A worldwide survey of genome sequence variation provides insight into the evolutionary history of the honeybee *Apis mellifera*

Andreas Wallberg[1], Fan Han[1,10], Gustaf Wellhagen[1,10], Bjørn Dahle[2], Masakado Kawata[3], Nizar Haddad[4], Zilá Luz Paulino Simões[5], Mike H Allsopp[6], Irfan Kandemir[7], Pilar De la Rúa[8], Christian W Pirk[9] & Matthew T Webster[1]

The honeybee *Apis mellifera* has major ecological and economic importance. We analyze patterns of genetic variation at 8.3 million SNPs, identified by sequencing 140 honeybee genomes from a worldwide sample of 14 populations at a combined total depth of 634×. These data provide insight into the evolutionary history and genetic basis of local adaptation in this species. We find evidence that population sizes have fluctuated greatly, mirroring historical fluctuations in climate, although contemporary populations have high genetic diversity, indicating the absence of domestication bottlenecks. Levels of genetic variation are strongly shaped by natural selection and are highly correlated with patterns of gene expression and DNA methylation. We identify genomic signatures of local adaptation, which are enriched in genes expressed in workers and in immune system– and sperm motility–related genes that might underlie geographic variation in reproduction, dispersal and disease resistance. This study provides a framework for future investigations into responses to pathogens and climate change in honeybees.

Insect pollination is necessary for one-third of our food and is a vital part of the ecosystem. The honeybee *A. mellifera* is a key pollinator, with its services to agriculture valued at >$200 billion per year worldwide[1]. It is therefore a major cause of concern that honeybees have faced huge and largely unexplained colony losses in recent decades[2]. However, little is known about global patterns of genomic variation in this species, which hold the key to an understanding of its evolutionary history, the biological basis of adaptation to different climates and mechanisms governing resistance to disease.

The native distribution of *A. mellifera* encompasses Africa, Europe and western Asia[3–8], and molecular dating suggests that the population expanded into this range around 1 million years ago[3,4]. Conflicting hypotheses have been proposed for the origin of this expansion[8], with analyses of limited numbers of genetic and morphometric markers supporting an origin in the Middle East[3–5] and a study of nuclear SNPs arguing for an African origin[7,9]. Honeybees show substantial phenotypic variation across their extensive geographic range. European bees exhibit morphological and behavioral adaptations to survive colder winters, whereas African colonies are more aggressive and show a greater tendency to swarm. African bees are also reported to have greater resistance to the pathogenic mite *Varroa destructor*[10–12], a major honeybee pathogen[13,14]. The genetic basis of this phenotypic variation is largely unknown.

Humans began harvesting wax and honey from honeybee colonies at least 7,000 years before the present[15]. Human activity has led to the transportation of honeybee colonies all over the world, artificial selection for desirable traits and gene flow between native subspecies[16], including the expansion of hybrid strains of Africanized bees, known for their highly aggressive stinging behavior, across the Americas[17] after their introduction to Brazil. The effects of these processes on the levels of genetic variation in honeybees have not been comprehensively evaluated. Here we investigate the evolution and genetic basis of adaptation in honeybees by performing whole-genome sequencing of 140 *A. mellifera* worker bees from 14 separate populations from a worldwide sample.

RESULTS

Global patterns of variation

We sampled *A. mellifera* from a total of 14 populations, which included 9 native subspecies chosen from across the native range of the species in addition to managed strains of mixed ancestry from apiaries in Europe and North America and Africanized bees from South America

[1]Department of Medical Biochemistry and Microbiology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden. [2]Norwegian Beekeepers Association, Kløfta, Norway. [3]Department of Ecology and Evolutionary Biology, Graduate School of Life Sciences, Tohoku University, Sendai, Japan. [4]Bee Research Department, National Center for Agricultural Research and Extension, Amman, Jordan. [5]Department of Biology, University of São Paulo, São Paulo, Brazil. [6]Plant Protection Research Institute, Agricultural Research Council, Stellenbosch, South Africa. [7]Department of Biology, Ankara University, Ankara, Turkey. [8]Department of Zoology and Physical Anthropology, University of Murcia, Murcia, Spain. [9]Department of Zoology and Entomology, University of Pretoria, Pretoria, South Africa. [10]These authors contributed equally to this work. Correspondence should be addressed to A.W. (andreas.wallberg@imbim.uu.se) or M.T.W. (matthew.webster@imbim.uu.se).

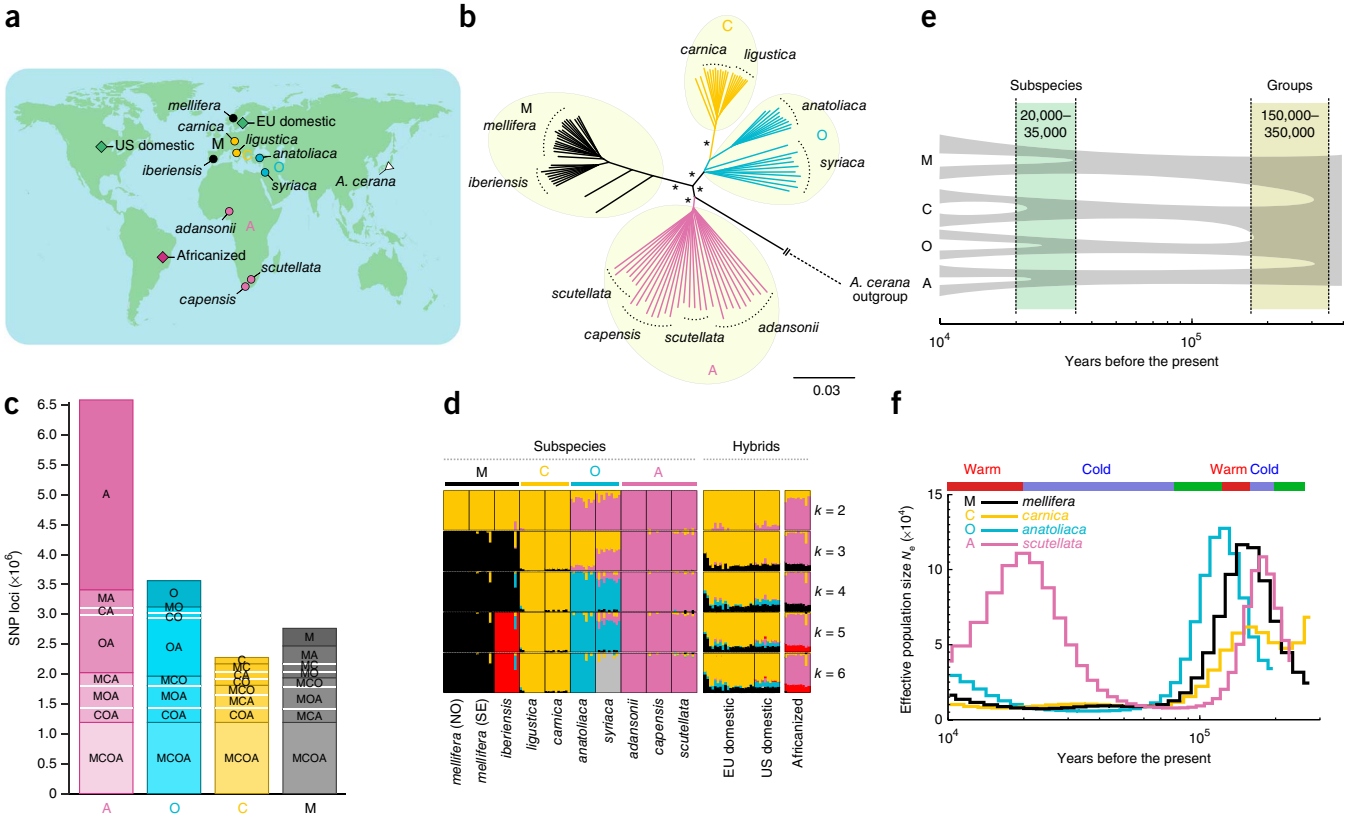

Figure 1 Geographic distribution of genetic variation and demographic history of honeybees. (**a**) Origin of analyzed samples. Worker bee samples from native subspecies from the 4 main geographic groups (colored circles) were collected from Europe (n = 50), Africa (n = 30) and the Middle East (n = 20). Samples from domestic strains (green diamonds) were collected from Europe (n = 20) and North America (n = 10). Africanized bees (n = 10) were collected from Brazil, and samples of a closely related species (*A. cerana*) were collected from Japan (n = 10). (**b**) Neighbor-joining tree constructed from allele-sharing distances between native subspecies. Nodes leading to the four geographic groups with 100% support are marked with an asterisk; the scale bar gives raw genetic distance per variable site. (**c**) Total numbers of variable SNPs in each of the four groups. For each group, SNPs are categorized according to the number of other groups in which they are polymorphic. (**d**) ADMIXTURE analysis showing clustering of samples into 2–6 groups, including native subspecies and hybrid strains. The inferred proportion of ancestry shared with each group is shown for each sample. EU, Europe; NO, Norway; SE, Sweden; US, United States. (**e**) Simplified model of population splits during honeybee evolution, with approximate dates of splits between groups and subspecies inferred by genealogical concordance. YBP, years before the present. (**f**) PSMC analysis performed on representatives of each group sequenced to high coverage showing inferred variation in N_e over time. Historical global temperature fluctuations are also marked. Generation time (g) = 1 year; transversion mutation rate (μ) = 0.15 × 10^{-8} mutations per bp per generation.

(n = 10 for each population; **Fig. 1a** and **Table 1**). We also sequenced *Apis cerana* workers from Japan (n = 10) and a haploid drone from the DH4 strain descended from the sample used to construct the genome assembly[9] for quality control (**Supplementary Fig. 1**). In total, we obtained a genomic data set with 634× coverage and called 8.3 million SNPs (**Table 1**, **Supplementary Fig. 2** and **Supplementary Note**).

An evolutionary tree of samples from native *A. mellifera* subspecies (**Fig. 1b** and **Supplementary Fig. 3**) inferred from all SNPs demonstrated strong clustering of samples according to four major groups previously delineated on the basis of morphometric and genetic classification[3–8]: group A (comprising subspecies from Africa), group M (comprising subspecies from western and northern Europe), group C (comprising subspecies from eastern and southern Europe) and group O (comprising subspecies from the Middle East and western Asia). A previous study of nuclear SNPs argued for an African origin on the basis of the position of the root of a phylogenetic tree[7,9], although a reanalysis of these data did not support this finding[8]. The root of our tree, defined by the *A. cerana* sequences, was placed unequivocally between the four clades. An African origin of *A. mellifera* is therefore not supported by our data, which did not identify any of the extant groups as being ancestral. Our analysis therefore does not explicitly support a specific model of *A. mellifera* origin but is most parsimonious with an origin in Asia, considering that all other extant *Apis* species are found there.

Levels of genetic variation were high in all samples. Among the native *A. mellifera* subspecies, those from Africa harbored the greatest variation. Watterson's estimator of the population mutation rate per base (θ_w) in African bees was 0.79%, whereas native European subspecies had lower levels of variation (average θ_w values of 0.30% and 0.33% for the C and M groups, respectively), and Middle-Eastern subspecies were intermediate (average θ_w value of 0.47%; **Table 1** and **Supplementary Note**), in concordance with previous studies based on a few loci[16]. We also note an extremely fast decay of linkage disequilibrium (LD) with physical distance, which reflects the high recombination rate in honeybees[18] (~50% reduction in the r^2 linkage statistic with only 500 bp; **Supplementary Fig. 4**). We estimated the effective population size (N_e) in European populations as ~200,000, whereas it was much higher in Africa (~500,000; **Table 1**). African populations also showed the lowest levels of LD, consistent with the higher N_e (**Supplementary Fig. 4**). Higher N_e estimates in Africa are consistent with other studies of genetic variation[16,19,20], and the current population of wild African bees is known to be larger than the corresponding population

Table 1 Genetic variation and effective population sizes

Sample	Number of samples	Variable SNPs	θ_w	N_e
A group				
adansonii	10	4,578,517	0.0072	457,253
capensis	10	4,193,692	0.0066	418,821
scutellata	10	4,005,286	0.0063	400,005
A group total	30	6,583,102	0.0079	500,184
O group				
anatoliaca	10	1,916,693	0.0030	191,419
syriaca	10	3,136,725	0.0049	313,262
O group total	20	3,580,686	0.0047	298,263
C group				
carnica	10	1,690,039	0.0027	168,783
ligustica	10	1,745,809	0.0028	174,353
C group total	20	2,275,598	0.0030	189,552
M group				
iberiensis	10	2,181,659	0.0034	217,881
mellifera (N)	10	1,578,044	0.0025	157,598
mellifera (S)	10	1,777,165	0.0028	177,484
M group total	30	2,764,459	0.0033	210,043
All native subspecies	100	7,928,360	0.0076	434,262
Other samples				
Africanized	10	4,021,673	0.0063	401,641
EU domestic 1	10	2,082,546	0.0033	207,982
EU domestic 2	10	2,424,202	0.0038	242,103
US domestic	10	2,633,877	0.0042	263,043
All	140	8,282,459	0.0075	472,537

in Europe[19]. However, the eusocial structure of honeybees is commonly believed to result in low N_e values, and our estimates of N_e are much higher than previous ones[21]. Our results suggest that mechanisms such as an extremely panmictic mating system and extensive geographic gene flow[22] maintain high levels of genetic variation in honeybee populations.

In general, there was a high degree of allele sharing among honeybee populations. About 1 million SNPs were polymorphic in all 4 major groups of honeybee (**Fig. 1c**). The higher genetic variation exhibited by mixed domestic beekeeping strains in both Europe and North America (**Table 1**) reflects their hybrid origins[16]. Honeybees are unusual among domestic species in that recent human management has increased genetic diversity in comparison to ancestral wild populations. Africanized bees from South America harbor similar levels of variation to those observed in African subspecies, which is a striking observation given that this additional variation is derived from a limited number (48) of mated queens from Africa[17].

Demographic history

Analyses of genetic co-ancestry partitioned native samples into the four known groups with high confidence[3–8] (**Fig. 1d**, **Supplementary Figs. 5** and **6**, and **Supplementary Note**). Subspecies from different groups had an average pairwise F_{ST} (allelic fixation index) of 0.42 and could be clearly distinguished, but there was extremely little genetic differentiation between subspecies within groups (average F_{ST} = 0.10; **Supplementary Fig. 7**). The domestic strains from both Europe and North America were strikingly similar and clustered primarily within the C group, likely owing to the dominant influence of the Italian bee *A. mellifera ligustica* in beekeeping[5]. The Africanized population from South America showed mostly African ancestry, with the contribution of European alleles from the M group. We detected evidence of admixture in the *A. mellifera syriaca* subspecies from Jordan, which we estimated to have derived ~18% of its genetic ancestry from African (A-group) bees.

The relationship between the four honeybee groups suggests an ancient split between them followed by the more recent divergence of subspecies within each group (**Fig. 1e**). Previous efforts at molecular dating have estimated that the four groups split from each other around 1 million years ago[3,4]. Here we used a genealogical concordance method[23] to estimate the divergence times between the A, C and M groups in the range of $0.59–0.98 \times 1.5N_e$ generations, which indicates that they split from each other around 300,000 years before the present. Although the European M and C clades were highly genetically differentiated, this variation seemed to be a consequence of increased genetic drift in smaller populations rather than an older split. The C and O clades appeared to have diverged more recently from each other ($0.58 \times 1.5N_e$ or ~165,000 years before the present). We estimated that the splits between subspecies within each of the four groups occurred $0.031–0.180 \times 1.5N_e$ generations ago, which corresponds to 13,000–38,000 years before the present, assuming a separate N_e for each group (**Fig. 1e** and **Supplementary Fig. 8**). An older divergence time was estimated between subspecies of the O group, which could be attributed to admixture in *A. mellifera syriaca*. These dates should be considered to represent minimum divergence times, as it is possible that honeybee clades diverged earlier but gene flow between them continued.

We performed a pairwise sequentially Markovian coalescent (PSMC) analysis[24] to infer historical changes in N_e, using single representatives of each group sequenced at higher coverage. We inferred striking fluctuations in N_e over time that mirrored glacial cycles[25] (**Fig. 1f** and **Supplementary Fig. 9**). African populations appeared to have peaked in size during periods of glaciation in temperate latitudes, whereas European populations expanded or reached their maxima during interglacial periods. Since the last glacial maximum 20,000 years ago, African populations have been declining, whereas non-African populations have been gradually expanding. Taken together, these analyses are consistent with *A. mellifera* colonizing its native geographic range >300,000 years ago, after which time the M and C lineages were confined to separate glacial refugia in southern Europe. These populations began to recolonize Europe after the last ice age, at a time when African populations were already abundant.

Pervasive influence of selection on the genome

We investigated the evolutionary forces affecting different functional classes of genes by analyzing the effects of selection on local variation. Genetic variation was reduced by ~50% within protein-coding exons and UTRs in comparison to introns and flanking noncoding regions (**Fig. 2a**), indicative of the effects of purifying selection on functional regions (**Supplementary Note**). Previous studies have shown that, in the honeybee genome[26] (and the genomes of a wide variety of invertebrates[27]), genes are divided into two distinct categories, which can be distinguished through a bimodal distribution of observed/expected CpG content ($CpG_{O/E}$). One low-CpG-content class, methylated in the germ line, is associated with housekeeping functions, and a second high-CpG-content class is associated with caste-specific functions[28,29]. We first sought to clarify this relationship by analyzing two gene expression data sets: one that contrasted expression levels in workers with those in queens[30] and one that contrasted expression in workers with that in drones[31] (**Fig. 2b**). Genes with increased expression in queens (average $CpG_{O/E}$ = 1.19) and workers (1.22) were slightly over-represented in the high-CpG-content category in comparison to those that were not biased toward expression in queens or workers (1.16). Worker-biased genes were strongly over-represented in the high-CpG-content category (average $CpG_{O/E}$ = 1.41) in comparison

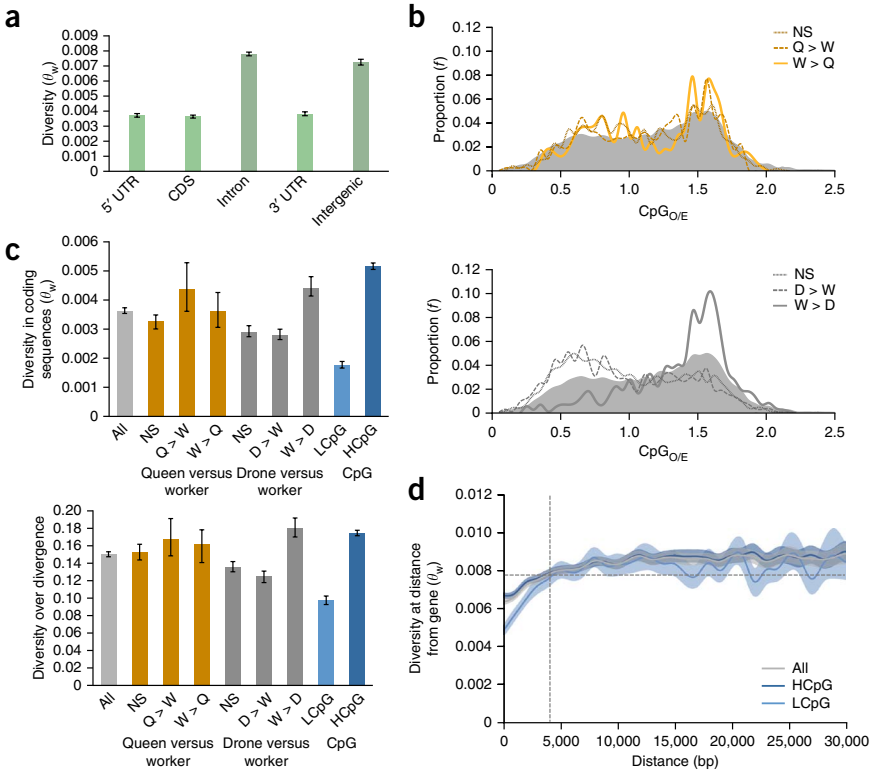

Figure 2 Genetic variation associated with gene function. (**a**) Mean levels of genetic variation in gene elements; 95% confidence intervals are estimated by bootstrap. CDS, coding sequence. (**b**) Correspondence between the $CpG_{O/E}$ content of genes and patterns of gene expression, comparing queens and workers (top) and workers and drones (bottom). (Q > W, W > Q, gene expression in queens significantly higher than workers or vice versa; D > W, W > D, gene expression in drones significantly higher than workers or vice versa; NS, unbiased gene expression.) (**c**) Mean levels of diversity (top) and diversity over divergence (bottom) in the coding region of categories of genes defined by expression patterns (as shown in **b**) and CpG content (LCpG, low CpG content; HCpG, high CpG content); 95% confidence intervals are estimated by bootstrap. (**d**) Levels of diversity in noncoding regions as a function of distance to the nearest protein-coding gene for high-CpG-content and low-CpG-content genes. The dotted horizontal line represents genome average levels of noncoding diversity. The dotted vertical line represents the average distance to a gene. 95% confidence intervals are estimated by bootstrap.

to both drone-biased genes (1.00) and genes whose expression was not biased toward either drones or workers (1.02). This finding suggests that germline-methylated genes tend to exhibit expression that is either unbiased or biased toward males (drones), whereas unmethylated genes tend to be biased toward expression in females (workers and queens).

We next examined genetic variation in genes according to these categories (**Fig. 2c**). The most striking pattern observed was that low-CpG-content genes had greatly reduced levels of variation in comparison to high-CpG-content genes (45% reduction). Consistent with this observation, we also found that genes with either unbiased or male-biased expression tended to have lower levels of variation. Levels of variation were reduced in these genes after correcting for the levels of divergence, which indicates that patterns of variation are also reduced by the effects of background selection (selection on linked variants). These results are consistent with the greater evolutionary conservation of germline-methylated genes and their role in housekeeping processes[26–29].

We also noted an average reduction in genetic variation in regions flanking genes. However, this effect extended only about 15,000 bp (or ~0.29 cM, in units of recombination frequency) (**Fig. 2d**), which is consistent with the effect of the extremely high recombination rates observed in honeybees reducing linkage with selected variants. Sites in the immediate vicinity of genes (<2 kb away) had a 16% reduction in diversity relative to those distant from genes (>20 kb away). However, the majority of the genome (~77%) was within 15 kb of a gene and showed an average reduction in variation of 9% in comparison with distant sites. It therefore seems that the majority of the honeybee genome is affected by linked selection, which is more pronounced around low-CpG-content genes. This finding suggests that, as in *Drosophila melanogaster*[32], selection has a pervasive impact on the honeybee genome and is not limited by a small effective population size.

Genomic signatures of local adaptation

To uncover genetic variants involved in local adaptation, we performed comparisons of the two European (M and C) groups and the African (A) group that each had independent histories from one another in different geographic regions. We excluded the O group because of its genetic proximity to the C group and recent admixture with the A group. We measured F_{ST} at every SNP for all three possible pairwise comparisons of two groups pooled together in comparison with the remaining group (**Fig. 3a** and **Supplementary Note**). In each comparison, there was a striking increase in the proportion of SNPs that were located within protein-coding regions at levels of F_{ST} greater than 0.9 (**Fig. 3b**), which is strong evidence for the action of positive selection. Among SNPs fixed for different alleles in Africa (A) versus Europe (MC), we found 43% of SNPs in protein-coding regions in comparison to 7% of SNPs in the data set as a whole ($P < 2 \times 10^{-16}$, chi-squared test). On average, however, the average F_{ST} of SNPs in protein-coding regions was not significantly different from that for SNPs in noncoding regions ($P = 0.545$, significance from bootstrap) (**Supplementary Fig. 10**).

Window-based F_{ST} decayed rapidly from high-F_{ST} SNPs to background levels, on average at distances of only 20–30 kb (**Fig. 3c**). We found significantly reduced levels of variation around SNPs with $F_{ST} > 0.9$, indicative of the effects of positive selection on linked variants, which extended an average of 100 kb (**Fig. 3d** and **Supplementary Fig. 11**). For the 194 SNPs that were fixed for alternate alleles in the A versus MC comparison, there was a 23% reduction in linked (≤20 kb) neutral diversity. Where possible, we categorized high-F_{ST} SNPs according to which population had a high frequency of the derived allele. Very few derived alleles were found at high frequency in one African and one European group in the C versus AM and the M versus AC comparisons. However, in the A versus MC comparison, about half of the derived alleles were at high frequency in both European (M and C) groups and half were at high frequency in the African (A) group (**Fig. 3e**). Among these two groups of variants are likely to be ones that are responsible for adaptation to temperate and tropical climates, respectively.

Genes under selection

We identified genes associated with the most differentiated SNPs taken from the top 0.1% of the F_{ST} distribution for each comparison as candidates for positive selection (**Fig. 4a**, **Supplementary Note** and **Supplementary Data Set 1**). We found that high-CpG-content genes and genes with worker-biased expression patterns were over-represented among these genes, whereas the low-CpG-content housekeeping class, as well as genes with drone-biased expression, were under-represented (**Fig. 4b** and **Supplementary Fig. 12a**). This finding indicates that, despite the fact that workers are sterile and do not directly pass on favorable alleles, their behavior and physiology are a major target of selection. In support of this idea, another

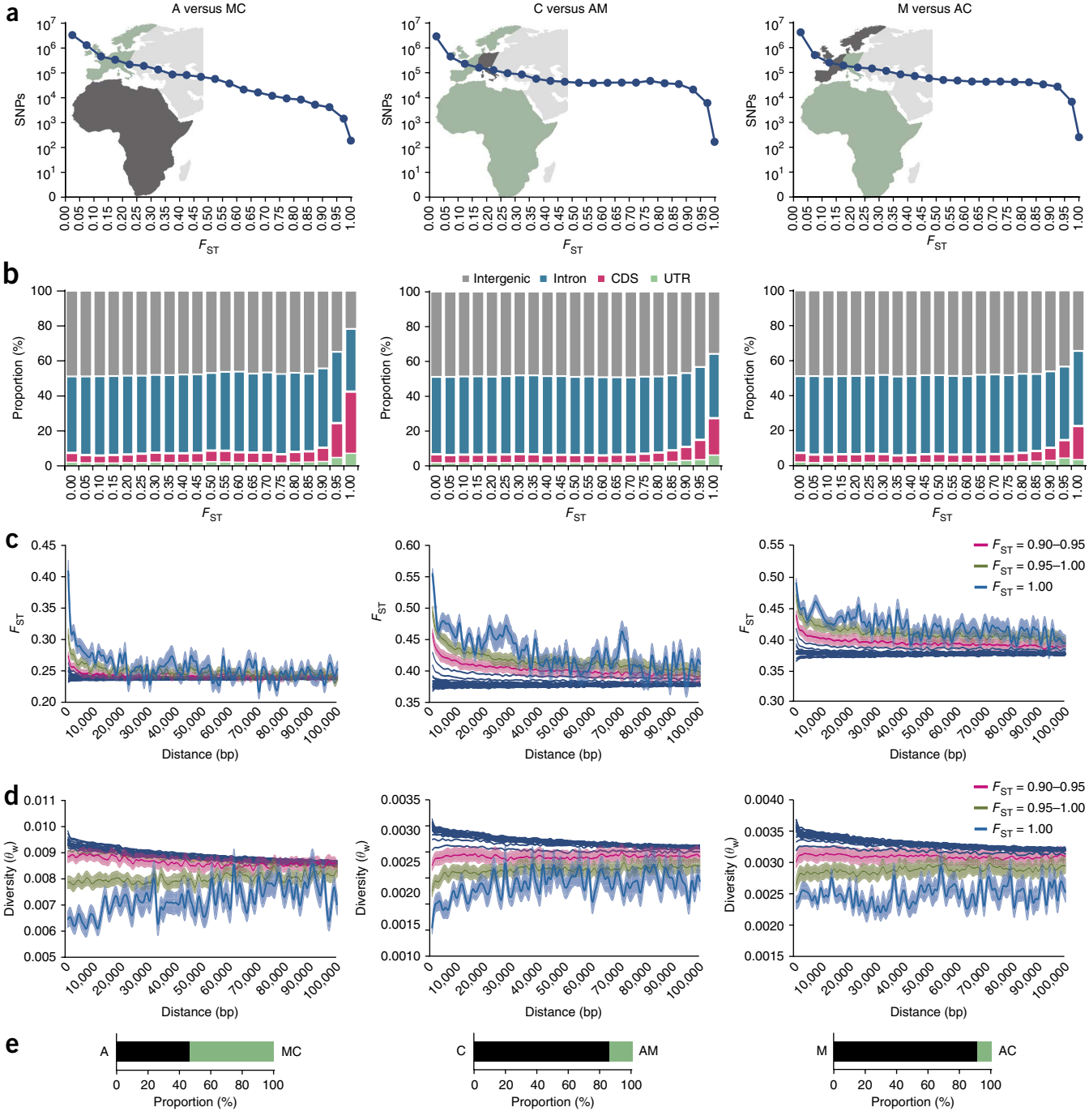

Figure 3 Genes under positive selection. (**a**) Number of SNPs divided by F_{ST} intervals observed in three pairwise comparisons of groups (A versus MC, C versus AM and M versus AC). The last interval contains the fixed SNPs. (**b**) Partitioning of SNPs according to the genomic element where they occur. The last interval contains the fixed SNPs. Levels of F_{ST} measured in windows around SNPs at different levels of F_{ST} as a function of physical distance; 95% confidence intervals are estimated by bootstrap. (**c**) Linked allelic differentiation (F_{ST}) measured in windows around SNPs at different levels of F_{ST} as a function of physical distance; 95% confidence intervals are estimated by bootstrap. (**d**) Levels of genetic diversity (θ_w) measured in windows around SNPs at different levels of F_{ST} as a function of physical distance; 95% confidence intervals are estimated by bootstrap. (**e**) Allocation of the derived allele in SNPs with F_{ST} of 0.95 and greater.

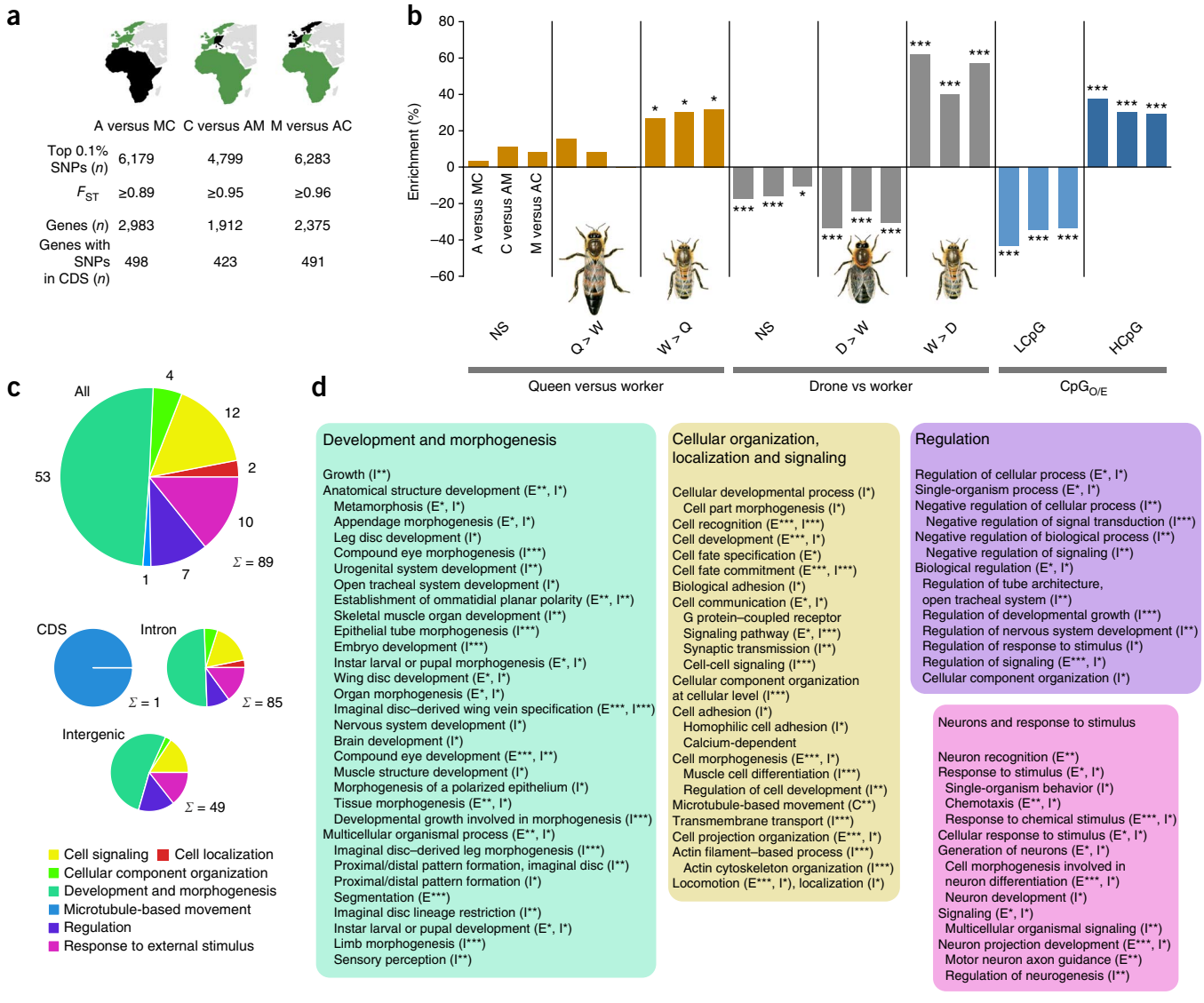

Figure 4 Function of selected genes. (**a**) Number and properties of SNPs in the top 0.1% of the F_{ST} distributions in each of the three pairwise comparisons of groups (A versus MC, C versus AM and M versus AC). Every SNP is classified according to its nearest gene. (**b**) Enrichment of genes according to $CpG_{O/E}$ category and expression pattern for SNPs within the top 0.1% of the F_{ST} distribution in the three comparisons (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, Fisher's exact test). (**c**) Enrichment of GO categories of genes containing highly differentiated SNPs in the A versus MC comparison summarized into broad categories by REViGO. The number of nonredundant GO terms in each category is shown around the top chart; the size of the slices represents the sum of the uniqueness scores of the GO terms defined by REViGO. Σ, total number of nonredundant GO terms. (**d**) Top-level and nested bottom-level biological process GO terms associated with genes with highly differentiated SNPs in the A versus MC comparison (C, coding; I, intronic; E, intergenic) (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$) and manually arranged according to the broad categories. Honeybee pictures in **b** were provided by courtesy of Encyclopaedia Britannica, Inc., copyright 2006; used with permission.

study identified an excess of genes with worker-biased expression under positive selection in the *A. mellifera* lineage and showed that positively selected genes were more likely to be taxonomically restricted[33]. We analyzed the Gene Ontology (GO) of the selection candidates, subdividing variants according to genomic location (protein-coding sequence, intron and intergenic region), and detected 262 significantly enriched GO terms ($P < 0.05$; **Supplementary Data Set 2**). Many highly differentiated noncoding SNPs, which are likely to have regulatory functions, were associated with devel opment and morphogenesis (**Fig. 4c**,**d** and **Supplementary Fig. 12b**). Candidate genes linked to abnormalities in muscles, bristles, trachea, appendages or mouth parts in *Drosophila* mutants might reflect selection on morphological variation across the geographic range of honeybees (see the **Supplementary Note** for details). We detected an increase in the proportion of nonsynonymous substitutions among high-F_{ST} SNPs in the coding regions of genes, evidence for positive selection on amino acid sequences in bees (**Supplementary Fig. 10c**).

We also identified selection in genes encoding proteins involved in cell signaling and response to stimulus (**Fig. 4c**,**d**) in addition to neuropeptides, protein hormones, glycolytic enzymes and G protein–coupled receptors, which control social behavior, development, feeding and reproduction[34]. African and European bees differed in the *AST* gene encoding the juvenile hormone–inhibiting allatostatin, an Ih-like dopamine regulator, and the genes for several odorant receptors. We detected a nonsynonymous variant in the neuronal gene *RhoGAP100F*

(Rho GTPase–activating protein at 100F), which evolves rapidly in highly social insect lineages[35]. We found highly differentiated SNPs in the exonic regions of key genes in the insulin-vitellogenin signaling pathway, which is important for queen longevity and for worker labor division[36], including a large number of highly differentiated SNPs in the 3′ UTR of *IhR* (encoding the insulin-like receptor) and nonsynonymous SNPs in the *foxo* (forkhead box, subgroup O), *Vg* (vitellogenin) and *yl* (vitellogenin receptor) genes (**Supplementary Fig. 13a**). It has been suggested that worker bees from temperate climates have increased capacity for vitellogenin storage, an adaptation that increases the longevity of overwintering bees[37]. This pathway might be dynamically evolving in bees, leading to geographic differences in queen longevity and fecundity[6].

The only significantly enriched GO biological process among all coding SNPs with signals of selection was microtubule-based movement, represented by several genes of the dynein sperm motor protein complex, including dynein intermediate chain 3 and the dynein heavy chain genes 1, 2, 3, 7, 10 and 16F (**Supplementary Fig. 13c**), whose orthologs are nearly all strongly expressed in *Drosophila* testis (data from FlyBase). We also identified *Est-6*, encoding seminal fluid enzyme esterase 6, which controls mating behavior in *Drosophila*[38], and the *Pex16* gene (peroxin 16), involved in spermatocyte maturation[39]. Queens often mate with 20 or more drones, a mating system that can be expected to induce sperm competition and rapid sperm evolution. Selection due to sperm competition might be stronger in African bees owing to a higher degree of polyandry[40,41]. The reproductive success of African over European drones when experimentally mated with queens[42] could potentially be explained by the differences we observed in sperm-related genes between bees from these geographic regions.

Honeybee immunity and response to infection range from the innate immune system to hygienic behaviors at the colony level. African bees differ from European bees in their capacity to tolerate and survive infection by *Varroa* mites[10,11] and, likely, other pathogens[43]. In the A versus MC comparison, nonsynonymous high-F_{ST} SNPs were significantly enriched in a GO category related to antibacterial peptides from the Imd pathway ($P < 0.008$), including *PGRP-LC*, *Rel* (relish) and *Iap2* (baculoviral IAP repeat–containing protein 4). Further highly differentiated coding SNPs were found in many innate immune defense genes within the Toll and JAK-STAT pathways, including *pll* (pelle) and several serine proteases. *pll* was earlier detected as being quickly evolving in social insects[35]. Among the top coding A versus MC SNPs, we also detect significant ($P = 0.012$) enrichment for genes involved in platelet plug formation, which is activated to heal wounds and form infection barriers in insects[44], including *Hml* (hemolectin) (**Supplementary Fig. 13d**). We also detected nonsynonymous variants in *Vps13*, encoding the autophagous vacuolar protein sorting 13, and *GMCOX12* and *GMCOX13*, encoding two encapsulating glucose dehydrogenase proteins, which are involved in the cellular response to pathogens. Functional characterization of these variants is likely to provide insights into the mechanisms of pathogen response in honeybees.

DISCUSSION

This study provides insights into the origins, evolution and genetic basis of adaptation in *A. mellifera*, a species of crucial importance to human society and the natural world. Our analysis indicates that honeybee population sizes have varied greatly in the past, likely owing to climatic changes. We find no evidence for an African origin of *A. mellifera*[7], and an origin closer to the only other *Apis* species, which are all restricted to Asia, is more consistent with our analysis[8]. Genes encoding worker traits are often involved in adaptation, which supports a role for kin selection. Our inference of strong selection on a number of genes with probable roles in sperm motility and maturation implicates sperm competition as a major driver of honeybee evolution. Differences in these genes between African and European bees might explain the reproductive advantage of Africanized bees[42]. We also identify a number of differences in genes involved in immunity between African and European honeybees, which might explain differences in disease resistance. Further studies of the genes and candidate mutations identified here will be useful for protecting honeybee populations from current and future challenges, including climate change and pathogens.

METHODS

Methods and any associated references are available at the end of this paper

Accession codes. All data from this study have been deposited at the NCBI Sequence Read Archive (SRA) under BioProject PRJNA236426.

ACKNOWLEDGMENTS

We thank I. Fries, J. Evans, M. Lodesani, M. Spivak, I. Arvidsson, A. Yusuf and P. Rosenkranz for providing samples. We also thank I. Jonasson, C. Tellgren-Roth, L. Nyberg and the Uppsala Genome Centre for performing the sequencing and mapping. Bioinformatic analyses were performed using resources at the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX). We thank I. Fries, D. Hultmark, A. Eyre-Walker, M. Jakobsson, L. Andersson and K. Lindblad-Toh for helpful discussions and comments on the manuscript. This study was funded by the Swedish Research Council Formas (grant 2010-1295).

AUTHOR CONTRIBUTIONS

M.T.W. conceived the study. A.W. and M.T.W. wrote the manuscript. A.W. performed all the analyses, including DNA extraction, quality control and variant calling, analysis of diversity and divergence, demographic analysis, gene annotation, selection analysis and GO analysis. M.T.W. contributed to all analyses. F.H. and G.W. contributed to the analysis of the levels of genetic variation and population substructure. B.D., M.K., N.H., Z.L.P.S., M.H.A., I.K., P.D.l.R. and C.W.P. provided samples and gave comments on the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Gallai, N., Salles, J.-M., Settele, J. & Vaissière, B.E. Economic valuation of the vulnerability of world agriculture confronted with pollinator decline. *Ecol. Econ.* **68**, 810–821 (2009).
2. Neumann, P. & Carreck, N. Honey bee colony losses. *J. Apic. Res.* **49**, 1–6 (2010).
3. Garnery, L., Cornuet, J.M. & Solignac, M. Evolutionary history of the honey bee *Apis mellifera* inferred from mitochondrial DNA analysis. *Mol. Ecol.* **1**, 145–154 (1992).
4. Arias, M.C. & Sheppard, W.S. Molecular phylogenetics of honey bee subspecies (*Apis mellifera* L.) inferred from mitochondrial DNA sequence. *Mol. Phylogenet. Evol.* **5**, 557–566 (1996).
5. Ruttner, F. *Biogeography and Taxonomy of Honeybees* (Springer-Verlag, Berlin, 1988).
6. Hepburn, H.R. & Radloff, S.E. *Honeybees of Africa* (Springer-Verlag, Berlin, 1998).
7. Whitfield, C.W. *et al.* Thrice out of Africa: ancient and recent expansions of the honey bee, *Apis mellifera*. *Science* **314**, 642–645 (2006).
8. Han, F., Wallberg, A. & Webster, M.T. From where did the Western honeybee (*Apis mellifera*) originate? *Ecol. Evol.* **2**, 1949–1957 (2012).

9. The Honeybee Genome Sequencing Consortium. Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature* **443**, 931–949 (2006).
10. Fazier, M. *et al.* A scientific note on *Varroa destructor* found in East Africa; threat or opportunity? *Apidologie (Celle)* **41**, 463–465 (2010).
11. Martin, S.J. & Medina, L.M. Africanized honeybees have unique tolerance to *Varroa* mites. *Trends Parasitol.* **20**, 112–114 (2004).
12. Strauss, U. *et al.* Seasonal prevalence of pathogens and parasites in the savannah honeybee (*Apis mellifera scutellata*). *J. Invertebr. Pathol.* **114**, 45–52 (2013).
13. Sammataro, D., Gerson, U. & Needham, G. Parasitic mites of honey bees: life history, implications, and impact. *Annu. Rev. Entomol.* **45**, 519–548 (2000).
14. Dietemann, V., Pirk, C.W.W. & Crewe, R. Is there a need for conservation of honeybees in Africa? *Apidologie (Celle)* **40**, 285–295 (2009).
15. Crane, E. *The World History of Beekeeping and Honey Hunting* (Routledge, New York, 1999).
16. Harpur, B.A., Minaei, S., Kent, C.F. & Zayed, A. Management increases genetic diversity of honey bees via admixture. *Mol. Ecol.* **21**, 4414–4421 (2012).
17. Scott Schneider, S., DeGrandi-Hoffman, G. & Smith, D.R. The African honey bee: factors contributing to a successful biological invasion. *Annu. Rev. Entomol.* **49**, 351–376 (2004).
18. Beye, M. *et al.* Exceptionally high levels of recombination across the honey bee genome. *Genome Res.* **16**, 1339–1344 (2006).
19. Moritz, R.F.A., Kraus, F.B., Kryger, P. & Crewe, R.M. The size of wild honeybee populations (*Apis mellifera*) and its implications for the conservation of honeybees. *J. Insect Conserv.* **11**, 391–397 (2007).
20. Jaffé, R. *et al.* Estimating the density of honeybee colonies across their natural range to fill the gap in pollinator decline censuses. *Conserv. Biol.* **24**, 583–593 (2010).
21. Baudry, E. *et al.* Relatedness among honeybees (*Apis mellifera*) of a drone congregation. *Proc. R. Soc. B–Biol. Sci.* **265**, 2009–2014 (1998).
22. Jaffé, R., Dietemann, V., Crewe, R.M. & Moritz, R.F.A. Temporal variation in the genetic structure of a drone congregation area: an insight into the population dynamics of wild African honeybees (*Apis mellifera scutellata*). *Mol. Ecol.* **18**, 1511–1522 (2009).
23. Wakeley, J. *Coalescent Theory: An Introduction* (Roberts & Co. Publishers, Greenwood Village, Colorado, USA, 2009).
24. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
25. Augustin, L. *et al.* Eight glacial cycles from an Antarctic ice core. *Nature* **429**, 623–628 (2004).
26. Elango, N., Hunt, B.G., Goodisman, M.A. & Yi, S.V. DNA methylation is widespread and associated with differential gene expression in castes of the honeybee, *Apis mellifera*. *Proc. Natl. Acad. Sci. USA* **106**, 11206–11211 (2009).
27. Sarda, S., Zeng, J., Hunt, B.G. & Yi, S.V. The evolution of invertebrate gene body methylation. *Mol. Biol. Evol.* **29**, 1907–1916 (2012).
28. Nanty, L. *et al.* Comparative methylomics reveals gene-body H3K36me3 in *Drosophila* predicts DNA methylation and CpG landscapes in other invertebrates. *Genome Res.* **21**, 1841–1850 (2011).
29. Lyko, F. *et al.* The honey bee epigenomes: differential methylation of brain DNA in queens and workers. *PLoS Biol.* **8**, e1000506 (2010).
30. Grozinger, C.M., Fan, Y., Hoover, S.E.R. & Winston, M.L. Genome-wide analysis reveals differences in brain gene expression patterns associated with caste and reproductive status in honey bees (*Apis mellifera*). *Mol. Ecol.* **16**, 4837–4848 (2007).
31. Zayed, A., Naeger, N.L., Rodriguez-Zas, S.L. & Robinson, G.E. Common and novel transcriptional routes to behavioral maturation in worker and male honey bees. *Genes Brain Behav.* **11**, 253–261 (2012).
32. Sella, G., Petrov, D.A., Przeworski, M. & Andolfatto, P. Pervasive natural selection in the *Drosophila* genome? *PLoS Genet.* **5**, e1000495 (2009).
33. Harpur, B.A. *et al.* Population genomics of the honey bee reveals strong signatures of positive selection on worker traits. *Proc. Natl. Acad. Sci. USA* **111**, 2614–2619 (2014).
34. Nässel, D.R. & Winther, A.M.E. *Drosophila* neuropeptides in regulation of physiology and behavior. *Prog. Neurobiol.* **92**, 42–104 (2010).
35. Woodard, S.H. *et al.* Genes involved in convergent evolution of eusociality in bees. *Proc. Natl. Acad. Sci. USA* **108**, 7472–7477 (2011).
36. Corona, M. *et al.* Vitellogenin, juvenile hormone, insulin signaling, and queen honey bee longevity. *Proc. Natl. Acad. Sci. USA* **104**, 7128–7133 (2007).
37. Amdam, G.V. *et al.* Higher vitellogenin concentrations in honey bee workers may be an adaptation to life in temperate climates. *Insectes Soc.* **52**, 316–319 (2005).
38. Fiumera, A.C., Dumont, B.L. & Clark, A.G. Associations between sperm competition and natural variation in male reproductive genes on the third chromosome of *Drosophila melanogaster*. *Genetics* **176**, 1245–1260 (2007).
39. Nakayama, M. *et al. Drosophila* carrying *Pex3* or *Pex16* mutations are models of Zellweger syndrome that reflect its symptoms associated with the absence of peroxisomes. *PLoS ONE* **6**, e22984 (2011).
40. Hernández-García, R., de la Rúa, P. & Serrano, J. Mating frequency in *Apis mellifera iberiensis* queens. *J. Apic. Res.* **48**, 121–125 (2009).
41. Franck, P., Koeniger, N., Lahner, G., Crewe, R.M. & Solignac, M. Evolution of extreme polyandry: an estimate of mating frequency in two African honeybee subspecies, *Apis mellifera monticola* and *A.m. scutellata*. *Insectes Soc.* **47**, 364–370 (2000).
42. DeGrandi-hoffman, G., Tarpy, D.R. & Schneider, S.S. Patriline composition of worker populations in honeybee (*Apis mellifera*) colonies headed by queens inseminated with semen from African and European drones. *Apidologie (Celle)* **34**, 111–120 (2003).
43. Fries, I. & Raina, S. American foulbrood and African honey bees (Hymenoptera: Apidae). *J. Econ. Entomol.* **96**, 1641–1646 (2003).
44. Galko, M.J. & Krasnow, M.A. Cellular and genetic analysis of wound healing in *Drosophila* larvae. *PLoS Biol.* **2**, E239 (2004).

ONLINE METHODS

Samples and DNA extraction. Samples of worker bees were collected from unrelated colonies (**Supplementary Note**). African samples were obtained from wild swarms now resident in apiaries. These comprised the Cape bee *A. mellifera capensis* ($n = 10$), *A. mellifera scutellata* ($n = 10$) from South Africa and *A. mellifera adansonii* ($n = 10$) from Nigeria (all from group A). European samples were collected from isolated apiaries that maintain pure subspecies. From the M group, these comprised *A. mellifera mellifera* from both Norway ($n = 10$) and Sweden ($n = 10$) and *A. mellifera iberiensis* from Spain ($n = 10$). From the C group, these comprised the Italian bee *A. mellifera ligustica* from Italy ($n = 10$) and the Carnolian bee *A. mellifera carnica* ($n = 10$) from Austria. From the O group, these comprised the Anatolian bee *A. mellifera anatoliaca* from Turkey ($n = 10$) and the Syrian bee *A. mellifera syriaca* from Jordan ($n = 10$). We also collected samples of North American domestic bees from Minnesota, USA ($n = 10$), and 2 samples of European domestic bees from Sweden (both $n = 10$) from apiaries that did not maintain specific subspecies. We included samples of the Asian bee *A. cerana* collected at various locations throughout Japan. We also included a drone from the DH4 line descended from the drone used to produce the original honeybee reference sequence[9].

We used a salt-ethanol precipitation protocol to extract high-quality DNA from the heads of individual *A. mellifera* worker bees. Each head was cut in half and put in preparation buffer (100 mM NaCl, 10 mM Tris-HCl, pH = 8.0, 0.5% SDS) together with proteinase K. Brain tissue was then dissolved by incubation at 50 °C for at least 4 h, after which time, the sample was frozen overnight (a freeze/thaw cycle was found to increase the final DNA yield). To precipitate DNA, we added saturated NaCl several times before adding 95% ethanol, and we spun the DNA into a pellet. The DNA pellet was suspended in TE buffer or double-distilled water. DNA concentration was determined using a NanoDrop ND-1000 spectrophotometer and an Invitrogen Qubit 2.0 fluorometer, and the fragment length distribution was assessed on an agarose gel. Clontech Chromaspin TE-400 or TE-1000 columns were used to remove pigments, salt, short DNA fragments and co-extracted RNA. The *A. cerana* samples were extracted using the Qiagen DNeasy Blood and Tissue kit and the manufacturer's recommended protocol. The average total amount of high-grade DNA in the extracted samples submitted for sequencing was 1.13 ± 0.53 µg.

Sequencing and mapping. We constructed barcoded fragment libraries for each individual sample according to the manufacturer's instructions (Life Technologies). The 151 libraries were then sequenced on a SOLiD 5500xl machine to produce 75-bp reads (**Supplementary Note**). We used 30 samples per flow chip, which were pooled and divided across available lanes, with the exception of the DH4 drone sample, which was run independently on a single lane. We next chose one sample from each group for sequencing at higher coverage. These samples comprised *A. mellifera mellifera*, *A. mellifera carnica*, *A. mellifera anatoliaca* and *A. mellifera scutellata*. These libraries were converted to WildFire libraries according to the manufacturer's instructions and were sequenced on a SOLiD 5500 WildFire instrument, also producing 75-bp reads across three lanes. Mapping of reads in color space to the Amel_4.5 reference was performed using LifeScope v2.5 or 2.5.1 (Life Technologies) with default settings.

Quality control and variant calling. We performed several steps to improve mapping quality (**Supplementary Note**). We first identified and marked PCR duplicates using Picard. We next identified regions of poorly and inconsistently mapped reads by realigning around indels. This was done with the Genome Analysis Toolkit (GATK)[45]. We performed Bayesian population-based SNP calling using FreeBayes across all *A. mellifera* samples. Biallelic SNPs with a quality score of 50 or greater were retained for further analysis. Several additional filters were then used to reduce the number of false positive SNPs (**Supplementary Table 1**). These filters were based on an abnormally low number of reads mapping to the SNP, an abnormally high read depth in a 100-bp window around the SNP, low genotype quality (average GQ < 20), a high number of repeat elements close to the SNP and a low number of samples with mapped reads in the region. Sites at which heterozygous calls were obtained in the haploid drone also exhibited an excess proportion of heterozygous calls in worker genotypes. These sites likely represent erroneous calls due to incorrect mapping, for example, in duplicated regions of the genome, and were removed from the analysis. After obtaining a high-quality set of genotypes from SNP calling (**Supplementary Table 2**), we conducted haplotype inference and imputation of missing genotypes using BEAGLE[46]. The mapped *A. cerana* sequences were used to infer the ancestral state of SNPs in cases where the most common *A. cerana* allele matched one of the *A. mellifera* alleles.

Analysis of genetic diversity and divergence. We measured raw genetic distance on the basis of the number of shared alleles between each individual sample from native subspecies and used a distance matrix to construct a tree using the neighbor-joining method[47] implemented in PHYLIP (**Supplementary Note**). We also estimated F_{ST} values between populations and used this distance matrix to construct a neighbor-joining tree using the same approach. A principal-component analysis was performed using multidimensional scaling in PLINK. We calculated Watterson's estimator (θ_w) of the population mutation rate (θ) per base using the number of SNPs segregating in each population[48]. We estimated N_e in each sample from our estimates of θ_w and an estimate of the mutation rate derived from divergence with *A. cerana*[49]. For a haplodiploid system, $\theta = 3N_e\mu$, where N_e is the effective population size and μ is the mutation rate per base. Two outlier samples from the M group were excluded from these calculations owing to potential hybridization (**Fig. 1b**). We used six publicly available sequences (~1 kb each) from noncoding regions of the *A. cerana* genome to estimate the level of divergence between the *A. mellifera* reference sequence representing a single copy of six different regions sequenced by ref. 49. We used these sequences because the more diverged 75-bp *A. cerana* reads produced by our sequencing cannot be unambiguously mapped to locations in the *A. mellifera* genome. This results in mapped reads being biased toward those with low divergence, which underestimates true levels of divergence between the two species. We identified sequences orthologous to each *A. cerana* sequence using a nucleotide BLAST search of the honeybee genome. The query and target sequences were then aligned using MAFFT[50]. Raw divergence was used to estimate the mutation rate per generation on the basis of a divergence time of 7 million years. We used this value to estimate N_e in each population from levels of variation.

Demographic analysis. We first ran ADMIXTURE[51] on the entire data set to estimate the genetic ancestry of each sample, specifying a range of 2–6 hypothetical ancestral populations (**Supplementary Note**). This tool is a fast implementation of an algorithm similar to STRUCTURE[52] that is suitable for large data sets. The analysis provided maximum-likelihood estimates of the proportion of each sequenced genome that was derived from each of K populations using a variety of values of K. We also ran STRUCTURE, and the results were not qualitatively different. We performed an analysis using ChromoPainter and fineSTRUCTURE[53], which uses a haplotype-based approach to estimate the ancestries of blocks of DNA across the genome of each sample, with these ancestries then summarized as a co-ancestry matrix that describes the ancestral relationships among samples. SNPs were treated as unlinked, and all individuals were compared against all other individuals to infer haplotype transmission. We used TreeMix[54] to infer the history of population splits and mixtures, allowing up to eight mixture events. This method constructs a bifurcating tree of populations and then identifies potential episodes of gene flow from the residual covariance matrix. We used genealogical concordance to estimate the times at which population splits occurred[23]. After a population splits in two, the number of loci with genealogies that match the population genealogy increases with time, whereas the number of discordant genealogies that support alternative population trees decreases. We calculated the time since the populations split (t) in units of $1.5N_e$ generations as $-\log((3 - 3P_C)/2)$, where P_C was the proportion of concordant loci (for haplodiploids). To convert the estimates into time, we used the N_e values estimated for each of the four main honeybee groups or from the average of two groups when dating splits between two groups (see the **Supplementary Note** for details). We used PSMC[24] to estimate variation in N_e over historical time. This method uses a hidden Markov model to partition a diploid genome into blocks with varying ancestries, which can be used to estimate the distribution of time to the most recent common ancestor (TMRCA) across the genome. We estimated the magnitude of piecewise constant ancestral N_e over 90 time intervals. We ran a PSMC analysis on one sample from each of the four ancestral groups sequenced at higher coverage—*A. mellifera mellifera* (M), *A. mellifera carnica* (C),

A. mellifera anatoliaca (O) and *A. mellifera scutellata* (A). SNPs were called on these samples using the standard pipeline recommended in ref. 24. SNPs designated as being of poor quality on the basis of the filters in the analysis of 140 samples were also filtered from this high-coverage data set. Thirty iterations of the algorithm were completed for each sample. A generation time of 1 year was used to convert estimates into time in years.

Analysis of methylation and gene expression. We measured $CpG_{O/E}$, which is a measure of the depletion of CpG dinucleotides, using the full predicted transcript sequences for all genes in Amel_4.5 (ref. 26; **Supplementary Note**). $CpG_{O/E}$ is the number of CpG dinucleotides normalized by GC content. It is lower in highly methylated regions owing to the elevated mutability of methylated CpG dinucleotides. We obtained two gene expression data sets based, respectively, on comparison of queens and sterile workers[30] and drones and workers[31]. These data sets were described in ref. 26. We identified 1,700 and 6,500 genes in each data set by mapping accession IDs and probe sequences to the Amel_4.5 genome sequences. These data sets were divided into those with increased expression in one caste and those with unbiased expression on the basis of original definitions. We analyzed the relationship between expression categories and $CpG_{O/E}$ distribution. Finding a clear bimodal distribution, we divided genes into categories with high and low $CpG_{O/E}$ on the basis of whether the $CpG_{O/E}$ was higher or lower than the mean (1.19).

We measured average genetic variation in different genomic categories defined by genome annotation using Watterson's estimator and estimated confidence intervals using bootstrapping. We next estimated genetic variation in the protein-coding regions of genes defined by expression and CpG categories. We corrected these measures using divergence with *A. cerana* reads. Levels of divergence with *A. cerana* reads were not estimated for noncoding regions because of poorer mapping of *A. cerana* reads. We also analyzed the levels of genetic variation in 1-kb windows at increasing distance from genes divided according to $CpG_{O/E}$ category.

Selection analysis. We estimated F_{ST} per SNP on the basis of three separate comparisons based on a standard formula (**Supplementary Note**). We omitted the O group owing to evidence for admixture with the A group. For each comparison, we pooled sequences from two groups to calculate F_{ST} with the remaining group (that is, AM versus C, AC versus M and MC versus A). We divided SNPs by F_{ST} interval and calculated the proportion in each annotation category (coding, noncoding, intronic and UTR). We also estimated the enrichment of CpG and expression categories as well as the levels of heterozygosity in windows (1 kb) of increasing distance from SNPs divided by F_{ST} category. We also estimated the proportion of SNPs in each F_{ST} category that were derived on the basis of the ancestral state as inferred from *A. cerana* reads.

Gene Ontology analysis. To identify homologs for the honeybee gene set, we used the honeybee gene set in BLAST analysis to query the *Drosophila* genome. This analysis was performed using the honeybee coding sequence from Amel_4.5 as the query and performing stand-alone BLASTX against *Drosophila* peptide sequences obtained from BioMart. We used the best *Drosophila* hit to determine the GO category for each honeybee gene. We used the REVIGO tool to produce summaries of nonredundant GO terms grouped into functional categories[55].

45. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
46. Browning, S.R. & Browning, B.L. 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).
47. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).
48. Watterson, G.A. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–276 (1975).
49. Cho, S., Huang, Z.Y., Green, D.R., Smith, D.R. & Zhang, J. Evolution of the complementary sex-determination gene of honey bees: balancing selection and trans-species polymorphisms. *Genome Res.* **16**, 1366–1375 (2006).
50. Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**, 511–518 (2005).
51. Alexander, D.H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
52. Pritchard, J.K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
53. Lawson, D.J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet.* **8**, e1002453 (2012).
54. Pickrell, J.K. & Pritchard, J.K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8**, e1002967 (2012).
55. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. REVIGO summarizes and visualizes long lists of Gene Ontology terms. *PLoS ONE* **6**, e21800 (2011).