



SCIENTIFIC REPORTS



OPEN

Whole genome SNP-associated signatures of local adaptation in honeybees of the Iberian Peninsula

Dora Henriques^{1,2}, Andreas Wallberg ³, Julio Chávez-Galarza^{1,4}, J. Spencer Johnston⁵, Matthew T. Webster³ & M. Alice Pinto ¹

The availability of powerful high-throughput genomic tools, combined with genome scans, has helped identifying genes and genetic changes responsible for environmental adaptation in many organisms, including the honeybee. Here, we resequenced 87 whole genomes of the honeybee native to Iberia and used conceptually different selection methods (Samβada, LFMM, PCAdapt, iHS) together with *in silico* protein modelling to search for selection footprints along environmental gradients. We found 670 outlier SNPs, most of which associated with precipitation, longitude and latitude. Over 88.7% SNPs laid outside exons and there was a significant enrichment in regions adjacent to exons and UTRs. Enrichment was also detected in exonic regions. Furthermore, *in silico* protein modelling suggests that several non-synonymous SNPs are likely direct targets of selection, as they lead to amino acid replacements in functionally important sites of proteins. We identified genomic signatures of local adaptation in 140 genes, many of which are putatively implicated in fitness-related functions such as reproduction, immunity, olfaction, lipid biosynthesis and circadian clock. Our genome scan suggests that local adaptation in the Iberian honeybee involves variations in regions that might alter patterns of gene expression and in protein-coding genes, which are promising candidates to underpin adaptive change in the honeybee.

In the current context of a global human-mediated environmental crisis, the long-standing goal of uncovering the genetic basis of adaptation has never been so important. Recent technological advances allow for major steps towards that goal. Increasingly powerful high-throughput sequencing and computational technologies, coupled with increasingly sophisticated analytical tools, have changed the scale of analysis from limited genomic regions and few loci to whole genomes, allowing thereby detection of signatures of selection at an unprecedented resolution and depth.

Most genome-wide analytical tools detect selection by searching for unusual patterns of genetic variation positing that population demographic history affects variation across all loci while natural selection operates at specific loci^{1–4}. Known as outlier tests, selection footprints are sought by scanning genomes using a population-based differentiation measure such as F_{ST} ^{5,6} or by an individual-based approach centred on Bayesian factor models⁷. Another class of increasingly popular analytical tools, known as genetic-environment association (GEA) methods, identify selection by finding strong associations between genetic and environmental data^{8–12}. By uncovering loci that are directly or indirectly correlated with the environmental factors, GEA methods can potentially identify selective pressures driving local adaptation^{13–15}. A drawback of both classes of tools is that demographic processes and complex spatial structuring may create patterns resembling selection, leading to false positives^{16–18}. However, a recently developed approach controls for population structure using latent factors estimated considering the statistical model and the data simultaneously. This approach has been incorporated into some outlier tests (e. g. the Bayesian factor model of PCAdapt⁷) and GEA methods (e. g. latent factor mixed model, LFMM⁹).

Studies using whole-genome scans have employed these analytical tools to identify hundreds of regions under selection in many model and non-model organisms^{19–29}. This study further contributes to the rapidly growing

¹Mountain Research Centre (CIMO), Polytechnic Institute of Bragança, Campus de Sta. Apolónia, 5300-253, Bragança, Portugal. ²Centre of Molecular and Environmental Biology (CBMA), University of Minho, Campus de Gualtar, 4710-057, Braga, Portugal. ³Department of Medical Biochemistry and Microbiology, Science for Life Laboratory, Uppsala University, SE -751 23, Uppsala, Sweden. ⁴Instituto Nacional de Innovación Agraria (INIA), Av. La Molina 1981, La Molina, Lima, Peru. ⁵Department of Entomology, Texas A&M University, College Station, TX, 77843-2475, USA. Correspondence and requests for materials should be addressed to M.A.P. (email: apinto@ipb.pt)

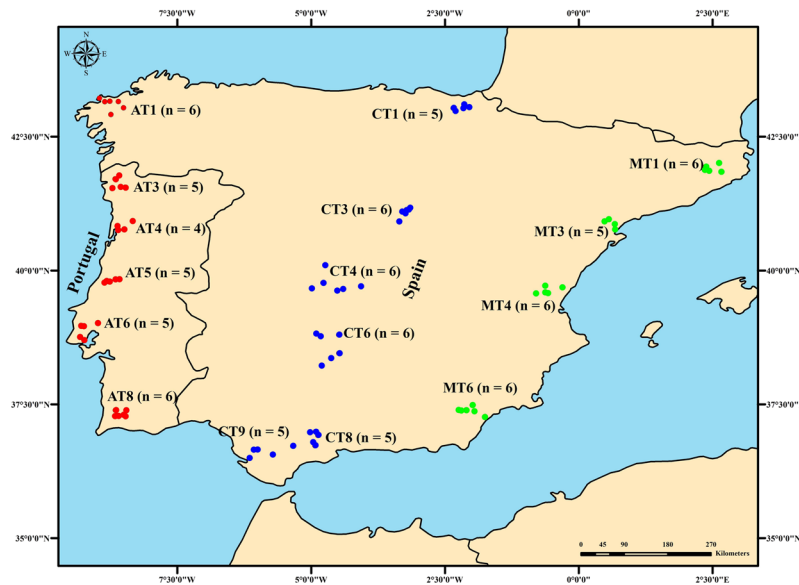


Figure 1. Location of sampling sites. The samples were distributed across three transects in the Iberian Peninsula, represented by three different colours: Atlantic in red (AT; N = 31), Central in blue (CT; N = 33), and Mediterranean in green (MT; N = 23). Each dot represents a single colony and apiary. Sampling site codes (AT1 to MT6) correspond to those reported by Chávez-Galarza, *et al.*⁴¹.

list of organisms by helping uncover genetic pathways underlying local adaptation of one of the most diverse and evolutionarily complex honeybee subspecies, the Iberian honeybee (hereafter IHB), *Apis mellifera iberiensis*.

The honeybee (*Apis mellifera* L.) evolved into 31 currently recognized subspecies^{30–34}, which have been grouped into four main evolutionary lineages: Northern and Western European, M; Southeastern European, C; African, A; and Middle Eastern, O³⁰. In this wide range of diversity, the M-lineage IHB is one of the most intriguing subspecies, exhibiting complex patterns of clinal variation as have many other organisms that evolved in the Iberian glacial refuge (reviewed by Weiss and Ferrand³⁵). Genetic surveys of the IHB have suggested that while evolutionarily neutral processes have played an important role in shaping the sharp northeastern-southwestern Iberian cline^{36–40}, selection is a force that cannot be ignored⁴¹. Iberia possesses high physiographic complexity, with several large mountain ranges, and due to its geographical position is under the influence of both the North Atlantic and the Mediterranean seas. These features have shaped a diverse array of climates (including desert, Mediterranean, Alpine, and Atlantic) and plant communities with variable flowering peaks to which the IHB had to adapt.

A previous selection scan of the IHB using an array of 383 SNPs (single nucleotide polymorphisms) identified 34 putatively adaptive SNPs located in genes involved in vision, xenobiotic detoxification, and innate immune response⁴¹. However, the 383 SNPs were widely spaced, and given the unusually high recombination rate in honeybees⁴² genomic regions important in local adaptation have certainly been missed, as suggested by whole-genome studies of other subspecies^{33,43–47}. In the present study, we employed a combination of outlier and GEA methods to identify genome-wide signatures of selection from 87 whole-genome sequences, thereby expanding the SNP-array scan of Chávez-Galarza, *et al.*⁴¹ by over 3 orders of magnitude (3367 fold). We approached local adaptation in the IHB by addressing the following questions: Does adaptation arise from mutations that change amino acids? Which genes are responsible for adaptation to different environments? Which environmental factors might act as selective pressures in IHBs? In answering these questions, major insights will be gained toward understanding the genetic pathways used by the IHB to adapt to the broad range of Iberian environments.

Results

A total of 1,289,449 SNPs were retained, after the filtering process and using a minor allele frequency (MAF) > 0.05, for 87 resequenced IHBs sampled from across the Iberian range (Fig. 1). Of these, 670,738 SNPs were located in intergenic regions (120,301 in intergenic regions < 2 Kb of exons, 37,058 in intergenic regions < 1 Kb upstream of exons), 557,334 in introns (23,092 < 50 bp of exons), 18,841 in UTRs (untranslated region), and 42,536 in exons (Supplementary Table S1). The average physical distance between SNPs was 170.262 bp varying between 1 bp and 136,266 bp (Supplementary Fig. S1).

Population Structure. Population structure and demographic history can create genomic patterns that mimic selection. Accordingly, population structure was analysed to prevent discovery of false positives^{16–18,48}. The genetic structure was inferred from the 1,289,449 SNPs with sNMF and PCAdapt, which identified one and two optimal number of clusters (K), respectively (Supplementary Fig. S2). Incongruent optimal K values can be obtained by different methods⁴⁹, especially in the presence of low levels of population differentiation⁵⁰, which is the case of the IHB with a global $F_{ST} = 0.021$. Despite the optimal $K = 1$ obtained by sNMF, further partitioning of

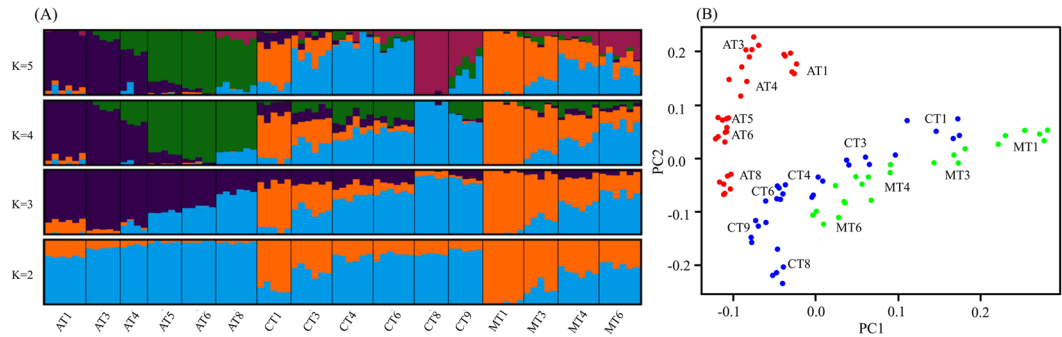


Figure 2. Population structure of *A. m. iberiensis*. **(A)** Structure estimated by sNMF from $K=2$ to $K=5$. The 16 sampling sites are arranged from north (AT1, CT1, MT1) to south (AT8, CT9, MT6) in each of the three transects. Plots represent each of the 87 individuals by a vertical bar partitioned into coloured segments (clusters) corresponding to membership proportions (Y-axis: 0-1) in each cluster. Vertical black lines separate the 16 sampling sites. **(B)** Score plot displaying the latent factors of each individual honeybee in PC1 and PC2 for $K=2$. Each colour represents a different transect.

the genome revealed a clinal pattern of variation, with the northern populations of the central and Mediterranean transects carrying an important genomic component assigned to the orange cluster (0.65 for $K=2$, Fig. 2A). This component decreased gradually towards the south and is absent in most Atlantic populations. Greater K values ($K \geq 3$) highlight the distinctiveness of the Atlantic populations. The clinal pattern of variation in the Mediterranean (MT) and central populations (CT) is captured by PC2, with the distinct Atlantic populations (AT) captured by the PC1 generated by PCAdapt fast (Fig. 2B). These genome-wide results confirm the Iberian cline captured by the 383 SNPs, and the claim that modern beekeeping has not disrupted the natural variation pattern in IHBs⁴⁰.

Signatures of Local Adaptation. Genetic-Environment Associations (GEA). To identify potential selective pressures driving local adaptation in the IHB, the GEA methods Samβada and LFMM were employed in the genome scan. A total of 38,683,470 univariate models (1,289,449 SNPs \times 2 alleles \times 15 environmental variables) were processed by Samβada. Over 1,305 SNPs were identified as outliers (false discovery rate, FDR < 0.05; Supplementary Table S2). The most frequently associated environmental variables were longitude (long; 1,071 models, 31%), precipitation in August (prec8; 758 models, 22%), May (prec5; 368 models, 11%), and January (prec1; 336 models, 10%). The 12 top-ranked models (Gscore > 50) identified 6 SNPs, which were located in genes GB40077 (1 SNP), GB54460 (1 SNP), GB45499 (2 SNPs) and GB48105 (2 SNP). Of the six SNPs, three SNPs were non-synonymous, with the strongest (Gscore = 59.0) tagging GB40077 and the other two tagging GB45499 (see further details in the protein modelling section), two were located in introns in the immediate vicinity of exons (between 85 and 225 bp), and one was in a synonymous position (Supplementary Table S2). The SNPs marking GB40077, GB54460 and GB45499 were associated with longitude whereas the two SNPs located in gene GB48105 were associated with precipitation in January.

A total of 1,416 (FDR < 0.05), 360 (FDR < 0.02), and 220 SNPs (FDR < 0.01) were identified by the LFMM method (Supplementary Table S3 and Fig. S3). The strongest 21 SNPs (defined by a cut-off level of $-\log_{10}(q\text{-value}) > 4$) were located in introns (11 in GB46620, 3 in GB43005, 1 in GB4810, 1 in GB54460), 1 in UTRs (GB48105), and 3 in exons (1 non-synonymous in GB46620, 1 non-synonymous in GB40077, 1 synonymous in GB48105). A single SNP mapped to an intergenic region, although close to a gene (207 bp upstream of GB46621). Most SNPs were associated with latitude (11 in GB46620, 3 in GB43005, 1 in GB46621) and/or precipitation in May (12 in GB46620, 1 in GB46621). The variables precipitation in January and longitude were associated with only 3 (GB48105) and 2 (1 in GB40077, 1 in GB54460) SNPs, respectively.

A total of 598 SNPs overlapped between Samβada and LFMM (Supplementary Tables S4 and S5). These SNPs mapped to 126 genes and 99 intergenic regions. The variables precipitation in August and precipitation in May showed the greatest number of associated SNPs (227 and 152, respectively; Table 1, Fig. 3, Supplementary Tables S2 and S3), although longitude (113), latitude (102) and precipitation in January (101) were also predominant variables (Table 1). The variable latitude shared 52% of the SNPs with precipitation in August whereas longitude shared 35.40% of the SNPs with precipitation in January and 14.16% with cloud cover in April (cld4) (Supplementary Table S6).

There was an enrichment of SNPs detected by both GEA methods in exons (P-value < 2.20×10^{-16}), UTRs (P-value = 8.8×10^{-4}), introns < 50 bp of exons (P-value = 8.01×10^{-5}), and intergenic regions < 1 Kb upstream of exons (P-value = 2.278×10^{-5} , χ^2 test).

PCAdapt fast. For further cross-validating selection and reducing detection of spurious signals, we combined the GEA methods with the differentiation-based PCAdapt fast. A total of 285 outlier SNPs were identified by PCAdapt (FDR < 0.05; Supplementary Fig. S4 and Tables S5 and S7), of which 266 (93.3%) were cross-detected by the GEA methods (Fig. 4). From the 285 SNPs, 84 were located in 36 intergenic regions and 201 in 61 genes. The genes containing the highest number of SNPs were GB49881 (40), GB49882 (27), and GB46620 (18).

Environmental variables	LFMM	Samβada	Overlapping
Precipitation August	124	152	227
Precipitation May	444	32	152
Longitude	0	423	113
Latitude	283	56	102
Precipitation January	63	67	101
Insolation April	18	55	50
Cloud cover April	165	8	32
Temperature min. January	58	27	18
Cloud cover July	24	4	10
Relative humidity January	27	21	6
Temperature min. June	1	16	6
Land cover	26	3	6
Relative humidity June	19	20	4
Relative humidity March	111	2	0
Altitude	0	3	0

Table 1. Environmental variables and number of associated SNPs identified exclusively by LFMM or Samβada and simultaneously by both methods.

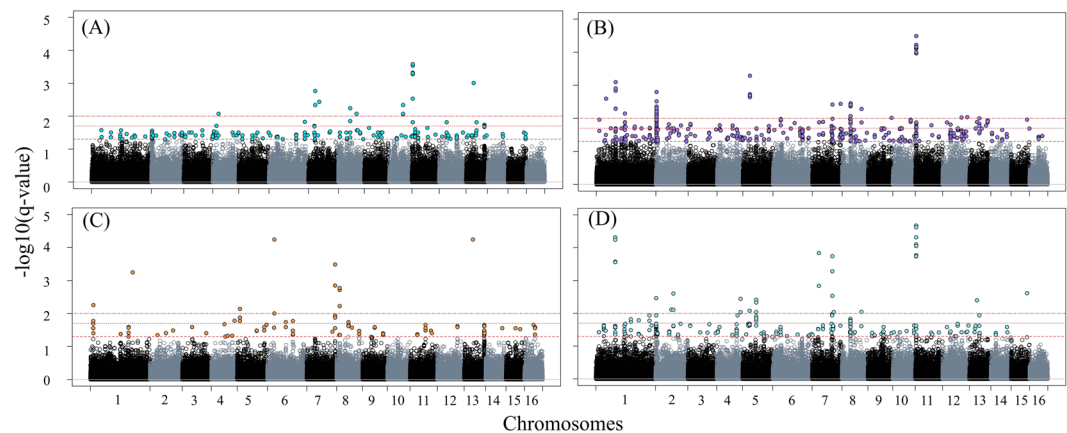


Figure 3. LFMM Manhattan plots. The plots represent the genome-wide distribution of significance values $-\log_{10}(\text{q-value})$ obtained by LFMM for the environmental variables with the strongest associations. (A) precipitation in August: 351 SNPs, (B) precipitation in May: 596 SNPs, (C) longitude: 113 SNPs, (D) latitude: 385 SNPs. The red lines indicate FDR values of 0.05, 0.02 and 0.01.

Putative targets of selection identified by PCAdapt fast were enriched in exons ($P\text{-value} < 2.20 \times 10^{-16}$, χ^2 test), in intergenic regions < 1 Kb upstream ($P\text{-value} < 2.20 \times 10^{-16}$, χ^2 test) of exons, and in introns < 50 bp of exons ($P\text{-value} = 3.1 \times 10^{-8}$, χ^2 test).

The Strongest Candidate SNPs. A total of 670 SNPs were detected by at least two selection methods, 11.3% were located in exons (41 non-synonymous and 35 synonymous SNPs), 3.0% in UTRs (20 SNPs), 46.1% in introns (309 SNPs, of which 28 were < 50 bp of exons), 18.7% in intergenic regions adjacent to (2–2,000 bp; 125 SNPs, of which 42 were < 1 Kb upstream) exons and 21% distant from (2,023–188,208 bp; 140 SNPs) exons.

The 670 SNPs exhibited |iHs| (integrated haplotype score) values ranging from 0.006 to 7.2 (Supplementary Tables S4 and S5). A total of 150 SNPs were strong candidates for recent ongoing selection as they showed a |iHs| > 2 (Supplementary Table S5). The two top-ranked SNPs displayed a |iHs| > 7 , standing out by a remarkably strong selection signature. One of these two is located 834 bp upstream of the undescribed gene GB54883 and the other is a longitude-associated non-synonymous SNP located in GB55263 (see further details in the protein modelling section).

The great majority (405 SNPs, 60.4%) of the 670 SNPs were located in exons, introns and UTRs of 140 genes. Of these, 8 genes carried > 10 SNPs (Supplementary Tables S4 and S5), mostly associated with precipitation in May and precipitation in August (Table 2). The aforementioned GB49881, GB49882, GB46620, and GB43005 are amongst the 8 genes and are highlighted by possessing 29, 19, 18 and 13 SNPs, respectively. Four genes were tagged by non-synonymous SNPs with GB48703 and GB48709 harbouring the most (Table 2).

<i>A. mellifera</i> gene	# SNPs	SNPs distribution across genomic regions	Environmental variables
GB49881	29	29 Intronic	Long, prec1, cld4
GB48698	20	1 Exonic (syn), 16 intronic, 3 UTR	Prec8
GB49882	19	5 Exonic (syn), 14 intronic	Long, cld4
GB46620	18	2 Exonic (non-syn), 16 intronic	Lat, prec1, prec5, prec8, ins4
GB48709	13	3 Exonic (non-syn), 1 exonic (syn), 9 intronic	Lat, prec5, prec8
GB43005	13	13 Intronic	Lat, prec1, prec5, tmin1, tmin6, ins4
GB48703	12	3 Exonic (non-syn), 1 exonic (syn), 6 intronic, 2 UTR	Prec5, prec8
GB48699	11	1 Exonic (non-syn), 10 intronic	Prec8

Table 2. Candidate genes containing more than 10 SNPs detected concurrently by at least two selection methods. Genes marked in bold carry SNPs that were cross-detected by iHS and/or PCAdapt. The correlated environmental variables are longitude (long); latitude (lat); cloud cover in April (cld4); insolation in April (ins4); precipitation in January (prec1), May (prec5) and August (prec8); minimum temperature in January (tmin1) and (tmin6).

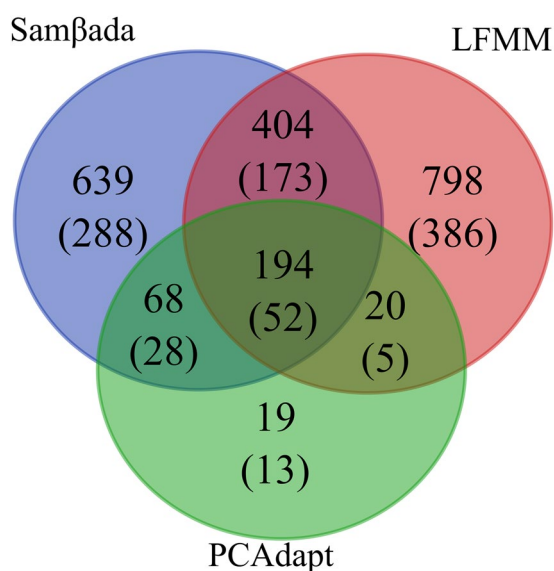


Figure 4. Overlapping SNPs identified by the three genome-scan methods. Numbers in the intersection regions represent overlapping SNPs among two or three methods. Numbers in parentheses show the corresponding genomic regions harbouring the SNPs.

The highest stringency cross-validation based on the three methods (Samβada, LFMM and PCAdapt fast) identified 194 overlapping SNPs (Fig. 4 and Supplementary Tables S4 and S5). These were located in 39 genes, including 6 of the 8 genes containing >10 SNPs. From the 194 SNPs, 68 displayed elevated |iHS| values (>2.0) representing 14 genes and 11 intergenic regions (Table 3). The genes with the highest number of SNPs and the uppermost |iHS| values were GB49881 (28 SNPs, |iHS| > 3.0) and GB49882 (6 SNPs, |iHS| > 5.4). Interestingly, these genes share transcript sequence associated with the SHAW protein, for which there are five alternative variants, and are only 1,864 bp apart. This remarkably short intergenic region contained 15 SNPs detected by at least two methods with |iHS| > 1.7 (Supplementary Table S5).

There was an enrichment of the 670 SNPs in exons (P-value < 2.20 × 10⁻¹⁶), UTRs (P-value = 0.0018), introns < 50 bp of exons (P-value = 6.393 × 10⁻⁶), and intergenic regions < 1 Kb upstream of exons (P-value = 2.73 × 10⁻⁷, χ² test).

Protein Modelling. To understand how SNPs causing amino acid changes could interfere with protein function, the 3D structure and stability were predicted for the different variants. A total of 41 non-synonymous SNPs were detected by at least two selection methods. The 41 SNPs were located in 29 genes. Protein prediction was available for only 11 of the 29 genes and 4 genes contained SNPs outside the 3D model (Supplementary Table S8). The remaining 7 genes (GB40077, GB45499, GB47279, GB48707, GB49875, GB51396, GB55263) were translated into a total of 37 protein variants (Supplementary Table S9). The gene GB49875 was the least diverse, with 3 variants, and gene GB40077 was the most diverse, with 9 variants (Fig. 5, Supplementary Fig. S5, Table S9). Most of these protein variants exhibited lower energy minimization than the reference (14 variants) and values

Gene	# SNPs	Genomic position	Environmental variables	Putative function
GB49881	28	Intronic	Long, prec1, cld4	Undescribed
GB49882	6	Intronic	Cld4	Sleep
GB49899	4	Intronic	Long, prec1	Pdz (post-synaptic density) domain
GB48696	4	Non-syn, syn, intergenic (<1235 bp)	Prec5, prec8	Inter-male aggressive behavior
GB48694	3	Intronic	Prec5	Undescribed
GB48105	2	Intronic (8 bp), UTR	Long, prec1	Neurogenesis
GB49874	2	Intergenic (<2060 bp)	Long, prec1	Undescribed
GB48702	2	Intronic (61 bp), intergenic (1, 111 bp)	Lat, prec5, prec8	Organism reproduction
GB51286	2	Intergenic (<17, 754 bp)	Long, prec1	Undescribed
GB48701	2	Intergenic (<804 bp)	Lat, prec5, prec8	Undescribed
GB48697	2	Syn, Intergenic (678 bp)	Lat, prec5, prec8	Undescribed
GB51427	1	Intergenic (952 pb)	Long	Response to fungus
GB49878	1	Intergenic (495 bp)	Long, prec1	Response to DDT
GB47226	1	Syn	Long, prec8	Undescribed
GB49879	1	Intronic (204 bp)	Long, prec1	Sleep
GB47281	1	Syn	Long, prec8	Ovarian nurse cell to oocyte transport;
GB47279	1	Non-syn	Long, prec8	Response to insecticide
GB48706	1	Intergenic (30 bp)	Prec5	Axoneme assembly
GB51401	1	Intergenic (180 bp)	Long	ATP-dependent RNA helicase activity
GB51396	1	Non-syn	Long	Oxidoreductase activity,
GB51422	1	Syn	Long	Undescribed
GB44109	1	Intergenic (2182 bp)	Prec1	Oxidation-reduction process

Table 3. Genomic information, and associated environmental variables, of candidate genes cross-detected by Sam β ada, LFMM, PCAdapt and $|iHs| > 2$. Putative functions were summarized from FLYBASE. The correlated environmental variables are longitude (long); cloud cover in April (cld4); precipitation in January (prec1), May (prec5) and August (prec8).

of Gibbs-free energy ($\Delta\Delta G$) > 0 (15 variants), with the highest $\Delta\Delta G$ values displayed by variants C and E of gene GB47279 (3.94 Kcal/mol and 2.29 Kcal/mol, respectively), indicating that the variants are less stable than the reference protein.

The non-synonymous SNPs tagging GB47279, GB48707, GB49875 and GB51396 produced amino acid changes outside of sites described as functionally important (see 3D structures in Supplementary Fig. S5). In contrast, GB40077, GB55263, and GB45499 contained non-synonymous SNPs that led to replacement of amino acids within or in the close vicinity of a functionally important site of the protein (Fig. 5). GB40077 and GB55263 encode proteins involved in lipid biosynthesis whereas GB45499 encodes a transport protein (Supplementary Table S4).

The single SNP detected inside the 3D prediction of GB40077 led to replacement of a proline (non-polar with restricted flexibility) by a serine (polar with low-flexibility) in position 363 (Fig. 5; Supplementary Table S9). GB55263 was also tagged by a single but strong SNP ($|iHs| = 7.05$; Supplementary Table S5), which causes a substitution of a threonine (polar with low flexibility) by an isoleucine (non-polar with moderate flexibility) at position 215 (Fig. 5; Supplementary Table S9). GB45499 was tagged by two non-synonymous SNPs. Of the two amino acid substitutions in this gene, only amino acid 74 was located inside the 3D prediction (Fig. 5). At position 74, the SNP leads to replacement of histidine (positive with moderate flexibility) by tyrosine (polar with moderate flexibility; Supplementary Table S9).

The geographical patterns exhibited by the variants of the amino acid under selection are shown in Fig. 5 and Supplementary Fig. S5. While variation of genes GB45499, GB55263, GB47279, GB48707, and GB51396 is oriented along a northeastern-southwestern axis, GB40077 and GB49875 display an eastern-western pattern with one the forms of the amino acid mostly confined to the Atlantic side of Iberia.

Gene Ontology and Annotation. The power of the gene ontology (GO) analysis for uncovering the biological significance of candidate regions identified in whole-genome selection scans depends on the number of annotated genes available for the focal organism⁵¹. Of the 140 candidate genes identified here by at least two selection methods, only 109 were retrieved from the DAVID database. Hence, the GO analysis should be interpreted with caution as it may reflect a biased representation of candidate genes and miss biological functions. The 109 genes showed a significant enrichment (P-value < 0.05 , before Bonferroni correction) for 6 functional terms (Supplementary Table S10), of which 4 formed one cluster (enrichment score = 2.58). The 4 terms were related with *membrane* of which only one (*integral component of membrane*) was significant after the Bonferroni correction. The remaining 2 functional terms (*olfactory learning* and *lucose/ribitol dehydrogenase*) were not clustered. While *olfactory learning* only included 2 genes, the fold enrichment was remarkably high (84.79).

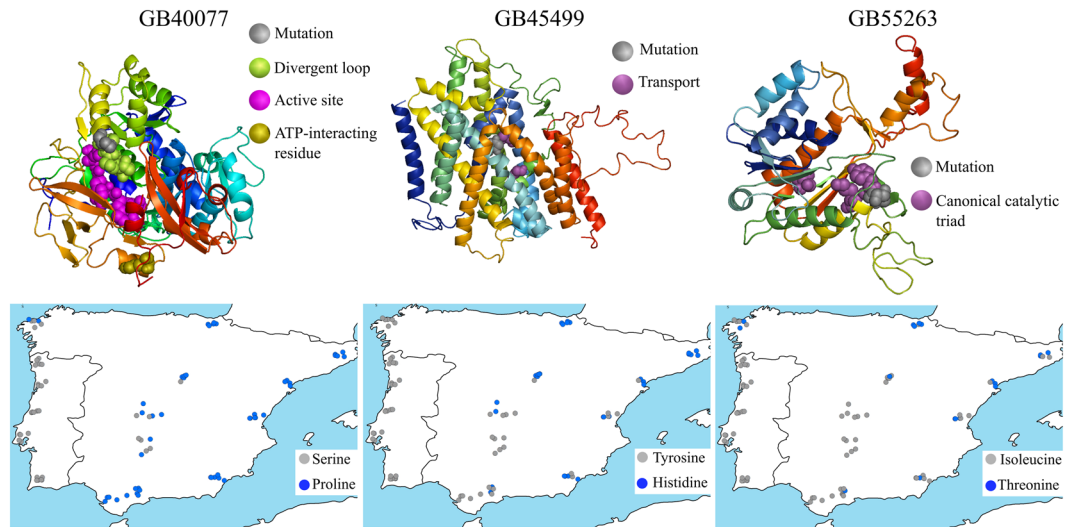


Figure 5. Predicted protein structures. The three genes harbour non-synonymous candidate SNPs, detected by three genome-wide methods, located nearby important places in the protein. The structures were predicted by Pymol considering the BeeBase reference amino acid sequences. The grey spheres represent the position and altered amino acids. The coloured spheres represent places with a known and important function in the protein. The maps depict the geographical patterns of the amino acids under selection.

The *membrane* cluster comprised 29 genes of which 16 could be grouped into three classes of proteins, including cell-surface receptors (7), transport (7), and cell-adhesion (1; Supplementary Tables S10 and S11). The cell-surface receptor genes (7 detected in GO analysis and GB45612 and GB48704) belong to four families, including the G-protein-coupled receptor family (GB40666, GB49166 GB48703 and its paralogue GB48704, GB49166 and GB51611), the ion-channel-coupled receptor (GB48639), the enzyme-coupled receptor (GB43446), and the CD36/scavenger receptor (GB49363). The transport proteins were represented by potassium channels (GB49879 and its paralogue GB49882) and transporters (GB45499, GB50262, GB49320, GB46597, GB53142, GB54678). Finally, the cell-adhesion proteins were represented by GB44159 and GB43719, being the latter undetected by the GO analysis (Supplementary Tables S4 and S10).

Although unrepresented in the GO enrichment analysis, many other genes are good candidates for local adaptation in the IHB as they are putatively implicated in the same biological function. These functions include reproduction with 7 genes, immunity with 11 genes, regulation of transcription with 7 genes, lipid storage and biosynthesis with 7 genes, olfaction with 8 genes, vision with 3 genes, and detoxification with 6 genes (Supplementary Table S11).

Discussion

In this study we employed conceptually different analytical tools to disentangle signatures of selection from genome-wide geospatial variation in the IHB. By scanning 87 whole genomes, we were able to refine inferences previously made from a limited number of pre-ascertained biased SNPs⁴¹ and provide further insights into the molecular basis of local adaptation in the IHB. In addition to providing unbiased information about the type of genes and biological processes putatively underlying local adaptation, we have never been so close to finding associated causal mutations.

The majority of the 41 non-synonymous outlier SNPs are likely causal mutations, especially those laying in genes GB40077, GB45499 and GB55263, as they could be linked to amino acid positions important for protein functioning^{52–55}. On the other hand, it is possible that many outlier SNPs are hitchhiked with the actual genetic target of the selective event as 80% of the outlier SNPs were <5 Kb apart. Yet, the hypothesis that many linked multiple causal mutations have a functional role cannot be ruled out^{56–58}.

A great proportion (88.7%) of cross-detected outlier SNPs are located in non-coding DNA, as opposed to the 11.3% exonic SNPs. A similar disproportionate fraction of non-coding to coding SNPs has been identified by whole-genome scans in other organisms, including humans⁵⁹, fishes⁵⁶, and fruit flies^{23,60}. This finding together with the significant enrichment of (i) outlier SNPs laying in <1 Kb upstream from the transcription start site of 42 genes, where the promoter is expected to be located, (ii) intronic regions in the immediate vicinity of exons (<50 bp), and (iii) UTRs suggest that regulatory sequences are an important source of adaptive change in the IHB. Further identification of causal mutations is a challenging endeavour that will require more accurate and comprehensive annotations of the honeybee genome, and especially annotation of the non-coding regulatory DNA, along with evidence from biochemical and functional assays.

Support for selection is provided by functional annotations of candidate genes that can be directly related to colony fitness and it is particularly compelling when multiple candidate genes are implicated in the same biological function. While the GO enrichment analysis only detected 6 significant terms (4 related with *membrane*), functional annotations indicate that many fitness-related functions are represented by multiple genes (e. g. as

reproduction with 7 genes or immunity with 11 genes). Other fitness-related biological functions were highlighted as they displayed strong selection signals. These include olfaction, circadian clock, and lipids biosynthesis and storage. Many of the candidate genes identified for the IHB were also detected by whole-genome selection scans for other honeybee subspecies^{43,44,47} (Supplementary Table S4), suggesting that they are adaptively important across diverse environments.

The GO analysis identified a cluster of 29 candidate genes encoding for membrane proteins in the IHB. The importance of membrane proteins in adaptation to new environments is evidenced by their rapid evolution compared with cytosolic proteins⁶¹. In this study, three candidate genes are highlighted in the group of membrane transport proteins. Gene GB45499 is one of strongest candidates for selection as it carries a mutation leading to an amino acid change in a site of the protein located in the transmembrane region and involved in transport activity⁵³. The replaced amino acid is located in the alpha-helix and, together with two other amino acids, it is important to maintain the open pathway from the intracellular space⁵³. Genes GB49879 and GB49882 are paralogous encoding voltage-gated K⁺ channels, which are putatively implicated in the circadian clock (see below). GB49879 was detected by all selection methods and exhibits a |iHS| = 2.19, indicative of strong signals of ongoing selection⁶². GB49882 is tagged by 19 SNPs, of which 5 are exonic. In addition to membrane transport protein, the selection scan identified three candidate genes in the group of membrane receptors all implicated in olfaction.

The adaptive relevance of olfaction is revealed by the significant enrichment of the GO term *olfactory learning* and identification of 8 candidate genes. GB48703 and GB48691 are amongst the most striking candidates deserving further investigation. GB48703 encodes an olfactory membrane receptor and carries 12 outlier SNPs, of which 3 are non-synonymous. GB48691 is implicated in olfactory learning and carries 9 outlier SNPs, of which one is non-synonymous. Unfortunately, protein prediction was not available for these genes hampering inferences on effects of the non-synonymous mutations in protein functioning. Colony fitness relies largely on olfactory perception. Olfaction is implicated in the learning process, which is crucial for the improvement of resources' acquisition, as well as in a wide array of behaviours, including detection of possible dangers, recognition of potential mates, and social interactions^{63,64}. Olfaction has also been shown to play a major role in the detection of brood cells infested by *Varroa destructor*⁶⁵, an invasive mite that has been challenging honeybee health at unprecedented levels.

The honeybee relies on a circadian clock to synchronize foraging behaviour and reproductive swarming with the maximum daily and seasonal availability of food resources^{66,67}. The importance of circadian rhythmicity in local adaption of IHBs is suggested by four candidate genes putatively operating in two functional components of the circadian clock: the oscillator and the output pathways. The core component "oscillator" is represented by GB52077. Its putative orthologue in *Drosophila* encodes for the transcription factor Period (*Per*). The honeybee *amPer* is an essential element of circadian rhythmicity, and its product is involved in a negative transcription/translational auto-regulatory feedback loop⁶⁸. The development of strong circadian rhythms in honeybee foragers has been shown to be associated with changes in brain *Per* expression⁶⁷. The component "output pathways" is represented by the striking candidates GB49879, GB49881 and GB49882. Genes GB49881 and GB49882 share transcript sequence associated with the SHAW voltage-gated K⁺ channel protein. GB49882 and GB49879 are paralogous encoding for SHAW and SHAW-like proteins, respectively. In *Drosophila*, the SHAW potassium channels regulate the intrinsic excitability in all neurons, being therefore important for output rhythms of the circadian clock⁶⁹. The four clock genes were mostly marked by intronic outlier SNPs, suggesting that gene regulation is an important molecular mechanism to meet functional demands of circadian rhythmicity.

Seven lipid-related candidate genes mostly implicated in lipids biosynthesis and storage were detected in the IHB, being GB55263 and GB40077 amongst the top-ranked candidates possibly playing a central role in IHB adaptation. The non-synonymous SNPs mapped to these genes are likely causal mutations as they lead to replacement of amino acids located in functionally important sites of the proteins. The mutation in GB55263 leads to an amino acid replacement in a canonical catalytic triad⁵⁴. The mutation in GB40077 leads to an amino acid replacement in a divergent loop⁵², which is important for mediating the protein-protein interaction or is part of the ATP binding site⁵⁵. The *Drosophila* ortholog of GB40077 is implicated in lipid homeostasis⁷⁰ and has been linked to the circadian clock^{71,72}.

Precipitation in August, May and January are the variables most frequently and strongly associated with SNPs. While precipitation may act as a selective pressure by interfering with foraging, winter mortality, behaviour in the nest, and mating flights^{73,74}, whether it is a direct cause of selection is unclear. It may very well be that precipitation operates indirectly by determining availability of pollen and nectar sources across space and time, which will not only influence foraging and colony build up but also reproduction. Due to the highly contrasting climates (e.g. average annual precipitation is 1336.3 mm in the northwest and 284.6 mm in the southeast), plant communities (wild plants or crops) and blooming seasons are very heterogeneous across Iberia⁷⁵. This could favour evolution of locally adapted populations to food resources. An interesting example of such adaptation is provided by the existence of an ecotype of *A. m. mellifera* (the other M-lineage subspecies in Europe) that has an annual brood cycle fine-tuned with the phenology of an abundant floral source in the Landes region in France^{76,77}.

Precipitation in August, May, and January covary with temperature, insolation, cloud cover, and precipitation in the other months. Multicollinearity may lead to incorrectly identifying a variable as causal when the true selective pressure is a correlated variable. However, it is also possible that selection is driven by composite environmental cues. For example, the mating behaviour of honeybee queens is influenced by a combination of temperature, wind, and cloud cover^{73,74}.

Longitude and latitude showed a large number of associations (113 and 102, respectively). While longitude and latitude do not act directly on organisms, they serve as composite variables representing multiple environmental factors, any one or a combination of which could be exerting parallel selective forces. Latitude has been found to be associated with circadian clock genes in *Drosophila*^{78–81} and humans^{82,83}, and now in honeybees. Clock genes are tagged by SNPs forming latitudinal and longitudinal gradients in Iberia. This finding suggests

that circadian rhythmicity is involved in local adaptation in IHBs by matching important behaviours, such as feeding and reproduction, with the diverse daily and seasonal environmental oscillations of Iberia.

Using both genetic and environmental data, we identified candidate genes putatively under climate-driven adaptation. This information is particularly important in the context of rapid global climate change, helping us to understand the mechanisms employed by organisms to adapt to varying environmental conditions.

Methods

Sampling. A total of 87 haploid *A. m. iberiensis* males were collected in 2010 from 16 sampling sites distributed across three north-south transects: one along the Atlantic coast (AT: N = 31), one along the centre (CT: N = 33), and another along the Mediterranean coast (MT: N = 23; see Chávez-Galarza, *et al.*⁴¹ for further sampling details). The sites cover a wide variety of climates ranging from the semi-arid in the southeastern part of Iberia to oceanic in the northwestern part (Fig. 1). Each of the 87 individuals represents a single colony and apiary.

Environmental Variables. Geographical coordinates, recorded for each apiary using a global positioning system (GPS), were used to obtain seven environmental variables from publicly available databases (WorldClim, Climatic Research Unit, OPENEI): precipitation (prec), minimum temperature (tmin), mean temperature (tmean), maximum temperature (tmax), cloud cover (cld), relative humidity (rh), and insolation (ins). These variables were integrated into a geographic information system (ArcGIS 9.3 from ESRI) to extract yearly, seasonal and monthly data. Arabic numerals appended to each environmental variable designate the month for which the variable was obtained; for example, prec5 refers to precipitation in May. In addition to climate, land cover was described for each apiary by calculating the percentage of level 3 land cover classes⁸⁴ within a 3 km radius circular area (for further details, see Chávez-Galarza, *et al.*⁴¹).

To prevent potential problems caused by non-independency, environmental variables were first organized into orthogonal vectors by performing a principal component analysis (PCA) using the *ade4* package⁸⁵. The strong correlation between many of the environmental variables in each vector means that they share a substantial amount of information and the relative importance of each variable is difficult to assess. Accordingly, variables that were correlated at $|r| > 0.8$ ⁸⁶ were removed from the data set. From an initial set of 123 environmental variables, 13 uncorrelated variables together with longitude (long) and latitude (lat), which are proxies for climatic diversity, were retained for further analysis (Supplementary Tables S12 and S13 and Fig. S6). Each retained variable is representative of a group of highly correlated variables, as listed in the Supplementary Table S13. The two largest groups comprise 33 variables; one of them is precipitation in May, which represents a wide array of variables, including precipitation, temperature, cloud and insolation; the other group is minimum temperature in June, which only represents temperature. Latitude is correlated with 27 variables, most of which represent insolation (ins), but also spring and summer precipitation (Supplementary Table S13).

Whole-Genome Sequencing and Filtering. Whole genome sequencing (WGS) was performed using the Illumina HiSeq 2500 platform, which produced a mean coverage of 11X, ranging from 3X to 23X (Supplementary Table S14). Sequencing libraries were generated using Illumina TruSeqTM Sample Preparation kits. The 2×150 paired end sequence reads were mapped against the reference honeybee genome Amel_4.5 using the Burrows-Wheeler Aligner (BWA)⁸⁷.

To improve the read mapping quality, PCR duplicates were identified and marked using Picard (<http://broadinstitute.github.io/picard/>) and realignment around indels was performed to correct inconsistently mapped reads using the Genome Analysis Toolkit (GATK)⁸⁸. To facilitate parallelization, the reads were split per chromosome using SAMtools (<http://samtools.sourceforge.net/>) and the readgroups information was modified with Picard. Bayesian population-based SNP calling was implemented using FreeBayes⁸⁹ across the 87 samples. To reduce poor mapping and spurious heterozygous positions, SNPs were removed that (1) had more than two alleles, (2) showed a quality score < 50 , (3) were present in less than 61 samples (70%), and (4) exhibited very high (> 3000) or very low (< 87) read depth (Supplementary Table S15). Haploid male data were intentionally misspecified to be diploid in the FreeBayes SNP calling process. Positions that showed more than 10 individuals as heterozygous were discarded, as they were unlikely to represent true SNPs. Missing genotypes were imputed by IMPUTE2⁹⁰. SNPs showing a minor allele frequency (MAF) < 0.05 were removed from the data set using PLINK⁹¹.

Genomic Information. Annotation information was obtained for all SNPs, including physical position, strand orientation and SNP functional state (non-synonymous, synonymous, intron or exon UTR, or intergenic regions), using the reference genome Amel_4.5, the Official Gene Set 3.2 (BEEBASE), and the Entrez Gene of NCBI. To have a complete functional annotation of each candidate gene, putative Gene Ontology classifications were obtained based on homology to *Drosophila melanogaster* using FLYBASE. The sequence alignments spanned at least 50 peptides with an e-score of 0.5 to assign orthologs. Approximately 7,103 *D. melanogaster* genes were linked to honeybee orthologs using these criteria. DAVID v.8.0 (the Database for Annotation, Visualization and Integrated Discovery) was accessed to determine if candidate genes were enriched for a specific functional annotation⁹². Genes were considered as candidates for selection if they were tagged by one or more SNP outliers laying in exons, introns, or UTRs.

Population Structure. Population structure was inferred from two different approaches: PCAdapt fast^{7,93} and sNMF⁹⁴. PCAdapt fast infers population structure using latent factors or scores. The approach sNMF is based on sparse non-negative matrix factorization to estimate the genetic ancestry components for each individual⁹⁴.

Ten runs were performed in sNMF with $\alpha = 8$ for each K value (1 to 10). Cross-entropy was used to guide the choice of the number of ancestral populations. To summarize and visualize the sNMF outputs, Q-plots were post-processed online with CLUMPAK⁹⁵. The results from the PCAdapt fast and sNMF were used to create latent factors in models (see the section below for further details).

Searches for Signatures of Local Adaptation. The whole genomes of the 87 IHBs were scanned for selection signals using three conceptually different methods (Sam β ada, LFMM and PCAdapt) and two data sets (a genomic data set and a combined genomic and environmental data set). The outlier SNPs detected by at least two methods were further examined using the haplotype-based method iHS and protein modelling for a secondary validation. Implementation of conceptually diverse approaches allows identification of potential false positives; by cross-validating outlier SNPs there is stronger evidence for selection^{96,97}. These SNPs are the most promising candidates for biochemical and functional follow up studies.

The significance levels of Sam β ada, LFMM and PCAdapt were assessed using the false discovery rate (FDR) procedure^{98,99}. To apply the FDR, the observed P-values should be uniformly distributed¹⁰⁰. When this assumption was not met, we applied the empirical null-hypothesis technique to recalibrate the distribution¹⁰⁰. Only SNPs with an FDR < 0.05 were considered as outliers.

Genetic-Environment Association Methods. Two GEA methods were employed to search for signatures of local adaptation. One implements mixed models (LFMM) and the other a logistic regression model (Sam β ada). LFMM uses an MCMC algorithm for regression analysis that models random effects, such as population history and isolation-by-distance, as unobserved (latent) factors¹⁰¹. This approach has proven to be efficient in screening genomes for signatures of local adaptation, performing well in cases of weak selection, complex hierarchical structure and polygenic selection^{13,97,102–104}. The program was run using 50,000 iterations and a burn-in of 25,000. Based on the ancestry estimates previously obtained with sNMF⁹⁴ and PCAdapt⁷, two latent factors were assumed. Since LFMM uses a stochastic algorithm, five runs with different seeds were performed. To increase the power of the LFMM test, the median z-score and adjustment of P-value were calculated.

The other GEA method Sam β ada is a spatial approach that uses univariate logistic regression models to identify locus-environment associations and at the same time measures spatial autocorrelation^{12,15}. Sam β ada was run for each of the 15 environmental variables. The analysis included global and local autocorrelation using a weighting factor based on the 25 nearest neighbours. The P-values were calculated from the Gscore.

Frequency-Based Method – PCAdapt fast. The frequency-based PCAdapt fast approach^{7,93} implements a genome scan to detect genes involved in local adaptation by taking into consideration population structure. PCAdapt fast infers population structure using latent factors or scores, and searches for loci that are atypically related to population structure measured by factor analysis (h). To calculate the best K, PCAdapt fast was run with K = 10. Given that the best K was 2, as determined by eigenvalues, the software was run for the second time to infer the loci under selection for K = 2. The latent factors, which describe population structure, were plotted in the first two PCA components (PC1 and PC2).

Haplotype-Based Method – iHS. The integrated haplotype score method, iHS, measures the strength of evidence for selection acting at or near a given SNP, tracking the decay of haplotype homozygosity for ancestral and derived haplotypes extending from a tested core^{62,105}. To determine the SNP variants state (ancestral or derived), we performed a pairwise alignment between *Apis mellifera* (v4.5¹⁰⁶) and *Apis cerana* reference genomes (v1.0¹⁰⁷) using the default settings of SATSUMA¹⁰⁸ whole-genome synteny package. Subsequently, the |iHS| values were estimated for candidate SNPs detected by at least two of the three previous methods using the Selscan package¹⁰⁵ with default parameters: –max-extend 1,000,000 (maximum EHH extension in bp), –max-gap 200,000 (maximum gap allowed between two SNPs in bp), –cutoff 0.05 (EHH decay cutoff). The script NORM, provided by Selscan, was implemented to frequency-normalize the output using the default parameter–bins 100 (number of frequency bins) over all chromosomes. Values of |iHS| > 2 are indicative of strong signals of recent positive selection⁶².

In Silico Analysis of 3D Protein Structure. Structures of related proteins were searched for on Phyre2¹⁰⁹ and the SWISS-MODEL servers¹¹⁰. The five best matches were aligned and compared with a reference protein using MEGA7¹¹¹; the structure with the best similarity and coverage was downloaded from the RCSB Protein Data Bank. The 3D structures of reference proteins and variants were modelled using SWISS-MODEL. FoldX¹¹² and the 3Drefine¹¹³ servers were used to refine the 3D structures. Protein stability of each variant was predicted using the Gibbs-free energy ($\Delta\Delta G$) calculated with the FoldX software. The minimum energy required for stable structure was estimated using GROMOS96 implemented in Swiss Pdb-viewer software¹¹⁴. Root-Median-Square Deviations (RMSD) between the reference protein and each variant were estimated using TM-score¹¹⁵. The 3D predicted protein structures were visualized in PyMol 0.99 (PyMOL Molecular Graphics System).

Data accessibility. Sequence data of *A. m. iberiensis* would be deposited at the ENA (www.ebi.ac.uk/ena) after the manuscript is accepted for publication.

References

1. Biswas, S. & Akey, J. M. Genomic insights into positive selection. *Trends in Genetics* **22**, 437–446, <https://doi.org/10.1016/j.tig.2006.06.005> (2006).
2. Guillot, G., Vitalis, R., le Rouzic, A. & Gautier, M. Detecting correlation between allele frequencies and environmental variables as a signature of selection. A fast computational approach for genome-wide studies. *Spatial Statistics* **8**, 145–155 (2014).
3. Luikart, G., England, P. R., Tallmon, D., Jordan, S. & Taberlet, P. The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics* **4**, 981–994 (2003).

4. Nielsen, R. *et al.* Genomic scans for selective sweeps using SNP data. *Genome Research* **15**, 1566–1575, <https://doi.org/10.1101/gr.4252305> (2005).
5. Foll, M. & Gaggiotti, O. A Genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. *Genetics* **180**, 977–993 (2008).
6. Excoffier, L., Hofer, T. & Foll, M. Detecting loci under selection in a hierarchically structured population. *Heredity (Edinb)* **103**, 285–298, <https://doi.org/10.1038/hdy.2009.74> (2009).
7. Duforet-Frebourg, N., Bazin, E. & Blum, M. G. Genome scans for detecting footprints of local adaptation using a Bayesian factor model. *Molecular biology and evolution* **31**, 2483–2495, <https://doi.org/10.1093/molbev/msu182> (2014).
8. Coop, G., Witonsky, D., Di Rienzo, A. & Pritchard, J. K. Using environmental correlations to identify loci underlying local adaptation. *Genetics* **185**, 1411–1423 (2010).
9. Frichot, E., Schoville, S. D., de Villemereuil, P., Gaggiotti, O. E. & Francois, O. Detecting adaptive evolution based on association with ecological gradients: orientation matters! *Heredity (Edinb)* **115**, 22–28, <https://doi.org/10.1038/hdy.2015.7> (2015).
10. Prunier, J., Laroche, J., Beaulieu, J. & Bousquet, J. Scanning the genome for gene SNPs related to climate adaptation and estimating selection at the molecular level in boreal black spruce. *Molecular Ecology* **20**, 1702–1716, <https://doi.org/10.1111/j.1365-294X.2011.05045.x> (2011).
11. Hoban, S. *et al.* Finding the genomic basis of local adaptation: Pitfalls, practical solutions, and future directions. *The American Naturalist* **188**, 379–397, <https://doi.org/10.1086/688018> (2016).
12. Stucki, S. *et al.* High performance computation of landscape genomic models integrating local indices of spatial association. *arXiv:1405.7658v1* (2014).
13. Lv, F. H. *et al.* Adaptations to climate-mediated selective pressures in sheep. *Molecular biology and evolution* **31**, 3324–3343, <https://doi.org/10.1093/molbev/msu264> (2014).
14. MacCallum, C. & Hill, E. Being Positive about Selection. *PLoS biology* **4**, e87, <https://doi.org/10.1371/journal.pbio.0040087> (2006).
15. Joost, S. *et al.* A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. *Molecular Ecology* **16**, 3955–3969, <https://doi.org/10.1111/j.1365-294X.2007.03442.x> (2007).
16. Forester, B. R., Jones, M. R., Joost, S., Landguth, E. L. & Lasky, J. R. Detecting spatial genetic signatures of local adaptation in heterogeneous landscapes. *Molecular Ecology* **25**, 104–120, <https://doi.org/10.1111/mec.13476> (2016).
17. Jensen, J. D., Kim, Y., DuMont, V. B., Aquadro, C. F. & Bustamante, C. D. Distinguishing between selective sweeps and demography using DNA polymorphism data. *Genetics* **170**, 1401–1410, <https://doi.org/10.1534/genetics.104.038224> (2005).
18. Manel, S. *et al.* Genomic resources and their influence on the detection of the signal of positive selection in genome scans. *Molecular Ecology* **25**, 170–184, <https://doi.org/10.1111/mec.13468> (2016).
19. Ko, A. *et al.* Amerindian-specific regions under positive selection harbour new lipid variants in Latinos. *Nature communications* **5**, 3983, <https://doi.org/10.1038/ncomms4983> (2014).
20. Triska, P. *et al.* Extensive admixture and selective pressure across the Sahel Belt. *Genome biology and evolution* **7**, 3484–3495, <https://doi.org/10.1093/gbe/evv236> (2015).
21. Xia, J. H. *et al.* Signatures of selection in tilapia revealed by whole genome resequencing. *Sci Rep* **5**, 14168, <https://doi.org/10.1038/srep14168> (2015).
22. Kang, L., Aggarwal, D. D., Rashkovetsky, E., Korol, A. B. & Michalak, P. Rapid genomic changes in *Drosophila melanogaster* adapting to desiccation stress in an experimental evolution system. *BMC Genomics* **17**, 233, <https://doi.org/10.1186/s12864-016-2556-y> (2016).
23. Božičević, V., Hutter, S., Stephan, W. & Wollstein, A. Population genetic evidence for cold adaptation in European *Drosophila melanogaster* populations. *Molecular Ecology* **25**, 1175–1191, <https://doi.org/10.1111/mec.13464> (2016).
24. Garud, N. R., Messer, P. W., Buzbas, E. O. & Petrov, D. A. Recent selective sweeps in North American *Drosophila melanogaster* show signatures of soft sweeps. *PLoS Genet* **11**, e1005004, <https://doi.org/10.1371/journal.pgen.1005004> (2015).
25. Benjelloun, B. *et al.* Characterizing neutral genomic diversity and selection signatures in indigenous populations of Moroccan goats (*Capra hircus*) using WGSdata. *Frontiers in Genetics* **6**, 107, <https://doi.org/10.3389/fgene.2015.00107> (2015).
26. Lai, F. N. *et al.* Whole-genome scanning for the litter size trait associated genes and SNPs under selection in dairy goat (*Capra hircus*). *Sci Rep* **6**, 12, <https://doi.org/10.1038/srep38096> (2016).
27. Wang, X. *et al.* Whole-genome sequencing of eight goat populations for the detection of selection signatures underlying production and adaptive traits. *Sci Rep* **6**, 38932, <https://doi.org/10.1038/srep38932> (2016).
28. Makinen, H., Vasemagi, A., McGinnity, P., Cross, T. F. & Primmer, C. R. Population genomic analyses of early-phase Atlantic Salmon (*Salmo salar*) domestication/captive breeding. *Evolutionary applications* **8**, 93–107, <https://doi.org/10.1111/eva.12230> (2015).
29. Sun, L. *et al.* Identification and analysis of genome-wide SNPs provide insight into signatures of selection and domestication in channel catfish (*Ictalurus punctatus*). *PLoS one* **9**, e109666, <https://doi.org/10.1371/journal.pone.0109666> (2014).
30. Ruttner, F. *Biogeography and taxonomy of honeybees*. 1–293 (Springer Verlag, 1988).
31. Sheppard, W. S. & Meixner, M. D. *Apis mellifera pomonella*, a new honey bee subspecies from Central Asia. *Apidologie* **34**, 367–375 (2003).
32. Meixner, M. D., Leta, M. A., Koeniger, N. & Fuchs, S. The honey bees of Ethiopia represent a new subspecies of *Apis mellifera* - *Apis mellifera simensis* n. ssp. *Apidologie* **42**, 425–437 (2011).
33. Chen, C. *et al.* Genomic analyses reveal demographic history and temperate adaptation of the newly discovered honey bee subspecies *Apis mellifera sinisxinyuan* n. ssp. *Molecular biology and evolution* **33**, 1337–1348, <https://doi.org/10.1093/molbev/msw017> (2016).
34. Engel, M. S. Fossil honey bees and evolution in the genus *Apis* (Hymenoptera: Apidae). *Apidologie* **29**, 265–281 (1998).
35. Weiss, S. & Ferrand, N. X. *Phylogeography of southern European refugia*. (Springer, 2007).
36. Franck, P., Garnery, L., Solignac, M. & Cornuet, J. M. The origin of west European subspecies of honeybees (*Apis mellifera*): New insights from microsatellite and mitochondrial data. *Evolution* **52**, 1119–1134 (1998).
37. Miguel, I., Iriando, 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**, 141–155, <https://doi.org/10.1051/apido:2007007> (2007).
38. Cánovas, F., Rúa, P., Serrano, J. & Galián, J. Microsatellite variability reveals beekeeping influences on Iberian honeybee populations. *Apidologie* **42**, 235–251, <https://doi.org/10.1007/s13592-011-0020-1> (2011).
39. Miguel, I. *et al.* Both geometric morphometric and microsatellite data consistently support the differentiation of the *Apis mellifera* M evolutionary branch. *Apidologie* **42**, 150–161, <https://doi.org/10.1051/apido/2010048> (2011).
40. Chávez-Galarza, J. *et al.* Revisiting the Iberian honey bee (*Apis mellifera iberiensis*) contact zone: maternal and genome-wide nuclear variations provide support for secondary contact from historical refugia. *Molecular ecology* **24**, 2973–2992, <https://doi.org/10.1111/mec.13223> (2015).
41. Chávez-Galarza, J. *et al.* Signatures of selection in the Iberian honey bee (*Apis mellifera iberiensis*) revealed by a genome scan analysis of single nucleotide polymorphisms. *Molecular ecology* **22**, 5890–5907, <https://doi.org/10.1111/mec.12537> (2013).
42. Beye, M. *et al.* Exceptionally high levels of recombination across the honey bee genome. *Genome Res* **16**, 1339–1344, <https://doi.org/10.1101/gr.5680406> (2006).

43. Harpur, B. A. *et al.* Population genomics of the honey bee reveals strong signatures of positive selection on worker traits. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 2614–2619, <https://doi.org/10.1073/pnas.1315506111> (2014).
44. Wallberg, A. *et al.* A worldwide survey of genome sequence variation provides insight into the evolutionary history of the honeybee *Apis mellifera*. *Nature genetics* **46**, 1081–1088, <https://doi.org/10.1038/ng.3077> (2014).
45. Wallberg, A., Pirk, C. W., Allsopp, M. H. & Webster, M. T. Identification of multiple loci associated with social parasitism in honeybees. *PLoS Genet* **12**, e1006097, <https://doi.org/10.1371/journal.pgen.1006097> (2016).
46. Mikheyev, A. S., Tin, M. M., Arora, J. & Seeley, T. D. Museum samples reveal rapid evolution by wild honey bees exposed to a novel parasite. *Nature communications* **6**, 7991, <https://doi.org/10.1038/ncomms8991> (2015).
47. Fuller, Z. L. *et al.* Genome-wide analysis of signatures of selection in populations of African honey bees (*Apis mellifera*) using new web-based tools. *BMC Genomics* **16**, 518, <https://doi.org/10.1186/s12864-015-1712-0> (2015).
48. Meirmans, P. G. The trouble with isolation by distance. *Molecular Ecology* **21**, 2839–2846, <https://doi.org/10.1111/j.1365-294X.2012.05578.x> (2012).
49. Campana, M. G., Hunt, H. V., Jones, H. & White, J. Corrsieve: software for summarizing and evaluating Structure output. *Molecular ecology resources* **11**, 349–352, <https://doi.org/10.1111/j.1755-0998.2010.02917.x> (2011).
50. Waples, R. S. & Gaggiotti, O. What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular Ecology* **15**, 1419–1439, <https://doi.org/10.1111/j.1365-294X.2006.02890.x> (2006).
51. Yon Rhee, S., Wood, V., Dolinski, K. & Draghici, S. Use and misuse of the gene ontology annotations. *Nature Reviews Genetics* **9**, 509–515, <https://doi.org/10.1038/nrg2363> (2008).
52. Hu, Y. Crystal structures and enzymatic mechanisms of a *Populus tomentosa* 4-coumarate–CoA ligase. <https://doi.org/10.2210/pdb3a9u/pdb> (2010).
53. Watanabe, A. *et al.* The mechanism of sodium and substrate release from the binding pocket of vSGLT. *Nature* **468**, 988–991, <https://doi.org/10.1038/nature09580> (2010).
54. Chhabra, A. *et al.* Nonprocessive [2 + 2]e⁻ off-loading reductase domains from mycobacterial nonribosomal peptide synthetases. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 5681–5686, <https://doi.org/10.1073/pnas.1118680109> (2012).
55. Campos, B. M. *et al.* A redox 2-Cys mechanism regulates the catalytic activity of divergent cyclophilins. *Plant physiology* **162**, 1311–1323, <https://doi.org/10.1104/pp.113.218339> (2013).
56. Martinez Barrio, A. *et al.* The genetic basis for ecological adaptation of the Atlantic herring revealed by genome sequencing. *eLife* **5**, e12081, <https://doi.org/10.7554/eLife.12081> (2016).
57. Anderson, E. C. Assessing the power of informative subsets of loci for population assignment: standard methods are upwardly biased. *Molecular ecology resources* **10**, 701–710, <https://doi.org/10.1111/j.1755-0998.2010.02846.x> (2010).
58. Lamichhaney, S. *et al.* Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature* **518**, 371–375, <https://doi.org/10.1038/nature14181> (2015).
59. Vernot, B. *et al.* Personal and population genomics of human regulatory variation. *Genome Research* **22**, 1689–1697, <https://doi.org/10.1101/gr.134890.111> (2012).
60. Andolfatto, P. Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* **437**, 1149–1152, <https://doi.org/10.1038/nature04107> (2005).
61. Sojo, V., Dessimoz, C., Pomiankowski, A. & Lane, N. Membrane proteins are dramatically less conserved than water-soluble proteins across the Tree of Life. *Molecular biology and evolution* **33**, 2874–2884, <https://doi.org/10.1093/molbev/msw164> (2016).
62. Voight, B. F., Kudravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS biology* **4**, e72, <https://doi.org/10.1371/journal.pbio.0040072> (2006).
63. Schowalter, T. D. *Insect ecology: an ecosystem approach*. (Academic Press, 2016).
64. Sandoz, J. C. Behavioral and neurophysiological study of olfactory perception and learning in honeybees. *Frontiers in systems neuroscience* **5**, 98, <https://doi.org/10.3389/fnsys.2011.00098> (2011).
65. Navajas, M. *et al.* Differential gene expression of the honey bee *Apis mellifera* associated with *Varroa destructor* infection. *BMC Genomics* **9**, 301, <https://doi.org/10.1186/1471-2164-9-301> (2008).
66. Simpson, J. The problem of swarming in beekeeping practice. *Bee World* **39**, 193–202, <https://doi.org/10.1080/0005772x.1958.11095063> (1958).
67. Bloch, G. The social clock of the honeybee. *Journal of Biological Rhythms* **25**, 307–317, <https://doi.org/10.1177/0748730410380149> (2010).
68. Eban-Rothschild, A. & Bloch, G. In *Honeybee neurobiology and behavior*. 31–45 (Springer, 2012).
69. Hodge, J. J. & Stanewsky, R. Function of the Shaw potassium channel within the *Drosophila* circadian clock. *PLoS one* **3**, e2274, <https://doi.org/10.1371/journal.pone.0002274> (2008).
70. Xu, X. *et al.* Insulin signaling regulates fatty acid catabolism at the level of CoA activation. *PLoS Genet* **8**, e1002478, <https://doi.org/10.1371/journal.pgen.1002478> (2012).
71. Claridge-Chang, A. *et al.* Circadian regulation of gene expression systems in the *Drosophila* head. *Neuron* **32**, 657–671 (2001).
72. Xu, K., DiAngelo, Justin, R., Hughes, Michael, E., Hogenesch, John, B. & Sehgal, A. The circadian clock interacts with metabolic physiology to influence reproductive fitness. *Cell Metabolism* **13**, 639–654, <https://doi.org/10.1016/j.cmet.2011.05.001> (2011).
73. Bol'shakova, M. D. The flight of honey bee drones, *Apis mellifera* L. (Hymenoptera, Apidae), to the queen in relation to various ecological factors. *Entomological Review* **56**, 53–56 (1978).
74. Lensky, Y. & Demter, M. Mating flights of the queen honeybee (*Apis mellifera*) in a subtropical climate. *Comparative Biochemistry and Physiology Part A: Physiology* **81**, 229–241, [https://doi.org/10.1016/0300-9629\(85\)90127-6](https://doi.org/10.1016/0300-9629(85)90127-6) (1985).
75. Loidi, J. *The vegetation of the Iberian Peninsula, Volume 2*. Vol. 2 (Springer, 2017).
76. Strange, J. P., Garnery, L. & Sheppard, W. S. Persistence of the Landes ecotype of *Apis mellifera mellifera* in southwest France: confirmation of a locally adaptive annual brood cycle trait. *Apidologie* **38**, 259–267 (2007).
77. Strange, J. P., Garnery, L. & Sheppard, W. S. Morphological and molecular characterization of the Landes honey bee (*Apis mellifera* L.) ecotype for genetic conservation. *Journal of Insect Conservation* **12**, 527–537, <https://doi.org/10.1007/s10841-007-9093-6> (2008).
78. Yerushalmi, S. & Green, R. M. Evidence for the adaptive significance of circadian rhythms. *Ecology Letters* **12**, 970–981, <https://doi.org/10.1111/j.1461-0248.2009.01343.x> (2009).
79. Kyriacou, C. P., Peixoto, A. A., Sandrelli, F., Costa, R. & Tauber, E. Clines in clock genes: fine-tuning circadian rhythms to the environment. *Trends in Genetics* **24**, 124–132, <https://doi.org/10.1016/j.tig.2007.12.003> (2007).
80. Costa, R., Peixoto, A. A., Barbujani, G. & Kyriacou, C. P. A latitudinal cline in a *Drosophila* clock gene. *Proceedings. Biological sciences* **250**, 43–49, <https://doi.org/10.1098/rspb.1992.0128> (1992).
81. Tauber, E. *et al.* Natural selection favors a newly derived timeless allele in *Drosophila melanogaster*. *Science* **316**, 1895–1898, <https://doi.org/10.1126/science.1138412> (2007).
82. Forni, D. *et al.* Genetic adaptation of the human circadian clock to day-length latitudinal variations and relevance for affective disorders. *Genome Biology* **15**, 499, <https://doi.org/10.1186/s13059-014-0499-7> (2014).

83. Dall'Ara, I. *et al.* Demographic history and adaptation account for clock gene diversity in humans. *Heredity (Edinb)* **117**, 165–172, <https://doi.org/10.1038/hdy.2016.39> (2016).
84. Heymann, Y., Steenmans, C., Croisille, G. & Bossard, M. Land cover. Technical Guide. Office for Official Publications of European Communities, Luxembourg (1994).
85. Thioulouse, J., Chessel, D., Doledec, S. & Olivier, J. M. ADE-4: a multivariate analysis and graphical display software. *Statistics and Computing* **7**, 75–83 (1997).
86. Manel, S. *et al.* Perspectives on the use of landscape genetics to detect genetic adaptive variation in the field. *Molecular Ecology* **19**, 3760–3772, <https://doi.org/10.1111/j.1365-294X.2010.04717.x> (2010).
87. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595, <https://doi.org/10.1093/bioinformatics/btp698> (2010).
88. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**, 491–498, <https://doi.org/10.1038/ng.806> (2011).
89. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907 [q-bio.GN]* (2012).
90. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529, <https://doi.org/10.1371/journal.pgen.1000529> (2009).
91. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* **81**, 559–575, <https://doi.org/10.1086/519795> (2007).
92. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research* **37**, 1–13, <https://doi.org/10.1093/nar/gkn923> (2009).
93. Duforet-Frebourg, N., Luu, K., Laval, G., Bazin, E. & Blum, M. G. Detecting genomic signatures of natural selection with Principal Component Analysis: Application to the 1000 genomes data. *Molecular biology and evolution*, <https://doi.org/10.1093/molbev/msv334> (2015).
94. Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G. & Francois, O. Fast and efficient estimation of individual ancestry coefficients. *Genetics* **196**, 973–983, <https://doi.org/10.1534/genetics.113.160572> (2014).
95. Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A. & Mayrose, I. Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. *Molecular ecology resources* **15**, 1179–1191, <https://doi.org/10.1111/1755-0998.12387> (2015).
96. Vasemagi, A. & Primmer, C. R. Challenges for identifying functionally important genetic variation: the promise of combining complementary research strategies. *Molecular Ecology* **14**, 3623–3642, <https://doi.org/10.1111/j.1365-294X.2005.02690.x> (2005).
97. de Villemereuil, P., Frichot, E., Bazin, E., Francois, O. & Gaggiotti, O. E. Genome scan methods against more complex models: when and how much should we trust them? *Molecular Ecology* **23**, 2006–2019, <https://doi.org/10.1111/mec.12705> (2014).
98. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate - a practical and powerful approach to multiple testing. *J Roy Stat Soc B Met* **57**, 289–300 (1995).
99. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 9440–9445, <https://doi.org/10.1073/pnas.1530509100> (2003).
100. François, O., Martins, H., Caye, K. & Schoville, S. D. Controlling false discoveries in genome scans for selection. *Molecular Ecology* **25**(454–469), 1365–1294X (2016).
101. Frichot, E., Schoville, S. D., Bouchard, G. & Francois, O. Testing for associations between loci and environmental gradients using latent factor mixed models. *Molecular biology and evolution* **30**, 1687–1699, <https://doi.org/10.1093/molbev/mst063> (2013).
102. Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M. & Holderegger, R. A practical guide to environmental association analysis in landscape genomics. *Molecular Ecology* **24**, 4348–4370, <https://doi.org/10.1111/mec.13322> (2015).
103. Zueva, K. J. *et al.* Footprints of directional selection in wild Atlantic salmon populations: evidence for parasite-driven evolution? *PloS one* **9**, e91672, <https://doi.org/10.1371/journal.pone.0091672> (2014).
104. Lotterhos, K. E. & Whitlock, M. C. The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology* **24**, 1031–1046, <https://doi.org/10.1111/mec.13100> (2015).
105. Szpiech, Z. A. & Hernandez, R. D. selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Molecular biology and evolution* **31**, 2824–2827, <https://doi.org/10.1093/molbev/msu211> (2014).
106. Elsik, C. G. *et al.* Finding the missing honey bee genes: lessons learned from a genome upgrade. *BMC genomics* **15**(86), 1471–2164 (2014).
107. Park, D. *et al.* Uncovering the novel characteristics of Asian honey bee, *Apis cerana*, by whole genome sequencing. *BMC genomics* **16**(1), 1471–2164 (2015).
108. Grabherr, M. G. *et al.* Genome-wide synteny through highly sensitive sequence alignment: Satsuma. *Bioinformatics* **26**(1145–1151), 1460–2059 (2010).
109. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature protocols* **10**, 845–858, <https://doi.org/10.1038/nprot.2015.053> (2015).
110. Biasini, M. *et al.* SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic acids research* **42**, W252–258, <https://doi.org/10.1093/nar/gku340> (2014).
111. Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular biology and evolution* **33**, 1870–1874, <https://doi.org/10.1093/molbev/msw054> (2016).
112. Schymkowitz, J. *et al.* The FoldX web server: an online force field. *Nucleic acids research* **33**, W382–388, <https://doi.org/10.1093/nar/gki387> (2005).
113. Bhattacharya, D., Nowotny, J., Cao, R. & Cheng, J. 3Drefine: an interactive web server for efficient protein structure refinement. *Nucleic acids research* **44**, W406–W409, <https://doi.org/10.1093/nar/gkw336> (2016).
114. Guex, N. & Peitsch, M. C. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* **18**, 2714–2723, <https://doi.org/10.1002/elps.1150181505> (1997).
115. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic acids research* **33**, 2302–2309, <https://doi.org/10.1093/nar/gki524> (2005).

Acknowledgements

John C. Patton, Phillip San Miguel, Paul Parker, Rick Westerman, University of Purdue, resequenced the 87 whole genomes of IHBs. José Rufino provided computational resources at IPB. Analyses were performed using the computational resources at the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX), Uppsala University. DH was supported by a PhD scholarship (SFRH/BD/84195/2012) from the Portuguese Science Foundation (FCT). MAP is a member of and receives support from the COST Action FA1307 (SUPER-B). This work was supported by FCT through the programs COMPETE/QREN/EU (PTDC/BIA-BEC/099640/2008) and the 2013–2014 BiodivERSA/FACCE-JPI (joint call for research proposals, with the national funders FCT, Portugal, CNRS, France, and MEC, Spain) to MAP.

Author Contributions

M.A.P., D.H., A.W., M.T.W. conceived the ideas and designed methodology. D.H. performed most of the analyses with assistance of A.W. and J.C.G., M.A.P. and D.H. wrote the manuscript with input from J.S.J. All the authors critically reviewed the manuscript for important intellectual content.

Additional Information

Supplementary information accompanies this paper at <https://doi.org/10.1038/s41598-018-29469-5>.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018