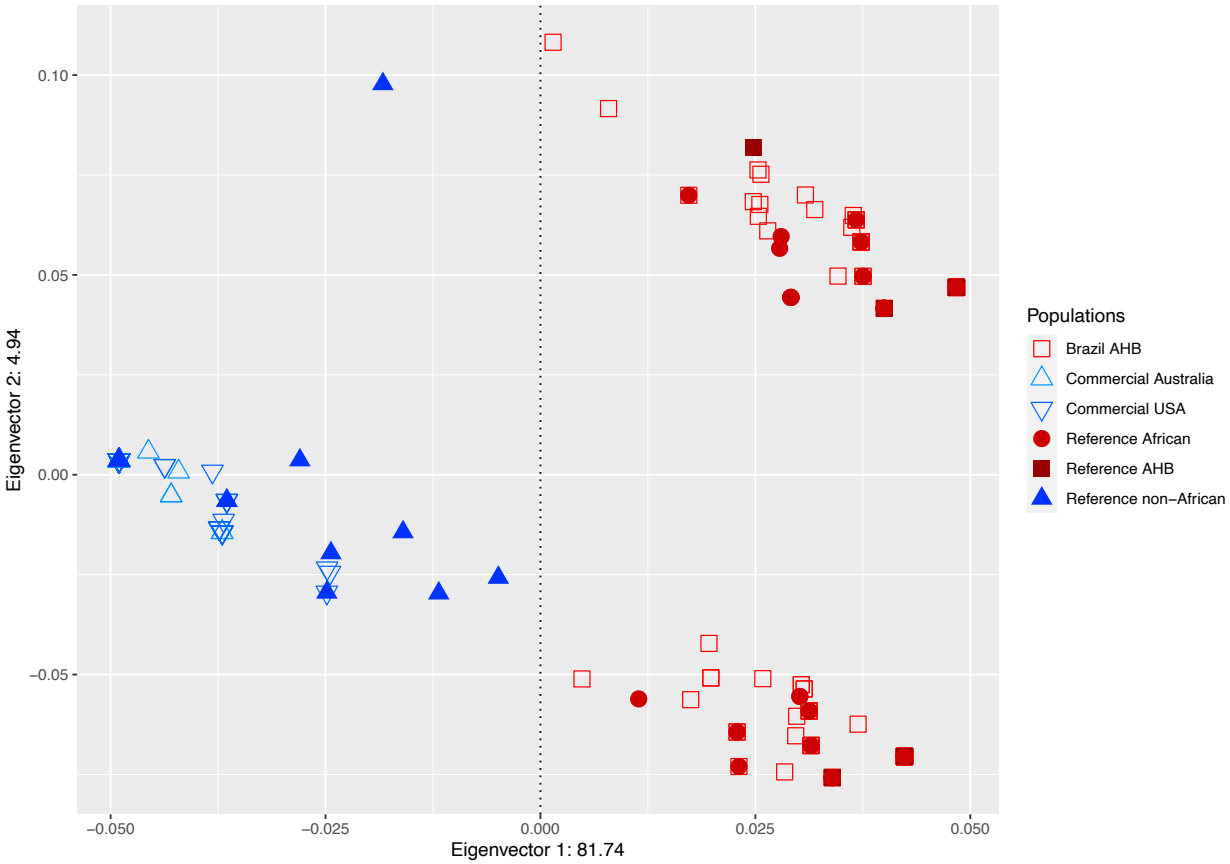

Fig. S5.1: Principal component analysis of reference and validation *Apis mellifera* samples. The first principal component explains the majority of variance in the dataset and clearly separates the African and Africanized samples from the non-African and non-Africanized samples. The second principal component explains only 5% of the variance, and separate samples among groups, rather than between ancestrally significant groups.

Chapter six: Conclusion and future work

The honey bee, *Apis mellifera*, is arguably the most important managed pollinator. The species has been translocated globally for apiculture and has positive effects on agriculture (Khalifa et al., 2021) and natural systems (Hung et al., 2018). Yet, despite the species importance, there are still several unknowns regarding the ancestral origin and ancestral complexity of contemporary populations. Understanding the genetic composition of honey bee colonies and how ancestry may influence phenotypes is important for mitigating declines, informing selective breeding practices, and improving the success of the beekeeping industry. In this dissertation, I delve into the evolutionary origin of *Apis mellifera* and learn how ancestral complexity has shaped the composition of contemporary populations (Fig. 6.1).

The ancestral origin of *Apis mellifera* has been hotly debated and revolves around two competing hypotheses, the out-of-Africa and out-of-Asia hypothesis (Dogantzis and Zayed, 2019) (Fig. 1.1). To settle this debate, I curated a comprehensive population genomic dataset representative of 18 different subspecies collected from the native distribution of *Apis mellifera*. Phylogenetic reconstruction of the honey bee samples emphasized an ancestral divide between West Asian lineages (Y and O), and phylogenies based on protein-coding regions resolved the Y lineage as the most basal branch (Dogantzis et al., 2021). Further, the biogeographic reconstruction supported an ancestral range in Asia with >70% probability (Dogantzis et al., 2021). The radiation of contemporary populations is suspected to have involved selection among 'hot spots', where 145 genes were found be associated with signatures of selection among all lineages. These finding provide the basis for tracking the evolution of derived mutations and their association with adaptive traits and genetically distinct lineages. This has important implications for conservation initiatives that wish to preserve the genetic underpinning of local adaption and can inform breeding practices that seek to enhance desirable colony phenotypes.

Though we have been able to settle the out-of-Africa and out-of-Asia debate, there are still some unknowns regarding the timeline and pattern of colonization events. Accurate divergence dating of *Apis mellifera* lineages and subspecies is needed to discern when colonization and divergence occurred. In Chapter 2, divergence dating suggests the evolution of lineages began c.6 Ma ago. However, other estimates place divergence between 1.3 Ma ago to as early as 300 Ka ago (Cornuet and Garnery, 1991, Arias and Sheppard, 1996, Wallberg et al., 2014). The discrepancy likely results from the use of different sequence data and different

divergence dating methods. Creating a standardized and reproducible approach is needed to accurately estimate timelines.

Colonization patterns are further complicated by the placement of the M lineage. We hypothesize that the M lineage, which forms a distinct evolutionary branch, likely colonized Europe via an independent northern route. However, there are alternative hypotheses that posit the M lineage expanded from Africa, which is supported by shared genetic similarity with subspecies from Northwest Africa. While these patterns are likely the result of recent introgression (Chávez-Galarza et al., 2017, Chávez-Galarza et al., 2015, Cánovas et al., 2008, Pinto et al., 2013, Boardman et al., 2020a), follow up studies are needed to track when admixture could have first occurred and if introgression is ongoing with Northwest African populations. Future work would benefit from honey bees collected from Northwest Africa that are free from admixture with the M lineage. Such samples would help conclude if the M lineage has close ties to Africa, or if patterns are confounded by recent introgression.

Finally, recent studies have focused on developing hypotheses about the evolution of *Apis mellifera* using nuclear variation but follow up studies using mitochondrial variation are needed to build the full picture. This is especially relevant for work done at the genus level where recent phylogenetic reconstructions are challenging the status quo of the expected topology (Boardman et al., 2020b).

Managed honey bees in North America are highly admixed, but the discovery of seven genetically distinct evolutionary lineages (Chapter 2) calls for a reassessment of the ancestral complexity of managed honey bee. Here, I used the extensive population genomics dataset from Chapter 2 to estimate ancestry proportions of Canadian honey bee colonies. Structure analyses revealed colonies to be primary composed of C-lineage ancestry with the remaining ancestry originating from the M and O-lineages. Interestingly, the degree of non-C-lineage admixture was found to be positively corelated with genetic diversity. Finally, I 'mapped' ancestry across the genome of Canadian honey bees and found that regions enriched for admixed ancestry (M and O) were concentrated on chromosomes 4 and 7 and were associated with genes that influence *Varroa* response and xenobiotic detoxification. These finding are important as they highlight the links between admixture, genetic diversity, and their association with important colony traits. These association can be used to trace the origin of beneficial loci to their ancestral lineage or

subspecies, which has significant implications for the conservation of derived traits and the selective breeding of commercial honey bees.

Colony traits such as those related to disease resistance and neonicotinoid tolerance are highly sought after among beekeepers. In particular, there has been intensive selective breeding programs targeting *Varroa* resistance (Saelao et al., 2020). It is possible that breeding efforts have selected for mutations that are present at a greater frequency among M and O lineages, which would consequently increase the associated ancestry around the target loci. Though selective breeding has not occurred in response to xenobiotic detoxification, since honey bees are often kept near agricultural areas for pollination (Tsvetkov et al., 2017), it is possible that exposure to agrochemicals has increased the frequency of M or O ancestry around associated loci. These are indeed intriguing results but require follow up studies to further explore the functional significance of admixture on honey bee traits.

Future studies could compare colonies that possess admixed ancestry among target outlier intervals, relative to a control group, to measures significant differences in the putatively selected phenotype. Additionally, if alleles associated with M and O lineage ancestry are indeed being selected for, scanning the genome for signatures of selection may reveal correlations with admixed ancestry around target loci. Such data are important for informing breeding programs and conservation initiatives to preserve the genetic origin of important traits. This is especially relevant for honey bee subspecies among the native range of *Apis mellifera,* where recent studies have shown some subspecies show a decline in genetic diversity (Espregueira Themudo et al., 2020). Additionally, native range subspecies can also be adversely affected by admixture, which may result in the loss of locally adaptive alleles (De la Rúa et al., 2013).

Though admixed ancestry is common among managed colonies, it is not always valued. This is especially true for Africanized honey bees (AHB), which are an invasive strain of honey bee that have rapidly spread throughout south America and into the southern United States. Though several molecular detection assays have been developed to identify AHB, some are not fully optimized, and some pose a financial barrier due to high costs. Here, I developed two new molecular assays to identify Africanized honey bees using nuclear SNP markers (Chapter 4) and insertion deletion markers (Chapter 5). These assays use a unique combination of markers that have been chosen based on their informativeness. Independently, both assays can accurately and consistently differentiate Africanized bees from commercial colonies. If implemented, these

assays are expected to be useful tools to identify and track the movement of Africanized bee populations and could help prevent the movement of AHB bees to unoccupied regions.

Though both assays have been optimized to maximize the differences between Africanized bees and commercial honey bees, future directions should include consistently updating the machine learning model used to analyze the SNP data. The analysis of the SNP genotypes uses an assisted machine learning model, while computationally intensive, has the added benefit of constantly being updated to improve classification. Thus, when classifying unknown samples, the updated model can make use of a larger dataset that captures the genetic variation of the target population to improve classification probabilities.

Additional follow-up steps are needed to implement both assays for AHB monitoring. The insertion deletion assay is easier to employ as most molecular genetics labs are equipped to run standard PCR reactions. Analysis of the genotyped samples can be easily done using any structure software that provides data on ancestry proportions. Costs incurred by labs would be associated with DNA extraction, PCR amplification, and gel electrophoresis screening. The estimated cost for genotypes one hundred samples is approximately $1300. In comparison, the cost of DNA extraction and genotyping for one hundred samples using the SNP panel is approximately $2600. There is also the added cost associated with the construction of the 80-marker panel, which has an upfront cost of approximately $2250 or more. When deciding on which assay to implement, researchers will need to consider the trade-offs. The insertion deletion analysis, while more cost effective, uses fewer informative markers and may be subject to sample loss if too many loci fail genotyping. Comparatively, the SNP panel is marker rich and is still effective when using as few as 30 markers, though the sequencing cost is markedly greater.

There are still several avenues available to improve our knowledge of *Apis mellifera* evolution and how ancestral complexity impacts contemporary populations. I am excited to see the expansion of the field and hope this collection of research will provide a comprehensive foundation for future work.

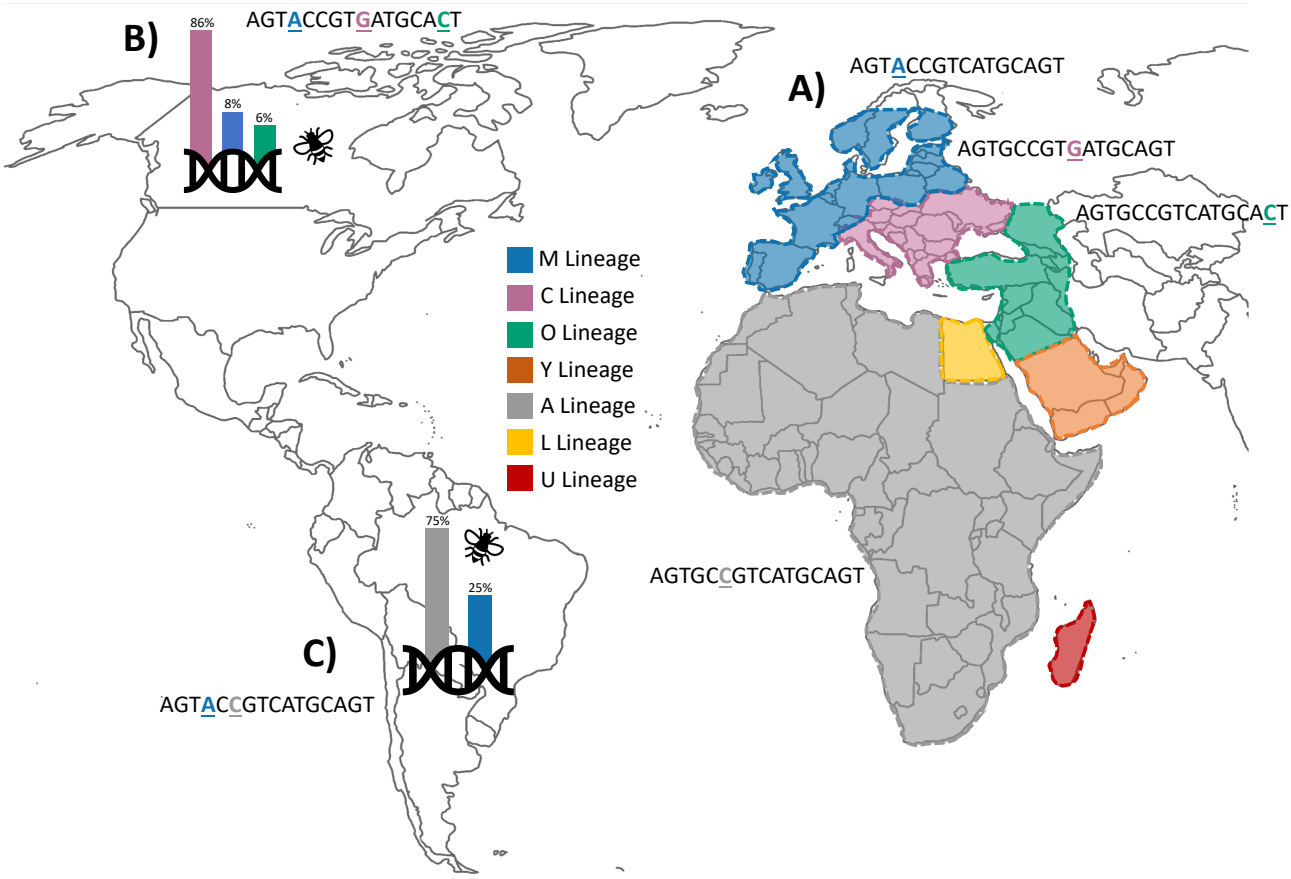

Fig. 6.1: Graphical summary of work. A) Chapter 2 explored the evolutionary relationships between *Apis mellifera* subspecies and discovered seven genetically distinct lineages that diversified out of Western Asia. Mutations associated with the adaptative radiation of lineages, while lineage specific (underlined loci), were associated with a common set of genes. Elucidating on the evolutionary origins of the species is important for identifying derived mutations (underlined loci) and their associated lineage of origin. **B)** In Chapter 3 the large genetic difference between lineages were used to determine the ancestral composition of managed honey bees, and to explore how ancestry is distributed across the genome. It was discovered that intervals enriched for admixed ancestry were concentrated on chromosome 4 and 7, and were associated with *Varroa* response and xenobiotic detoxification. These associations are important as they can be used to track the origin of beneficial loci back to their respective lineage. This has significant implications for subspecies conservation and selective breeding practices. **C)** In Chapter 4 and 5 the ancestral composition of Africanized bees was used to target loci that differentiate the population from commercial colonies (such as in B). Knowledge of the

genetic composition of honey bee lineages is imperative for accurately categorizing populations and identifying ancestry informative markers that can be used to differentiate groups.

References:

ABBOTT, R., ALBACH, D., ANSELL, S., ARNTZEN, J. W., BAIRD, S. J., BIERNE, N., BOUGHMAN, J., BRELSFORD, A., BUERKLE, C. A. & BUGGS, R. 2013. Hybridization and speciation. *Journal of Evolutionary Biology,* 26**,** 229-246.

ABBOTT, R. J., BARTON, N. H. & GOOD, J. M. 2016. Genomics of hybridization and its evolutionary consequences. *Molecular ecology,* 25**,** 2325-2332.

ALEXANDER, D. H. & LANGE, K. 2011. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC bioinformatics,* 12**,** 246.

ALPTEKIN, S., BASS, C., NICHOLLS, C., PAINE, M. J., CLARK, S. J., FIELD, L. & MOORES, G. D. 2016. Induced thiacloprid insensitivity in honeybees (A pis mellifera L.) is associated with up-regulation of detoxification genes. *Insect Molecular Biology,* 25**,** 171-180.

AMIRI, E., HERMAN, J. J., STRAND, M. K., TARPY, D. R. & RUEPPELL, O. 2020. Egg transcriptome profile responds to maternal virus infection in honey bees, *Apis mellifera*. *Infection, Genetics and Evolution,* 85**,** 104558.

ARECHAVALETA-VELASCO, M. E., ALCALA-ESCAMILLA, K., ROBLES-RIOS, C., TSURUDA, J. M. & HUNT, G. J. 2012. Fine-scale linkage mapping reveals a small set of candidate genes influencing honey bee grooming behavior in response to Varroa mites. *PLoS One,* 7**,** e47269.

ARIAS, M. C. & SHEPPARD, W. S. 2005. Phylogenetic relationships of honey bees (Hymenoptera: Apinae: Apini) inferred from nuclear and mitochondrial DNA sequence data. *Molecular phylogenetics and evolution,* 37**,** 25-35.

ASHBY, R., FORÊT, S., SEARLE, I. & MALESZKA, R. 2016. MicroRNAs in honey bee caste determination. *Scientific Reports,* 6**,** 18794.

BADAOUI, B., FOUGEROUX, A., PETIT, F., ANSELMO, A., GORNI, C., CUCURACHI, M., CERSINI, A., GRANATO, A., CARDETI, G. & FORMATO, G. 2017. RNA-sequence analysis of gene expression from honeybees (Apis mellifera) infected with Nosema ceranae. *PloS one,* 12**,** e0173438.

BASHIR, M. & HASSAN, N. H. B. 2016. Analysis of 30 Biallelic INDEL Markers Using the Investigator DIPplex® Kit. *Forensic DNA typing protocols.* Springer.

BEGUN, D. J., HOLLOWAY, A. K., STEVENS, K., HILLIER, L. W., POH, Y.-P., HAHN, M. W., NISTA, P. M., JONES, C. D., KERN, A. D., DEWEY, C. N., PACHTER, L., MYERS, E. & LANGLEY, C. H. 2007. Population Genomics: Whole-Genome Analysis of Polymorphism and Divergence in Drosophila simulans. *PLoS Biology,* 5**,** e310.

BEHRENS, D., HUANG, Q., GESSNE, C., ROSENKRANZ, P., FREY, E., LOCKE, B., MORITZ, R. F. & KRAUS, F. 2011. Three QTL in the honey bee Apis mellifera L. suppress reproduction of the parasitic mite Varroa destructor. *Ecology and evolution,* 1**,** 451-458.

BERENBAUM, M. R. & JOHNSON, R. M. 2015. Xenobiotic detoxification pathways in honey bees. *Current opinion in insect science,* 10**,** 51-58.

BISWAS, S., RUSSELL, R. J., JACKSON, C. J., VIDOVIC, M., GANESHINA, O., OAKESHOTT, J. G. & CLAUDIANOS, C. 2008. Bridging the synaptic gap: neuroligins and neurexin I in *Apis mellifera*. *PloS one,* 3**,** e3542.

BOARDMAN, L., EIMANIFAR, A., KIMBALL, R., BRAUN, E., FUCHS, S., GRÜNEWALD, B. & ELLIS, J. D. 2020a. The mitochondrial genome of the Spanish honey bee, *Apis mellifera iberiensis* (Insecta: Hymenoptera: Apidae), from Portugal. *Mitochondrial DNA Part B,* 5**,** 17-18.

BOARDMAN, L., EIMANIFAR, A., KIMBALL, R. T., BRAUN, E. L., FUCHS, S., GRÜNEWALD, B. & ELLIS, J. D. 2020b. The complete mitochondrial genome of Apis mellifera jemenitica (Insecta: Hymenoptera: Apidae), the Arabian honey bee. *Mitochondrial DNA Part B,* 5**,** 875-876.

BOLGER, A. M., LOHSE, M. & USADEL, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics***,** btu170.

BORBA, R. S., HOOVER, S. E., CURRIE, R. W., GIOVENAZZO, P., GUARNA, M. M., FOSTER, L. J., ZAYED, A. & PERNAL, S. F. 2022. Phenomic analysis of the honey bee pathogen-web and its dynamics on colony productivity, health and social immunity behaviors. *Plos one,* 17**,** e0263273.

BREIMAN, L. 2001. Random forests. *Machine learning,* 45**,** 5-32.

BROWNING, B. L., ZHOU, Y. & BROWNING, S. R. 2018. A one-penny imputed genome from next-generation reference panels. *The American Journal of Human Genetics,* 103**,** 338-348.

BROWNING, S. R. & BROWNING, B. L. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics,* 81**,** 1084-1097.

BUERKLE, C. A. & LEXER, C. 2008. Admixture as the basis for genetic mapping. *Trends in ecology & evolution,* 23**,** 686-694.

CALFEE, E., AGRA, M. N., PALACIO, M. A., RAMÍREZ, S. R. & COOP, G. 2020. Selection and hybridization shaped the Africanized honey bee invasion of the Americas. *bioRxiv*.

CÁNOVAS, F., DE LA RÚA, P., SERRANO, J. & GALIÁN, J. 2008. Geographical patterns of mitochondrial DNA variation in A*pis mellifera iberiensis* (Hymenoptera: Apidae). *Journal of Zoological Systematics and Evolutionary Research,* 46**,** 24-30.

CARPENTER, M. H. & HARPUR, B. A. 2021. Genetic past, present, and future of the honey bee (Apis mellifera) in the United States of America. *Apidologie,* 52**,** 63-79.

CHÂLINE, N., RATNIEKS, F. L., RAINE, N. E., BADCOCK, N. S. & BURKE, T. 2004. Non-lethal sampling of honey bee, Apis mellifera, DNA using wing tips. *Apidologie,* 35**,** 311-318.

CHAN, Q. W., CHAN, M. Y., LOGAN, M., FANG, Y., HIGO, H. & FOSTER, L. J. 2013. Honey bee protein atlas at organ-level resolution. *Genome research,* 23**,** 1951-1960.

CHAPMAN, N. C., BOURGEOIS, A. L., BEAMAN, L. D., LIM, J., HARPUR, B. A., ZAYED, A., ALLSOPP, M. H., RINDERER, T. E. & OLDROYD, B. P. 2017. An abbreviated SNP panel for ancestry assignment of honeybees (Apis mellifera). *Apidologie,* 48**,** 776-783.

CHAPMAN, N. C., HARPUR, B. A., LIM, J., RINDERER, T. E., ALLSOPP, M. H., ZAYED, A. & OLDROYD, B. P. 2015. A SNP test to identify Africanized honeybees via proportion of 'African' ancestry. *Molecular ecology resources,* 15**,** 1346-1355.

CHAPMAN, N. C., HARPUR, B. A., LIM, J., RINDERER, T. E., ALLSOPP, M. H., ZAYED, A. & OLDROYD, B. P. 2016. Hybrid origins of Australian honeybees (Apis mellifera). *Apidologie,* 47**,** 26-34.

CHÁVEZ-GALARZA, J., GARNERY, L., HENRIQUES, D., NEVES, C. J., LOUCIF-AYAD, W., JONHSTON, J. S. & PINTO, M. A. 2017. Mitochondrial DNA variation of *Apis mellifera iberiensis*:

further insights from a large-scale study using sequence data of the tRNAleu-cox2 intergenic region. *Apidologie,* 48**,** 533-544.

CHÁVEZ-GALARZA, J., HENRIQUES, D., JOHNSTON, J. S., CARNEIRO, M., RUFINO, J., PATTON, J. C. & PINTO, M. A. 2015. Revisiting the Iberian honey bee (*Apis mellifera iberiensis*) contact zone: maternal and genome-wide nuclear variations provide support for secondary contact from historical refugia. *Molecular ecology,* 24**,** 2973-2992.

CHEN, C., LIU, Z., PAN, Q., CHEN, X., WANG, H., GUO, H., LIU, S., LU, H., TIAN, S. & LI, R. 2016. Genomic analyses reveal demographic history and temperate adaptation of the newly discovered honey bee subspecies *Apis mellifera sinisxinyuan* n. ssp. *Molecular biology and evolution,* 33**,** 1337-1348.

CHEN, C., WANG, H., LIU, Z., CHEN, X., TANG, J., MENG, F. & SHI, W. 2018. Population genomics provide insights into the evolution and adaptation of the eastern honey bee (*Apis cerana*). *Molecular biology and evolution,* 35**,** 2260-2271.

CHEN, Y.-R., TZENG, D. T. & YANG, E.-C. 2021. Chronic Effects of Imidacloprid on Honey Bee Worker Development—Molecular Pathway Perspectives. *International journal of molecular sciences,* 22**,** 11835.

CHMIEL, J. A., DAISLEY, B. A., PITEK, A. P., THOMPSON, G. J. & REID, G. 2020. Understanding the effects of sublethal pesticide exposure on honey bees: a role for probiotics as mediators of environmental stress. *Frontiers in Ecology and Evolution,* 8**,** 22.

CINGOLANI, P., PLATTS, A., WANG, L. L., COON, M., NGUYEN, T., WANG, L., LAND, S. J., LU, X. & RUDEN, D. M. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly,* 6**,** 80-92.

CONTE, G. L., ARNEGARD, M. E., PEICHEL, C. L. & SCHLUTER, D. 2012. The probability of genetic parallelism and convergence in natural populations. *Proceedings of the Royal Society B: Biological Sciences,* 279**,** 5039-5047.

CORSETTI, E. & AZPIAZU, N. 2013. Functional dissection of the splice variants of the Drosophila gene homothorax (hth). *Developmental biology,* 384**,** 72-82.

CORTES, C. & VAPNIK, V. 1995. Support-vector networks. *Machine learning,* 20**,** 273-297.

CRIDLAND, J. M., TSUTSUI, N. D. & RAMÍREZ, S. R. 2017. The complex demographic history and evolutionary origin of the western honey bee, *Apis mellifera*. *Genome biology and evolution,* 9**,** 457–472.

DANECEK, P., AUTON, A., ABECASIS, G., ALBERS, C. A., BANKS, E., DEPRISTO, M. A., HANDSAKER, R. E., LUNTER, G., MARTH, G. T., SHERRY, S. T., MCVEAN, G., DURBIN, R. & GROUP, G. P. A. 2011. The variant call format and VCFtools. *Bioinformatics,* 27**,** 2156-2158.

DARVASI, A. & SHIFMAN, S. 2005. The beauty of admixture. *Nature genetics,* 37**,** 118-119.

DE LA RÚA, P., JAFFÉ, R., MUÑOZ, I., SERRANO, J., MORITZ, R. F. & KRAUS, F. B. 2013. Conserving genetic diversity in the honeybee: Comments on Harpur et al.(2012). Wiley Online Library.

DEGRANDI-HOFFMAN, G., GAGE, S. L., CORBY-HARRIS, V., CARROLL, M., CHAMBERS, M., GRAHAM, H., DEJONG, E. W., HIDALGO, G., CALLE, S. & AZZOUZ-OLDEN, F. 2018. Connecting the nutrient composition of seasonal pollens with changing nutritional needs of honey bee (Apis mellifera L.) colonies. *Journal of insect physiology,* 109**,** 114-124.

DELANEY, D. A., KELLER, J. J., CAREN, J. R. & TARPY, D. R. 2011. The physical, insemination, and reproductive quality of honey bee queens (Apis mellifera L.). *Apidologie,* 42**,** 1-13.
DIAS-ALVES, T., MAIRAL, J. & BLUM, M. G. B. 2018. Loter: A Software Package to Infer Local Ancestry for a Wide Range of Species. *Molecular Biology and Evolution,* 35**,** 2318-2326.
DOBIN, A., DAVIS, C. A., SCHLESINGER, F., DRENKOW, J., ZALESKI, C., JHA, S., BATUT, P., CHAISSON, M. & GINGERAS, T. R. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics,* 29**,** 15-21.
DOGANTZIS, K. A., HARPUR, B. A., RODRIGUES, A., BEANI, L., TOTH, A. L. & ZAYED, A. 2018. Insects with similar social complexity show convergent patterns of adaptive molecular evolution. *Scientific reports,* 8**,** 1-8.
DOGANTZIS, K. A., TIWARI, T., CONFLITTI, I. M., DEY, A., PATCH, H. M., MULI, E. M., GARNERY, L., WHITFIELD, C. W., STOLLE, E., ALQARNI, A. S., ALLSOPP, M. H. & ZAYED, A. 2021. Thrice out of Asia and the adaptive radiation of the western honey bee. *Science advances,* 7**,** eabj2151.
DOGANTZIS, K. A. & ZAYED, A. 2019. Recent advances in population and quantitative genomics of honey bees. *Current opinion in insect science,* 31**,** 93-98.
EL-NIWEIRI, M. A. & MORITZ, R. F. 2008. Mitochondrial discrimination of honeybees (*Apis mellifera*) of Sudan. *Apidologie,* 39**,** 566-573.
ELLIS, J. S., SOLAND-RECKEWEG, G., BUSWELL, V. G., HUML, J. V., BROWN, A. & KNIGHT, M. E. 2018. Introgression in native populations of Apis mellifera mellifera L: implications for conservation. *Journal of Insect Conservation,* 22**,** 377-390.
ELSIK, C. G., TAYAL, A., DIESH, C. M., UNNI, D. R., EMERY, M. L., NGUYEN, H. N. & HAGEN, D. E. 2016. Hymenoptera Genome Database: integrating genome annotations in HymenopteraMine. *Nucleic acids research,* 44**,** D793-D800.
ELSIK, C. G., WORLEY, K. C., BENNETT, A. K., BEYE, M., CAMARA, F., CHILDERS, C. P., DE GRAAF, D. C., DEBYSER, G., DENG, J. & DEVREESE, B. 2014. Finding the missing honey bee genes: lessons learned from a genome upgrade. *Bmc Genomics,* 15**,** 1.
ESPREGUEIRA THEMUDO, G., REY-IGLESIA, A., ROBLES TASCÓN, L., BRUUN JENSEN, A., DA FONSECA, R. R. & CAMPOS, P. F. 2020. Declining genetic diversity of European honeybees along the twentieth century. *Scientific reports,* 10**,** 1-12.
FAOSTAT 2022. Number of beehives worldwide from 2010 to 2020.
FORTUNE BUSINESS INSIGHTS 2022. Market value of honey worldwide from 2019 to 2028.
FRANCK, P., GARNERY, L., LOISEAU, A., OLDROYD, B., HEPBURN, H., SOLIGNAC, M. & CORNUET, J. M. 2001. Genetic diversity of the honeybee in Africa: microsatellite and mitochondrial data. *Heredity,* 86**,** 420-430.
FRANKHAM, R. 2015. Genetic rescue of small inbred populations: meta-analysis reveals large and consistent benefits of gene flow. *Molecular Ecology,* 24**,** 2610-2618.
FULLER, Z. L., NIÑO, E. L., PATCH, H. M., BEDOYA-REINA, O. C., BAUMGARTEN, T., MULI, E., MUMOKI, F., RATAN, A., MCGRAW, J., FRAZIER, M., MASIGA, D., SCHUSTER, S., GROZINGER, C. M. & MILLER, W. 2015. Genome-wide analysis of signatures of selection in populations of African honey bees (Apis mellifera) using new web-based tools. *BMC genomics,* 16**,** 518.

GALLAI, N., SALLES, J.-M., SETTELE, J. & VAISSIÈRE, B. E. 2009. Economic valuation of the vulnerability of world agriculture confronted with pollinator decline. *Ecological economics,* 68**,** 810-821.

GARNERY, L., CORNUET, J. M. & SOLIGNAC, M. 1992. Evolutionary history of the honey bee *Apis mellifera* inferred from mitochondrial DNA analysis. *Molecular ecology,* 1**,** 145-154.

GILL, N. S. & SANGERMANO, F. 2016. Africanized honeybee habitat suitability: a comparison between models for southern Utah and southern California. *Applied geography,* 76**,** 14-21.

GRAY, A., ADJLANE, N., ARAB, A., BALLIS, A., BRUSBARDIS, V., CHARRIÈRE, J.-D., CHLEBO, R., COFFEY, M. F., CORNELISSEN, B. & AMARO DA COSTA, C. 2020. Honey bee colony winter loss rates for 35 countries participating in the COLOSS survey for winter 2018–2019, and the effects of a new queen on the risk of colony winter loss. *Journal of Apicultural Research,* 59**,** 744-751.

GROZINGER, C. M. & ZAYED, A. 2020. Improving bee health through genomics. *Nature Reviews Genetics,* 21**,** 277-291.

GUARNA, M. M., MELATHOPOULOS, A. P., HUXTER, E., IOVINELLA, I., PARKER, R., STOYNOV, N., TAM, A., MOON, K.-M., CHAN, Q. W. & PELOSI, P. 2015. A search for protein biomarkers links olfactory signal transduction to social immunity. *BMC genomics,* 16**,** 1-16.

GUZMÁN-NOVOA, E., PAGE, R. E. & FONDRK, M. K. 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.

HADDAD, N. J., LOUCIF-AYAD, W., ADJLANE, N., SAINI, D., MANCHIGANTI, R., KRISHNAMURTHY, V., ALSHAGOOR, B., BATAINH, A. M. & MUGASIMANGALAM, R. 2015. Draft genome sequence of the Algerian bee Apis mellifera intermissa. *Genomics data,* 4**,** 24-25.

HAILU, T. G., D'ALVISE, P., TOFILSKI, A., FUCHS, S., GREILING, J., ROSENKRANZ, P. & HASSELMANN, M. 2020. Insights into Ethiopian honey bee diversity based on wing geomorphometric and mitochondrial DNA analyses. *Apidologie,* 51**,** 1182–1198.

HAMILTON, J. A. & MILLER, J. M. 2016. Adaptive introgression as a resource for management and genetic conservation in a changing climate. *Conservation Biology,* 30**,** 33-41.

HAN, F., WALLBERG, A. & WEBSTER, M. T. 2012. From where did the Western honeybee (*Apis mellifera*) originate? *Ecology and evolution,* 2**,** 1949-1957.

HANLEY, N., BREEZE, T. D., ELLIS, C. & GOULSON, D. 2015. Measuring the economic value of pollination services: Principles, evidence and knowledge gaps. *Ecosystem Services,* 14**,** 124-132.

HARPUR, B., CHAPMAN, N., KRIMUS, L., MACIUKIEWICZ, P., SANDHU, V., SOOD, K., LIM, J., RINDERER, T., ALLSOPP, M. & OLDROYD, B. 2015. Assessing patterns of admixture and ancestry in Canadian honey bees. *Insectes sociaux,* 62**,** 479-489.

HARPUR, B. A., DEY, A., ALBERT, J. R., PATEL, S., HINES, H. M., HASSELMAN, M., PACKER, L. & ZAYED, A. 2017. Queens and workers contribute differently to adaptive evolution in bumble bees and honey bees. *Genome Biology and Evolution,* evx182.

HARPUR, B. A., GUARNA, M. M., HUXTER, E., HIGO, H., MOON, K.-M., HOOVER, S. E., IBRAHIM, A., MELATHOPOULOS, A. P., DESAI, S., CURRIE, R. W., PERNAL, S. F., FOSTER, L. J. & ZAYED, A. 2019. Integrative Genomics Reveals the Genetics and Evolution of the Honey Bee's Social Immune System. *Genome Biology and Evolution,* 11**,** 937-948.

HARPUR, B. A., KADRI, S. M., ORSI, R. O., WHITFIELD, C. W. & ZAYED, A. 2020. Defense Response in Brazilian Honey Bees (*Apis mellifera scutellata* × spp.) Is Underpinned by Complex Patterns of Admixture. *Genome biology and evolution,* 12**,** 1367-1377.

HARPUR, B. A., KENT, C. F., MOLODTSOVA, D., LEBON, J. M. D., ALQARNI, A. S., OWAYSS, A. A. & ZAYED, A. 2014. Population genomics of the 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**,** 2614-9.

HARPUR, B. A., MINAEI, S., KENT, C. F. & ZAYED, A. 2012. Management increases genetic diversity of honey bees via admixture. *Molecular Ecology,* 21**,** 4414-4421.

HARPUR, B. A., MINAEI, S., KENT, C. F. & ZAYED, A. 2013. Admixture increases diversity in managed honey bees: Reply to De la Rúa et al.(2013). *Molecular ecology,* 22**,** 3211-3215.

HARRIS, A. M. & DEGIORGIO, M. 2017. Admixture and ancestry inference from ancient and modern samples through measures of population genetic drift. *Human Biology,* 89**,** 21-46.

HEDRICK, P. W. 2013. Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. *Molecular ecology,* 22**,** 4606-4618.

HENRIQUES, D., BROWNE, K. A., BARNETT, M. W., PAREJO, M., KRYGER, P., FREEMAN, T. C., MUÑOZ, I., GARNERY, L., HIGHET, F. & JONHSTON, J. S. 2018a. High sample throughput genotyping for estimating C-lineage introgression in the dark honeybee: an accurate and cost-effective SNP-based tool. *Scientific Reports,* 8**,** 8552.

HENRIQUES, D., PAREJO, M., VIGNAL, A., WRAGG, D., WALLBERG, A., WEBSTER, M. T. & PINTO, M. A. 2018b. Developing reduced SNP assays from whole-genome sequence data to estimate introgression in an organism with complex genetic patterns, the Iberian honeybee (*Apis mellifera iberiensis*). *Evolutionary Applications,* 0**,** 1-13.

HOBAN, S., KELLEY, J. L., LOTTERHOS, K. E., ANTOLIN, M. F., BRADBURD, G., LOWRY, D. B., POSS, M. L., REED, L. K., STORFER, A. & WHITLOCK, M. C. 2016. Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. *The American Naturalist,* 188**,** 379-397.

HU, H., BIENEFELD, K., WEGENER, J., ZAUTKE, F., HAO, Y., FENG, M., HAN, B., FANG, Y., WUBIE, A. J. & LI, J. 2016. Proteome analysis of the hemolymph, mushroom body, and antenna provides novel insight into honeybee resistance against varroa infestation. *Journal of proteome research,* 15**,** 2841-2854.

HUANG, D. W., SHERMAN, B. T. & LEMPICKI, R. A. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols,* 4**,** 44-57.

HUNG, K.-L. J., KINGSTON, J. M., ALBRECHT, M., HOLWAY, D. A. & KOHN, J. R. 2018. The worldwide importance of honey bees as pollinators in natural habitats. *Proceedings of the Royal Society B: Biological Sciences,* 285**,** 20172140.

HUXEL, G. R. 1999. Rapid displacement of native species by invasive species: effects of hybridization. *Biological conservation,* 89**,** 143-152.

ILYASOV, R. A., LEE, M.-L., TAKAHASHI, J.-I., KWON, H. W. & NIKOLENKO, A. G. 2020. A revision of subspecies structure of western honey bee Apis mellifera. *Saudi Journal of Biological Sciences,* 27**,** 3615.

JARNEVICH, C. S., ESAIAS, W. E., MA, P. L., MORISETTE, J. T., NICKESON, J. E., STOHLGREN, T. J., HOLCOMBE, T. R., NIGHTINGALE, J. M., WOLFE, R. E. & TAN, B. 2014. Regional distribution models with lack of proximate predictors: A fricanized honeybees expanding north. *Diversity and distributions,* 20**,** 193-201.

JI, Y. 2021. The geographical origin, refugia, and diversification of honey bees (Apis spp.) based on biogeography and niche modeling. *Apidologie,* 52**,** 367-377.

JI, Y., LI, X., JI, T., TANG, J., QIU, L., HU, J., DONG, J., LUO, S., LIU, S. & FRANDSEN, P. B. 2020. Gene reuse facilitates rapid radiation and independent adaptation to diverse habitats in the Asian honeybee. *Science Advances,* 6**,** eabd3590.

JOMBART, T. 2008. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics,* 24**,** 1403-1405.

KADRI, S. M., HARPUR, B. A., ORSI, R. O. & ZAYED, A. 2016. A variant reference data set for the Africanized honeybee, Apis mellifera. *Scientific data,* 3**,** 160097.

KENT, R. B. 1988. The introduction and diffusion of the African honeybee in South America. *Yearbook of the Association of Pacific Coast Geographers,* 50**,** 21-43.

KERR, W. E. 1967. The history of the introduction of African bees to Brazil. *South African Bee Journal,* 39**,** 3-5.

KHALIFA, S. A., ELSHAFIEY, E. H., SHETAIA, A. A., EL-WAHED, A. A. A., ALGETHAMI, A. F., MUSHARRAF, S. G., ALAJMI, M. F., ZHAO, C., MASRY, S. H. & ABDEL-DAIM, M. M. 2021. Overview of bee pollination and its economic value for crop production. *Insects,* 12**,** 688.

KIM, B. Y., HUBER, C. D. & LOHMUELLER, K. E. 2018. Deleterious variation shapes the genomic landscape of introgression. *PLoS Genetics,* 14**,** e1007741.

KOLBE, J. J., LARSON, A., LOSOS, J. B. & DE QUEIROZ, K. 2008. Admixture determines genetic diversity and population differentiation in the biological invasion of a lizard species. *Biology letters,* 4**,** 434-437.

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

KOTTHOFF, U., WAPPLER, T. & ENGEL, M. S. 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.

LAMM, K. S. & REDELINGS, B. D. 2009. Reconstructing ancestral ranges in historical biogeography: properties and prospects. *Journal of Systematics and Evolution,* 47**,** 369-382.

LANDIS, M. J., MATZKE, N. J., MOORE, B. R. & HUELSENBECK, J. P. 2013. Bayesian analysis of biogeography when the number of areas is large. *Systematic biology,* 62**,** 789-804.

LATTORFF, H. M. G., BUCHHOLZ, J., FRIES, I. & MORITZ, R. F. 2015. A selective sweep in a Varroa destructor resistant honeybee (*Apis mellifera*) population. *Infection, Genetics and Evolution,* 31**,** 169-176.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G. & DURBIN, R. 2009. 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and samtools. *Bioinformatics,* 25**,** 2078-2079.

LI, Z., HE, J., YU, T., CHEN, Y., HUANG, W.-F., HUANG, J., ZHAO, Y., NIE, H. & SU, S. 2019. Transcriptional and physiological responses of hypopharyngeal glands in honeybees (Apis mellifera L.) infected by Nosema ceranae. *Apidologie,* 50**,** 51-62.

LIAO, Y., SMYTH, G. K. & SHI, W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics,* 30**,** 923-930.
LIN, W., MCBROOME, J., REHMAN, M. & JOHNSON, B. R. 2018. Africanized bees extend their distribution in California. *PloS one,* 13**,** e0190604.
LINCK, E. & BATTEY, C. 2019. Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Molecular Ecology Resources,* 19**,** 639-647.
LIU, H., JIA, Y., SUN, X., TIAN, D., HURST, L. D. & YANG, S. 2016. Direct determination of the mutation rate in the bumblebee reveals evidence for weak recombination-associated mutation and an approximate rate constancy in insects. *Molecular biology and evolution,* 34**,** 119-130.
LIU, Y., NYUNOYA, T., LENG, S., BELINSKY, S. A., TESFAIGZI, Y. & BRUSE, S. 2013. Softwares and methods for estimating genetic ancestry in human populations. *Human genomics,* 7**,** 1.
LO, N., GLOAG, R. S., ANDERSON, D. L. & OLDROYD, B. P. 2010. A molecular phylogeny of the genus *Apis* suggests that the Giant Honey Bee of the Philippines, *A. breviligula* Maa, and the Plains Honey Bee of southern India, *A. indica* Fabricius, are valid species. *Systematic Entomology,* 35**,** 226-233.
LOVE, M. I., HUBER, W. & ANDERS, S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology,* 15**,** 550.
MALLET, J. 2005. Hybridization as an invasion of the genome. *Trends in ecology & evolution,* 20**,** 229-237.
MANJON, C., TROCZKA, B. J., ZAWORRA, M., BEADLE, K., RANDALL, E., HERTLEIN, G., SINGH, K. S., ZIMMER, C. T., HOMEM, R. A. & LUEKE, B. 2018. Unravelling the molecular determinants of bee sensitivity to neonicotinoid insecticides. *Current biology,* 28**,** 1137-1143. e5.
MATZKE, N. J. 2013a. BioGeoBEARS: BioGeography with Bayesian (and likelihood) evolutionary analysis in R Scripts. *R package, version 0.2,* 1**,** 2013.
MATZKE, N. J. 2013b. Probabilistic historical biogeography: new models for founder-event speciation, imperfect detection, and fossils allow improved accuracy and model-testing. *Frontiers of Biogeography,* 5.
MCKEIGUE, P. M. 2005. Prospects for Admixture Mapping of Complex Traits. *The American Journal of Human Genetics,* 76**,** 1-7.
MEIRMANS, P. G. & HEDRICK, P. W. 2011. Assessing population structure: FST and related measures. *Molecular ecology resources,* 11**,** 5-18.
MITREČIĆ, D., MAVRIĆ, S. & GAJOVIĆ, S. 2008. PCR-based identification of short deletion/insertions and single nucleotide substitutions in genotyping of splotch (Pax3sp) and truncate (Nototc) mouse mutants. *Molecular and cellular probes,* 22**,** 110-114.
MOLODTSOVA, D., HARPUR, B. A., KENT, C. F., SEEVANANTHAN, K. & ZAYED, A. 2014. Pleiotropy constrains the evolution of protein but not regulatory sequences in a transcription regulatory network influencing complex social behaviors. *Frontiers in Genetics,* 5**,** 1-7.
MOMENI, J., PAREJO, M., NIELSEN, R. O., LANGA, J., MONTES, I., PAPOUTSIS, L., FARAJZADEH, L., BENDIXEN, C., CĂUIA, E. & CHARRIÈRE, J.-D. 2021. Authoritative subspecies diagnosis tool for European honey bees based on ancestry informative SNPs. *BMC genomics,* 22**,** 1-12.

MONDET, F., BEAUREPAIRE, A., MCAFEE, A., LOCKE, B., ALAUX, C., BLANCHARD, S., DANKA, B. & LE CONTE, Y. 2020a. Honey bee survival mechanisms against the parasite Varroa destructor: a systematic review of phenotypic and genomic research efforts. *International journal for parasitology,* 50**,** 433-447.

MONDET, F., BEAUREPAIRE, A., MCAFEE, A., LOCKE, B., ALAUX, C., BLANCHARD, S., DANKA, B. & YVES, L. C. 2020b. Honey bee survival mechanisms against the parasite Varroa destructor: a systematic review of phenotypic and genomic research efforts. *International journal for parasitology,* 50**,** 433-447.

MUHLFELD, C. C., KALINOWSKI, S. T., MCMAHON, T. E., TAPER, M. L., PAINTER, S., LEARY, R. F. & ALLENDORF, F. W. 2009. Hybridization rapidly reduces fitness of a native trout in the wild. *Biology Letters***,** rsbl. 2009.0033.

MULLEN, E. K. & THOMPSON, G. J. 2015. Understanding honey bee worker self-sacrifice: a conceptual–empirical framework. *Advances in insect physiology,* 48**,** 325-354.

MUÑOZ, I., HENRIQUES, D., JARA, L., JOHNSTON, J. S., CHÁVEZ-GALARZA, J., DE LA RÚA, P. & PINTO, M. A. 2017. SNPs selected by information content outperform randomly selected microsatellite loci for delineating genetic identification and introgression in the endangered dark European honeybee (*Apis mellifera mellifera*). *Molecular ecology resources,* 17**,** 783-795.

MUÑOZ, I., HENRIQUES, D., JOHNSTON, J. S., CHÁVEZ-GALARZA, J., KRYGER, P. & PINTO, M. A. 2015. Reduced SNP panels for genetic identification and introgression analysis in the dark honey bee (*Apis mellifera mellifera*). *PloS one,* 10**,** e0124365.

NEI, M. & LI, W.-H. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences,* 76**,** 5269-5273.

NELSON, R. M., WALLBERG, A., SIMÕES, Z. L. P., LAWSON, D. J. & WEBSTER, M. T. 2017. Genome-wide analysis of admixture and adaptation in the Africanized honeybee. *Molecular Ecology,* 26.

NEOV, B., SHUMKOVA, R., PALOVA, N. & HRISTOV, P. 2021. The health crisis in managed honey bees (Apis mellifera). Which factors are involved in this phenomenon? *Biologia,* 76**,** 2173-2180.

PARADIS, E., CLAUDE, J. & STRIMMER, K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics,* 20**,** 289-290.

PAREJO, M., WRAGG, D., GAUTHIER, L., VIGNAL, A., NEUMANN, P. & NEUDITSCHKO, M. 2016. Using Whole-genome Sequence Information to foster Conservation Efforts for the European Dark Honey Bee, *Apis mellifera mellifera*. *Frontiers in Ecology and Evolution,* 4**,** 140.

PATTERSON, N., MOORJANI, P., LUO, Y., MALLICK, S., ROHLAND, N., ZHAN, Y., GENSCHORECK, T., WEBSTER, T. & REICH, D. 2012. Ancient admixture in human history. *Genetics,* 192**,** 1065-1093.

PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R. & DUBOURG, V. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research,* 12**,** 2825-2830.

PEREIRA, R., PHILLIPS, C., PINTO, N., SANTOS, C., SANTOS, S. E. B. D., AMORIM, A., CARRACEDO, Á. & GUSMÃO, L. 2012. Straightforward inference of ancestry and admixture

proportions through ancestry-informative insertion deletion multiplexing. *PloS one,* 7**,** e29684.
PETTIS, J. S. & DELAPLANE, K. S. 2010. Coordinated responses to honey bee decline in the USA. *Apidologie,* 41**,** 256-263.
PICKRELL, J. K. & PRITCHARD, J. K. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet,* 8**,** e1002967.
PINTO, M. A., HENRIQUES, D., CHÁVEZ-GALARZA, J., KRYGER, P., GARNERY, L., VAN DER ZEE, R., DAHLE, B., SOLAND-RECKEWEG, G., DE LA RÚA, P. & DALL'OLIO, R. 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., HENRIQUES, D., NETO, M., GUEDES, H., MUÑOZ, I., AZEVEDO, J. C. & DE LA RÚA, P. 2013. Maternal diversity patterns of Ibero-Atlantic populations reveal further complexity of Iberian honeybees. *Apidologie,* 44**,** 430-439.
PINTO, M. A., RUBINK, W. L., PATTON, J. C., COULSON, R. N. & JOHNSTON, J. S. 2005. Africanization in the United States: replacement of feral European honeybees (Apis mellifera L.) by an African hybrid swarm. *Genetics,* 170**,** 1653-1665.
POPLIN, R., RUANO-RUBIO, V., DEPRISTO, M. A., FENNELL, T. J., CARNEIRO, M. O., VAN DER AUWERA, G. A., KLING, D. E., GAUTHIER, L. D., LEVY-MOONSHINE, A. & ROAZEN, D. 2017. Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv***,** 201178.
PORRINI, L. P., QUINTANA, S., BRASESCO, C., PORRINI, M. P., GARRIDO, P. M., EGUARAS, M. J., MÜLLER, F. & FERNANDEZ IRIARTE, P. 2019. Southern limit of Africanized honey bees in Argentina inferred by mtDNA and wing geometric morphometric analysis. *Journal of Apicultural Research***,** 1-10.
POTTS, S. G., ROBERTS, S. P., DEAN, R., MARRIS, G., BROWN, M. A., JONES, R., NEUMANN, P. & SETTELE, J. 2010. Declines of managed honey bees and beekeepers in Europe. *Journal of apicultural research,* 49**,** 15-22.
R CORE TEAM 2013. R: A language and environment for statistical computing. 201.
RAFFIUDIN, R. & CROZIER, R. H. 2007. Phylogenetic analysis of honey bee behavioral evolution. *Molecular phylogenetics and evolution,* 43**,** 543-552.
RAGHAVAN, M., STEINRÜCKEN, M., HARRIS, K., SCHIFFELS, S., RASMUSSEN, S., DEGIORGIO, M., ALBRECHTSEN, A., VALDIOSERA, C., ÁVILA-ARCOS, M. C. & MALASPINAS, A.-S. 2015. Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science,* 349**,** aab3884.
RANGEL, J., GIRESI, M., PINTO, M. A., BAUM, K. A., RUBINK, W. L., COULSON, R. N. & JOHNSTON, J. S. 2016. Africanization of a feral honey bee (Apis mellifera) population in South Texas: does a decade make a difference? *Ecology and evolution,* 6**,** 2158-2169.
REE, R. H. 2005. Detecting the historical signature of key innovations using stochastic models of character evolution and cladogenesis. *Evolution,* 59**,** 257-265.
REE, R. H. & SMITH, S. A. 2008. Maximum likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis. *Systematic biology,* 57**,** 4-14.
RINDERER, T. E., HARRIS, J. W., HUNT, G. J. & DE GUZMAN, L. I. 2010. Breeding for resistance to Varroa destructor in North America. *Apidologie,* 41**,** 409-424.

RONAI, I., OLDROYD, B. & VERGOZ, V. 2016. Queen pheromone regulates programmed cell death in the honey bee worker ovary. *Insect Molecular Biology,* 25**,** 646-652.
RONQUIST, F. 1997. Dispersal-vicariance analysis: a new approach to the quantification of historical biogeography. *Systematic Biology,* 46**,** 195-203.
RUTTNER, F. 1988. *Biogeography and Taxonomy of Honeybees,* Germany, Springer Berlin Heidelberg.
RUTTNER, F., TASSENCOURT, L. & LOUVEAUX, J. 1978. Biometrical-statistical analysis of the geographic variability of *Apis mellifera* L. *Apidologie,* 9**,** 363-381.
SAELAO, P., SIMONE-FINSTROM, M., AVALOS, A., BILODEAU, L., DANKA, R., DE GUZMAN, L., RINKEVICH, F. & TOKARZ, P. 2020. Genome-wide patterns of differentiation within and among US commercial honey bee stocks. *BMC genomics,* 21**,** 1-12.
SAKAGUCHI, S., TAKEUCHI, Y., YAMASAKI, M., SAKURAI, S. & ISAGI, Y. 2011. Lineage admixture during postglacial range expansion is responsible for the increased gene diversity of Kalopanax septemlobus in a recently colonised territory. *Heredity,* 107**,** 338-348.
SANTOS, C., PHILLIPS, C., OLDONI, F., AMIGO, J., FONDEVILA, M., PEREIRA, R., CARRACEDO, Á. & LAREU, M. V. 2015. Completion of a worldwide reference panel of samples for an ancestry informative Indel assay. *Forensic Science International: Genetics,* 17**,** 75-80.
SCHNEIDER, S. S., DEGRANDI-HOFFMAN, G. & SMITH, D. R. 2004. The African honey bee: factors contributing to a successful biological invasion. *Annual Reviews in Entomology,* 49**,** 351-376.
SCHRIDER, D. R. & KERN, A. D. 2018. Supervised machine learning for population genetics: a new paradigm. *Trends in Genetics,* 34**,** 301-312.
SEDLAZECK, F. J., RESCHENEDER, P. & VON HAESELER, A. 2013. NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics***,** btt468.
SEEHAUSEN, O. 2004. Hybridization and adaptive radiation. *Trends in ecology & evolution,* 19**,** 198-207.
SEELEY, T. D. 1983. The ecology of temperate and tropical honeybee societies: Ecological studies complement physiological ones, offering a new perspective on patterns of honeybee adaptation. *American Scientist,* 71**,** 264-272.
SHEPPARD, W. 1989. A history of the introduction of honey bee races into the United States. Part 1. *American Bee Journal,* 129**,** 617-619.
SHEPPARD, W. S. & SMITH, D. R. 2000. Identification of African-derived bees in the Americas: a survey of methods. *Annals of the Entomological society of America,* 93**,** 159-176.
SHRINER, D. 2013. Overview of admixture mapping. *Current protocols in human genetics***,** 1.23. 1-1.23. 8.
SMITH, K. M., LOH, E. H., ROSTAL, M. K., ZAMBRANA-TORRELIO, C. M., MENDIOLA, L. & DASZAK, P. 2013. Pathogens, pests, and economics: drivers of honey bee colony declines and losses. *EcoHealth,* 10**,** 434-445.
STAMATAKIS, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics,* 30**,** 1312-1313.
STOHLGREN, T. J., SZALANSKI, A. L., GASKIN, J., YOUNG, N., WEST, A., JARNEVICH, C. S. & TRIPODI, A. 2014. From hybrid swarms to swarms of hybrids. *Environment and Ecology Research,* 2**,** 1-318.

TANG, H., CORAM, M., WANG, P., ZHU, X. & RISCH, N. 2006. Reconstructing genetic ancestry blocks in admixed individuals. *The American Journal of Human Genetics,* 79**,** 1-12.

TARPY, D. R., VANENGELSDORP, D. & PETTIS, J. S. 2013. Genetic diversity affects colony survivorship in commercial honey bee colonies. *Naturwissenschaften,* 100**,** 723-728.

TAVARES, H., WHIBLEY, A., FIELD, D. L., BRADLEY, D., COUCHMAN, M., COPSEY, L., ELLEOUET, J., BURRUS, M., ANDALO, C. & LI, M. 2018. Selection and gene flow shape genomic islands that control floral guides. *Proceedings of the National Academy of Sciences,* 115**,** 11006-11011.

TSVETKOV, N., BAHIA, S., CALLA, B., BERENBAUM, M. R. & ZAYED, A. In review Genetics of neonicotinoid tolerance in honey bees. *Current Biology*.

TSVETKOV, N., SAMSON-ROBERT, O., SOOD, K., PATEL, H., MALENA, D., GAJIWALA, P., MACIUKIEWICZ, P., FOURNIER, V. & ZAYED, A. 2017. Chronic exposure to neonicotinoids reduces honey bee health near corn crops. *Science,* 356**,** 1395-1397.

UNTERGASSER, A., CUTCUTACHE, I., KORESSAAR, T., YE, J., FAIRCLOTH, B. C., REMM, M. & ROZEN, S. G. 2012. Primer3—new capabilities and interfaces. *Nucleic acids research,* 40**,** e115-e115.

VÄLI, Ü., BRANDSTRÖM, M., JOHANSSON, M. & ELLEGREN, H. 2008. Insertion-deletion polymorphisms (indels) as genetic markers in natural populations. *BMC genetics,* 9**,** 1-8.

VAN ALPHEN, J. J. & FERNHOUT, B. J. 2020. Natural selection, selective breeding, and the evolution of resistance of honeybees (Apis mellifera) against Varroa. *Zoological letters,* 6**,** 1-20.

VAN DER AUWERA, G. A., CARNEIRO, M. O., HARTL, C., POPLIN, R., DEL ANGEL, G., LEVY-MOONSHINE, A., JORDAN, T., SHAKIR, K., ROAZEN, D. & THIBAULT, J. 2013. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current protocols in bioinformatics,* 43**,** 11.10. 1-11.10. 33.

VAN DER AUWERA, G. A. & O'CONNOR, B. D. 2020. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*, O'Reilly Media.

VAN DOREN, B. M., CAMPAGNA, L., HELM, B., ILLERA, J. C., LOVETTE, I. J. & LIEDVOGEL, M. 2017. Correlated patterns of genetic diversity and differentiation across an avian family. *Molecular Ecology,* 26**,** 3982-3997.

VITTI, J. J., GROSSMAN, S. R. & SABETI, P. C. 2013. Detecting natural selection in genomic data. *Annual review of genetics,* 47**,** 97-120.

VONHOLDT, B. M., BRZESKI, K. E., WILCOVE, D. S. & RUTLEDGE, L. Y. 2018. Redefining the role of admixture and genomics in species conservation. *Conservation Letters,* 11**,** e12371.

WALLBERG, A., BUNIKIS, I., PETTERSSON, O. V., MOSBECH, M.-B., CHILDERS, A. K., EVANS, J. D., MIKHEYEV, A. S., ROBERTSON, H. M., ROBINSON, G. E. & WEBSTER, M. T. 2019. A hybrid de novo genome assembly of the honeybee, *Apis mellifera*, with chromosome-length scaffolds. *BMC genomics,* 20**,** 275.

WALLBERG, A., HAN, F., WELLHAGEN, G., DAHLE, B., KAWATA, M., HADDAD, N., SIMÕES, Z. L. P., ALLSOPP, M. H., KANDEMIR, I., DE LA RÚA, P., PIRK, C. W. & WEBSTER, M. T. 2014. A worldwide survey of genome sequence variation provides insight into the evolutionary history of the honeybee *Apis mellifera*. *Nature genetics,* 46**,** 1081-1088.

WALLBERG, A., SCHÖNING, C., WEBSTER, M. T. & HASSELMANN, M. 2017. Two extended haplotype blocks are associated with adaptation to high altitude habitats in East African honey bees. *PLoS genetics,* 13**,** e1006792.

WATTERSON, G. 1975. On the number of segregating sites in genetical models without recombination. *Theoretical population biology,* 7**,** 256-276.

WEIR, B. S. & COCKERHAM, C. C. 1984. Estimating F-statistics for the analysis of population structure. *evolution,* 38**,** 1358-1370.

WHITFIELD, C. W., BEHURA, S. K., BERLOCHER, S. H., CLARK, A. G., JOHNSTON, J. S., SHEPPARD, W. S., SMITH, D. R., SUAREZ, A. V., WEAVER, D. & TSUTSUI, N. D. 2006. Thrice out of Africa: ancient and recent expansions of the honey bee, *Apis mellifera*. *Science,* 314**,** 642-645.

WILLING, E.-M., DREYER, C. & VAN OOSTERHOUT, C. 2012. Estimates of Genetic Differentiation Measured by FST Do Not Necessarily Require Large Sample Sizes When Using Many SNP Markers. *PLoS ONE,* 7**,** e42649.

WILM, A., AW, P. P. K., BERTRAND, D., YEO, G. H. T., ONG, S. H., WONG, C. H., KHOR, C. C., PETRIC, R., HIBBERD, M. L. & NAGARAJAN, N. 2012. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic acids research,* 40**,** 11189-11201.

WILSON, E. O. 1971. *The insect societies,* Cambridge, Massachusetts, Harvard University Press.

WINKLER, C. A., NELSON, G. W. & SMITH, M. W. 2010. Admixture mapping comes of age. *Annual review of genomics and human genetics,* 11**,** 65-89.

WINSTON, M. L. 1992. The biology and management of Africanized honey bees. *Annual review of entomology,* 37**,** 173-193.

WINSTON, M. L., TAYLOR, O. R. & OTIS, G. W. 1983. Some differences between temperate European and tropical African and South American honeybees. *Bee World,* 64**,** 12-21.

WRAGG, D., MARTI-MARIMON, M., BASSO, B., BIDANEL, J.-P., LABARTHE, E., BOUCHEZ, O., LE CONTE, Y. & VIGNAL, A. 2016. Whole-genome resequencing of honeybee drones to detect genomic selection in a population managed for royal jelly. *Scientific reports,* 6**,** 27168.

YANG, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution,* 24**,** 1586-1591.

ZANNI, V., GALBRAITH, D. A., ANNOSCIA, D., GROZINGER, C. M. & NAZZI, F. 2017. Transcriptional signatures of parasitization and markers of colony decline in Varroa-infested honey bees (*Apis mellifera*). *Insect biochemistry and molecular biology,* 87**,** 1-13.

ZAUMSEGEL, D., ROTHSCHILD, M. A. & SCHNEIDER, P. M. 2013. A 21 marker insertion deletion polymorphism panel to study biogeographic ancestry. *Forensic Science International: Genetics,* 7**,** 305-312.

ZAYED, A. & WHITFIELD, C. W. 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,* 105**,** 3421-3426.

ZHENG, X., LEVINE, D., SHEN, J., GOGARTEN, S. M., LAURIE, C. & WEIR, B. S. 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics,* 28**,** 3326-3328.

Appendix A: Statement on contributions

Chapter 1: Introduction
Authors: Kathleen A. Dogantzis

The introductory chapter was written by K.A.D

Chapter 2: Thrice out of Asia and the adaptative radiation of the western honey bee
Kathleen A. Dogantzis, Tanushree Tiwari, Ida M. Conflitti, Alivia Dey, Harland M. Patch, Elliud M. Muli, Lionel Garnery, Charles W. Whitfield, Eckart Stolle, Abdulaziz S. Alqarni, Michael H Allsopp, Amro Zayed

K.A.D. and A.Z. prepared and wrote the manuscript. K.A.D. carried out formal analyses and data visualization. K.A.D. and T.T. finalized sequenced data processing. I.M.C. and A.D. preformed resource and sequencing preparation. H.M.P., E.M.M., L.G., C.W.W., E.S., A.S.A., and M.H.A. provided sample resources. All authors provided comments on the manuscript.

Chapter 3: Patterns of admixture in Canadian honey bees are associated with genetic diversity and colony phenotypes.
Authors: Kathleen A. Dogantzis, Tanushree Tiwari, Rodney Richardson, Clement Kent, Stephen Rose, Alivia Dey, Ida Conflitti, Abdulaziz S. Alqarni, Harland M. Patch, Shelley E. Hoover, Robert W. Currie, Pierre Giovenazzo, M. Marta Guarna, Stephen F. Pernal, Leonard J. Foster, Amro Zayed.

K.A.D. prepared and wrote the manuscript. K.A.D. carried out formal analyses and data visualization. K.A.D., T.T., R.R., C.K., and S.R. finalized sequenced data processing. A.D. and I.M.C preformed resource and sequencing preparation. A.S.A., H.M.P. provided sample resources. S.E.H., R.W.C., P.G., M.M.G., S.F.P., L.J.F., and A.Z. provided sample resources, and acquired funding for the BeeOmics project. A.Z. provided comments on the manuscript.

Chapter 4: Accurate detection of Africanized bees using a SNP-based diagnostic assay.
Authors: Kathleen A. Dogantzis, Ida Conflitti, Alivia Dey, Tanushree Tiwari, Stephen Rose, Nadine Chapman, Samir Kadri, Harland M. Patch, Elliud M. Muli, Abdulaziz S. Alqarni, Michael H Allsopp, Amro Zayed.

K.A.D. and prepared and wrote the manuscript. K.A.D. carried out formal analyses and data visualization. K.A.D., and S.R. developed machine learning models. K.A.D. and T.T. finalized sequenced data processing. I.M.C. and A.D. preformed resource and sequencing preparation. N.C., S.K., H.M.P., E.M.M., A.S.A., and M.H.A. provided sample resources. A.Z. provided comments on the manuscript and acquired funding.

Chapter 5: Developing a collection of insertion-deletion markers to identify Africanized honey bees
Authors: Kathleen A. Dogantzis, Dar'ya Semenova, Ida Conflitti, Alivia Dey, Tanushree Tiwari, Stephen Rose, Nadine Chapman, Samir Kadri, Harland M. Patch, Elliud M. Muli, Abdulaziz S. Alqarni, Michael H Allsopp, Amro Zayed

K.A.D. and prepared and wrote the manuscript. K.A.D. and D.S., carried out formal analyses and data visualization. K.A.D., and D.S., developed markers, primers, and PCR protocol. K.A.D., and S.R. developed machine learning models. I.C., and D.S performed resource preparation and genotyping. A.D preformed resource and sequencing preparation. N.C., S.K., H.M.P., E.M.M., A.S.A., and M.H.A. provided sample resources. A.Z. provided comments on the manuscript and acquired funding.

Chapter 6: Conclusion and future work
Authors: Kathleen A. Dogantzis

The concluding chapter was written by K.A.D

Sincerely,

Kathleen A. Dogantzis

Approved by,

Amro Zayed

Amro Zayed, PhD
Associate Professor of Biology