Saelao *et al. BMC Genomics* (2020) 21:704
https://doi.org/10.1186/s12864-020-07111-x

BMC Genomics

RESEARCH ARTICLE Open Access

Genome-wide patterns of differentiation within and among U.S. commercial honey bee stocks

Perot Saelao[1,2], Michael Simone-Finstrom[1*], Arian Avalos[1], Lelania Bilodeau[1], Robert Danka[1], Lilia de Guzman[1], Frank Rinkevich[1] and Philip Tokarz[1]

Abstract

Background: The population genetics of U.S. honey bee stocks remain poorly characterized despite the agricultural importance of *Apis mellifera* as the major crop pollinator. Commercial and research-based breeding programs have made significant improvements of favorable genetic traits (e.g. production and disease resistance). The variety of bees produced by artificial selection provides an opportunity to characterize the genetic diversity and regions of the genome undergoing selection in commonly managed stocks.

Results: Pooled sequencing of eight honey bee stocks found strong genetic similarity among six of the stocks. Two stocks, Pol-line and Hilo, showed significant differentiation likely due to their intense and largely closed breeding for resistance to the parasitic *Varroa* mite. Few variants were identified as being specific to any one stock, indicating potential admixture among the sequenced stocks. Juxtaposing the underlying genetic variation of stocks selected for disease- and parasite-resistance behavior, we identified genes and candidate regions putatively associated with resistance regulated by hygienic behavior.

Conclusion: This study provides important insights into the distinct genetic characteristics and population diversity of honey bee stocks used in the United States, and provides further evidence of high levels of admixture in commercially managed honey bee stocks. Furthermore, breeding efforts to enhance parasite resistance in honey bees may have created unique genetic profiles. Genomic regions of interest have been highlighted for potential future work related to developing genetic markers for selection of disease and parasite resistance traits. Due to the vast genomic similarities found among stocks in general, our findings suggest that additional data regarding gene expression, epigenetic and regulatory information are needed to more fully determine how stock phenotypic diversity is regulated.

Keywords: *Apis mellifera*, Pool-seq, F_{ST}, Single nucleotide polymorphism, Genetic diversity, Stock identification

Background

Managing genetic diversity is critical for the success and sustainability of breeding programs of domesticated organisms. Retaining enough genetic diversity while simultaneously applying selection for desirable traits is a balance that ensures future success through reservoirs of standing variation and minimizes the likelihood of genetic bottlenecks [1–3]. Honey bees (*Apis mellifera*) are the predominant managed insect pollinator for a majority of food crop species and depend on high genetic diversity within colonies and between populations. High intracolony diversity in honey bees promotes fitness [4–7], productivity [8, 9] and colony survival [10]. At the population level, low genetic diversity may contribute to the long-term declines of

* Correspondence: michael.simonefinstrom@usda.gov
[1]USDA-ARS, Honey Bee Breeding, Genetics and Physiology Laboratory, Baton Rouge, LA 70820, USA
Full list of author information is available at the end of the article

managed colony numbers, and to the more recent trends of high annual mortality and poor health of colonies [5, 11, 12].

There are currently a number of honey bee stocks (supposedly isolated populations bred for desirable traits) produced by queen breeders for widespread use across the United States. The genetic background favored by the majority of U.S. beekeepers is derived primarily from the subspecies *A. m. ligustica*; such bees are generally referred to as Italian honey bees. Over the years, Italian stocks have been selected primarily for honey production, colony size, and body color [13]. Other widely used stocks are derived from the related subspecies *A. m. carnica*, known as Carniolan bees. The majority of queens that are commercially produced in the U.S. originate from a relatively small number of queen producers in California, Hawaii and the southern U.S. It has been suggested by previous work that a combined 500 breeder queens across operations are used to rear over a million daughter queens that are then distributed throughout the U.S. [13]. While these do not all represent genetically distinct lines, the varied sources contribute to genetic diversity among stocks. In addition, there are also less frequently used stocks developed for more specialized traits or as endpoints of research-based breeding programs such as Russian, Minnesota Hygienic, Pol-line, Mite-biter and Saskatraz [14–19]. The breeding schemes for these stocks vary. For example, Russian bees are bred in a highly structured closed breeding program, consisting of 16 distinct lines and directed crosses (via open mating) among those lines [20]. The Pol-line breeding program is also a closed system, propagating 12–16 queen lines annually, however mating occurs via instrumental insemination of pooled drone semen [21]. In contrast, Minnesota Hygienic bees are propagated based on colony phenotype, independently by commercial beekeepers [14]. The variance in breeding and population structure among managed honey bee colonies has, as of yet, lacked a comprehensive high resolution understanding of the genetic diversity and structure in managed U.S. honey bee populations. A population genetic approach to the analysis of honey bee genetic variation could prove useful by identifying regions that have undergone selection through breeding efforts which can be used as a basis for novel breeding programs. Previous work has provided genomic information about aggressive behavior, swarming, and Africanization in various honey bee populations globally [22–25]. Several studies have described the overall diversification of U.S. honey bees, with a focus on the origins of managed and feral populations [10, 13, 26–30]. A more comprehensive understanding of the genetic differentiation among commercially relevant stocks, however, is an important first step toward efforts to begin implementing genomic-based marker-assisted selection to rapidly and efficiently improve honey bee breeding programs.

We investigated the current patterns of genetic diversity of eight U.S. honey bee stocks using whole genome sequencing of pooled individuals (Pool-seq) to determine genome-wide allele frequencies. This cost-effective approach facilitates the sampling of a large number of individuals while effectively calling conserved alleles segregating within the population to better understand the wider population genetic structure [31]. We characterize the genetic diversity found within and between these stocks and identify single nucleotide polymorphisms (SNPs) that may be useful in stock identification. Furthermore, we used this information along with population demographic data to discover genomic regions likely under selection for enhanced resistance to pathogens and parasites. The overall aim was to provide an initial genetic screen to strengthen future efforts toward incorporating the use of marker-assisted selection in honey bee breeding, especially for traits that are challenging to phenotype.

Results

We evaluated seven U.S. honey bee stocks collected in 2016 and one collected in 2019 (Hilo). Four of the stocks are widely used in the beekeeping industry including three Italian (designated as Italian1, Italian2 and Italian3) and one Carniolan, representing some of the major bee breeding programs in North America. The remaining four stocks (Minnesota Hygienic, Pol-line, Hilo and Russian), are products of specialized breeding efforts. Minnesota Hygienic bees were originally derived from Italian bees and were selected for enhanced expression of hygienic behavior via a freeze-killed brood assay [32]. This trait affords resistance to several pathogens and, to a lesser extent, the parasitic mite *Varroa destructor* [32–34]. Proteomic and QTL analysis have identified several promising candidate genomic regions that may assist in future biomarker identification related to general hygienic behavior [35–37]. Pol-line honey bees were derived from Italian bees with high *Varroa* resistance largely derived from the expression of the '*Varroa*-sensitive hygiene' (VSH) trait [38]. Several studies have attempted to profile the genetic and biochemical mechanisms regulating hygienic behavior [39, 40], though the precise regulatory mechanisms managing this trait are still unknown. Hilo stock is from an ongoing breeding program that combines VSH behavior from Pol-line with Italian stocks for improved honey production and colony size. Russian bees were selected for low mite population growth of *Varroa* [19, 41], and display VSH as one mechanism of resistance [19, 42]. We considered stock-specific differences individually in our assessments of variation. However, in instances where we examined variation derived from targeted breeding efforts in the research stocks, we combined all three Italian stocks to form a common background under the assumption that their combined profile is representative of the broader genetic diversity of managed Italian bees across the U.S.

Within-stock genome-wide variation

We found ~ 1.2–1.8 million polymorphic SNPs within each stock (Table 1). Estimates of π and θ averaged over the whole genome were lowest in Italian3 (π = 4.226E-03; θ = 4.306E-03) followed closely by Pol-line (π =4.361E-03; θ = 4.377E-03). Highest estimates of π and θ diverged, with Italian2 having the highest estimate of π (5.084E-03) and Carniolan having the highest estimate of θ (5.099E-03, Table 1). Estimates of π over the *csd* locus suggest that Italian1 is the most genetically diverse across this region, while the Hilo is the least genetically diverse (Fig. 1, Supplemental Table 2).

The covariation of allele frequencies of a subset of 477,735 SNPs, where at least one variant existed among the eight stocks at a frequency > 0.1, were used to examine the genetic structure among stocks. Two analyses were used for this process: a principal component analysis to dimensionally reduce the genetic variation per sample, and a k-nearest neighbor (KNN) clustering analysis to identify broad genetic clusters among the eight stocks. The first principal component (PC1) accounted for 5.1% of the variation in allele frequency data, while PC2 and PC3 accounted for 3.9 and 2.6% of the variation, respectively. Broadly, PC1 appeared to differentiate among the Carniolan-Italian spectrum of genetic diversity while PC2 primarily differentiated Pol-line and Hilo stocks (Fig. 2a). PC3 differentiated the Italian2, Hilo, and Carniolan stocks (Fig. 2b-c). The KNN analysis identified three genetic subgroups among the eight sequenced stocks. Stocks within these subgroups were likely experiencing continued gene flow or lacking the generational time to have diverged. The KNN analysis further highlighted Pol-line and Hilo as highly differentiated from other commercial stocks (Fig. 2d).

Table 1 Pi, Theta, and the average total number of SNPs measured among the eight honey bee stocks

Stock	π		Average Total No. SNPs
Italian 1	4.998E-03	5.034E-03	1,458,101
Italian 2	5.084E-03	4.819E-03	1,465,571
Italian 3	4.226E-03	4.306E-03	1,439,926
Carniolan	4.712E-03	5.099E-03	1,422,100
Minnesota Hygienic	5.073E-03	5.035E-03	1,538,237
Pol-line	4.361E-03	4.377E-03	1,423,510
Hilo	4.848E-03	4.840E-03	1,447,778
Russian	4.700E-03	4.681E-03	1,423,499

Population differentiation

In order to investigate stock differentiation between populations from different genetic backgrounds, we estimated pairwise F_{ST} at each polymorphic site. Average genome-wide measures of F_{ST} found the greatest differentiation between Minnesota Hygienic and Hilo stocks ($F_{ST} = 0.031$), while Russian and Minnesota Hygienic stocks were all similarly less differentiated from the generalized Italian stock ($F_{ST} = 0.011$, Table 2). Pairwise F_{ST} comparisons between the generalized Italian stock and those selected for parasite and pathogen resistance (i.e., Minnesota Hygienic, Pol-line, Hilo, and Russian stocks) found that only Pol-line stock had a wider range of F_{ST} demonstrating greater differentiation from the Italian stock (Fig. 3). We found a total of 422 candidate SNPs that may be used to differentiate pairs of stocks, based on pair-wise comparisons for SNPs with a high degree of fixed differentiation ($F_{ST} > 0.65$) (Table 2) [37, 43].

To establish unique stock-specific SNPs among all of the stocks examined as possible candidates for stock identification, we analyzed SNPs with the top 1.0 and 0.1% F_{ST} values and extracted only the unique SNPs. Within the top 1.0% F_{ST} value SNPs we found a range of 5–405 SNPs that are putatively stock-specific (Table 3). However, there were 0–12 stock-specific SNPs when only the top 0.1% of FST value SNPs were considered. A full list of the stock-specific SNPs in the top 0.1% can be found in Supplemental Table 3.

Genomic regions under selection

A total of 46 haplotype blocks containing the most significant Composite Selection Signal (CSS) scores (top 0.1%) among generalized Italian, Minnesota Hygienic, Pol-line, Hilo, and Russian stocks with more than five SNPs were identified. Several regions demonstrated overlap with regions previously associated with general hygienic behavior or VSH (Table 4). A full list of the top 0.1% CSS SNPs can be found in the Supplemental Table 4 and the distribution of CSS scores can be found in Supplemental Figure 1. In the most significant CSS score regions, gene enrichment was found in categories of *integral component of membrane* (GO:0016021); and UniProtKB keywords *transmembrane, transmembrane helix*, and *membrane associated clustering*.

Discussion

This study represents the first genetic characterization of U.S. honey bee stocks. We show that within-stock genetic diversity is limited. The genetic variation that exists was largely shared among stocks with few markers found to be unique to each. We show that Pol-line and Hilo are the most distinct among the stocks, with the largest number of differentiated markers. These results confirm previous work indicating a high degree of admixture among honey

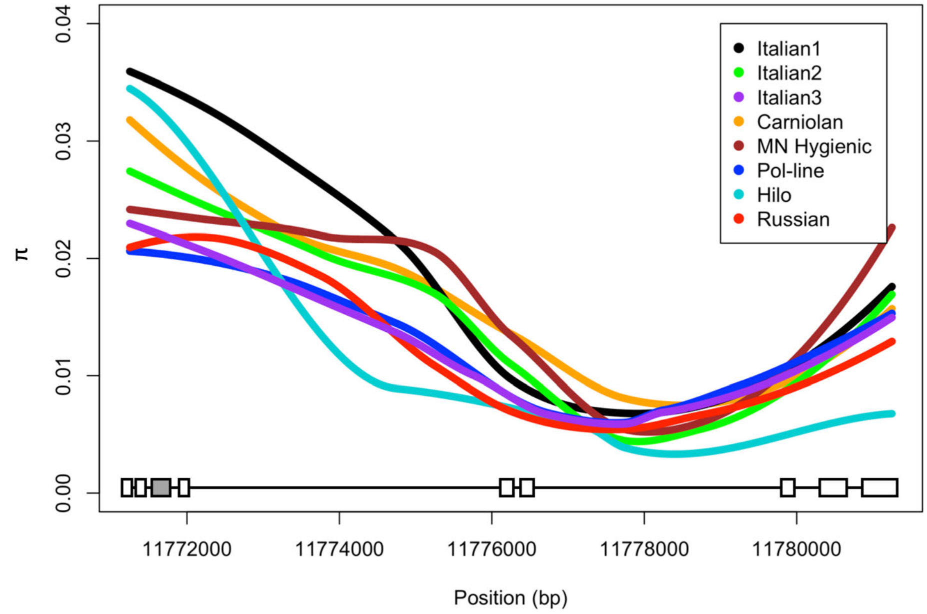

Fig. 1 Smoothed fit of nucleotide diversity (*π*) across the complementary sex determiner, *csd*, locus estimated over 250 nonoverlapping windows for each stock. *Csd* exons are depicted across the base of the figure. Exon 7, the hypervariable region, is depicted in grey

bee populations in North America [44] and that targeted breeding for very specific traits can quickly differentiate populations.

Genetic diversity metrics: within stocks

We found relatively similar, low levels of within-population genetic diversity measured by π and θ across the genome for the eight stocks. It should be noted that a few factors may potentially be influencing our findings such as the matrilineal bias influenced by honey bee multiple mating, our pooling approach, and potential non-random sampling of individuals by bee breeders. However, the differences in within-stock variation examined among stocks provides an informative, relative measure of genetic diversity. In regard to possible over-sampling of related individuals within a population, we do not observe any distinguishing effects between the commercial and highly controlled Hilo, Pol-line, and Russian populations. This suggests that within sample relationships may not be contributing significantly to the reduced genetic diversity.

Admixture appears to cause commercially managed honey bees to have greater genetic diversity than their progenitor populations [45]; we thus expected the commercial Italian and Carniolan stocks to have the highest levels of π and θ when compared to the specialized research-based stocks. However, we found inconsistent levels of genetic diversity both in the commercial and research-based stocks. Although our genome-wide estimates of π and θ show low genetic diversity within stocks, analysis of the *csd* locus provided greater resolution of recent population changes. Honey bees possess a single-locus sex determination system where homozygosity at the *csd* locus results in functionally non-viable diploid males [46]. Genetic variation at this locus has been demonstrated to be critical for colony health because of its effect on brood viability [47, 48], and has been used to calculate estimates of mating population size and broad genetic diversity [23, 47]. It is interesting to note that although the Hilo stock exhibited moderately high genome-wide π and estimates, it had the lowest such estimates at the *csd* locus. This suggests a disconnect between genomic level nucleotide diversity and that of the *csd* locus, which is possible in honey bees given their high rate of recombination and low levels of linkage-disequilibrium [46]. Further exploration of this will be informative for adequate determination and mitigation of inbreeding depression in honey bee breeding operations.

Population variation: among stocks

Differences in genetic variation among stocks identified three general clusters which principally grouped as Hilo-Pol-line, Italian1-Italian3-Russian-Carniolan, and Italian2-Minnesota Hygienic. This provides further evidence of the genetic differentiation of Hilo and Pol-line from other stocks. We posit this was largely driven by a combination of continued selective pressure for a specific, shared mite-resistant phenotype (VSH) along with relative reproductive isolation. Interestingly, though Hilo and Italian2 stocks have some shared ancestry (4 of the 12 Hilo colonies had an Italian2 maternal lineage), Hilo primarily clusters with Pol-line. This may indicate a greater contribution of the Pol-line ancestry to Hilo, with F_{ST} estimates suggesting that Hilo is more differentiated from the combined Italian

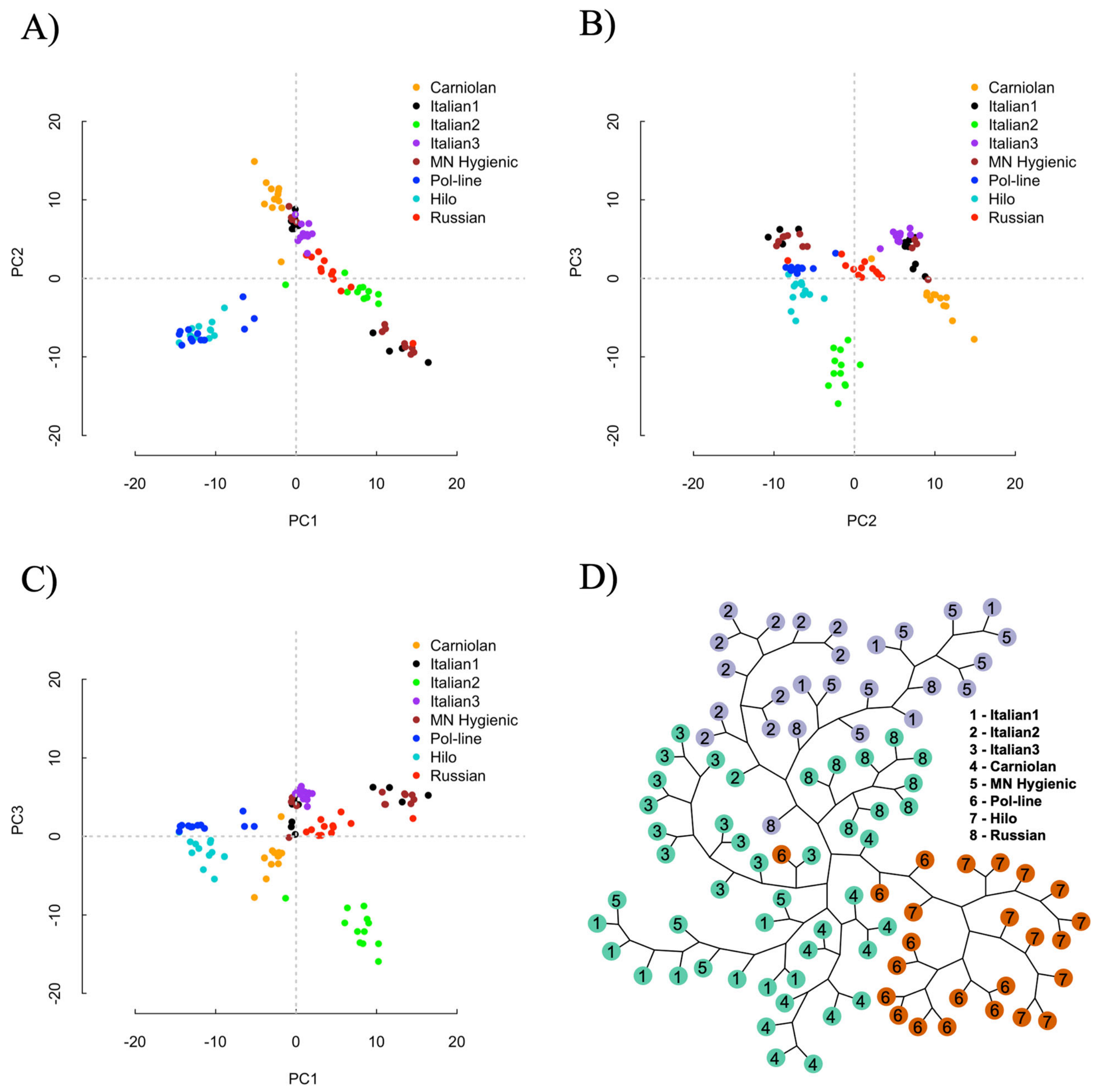

Fig. 2 Principal component analysis of allele frequencies by principal components (PC) and k-nearest neighbor (KNN) analysis of the eight sequenced honey bee stocks demonstrating the distribution of genetic relatedness. PC 1 and 2 (**a**); PC 2 and 3 (**b**), PC 1 and 3 (**c**), and KNN (**d**)

stocks ($F_{ST} = 0.024$) than they are from Pol-line ($F_{ST} = 0.016$). Additionally, these findings may reflect a paternal effect as the relationship between these stocks is through some sharing of patrilineal lineages. Other stocks that were either selected for hygienic behavior (Minnesota Hygienic) or express it at a high level without targeted selection for the trait (Russian) did not show a similar degree of differentiation from the common Italian background, potentially highlighting that VSH behavior has at least some partially unique genetic features. This may be a function of increased independent selection for hygienic behavior by bee breeders [49] in the Italian based stocks and the high likelihood of introgression from surrounding populations. This may also explain the distribution of the Russian and Minnesota Hygienic stocks across two genetic clusters determined through the KNN analysis as a result of shared selection pressure among operations.

The majority of SNPs identified in this study did not demonstrate a significant degree of differentiation between stocks. Hilo and Pol-line had the most SNPs at the 1.0% ($n = 405$) and 0.1% ($n = 12$) cutoffs, respectively. This further highlights the genetic differentiation of these

Table 2 Stock specific measures of genetic diversity within honey bees. Shaded values indicate the number of highly differentiated pair-wise ($F_{ST} > 0.65$) SNPs, while unshaded cells show the average genome-wide F_{ST} values

	Number of SNPs unique between lines ($F_{ST} > 0.65$)					
	Italian	Carniolan	MN Hygienic	Pol-line	Hilo	Russian
Italian	■	0	0	5	35	0
Carniolan	0.016	■	10	39	21	28
MN Hygienic	0.011	0.024	■	33	121	1
Pol-line	0.021	0.025	0.023	■	21	28
Hilo	0.024	0.028	0.031	0.016	■	80
Russian	0.011	0.017	0.014	0.023	0.027	■
	Average Genome-wide F_{ST}					

populations compared to other stocks, likely in response to targeted selection focused largely on one trait (VSH). The strong genetic similarity among Italian, Russian, Minnesota Hygienic, and Carniolan stocks yielded few unique SNPs that could effectively segregate the stocks at the highest F_{ST}. Interestingly, Carniolan and Russian stocks had very few SNPs differentiating them from the generalized Italian stock despite the controlled nature of their breeding programs. This supports the case that introgression from external Italian honey bee populations exists either from their population of origin (e.g. [50]) or after their introduction in the U.S. For example, while Russian honey bee stock is produced as part of a closed network of breeders [44], queens are produced via natural mating. As such, there is an increased likelihood of admixture, as has been seen in recent stock assessment [51]. Despite these populations appearing relatively genetically similar to one another, there remain significant and observable phenotypic differences between stocks. The underlying variation driving the phenotypic differences may be caused by mutations not detected in this study such as indels, inversions, or low frequency variation among polygenic traits. However, differences in and between populations were still detected despite the conservative nature of our analyses. Furthermore, differences in gene regulatory or epigenetic profiles may also be influencing differences between stocks and are potential avenues for further investigation. The findings summarized above provide several lines of evidence consistent with the genetic effects of selective breeding in combination with high levels of admixture within U.S. honey bee populations.

Signatures of selection and identification of candidate genes

CSS scores improve the power to detect and resolve selection signals and localize candidate regions involved in traits experiencing selection pressure. Regions under selection shared by Minnesota Hygienic, Pol-line, Hilo, and Russian stocks provide actionable targets for future research and breeding. Our annotation using haplotype blocks identified 46 of the 58,333 [23] that were shared among the four stocks with a strong signal of selection

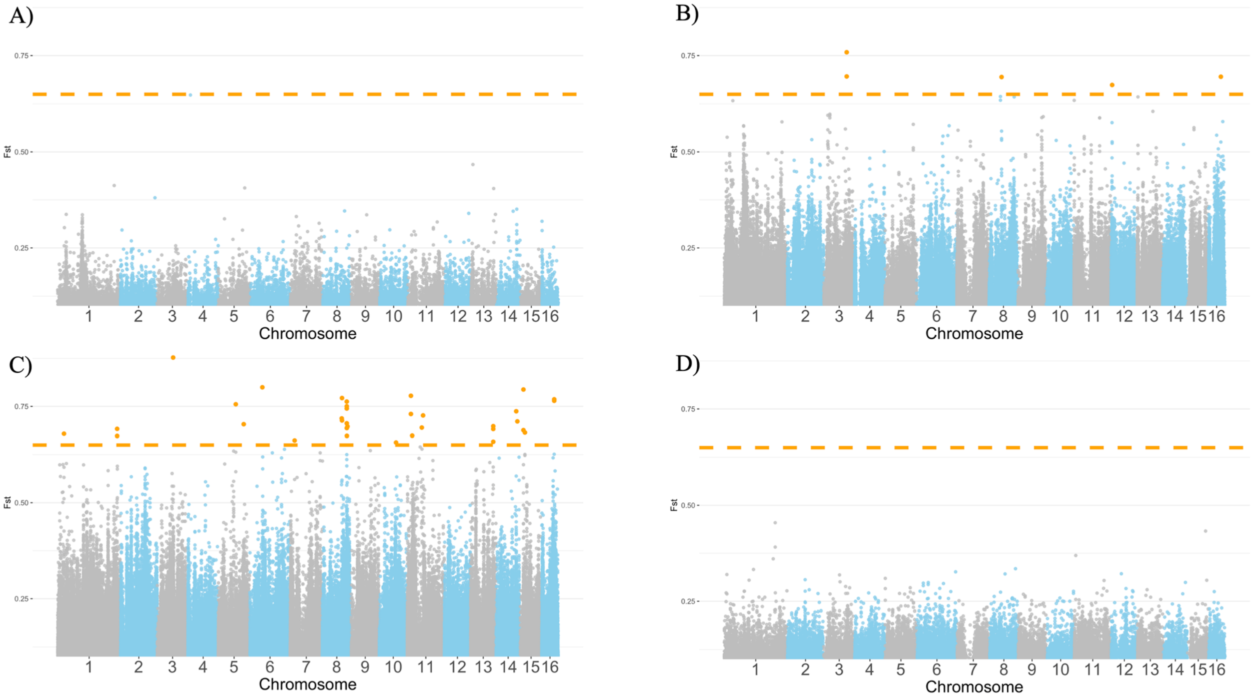

Fig. 3 Pair-wise F_{ST} measures between combined Italian stocks and **a**) Minnesota, **b**) Pol-line, **c**) Hilo, and **d**) Russian stocks bred for resistance traits. $F_{ST} > 0.65$, highlighted in orange, depicts significant fixed differentiation between the comparisons

providing evidence of a common selection signal among stocks associated with mite- and disease-resistance traits. However, there remains an unlikely possibility that a CSS signal may arise of a specific region that is highly selected in only the Italian stocks and not of the other populations. Though we feel that it is more likely that the research stocks are arriving at a shared resistance given the intent of their respective programs. Overlap between our CSS signal and previous QTLs associated with hygienic behavior and social immunity was found across LG 1, 3, 7, 8, and 9 [37, 52, 53]. Our findings further support that the active selection for behavior conferring improved resistance to parasites and pathogens, along with the increased fitness of these traits, may be shared among these lines. Further investigations to determine the functional mechanisms regulating social immunity will aid in the development of molecular tools to apply for more full integration of resistance traits among honey bee breeding programs and populations. Certainly, the successful use of proteomic markers for the selection of hygienic behavior highlights the value for the integration of functional genomics in work related to genomic markers for selection [35, 54, 55]. Given the complex regulatory network influencing disease and parasite resistance traits, significant work remains before marker assisted selection can be employed on a large-scale and breeding efforts focusing on resistance phenotypes must continue to be supported.

Table 3 Number of Stock-Specific SNPs among the Top 1.0 and 0.1% of F_{ST} values

Stock	Unique SNPs Among the Top 1.0% of F_{ST} SNPs	Unique SNPs Among the Top 0.1% of F_{ST} SNPs
Italian1	5	0
Italian2	275	10
Italian3	91	9
All Italian	0	0
Carniolan	127	4
Minnesota Hygienic	106	1
Pol-line	157	12
Hilo	405	1
Russian	26	1

Table 4 CSS regions identified to be under high positive selective pressure among Minnesota Hygienic, Pol-line, Hilo, and Russian stocks putatively for general hygienic behavior or VSH (Tsuruda et al. 2012, Splotter et al. 2012). Position information of enriched blocks, genes and overlap with previously identified QTLs are shown

Chr	Starts	Ends	SNPcount	Gene	Previously Identified Hygienic Association
LG1	2,345,989	2,346,935	6	LOC724238	
LG1	2,909,659	2,911,035	6	LOC410758	
LG1	3,761,417	3,762,423	6		Spötter et al. 2012
LG1	7,819,348	7,820,225	7		
LG1	8,889,621	8,890,832	7	LOC410732	
LG1	8,891,866	8,892,300	6	LOC410732	
LG1	8,927,676	8,928,305	6		
LG1	9,466,743	9,467,416	6	LOC409608	
LG1	9,491,950	9,492,350	10	LOC100578747, LOC409608	
LG1	21,158,620	21,159,496	6	LOC413968	
LG2	6,660,797	6,661,595	8	LOC725588	
LG2	9,716,385	9,717,106	11	LOC551772	Spötter et al. 2012
LG3	9,550,121	9,550,612	8		
LG3	10,115,687	10,116,744	10	LOC413829	Spötter et al. 2012
LG3	10,530,176	10,530,589	10	LOC410926	Spötter et al. 2012
LG3	12,691,622	12,696,809	9	LOC552149	
LG4	13,011,199	13,011,553	8	LOC551628	
LG5	6,479,299	6,479,897	8	LOC113218552, LOC411079	
LG6	13,890,152	13,890,546	7	LOC102656661	
LG6	13,905,063	13,906,782	6	LOC102656661	
LG6	14,675,405	14,675,960	6	LOC107964545	
LG7	674,945	675,266	7	LOC412896	
LG7	12,686,713	12,687,509	10	LOC408915	
LG8	2,114,043	2,120,983	6	LOC102654371	
LG8	2,188,402	2,193,281	6	crh-BP	
LG8	10,973,296	10,974,126	6		Spötter et al. 2012
LG9	10,466,886	10,467,800	13		Tsuruda et al. 2012
LG10	6,334,326	6,335,519	10		
LG10	10,804,935	10,805,598	6	LOC100577980, LOC411288	
LG11	710,227	727,890	6		
LG11	3,445,285	3,456,385	7		
LG11	11,299,503	11,300,480	6	LOC724460, LOC726180	
LG11	14,537,227	14,539,063	11	LOC102653678	
LG12	1,549,766	1,551,073	6		
LG12	2,294,313	2,298,222	7	LOC411813	
LG12	3,715,250	3,715,959	7	LOC100578102	
LG12	5,861,319	5,861,711	7	LOC724292	
LG13	849,828	851,798	6		
LG14	7,837,851	7,838,138	6	LOC410515	
LG14	8,891,482	8,892,699	10	LOC410535	
LG14	9,640,842	9,641,372	10		
LG16	261,473	262,437	6	LOC107965756	

Table 4 CSS regions identified to be under high positive selective pressure among Minnesota Hygienic, Pol-line, Hilo, and Russian stocks putatively for general hygienic behavior or VSH (Tsuruda et al. 2012, Splotter et al. 2012). Position information of enriched blocks, genes and overlap with previously identified QTLs are shown *(Continued)*

Chr	Starts	Ends	SNPcount	Gene	Previously Identified Hygienic Association
LG16	874,080	874,571	6		
LG16	2,345,631	2,346,711	11	LOC724468	
LG16	2,634,030	2,634,557	8	LOC413789	
LG16	4,418,236	4,419,113	6		

Conclusions

This study categorizes the underlying genetic variation that exists in common honey bee stocks through the use of genomics and population genetic approaches. The findings suggest the need for future efforts to integrate additional forms of data including gene expression, epigenetic, and regulatory information to provide a more complete understanding of the mechanisms regulating stock phenotypic diversity. Validation of QTLs for complex traits associated with disease and parasite resistance among honey bee populations has been extremely difficult to achieve [42]. Understanding of which genomic regions are shared or are particularly differentiated is an important step forward. Through the incorporation of the results presented here and future efforts into data repository resources such as Hymenopteramine [56], beekeepers and honey bee researchers can begin to work towards the development of marker-assisted selection to "build a better bee."

Methods

Sample preparation and sequencing

Pool-seq was used to widely sample each population and identify conserved patterns of genetic variation within each of the 8 sampled stocks. To sample, 96 worker bees were collected from 12 breeder colonies of each stock. The Carniolan, Minnesota Hygienic and three Italian stocks were randomly sampled by the bee breeders to be sent for sequencing. Given that Hilo, Pol-line, and Russian stocks are either maintained by the USDA-ARS Baton Rouge research unit or close collaborators, they were carefully sampled to ensure independent sampling within each of the populations so as to minimize relatedness between colonies. Colonies from the Russian stock were selected from 12 of the 18 breeding lines, Pol-line colonies represented 12 different lines and Hilo represents 10 different lines. Eight pools of 12 heads per colony were prepared by first excising the heads using sterile surgical scalpels. The compound eye lenses were then removed to limit introduction of magnesium residues into downstream reactions [57]. Samples were held at -20 °C prior to DNA extraction. Each pool was first homogenized using Omni BeadRuptor™ – Elite 2.0 mL Soft Tissue Kit vials pre-fitted with 1.4 mm ceramic disruption beads. Final DNA extraction was completed using either Qiagen DNeasy Blood and Tissue Kit (Qiagen USA, CAT#69506) following manufacturer recommended protocol for single tube extraction or Sbeadex® purification kits (Biosearch Technologies UK, CAT#NAP41450). The extraction method for each sample is listed in Supplementary File Table 1. Final quality check was performed using Quant-iT™ PicoGreen™ dsDNA assay kit (Invitrogen USA, CAT#P11496).

Sequencing was performed at the Institute for Genome Science at the University of Maryland School of Medicine using the Illumina HiSeq4000 150 paired-end reads. The 96 pools of samples were spread across eight lanes for a mean coverage depth of 30x per sample. Raw sequence files were trimmed using Trim Galore v0.6.0 and then aligned to the honey bee reference genome (Amel_HAv3.1) using bwa v0.7.16a to generate bam files [58]. Duplicates were marked using picard tools's MarkDuplicates (http://broadinstitute.github.io/picard/) on the trimmed fastq files followed by SortSam to sort reads. Sorted reads were then processed by GATK's BaseRecalibrator to correct for patterns of systematic errors in base quality scores using base settings [59–61].

Variant calling

SNPs were derived for our data set by using using Lofreq's "call" function [62] which produces a VCF file used in the subsequent downstream analysis. Allele frequency information was extracted and visualized from Lofreq produced VCF files using SNPRelate [63]. Variants were called for each population separately, filtered by the above quality measures.

Summary of genetic variation

A joined set of genetic variants with at least one overlapping SNP was used to conduct a principal component analysis (PCA) to summarize the covariation in allele frequencies between samples. The code used to generate the overlapping SNP set is provided in Supplemental File 1. An iterative k-means hierarchical analysis was applied to the resulting principal components (PC) to identify the optimal number of genetic clusters in our data set agnostic of population membership [64]. The Akaike Information Criterion (AIC) was used on a range of 1–10 possible clusters and the optimal number of clusters was selected using the elbow method.

Estimation of population genetic parameters and stock specific SNPs

Genome-wide patterns of differentiation between the sequenced stocks was estimated with the population genetic parameters pi (π), theta (θ), and F_{ST} [65–67]. For the within and between pairwise comparisons of the stocks, the software packages Popoolation was used to estimate π and θ [68] and Popoolation2 was used to estimate F_{ST} [69]. Both these packages used mpileup files generated though samtools from the recalibrated bam files. In estimates of π and θ, we adjusted the run parameters of Popoolation and Popoolation2 to include a minimum coverage of 20 reads, a minimum count of 10 reads, and a minimum quality threshold of 20. We also used a max-coverage threshold of no more than 2% of the total number of reads. These filters frame assured the consideration of the most conservative and consistent SNPs among our data set. A generalized Italian population was generated by merging the commercial stocks advertised as Italian in origin. SNPs were considered to have a high degree of fixation difference between pairwise comparisons of generalized Italians, Pol-line, Hilo, Russian, and Minnesota Hygienic stocks when F_{ST} was > 0.65 [43].

Stock specific SNPs were identified using the top 1.0 and 0.1% F_{ST} values from each of the pairwise comparisons of stocks. These most highly differentiated SNPs were then compared among the six stocks (Carniolan, Minnesota Hygienic, Pol-line, Hilo, Russian, and generalized Italian) to identify SNPs unique to that stock among these top candidates.

Composite selection signal score to identify signals of selection

A composite selection signal (CSS) score was used to refine and detect signatures of selection [70]. This method unifies multiple estimates of population differentiation measurements in order to capture highly differentiated loci and their respective genetic variants being enriched within the population. Briefly, a fractional rank of the F_{ST} calculations for Minnesota Hygienic-Italian, Pol-line-Italian, Hilo-Italian, and Russian-Italian comparisons were calculated and used to derive a Z-score. The Z-score was averaged for each corresponding SNP, compared to a standard distribution to derive a *P* value, and transformed using a -Log_{10}(*P* value) representing the CSS [70]. CSS outliers were identified as having a CSS score in the top 0.1% of the distribution. To assist in functional annotation, we capitalized on the small span of linkage-disequilibrium across the honey bee genome. Candidate SNPs from our CSS values were localized to previously described haplotype blocks [23]. These haplotype blocks are genomic spans with conservative estimates of linkage-disequilibrium derived across three genetically distinct populations. We considered only haplotype blocks found to have > 5 SNPs/block. The lengths of remaining haplotype blocks ranged from 287 to 17,663 base pairs. After overlap with our CSS results we arrived at 58,333 haplotype blocks with at least one SNP from our CSS analysis. This subset was used to identify genes located within genomic blocks with high CSS scores. We used DAVID functional annotation tool v6.8 and HymenopteraMine to identify clusters of gene ontology terms enriched in highly selected regions [56, 71, 72] with a minimum enrichment score of 1.3 for gene clusters and the Benjamini corrected *P*-value as the cutoff using the *Apis mellifera* background. All statistical analyses were performed with R v3.6.0 [73].

Additional ffiles

Abbreviations

SNP: Single nucleotide polymorphism; CSS: Composite selection score; QTL: Quantitative trait loci; VSH: Varroa sensitive hygiene; KNN: k-nearest neighbors; AIC: Akaike information criterion

Acknowledgements

We thank Lorraine Beaman, Sharon O'Brien, RaeDiance Fuller, and Christian Moreau for sample processing. We thank five commercial queen breeders for contributing stocks used in this study, and BartJan Fernhout, David Thomas, and Danielle Downey of the Hilo Bee breeding project for their support. We acknowledge Matt Tarver and Beth Holloway for the initial conceptualization of the study.

Authors' contributions

PS performed the bioinformatics, interpreted the results, generated the figures and drafted the manuscript. MSF designed the final experimental plan, coordinated the project, interpreted the results and drafted the manuscript. AA advised on the analysis and interpreted the results. LB advised on the analysis and co-conceived the project. RGD advised on the analysis and collected samples. LDG advised on the analysis and collected samples. FR advised on the analysis. PT carried out the molecular experiments and generated the data. All authors read, edited and approved the manuscript for submission.

Author information

Lelania Bilodeau has previously published under the name A. Lelania Bourgeois.

Funding

This research was funded by the USDA-ARS, research plan 6050–21000-014-00D.

Availability of data and materials

Sequencing data presented in this study has been deposited at NCBI under the accession PRJNA605407.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors have no competing interests. Any mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply a recommendation or endorsement by the U.S. Department of Agriculture. USDA is an equal opportunity employer.

Author details

[1]USDA-ARS, Honey Bee Breeding, Genetics and Physiology Laboratory, Baton Rouge, LA 70820, USA. [2]Present Address: USDA-ARS Knipling-Bushland U.S. Livestock Arthropod Pests Research Unit, Kerrville, TX 78028, USA.

Received: 1 May 2020 Accepted: 29 September 2020
Published online: 08 October 2020

References

1. Frankham R. Quantitative genetics in conservation biology. Genet Res. 1999; 74(3):237–44.
2. Meffe GK. Conservation genetics and the Management of Endangered Fishes. Fisheries. 1986;11(1):14–23.
3. Hailer F, Helander B, Folkestad AO, Ganusevich SA, Garstad S, Hauff P, Koren C, Nygård T, Volke V, Vilà C, et al. Bottlenecked but long-lived: high genetic diversity retained in white-tailed eagles upon recovery from population decline. Biol Lett. 2006;2(2):316–9.
4. Jones JC, Myerscough MR, Graham S, Oldroyd BP. Honey bee Nest thermoregulation: diversity promotes stability. Science. 2004;305(5682):402–4.
5. Oldroyd BP, Fewell JH. Genetic diversity promotes homeostasis in insect colonies. Trends Ecol Evol. 2007;22(8):408–13.
6. Seeley TD, Tarpy DR. Queen promiscuity lowers disease within honeybee colonies. Proc R Soc B Biol Sci. 2007;274(1606):67–72.
7. Tarpy DR. Genetic diversity within honeybee colonies prevents severe infections and promotes colony growth. Proc R Soc Lond Ser B Biol Sci. 2003;270(1510):99–103.
8. Oldroyd BP. What's killing American honey bees? PLoS Biol. 2007;5(6):e168.
9. Mattila HR, Seeley TD. Genetic diversity in honey bee colonies enhance productivity and fitness. Science. 2007;317(5836):362–4.
10. Tarpy DR, van Engelsdorp D, Pettis JS. Genetic diversity affects colony survivorship in commercial honey bee colonies. Naturwissenschaften. 2013; 100(8):723–8.
11. Vanengelsdorp D, Meixner MD. A historical review of managed honey bee populations in Europe and the United States and the factors that may affect them. J Invertebr Pathol. 2010;103(Suppl 1):S80–95.
12. Sheppard WS. Honey bee genetic diversity and breeding - towards the reintroduction of European germplasm. American Bee J. 2012;152:155–8.
13. Delaney DA, Meixner MD, Schiff NM, Sheppard WS. Genetic characterization of commercial honey bee (Hymenoptera: Apidae) populations in the United States by using mitochondrial and microsatellite markers. Ann Entomol Soc Am. 2009;102(4):666–73.
14. Spivak M, Reuter GS. Resistance to American foulbrood disease by honey bee colonies *Apis mellifera* bred for hygienic behavior. Apidologie. 2001;32: 555–65.
15. Danka RG, Harris JW, Villa JD. Hygienic responses to Varroa destructor by commercial and feral honey bees from the big island of Hawaii before exposure to mites. Sci Bee Culture. 2010;2(1):11–4.
16. Robertson AJ, Trost B, Scruten E, Robertson T, Mostajeran M, Connor W, Kusalik A, Griebel P, Napper S. Identification of developmentally-specific kinotypes and mechanisms of Varroa mite resistance through whole-organism, kinome analysis of honeybee. Front Genet. 2014;5:139.
17. Morfin N, Given K, Evans M. et al. Grooming behavior and gene expression of the Indiana "mite-biter" honey bee stock. Apidologie. 2020;51:267–75. https://doi.org/10.1007/s13592-019-00710-y.
18. Kwong WK, Engel P, Koch H, Moran NA. Genomics and host specialization of honey bee and bumble bee gut symbionts. Proc Natl Acad Sci U S A. 2014;111(31):11509–14.
19. de Guzman LI, Rinderer TE, Frake AM. Growth of *Varroa destructor* (Acari: Varroidae) populations in Russian honey bee (Hymenoptera: Apidae) colonies. Ann Entomol Soc Am. 2007;100(2):187–95.
20. Bourgeois L, Rinderer T. Genetic characterization of Russian honey bee stock selected for improved resistance to Varroa destructor. J Econ Entomol. 2009; 102:1233–8.
21. Danka RG, Harris JW, Dodds GE. Selection of VSH-derived "pol-line" honey bees and evaluation of their Varroa-resistance characteristics. Apidologie. 2015;47(3):483–90.
22. Nelson RM, Wallberg A, Simões ZLP, Lawson DJ, Webster MT. Genomewide analysis of admixture and adaptation in the Africanized honeybee. Mol Ecol. 2017;26(14):3603–17.
23. Avalos A, Pan H, Li C, Acevedo-Gonzalez JP, Rendon G, Fields CJ, Brown PJ, Giray T, Robinson GE, Hudson ME, et al. A soft selective sweep during rapid evolution of gentle behaviour in an Africanized honeybee. Nat Commun. 2017;8(1):1550.
24. Wallberg A, Pirk CW, Allsopp MH, Webster MT. Identification of multiple loci associated with social parasitism in honeybees. PLoS Genet. 2016;12(6): e1006097.
25. Wallberg A, Schöning C, Webster MT, Hasselmann M. Two extended haplotype blocks are associated with adaptation to high altitude habitats in east African honey bees. PLoS Genet. 2017;13(5):–e1006792.
26. Cridland JM, Tsutsui ND, Ramírez SR. The complex demographic history and evolutionary origin of the Western honey bee, *Apis Mellifera*. Genome Biol Evol. 2017;9(2):457–72.
27. Wallberg A, Han F, Wellhagen G, Dahle B, Kawata M, Haddad N, Simoes ZL, Allsopp MH, Kandemir I, De la Rua P, et al. A worldwide survey of genome sequence variation provides insight into the evolutionary history of the honeybee *Apis mellifera*. Nat Genet. 2014;46(10):1081–8.
28. Franck P, Garnery L, Celebrano G, Solignac M, Cornuet J-M. Hybrid origins of honeybees from Italy (*Apis mellifera ligustica*) and Sicily (*A. m. sicula*). Mol Ecol. 2000;9(7):907–21.
29. Ruttner F: Biogeography and taxonomy of honeybees. Berlin: Springer-Verlag; 2013.
30. Sheppard WS. Comparative study of enzyme polymorphism in United States and European honey bee (Hymenoptera: Apidae) populations. Ann Entomol Soc Am. 1988;81(6):886–9.
31. Schlotterer C, Tobler R, Kofler R, Nolte V. Sequencing pools of individuals - mining genome-wide polymorphism data without big funding. Nat Rev Genet. 2014;15(11):749–63.
32. Spivak M, Reuter GS. Performance of hygienic honey bee colonies in a commercial apiary. Apidologie. 1998;29:291–302.
33. Spivak M. Honey bee hygienic behavior and defense against *Varroa jacobsoni*. Apidologie. 1996;27(4):245–60.
34. Spivak M, Reuter GS. *Varroa destructor* infestation in untreated honey bee (Hymenoptera: Apidae) colonies selected for hygienic behavior. J Econ Entomol. 2001;94(2):326–31.
35. Parker R, Guarna MM, Melathopoulos AP, Moon K-M, White R, Huxter E, Pernal SF, Foster LJ. Correlation of proteome-wide changes with social immunity behaviors provides insight into resistance to the parasitic mite, *Varroa destructor*, in the honey bee (*Apis mellifera*). Genome Biol. 2012;13(9): R81.
36. Oxley PR, Spivak M, Oldroyd BP. Six quantitative trait loci influence task thresholds for hygienic behaviour in honeybees (*Apis mellifera*). Mol Ecol. 2010;19(7):1452–61.
37. Harpur BA, Guarna MM, Huxter E, Higo H, Moon KM, Hoover SE, Ibrahim A, Melathopoulos AP, Desai S, Currie RW, et al. Integrative genomics reveals the genetics and evolution of the honey bee's social immune system. Genome Biol Evol. 2019;11(3):937–48.
38. Danka RG, Harris JW, Villa JD, Dodds GE. Varying congruence of hygienic responses to *Varroa destructor* and freeze-killed brood among different types of honeybees. Apidologie. 2013;44(4):447–57.
39. Boutin S, Alburaki M, Mercier P-L, Giovenazzo P, Derome N. Differential gene expression between hygienic and non-hygienic honeybee (*Apis mellifera* L.) hives. BMC Genomics. 2015;16(1):500.
40. Scannapieco AC, Mannino MC, Soto G, Palacio MA, Cladera JL, Lanzavecchia SB. Expression analysis of genes putatively associated with hygienic behavior in selected stocks of *Apis mellifera* L. from Argentina. Insect Soc. 2017;64(4):485–94.
41. Rinderer TE, de Guzman LI, Delatte GT, Stelzer JA, Lancaster VA, Kunznetsov V, Beaman L, Watt R, Harris JW. Resistance to the parasitic mite *Varroa destructor* in honey bees from far-eastern Russia. Apidologie. 2001;32:381–94.
42. Kirrane MJ, de Guzman LI, Holloway B, Frake AM, Rinderer TE, Whelan PM. Phenotypic and genetic analyses of the Varroa sensitive hygienic trait in Russian honey bee (Hymenoptera: Apidae) colonies. PLoS One. 2015;10(4): e0116672.
43. Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. Natural selection has driven population differentiation in modern humans. Nat Genet. 2008; 40(3):340–5.
44. Bourgeois L, Sheppard WS, Sylvester HA, Rinderer TE. Genetic stock identification of Russian honey bees. J Econ Entomol. 2010;103(3):917–24.
45. Harpur BA, Minaei S, Kent CF, Zayed A. Management increases genetic diversity of honey bees via admixture. Mol Ecol. 2012;21(18):4414–21.

46. Beye M, Gattermeier I, Hasselmann M, Gempe T, Schioett M, Baines JF, Schlipalius D, Mougel F, Emore C, Rueppell O, et al. Exceptionally high levels of recombination across the honey bee genome. Genome Res. 2006;16(11): 1339–44.
47. Gloag R, Ding G, Christie JR, Buchmann G, Beekman M, Oldroyd BP. An invasive social insect overcomes genetic load at the sex locus. Nat Ecol Evol. 2016;1(1):11.
48. Heimpel GE, de Boer JG. Sex determination in the Hymenoptera. Annu Rev Entomol. 2008;53:209–30.
49. McNeil MEA. Boots on the ground: a radical shift in the Interface between research and real world. Bee Culture. 2012;140:37–40.
50. Muñoz I, Dall'Olio R, Lodesani M, De la Rúa P. Population genetic structure of coastal Croatian honeybees (*Apis mellifera carnica*). Apidologie. 2009;40(6): 617–26.
51. Bourgeois L, Beaman L. Tracking the genetic stability of a honey bee (Hymenoptera: Apidae) breeding program with genetic markers. J Econ Entomol. 2017;110(4):1419–23.
52. Tsuruda JM, Harris JW, Bourgeois L, Danka RG, Hunt GJ. High-resolution linkage analyses to identify genes that influence Varroa sensitive hygiene behavior in honey bees. PLoS One. 2012;7(11):e48276.
53. Spötter A, Gupta P, Nürnberg G, Reinsch N, Bienefeld K. Development of a 44K SNP assay focussing on the analysis of a varroa-specific defence behaviour in honey bees (*Apis mellifera carnica*). Mol Ecol Resour. 2012;12(2): 323–32.
54. Guarna MM, Hoover SE, Huxter E, Higo H, Moon KM, Domanski D, Bixby MEF, Melathopoulos AP, Ibrahim A, Peirson M, et al. Peptide biomarkers used for the selective breeding of a complex polygenic trait in honey bees. Sci Rep. 2017;7(1):8381.
55. Guarna MM, Melathopoulos AP, Huxter E, Iovinella I, Parker R, Stoynov N, Tam A, Moon K-M, Chan QWT, Pelosi P, et al. A search for protein biomarkers links olfactory signal transduction to social immunity. BMC Genomics. 2015;16(1):63.
56. Elsik C, Tayal A, Diesh C, Unni D, Emery M, Nguyen H, Hagen D. Hymenoptera genome database: integrating genome annotations in HymenopteraMine. Nucleic Acids Res. 2015;44.
57. Boncristiani H, Li J, Evans JD, Pettis J, Chen Y. Scientific note on PCR inhibitors in the compound eyes of honey bees, *Apis mellifera*. Apidologie. 2011;42(4):457–60.
58. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
59. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, et al. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013;43(1):11.10.11–33.
60. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491–8.
61. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297–303.
62. Wilm A, Aw PPK, Bertrand D, Yeo GHT, Ong SH, Wong CH, Khor CC, Petric R, Hibberd ML, Nagarajan N. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res. 2012;40(22):11189–201.
63. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics. 2012;28(24):3326–8.
64. Mucherino A, Papajorgji PJ, Pardalos PM. k-Nearest Neighbor Classification. In: Data Mining in Agriculture. Springer Optimization and Its Applications. 2009;34:83–106. https://doi.org/10.1007/978-0-387-88615-2_4.
65. Nei M, Li WH. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci U S A. 1979;76(10):5269–73.
66. Hartl DL, Clark AG, Clark AG. Principles of population genetics, vol. 116. Sunderland: Sinauer associates; 1997.
67. Watterson GA. On the number of segregating sites in genetical models without recombination. Theor Popul Biol. 1975;7(2):256–76.
68. Kofler R, Orozco-terWengel P, De Maio N, Pandey RV, Nolte V, Futschik A, Kosiol C, Schlötterer C. PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. PLoS One. 2011;6(1):e15925.
69. Kofler R, Pandey RV, Schlotterer C. PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). Bioinformatics. 2011;27(24):3435–6.
70. Randhawa IAS, Khatkar MS, Thomson PC, Raadsma HW. Composite selection signals can localize the trait specific genomic regions in multi-breed populations of cattle and sheep. BMC Genet. 2014;15(1):34.
71. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44–57.
72. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37(1):1–13.
73. Team RC. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2019.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.