EVOLUTIONARY BIOLOGY

Thrice out of Asia and the adaptive radiation of the western honey bee

Kathleen A. Dogantzis[1], Tanushree Tiwari[1], Ida M. Conflitti[1], Alivia Dey[1], Harland M. Patch[2], Elliud M. Muli[3], Lionel Garnery[4], Charles W. Whitfield[5]†, Eckart Stolle[6], Abdulaziz S. Alqarni[7], Michael H. Allsopp[8], Amro Zayed[1]*

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. The adaptive radiation of honey bees involved selection on a few genomic "hotspots." We found 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 12 extant species that form three distinct groups: giant honey bees, dwarf honey bees, and cavity-nesting honey bees (*1*–*3*). 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, *A. mellifera* has diversified into several subspecies (*4*, *5*), of which there are approximately 10 subspecies in Africa, 9 in Asia, and potentially as many as 13 subspecies in Europe (*6*). 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 (*4*, *5*). Although it is reasonably accepted that the genus emerged in Asia, the ancestral origin and adaptive radiation of contemporary *A. mellifera* lineages and subspecies remain unresolved.

Early fossil records from the Oligocene (34 to 23 Ma ago) place ancestral *Apis* within Europe, followed by a migration of the genus during the Late Oligocene or during the Miocene (23 to 5.5 Ma ago) (*5*, *7*). It has been hypothesized that ancestral *Apis* migrated from Europe into Asia, where it diversified into the three modern lineages of *Apis*, including *A. mellifera* of the cavity-nesting bees (*5*, *7*). 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 *A. mellifera*, while the remaining extent *Apis* species descended from ancestors in Asia (*7*). These different hypotheses about the biogeography and diversification of the *Apis* genus are important for understanding the two competing hypotheses regarding the origin of *A. mellifera* in Asia (*8*–*11*) or Africa (*12*, *13*). 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) (*11*). However, it has also been proposed that the route into Africa could have acted as a western expansion of the M lineage into Europe (*8*). 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 lineages through Northeast Africa and Western Asia (*12*).

Resolving the ancestral origin and evolutionary expansion of *A. mellifera* will enhance our ability to identify derived and ancestral genetic mutations. This is especially relevant for tracing the evolution of derived phenotypes and for discerning how locally adapted subspecies may contribute to the fitness and diversity of managed colonies (*14*). Recent genomic studies of *A. mellifera* have shown that with the addition of new subspecies and enhanced datasets (*9*, *15*), estimates of evolutionary origin can change. As such, the increased representation of samples from Africa and Western Asia—two historically undersampled regions (*4*, *16*)—may be the key to resolving the out-of-Africa and out-of-Asia debate.

Here, we used an extensive population genomic dataset composed 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 *A. mellifera*. These samples were collected throughout the native distribution of *A. 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. Last, we assessed patterns of selection among lineages to identify and categorize the genomic regions associated with the adaptative radiation of the species.

RESULTS

Sequencing and variant detection

We curated a genomic dataset of 251 individual *A. mellifera* samples representing 18 putative subspecies, of which 14 representative groups

[1]Department of Biology, York University, 4700 Keele Street, Toronto, M3J 1P3 Ontario, Canada. [2]Department of Entomology, The Pennsylvania State University, State College, PA, USA. [3]Department of Life Science, South Eastern Kenya University (SEKU), P.O. Box 170-90200, Kitui, Kenya. [4]Laboratoire Evolution Génome Comportement Ecologie (EGCE) UMR 9191, Gif sur-Yvette, France. [5]Department of Entomology, University of Illinois at Urbana-Champaign, Urbana, IL, USA. [6]LIB–Leibniz Institute for the Analysis of Biodiversity Change Museum Koenig, Center of Molecular Biodiversity Research Adenauerallee 160, 53113 Bonn, Germany. [7]Department of Plant Protection, College of Food and Agriculture Sciences, King Saud University, Riyadh, Saudi Arabia. [8]Plant Protection Research Institute, Agricultural Research Council, Stellenbosch, South Africa.
*Corresponding author. Email: zayed@yorku.ca
†Now retired.

were retained (Supplementary Text and data S1). The dataset is composed of several previously published samples (*17*–*21*) and 160 newly sequenced 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 ± 17.7*x*. After filtering raw variants, we retained a working dataset of 11.8 million single-nucleotide polymorphisms (SNPs).

Population structure and admixture

Using several population structure analyses, we investigated patterns of clustering and admixture among honey bee samples (Fig. 1B and figs. S1 to S3). 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 lineages), Europe (C lineage), and Eurasia (M lineage) (Fig. 1, A and B). 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. 1, A and B). At $K = 8$, *A. m. intermissa* (North Africa), a highly admixture subspecies (27%), is identified as an independent genetic cluster (fig. S1). However, this cluster is not consistent with other K values (fig. S1) 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. S1). 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. 1A), which is likely the contributor to high levels of hybridization (*9*, *10*). Introgression of the C lineage into the L and M lineages was detected (fig. S1). These admixture patterns are expected given the close geographic proximity of these lineages, and M and C lineage admixture has been documented extensively (*22*–*27*). Last, we detect varying levels of admixture in samples from Kyrgyzstan (*A. m. pomonella*), with some samples displaying high levels of C

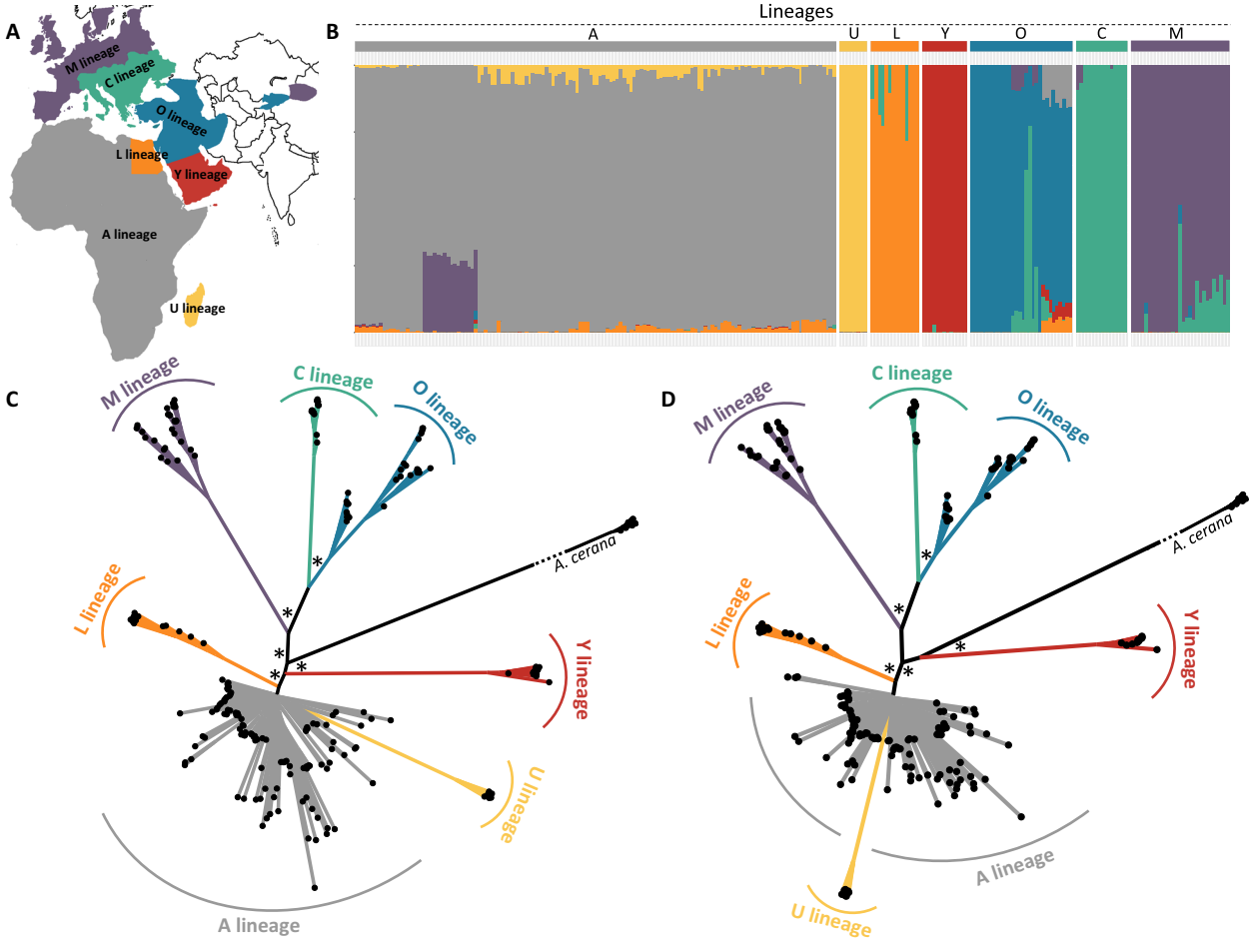

Fig. 1. Population structure and phylogenetic reconstruction of *A. mellifera*. (**A**) Map of the native distribution of the seven genetically distinct lineages. (**B**) Patterns of ancestry and population structure identified with ADMIXTURE when $K = 7$. Vertical bars represent individual bees, and colored segments represent the proportion of ancestry to the different clusters. (**C**) Evolutionary relationships among *A. mellifera* samples reconstructed with a neighbor-joining tree using SNPs located genome-wide. Asterisks represent node support of 100%. (**D**) Evolutionary relationships among *A. 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.

lineage ancestry (fig. S1) 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 *A. mellifera* samples (Fig. 1, C and D, and figs. S4 to S7). Our analyses resolved two topologies that differed slightly with respect to the placement of the Y lineage (Fig. 1, C and D). Subspecies are consistently clustered into previously recognized lineages, and two definitive clades are 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 *A. mellifera* lineages may have begun to diverge as early as c. 6 Ma ago (fig. S8).

To predict the most likely ancestral range of the species and major clades, we applied a biogeographic reconstruction to both resolved topologies (table S1). The ancestral range for the most recent common ancestor of the species was predicted to be in Asia with 64.5 to 71.4% probability, while probabilities for an African or European ancestral range were much lower (<6%) (Fig. 2, fig. S9, and table S2). This finding complements a recent independent study that predicted the ancestral range for cavity-nesting bees to be in Southeast Asia (*28*). The ancestral range of the most recent common ancestor of the M, C, and O clades was predicted to be in Asia with a 70% probability, while the ancestral range of the L, A, and U clades varied (70% Asia or 100% Africa) depending on the topology (Fig. 2, fig. S9, and table S2). Microgeographic classification of subspecies likewise places the ancestral range of the species in Western Asia (fig. S10). The use of an outgroup in biogeographic reconstructions is recommended to prevent the inference of wide ancestral ranges (*29*). We used *Apis 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 (*28*).

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 (*10*), have likely shaped patterns of genetic diversity and effective population size among contemporary populations. For instance, genetic diversity is highest among the A lineage ($\pi = 3.54 \times 10^{-3}$, $\theta_w = 1.01 \times 10^{-2}$), relative to European (average $\pi = 1.48 \times 10^{-3}$, $\theta_w = 1.84 \times 10^{-3}$) and Asian (average $\pi = 1.84 \times 10^{-3}$, $\theta_w = 1.83 \times 10^{-3}$) lineages (table S3). Likewise, estimates of N_e were considerably larger for the A lineage (~640,000), relative to European or Asian lineages (~116,000) (table S3), as previously documented (*10*). The U lineage of Madagascar had relatively high levels of diversity ($\pi = 2.33 \times 10^{-3}$, $\theta_w = 1.70 \times 10^{-3}$) 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 S7). In addition, 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 (figs. S11 and S12).

Measures of pairwise F_{ST} between lineages were high (0.528 ± 0.149) (table S5), while estimates between subspecies within the same lineage were low (0.163 ± 0.073) (table S6). Outgroup f_3 statistics, which are less sensitive to lineage-specific drift (*30*), were used to quantify the genetic distance between lineages relative to an outgroup

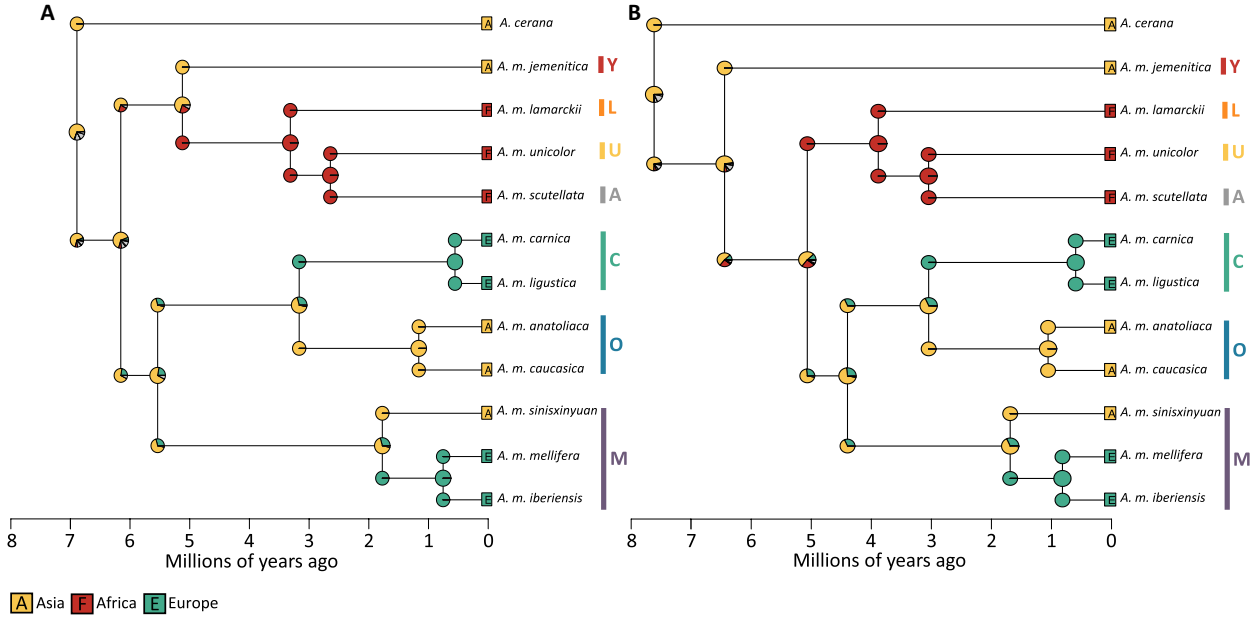

Fig. 2. Ancestral biogeographic range reconstruction of *A. mellifera* using two resolved topologies. The current geographic range of subspecies is indicated at branch tips by letters A (Asia), F (Africa), and E (Europe). Colored bars to the right of the trees indicate the lineage association of the 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 to 73%. (**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.

(*A. cerana*). The analysis identified high f_3 between the O and C lineages, affirming a longer shared evolutionary history (fig. S13). Pairwise f_3 values between the A, L, and U lineages were also high, suggesting a close evolutionary relationship between African lineages (fig. S13). Last, we observed high f_3 values between the M and C lineages (fig. S13), despite having high genetic differentiation ($F_{ST} = 0.66$) (table S5), 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 S7 and 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 that they likely experience disparate selective pressures (*20*). In addition, the A lineage was excluded from analyses because of a relatively low number of lineage-specific outlier markers (table S7 and Supplementary Text). While we and others (*10*, *17*) have found 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 (Supplementary Text).

Outlier SNPs were enriched within protein coding (χ^2, $P < 6.43 \times 10^{-11}$) and putative promoter regions (χ^2, $P < 3.12 \times 10^{-2}$) of most lineages (table S8). Comparatively, introns were deficient of outlier SNPs, which was significant among four lineages (χ^2, $P < 2.79 \times 10^{-2}$) (table S8). Although each outlier SNP is distinct to a lineage, we found that their distribution was concentrated among a relatively small set of genomic hotspots as evidenced by a significant overlap of genes with outlier SNPs between pairwise lineage comparisons (781 ± 263 genes; $P < 3.57 \times 10^{-50}$) (Fig. 3 and table S9). 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 were 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 S10). 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

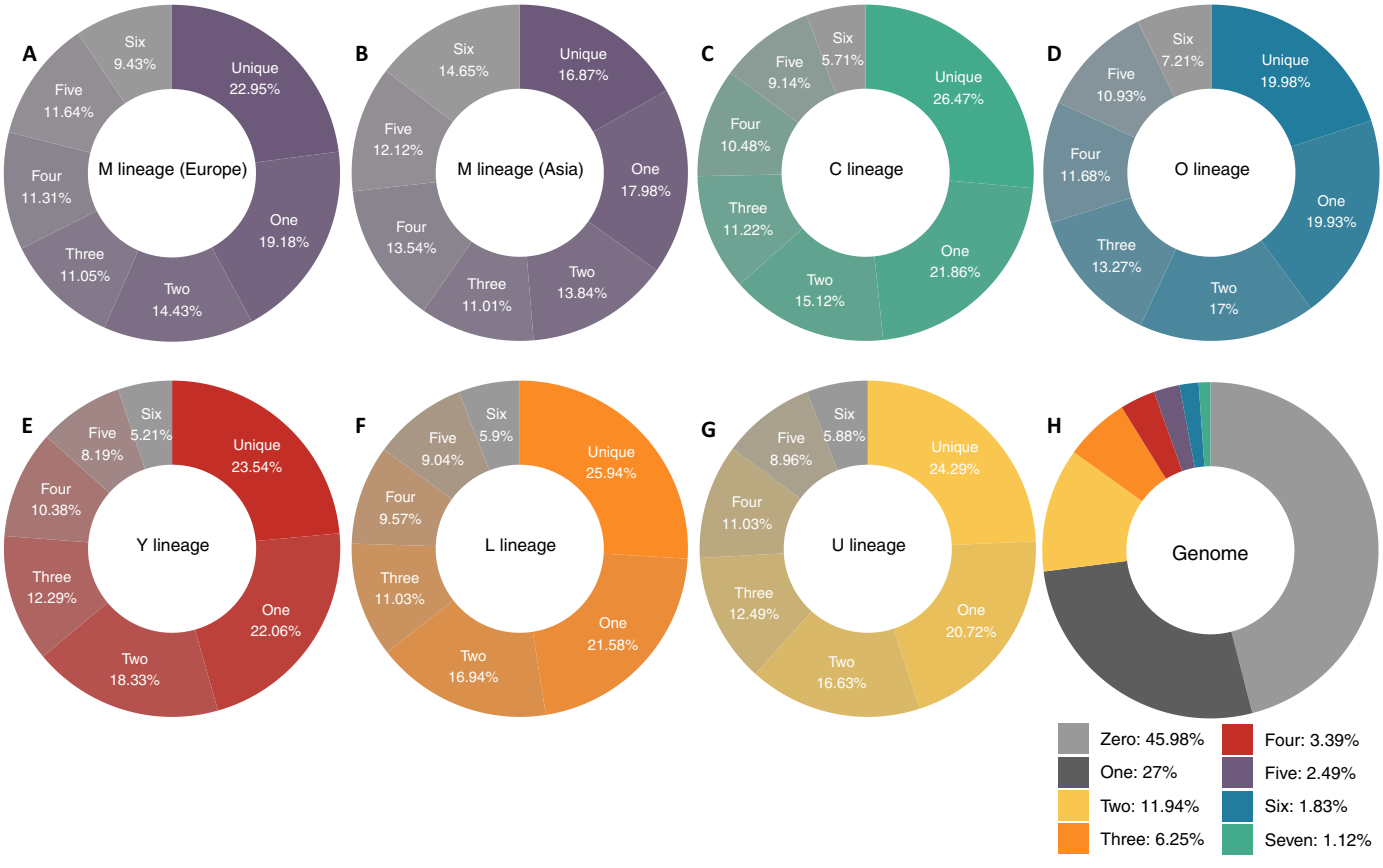

Fig. 3. Proportion of genes that overlap among lineages and across the *A. mellifera* genome. (**A**) to (**G**) illustrate the proportion of genes that are either unique to a lineage or share signs of selection among one to six other lineages. (**H**) illustrates the proportion of genes across the genome that have outlier SNPs among no lineages to all seven lineages.

lower (29.6%; Mann-Whitney test, $P < 7.17 \times 10^{-48}$), 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 *A. 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 *Drosophila melanogaster*, which is important for antennal development, appendage patterning, and cell division of the eye field (*31*). We also found enrichment of GO terms related to neuron development and receptor and signaling activity. There was also evidence for enrichment of genes related to learning and memory, as well as behavior, 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 (*32*). In addition, among the 145 genes under selection across all lineages, there were several genes that have been found to be associated with colony behavior traits including colony defense (*33*), immunity (*34*–*37*), and the production of honey and royal jelly (data S2) (*38*). Intriguingly, we find that several genes overlap among colony traits. For example, three genes (GB54493, GB51389, and 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. Last, 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 (*39*) and proteins in adults (*17*, *40*) to define genes associated with queen and worker traits (i.e., queen-biased versus worker-biased expression). We found, relative to expected values, that genes associated with local adaptation of honey bee lineages were significantly elevated in the worker caste (χ^2, $P < 7.16 \times 10^{-4}$) but often significantly underrepresented in the queen caste (χ^2, $P < 3.99 \times 10^{-7}$ larvae; not statistically significant among adults) (Fig. 4A). Likewise, the proportion of genes with outlier SNPs was significantly higher in worker-biased genes, relative to queen-biased genes (χ^2, $P < 3.11 \times 10^{-2}$) (Fig. 4B). Last, 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, $P = 1.35 \times 10^{-13}$).

DISCUSSION

Deciphering the ancestral origin of contemporary *A. mellifera* lineages is a major unsolved question with implications for understanding the evolution of this model eusocial species. There are currently two hypotheses that place the origin of *A. mellifera* in either Africa (*12*, *13*) or Asia (*8*–*11*). Our analysis supports the hypothesis of an Asian origin of *A. mellifera*. *A. mellifera* likely diverged from other cavity-nesting bees in Southeast Asia (*28*) 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. 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 *A. mellifera* originating independently in Africa.

Once diverged from other cavity-nesting bees, our biogeographic reconstruction provides several hypotheses for how *A. mellifera* expanded to its current distribution. The M lineage, which forms a distinct evolutionary branch, likely colonized Europe via an independent northern route. Although previous studies hypothesized that the M lineage expanded from Africa (*9*, *12*, *13*), which was supported by shared genetic similarity with the A lineage, these patterns are likely the result of recent nuclear (*12*, *15*, *41*) and mitochondrial (*41*–*45*) 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. Last, 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 (*46*, *47*) and Western Asia (*48*), notably the Y lineage. In contrast, the A lineage, which comprises the remainder of Africa, has distinct nuclear and mitochondrial (*48*) variants and is ancestral to the U lineage.

The adaptive radiation of *A. 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 (*49*). However, recent studies with *A. cerana* have also uncovered patterns of gene reuse, which may be linked to radiation among diverse habitats (*50*). In our study, we find that genomic hotspots are prevalent among genes related to development, morphogenesis, and behavior. This pattern of selection is consistent with the extensive morphological and behavioral adaptations that have occurred among the species, especially between tropically and temperately adapted bees (*51*).

Last, 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 (*17*) and is hypothesized to be related to the eusocial nature of honey bee colonies (*52*, *53*). Honey bee colonies are composed of several thousand workers who contribute to important colony tasks such a brood rearing and resource provisioning. Although workers do not lay eggs, natural selection may indirectly select for worker phenotypes to optimize colony fitness (*17*). This is relevant given the diverse colony adaptations that have arisen in response to environmental variables, including traits directly related to colony defense (*33*), immunity (*34*–*37*), and the production of honey and royal jelly (*38*). We also find signs of pleiotropy between worker phenotypes, indicating that not only is repeated selection among a common set of genes prevalent across *A. mellifera* lineages, but also the same genes are modulating fitness by influencing several different phenotypes.

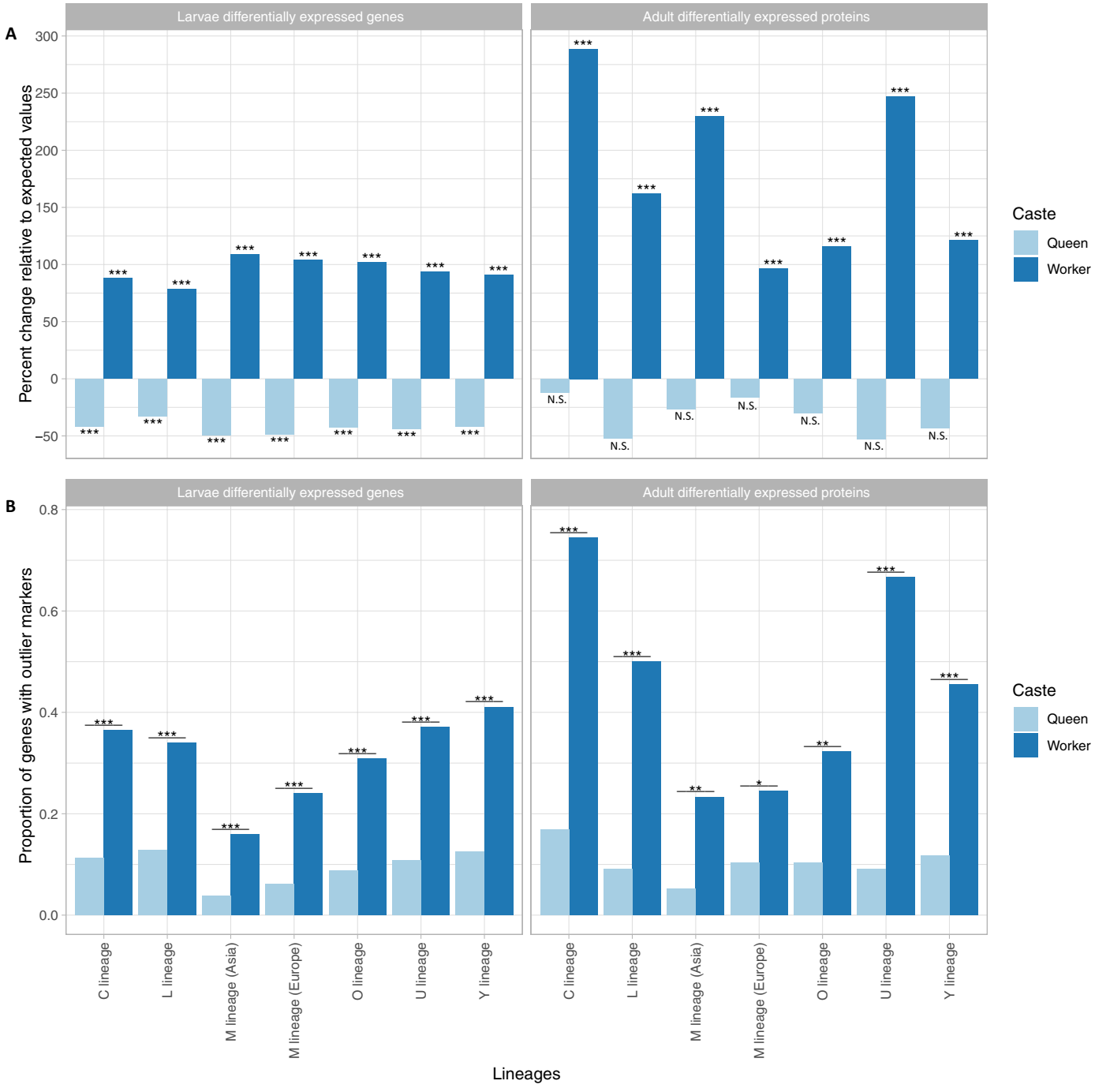

Fig. 4. Association of genes with outlier loci among the worker and queen caste. (**A**) Percent change in the observed number of genes with outlier SNPs among queen- and worker-biased genes for larvae and adults, relative to expected values. For example, a negative change suggests an underrepresentation 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 (*$P < 0.05$, **$P < 0.01$, and ***$P < 0.001$), while N.S. is not statistically significant. (**B**) Proportion of genes with outlier SNPs among worker caste– and queen caste–biased genes for larvae and adults for each lineage. Asterisks represent the degree of significant difference between the proportions (*$P < 0.05$, **$P < 0.01$, and ***$P < 0.001$).

In conclusion, we have presented compelling evidence that *A. 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 *A. mellifera* maintain high genetic diversity, which has allowed the species to adapt to diverse environments through repeated selection among a common set of genes. These genes are often related to worker phenotypes, supporting that the worker caste is key to the adaptative radiation of the species.

METHODS

Data processing

Methods for DNA extraction, genome alignment, and SNP detection are described in detail in the Supplementary Materials. In brief, we generated a dataset of 251 individual *A. mellifera* samples, of which 160 samples are newly sequenced, with the remaining samples downloaded from the Sequence Read Archive (SRA) in addition to 15 *A. cerana* genomes (data S1) (*54*). Sequence reads were trimmed of adapters, and low-quality bases use Trimmomatic v0.36 (*55*). Trimmed reads were aligned to the honey bee reference genome (*56*) using NextGenMap aligner v0.4.12 (*57*), and duplicate reads were marked with Picard v2.1.0 (https://broadinstitute.github.io/picard/). SNPs were identified and filtered using GATK v3.7 (*58*, *59*).

Population structure

The program ADMIXTURE v1.3.0 (*60*) was used to estimate ancestry proportions and population structure among the 251 *A. mellifera* samples. To reduce the effects of uninformative and low-frequency variants (*61*), 1 million variants were selected among a pool of SNPs pruned for biallelic loci with a minor allele frequency (MAF) >0.05. To account for LD, SNPs were further pruned (38,493) for a minimum distance of 5000 base pairs (bp)—a distance where LD typically decays to background level in the honey bee genome (*10*). Both analyses were run with predicted *K* values of 1 to 18 and used the 10× cross-validation procedure to estimate the optimal number of ancestral groups (*K*). A principal components analysis was generated to examine the genetic relatedness and clustering patterns among *A. mellifera* samples using the SNPRelate (*62*) package in R (*63*) with all available SNP markers. Last, we performed a hierarchal structure analysis using identity by state with the SNPRelate (*62*) package in R (*63*) to qualitatively determine lineage assignment using the snpgdsCutTree function.

Phylogenetic reconstruction

We produced several phylogenetic trees using an SNP dataset pruned of ambiguous loci, as implemented by RaxML v8.2.12 (*64*), and loci with low coverage (<0.8) in *A. cerana*. Trees were constructed using three different datasets: (i) SNPs located genome-wide (2,126,091), (ii) SNPs within coding regions (276,602), and (iii) randomly selected SNPs located among intrageneric and intergenic regions (276,602). Neighbor-joining trees were constructed with all three SNP sets using allele-sharing distance with Adegenet (*65*) and Ape (*66*) in R (*63*). Confidence levels for bipartitions in the neighbor-joining tree were calculated using 100 bootstrap replicates as implemented in Ape (*66*). Maximum likelihood trees were constructed using SNP sets 2 and 3 using the program RaxML v8.2.12 (*64*). 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 were performed in a single program run. Last, the program TreeMix v1.13 (*67*) was used to produce maximum likelihood trees using dataset 1. 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 (*68*), 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 and Supplementary Text).

Divergence time estimation

Divergence times were estimated with PAML 4.9 (*69*) using both resolved phylogenetic topologies. We used the putative coding regions of nonoverlapping genes, with high (>0.9) sequence coverage among the outgroup (*A. 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 (*69*). We used the REV [general-time-reversible (GTR)] model with the strict molecular clock and calibrated the divergence time between *A. mellifera* and *A. cerana* at 7.5 Ma ago (*1*, *10*, *20*) using a time unit of 100 Ma ago (@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 (*69*). Divergence times were estimated following the two-step approximate likelihood calculation with the MCMCtree package in PAML 4.9 (*69*). We used the REV (GTR) model with independent clock rates, and root age was bound between >0.06<0.09 (*1*, *10*, *20*) using a time unit of 100 Ma ago. 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 *A. mellifera*, we estimated the most probable model of geographic range expansion on the divergence time tree of both topologies using the R package BioGeoBEARS (*70*, *71*). BioGeoBEARS uses three different models of geographic range evolution: Extinction Cladogenesis (DEC) (*72*, *73*), a likelihood version of dispersal-vicariance analysis (DIVA) (*74*), and a likelihood version of BayArea (BAYAREA) (*75*). In addition, 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 *A. mellifera* distribution: Europe (E), Africa (F), and Asia (A). We tested all six biogeographic models provided by BioGeoBEARS and used the Akaike information criterion 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 500-bp sliding windows with a 250-bp step size using VCFtools v0.1.17 (*76*). Segregating sites (*S*) were calculated by counting the number of polymorphic loci, and singletons ($S_{singletons}$) 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 (*77*) $\theta_w = 3N_e\mu$ (3 is used because *A. mellifera* is haplodiploid), where μ is the mutation rate. N_w was calculated using two estimates of mutation rate 5.27×10^{-9} (*10*) and 3×10^{-9} (*78*). We calculated LD as a measure of the squared correlation coefficient between variants (r^2). LD was measured within 5000-bp windows using SNP variants that had ≥0.8 coverage and an MAF >0.05

using VCFtools (*76*). The pairwise F_{ST} matrix was calculated using Weir and Cockerham's weighted F_{ST} statistic with VCFtools v 0.1.17 (*76*). Last, the program ADMIXTOOLS (*79*) was used to calculate outgroup f_3 statistics (*80*), which can be used to quantify the genetic distance between populations relative to an outgroup, *A. 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 v0.1.17 (*76*). 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 an 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 (*20*). We used the program SNPeff v4.3t (*81*) to annotate SNPs at the gene and functional category levels, including exons, introns, and promoter regions, which were defined as the sequence 1000 bp upstream of the start codon of a gene (*82*) and excluded regions that overlapped with neighboring genes. In addition, SNPeff v4.3t was used to predict mutation effects on genes, such as amino acid changes. Last, GO enrichment was conducted with DAVID v6.8 (*83*) using *D. melanogaster* orthologs (*84*). GO functional annotation clusters with an enrichment score ≥1.3 and GO 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 and 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 that 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 here (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 and then compared the null distribution 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-hour-old queen and workers (*39*). Reads were downloaded from the SRA (BioProject PRJNA260604) and trimmed of adapters and low-quantity bases (<20) using Trimmomatic v0.36 (*55*). Trimmed reads were then aligned to the honey bee reference genome (*56*) using multisample two-pass mapping with STAR v2.7 (*85*). Using the aligned RNA sequencing data, a matrix of unnormalized read counts was constructed for annotated gene regions using featureCounts in Subread v2.0.1 (*86*). Last, DESeq2 (*87*) in R (*63*) 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, false discovery rate <0.05 after applied SE, and gene-level read counts ≥10 per individual in the up-regulated caste. In addition, we used a protein atlas, which examined protein expression across 26 tissues in queen and worker honey bees (*40*). Harpur *et al.* (*17*) had previously generated mutually exclusive queen- and worker-biased proteins from this resource (1582 proteins), which we used for our analysis.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at https://science.org/doi/10.1126/sciadv.abj2151

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. M. C. Arias, W. S. Sheppard, Phylogenetic relationships of honey bees (Hymenoptera: Apinae: Apini) inferred from nuclear and mitochondrial DNA sequence data. *Mol. Phylogenet. Evol.* **37**, 25–35 (2005).
2. N. Lo, R. S. Gloag, D. L. Anderson, B. P. Oldroyd, 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. *Syst. Entomol.* **35**, 226–233 (2010).
3. R. Raffiudin, R. H. Crozier, Phylogenetic analysis of honey bee behavioral evolution. *Mol. Phylogenet. Evol.* **43**, 543–552 (2007).
4. K. A. Dogantzis, A. Zayed, Recent advances in population and quantitative genomics of honey bees. *Curr. Opin. Insect Sci.* **31**, 93–98 (2019).
5. F. Ruttner, *Biogeography and Taxonomy of Honeybees* (Springer Berlin Heidelberg, 1988).
6. R. A. Ilyasov, M.-I. Lee, J.-i. Takahashi, H. W. Kwon, A. G. Nikolenko, A revision of subspecies structure of western honey bee Apis mellifera. *Saudi J. Biol. Sci.* **27**, 3615–3621 (2020).
7. U. Kotthoff, T. Wappler, M. S. Engel, Greater past disparity and diversity hints at ancient migrations of European honey bee lineages into Africa and Asia. *J. Biogeogr.* **40**, 1832–1838 (2013).
8. F. Ruttner, L. Tassencourt, J. Louveaux, Biometrical-statistical analysis of the geographic variability of *Apis mellifera* L. I. Material and Methods. *Apidologie* **9**, 363–381 (1978).
9. J. M. Cridland, N. D. Tsutsui, S. R. Ramírez, The complex demographic history and evolutionary origin of the western honey Bee, *Apis Mellifera*. *Genome Biol. Evol.* **9**, 457–472 (2017).
10. A. Wallberg, F. Han, G. Wellhagen, B. Dahle, M. Kawata, N. Haddad, Z. L. P. Simões, M. H. Allsopp, I. Kandemir, P. De la Rúa, C. W. Pirk, M. T. Webster, A worldwide survey of genome sequence variation provides insight into the evolutionary history of the honeybee *Apis mellifera*. *Nat. Genet.* **46**, 1081–1088 (2014).
11. L. Garnery, J. M. Cornuet, M. Solignac, Evolutionary history of the honey bee *Apis mellifera* inferred from mitochondrial DNA analysis. *Mol. Ecol.* **1**, 145–154 (1992).
12. C. W. Whitfield, S. K. Behura, S. H. Berlocher, A. G. Clark, J. S. Johnston, W. S. Sheppard, D. R. Smith, A. V. Suarez, D. Weaver, N. D. Tsutsui, Thrice out of Africa: Ancient and recent expansions of the honey bee, *Apis mellifera*. *Science* **314**, 642–645 (2006).
13. E. O. Wilson, *The Insect Societies* (Harvard Univ. Press, 1971).

14. B. A. Harpur, S. Minaei, C. F. Kent, A. Zayed, Management increases genetic diversity of honey bees via admixture. *Mol. Ecol.* **21**, 4414–4421 (2012).
15. F. Han, A. Wallberg, M. T. Webster, From where did the Western honeybee (*Apis mellifera*) originate? *Ecol. Evol.* **2**, 1949–1957 (2012).
16. C. M. Grozinger, A. Zayed, Improving bee health through genomics. *Nat. Rev. Genet.* **21**, 277–291 (2020).
17. B. A. Harpur, C. F. Kent, D. Molodtsova, J. M. D. Lebon, A. S. Alqarni, A. A. Owayss, A. Zayed, Population genomics of the honey bee reveals strong signatures of positive selection on worker traits. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 2614–2619 (2014).
18. Z. L. Fuller, E. L. Niño, H. M. Patch, O. C. Bedoya-Reina, T. Baumgarten, E. Muli, F. Mumoki, A. Ratan, J. McGraw, M. Frazier, D. Masiga, S. Schuster, C. M. Grozinger, W. Miller, Genome-wide analysis of signatures of selection in populations of African honey bees (Apis mellifera) using new web-based tools. *BMC Genomics* **16**, 518 (2015).
19. N. J. Haddad, W. Loucif-Ayad, N. Adjlane, D. Saini, R. Manchiganti, V. Krishnamurthy, B. AlShagoor, A. M. Batainh, R. Mugasimangalam, Draft genome sequence of the Algerian bee Apis mellifera intermissa. *Genom. Data* **4**, 24–25 (2015).
20. C. Chen, Z. Liu, Q. Pan, X. Chen, H. Wang, H. Guo, S. Liu, H. Lu, S. Tian, R. Li, W. Shi, Genomic analyses reveal demographic history and temperate adaptation of the newly discovered honey bee subspecies *Apis mellifera sinisxinyuan* n. ssp. *Mol. Biol. Evol.* **33**, 1337–1348 (2016).
21. A. Wallberg, C. Schöning, M. T. Webster, M. Hasselmann, Two extended haplotype blocks are associated with adaptation to high altitude habitats in East African honey bees. *PLOS Genet.* **13**, e1006792 (2017).
22. D. Henriques, K. A. Browne, M. W. Barnett, M. Parejo, P. Kryger, T. C. Freeman, I. Muñoz, L. Garnery, F. Highet, J. S. Jonhston, G. P. McCormack, M. A. Pinto, High sample throughput genotyping for estimating C-lineage introgression in the dark honeybee: An accurate and cost-effective SNP-based tool. *Sci. Rep.* **8**, 8552 (2018).
23. D. Henriques, M. Parejo, A. Vignal, D. Wragg, A. Wallberg, M. T. Webster, M. A. Pinto, 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*). *Evol. Appl.* **11**, 1270–1282 (2018).
24. M. Parejo, D. Wragg, L. Gauthier, A. Vignal, P. Neumann, M. Neuditschko, Using whole-genome sequence information to foster conservation efforts for the european dark honey bee, *Apis mellifera mellifera*. *Front. Ecol. Evol.* **4**, 140 (2016).
25. I. Muñoz, D. Henriques, L. Jara, J. S. Johnston, J. Chávez-Galarza, P. De La Rúa, M. A. Pinto, 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*). *Mol. Ecol. Resour.* **17**, 783–795 (2017).
26. I. Muñoz, D. Henriques, J. S. Johnston, J. Chávez-Galarza, P. Kryger, M. A. Pinto, Reduced SNP panels for genetic identification and introgression analysis in the dark honey bee (*Apis mellifera mellifera*). *PLOS ONE* **10**, e0124365 (2015).
27. M. A. Pinto, D. Henriques, J. Chávez-Galarza, P. Kryger, L. Garnery, R. van der Zee, B. Dahle, G. Soland-Reckeweg, P. De la Rúa, R. Dall'Olio, N. L. Carreck, J. S. Johnston, Genetic integrity of the Dark European honey bee (*Apis mellifera mellifera*) from protected populations: A genome-wide assessment using SNPs and mtDNA sequence data. *J. Apic. Res.* **53**, 269–278 (2014).
28. Y. Ji, The geographical origin, refugia, and diversification of honey bees (*Apis spp.*) based on biogeography and niche modeling. *Apidologie* **52**, 367–377 (2021).
29. K. S. Lamm, B. D. Redelings, Reconstructing ancestral ranges in historical biogeography: Properties and prospects. *J. Syst. Evol.* **47**, 369–382 (2009).
30. A. M. Harris, M. DeGiorgio, Admixture and ancestry inference from ancient and modern samples through measures of population genetic drift. *Hum. Biol.* **89**, 21–46 (2017).
31. E. Corsetti, N. Azpiazu, Functional dissection of the splice variants of the Drosophila gene homothorax (hth). *Dev. Biol.* **384**, 72–82 (2013).
32. S. Biswas, R. J. Russell, C. J. Jackson, M. Vidovic, O. Ganeshina, J. G. Oakeshott, C. Claudianos, Bridging the synaptic gap: Neuroligins and neurexin I in *Apis mellifera*. *PLOS ONE* **3**, e3542 (2008).
33. B. A. Harpur, S. M. Kadri, R. O. Orsi, C. W. Whitfield, A. Zayed, Defense response in Brazilian honey bees (*Apis mellifera scutellata* × spp.) is underpinned by complex patterns of admixture. *Genome Biol. Evol.* **12**, 1367–1377 (2020).
34. F. Mondet, A. Beaurepaire, A. McAfee, B. Locke, C. Alaux, S. Blanchard, B. Danka, L. C. Yves, Honey bee survival mechanisms against the parasite Varroa destructor: A systematic review of phenotypic and genomic research efforts. *Int. J. Parasitol.* **50**, 433–447 (2020).
35. H. M. G. Lattorff, J. Buchholz, I. Fries, R. F. Moritz, A selective sweep in a Varroa destructor resistant honeybee (*Apis mellifera*) population. *Infect. Genet. Evol.* **31**, 169–176 (2015).
36. V. Zanni, D. A. Galbraith, D. Annoscia, C. M. Grozinger, F. Nazzi, Transcriptional signatures of parasitization and markers of colony decline in Varroa-infested honey bees (*Apis mellifera*). *Insect Biochem. Mol. Biol.* **87**, 1–13 (2017).
37. E. Amiri, J. J. Herman, M. K. Strand, D. R. Tarpy, O. Rueppell, Egg transcriptome profile responds to maternal virus infection in honey bees, *Apis mellifera*. *Infect. Genet. Evol.* **85**, 104558 (2020).
38. D. Wragg, M. Marti-Marimon, B. Basso, J.-P. Bidanel, E. Labarthe, O. Bouchez, Y. Le Conte, A. Vignal, Whole-genome resequencing of honeybee drones to detect genomic selection in a population managed for royal jelly. *Sci. Rep.* **6**, 27168 (2016).
39. R. Ashby, S. Forêt, I. Searle, R. Maleszka, MicroRNAs in honey bee caste determination. *Sci. Rep.* **6**, 18794 (2016).
40. Q. W. Chan, M. Y. Chan, M. Logan, Y. Fang, H. Higo, L. J. Foster, Honey bee protein atlas at organ-level resolution. *Genome Res.* **23**, 1951–1960 (2013).
41. J. Chávez-Galarza, D. Henriques, J. S. Johnston, M. Carneiro, J. Rufino, J. C. Patton, M. A. Pinto, Revisiting the Iberian honey bee (*Apis mellifera iberiensis*) contact zone: Maternal and genome-wide nuclear variations provide support for secondary contact from historical refugia. *Mol. Ecol.* **24**, 2973–2992 (2015).
42. J. Chávez-Galarza, L. Garnery, D. Henriques, C. J. Neves, W. Loucif-Ayad, J. S. Jonhston, M. A. Pinto, Mitochondrial DNA variation of *Apis mellifera iberiensis*: Further insights from a large-scale study using sequence data of the $tRNA^{leu}$-cox2 intergenic region. *Apidologie* **48**, 533–544 (2017).
43. F. Cánovas, P. De la Rúa, J. Serrano, J. Galián, Geographical patterns of mitochondrial DNA variation in A*pis mellifera iberiensis* (Hymenoptera: Apidae). *J. Zool. Syst. Evol. Res.* **46**, 24–30 (2008).
44. M. A. Pinto, D. Henriques, M. Neto, H. Guedes, I. Muñoz, J. C. Azevedo, P. De la Rúa, Maternal diversity patterns of Ibero-Atlantic populations reveal further complexity of Iberian honeybees. *Apidologie* **44**, 430–439 (2013).
45. L. Boardman, A. Eimanifar, R. Kimball, E. Braun, S. Fuchs, B. Grünewald, J. D. Ellis, The mitochondrial genome of the Spanish honey bee, *Apis mellifera iberiensis* (Insecta: Hymenoptera: Apidae), from Portugal. *Mitochondrial DNA Part B* **5**, 17–18 (2020).
46. M. A. El-Niweiri, R. F. Moritz, Mitochondrial discrimination of honeybees (*Apis mellifera*) of Sudan. *Apidologie* **39**, 566–573 (2008).
47. T. G. Hailu, P. D'Alvise, A. Tofilski, S. Fuchs, J. Greiling, P. Rosenkranz, M. Hasselmann, Insights into Ethiopian honey bee diversity based on wing geomorphometric and mitochondrial DNA analyses. *Apidologie* **51**, 1182–1198 (2020).
48. P. Franck, L. Garnery, A. Loiseau, B. Oldroyd, H. Hepburn, M. Solignac, J. M. Cornuet, Genetic diversity of the honeybee in Africa: Microsatellite and mitochondrial data. *Heredity* **86**, 420–430 (2001).
49. G. L. Conte, M. E. Arnegard, C. L. Peichel, D. Schluter, The probability of genetic parallelism and convergence in natural populations. *Proc. Biol. Sci.* **279**, 5039–5047 (2012).
50. Y. Ji, X. Li, T. Ji, J. Tang, L. Qiu, J. Hu, J. Dong, S. Luo, S. Liu, P. B. Frandsen, X. Zhou, S. H. Parey, L. Li, Q. Niu, X. Zhou, Gene reuse facilitates rapid radiation and independent adaptation to diverse habitats in the Asian honeybee. *Sci. Adv.* **6**, eabd3590 (2020).
51. M. L. Winston, O. R. Taylor, G. W. Otis, Some differences between temperate European and tropical African and South American honeybees. *Bee World* **64**, 12–21 (1983).
52. B. A. Harpur, A. Dey, J. R. Albert, S. Patel, H. M. Hines, M. Hasselman, L. Packer, A. Zayed, Queens and workers contribute differently to adaptive evolution in bumble bees and honey bees. *Genome Biol. Evol.* **9**, 2395–2402 (2017).
53. K. A. Dogantzis, B. A. Harpur, A. Rodrigues, L. Beani, A. L. Toth, A. Zayed, Insects with similar social complexity show convergent patterns of adaptive molecular evolution. *Sci. Rep.* **8**, 10388 (2018).
54. C. Chen, H. Wang, Z. Liu, X. Chen, J. Tang, F. Meng, W. Shi, Population genomics provide insights into the evolution and adaptation of the eastern honey bee (*Apis cerana*). *Mol. Biol. Evol.* **35**, 2260–2271 (2018).
55. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
56. C. G. Elsik, K. C. Worley, A. K. Bennett, M. Beye, F. Camara, C. P. Childers, D. C. de Graaf, G. Debyser, J. Deng, B. Devreese, E. Elhaik, J. D. Evans, L. J. Foster, D. Graur, R. Guigo; HGSC production teams, K. J. Hoff, M. E. Holder, M. E. Hudson, G. J. Hunt, H. Jiang, V. Joshi, R. S. Khetani, P. Kosarev, C. L. Kovar, J. Ma, R. Maleszka, R. F. A. Moritz, M. C. Munoz-Torres, T. D. Murphy, D. M. Muzny, I. F. Newsham, J. T. Reese, H. M. Robertson, G. E. Robinson, O. Rueppell, V. Solovyev, M. Stanke, E. Stolle, J. M. Tsuruda, M. Van Vaerenbergh, R. M. Waterhouse, D. B. Weaver, C. W. Whitfield, Y. Wu, E. M. Zdobnov, L. Zhang, D. Zhu, R. A. Gibbs, Finding the missing honey bee genes: Lessons learned from a genome upgrade. *BMC Genomics* **15**, 86 (2014).
57. F. J. Sedlazeck, P. Rescheneder, A. Von Haeseler, NextGenMap: Fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics* **29**, 2790–2791 (2013).
58. R. Poplin, V. Ruano-Rubio, M. A. De Pristo, T. J. Fennell, M. O. Carneiro, G. A. Van der Auwera, D. E. Kling, L. D. Gauthier, A. Levy-Moonshine, D. Roazen, K. Shakir, J. Thibault, S. Chandran, C. Whelan, M. Lek, S. Gabriel, M. J. Daly, B. Neale, D. G. Mac Arthur, E. Banks, Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv* , 201178 (2017).
59. G. A. Van der Auwera, M. O. Carneiro, C. Hartl, R. Poplin, G. Del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, From FastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–11.10.33 (2013).

60. D. H. Alexander, K. Lange, Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* **12**, 246 (2011).
61. E. Linck, C. Battey, Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Mol. Ecol. Resour.* **19**, 639–647 (2019).
62. X. Zheng, D. Levine, J. Shen, S. M. Gogarten, C. Laurie, B. S. Weir, A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).
63. R Core Team, R: A language and environment for statistical computing. **201**, (2013).
64. A. Stamatakis, RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
65. T. Jombart, adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403–1405 (2008).
66. E. Paradis, J. Claude, K. Strimmer, APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290 (2004).
67. J. K. Pickrell, J. K. Pritchard, Inference of population splits and mixtures from genome-wide allele frequency data. *PLOS Genet.* **8**, e1002967 (2012).
68. S. R. Browning, B. L. Browning, Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).
69. Z. Yang, PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
70. N. J. Matzke, BioGeoBEARS: BioGeography with Bayesian (and likelihood) evolutionary analysis in R Scripts. *R package, version 0.2* **1**, 2013 (2013).
71. N. J. Matzke, Probabilistic historical biogeography: New models for founder-event speciation, imperfect detection, and fossils allow improved accuracy and model-testing. *Front. Biogeogr.* **5**, (2013).
72. R. H. Ree, S. A. Smith, Maximum likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis. *Syst. Biol.* **57**, 4–14 (2008).
73. R. H. Ree, Detecting the historical signature of key innovations using stochastic models of character evolution and cladogenesis. *Evolution* **59**, 257–265 (2005).
74. F. Ronquist, Dispersal-vicariance analysis: A new approach to the quantification of historical biogeography. *Syst. Biol.* **46**, 195–203 (1997).
75. M. J. Landis, N. J. Matzke, B. R. Moore, J. P. Huelsenbeck, Bayesian analysis of biogeography when the number of areas is large. *Syst. Biol.* **62**, 789–804 (2013).
76. P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin; 1000 Genomes Project Analysis Group, The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
77. G. Watterson, On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–276 (1975).
78. H. Liu, Y. Jia, X. Sun, D. Tian, L. D. Hurst, S. Yang, Direct determination of the mutation rate in the bumblebee reveals evidence for weak recombination-associated mutation and an approximate rate constancy in insects. *Mol. Biol. Evol.* **34**, 119–130 (2017).
79. N. Patterson, P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, D. Reich, Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
80. M. Raghavan, M. Steinrücken, K. Harris, S. Schiffels, S. Rasmussen, M. De Giorgio, A. Albrechtsen, C. Valdiosera, M. C. Ávila-Arcos, A.-S. Malaspinas, A. Eriksson, I. Moltke, M. Metspalu, J. R. Homburger, J. Wall, O. E. Cornejo, J. V. Moreno-Mayar, T. S. Korneliussen, T. Pierre, M. Rasmussen, P. F. Campos, P. de Barros Damgaard, M. E. Allentoft, J. Lindo, E. Metspalu, R. Rodríguez-Varela, J. Mansilla, C. Henrickson, A. Seguin-Orlando, H. Malmström, T. Stafford Jr., S. S. Shringarpure, A. Moreno-Estrada, M. Karmin, K. Tambets, A. Bergström, Y. Xue, V. Warmuth, A. D. Friend, J. Singarayer, P. Valdes, F. Balloux, I. Leboreiro, J. L. Vera, H. Rangel-Villalobos, D. Pettener, D. Luiselli, L. G. Davis, E. Heyer, C. P. E. Zollikofer, M. S. Ponce de León, C. I. Smith, V. Grimes, K.-A. Pike, M. Deal, B. T. Fuller, B. Arriaza, V. Standen, M. F. Luz, F. Ricaut, N. Guidon, L. Osipova, M. I. Voevoda, O. L. Posukh, O. Balanovsky, M. Lavryashina, Y. Bogunov, E. Khusnutdinova, M. Gubina, E. Balanovska, S. Fedorova, S. Litvinov, B. Malyarchuk, M. Derenko, M. J. Mosher, D. Archer, J. Cybulski, B. Petzelt, J. Mitchell, R. Worl, P. J. Norman, P. Parham, B. M. Kemp, T. Kivisild, C. Tyler-Smith, M. S. Sandhu, M. Crawford, R. Villems, D. G. Smith, M. R. Waters, T. Goebel, J. R. Johnson, R. S. Malhi, M. Jakobsson, D. J. Meltzer, A. Manica, R. Durbi, Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science* **349**, aab3884 (2015).
81. P. Cingolani, A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, D. M. Ruden, 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. *Flying* **6**, 80–92 (2012).
82. D. Molodtsova, B. A. Harpur, C. F. Kent, K. Seevananthan, A. Zayed, Pleiotropy constrains the evolution of protein but not regulatory sequences in a transcription regulatory network influencing complex social behaviors. *Front. Genet.* **5**, 431 (2014).
83. D. W. Huang, B. T. Sherman, R. A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).
84. C. G. Elsik, A. Tayal, C. M. Diesh, D. R. Unni, M. L. Emery, H. N. Nguyen, D. E. Hagen, Hymenoptera Genome Database: Integrating genome annotations in HymenopteraMine. *Nucleic Acids Res.* **44**, D793–D800 (2016).
85. A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T. R. Gingeras, STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
86. Y. Liao, G. K. Smyth, W. Shi, featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
87. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
88. A. Wallberg, I. Bunikis, O. V. Pettersson, M.-B. Mosbech, A. K. Childers, J. D. Evans, A. S. Mikheyev, H. M. Robertson, G. E. Robinson, M. T. Webster, A hybrid de novo genome assembly of the honeybee, *Apis mellifera*, with chromosome-length scaffolds. *BMC Genomics* **20**, 275 (2019).
89. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin; 1000 Genome Project Data Processing Subgroup, 2009. The sequence alignment/map format and samtools. *Bioinformatics* **25**, 2078–2079 (2009).
90. B. A. Harpur, M. M. Guarna, E. Huxter, H. Higo, K.-M. Moon, S. E. Hoover, A. Ibrahim, A. P. Melathopoulos, S. Desai, R. W. Currie, S. F. Pernal, L. J. Foster, A. Zayed, Integrative genomics reveals the genetics and evolution of the honey bee's social immune system. *Genome Biol. Evol.* **11**, 937–948 (2019).
91. P. G. Meirmans, P. W. Hedrick, Assessing population structure: FST and related measures. *Mol. Ecol. Resour.* **11**, 5–18 (2011).

Acknowledgments: We thank L. Packer for advice on ancestral range reconstruction and L. Foster, M. Guarna, S. Pernal, S. Hoover, R. Currie, and P. Giovenazzo for help in securing funding. We thank B. Oldroyd for providing samples and for helpful comments on the manuscript. **Funding:** This research was supported by funding from Genome Canada, Genome British Columbia, the Ontario Research Fund, through the Bee'Omics Project (227BEE) (to A.Z.), a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (to A.Z.), Compute Canada (eck-062-ab to A.Z.), a York University Research Chair in Genomics (to A.Z.), an Ontario Graduate Scholarship (to K.A.D.), a PAm Costco Fellowship (to K.A.D.), and the Deanship of Scientific Research at King Saud University (RGP-189, ASA). German Academic Exchange Service (DAAD) funded a scientific exchange visit to Kyrgyzstan (to E.S.). **Author contributions:** 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. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. All new honey bee genomes have been deposited on NCBI's Short Read Archive. BioProject: PRJNA729035.

Submitted 27 April 2021
Accepted 14 October 2021
Published 3 December 2021
10.1126/sciadv.abj2151

Thrice out of Asia and the adaptive radiation of the western honey bee

Kathleen A. DogantzisTanushree TiwariIda M. ConflittiAlivia DeyHarland M. PatchElliud M. MuliLionel GarneryCharles W. WhitfieldEckart StolleAbdulaziz S. AlqarniMichael H. AllsoppAmro Zayed

Sci. Adv., 7 (49), eabj2151. • DOI: 10.1126/sciadv.abj2151