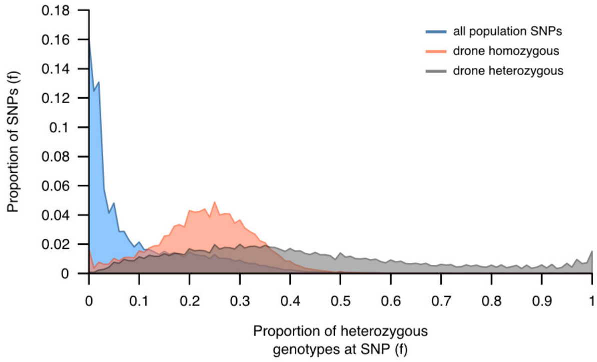
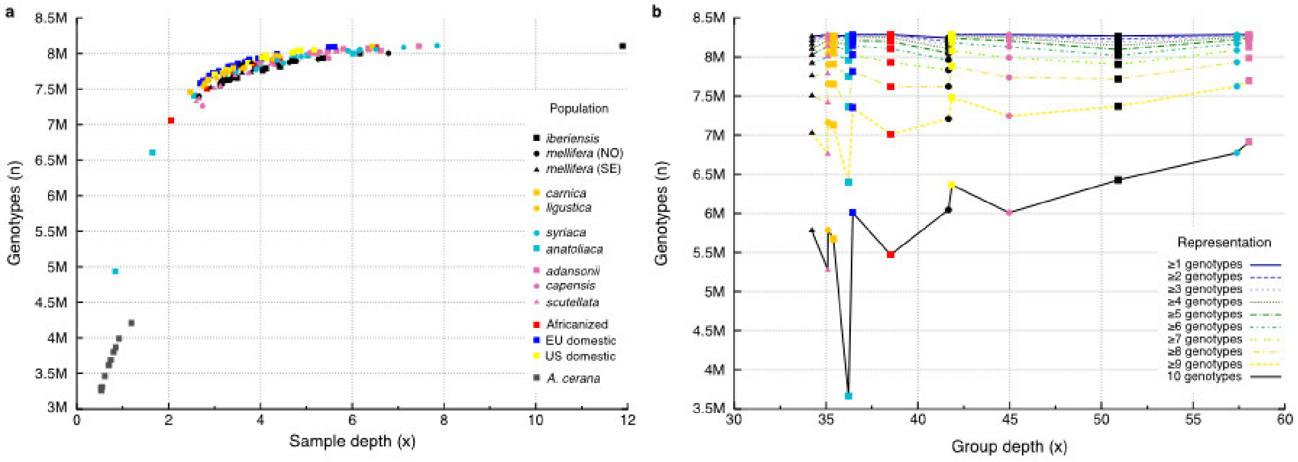


Supplementary Figure 1

Diploid genotype quality control using a haploid drone sequenced to 20× depth of coverage.

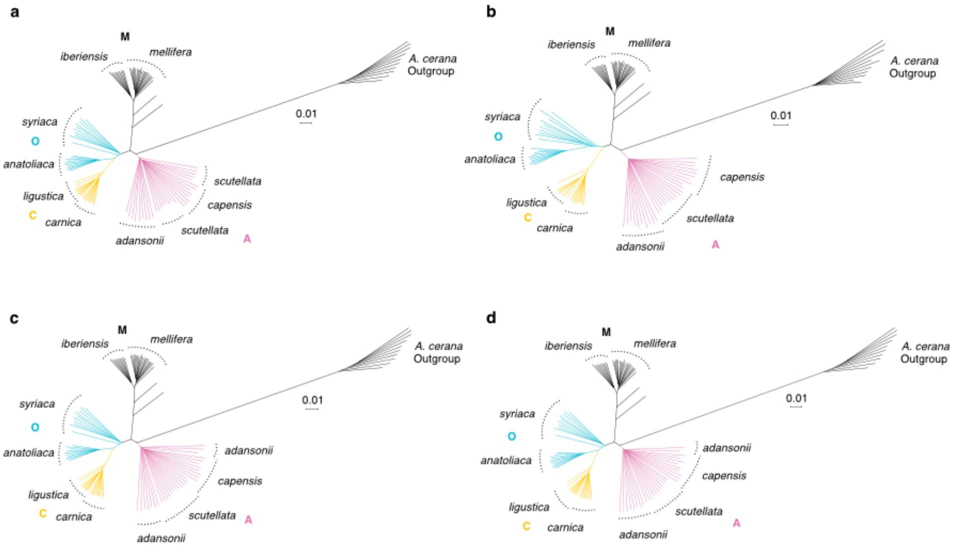
Haploid drone data were intentionally misspecified to be diploid in the FreeBayes SNP calling process, and variable positions shared by the drone and the population data set were compared. This figure shows the distributions of heterozygous genotypes in the population data set conditional on their genotype in the resequenced drone (blue, all population data set SNPs; pink, SNPs where the drone had a homozygous difference compared to the reference genome; gray, SNPs where the drone had a heterozygous genotype). Chromosomal positions with heterozygous drone genotypes demonstrated extremely elevated levels of heterozygous genotypes in the population data set (gray) compared to the L-shaped distribution of all population SNPs (blue) and the positions where the drone had homozygous differences relative to the reference sequence (red). SNPs overlapping the heterozygous drone positions were therefore removed from the population data set, as they are unlikely to represent true SNPs.



Supplementary Figure 2

Genotype representation.

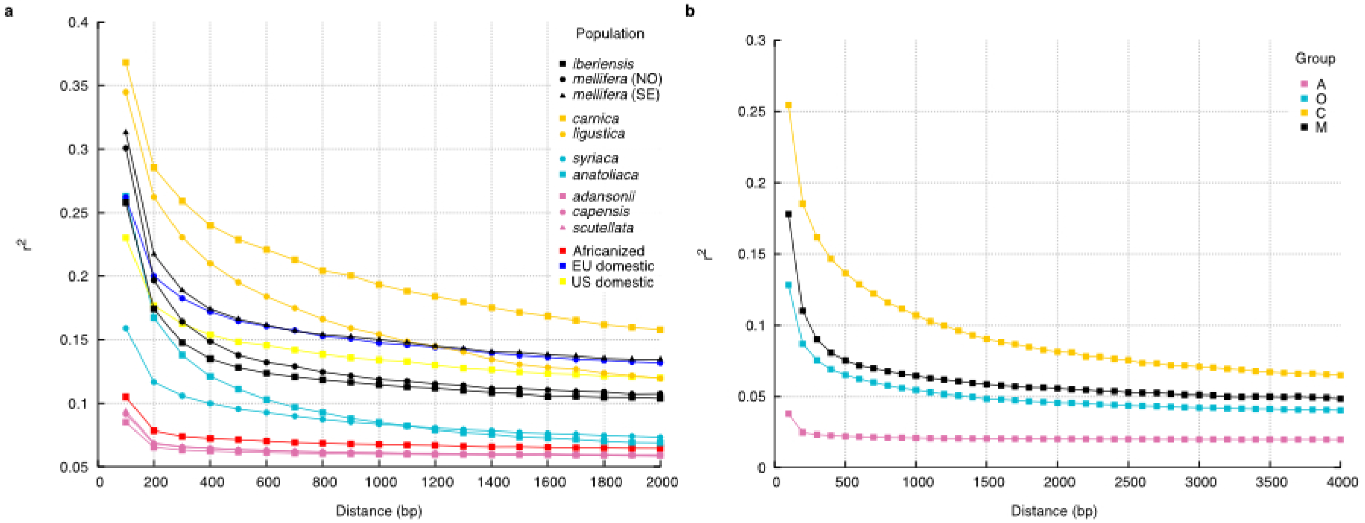
(**a**) The number of genotypes called per sample (*y* axis) compared to the average depth of coverage of the sample (*x* axis). The number of SNP positions at which a sample had its genotype inferred varied between 5–8 million and is correlated with its depth of coverage (i.e., poorly represented samples had more unknown genotypes before alleles were imputed with BEAGLE). (**b**) The number of SNPs where each population had one or more genotypes called out of all ten samples (*y* axis) compared to the average depth of coverage for the population (*x* axis). Populations sequenced at lower depth generally had more missing genotypes than those sequenced at higher depth. The missing genotypes appeared to be partly randomly distributed among samples and were smoothened at the population level. Of the 8.3 million SNPs in the data set, all populations (see legend in **a**) are represented by at least 5 genotypes at 8 million SNPs and by 8 genotypes at 7 million SNPs. The *A. mellifera anatolica* population had two samples sequenced at low coverage (A), resulting in the lowest genotype representation at the population level. In total, about 6% of the genotypes in the SNP matrix were missing and were imputed with BEAGLE.



Supplementary Figure 3

Neighbor-joining trees based on allele sharing distances.

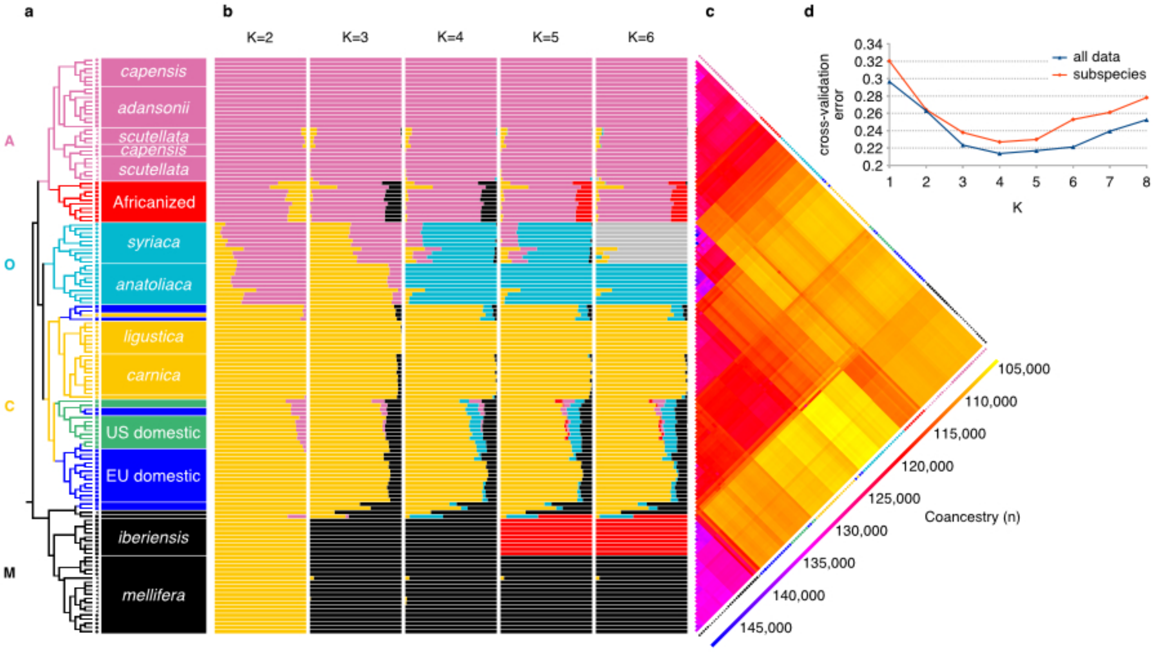
Trees were constructed using different subsets of the data using samples from native subspecies. (**a**) All data. (**b**) Coding sequences. (**c**) Genic sequences (introns, exons and UTRs). (**d**) Intergenic sequences. All trees show consistent topologies.



Supplementary Figure 4

Linkage disequilibrium decays rapidly in honeybees.

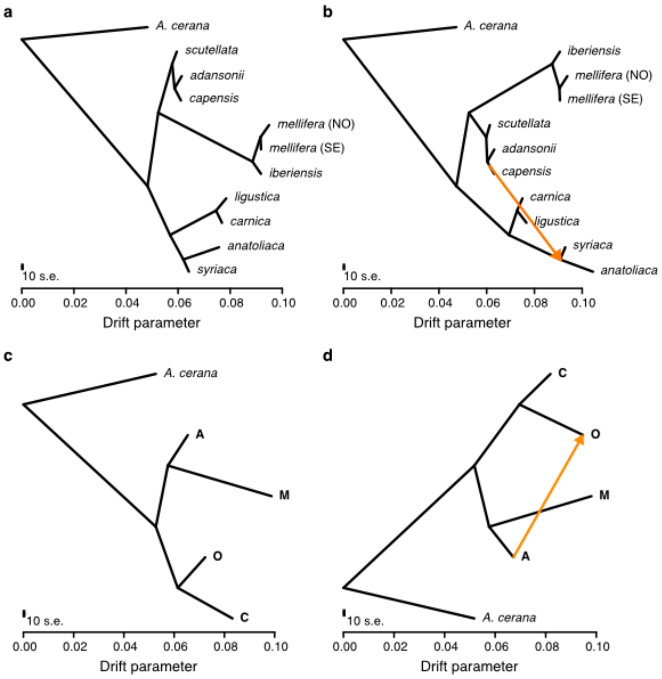
The linkage disequilibrium (LD) between SNPs as measured using the r^2 estimator (*y* axes) is shown over increasing distance between the SNPs (*x* axes). (**a**) LD was found to decay quickly across the natural, domestic and Africanized populations. All populations have an r^2 value of <0.2 at 1 kb. The rate of decay of LD reflected population size and was most rapid in the African populations. The highest LD was found among European bees. The domestic lineages had an intermediate decay of LD. (**b**) The corresponding decay of LD after grouping the samples according to major continental groups—A (*n* = 30), C (*n* = 20), M (*n* = 30) and O (*n* = 20)—shows that r^2 is below 0.1 in all groups at distances of greater than 1,500 bp.



Supplementary Figure 5

Model-based clustering and ancestral inferences.

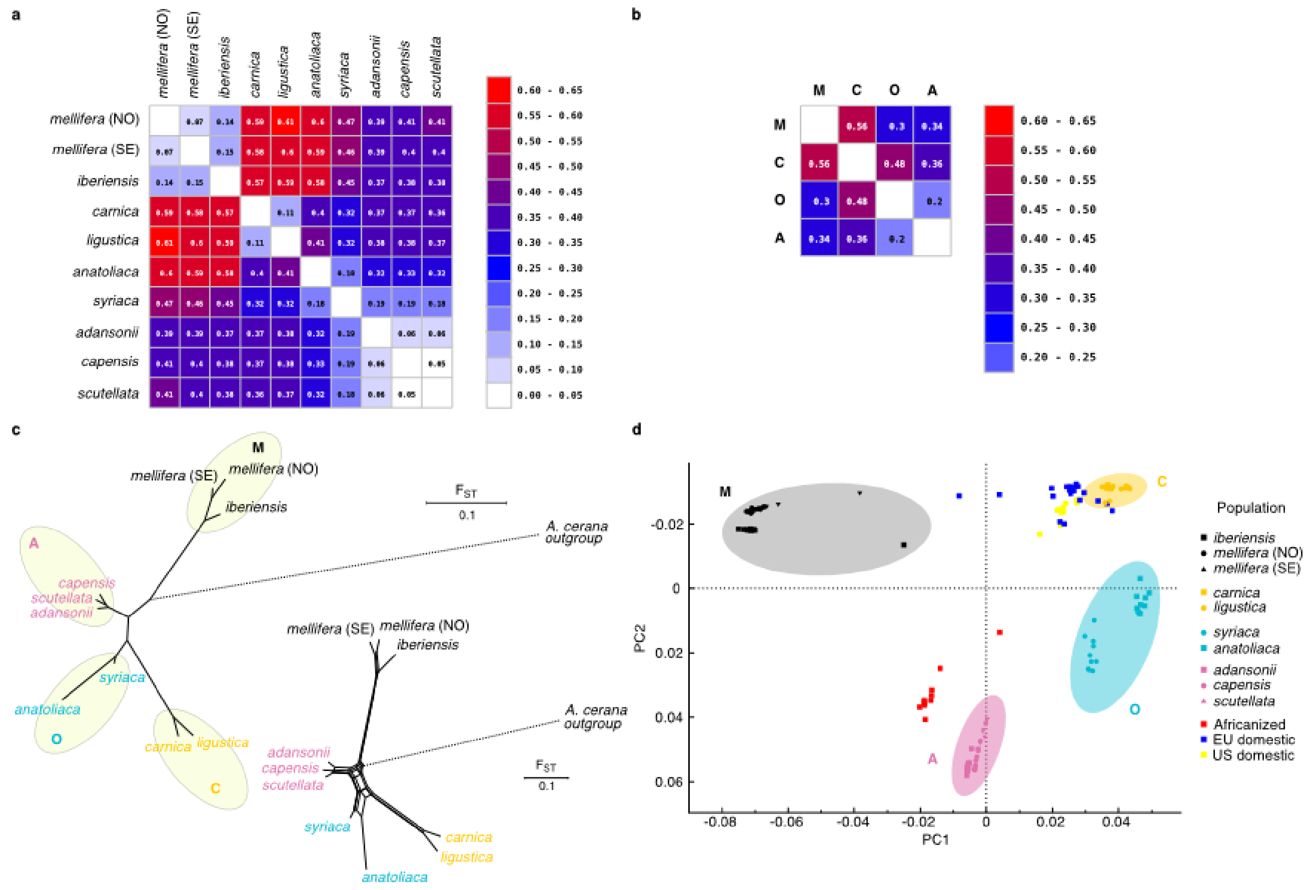
(**a**) A tree representation of the genetic relationships between the individual honeybee samples. fineSTRUCTURE was used to infer interrelationships among the samples, recovering a topology where most populations are identified as monophyletic clades (with a few outliers) and structured according to the four continental groups. South African *A. mellifera capensis* samples were recovered inside the *A. mellifera scutellata* population. The domestic bees clustered mainly with the bees from central Europe (C group). Africanized bees from Brazil were grouped with African bees (A group). (**b**) Allelic ancestry of each sampled honeybee. ADMIXTURE was used to infer allelic origins, and the samples were sorted according to fineSTRUCTURE topology. At *K* = 4, ADMIXTURE subdivides samples according to the four major groups inferred with fineSTRUCTURE. The domestic bees have mainly C-type alleles but also show some proportion of alleles with other origins. The Africanized bees appear to have mainly African alleles but also carry a remainder proportion of alleles originating from western Europe (M group). ADMIXTURE detects introgression of African alleles into the *A. mellifera syriaca* population. About 18% of *A. mellifera syriaca* alleles appear to be of African origin. (**c**) The co-ancestry matrix estimated in fineSTRUCTURE shows high levels of co-ancestry within the populations and groups and also detects introgression of African alleles into the *A. mellifera syriaca* population. (**d**) The cross-validation error according to the number of hypothetical ancestral groups (*K*). The cross-validation (CV) procedure implemented in ADMIXTURE was applied to all data (140 samples, including subspecies, domestic and Africanized samples; blue) and a data set including only the natural subspecies (100 samples; red). In both cases, the optimal value for *K* was found to be 4 (lowest CV errors), suggesting that the variation is best summarized as being subdivided according to the four ancestral groups.



Supplementary Figure 6

TreeMix analysis.

The relationships, divergence and major mixtures between populations were inferred and illustrated as trees using TreeMix. (**a**) The TreeMix ML tree explains 98.2% of the variation. It identifies the four major groups and places the root between the A + M and C + O clusters. (**b**) The most important mixing event (arrow) based on the residuals involves the migration of African alleles into the O group and explains an additional 0.6% of the variation. (**c**,**d**) The inferred interrelationships and major mixing event (arrow) after pooling the samples according to the continental groups are consistent with the population tree.



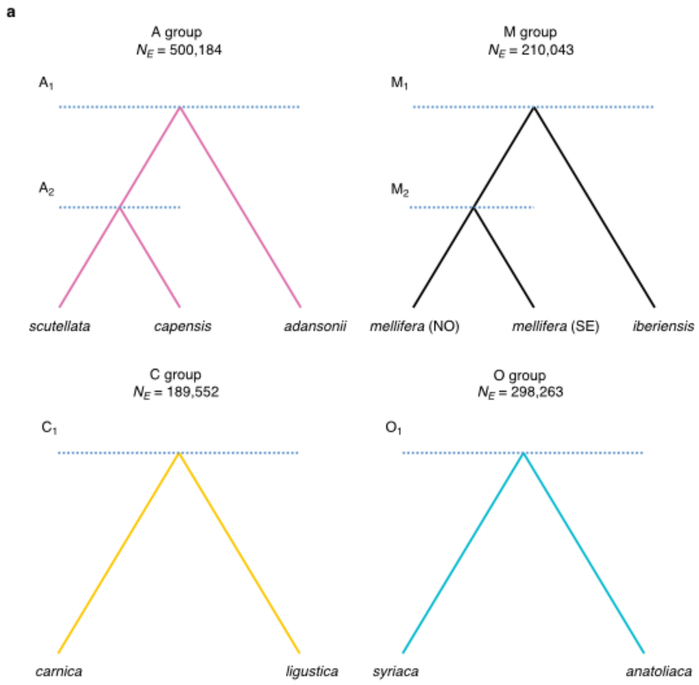
	mellifera (NO)	mellifera (SE)	iberiensis	carnica	ligustica	anatoliaca	syriaca	adansonii	capensis	scutellata
mellifera (NO)		0.07	0.14	0.59	0.61	0.6	0.47	0.39	0.41	0.41
mellifera (SE)	0.07		0.15	0.58	0.6	0.59	0.46	0.39	0.4	0.4
iberiensis	0.14	0.15		0.57	0.59	0.58	0.45	0.37	0.38	0.38
carnica	0.59	0.58	0.57		0.11	0.4	0.32	0.37	0.37	0.36
ligustica	0.61	0.6	0.59	0.11		0.41	0.32	0.38	0.38	0.37
anatoliaca	0.6	0.59	0.58	0.4	0.41		0.18	0.32	0.33	0.32
syriaca	0.47	0.46	0.45	0.32	0.32	0.18		0.19	0.19	0.18
adansonii	0.39	0.39	0.37	0.37	0.38	0.32	0.19		0.06	0.06
capensis	0.41	0.4	0.38	0.37	0.38	0.33	0.19	0.06		0.05
scutellata	0.41	0.4	0.38	0.36	0.37	0.32	0.18	0.06	0.05	

	M	C	O	A
M		0.56	0.3	0.34
C	0.56		0.48	0.36
O	0.3	0.48		0.2
A	0.34	0.36	0.2	

Supplementary Figure 7

***F_{ST}* and MDS analyses of population divergence.**

(**a**) The fixation index (F_{ST}) matrix reconstructed from estimates of pairwise F_{ST} between all natural populations. There is evidence of large amounts of drift separating the populations of western and northern Europe (*A. mellifera mellifera* and *A. mellifera iberiensis*; M group) from those of central/southeastern Europe (*A. mellifera carnica* and *A. mellifera ligustica*; C group) and the introgression of African alleles into the *A. mellifera syriaca* population. (**b**) The corresponding F_{ST} matrix of pairwise distances between samples when aggregated according to the major groups is similar to that among populations. (**c**) Upper left, neighbor-joining (NJ) tree reconstructed from the pairwise F_{ST} distance matrix shows that the outgroup connects to the middle of the tree rather than within any of the groups. Lower right, the corresponding splits network reconstructed using SplitsTree indicates that the tree-like resolution is lowest at the basal positions among the groups. (**d**) The principal-component analysis (PCA) plot of the first two components (PC1 and PC2) as recovered by multidimensional scaling (MDS) in PLINK further supports the four major groups and the dominance of C-group alleles in domestic bees. The Africanized hybrid bees appear most similar to the native African samples.



b

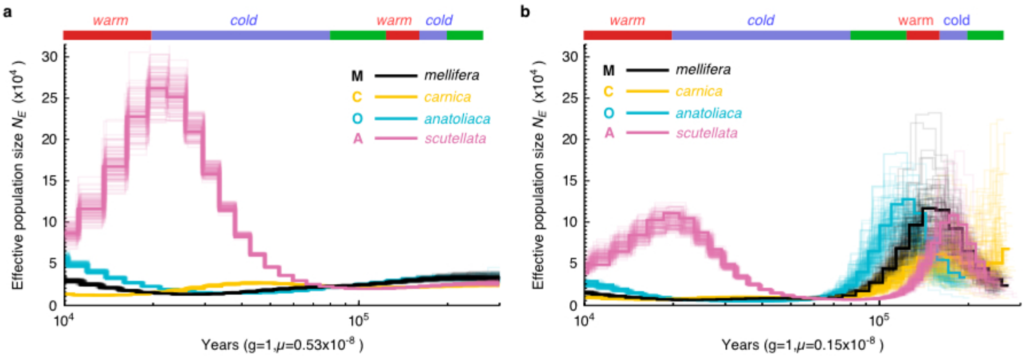
Comparison		N_E	T	YBP	SE
within	A_1	500,184	0.04	32,709	2,251
	A_2	500,184	0.03	23,187	2,251
	C_1	189,552	0.09	24,633	1,166
	M_1	210,043	0.12	37,845	1,229
	M_2	210,043	0.04	13,069	1,166
	O_1	298,263	0.18	80,646	1,611
between	A-C	344,868	0.59	303,750	2,324
	A-M	355,114	0.62	329,695	2,423
	A-O*	345,801	0.49	252,586	1,743
	C-M	199,798	0.98	294,623	2,740
	C-O*	190,485	0.58	164,744	1,334
	M-O*	200,731	1.12	337,553	3,342

*O-group included only *anatoliaca*

Supplementary Figure 8

Genealogical concordance analyses.

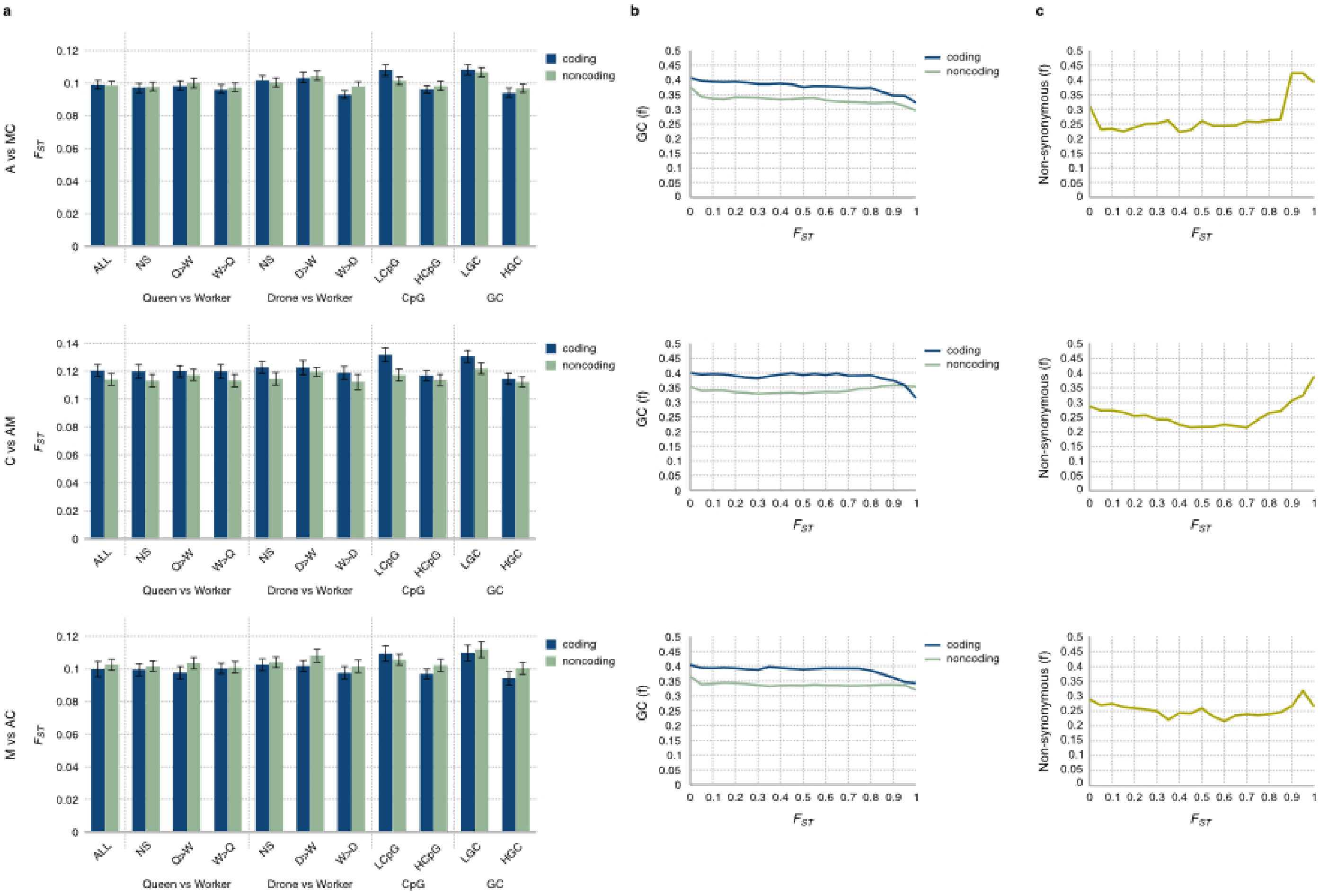
(**a**) Population split events (blue dashed lines) in each group (A, M, C, O) as represented by the sampled honeybees were assumed to represent the origin of each population. (**b**) Population ages estimated by genealogical concordance (within, splits within the major groups, as shown in **a**; between, splits between groups). The proportion of concordant sites was computed from whole-genome patterns of allele sharing among samples and used to estimate the divergence times *T* in units of 1.5N_e. The effective population size was computed from the levels of sequence variation in each population. The mutation rate based on the divergence between *A. mellifera* and *A. cerana* (5.27 × 10^{-9}) was used to transform ages into units of time, assuming a generation turnover of 1 year. Modern subspecies diverged around 20,000 YBP, whereas the major continental groups appear to have diverged around 200,000–300,00 YBP. The method is sensitive to external gene flow, as shown by the introgression of African alleles into the *A. mellifera syriaca* population, which overestimates the age of its split from *A. mellifera anatoliaca*. Because of admixture, the *A. mellifera syriaca* population was not included when estimating the ages of the splits between the O group and the other three groups.



Supplementary Figure 9

PSMC analysis.

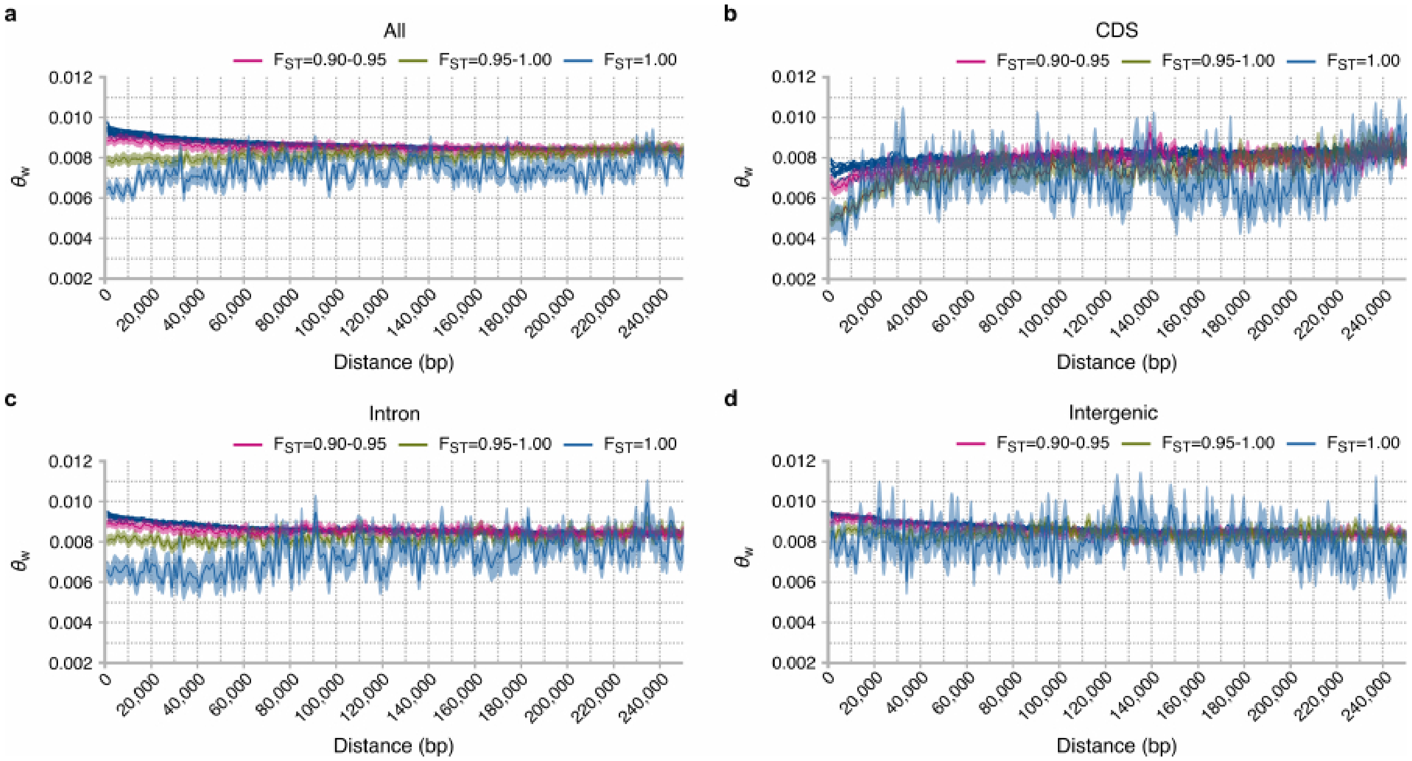
The historical effective population size (N_e) (y axes) was traced back in time (x axes) using the Pairwise Sequentially Markovian Coalescent (PSMC) model applied to four samples sequenced at ~20× depth of coverage. PSMC reconstructs the variation of historical N_e values from patterns of heterozygosity along the chromosomes. (**a**) Analysis of the complete data set. Generation turnover time was specified to be 1 year and the mutation rate was estimated to be 0.53×10^{-8} mutations per base and generation, computed from divergence observed with the outgroup species *A. cerana* (confidence intervals computed from 125 bootstrap replicates). (**b**) Analysis of only transversions. Old coalescent events and historical population sizes were reconstructed with greater resolution when the genotype data were thinned to include only transversions, suggesting alternating patterns of population expansion and contraction that match the last two glaciation cycles (errors bars are computed as in **a**; the transversion mutation rate was estimated to be 0.15×10^{-8}).



Supplementary Figure 10

Genome-wide patterns of population differentiation.

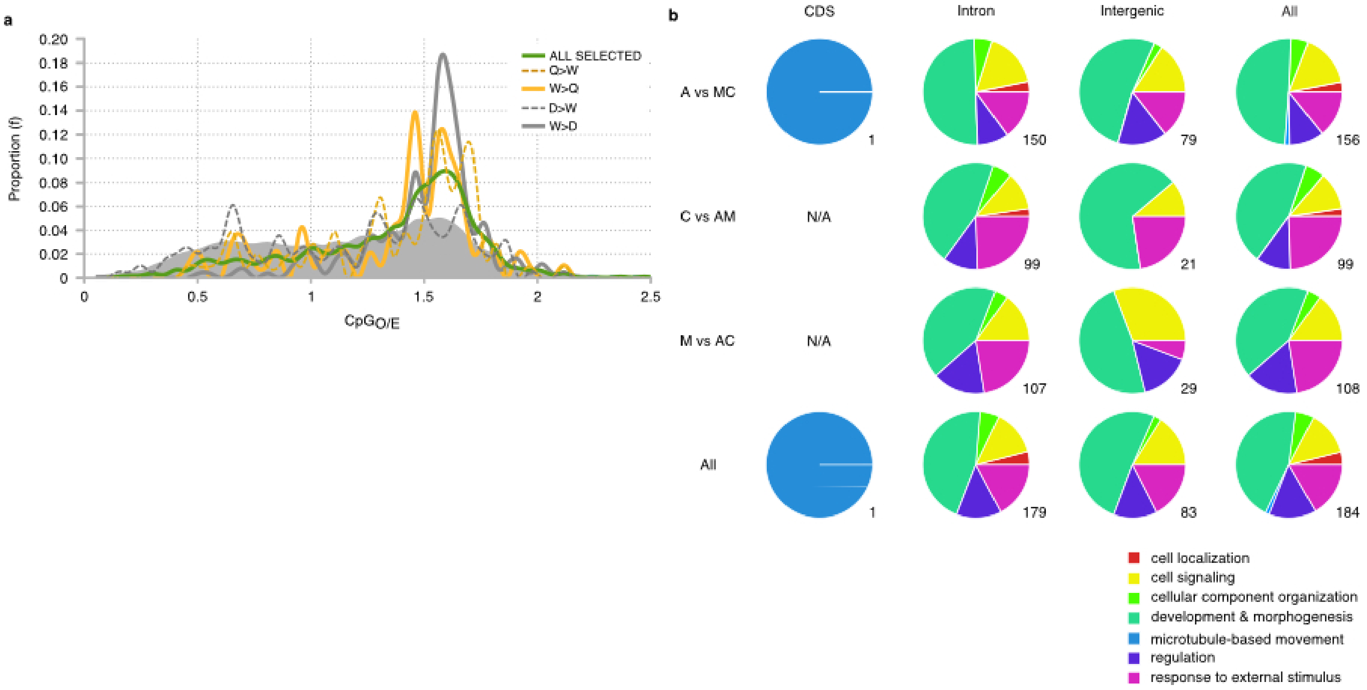
(**a**) The mean fixation index (F_{ST}) of coding and noncoding SNPs detected in three comparisons (top, A versus MC; middle, C versus AM; bottom, M versus AC; A, Africa; M, northern/western Europe; C, southern/eastern Europe) as estimated for genes of different caste expression and $\mathrm{CpG}_{\mathrm{O/E}}$ categories. The F_{ST} of SNPs in coding regions (CDS + UTRs) is not significantly higher than the level of differentiation detected in noncoding (intron + intergenic) regions. There are no signs of genome-wide signals of positive selection as detected by F_{ST} on coding regions across all genes in the genome or among genes with caste-biased overexpression. The coding regions of low-GC/CpG genes have a small (~12%) increase in mean F_{ST} compared to high-GC/CpG genes, a signal that is less clear in noncoding regions (NS, genes with unbiased expression; Q, queen-biased genes; W, worker-biased genes; D, drone-biased genes; LCpG, low-CpG genes; HCpG, high-CpG genes; LGC, low-GC genes; HGC, high-GC genes; 95% confidence intervals were estimated by bootstrap). (**b**) GC content in a 100-bp window around each SNP (*y* axes) is weakly correlated with F_{ST} (*x* axes; $R^2 < 2\%$) and mainly detected as a 10–20% drop in the extreme tail along the F_{ST} distribution for SNPs in coding regions. (**c**) Substitutions in coding regions were analyzed for codon changes (proportion of nonsynonymous SNPs; *y* axes) across the F_{ST} spectrum (*x* axes). High-F_{ST} substitutions ($F_{\mathrm{ST}} \geq 0.9$; $f_{\mathrm{non}} = 0.42$) were found to be significantly enriched for nonsynonymous changes between African and European bees (A versus MC) compared to low-F_{ST} substitutions ($F_{\mathrm{ST}} < 0.9$; $f_{\mathrm{non}} = 0.28$) and in the C versus AM comparison ($F_{\mathrm{ST}} \geq 0.9$; $f_{\mathrm{non}} = 0.31$ versus $F_{\mathrm{ST}} < 0.9$; $f_{\mathrm{non}} = 0.26$) but not the M versus AC comparison ($F_{\mathrm{ST}} \geq 0.9$; $f_{\mathrm{non}} = 0.28$ versus $F_{\mathrm{ST}} < 0.9$; $f_{\mathrm{non}} = 0.28$) by bootstrapping the SNPs and assessing the variation around the mean at the 5% level. The large enrichment of highly differentiated nonsynonymous SNPs suggests that protein sequences are under divergent positive selection in African and European bees.



Supplementary Figure 11

Levels of nucleotide diversity around SNPs conditional on F_{ST}.

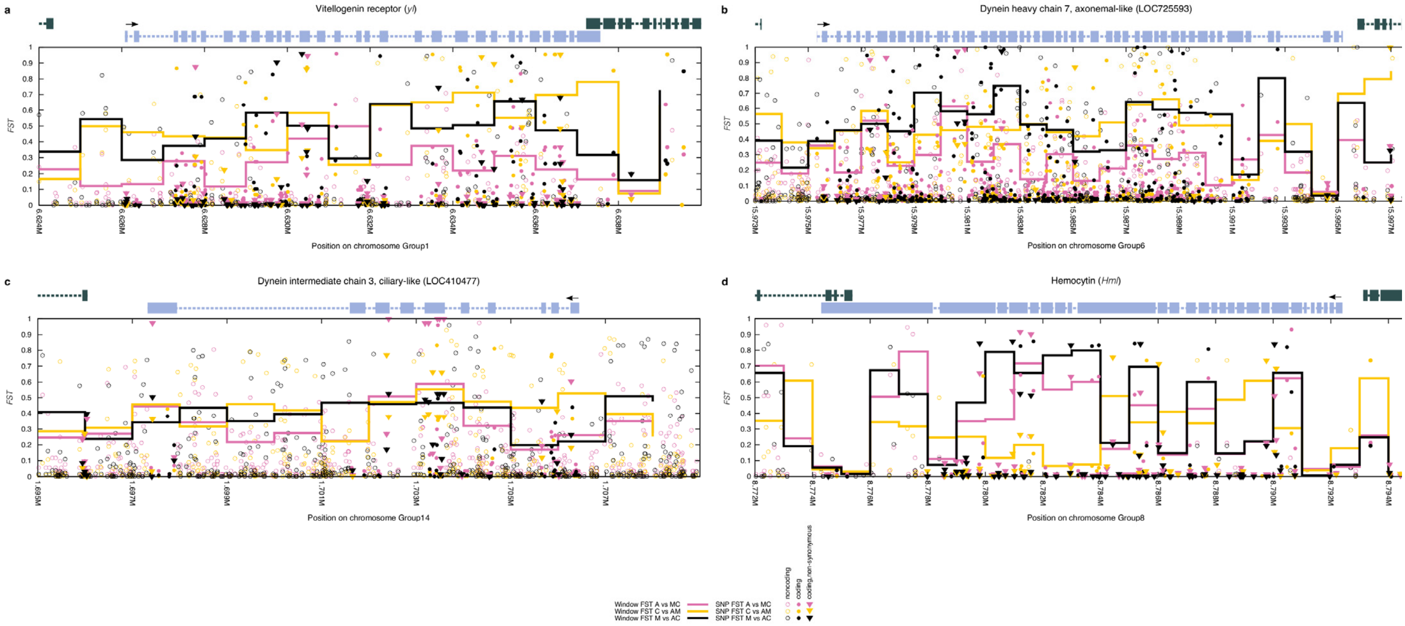
Neutral diversity (Watterson's theta, θ_W) in the African bees was measured in 1-kb windows up to a distance of 250 kb on either side of every SNP detected between African and European bees and grouped into F_{ST} intervals (size = 0.05; the last interval contains the fixed SNPs). The mean diversity close to a SNP of a particular F_{ST} class (1–20 kb away) was compared to the diversity of all SNPs (at 100–250 kb away, matching the average diversity computed across the whole genome). (**a**) All SNPs. Diversity is reduced by ~8% around SNPs with F_{ST} = 0.95–1.00 and by ~23% around fixed SNPs with F_{ST} = 1 (pink, F_{ST} = 0.9–0.95; green, F_{ST} = 0.95–1.00; blue, F_{ST} = 1; 95% bootstrap confidence intervals are shown; average diversities computed for all lower F_{ST} classes shown in dark blue). (**b**) Coding SNPs. SNPs in coding regions are associated with the greatest reduction in diversity ($F_{ST(1)}$ = 31%; $F_{ST(0.95–1.00)}$ = 32%; $F_{ST(0.9–0.95)}$ = 16%; $F_{ST(0.85–0.90)}$ = 15%). (**c**) Intronic SNPs. It is mainly the fixed SNPs in introns that show reduced levels of variation (23%). (**d**) Intergenic SNPs. These SNPs are only weakly associated with reduced variation ($F_{ST(1)}$ = 6%).



Supplementary Figure 12

Properties of highly differentiated SNPs.

The highly differentiated (top 0.1%) SNPs in the 3 pairwise comparisons between the African and European groups (6,179 SNPs with F_{ST} ≥ 0.89 in A versus MC; 4,799 SNPs with F_{ST} ≥ 0.95 in C versus AM; 6,283 SNPs with F_{ST} ≥ 0.96 in M versus AC) were subdivided according to honeybee genes, gene regions and caste-biased expression. (**a**) The CpG$_{O/E}$ levels of all putatively selected genes (green) associated with the highly differentiated SNPs between African and European bees are elevated compared to all genes (gray background). Likewise, the subsets of potentially selected caste-biased genes converge on high CpG$_{O/E}$ levels relative to all genes of the caste (**Fig. 4b**) (Q, queen; W, worker; D, drone), suggesting that unmethylated or differentially methylated genes are important targets of selection and adaptation in honeybees. (**b**) The functional associations of the *Drosophila* orthologs of the honeybee genes were analyzed with g:Profiler for functional enrichments. Across all contrasts and the 3 gene regions, 184 biological process (BP) terms were found to be significantly enriched in the SNP data sets, mostly driven by strong enrichments in intronic SNPs. The BP-associated noncoding regions were clustered into major categories using the REVIGO web service, spanning cellular development and signaling, regulation, response to stimulus and morphogenic development. SNPs in coding regions were significantly enriched for microtubule-based movement, detecting the large number of sperm motor proteins undergoing selection between African and European bees.



Supplementary Figure 13

Examples of genes with highly differentiated SNPs.

About 2,000–3,000 genes were associated with the top 0.1% of the most differentiated SNPs across three different comparisons (A versus MC, C versus AM, M versus AC; A, African; M, northern/western Europe; C, southern/eastern Europe; see **Fig. 4a** for details), including genes involved in the social, reproductive and immune systems. (**a**) Model of the vitellogenin receptor (*yl*) gene (light blue above graph; exons are shown as blocks; introns are shown as dashed lines; the arrow indicates the strand; neighboring genes shown in gray) involved in honeybee worker labor division and behavior. The allelic differentiations of SNPs among bees from different groups were estimated using the F_{ST} fixation index (*y* axis; circles, noncoding SNPs; filled circles, coding SNPs; triangles, nonsynonymous coding SNPs; F_{ST} measured in 1-kb windows shown as lines) and are illustrated at their coordinates along the gene (*x* axis). All honeybee groups (A, C, M) are associated with nonsynonymous SNPs at $F_{ST} > 0.9$, but the exons along the last 4 kb of the gene in C-group bees are the most differentiated. (**b**) Model of the microtubular dynein heavy chain 7 sperm motor protein gene (symbols as in **a**) showing highly differentiated SNPs and regions, found mainly in the A and M groups. The African bees have nonsynonymous SNPs distributed across several exons. (**c**) Model of the microtubular dynein intermediate chain 3 sperm motor gene (symbols as in **a**) showing a large cluster of nonsynonymous SNPs in an exon separating African and European bees. (**d**) Model of the immune gene hemocytin/hemolectin (*Hml*) showing several highly differentiated SNPs ($F_{ST} > 0.9$) in the African bees and a haplotype-like signature in M-group bees at lower fixation.

Supplementary Tables

Supplementary Table 1. SNP filtering. A base filter accepting biallelic SNPs with a QUAL score of ≥50 was applied to the FreeBayes SNP dataset, after which several quality control filters were applied to the remaining SNPs to reduce the influence of poor mapping and spurious heterozygosity. Multiple filters often identified the same problematic sites and together removed about 12% (1M) of the SNPs.

Filter criterion to pass	SNPs removed (%)
No drone heterozygosity	2
Dataset depth ≥140 (1X per bee)	5
Dataset depth ≤1200 in 100bp window	1
Average genotype quality ≥20	4
Repeat masked sites ≤0.75 in 1000 bp window	2
≥70 (50%) samples genotyped	5
≥80 samples genotyped in a 100 bp window	5

Supplementary Table 2. Depth of coverage.

Technology	Group or lineage	Population	Source	N	Coverage (x)	SNPS
SOLiD						
	A	*adansonii*	Nigeria	10	61	4,578,517
		capensis	South Africa	10	47	4,193,692
		scutellata	South Africa	10	37	4,005,286
		Subtotal		30	145	6,583,102
	C	*carnica*	Germany	10	37	1,690,039
		ligustica	Italy	10	37	1,745,809
		Subtotal		20	74	2,275,598
	O	*anatoliaca*	Turkey	10	38	1,916,693
		syriaca	Jordan	10	59	3,136,725
		Subtotal		20	98	3,580,686
	M	*iberiensis*	Spain	10	54	2,181,659
		mellifera (NOR)	Norway	10	43	1,578,044
		mellifera (SE)	Sweden	10	36	1,777,165
		Subtotal		30	133	2,764,459
	Hybridized	Africanized bees	Brazil	10	40	4,021,673
	Domestic	EU domestic	Sweden	20	36	2,424,202
		US domestic	USA	10	44	2,633,877
		Subtotal		20†	40	3,189,161
	Total			140	610	8,284,886
Wildfire						
	QC drone	DH4 drone	USA	1	20	-
	A	*scutellata*	South Africa	1	23	759,930*
	C	*carnica*	Germany	1	18	329,759*
	O	*anatoliaca*	Turkey	1	21	505,061*
	M	*mellifera*	Norway	1	16	470,991*

*Heterozygous genotypes

†Including 10 EU domestic bees

Supplementary Note

Origin of samples

Our sampling scheme was designed to include at least two subspecies from each of the four honeybee lineages (A=Africa, C=central or southeast Europe, M=northern or western Europe, and O=Middle East/western Asia) previously described based on morphology and genetics[1–5]. We also surveyed domestic bees of no specific subspecies from Europe and USA, and Africanized bees from Brazil. Finally we sequenced a single haploid drone sample from the DH4 lineage[6] and the Asian honeybee *Apis cerana* was included as an outgroup. For each population, we sequenced 10 samples of worker bees from different colonies. The range of native subspecies groups is shown in Fig. 1a.

Samples were collected into 90-100% ethanol. In all cases single samples of worker bees were used, each from different colonies. Care was taken to ensure colonies were unrelated. Details of each collection are provided below. The non-African native samples were taken from apiaries that maintain native subspecies, whereas the African native samples were obtained from apiaries containing caught swarms, classified on the basis of morphology and location.

African subspecies (A group)

African samples were obtained from wild swarms now resident in apiaries. These comprise the Cape bee, *A. m. capensis* (n=10) and *A. m. scutellata* (n=10) from South Africa, and *A. m. adansonii* (n=10) from Nigeria (all from group A).

Subspecies: A. m. scutellata
Collector: Christian Pirk

The African bee *A. m. scutellata* has a wide distribution across east and southern Africa[4]. It inhabits thorn woodland and tall grass savannah, found between 500m and 2400m altitude, from Ethiopia to the Cape Province. It is the most studied of the African subspecies of *A. mellifera* and is small and yellow, as is typical for most African honeybee subspecies. We sequenced ten *A. m. scutellata* worker bees collected from different colonies in South Africa. They were collected from two apiaries in 2011: one in the University of Pretoria, and one in Bronkhorstspruit, 60 km east of Pretoria.

Subspecies: A. m. capensis
Collector: Christian Pirk, Mike Allsopp

The Cape bee *A. m. capensis* is characterized by the ability of worker bees to lay eggs under certain conditions, which are produced by thelytoky[4]. This is enabled by the faster activation of worker ovaries in this subspecies. They are prone to parasitize and lay eggs in other colonies. *A. m. capensis* is dark in color and found in the far south of South Africa. Five of the worker bee samples were collected in 2011 from colonies derived from five wild swarms collected from Eastern Cape around Port Elizabeth area, next to the Kragga Kamma Game reserve. They were initially trapped in three different years: one in 2008, another in 2009 and three in 2010 and placed in two apiaries (CP).

Another five samples were trapped and collected in the vicinity of Stellenbosch in 2011 (MA).

Subspecies: A. m. adansonii
Collector: Abdullahi Yusuf, Christian Pirk

The subspecies *A. m. adansonii* is widely distributed across central and west Africa[4]. It is small and yellow. We collected from managed colonies in Kaduna state, Nigeria: two from a garden apiary, five from beehives in an orchard and three from a small rural apiary.

European subspecies (C group)

European samples were collected from isolated apiaries that maintain pure subspecies. From the C group, these comprise the Italian bee *A. m. ligustica* from Italy (n=10) and the Carnolian bee *A. m. carnica* (n=10) from Austria.

Subspecies: A. m. ligustica
Collector: Marco Lodesani

The Italian bee *A. m. ligustica* is considered as one of the most important for modern beekeeping[1]. It is a yellow bee commonly believed to be adaptable to a range of climates and amenable to modern beekeeping due to docility, productivity and tendency to build big colonies without swarming. We collected ten samples of *A. m. ligustica* from different colonies from apiaries at CRA-API (Reggio Emilia, Italy), from colonies with queens belonging from ligustica queen bee breeders from 8 different Italian regions (from Piedmont in the north to Puglia in the south and Sardinia).

Subspecies: A. m. carnica
Collector: Peter Rosencrantz

The Carnolian bee *A. m. carnica* was originally distributed across central Europe[1]. Its range stretched from southern Poland in the north through Austria to former Yugoslavia and Romania in the south. Like the Italian bee, it is highly valued as a beekeeping strain. Ten samples of *A. m. carnica* collected from different colonies in the apiaries of the Carnica Singer bee breeding station, Purgstall an der Erlauf, Austria.

European subspecies (M group)

Bees from the M-group comprise *A. m. mellifera* from both Norway (n=10) and Sweden (n=10) and *A. m. iberiensis* from Spain (n=10).

Subspecies: A. m. mellifera
Collector: Ingvar Arvidsson, Bjørn Dahle

The European dark bee, *A. m. mellifera* is characterized by long abdominal cover hair and a large body size with a broad abdomen[1]. It was originally distributed across northern Europe, from UK and France in the west, through southern Scandinavia and Germany to Poland and Russia in the east. The bees were sampled from colonies managed by members of the Nordbi conservation project (http://www.nordbi.org/). We collected ten samples of *A. m. mellifera* from Nordic bee breeders in Sweden (IA), spanning colonies in Halland, Dalsland and Uppland in the south to Jämtland and Västerbotten in the north. Ten samples were collected from breeders in Norway (BD), including colonies from the Norges Birøkterlag breeding stations and protected populations in the Flekkefjord area.

Subspecies: A. m. iberiensis
Collector: Pilar de la Rua

A. m. iberiensis is another dark bee from the M lineage, which is restricted to the Iberian Peninsula . We collected ten samples from an apiary at the University of Murcia, Spain.

Middle Eastern subspecies (O group)

Samples from the O-group comprise the Anatolian bee *A. m. anatoliaca* from Turkey (n=10) and the Syrian bee *A. m. syriaca* from Jordan (n=10).

Subspecies: A. m. syriaca
Collector: Nizar Haddad

The Syrian bee *A. m. syriaca* is found on the mountains and valleys east of the Mediterranean Sea. It is described as a small yellow bee with fierce colony defense[1]. We collected samples of *A. m. syriaca* from ten colonies around Amman (Maro station) and Wadi Ben Hammad, Jordan.

Subspecies: A. m. anatoliaca
Collector: Irfan Kandemir

The Anatolian bee *A. m. anatoliaca* is found widely in Turkey and described as similar in size and color to the Italian bee although slightly larger and darker[1]. We collected *A. m. anatoliaca* samples from 10 colonies in 3 different apiaries in central Anatolia, Turkey.

Domestic or hybrid strains

We also collected samples of North American domestic bees from Minnesota, USA (n=10) and two samples of European domestic bees from Sweden (both n=10) from apiaries that did not maintain specific subspecies. We also included a drone from the DH4 line descended from the drone used to produce the original honeybee reference sequence[6] and population samples from Africanized bees in Brazil (n=10).

Strain: US domestic
Collector: Marla Spivak

Samples of US domestic honeybees were collected from 10 colonies in a commercial apiary near St. Bonifacius, Minnesota, USA.

Strain: EU domestic
Collector: Ingemar Fries

The EU domestic samples were collected on two separate occasions (in 2000 and 2010/11), each time from 10 different colonies from beehives distributed across southern Gotland (Sudret och Näsudden), Sweden. These colonies had a mixed genetic background, with known recent gene flow between domestic *A. m. mellifera*, *A. m. ligustica*, *A. m. carnica* and Buckfast bees.

Strain: Africanized bees
Collector: Zila Luz Paulino Simoes

Africanized bees[7] were collected from apiaries in the vicinity of Sao Paolo, Brazil.

Strain: DH4 reference strain
Collector: Jay Evans

We sequenced a drone from the DH4 strain (BeeWeaver Apiaries), descended from the drone that was sequenced by the honeybee genome consortium to construct the *A. mellifera* reference genome assembly[6].

Outgroup

We included samples of the Asian bee *A. cerana* collected at various locations throughout Japan.

Species: Apis cerana
Collector: Masakado Kawata

The Asian honeybee *Apis cerana* represents the most closely related extant lineage outgroup to *A. mellifera*[8]. We collected *Apis cerana japonica* samples from several locations in Japan. They were collected from managed beehives, but derived from wild caught swarms. The sampling locations were distributed from the north of Honshu to the south of Kyushu. The *A. cerana* samples were included to enable inference of ancestral alleles at the variable positions in the *A. mellifera* genome.

Sequencing, mapping and variant processing

Sequencing

We constructed barcoded fragment libraries for each individual sample according to the manufacturer's instructions (Life Technologies) and performed sequencing using the AB SOLiD™ technology at the Uppsala Genome Centre, aiming for 4-6x coverage per sample. Our strategy was to sequence a large number of individually barcoded worker bees at relatively low coverage, a strategy similar to that implemented by the 1000 genomes project[9]. This enabled us to maximize the amount of genetic diversity surveyed and still obtain reliable individual genotype calls using population-based SNP calling. We first performed a pilot study using the sample of 10 *A. m. ligustica* bees, in order to assay the coverage generated per lane. This prompted us to decide on a strategy to sequence 150 barcoded worker samples using 30 libraries per flow cell (equivalent to 5 libraries per lane). Fragment libraries were constructed according to the manufacturer's instructions (LT) and sequenced on a SOLID5500xl machine. For each run, all of the samples were mixed and sequenced across all available lanes. The average coverage per sample was 4.4±1.5x across the genome and 5.2±1.5x per sample at the SNP positions of the final genotype matrix. We refer to this dataset as the population dataset.

Some samples were sequenced at higher coverage to facilitate individual genotype calling in the absence of the population dataset. We sequenced the DH4 reference drone strain to 20x by running the sample on a single lane. We next selected four samples, one from each of the four main lineages, for deeper sequencing for more accurate calling of rare variants necessary for the pairwise sequentially Markovian coalescent (PSMC) analysis[10], aiming at ~20x per sample. The sample selection was based on earlier library quality and barcode compatibility. We selected samples from the *A. m. mellifera* (M group), *A. m. carnica* (C group), *A. m. anatoliaca* (O group) and *A. m. scutellata* (A group) populations. The four samples were converted to WildFire libraries according to the manufacturer's instructions and sequenced on a SOLID 5500 WildFire instrument (three lanes). The previously produced read libraries were combined with the WildFire reads in order to maximize the data available for each sample. The combined libraries had an average depth per sample of 19.6±3.2x. We refer to this dataset as the PSMC dataset.

Mapping and quality control

We developed a pipeline for read mapping and quality control. We first mapped the 75bp color space reads to the honeybee apiMel4.5 reference genome using LifeScope™ v2.5.0/v2.5.1 with the default settings[11]. About 83% of the produced *A. mellifera* reads mapped against the reference sequence, whereas we could only map 48% of the more divergent *A. cerana* reads. Picard v1.41 (http://picard.sourceforge.net/) was used to mark PCR duplicates (~17%). The honeybee reference genome spans ~229MB distributed across 16 nuclear chromosomes (Group1-16), the mitochondrial chromosome (MT) and a number of currently unplaced scaffolds or contigs (GroupUN). To help downstream pipeline parallelisation, reads were rearranged into one BAM file per individual and chromosome using SAMtools v0.1.18 (http://samtools.sourceforge.net/) and Picard. Read group information was modified across the BAM files using Picard. Putative indel regions were realigned using GATK v2.0[12] and the FreeBayes v0.9.6[13] bamleftalign tool. We noticed very little alignment differences after the realignment steps. Per-base alignment quality (BAQ) scores were added to the reads using GATK.

Genotyping, filtering and imputation

Population dataset

We performed SNP calling with FreeBayes v0.9.6. The chromosomes were further subdivided into 1Mb or 2Mb segments and analyzed in parallel. We initially called SNPs using population-based SNP calling across all *A. mellifera* worker samples (n=140). Calling was performed separately for the re-sequenced drone (DH4). The ten *A. cerana* samples were also SNP called separately from the population dataset. We mostly used the default settings in FreeBayes, together with the binomial observation priors on read placement, strand balance and read position probabilities (-V flag), while ignoring multi-nucleotide polymorphisms (MNPs) (-X flag) and contiguous alleles (-E 0). Duplicate reads were excluded from the analysis. We used the haploid setting (-P 1) for the mitochondrial chromosome. The population priors were switched off (-k) for the samples that were called individually. Sequencing and read mapping errors can inflate sequence variation and interfere with SNP inference. As a base filter, only biallelic SNPs with a minimum quality score (QUAL) of 50 were accepted for further analysis. We then applied an additional set of filters to further remove potentially erroneous SNPs from the dataset (summarized in Supplementary Table 1).

The first filter was based on nuclear SNPs identified from re-sequencing the haploid drone sequence, which we specified to be diploid when calling the SNPs. Using the previously specified FreeBayes settings and base filter, we identified 944K SNPs at an average read depth (DP) of 19.7x in the drone, out of which 756K were homozygous for the alternate allele and 188K (20%) were heterozygous genotypes. The heterozygous SNPs are clearly spurious due to the haploid nature of the drone sample. The DP at the homozygous SNPs was 17.4x. At the heterozygous positions it was nearly the double (DP=29x), indicative of mapping artifacts that could also affect the population dataset. Having applied the base filters, we next queried the population dataset at these positions. The DP at the SNPs in the population dataset at this filter stage was 673x. At the positions where the drone was homozygous for the alternate allele, the DP was 536x in the population dataset and the proportion of heterozygous genotypes was 29% on average. In contrast, at the positions where the drone had been found to be heterozygous, we found the DP to be unusually high (1020x) and the proportion of heterozygous genotypes to be drastically elevated (56%) (Supplementary Fig. 1), suggesting a shared mapping bias across the two analyses. These positions most likely represent errors in the assembly or mapping, which may be caused by segmental duplications or copy number variants. We therefore removed these positions from the population dataset.

We applied a minimum depth filter of 140x (1x per bee) for accepting a SNP as well as a maximum depth of 1200x (about 2 times the dataset average) in a 100bp window around the SNP. We removed SNPs with >75% repeat masked bases in the vicinity (1kb window). The average genotype quality score (GQ) of genotypes was ~34. We therefore removed all SNPs with a mean GQ was lower than 20. Lastly, sample representation at SNPs was used as a filter criterion. We removed all SNPs where less than 70 (half) of the samples had had their genotype called (due to no or low quality coverage) and where less than 80 samples had coverage in 100 bp windows around the SNP. These filter criteria often overlapped and identified the same problematic sites. Two or more filters typically identified the same unreliable SNPs that were removed from the population dataset. In addition, sites that did not pass these filters typically exhibited an elevated proportion of heterozygous genotypes compared with other SNPs, suggestive of them being artifacts. In total, 1,018,515 SNPs (~12%) were filtered away from the population dataset by applying these second stage filters (Supplementary Table 1). The canonical version of the population dataset used for downstream analyses spanned 8,284,886 SNPs.

Since there was no possibility to detect SNPs at unmapped positions, as well as at positions and windows which did not pass our SNP filter criteria, these positions where recorded as unsurveyed. Downstream population parameter estimates such as region-based or full-genome levels of variation were corrected for available sites whenever applicable. This translated into 78% (179Mb) of the honeybee genome being surveyed and available for SNP discovery in the population dataset.

We used the mapped sequences of the ten *A. cerana* outgroup samples to infer the ancestral state of *A. mellifera* SNPs. For each position in the *A. mellifera* ingroup population dataset, we first collected the corresponding alleles of each individual *A. cerana* sample from its original BAM file. Ambiguous or potentially heterozygous *A. cerana* genotypes were resolved by overlaying the genotype inferred with FreeBayes after applying the biallelic and QUAL≥50 base filters. If there was no sample genotype available at such an ambigous site we considered it missing and did not include it in the subsequent ancestry inference. Most often, we observed no shared variation (variation involving the same two alleles) between *A. mellifera* and *A. cerana*. The ancestral allele was then simply taken to be the allele shared between the ingroup and the outgroup. The other allele in the ingroup was then assigned to be the derived state. In cases where we observed shared variation between the ingroup and outgroup we did not infer ancestral/derived states. We assigned allelic ancestry to about 63% of the 8.3M SNPs.

It is clear that average coverage and called genotypes per individual varies according to total number of mapped reads and that some honeybee samples were sequenced at a higher mean depth than others (Supplementary Fig. 2a). The most poorly sequenced sample had ~4.9M genotypes called whereas the best-sequenced samples had about 8.1M genotypes called. However, population genotype representation at SNPs was similar, despite the fact that coverage was variable between individuals and populations. Each population typically had at least one out of ten genotypes called at 99.9% of the 8.3M SNPs, at least five genotypes called at ~98.4% of the SNPs and all ten genotypes called on average at 71% of the SNPs (Supplementary Fig. 2b). This suggests that we have good and similar power to genotype common SNPs in each population.

The overall completeness of the genotype matrix was 94%. We imputed the missing 6% of the genotypes and phased haplotypes using Beagle v3.3.2. We used the genotype likelihoods reported by FreeBayes and the following settings: niterations=30, nsamples=10 and lowmem=true.

PSMC dataset

Since each individual sample in the population data is sequenced at a low depth, it is clear that the power to detect rare alleles that only occur in a few copies is reduced: they are difficult to distinguish from technical artifacts and the individual low coverage sample may not even be mapped at the positions. For the PSMC analysis it is necessary to obtain a diploid sample that represents all variation, including rare variants, as well as accurate genotypes. This is because the analysis relies on the distribution of heterozygous positions along the chromosomes to infer and compile coalescent times. These rare variants are detected with low power in the population sample but with greater power and accuracy in the samples sequenced at high depth of coverage since all heterozygous genotype can be expected to be equally well covered, regardless of the allelic frequency in the populations.

We had previously observed excess heterozygosity at sites with unusually low or high coverage. This matched the usage guidelines of the PSMC implementation, which cautions against including SNPs inferred from too low (less than a third of the genomic average) or too high (more than twice the genomic average) depth of coverage. We therefore performed individual FreeBayes SNP calling and filtering of the four PSMC samples within a restricted depth of coverage interval. First, we applied a minimum coverage threshold of one third of the genomic average and a maximum threshold of twice the genomic average of each sample, according to the recommendations. Moreover, we expected truly heterozygous positions sequenced at high depths of coverage to have the reference and alternate alleles mapped at about the same read depth. We therefore next included only the biallelic heterozygous genotypes where the minor allele was supported by at least 25% of the observed reads. Lastly, we removed the SNPs for which we had earlier detected elevated levels of heterozygosity (such as the heterozygous genotypes in the haploid drone) in the population dataset. This approach enabled high-confidence calling of common and rare variants in well-mapped areas along the 16 assembled chromosomes (200Mb). For each sample, read data spanning about ~62% of the genome was found to pass these filters.

In the *A. m. mellifera* (M group) sample, we recovered 334,215 heterozygous genotypes (336 b/SNP; 0.0027 SNPs/b). In the *A. m. carnica* (C group) sample, we found 304,644 such SNPs (407 b/SNP; 0.0025 SNPs/b). We detected 358,176 SNPs (344 b/SNP; 0.0029 SNPs/b) in the *A. m. anatoliaca* (O group) sample and 685,808 SNPs (179b/SNP; 0.0056 SNPs/b) in the *A. m. scutellata* (A group) sample. These levels of variation in each sample closely matched the genetic variation measured in the populations (see below; Table 1, Supplementary Table 2).

Using the same SNP call settings and filters, we detected 547,503 homozygotic differences between the haploid drone and the reference genome sequence distributed across 166Mb (73% coverage; 334b/SNP). These SNPs most likely represent true variants within the DH4 line that differ from the drone we sequenced and the one used to make the reference assembly.

Analysis of genetic variation and population history

Tree construction

We used several methods to draw trees from the data. We first constructed a neighbor-joining tree using allele-sharing distances from the individual samples of the native species (Fig. 1b). This confirms the clustering by group (A, C, M, and O). In all groups with the exception of Africa, samples for each individual subspecies are clustered together, with the exception of two the two outlier samples from the M group (see above). The outgroup was found to connect at the center of the tree, rather than within African branches as found in a previous SNP analysis[5]. It is worth noting that the previous dataset suffered from ascertainment bias (the SNPs were mainly derived from comparisons of the reference assembly with Africanized bees) and also that alternative topologies are supported when certain lineages with evidence for admixture are removed[14]. In order to test whether our tree topology was robust to selection on coding sequences, we constructed trees using 1) only the exon sequences, 2) all genic sequences (including introns) and 3) intergenic regions (Supplementary Fig. 3). All trees have nearly identical topology.

Our results do not explicitly support any model of migration of *A. mellifera*. However, the fact that the root of our tree of *A. mellifera* subspecies is not found in Africa, as presented by Whitfield et al.[5], prompts us to explore other hypotheses. Given that all other *Apis* species are restricted to Asia, it is most parsimonious to assume the origins of *A. mellifera* also lie in this continent. Whitfield et al.[5] suggest a scenario whereby adaptation to cold is a recent adaptation that allowed *A. mellifera* to leave Africa after a more ancient migration from Asia. However, Ruttner[1] emphasized that both *A. mellifera* and *A. cerana* are adapted to cold, suggesting that cold-adaptation is an ancestral trait in *A. mellifera*. Hence, there is no reason why *A. mellifera* could not colonize Europe immediately after a migration from Asia.

Genetic variation and effective population size

In order to determine whether SNPs are private or shared among continents and the four major groups (M, C, O, and A), we recorded the groups and combinations of groups that each SNP was polymorphic in. We found that in about 1M SNPs, the variation is shared among all four groups and that the African bees by far have the largest amount of private variation (Fig. 1c).

We measured genetic variation using Watterson's theta estimator of the population mutation rate per base (θ_{w}), using the standard formula[15]. Genetic variation per bp was estimated by dividing by the number of sites in the genome with sufficient coverage to call SNPS. We estimated effective population size (N_E) using the standard equation $\theta_{\mathrm{w}} = 3N_E\mu$ (a factor 3 is used because of honeybee haplodiploidy), where μ is mutation rate per site per generation (see below). Our estimation of N_E >100,000 is much higher than previous estimates of short term N_E obtained by counting total numbers and relatedness of samples from drone congregation areas, where thousands of males aggregate to mate with virgin queens. Each queen mates with several drones, and it is estimated that each congregation contains drones from 250 colonies[16], Although this number is much lower than N_E we estimate, there are many factors that could make the long term N_E higher and maintain high genetic variation. The genetic composition of drone congregation areas is highly variable over time[17], and showing very high queen turnover. This suggests that there are high levels of migration and gene flow, resulting in large interconnected populations with very little substructure. Furthermore queens mate with up to 25 drones[18], and relatedness among drones is low[17].

Mutation rate

We estimated mutation rate using divergence from *A. cerana*. As we noticed that the *A. cerana* reads had poor mapping in non-coding regions due to strict mapper cutoffs, we suspected that divergence would be underestimated from the mapped reads. We therefore chose to align six 1 kb sequences from non-coding "control regions" sequenced in *A. cerana* as part of another study[19]. We identified their orthologous regions by performing a BLAST search against the honeybee genome. We then extracted the matching sequences and aligned them with MAFFT[20] using the default settings. We estimated divergence as 7.37% in the control regions, compared to ~2.6% across our mapped *A. cerana* data (~2.2% in coding regions and ~2.8% in non-coding regions). We next estimated mutation rate per base per year assuming a divergence time of 7 million years and a generation time of 1 year[21]. This results in a mutation rate as 5.27 x 10^{-9} per year. This is comparable to estimates of mutation rates in *Drosophila*, from which the most accurate measures of mutation rates in insects are available. A direct estimate from the *Drosophila* mutation accumulation lines is 5.8 x 10^{-9} (ref. [22]) and a phylogenetic estimate from *Drosophila* divergence 2.6 x 10^{-9} (ref. [23]).

Linkage disequilibrium

We used the imputed and phased data to determine the average decay of linkage disequilibrium (LD) in different groups and subspecies. LD decays rapidly with physical distance in all of the subspecies consistent with the extremely high recombination rate observed in honeybee (19 cM/Mb)[24]. At distance >1 kb the average r^2 is <0.2 for all subspecies, the domestic populations and the Africanized bees (Supplementary Fig. 4). However, there are differences in the rate of decay, which reflect population history and size. LD is extremely low in Africa (A) and highest in the central/SE European group (C), again reflecting higher long-term effective population size in Africa. Interestingly, we observe nearly as rapid decay of LD in the Africanized bees from Brazil (and effective population size) as in the African subspecies, even though the former is known to have been founded from a very small population[7]. Honeybees have high levels of genetic variation both in terms of number of SNPs and in haplotype diversity, suggesting they are able to efficiently maintain high levels of variation while restricting reproduction to the queen and drone castes (see above).

Population structure

We applied both SNP and haplotype-based Maximum Likelihood and Bayesian model-based analyses to infer population ancestry across the full 8.3M SNP dataset (Supplementary Fig. 5; Supplementary Fig. 6).

ADMIXTURE

We used ADMIXTURE v1.22[25] to identify allelic origin and population coancestry in the natural subspecies and domestic and hybrid lineages. This is a fast implementation of a algorithm similar to STRUCTURE[26], that is suitable for large datasets. This provides 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 explored coancestry spanning 2-6 ancestral populations (K=2-6) (Supplementary Fig. 5b). The algorithm was run for 30 iterations in each analysis. The results were not qualitatively different to those obtained in a similar analysis with STRUCTURE[26]. The optimal number of hypothetical ancestral groups (K) was inferred using the cross-validation (CV) error estimation method implemented in ADMIXTURE. The algorithm estimates the CV error for each value of K by masking and re-inferring genotypes. The optimal value for K is taken as that with the lowest CV error. The cross-validation (CV) procedure implemented in ADMIXTURE was applied to all data (140 samples including subspecies, domestic and Africanized samples) and a dataset including only the natural subspecies (100 samples) (Supplementary Fig. 5d). In both cases the optimal value for K was found to be 4, suggesting that the variation is best summarized as being subdivided according four ancestral groups. This is in agreement with other reconstructions identifying the four major continental groups (M, C, O, and A).

At K=4, the natural subspecies were subdivided into the four major continental groups (M, C, O, and A), with some level introgression in a few samples (most notably two presumably outbred M-group bees). We found the domestic lineages to harbor mostly C-group alleles but also that about 10-20% of alleles appear to be of other origins. The Africanized bees from Brazil were found to be mostly African with about 15-20% alleles being of from the M group, with indication of affinity with *A. m. iberiensis* from the Iberian peninsula. It has previously been suggested to that Africanized bees have experienced a selective removal of C alleles[27], but the absence of C alleles could also reflect ancestry or gene flow from bees of Iberian origin into southern America, which is consistent with recent human migration patterns (Supplementary Fig. 5b). We detect admixture between *A. m. syriaca* bees from Jordan and bees from the A group. A zone of admixture between African and European bees has been previously detected in Iberian peninsula[28–31]. Jordan could therefore represent an additional zone of admixture. The presence of a mitochondrial haplotype different to most A and O-group bees was recently used to suggest that this subspecies forms a fifth major group[32]. We interpret our results as evidence for an Eastern zone of hybridization with African bees rather than a new distinct group. One *A. m. iberiensis* sample from Spain and one *A. m. mellifera* sample from Sweden were found to have unusually high levels of mixed allelelic ancestry (Fig. 2a). These could potentially be highly admixed and were therefore excluded from downstream calculations of allelic differentiation and heterozygosity involving the M-group.

fineSTRUCTURE

We next applied the haplotype similarity analysis as implemented in fineSTRUCTURE v0.0.4 to further assess population ancestry and clustering. The method uses a haplotype-based approach to estimate ancestry of blocks of DNA across the genome of each sample that are then summarized as a coancestry matrix that describes ancestral relationships between samples. The initial analysis was carried out with the chromopainter tool using the following settings: data unlinking in the absence of recombination rates among SNPs (-u), chunk size of 1000 (-k 1000), 10 samples per recipient haplodiploid (-s 10) and all individuals against all in terms of haplotype transmission (-a 0 0). The chromosomes were analyzed in parallel and the results were merged using the fineSTRUCTURE GUI application.

The population structure inferred with fineSTRUCTURE analyses was highly consistent with the trees and clusters recovered using a neighbor-joining tree based on allele sharing and the multidimensional scaling analysis (see below; Supplementary Fig. 5a). The four major groups were again recovered, with closest affinity between the C and O groups. Most of the samples of the natural subspecies clustered within their expected groups, although there were exceptions, such as the clustering of *A. m. capensis* bees within the *A. m. scutellata* population and the previously recovered basal position of the two outbred M-group bees. Most of the domestic samples from both the US and EU formed a sister clade to the C-group bees (although some clustered with the outbred M-bees), illustrating the influence of C/SE European alleles and origins of current beekeeping strains. The Africanized bees from Brazil clustered with the A-group, again demonstrating their similarity to African bees. fineSTRUCTURE detected a high amount of haplotype sharing between the *A. m. syriaca* bees from Jordan and the African subspecies, as recovered also by ADMIXTURE (see above; Supplementary Fig. 5a,c).

TreeMix

TreeMix was used to infer population splits and migration events among the natural populations (Supplementary Fig. 6). This method constructs a bifurcating tree of populations and then identifies potential episodes of gene flow from the residual covariance matrix. We smoothened the allelic variation between populations by pooling them according to their major continental group and repeated the analysis on the four groups. The runtime mixture/migration parameter (-m) was set to increase from 0 to 8 events for natural subspecies and 0 to 6 events for the groups. We found the method to recover a population tree mostly similar to our other inferences: the subspecies clustered within their respective continental group. The two European groups were again found to be genetically distant; C-group bees clustering with the Middle Eastern O-group whereas the M-group bees clustered with the African bees. The majority (98.2%) of the variation was explained by the ML topology and the most important migration events based on the residuals (explaining an additional 0.6% of the variation) involved African migration of alleles into the O-group, as supported by other assessments, followed by minor gene flow between M↔C and C→A in accordance with the limited levels of introgression detected by ADMIXTURE.

Evolutionary relationships between populations

Population tree reconstruction

We calculated genetic distances between native populations and groups using the Reynolds et al (1983) F_{ST} estimator[33] (Supplementary Fig. 7a-b). The pairwise F_{ST} distances were used to produce NJ trees (Supplementary Fig. 7c) with PHYLIP. The trees recapitulate the relationship between the groups, and also place the root in the center of the subspecies groups. It is interesting to note that the highest degree of population differentiation is observed between the two European (M and C) groups (F_{ST} = 0.56), indicative of a large effect of genetic drift in these populations with small N_E. Within groups, levels of population differentiation are generally very low (F_{ST} = 0.05 - 0.18) suggesting little or no genetic isolation. Population differentiation among African subspecies is particularly low, and comparable to that observed between populations of the same subspecies (*A. m. mellifera*) sampled from Norway and Sweden. We also applied the NeighbourNet algorithm in SplitsTree to reconstruct a splits network from the pairwise F_{ST} distances, the structure of which was highly similar to the NJ tree but indicating some level of ambiguity among the most basal splits and the position of the outgroup and root (Supplementary Fig. 7c).

Multidimensional scaling

We analyzed relationships between samples using a standard multidimensional scaling plot methods as implemented in PLINK[34]. As found previously using both genetic and morphometric analysis, and from the analyses above, the samples clearly cluster into 4 major groups corresponding to A, C, M, and O. Both the EU and US domestic beekeeping strains cluster close to C, and the Africanized bees from Brazil cluster closely with A (Supplementary Fig. 7d).

Concordance analysis and molecular dating

We used a coalescent approach to estimate the time of splits between populations based on the concordance between population trees and gene trees[35]. 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 can estimate the amount of genetic drift that has occurred on specific branches of the population tree, and hence the times of divergence, by quantifying the amount of genealogical concordance[35]. There are three possible genealogical topologies between four genetic lineages: one concordant topology matching the population topology and two discordant topologies. If we consider 4 lineages with topology (((1,2),3),4), there are only three site patterns that are informative about the topology - AABB (concordant), ABBA and BABA, (both discordant):

$P_{\text{concordant}} = N_{\text{AABB}}/(N_{\text{AABB}}+N_{\text{ABBA}}+N_{\text{BABA}})$

The proportion of concordant topologies increases with time from the divergence event. This proportion can therefore be used to estimate the time since divergence:

$E(P_{\text{concordant}}) = 2e^{-T}/3$ and the estimator of T is then: $T = -\log((3-3P_{\text{concordant}})/2)$

Standard errors were inferred as 95% confidence intervals (CI) around the maximum likelihood estimate of T (TMLE, as computed above) using the relative log likelihood procedure:

$\log(L(T)) = N_c \times \log(1 - 2/3e^{-T}) + N_d \times \log(2/3e^{-T})$,

where N_c is the number of concordant sites, N_d is the number of discordant sites and the 95% CI of T is that where $\log(L(T)) - T_{\text{MLE}} > -2$. The ΔT representing this level of variation at either side of the maximum estimate is then: $\Delta T = | T_{\text{MLE}} - T |$. Due to high number of SNPs included in the analyses, our estimates of the CI was usually very small (~1% of the absolute value of T).

When there is no directional gene flow, the proportion of ABBA and BABA are expected to be the same. The difference between ABBA and BABA is the basis of the Patterson's D statistic to detect biased gene flow[36,37]. When all lineages come from different populations, then the proportion of concordant loci is informative of the internode distance between the node leading to 1 and 2 and the node leading to 3. When 1 and 2 are sampled from the same population, then the divergence time is equal to the total distance between 1 and 2.

The honeybee evolutionary tree is characterized by two major events. First, each of the four groups are comprised of closely related subspecies that split only recently from each other (Fig. 1b,e). Second, the four groups split from each other some time in the past, although the branching pattern is unclear. However, levels of genetic drift, which are dependent on N_E, affect the branch lengths and it is clear that African populations in particular have had larger effective population sizes over longer periods of time.

To investigate within-group population splits, for each SNP we randomly sampled chromosome copies from populations descended from a particular node in the tree, where samples 1 and 2 came from the same subspecies on one descendant branch and sample 3 came from a subspecies on the other descendant branch. Sample 4 was always represented by the outgroup state (*A. cerana*). Which of the two descendant branches from which to choose samples 1 and 2 was chosen at random for each draw, but samples 1 and 2 were always drawn from the same subspecies. Sample 3 was always drawn from a population on the alternate descendant branch. The purpose of this randomization algorithm was to give an unbiased estimate of population split times not affected by biased gene flow or demographic changes in a specific population. Sampling within a population was done by first randomly selecting chromosome copies and then recording the allelic variant present across the full genome. Only cases that match one of the three topology-informative site patterns were retained. We counted the number of concordant and discordant sites from this analysis and used them to estimate T in units of $1.5N_E$ generations (because of haplodiploidy). In order to eventually convert these into years we estimated average N_E in each lineage from θ_W estimated across all samples in the lineage. We investigated timing of the six within-group splits as shown in Supplementary Fig. 8a. The two outlier samples from M group bees that showed unusually high levels of admixture and did not cluster with other samples in the NJ tree were excluded from the analysis due to them being potential hybrids (one *A. m. iberiensis* from Spain and one *A. m. mellifera* from Sweden).

Sampling between pairs of groups was performed in a similar fashion, where samples 1 and 2 came from the same subspecies in one group, sample 3 came from a subspecies in another group and sample 4 was the outgroup state. Which of the two groups that was represented by 1 and 2 was chosen at random in order to eliminate the effects of directional gene flow (see above). We performed all possible pairwise comparisons of two groups. Results of the within-group and between-group comparisons are shown in Supplementary Fig. 8b. All samples of *A. m. syriaca* were excluded from this analysis due to signals of introgression and the O group was therefore represented solely by the *A. m. anatoliaca* samples.

We converted measures of T into years by correcting for N_E. For the within-group estimates we used N_E estimated for the entire group in question and for between-group estimates we used the average N_E of the two groups being compared, taking it as estimate of ancestral population size. This allows us to estimate the timing of splits between subspecies within groups and between groups. Divergence between subspecies within the same group was estimated to have occurred between 13-38,000 YBP (with the exception of *A m. syriaca* and *A. m. anatoliaca,* which have an older divergence time, presumably because of African admixture in *A. m. syriaca*). These dates roughly correspond to the last glacial maximum, and suggest that honeybee subspecies within each of the four groups began diverging from each other at this time[38]. This scenario is consistent with the pattern of postglacial colonization inferred in many other animal species[39]. Pairwise divergence between the A, C and M lineages are all estimated to have happened ~300,000 YBP, suggesting that the ancestral honeybee population split and expanded to different geographic regions around this time. This estimate is more recent than estimates based on divergence of mtDNA lineages of 0.67 million YBP[2], 1 million YBP[3] and 0.3-1.3 million YBP[40] assuming an mtDNA mutation rate of 2% per year. Our more recent date suggests extant mtDNA lineages diverged before the split between the four main groups, and were segregating in the ancestral population. In general, data from single loci such as mtDNA are not suitable for estimating the age of within-species population splits[41].

PSMC analysis

We used the Li and Durbin pairwise sequentially Markovian coalescent (PSMC) method to reconstruct past changes in effective population size. This method uses variation in the density of heterozygous sites in a single diploid individual to infer the distribution of time to most recent common ancestor (TMRCA) of haplotypes across the genome by applying a hidden Markov model to partition the genome into blocks with varying ancestries across the genome[10]. The SNP selection and filtering procedure is detailed above. The SNPs of each of the four samples in the PSMC dataset were overlaid onto the reference chromosome sequence and filtered positions outside the evaluated depth interval were masked as unsurveyed/missing. To avoid estimating coalescent times across scaffold borders (marked by stretches of 50,000 consecutive Ns in the apiMel4.5 reference chromosomes), we subdivided the 16 nuclear chromosomal sequences into scaffolds. The scaffolds were either treated as separate sequences (many short accessions), or, as an alternative method, merged back into full chromosomes (few long accessions; w/o scaffold borders) when converting the nucleotide sequence data into the native PSMC format. The results were highly consistent across the two approaches. We first used the default block size (100b) and SNPs representing transversions for encoding the heterozygosity. Due to the high level of heterozygosity and rapid decay of LD in honeybees compared to humans, we also used a small block size of 20bp compared to the default originally specified for human sequence data (100b) to encode the heterozygosity using all SNPs. We found that resolution far back in time increased when the heterozygous genotypes were thinned to include only transversions. At this block size and using only transversions, we found that about 0.6-0.7% of blocks contained at least one heterozygous genotype in *A. m. mellifera* (M group), *A. m. carnica* (C group) and *A. m. anatoliaca* (O group) samples and that about 1.2-1.4% of the blocks were doing so in the *A. m. scutellata* (A group) sample. When all SNPs were included the proportion of blocks with heterozygous genotypes was about five times larger. We used a model based on 90 time intervals with 45 free interval parameters (-p "45*2"), a maximum $2N_0$ coalescent time of 15 (-t15) and ran the analyses for 30 iterations (-N30) (Supplementary Fig. 9).

Using our approach we found evidence that effective population sizes have fluctuated greatly. The present day European (M+C) and Middle-Eastern (O) populations demonstrate aligned expansions and contractions that may have happened during the same time periods in the past, whereas the fluctuations in the African bees are shifted in time relative to the other bees. Assuming a generation time of 1 year per bee and our estimated mutation rate of 5.27×10^{-9} (or 1.49×10^{-9}, including only transversions), the fluctuations match postglacial expansions in the European and Middle-Eastern bees and postglacial contractions in African bees, suggesting a link between climate, habitat and population demography[38,39].

Bootstrapping analyses were performed to describe the uncertainty in the PSMC evidence. We set the splitfa trunk size to correspond to 1Mb of nucleotides and analysed 125 bootstrap replicates per honeybee sample. We generally found uncertainty in the estimation of historical population size to increase with time (Supplementary Fig. 9b).

Signatures of selection

Gene annotation and function

The Honey Bee Genome Sequencing Consortium has continuously revised the honeybee reference genome since its first publication in 2006. In this study, we used version 4.5 of the reference genome (apiMel4.5) for nucleotide sequences and scaffold coordinates and the current draft version of the gene annotations and gene coordinates to assess gene function and selection. We focused on the genes of 16 assembled chromosomes (13,285 genes). We computed rates of $CpG_{O/E}$ from the gene regions and along chromosome window intervals as proxies of methylation levels. $CpG_{O/E}$ is the number of CpG dinucleotides corrected for GC content and was computed as:

$CpG_{O/E} = N_b \text{ x } N_{CpG} / (N_C \text{ x } N_G)$,

where N_b are the number of bases in the sequence, N_{CpG} are the number of CpG sites, N_c is the number of Cs and N_G is the number of Gs. Due to clear bimodal distribution, we subdivided the genes into high and low $CpG_{O/E}$ classes depending on their $CpG_{O/E}$ level compared to the mean $CpG_{O/E}$ (1.19) as well as into high and low GC classes relative to the mean GC (0.31).

We analyzed two sets of honeybee caste-biased gene expression data assessing differences in brain gene expression between: i) queens vs workers (ref. [42]; ~1770 accessions) and; ii) drones vs workers (ref. [43]; ~6500 accessions). These datasets were divided into those with significantly increased expression in one caste and unbiased genes based on original definitions. Whenever available, gene accession IDs listed in the expression studies were matched against previous (v4.0) or current (v4.5) versions of the reference genome. Microarray probes used in the expression studies were also assigned to genes using BLAST when we were unable to link accession IDs. Such BLAST alignments of short probes against genes had to span at least 40 nucleotides and have an e-score of at least 0.5 to be significant and included. Since the gene regions themselves have been revised over time, we also used BLAST to establish homology among revisions and carry the information over to our gene list. BLAST gene alignments had to span at least 100 nucleotides and have an e-score of at least 0.5 to link annotations. Gene names linked to previously hypothesized orthology in version apiMel4.5 were inserted into the gene list.

In order to identify homologs of the honeybee gene set, we blasted the honeybee gene set against the *Drosophila melanogaster* reference gene set (BDGP5) to identify homologs. This was performed using the honeybee CDS sequence as query and running BLASTx against *Drosophila* peptide sequences obtained from BioMart (http://www.ensembl.org/biomart/martview/). We used the best *Drosophila* hit score to determine the gene ontology category for each honeybee gene. The pairwise sequence alignment had to span at least 50 peptides and have an e-score of 0.5 to assign orthologs. About ~7100 *Drosophila* genes were linked to honeybee orthologs using this method.

We thus constructed a comprehensive annotation resource that cross-referenced honeybee reference genome versions, gene regions and CDS reading frames, rates of $\mathrm{CpG}_{\mathrm{O/E}}$, caste-biased expression and *Drosophila* orthology. The database was used to relate patterns of genetic variation and selection to biological properties and identify functional enrichments.

Patterns of genetic variation along chromosomes

The SNPs detected in the population dataset were overlaid across the chromosomes and gene coordinates. The genes were subdivided into regions (5'-UTR, CDS, intron, 3'-UTR and downstream/upstream intergenic intervals). The genes were also partitioned according to low/high CpG classes and caste-biased expression categories[43,44]. We then estimated the mean and variation in diversity (as estimated using θ_W), partitioned according to gene region, CpG and expression properties. The mean and variation was computed by bootstrapping (1000 replicates) the observations of each category. For the CDS region we computed diversity over to divergence to assess signals of positive selection in *A. mellifera* (i.e. low diversity relative to high divergence) by comparing the levels of diversity to the divergence computed from the mapped *A. cerana* reads. We had earlier noticed that we underestimated the divergence between the two species in neutral regions and therefore did not apply the test to non-coding intervals. We also analysed levels of genetic variation in 1Kb windows at increasing distance from genes divided according to $\mathrm{CpG}_{\mathrm{O/E}}$ category.

Analysis of population differentiation

The main aims of the study were to describe global patterns of variation and detect selection in the honeybee. We therefore chose to identify alleles under selection among the major continental groups (M, C, O, and A) and in particular selection generating allelic divergence between the European and African bees. This was done by identifying SNPs with high F_{ST} in three comparisons, in turn clustering two groups together and comparing them to the third group: A vs MC, C vs AM and M vs AC. We omitted the O group due to evidence for admixture with the A group. We used the Weir & Cockerham (1984) F_{ST} estimator[45] and scanned the full genome and cross-referenced the contrast-specific F_{ST} values with our expression and gene regions annotation resources (see above). Intergenic SNPs were assigned to the closest gene (we did not assign such SNPs that were upstream of the first gene on a scaffold and downstream of the last gene since cross-scaffold distances were impossible to compute). The two potentially admixed M-group samples (see above) were excluded from the F_{ST} analyses.

We measured the average F_{ST} of SNPs in different functional categories. We divided SNPs by F_{ST} intervals and calculated the proportion in each annotation category (coding, noncoding, intronic, UTR). We do not find significant differences between average F_{ST} of SNPs on coding or noncoding regions (P = 0.545, bootstrap; Supplementary Fig. 10a). This contrasts to an earlier study using a smaller dataset that genotyped a panel of 444

SNPs and found significantly elevated F_{ST} in exons[27]. Our analysis suggests that although SNPs with $F_{ST} > 0.9$ are enriched in genes, indicating that genes are enriched for targets of positive selection, this does not have an overall effect on F_{ST} in genes. We also estimated enrichment of CpG and expression categories. There are slight differences in average F_{ST} when SNPs are classified according to gene expression, $CpG_{O/E}$ and surrounding GC, as computed from bootstrapping the dataset. In particular, genes with lower GC content and low $CpG_{O/E}$ tend to have slightly higher F_{ST} than high GC and high $CpG_{O/E}$ genes. We also analyzed the average GC content of SNPs with different F_{ST} between the African and European bees. We find weakly correlated relationships (r^2=0.01681; $p < 2.2 \times 10^{-16}$) between GC and F_{ST} at noncoding SNPs. For coding SNPs, we similarly observe a very slight reduction in GC content for high-F_{ST} SNPs ($F_{ST} > 0.9$; $r^2 = 0.005469$; $p < 2.2 \times 10^{-16}$) (Supplementary Fig. 10b). This contrasts with an earlier finding suggesting that GC rich regions are faster evolving[46]. A negative correlation between F_{ST} and GC in coding SNPs was observed by[27] although the trend reported here is much less pronounced. The fact that we also do not observe a strong correlation between GC content, a strong correlate of recombination rate in honeybees, and F_{ST} in either coding or noncoding regions, suggests that any potential effect of recombination on F_{ST} (e.g. due to GC-biased gene conversion[46,47]) in honeybees is negligible (Supplementary Fig. 10a-b.

We assessed the proportion of non-synonymous over synonymous substitutions encoded by SNPs in the CDS regions, expecting highly differentiated SNPs to represent selection on the amino acid sequence of proteins and more frequently substitute amino acids than SNPs at low or intermediate F_{ST} values. This pattern was true in two of the contrasts (A vs MC and C vs AM; Supplementary Fig. 10c), suggesting that protein sequences are under divergent positive selection among the major continental groups of honeybees.

We sought to identify the effect of selection around high-F_{ST} SNPs by quantifying levels of diversity around SNPs according to F_{ST}, which was performed by measuring levels of heterozygosity using θ_w in windows of 1Kb at increasing distance from SNPs subdivided by F_{ST} intervals of 0.05. Fixed SNPs (F_{ST}=1) were placed in a separate interval. We find that high-F_{ST} SNPs are associated with reduction in linked genetic variation. This reduction is most pronounced around SNPs fixed for alternate alleles and extends about 100 kb. For the 194 SNPs that are fixed for alternate alleles in the A vs MC comparison, there is a 23% reduction in linked (≤20kb) neutral diversity compared to average diversity further away (100-250 kb) (Supplementary Fig. 11a). When the analysis is restricted to SNPs in coding regions the reduction is more pronounced (31% for fixed SNPs; Supplementary Fig. 11b). For intronic SNPs, there is a modest reduction (23% for fixed SNPs; Supplementary Fig. 11c) and intergenic SNPs are only weakly associated with reduced variation (6% reduction for fixed SNPs; Fig. 10D). The signature of selection is therefore strongest for coding SNPs.

We also estimated the proportion of SNPs in each F_{ST} interval that were derived based on ancestral state of inferred from *A. cerana* reads. The influence of allelic ancestry on levels of variation was not straightforward. Averaging across all 131 fixed SNPs in the A vs MC comparison where the ancestral allele could be inferred, we observed greater reduction in African diversity when associated with the derived allele (20%; 71 SNPs) compared to the ancestral allele (10%; 60 SNPs). However, the levels of diversity were generally correlated between European and African rather than diametrically opposed. Episodes of adaptation in honeybees may therefore not always conform to classic sweep model where a new mutation rapidly becomes fixed in a population, but rather involve selection on standing variation[48].

We identified SNPs in the top 0.1% of F_{ST} in each comparison. We categorized these SNPs according to patterns of gene expression[43,44] and $CpG_{O/E}$. These highly differentiated SNPs were significantly enriched in high-CpG genes (p<0.001 across all three comparisons, Fisher's Exact Test) and in genes with biased overexpression in workers relative to the queens (p<0.05 across all three comparisons, Fisher's Exact Test) and the drones (p<0.001 across all three comparisons, Fisher's Exact Test). Genes with biased overexpression in queens were found to be somewhat enriched among the highly differentiated SNPs between Africa and Europe although the enrichment was not significant at the 5% level (p=0.10). Drone biased genes were significantly underrepresented (p<0.001 in all three comparisons). In order to test whether these effects were independent, we divided the caste gene data into high vs low CpG subsets. Among high CpG genes, there is still a significant enrichment for genes overexpressed in workers vs drones (p<0.001 in all three comparisons) but mainly trending enrichment among low CpG genes (p=0.04 in A vs MC, p=0.08 in C vs AM, p=0.10 in M vs AC). However, high CpG genes with worker-biased expression compared to queens are only significantly enriched in one of the comparisons (M vs AC; p=0.02). There is a significant enrichment of high CpG genes vs low CpG genes within most expression classes in all comparisons (p<0.05 for all comparisons except for queen genes in the C vs AM comparison where p=0.054). The CpG distributions of putatively selected genes in the African vs European comparison are shown in Supplementary Fig. 12a.

Our results are broadly consistent with a recent analysis of genomic variation and divergence by Harpur et al.[49]. This study identified genes under positive selection since the divergence of *A. mellifera* and *A. cerana* using an adapted McDonald-Kreitman test[50] that compares the ratio of nonsynonymous to synonymous changes in divergence and diversity data. They found that genes that are more taxonomically restricted have higher rates of adaptive evolution. This is consistent with our finding that caste-specific genes are enriched for highly differentiated SNPs whereas housekeeping genes (that are likely to be taxonomically conserved) are underrepresented. This implies that taxonomically restricted genes are more likely to be involved in the differences between geographic groups of *A. mellifera*. They also find that worker-biased genes show greater signs of positive selection on the lineage connecting *A. mellifera* and *A. cerana* compared with queen-biased and unbiased genes[49]. This is also consistent with our finding that worker-biased genes are more often highly differentiated between populations of *A. mellifera* and suggests that similar selective forces have been acting during the evolution of the *Apis* lineage to between populations of *A. mellifera*.

Function of genes under selection

The SNPs detected in each of the three contrasts were sorted according to F_{ST} and gene ontology (GO) analyses were carried out on the top 0.1% outlier SNPs: A vs MC (6179 SNPs with $F_{ST} \geq 0.89$; associated with 2983 genes), C vs AM (4799 SNPs with $F_{ST} \geq 0.95$; associated with 1912 genes) and M vs AC (6283 SNPs with $F_{ST} \geq 0.96$; 2375 genes). Each

SNP was associated with each nearest gene. A full list of genes and variants is found in the supplementary file 1. The available *Drosophila* orthologous of the honeybee genes associated with the SNPs were analyzed using the g:GOSt tool provided by the g:Profiler web service[51] searching for significantly over-represented functional GO terms. The version of the service used for the analyses has been archived as version r1185_e69_eg16. The gene lists were either analyzed in full or subdivided into CDS (all and only non-synonymous), intron or intergenic lists according to SNP position. We compared the lists of putatively selected genes to the background list of 7100 *Drosophila* genes that we had previously linked to the complete honeybee gene set and recorded the significantly enriched GO terms. The total number of enriched GO terms in each comparison and summary across all three comparisons are shown in Supplementary Fig. 12b and all terms, gene numbers and levels of significance are listed in the Supplementary File 2.

We used the REVIGO tool[52] to produce summaries of non-redundant GO terms grouped into functional categories. We first compiled a list of all of the GO terms that were enriched in each of the three comparisons (A vs MC, M vs AC, and C vs AM) in each of the three genomic locations (intron, exon, intergenic), in addition to combined lists taking the union of enriched GO terms across categories and across comparisons (16 lists in total). REViGO takes long lists of gene ontology and summarizes them by removing redundant GO terms and then groups GO terms into similar categories. We used the REViGO web server to produce categorized lists of each of these comparisons. We used the TreeMap option and then exported the list of grouped GO terms for each comparison in the biological process category. We edited this list by choosing terms for each group defined by REViGO so that they were consistent across categories and then displayed the results graphically (Fig. 4c) by summing the "uniqueness" score across GO terms in each category defined by REViGO. In general, the set of nonredundant terms found in each of the three comparisons was consistent. The intronic and intergenic categories predominantly contained genes involved in development and morphogenesis across all three comparisons, in addition to many terms related to cell signaling and response to external stimulus. The only significantly enriched biological process GO category in coding regions were microtubule-based movement (only found in the A vs MC comparison), which includes the dynein complex of sperm motor proteins. However, we also recovered significantly enriched immune defense-related phenotype (HP) and reaction (REAC) terms in the subclass of non-synonymous coding SNPs.

Many spermatogenic, sperm motor protein and axoneme structure genes in the honeybee genome contain non-synonymous variants (NSVs) at or close to fixation for alternative alleles in the population comparisons evaluated here. These include the seminal fluid enzyme esterase 6 (*Est-6*), which is associated with mating behavior in *D. melanogaster*[53], the *Pex16* gene, involved in spermatocyte maturation[54] and 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[55] whose orthologs are nearly all strongly expressed in *D. melanogaster* testis (data from FlyBase). This indicates that positive selection on sperm function has been important in recent honeybee evolution. The queen bee mates with up to twenty or more drones during her nuptial flight and stores a fraction of the sperm from each drone for the rest of her life, although the degree of polyandry differs between subspecies and is higher in African subspecies[18,56-59]. This mating system can be expected to induce sperm competition and rapid sperm evolution due to selection on male reproductive performance. Most of the NSVs are different between African and European bees, and could potentially account for the reproductive success of African over European drones when experimentally mated with queens under conditions of sperm competition[60].

We found several genes acting in pathways influencing cell proliferation, growth and development to carry strongly segregating or fixed NSVs, including, trithorax related (*trr*)[61], the epidermal growth factor receptor substrate 15 (*Eps-15*), adenylate kinase-2 (*Adk2*)[62], dumpy (*dp*)[63], papilin (*Ppn*)[64], pannier (*pnr*)[65], integrin alpha-PS2/inflated (*if*)[66], grunge (*Gug*)[67], and the sallimus/zormin (*sls*) complex[68]. These genes are linked to physical abnormalities in salivary glands, muscles, bristles, trachea, appendages, mouth parts or wings in Drosophila mutants and may affect phenotypes and reflect selection on morphological variation across the geographic range of honeybees.

A number of putatively selected genes are important signaling molecules, including neuropeptides, protein hormones and G-coupled receptors, which are also important in several key biological processes including behavior, development, feeding and reproduction[69]. African and European bees differ in several such genes, including the neural futsch and sulfateless (*sfl*) neurodevelopmental genes, the juvenile hormone inhibiting allatostatin (*ast*), a Ih-like dopamin regulator as well as several odorant receptors, that regulate feeding and social interactions. We detect NSVs in the neural *RhoGAP100F*, involved in axogenesis and signaling, which was earlier detected as fast evolving in highly social lineages[70]. Other genes involved in neural development with highly differentiated NSVs include the FMRFamide neuropeptide receptor (*fr*), the Neurexin IV (*Nrx-IV*) presynaptic protein, the wntless (*wls*) membrane gprotein, the silver (*svr*) carboxypeptidase gene, and ankyrin 2 (*Ank2*).

The insulin-vitellogenin signaling pathway is important for queen longevity and for worker labor division[71]. We found segregating SNPs in the exonic regions of key genes in this pathway including an large number of SNPs in the 3'-UTR of the insulin-like receptor (*InR*) and highly differentiated NSVs in *FOX*, vitellogenin (*Vg*) and the vitellogenin receptor (*yl*) genes. This suggests this pathway is dynamically evolving in bees and that queen longevity and fecundity differs among geographic regions. These mutations could explain the faster turnover of queen bees in African honeybee colonies[4]. It has been suggested that worker bees from temperate climates have increased capacity for vitellogenin storage, an adaptation that increases longevity of overwintering bees[72]. Interestingly, *Vg* also seems to be under strong positive selection along the lineage separating *A. mellifera* and *A. cerana*[49]. Examples of genes with highly differentiated SNPs are shown in Supplementary Fig. 13.

Honeybee immunity and response to infection range from the innate immune system at the cellular level to hygienic behaviors at the colony level. African bees differ from European bees in their capacity to tolerate and survive *Varroa* mite infestation[73,74] and likely other pathogens[75]. We detected highly differentiated NSVs in a large number of innate immune defense genes, many of which are part of the Toll signaling pathway, including pelle (*pll*) and the glucan gram-negative bacteria-binding protein 1-2 (*GNBP1*). This signal of selection was detected on three serine proteases with implied roles in honeybee immune response: *SP1*, *SP10/snake* (*snk*) and *SP49*[76]. *pll* was also detected as fast evolving in highly social lineages[70]. We also detected selection in the coding regions of *TEPB* in the JAK/STAT pathway and peptidoglycan-recognition protein LC (*PGRP-LC*) and Relish (*Rel*) from the Imd pathway[77–80]. The autophagous vacuolar proteins sorting 13 (*Vps13*) and two encapsulating glucose dehydrogenase (*Gld*) proteins (*GMCOX12* & *GMCOX13*) were also identified as being targets of selection, genes involved in cellular response and pathogen removal[81–83]. Among the top coding A vs MC SNPs, we also detect significant (P=0.012) enrichment for genes involved in platelet plug formation and coagulation, which are activated in the wound healing process and helps forming infection barriers in insects[84], including NSVs in Hemolectin (*Hml*)[85,86] and integrin alpha-PS2/inflated (*if*). The GO term "positive regulation of biosynthetic process of antibacterial peptides active against Gram-negative bacteria" is significantly enriched (p<0.00818) in genes containing high-F_{ST} nonsynonymous SNPs in the A vs MC comparison. This enrichment is due to the genes peptidoglycan-recognition protein LC (*PGRP-LC*) and Relish (*Rel*) from the Imd pathway and the gene *Iap2* (baculoviral IAP repeat-containing protein 4)[87].

References

1. Ruttner, F. *Biogeography and Taxonomy of Honeybees*. (Springer-Verlag, 1988).

2. Arias, M. C. & Sheppard, W. S. Molecular phylogenetics of honey bee subspecies (Apis mellifera L.) inferred from mitochondrial DNA sequence. *Mol. Phylogenet. Evol.* **5,** 557–66 (1996).

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. Hepburn, H. R. & Radloff, S. E. *Honeybees of Africa*. (Springer-Verlag, 1998).

5. Whitfield, C. W. *et al.* Thrice out of Africa: ancient and recent expansions of the honey bee, Apis mellifera. *Science* **314,** 642–5 (2006).

6. HGSC. Insights into social insects from the genome of the honeybee Apis mellifera. *Nature* **443,** 931–49 (2006).

7. 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).

8. Hepburn, H. R. & Radloff, S. E. *Honeybees of Asia*. (Springer, 2011).

9. Durbin, R. M. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).

10. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475,** 493–496 (2011).

11. LifeScope™ Genomic Analysis Software for SOLiD® Next Generation Sequencing. at <http://www.lifetechnologies.com/se/en/home/life-science/sequencing/next-generation-sequencing/solid-next-generation-sequencing/solid-next-generation-sequencing-data-analysis-solutions/lifescope-data-analysis-solid-next-generation-sequencing/lifescope-genomic-analysis-software-solid-next-generation-sequencing.html>

12. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).

13. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv:1207.3907* (2012). at <http://arxiv.org/abs/1207.3907>

14. Han, F., Wallberg, A. & Webster, M. T. From where did the Western honeybee (Apis mellifera) originate? *Ecol Evol* **2,** 1949–57 (2012).

15. Watterson, G. A. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7,** 256–276 (1975).

16. Baudry, E. *et al.* Relatedness among honeybees (Apis mellifera) of a drone congregation. *Proc. R. Soc. B-Biol. Sci.* **265,** 2009–2014 (1998).

17. 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).

18. 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 Sociaux* **47,** 364–370 (2000).

19. 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).

20. 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).

21. Arias, M. C. & Sheppard, W. S. Phylogenetic relationships of honey bees (Hymenoptera:Apinae:Apini) inferred from nuclear and mitochondrial DNA sequence data. *Mol. Phylogenet. Evol.* **37,** 25–35 (2005).

22. Haag-Liautard, C. *et al.* Direct estimation of per nucleotide and genomic deleterious mutation rates in Drosophila. *Nature* **445,** 82–85 (2007).

23. Eyre-Walker, A., Keightley, P. D., Smith, N. G. C. & Gaffney, D. Quantifying the Slightly Deleterious Mutation Model of Molecular Evolution. *Mol Biol Evol* **19,** 2142–2149 (2002).

24. Beye, M. *et al.* Exceptionally high levels of recombination across the honey bee genome. *Genome Res.* **16,** 1339–44 (2006).

25. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19,** 1655–1664 (2009).

26. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of Population Structure Using Multilocus Genotype Data. *Genetics* **155,** 945–959 (2000).

27. Zayed, A. & Whitfield, C. W. A genome-wide signature of positive selection in ancient and recent invasive expansions of the honey bee Apis mellifera. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 3421–6 (2008).

28. Smith, D. R. *et al.* Geographical Overlap of Two Mitochondrial Genomes in Spanish Honeybees (Apis mellifera iberica). *J. Hered.* **82,** 96 –100 (1991).

29. Garnery, L., Mosshine, E. H., Oldroyd, B. P. & Cornuet, J. M. Mitochondrial-DNA Variation in Moroccan and Spanish Honey-Bee Populations. *Mol. Ecol.* **4,** 465–471 (1995).

30. Arias, M. C., Rinderer, T. E. & Sheppard, W. S. Further characterization of honey bees from the Iberian Peninsula by allozyme, morphometric and mtDNA haplotype analyses. *J. Apic. Res.* **47,** 188–196 (2006).

31. Miguel, I., Iriondo, M., Garnery, L., Sheppard, W. S. & Estonba, A. Gene flow within the M evolutionary lineage of Apis mellifera: role of the Pyrenees, isolation by distance and post-glacial re-colonization routes in the western Europe. *Apidologie* **38,** 15 (2007).

32. Alburaki, M. *et al.* A fifth major genetic group among honeybees revealed in Syria. *BMC Genet.* **14,** 117 (2013).

33. Reynolds, J., Weir, B. S. & Cockerham, C. C. Estimation of the Coancestry Coefficient: Basis for a Short-Term Genetic Distance. *Genetics* **105,** 767–779 (1983).

34. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).

35. Wakeley, J. *Coalescent theory: an introduction*. (Roberts & Co. Publishers, 2009).

36. Durand, E. Y., Patterson, N., Reich, D. & Slatkin, M. Testing for Ancient Admixture between Closely Related Populations. *Mol. Biol. Evol.* **28,** 2239–2252 (2011).

37. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328,** 710–722 (2010).

38. Augustin, L. *et al.* Eight glacial cycles from an Antarctic ice core. *Nature* **429,** 623–628 (2004).

39. Hewitt, G. The genetic legacy of the Quaternary ice ages. *Nature* **405,** 907–913 (2000).

40. Cornuet, J. M. & Garnery, L. Mitochondrial DNA variability in honeybees and its phylogeographic implications. *Apidologie* **22,** 16 (1991).

41. Rosenberg, N. A. & Nordborg, M. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat. Rev. Genet.* **3,** 380–390 (2002).

42. 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. U. S. A.* **106,** 11206–11 (2009).

43. 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).

44. 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).

45. Weir, B. S. & Cockerham, C. C. Estimating F-statistics for the analysis of population structure. *Evolution* **38,** 1358–1370 (1984).

46. Kent, C. F., Minaei, S., Harpur, B. A. & Zayed, A. Recombination is associated with the evolution of genome structure and worker behavior in honey bees. *Proc. Natl. Acad. Sci.* **109,** 18012–18017 (2012).

47. Duret, L. & Galtier, N. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu. Rev. Genomics Hum. Genet.* **10,** 285–311 (2009).

48. Messer, P. W. & Petrov, D. A. Population genomics of rapid adaptation by soft selective sweeps. *Trends Ecol. Evol.* **28,** 659–669 (2013).

49. Harpur, B. A. *et al.* Population genomics of the honey bee reveals strong signatures of positive selection on worker traits. *Proc. Natl. Acad. Sci.* 201315506 (2014). doi:10.1073/pnas.1315506111

50. Eilertson, K. E., Booth, J. G. & Bustamante, C. D. SnIPRE: Selection Inference Using a Poisson Random Effects Model. *PLoS Comput Biol* **8,** e1002806 (2012).

51. Reimand, J., Kull, M., Peterson, H., Hansen, J. & Vilo, J. g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.* **35,** W193–W200 (2007).

52. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PLoS ONE* **6,** e21800 (2011).

53. 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).

54. 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).

55. Mencarelli, C., Lupetti, P. & Dallai, R. New insights into the cell biology of insect axonemes. *Int. Rev. Cell Mol. Biol.* **268,** 95–145 (2008).

56. Hernández-García, R., Rúa, P. de la & Serrano, J. Mating frequency in Apis mellifera iberiensis queens. *J. Apic. Res.* **48,** 121–125 (2009).

57. Estoup, A., Solignac, M. & Cornuet, J.-M. Precise Assessment of the Number of Patrilines and of Genetic Relatedness in Honeybee Colonies. *Proc. Biol. Sci.* **258,** 1–7 (1994).

58. Schlüns, H., Schlüns, E. A., van Praagh, J. & Moritz, R. F. A. Sperm numbers in drone honeybees (*Apis mellifera*) depend on body size. *Apidologie* **34,** 577–584 (2003).

59. Kraus, B., Neumann, P., Praagh, J. van, Moritz, R. F. A. & Neumann, P. *Sperm limitation and the evolution of extreme polyandry in honeybees (Apis mellifera L.).* (2004).

60. 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* **34,** 111–120 (2003).

61. Kanda, H., Nguyen, A., Chen, L., Okano, H. & Hariharan, I. K. The Drosophila Ortholog of MLL3 and MLL4, trithorax related, Functions as a Negative Regulator of Tissue Growth. *Mol. Cell. Biol.* **33,** 1702–1710 (2013).

62. Chen, R.-P. *et al.* Adenylate kinase 2 (AK2) promotes cell proliferation in insect development. *BMC Mol. Biol.* **13,** 31 (2012).

63. Wilkin, M. B. *et al.* Drosophila dumpy is a gigantic extracellular protein required to maintain tension at epidermal-cuticle attachment sites. *Curr. Biol. CB* **10,** 559–567 (2000).

64. Kramerova, I. A. *et al.* Papilin in development; a pericellular protein with a homology to the ADAMTS metalloproteinases. *Dev. Camb. Engl.* **127,** 5475–5485 (2000).

65. Calleja, M. *et al.* Generation of medial and lateral dorsal body domains by the pannier gene of Drosophila. *Dev. Camb. Engl.* **127,** 3971–3980 (2000).

66. Walsh, E. P. & Brown, N. H. A screen to identify Drosophila genes required for integrin-mediated adhesion. *Genetics* **150,** 791–805 (1998).

67. Erkner, A. *et al.* Grunge, related to human Atrophin-like proteins, has multiple functions in Drosophila development. *Dev. Camb. Engl.* **129,** 1119–1129 (2002).

68. Burkart, C. *et al.* Modular proteins from the Drosophila sallimus (sls) gene and their expression in muscles with different extensibility. *J. Mol. Biol.* **367,** 953–969 (2007).

69. Nässel, D. R. & Winther, A. M. E. Drosophila neuropeptides in regulation of physiology and behavior. *Prog. Neurobiol.* **92,** 42–104 (2010).

70. Woodard, S. H. *et al.* Genes involved in convergent evolution of eusociality in bees. *Proc. Natl. Acad. Sci.* **108,** 7472–7477 (2011).

71. Corona, M. *et al.* Vitellogenin, juvenile hormone, insulin signaling, and queen honey bee longevity. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 7128–7133 (2007).

72. Amdam, G. V. *et al.* Higher vitellogenin concentrations in honey bee workers may be an adaptation to life in temperate climates. *Insectes Sociaux* **52,** 316–319 (2005).

73. Fazier, M. *et al.* A scientific note on *Varroa destructor* found in East Africa; threat or opportunity? *Apidologie* **41,** 463–465 (2010).

74. Martin, S. J. & Medina, L. M. Africanized honeybees have unique tolerance to Varroa mites. *Trends Parasitol.* **20,** 112–114 (2004).

75. Fries, I. & Raina, S. American Foulbrood and African Honey Bees (Hymenoptera: Apidae). *J. Econ. Entomol.* **96,** 1641–1646 (2003).

76. Zou, Z., Lopez, D. L., Kanost, M. R., Evans, J. D. & Jiang, H. Comparative analysis of serine protease-related genes in the honey bee genome: possible involvement in embryonic development and innate immunity. *Insect Mol. Biol.* **15,** 603–614 (2006).
77. Rus, F. *et al.* Ecdysone triggered PGRP-LC expression controls Drosophila innate immunity. *EMBO J.* **32,** 1626–1638 (2013).
78. Schmidt, R. L., Trejo, T. R., Plummer, T. B., Platt, J. L. & Tang, A. H. Infection-induced proteolysis of PGRP-LC controls the IMD activation and melanization cascades in Drosophila. *FASEB J.* **22,** 918–929 (2008).
79. De Gregorio, E., Spellman, P. T., Tzou, P., Rubin, G. M. & Lemaitre, B. The Toll and Imd pathways are the major regulators of the immune response in Drosophila. *EMBO J.* **21,** 2568–2579 (2002).
80. Silverman, N. *et al.* A Drosophila IkB kinase complex required for Relish cleavage and antibacterial immunity. *Genes Dev.* **14,** 2461–2471 (2000).
81. Deretic, V., Saitoh, T. & Akira, S. Autophagy in infection, inflammation and immunity. *Nat. Rev. Immunol.* **13,** 722–737 (2013).
82. Antúnez, K. *et al.* Immune suppression in the honey bee (Apis mellifera) following infection by Nosema ceranae (Microsporidia). *Environ. Microbiol.* **11,** 2284–2290 (2009).
83. Yang, X. & Cox-Foster, D. L. Impact of an ectoparasite on the immunity and pathology of an invertebrate: evidence for host immunosuppression and viral amplification. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 7470–7475 (2005).
84. Galko, M. J. & Krasnow, M. A. Cellular and Genetic Analysis of Wound Healing in Drosophila Larvae. *PLoS Biol.* **2,** (2004).
85. Goto, A. *et al.* A Drosophila haemocyte-specific protein, hemolectin, similar to human von Willebrand factor. *Biochem. J.* **359,** 99–108 (2001).
86. Lesch, C. *et al.* A role for Hemolectin in coagulation and immunity in Drosophila melanogaster. *Dev. Comp. Immunol.* **31,** 1255–1263 (2007).
87. Valanne, S., Kleino, A., Myllymäki, H., Vuoristo, J. & Rämet, M. Iap2 is required for a sustained response in the Drosophila Imd pathway. *Dev. Comp. Immunol.* **31,** 991–1001 (2007).