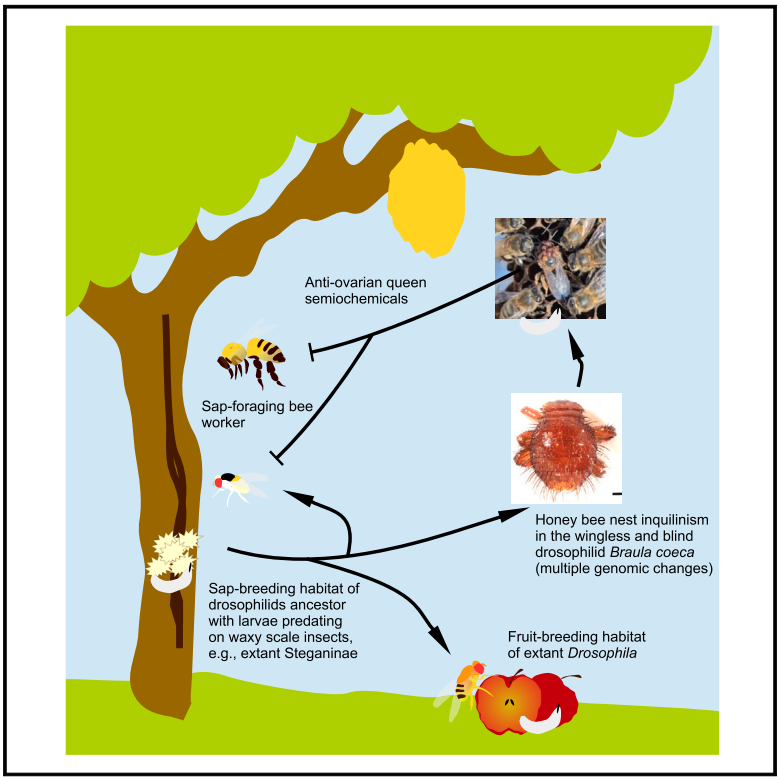
Report

Current Biology

The genome of the blind bee louse fly reveals deep convergences with its social host and illuminates *Drosophila* origins

Graphical abstract



Highlights

- The bee louse fly (*Braula coeca*) is a drosophilid inquiline of honey bee nests
- *B. coeca* likely evolved from a sap-breeding ancestor associated with scale insects
- *B. coeca* has a large genome, including a horizontal transposon transfer with bees
- Many chemosensory and detoxification genes were lost, mirroring trends in bees

Authors

Héloïse Bastide, Hélène Legout, Noé Dogbo, ..., Fabrice Requier, Jean-Christophe Sandoz, Amir Yassin

Correspondence

heloise.bastide@universite-paris-saclay.fr

In brief

Bastide et al. analyze the genome of the blind and wingless bee louse fly *Braula coeca*, an inquiline of the honey bee. *B. coeca* likely descends from a sap-breeder drosophilid ancestor. Inquilinism involves horizontal transposon transfer and convergent losses in relevant chemosensory and detoxification genes with the social host.

Bastide et al., 2024, Current Biology *34*, 1–11
March 11, 2024
https://doi.org/10.1016/j.cub.2024.01.034

CellPress

Current Biology

CellPress

Report

The genome of the blind bee louse fly reveals deep convergences with its social host and illuminates *Drosophila* origins

Héloïse Bastide,[1,3,*] Hélène Legout,[1] Noé Dogbo,[1] David Ogereau,[1] Carolina Prediger,[1] Julie Carcaud,[1] Jonathan Filée,[1] Lionel Garnery,[1] Clément Gilbert,[1] Frédéric Marion-Poll,[1,2] Fabrice Requier,[1] Jean-Christophe Sandoz,[1] and Amir Yassin[1]

[1]Université Paris-Saclay, CNRS, IRD, UMR Évolution, Génomes, Comportement et Écologie, 91198 Gif-sur-Yvette, France

[2]Université Paris-Saclay, AgroParisTech, 91123 Palaiseau Cedex, France

[3]Lead contact

*Correspondence: heloise.bastide@universite-paris-saclay.fr

https://doi.org/10.1016/j.cub.2024.01.034

SUMMARY

Social insects' nests harbor intruders known as inquilines,[1] which are usually related to their hosts.[2,3] However, distant non-social inquilines may also show convergences with their hosts,[4,5] although the underlying genomic changes remain unclear. We analyzed the genome of the wingless and blind bee louse fly *Braula coeca*, an inquiline kleptoparasite of the western honey bee, *Apis mellifera*.[6,7] Using large phylogenomic data, we confirmed recent accounts that the bee louse fly is a drosophilid[8,9] and showed that it had likely evolved from a sap-breeder ancestor associated with honeydew and scale insects' wax. Unlike many parasites, the bee louse fly genome did not show significant erosion or strict reliance on an endosymbiont, likely due to a relatively recent age of inquilinism. However, we observed a horizontal transfer of a transposon and a striking parallel evolution in a set of gene families between the honey bee and the bee louse fly. Convergences included genes potentially involved in metabolism and immunity and the loss of nearly all bitter-tasting gustatory receptors, in agreement with life in a protective nest and a diet of honey, pollen, and beeswax. Vision and odorant receptor genes also exhibited rapid losses. Only genes whose orthologs in the closely related *Drosophila melanogaster* respond to honey bee pheromone components or floral aroma were retained, whereas the losses included orthologous receptors responsive to the anti-ovarian honey bee queen pheromones. Hence, deep genomic convergences can underlie major phenotypic transitions during the evolution of inquilinism between non-social parasites and their social hosts.

RESULTS AND DISCUSSION

The bee louse fly *Braula coeca* is an aberrant member of the Drosophilidae

Among the several parasites and inquilines that are attracted by the rich resources and clean and protective shelter of the western honey bee, *Apis mellifera*, nest, none has undergone as profound morphological changes as the apterous and quasi-blind bee louse fly, *Braula coeca* (Figures 1A–1C). The female lays eggs in honey (not brood) cells, and the hatched larvae eat pollen and wax, where they burrow tunnels in which they pupate without forming true puparia.[6,7] Following emergence, the adults attach to the bodies of worker bees, migrating from one individual to another until reaching the queen (Figure 1A). There, they move to the queen's head, stimulate regurgitation, and imbibe honey and nectar from her mouth.[6,7] The bee louse fly is considered an inquiline kleptoparasite, with potential negative effects on honey bee colony health due to the galleries it makes in bee combs and its facilitation of the transmission of serious pathogenic viruses to the bees.[10]

Ever since Réaumur's first description of the bee louse fly in 1740, and Nitzsch's creation of the genus *Braula* in 1818,[11,12] the positioning within the Diptera of the bee louse fly and affiliated species that were classified under the family Braulidae has been puzzling because of its aberrant morphology and unique adaptations to a social host. This family contains seven species belonging to the genera *Braula* and *Megabraula*, which are all inquilines to honey bee species of the genus *Apis*. Recent phylogenetic analyses based on a transcriptome assembled from one adult fly and using 1,130 loci interestingly showed *Braula coeca*, the most widespread braulid, to constitute a basal lineage within the Drosophilidae that was sister to four genera of the subfamily Steganinae.[8,9] To reassess this hypothesis using a larger dataset, we sequenced the whole genome from a pooled sample of 15 unsexed *B. coeca* flies, all collected on Ouessant Island in western France. We used a hybrid approach to assemble a genome using long-read Oxford Nanopore Technology (ONT) and short-read Illumina sequencing (see STAR Methods). Benchmarking universal single-copy orthologs (BUSCO)[13] gave a score of 95.8% of the Dipteran conserved single-copy orthologs with 1.3% of duplicated genes. This value is higher than the recommended score of 90% for reference genomes.[14] Merqury[15] estimated an assembly completeness of 93.6% and a consensus quality value (QV) of 41, which exceeds

Please cite this article in press as: Bastide et al., The genome of the blind bee louse fly reveals deep convergences with its social host and illuminates *Drosophila* origins, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.01.034

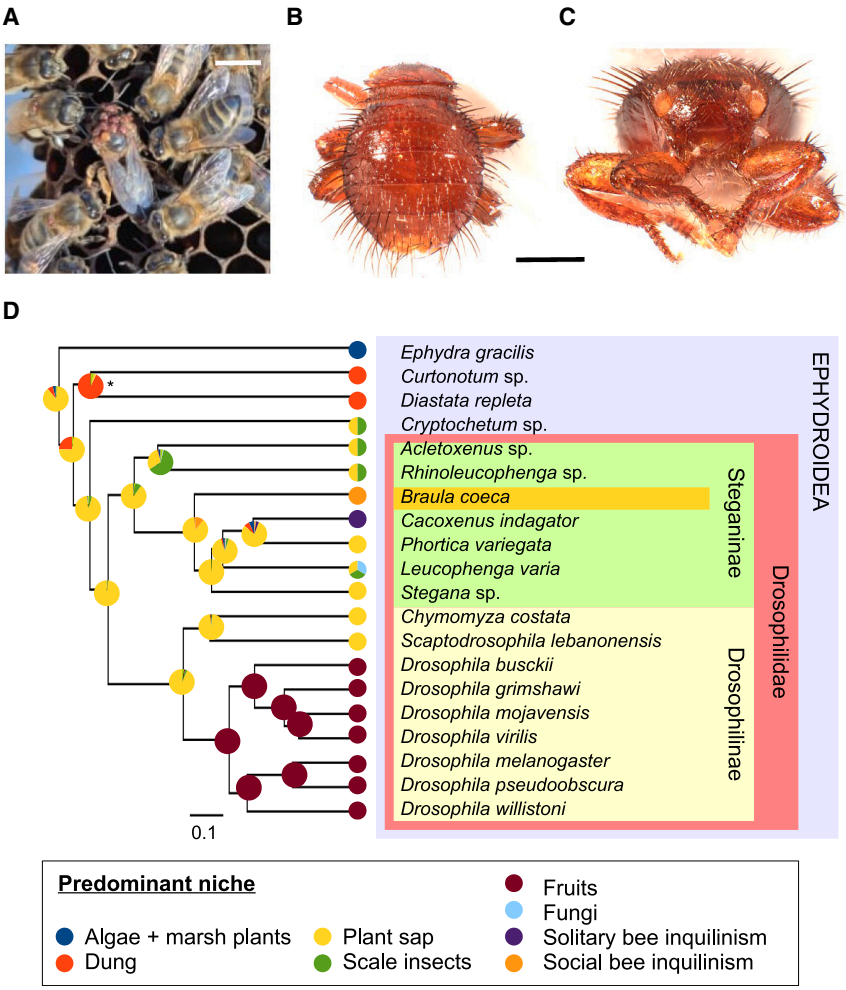


Figure 1. The bee louse fly (*Braula coeca*) is an inquiline of the western honey bee (*Apis mellifera*) and has likely evolved from a sap-breeding drosophilid associated with scale insects

(A) Tens of *B. coeca* adults preferentially attached to the honey bee queen (© Etienne Minaud). Scale bars, 5 mm.

(B) Dorsal view of an adult showing the loss of the wings, halters, and scutum, mesonotum reduction and the legs' robustness. Scale bars, 0.5 mm.

(C) Frontal view of an adult showing the reduction of the eyes and the loss of the ocelli. Scale bars, 0.5 mm.

(D) Maximum-likelihood phylogeny inferred from 3,100 conserved single-copy proteins (2,557,349 amino acids), showing the position of *B. coeca* (orange) in the subfamily Steganinae (light green) of the Drosophilidae (red). Outgroup species belong to the superfamily Ephydroidea (light blue). All internal nodes had an ultra-fast bootstrap value of 100% except * = 73%. Pie charts at internal nodes indicate the likelihood of ancestral breeding niches inferred from the predominant niches of terminal taxa.

See also Table S1.

the recommended threshold of QV40 for reference genomes.[16] We assembled two genomes and one transcriptome of three additional steganine genera. We then built a supermatrix of 3,100 BUSCO genes (2,557,349 amino acids) that included 15 drosophilid species (representative members of the 4 main radiations in the family),[17] and 5 species belonging to the superfamily Ephydroidea to which both the Drosophilidae and Braulidae belong (Table S1). The maximum-likelihood phylogenetic analysis of this large dataset reconfirmed the close-relatedness of *B. coeca* to the Drosophilidae. It further showed that it is a full member of the subfamily Steganinae (Figure 1D). The taxonomic priority principle should consider the family Drosophilidae, described in 1856,[18] a junior synonym for the family Braulidae, described in 1853.[19] However, the asymmetric size and scientific relevance of the two families argue against such a decision. We, therefore, opt for synonymizing the Braulidae with the Drosophilidae, referring hereafter to *Braula* and *Megabraula* as members of the subfamily Steganinae.

Inquilinism in the bee louse fly likely evolved from sap breeders associated with scale insects

To gain further insight into the history of the association between Braula and Apis, we mapped the predominant ecological habitats of ephydroid families on the phylogeny. The ancestral habitat of ephydroids was presumed to be rotting leaf molds.[20] From this, multiple specializations took place, including the exploitation of aquatic molds (and eventually algae) in the Ephydridae,[21] mammal dung in Curtonotidae and Diastatidae,[21,22] and fermenting vegetables, fruits, sap, and fungi, with specialization mostly on yeasts in the Drosophilidae[23] (Figure 1D). Bayesian reconstruction suggests the ancestral habitat of the Drosophilidae to be tree sap breeding (Figure 1D), with fungus and fruit breeding subsequently deriving and predominating in the genera *Leucophenga* and *Drosophila*, respectively. Remarkably, the deepest branches in the Steganinae and the Cryptochaetidae (the closest relative to the Drosophilidae) represent lineages whose larvae are predatory of scale insects and mealy bugs, e.g., *Acletoxenus formosus* and *A. indicus* on aleyrodoids, *Rhinoleucophenga brasiliensis*, and *R. obesa*, as well as *Cryptochaetidum iceryae* and *C. grandicorne* on coccoids.[24] In those lineages, adults are often seen to feed on the honeydew produced by the bugs, an abundant sugar-rich substrate sucked from plants' sap, while larvae take shelter and develop in the waxy secretions of these insects. This dependence on sugary substrate (honeydew) and development in a waxy environment could have predisposed *Braula*'s inquilinism in bee nests.

The bee louse fly inquilinism is relatively recent

To date *Braula* inquilinism, we inferred a fossil-calibrated phylogeny using 79 single-copy orthologs (63,192 amino acids) in 17 Acalyptrate dipteran and 25 Apocrite hymenopteran species (see STAR Methods; Table S1; Figure 2A). Five non-ephydroid dipteran species with RefSeq assemblies were included in this analysis to correct for tree imbalance.[25] The divergence of *B. coeca* from its closest steganine relatives (node 1 in

Please cite this article in press as: Bastide et al., The genome of the blind bee louse fly reveals deep convergences with its social host and illuminates *Drosophila* origins, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.01.034

CellPress

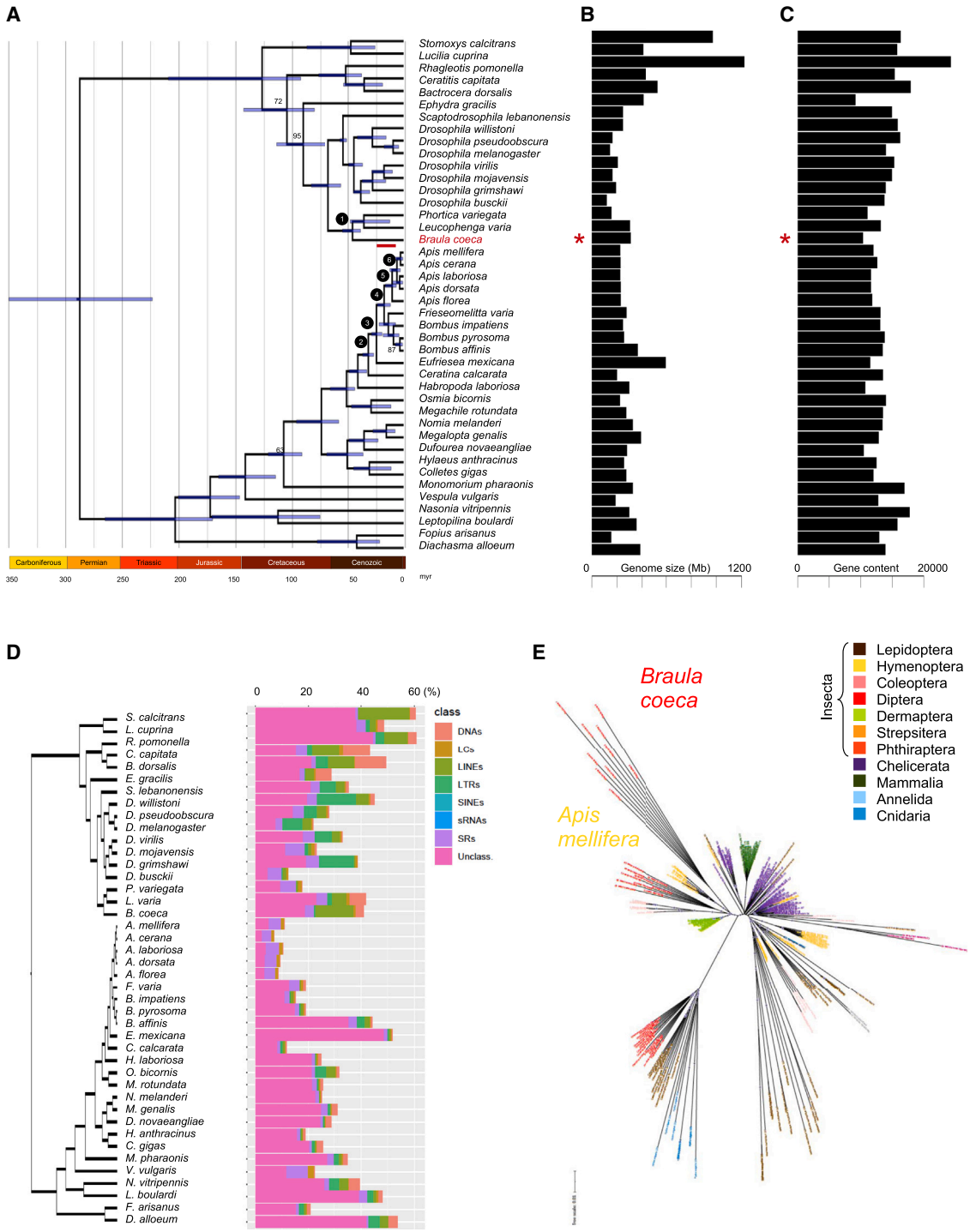


Figure 2. Evolution of the bee louse fly inquilinism, its genome size, gene content, and transposable elements in the bee louse fly, with evidence for horizontal transfer between the inquiline and its host

(A) Fossil-calibrated maximum-likelihood phylogeny inferred from 79 conserved single-copy proteins (63,192 amino acids) demonstrating major stages in the evolution of the inquiline and its social host. All internal nodes had an ultra-fast bootstrap value of 100 (except when given), with a blue interval indicating a 95% confidence level of divergence time estimate inferred by MCMCTree. The red bar indicates the likely interval of the origin of the bee louse fly-*Apis* association. Labels 1–6 refer to the major stages mentioned in the text.

(B) Genome size evolution. Red asterisk indicates the estimate for *B. coeca*.

(C) Gene content evolution. Red asterisk indicates the estimate for *B. coeca*.

(legend continued on next page)

Please cite this article in press as: Bastide et al., The genome of the blind bee louse fly reveals deep convergences with its social host and illuminates *Drosophila* origins, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.01.034

Figure 2A: 44.9 [37.8–53.8] million years [Ma] [95% confidence interval]) overlapped with the origin of the Apidae (50.12 [42.9–64.5] Ma) and with the transition from solitary to subsocial (2: 40.25 [32.4–47.8] Ma) and primitively social habits (3: 30.98 [26.3–36.0] Ma).[26] It is possible that the origin of the bee louse fly-apid interactions occurred at sap-breeding sites, when early subsocial apids started to gather resin and other plant exudates, as well as scale insects' honeydew, and stored them in their nests. As eusociality evolved (4: 23.8 [18.9–28.2] Ma), the proportion of resin to secreted wax diminished and some cells were also used to store nectar and honey for the brood.[27] A shift from the putatively ancestral dependence on honey and wax produced by scale insects to those produced by bees might have evolved by then. The transition to eusociality in the genus *Apis* required an important division of labor that involved the evolution of pheromonal control of the reproductive capacity of worker females by the queen and the evolution of trophallaxis.[27] Adaptation of *Braula* to the queen pheromone compounds that have anti-ovarian effects on a wide range of insects, including *Drosophila melanogaster*,[28] and the exploitation of trophallaxis[7] could not have evolved before the advancement of eusociality (5: 17.1 [11.6–23.0] Ma). The evolution of blindness and apterism should have constrained the dispersal of the bee lice, relating their speciation history to that of their hosts. Indeed, only seven bee louse fly species are known, of which five *Braula* species are restricted to the western honey bee, *A. mellifera*, and two *Megabraula* species are restricted to the giant honey bee, *A. laboriosa*, in the Himalayas.[29,30] The divergence between these *Apis* species, and presumably between *Braula* and *Megabraula*, is estimated at 6: 5.8 [2.8–12.0] million years ago (mya). Therefore, the evolution of the bee louse fly inquilinism likely took place during the Mid- to Late Miocene period, between 5.8 and 17.1 mya (Figure 2A). We cannot rule out an even more recent origin if the ancestor of *Braula* or *Megabraula* has shifted from one *Apis* host to another, i.e., <5.8 mya.

The bee louse fly inquilinism was accompanied by a reduction in gene content but not genome size

Loss of significant portions of genomic and gene contents is a characteristic of obligate parasites specializing on specific hosts or inhabiting extreme environments. For example, the human body louse, *Pediculus humanus*, has one of the smallest genomes and the lowest numbers of genes in insects (108 megabases [Mb] and 10,773 protein-coding genes).[31] For *B. coeca*, we obtained a final assembly size of 309.35 Mb shared by 2,477 contigs, with an N50 of 347,211 bp. This N50 estimate is typical of hybrid genome assemblies obtained using a pooled sample of wild-caught drosophilid flies from species with large genome sizes (>300 Mb).[32] No evidence for polyploidy or other endosymbiont that could have biased the genome size estimate was detected (Figure S1). Genome size prediction using k-mers distribution spectra predicted a genome of 308 Mb, concordant with the assembly size (Figure S1). Such a genome size is significantly larger than the remaining drosophilid species (Student's t one-sample test $p < 1.3 \times 10^{-4}$; Shapiro-Wilk normality test $p = 0.55$). Phylogenetic analysis of genome size evolution indicates that the *B. coeca* genome likely retained the size of the ancestral Steganinae, i.e., a stronger reduction occurred in the Drosophilinae lineage containing *D. melanogaster* (Figures 2B and S2; Table S1).

To determine the number of protein-coding genes, we used four rounds of Maker,[33] supported by the training of the gene finding and prediction tools SNAP[34] and Augustus.[35] The annotation, made on the repeat-masked genome, yielded 10,349 protein-coding genes with an annotation edit distance (AED) ≤ 0.5 for 96.4% of our gene models and a Pfam domain found in 83.66% of the proteins (BUSCO score = 91%). Using the same strategy, we annotated two steganine genomes, namely *Phortica variegata* and *Leucophenga varia*. The annotation yielded 11,067 (BUSCO score = 91%) and 13,160 (BUSCO score = 90.8%) protein-coding genes, respectively. The annotation of the ephydrid *Ephydra gracilis* genome yielded 9,154 protein-coding genes (BUSCO score = 68.9%) (Figure S2). *Ephydra* is particular among Ephydroidea in adapting to hypersaline waters and associated algal flora.[36] Given the current lack of knowledge of ephydrid genetics, whether their low gene content is due to their high specialization or an artifact of incomplete annotation is hard to know. Regardless, the bee louse fly has the lowest number of protein-coding genes compared with other drosophilids (Student's t one-sample test $p < 1.5 \times 10^{-5}$; Shapiro-Wilk normality test $p = 0.48$) despite having a total genome size that is among the largest genomes in the family. Whereas the near-completeness of our *B. coeca* genome (~95%) might have reduced the number of annotated genes, low-complexity, hard-to-assemble genomic regions are usually mostly heterochromatic and poor in genes, e.g., centromeres, Y chromosomes, etc. A low gene content is also characteristic of bee genomes, compared with ants and wasps, with a remarkable trend of gene reduction within the family Apidae during the evolution of the genus *Apis* (Figures 2C and S2; Table S1).

TEs expanded in the bee louse fly, with one element horizontally transferred with the host

The bee louse fly's large genome size and low gene content suggest an increase in repetitive sequences. RepeatModeler and RepeatMasker analyses[37,38] indicated that nearly 41.34% of the *B. coeca* genome consists of such sequences, compared with 22.05% and 10.98% in *D. melanogaster* and *A. mellifera*, respectively (Figure 2D). Remarkably, half of the bee louse fly repetitive sequences consisted of long interspersed nuclear elements (LINEs) retrotransposons (14.94%). Although LINEs are usually among the most abundant transposable elements (TEs) after long terminal repeats (LTRs) within the Drosophilidae,[39] their values did not exceed what was found in *B. coeca* (we found the highest percentage in *Leucophenga varia* with 5.54%). It is at

(D) Proportions of transposable elements in the genomes of 42 dipteran and hymenopteran species. DNAs, DNA transposons; LCs, low-complexity elements; LINEs, long interspersed nuclear elements; LTRs, long terminal repeats; SINEs, short interspersed nuclear elements; sRNAs, small RNAs; SRs, single repeats; and Unclass., unclassified.

(E) Maximum-likelihood phylogeny of *Famar1*-like copies from 38 animal species. Filled circles indicate ultra-fast bootstrap values higher than 90%. See also Table S1 and Figure S2.

Please cite this article in press as: Bastide et al., The genome of the blind bee louse fly reveals deep convergences with its social host and illuminates *Drosophila* origins, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.01.034

present unclear what factors influence the diversity of TE landscapes among eukaryote species.[40] Nonetheless, this difference means that, whereas the bee louse fly has likely retained the ancestrally large genome size of the Drosophilidae, its TE constitution has largely evolved.

Because host-parasite relationships have repeatedly been invoked as a factor that may favor horizontal transfer of TEs,[41] we searched for evidence of such transfers between *B. coeca* and *A. mellifera*. We found one TE, a DNA transposon *Famar1-like* element, previously described in the earwig *Forficula auricularia*,[42] that belongs to the Tc1-mariner superfamily. This element showed a high similarity between *B. coeca* and *A. mellifera* but was absent in all other drosophilid species for which a genome is available in GenBank, which is highly suggestive of an acquisition through horizontal transfer (Figure S2). Indeed, phylogenetic analysis of multiple copies of this TE extracted from 37 widely divergent animal species (Figure 2E) supported a direct transfer event between *B. coeca* and *A. mellifera*, although the directionality of the transfer cannot be inferred because the elements from the two species form mutually exclusive monophyletic clades. Remarkably, all elements found in the genomes of four *A. mellifera* subspecies, including *A. m. carnica*, *A. m. caucasia*, *A. m. mellifera*, and *A. m. ligustica*, formed an exclusively monophyletic clade. The transfer time between *B. coeca* and *A. mellifera* likely preceded the dispersion of this element among the subspecies or even their differentiation 0.77 mya[43] if the element was ancestral in *A. mellifera*. On the other hand, we did not find any trace of this element or any other related element in any other *Apis* species, indicating that the maximal time of horizontal transfer likely does not surpass 2.73 (0.70–8.9) mya, i.e., the time of divergence between *A. mellifera* and its closest relative, *A. cerana* (Figure 2A). The tight ecological connection between the bee louse fly and its host may have favored this transfer, as was suggested for blood- or sap-sucking insects.[44,45]

Gene families with excess losses show striking cross-order parallelism

Despite their deep divergence, we tested whether parallel changes could explain the reduction of protein-coding genes in both the honey bee and the bee louse fly. We used OrthoFinder[46] to cluster orthologous proteins from the 25 hymenopteran and 17 dipteran species. We identified 19,010 orthogroups. Of these, 935 showed significant size evolution among the 42 species when analyzed using CAFE5[47] and after applying an error model that accounted for misassemblies and misannotations. To classify those orthogroups into functional categories, we extracted groups that contained *D. melanogaster* orthologs for which a molecular function, i.e., a gene group, was assigned in the FlyBase database[48] (see STAR Methods). Of 1,078 gene groups, 136 significantly deviated from the birth-death model estimated by CAFE5.

After correction for multiple testing, 17 gene groups had significant losses in the bee louse fly, with no group showing significant gains (Table 1). The reduction of most of these groups showed a striking parallelism with bees (Anthophila) in particular and hymenopterans in general (Table 1; Figure S3). The most significant groups were those involved in the chemical detection of taste (gustatory receptors [GRs] and divergent ionotropic receptors [IR-DIVs]) and odors (odorant receptors [ORs] and odorant binding proteins). The remaining groups included those involved in recognition and signaling, with a potential role in metabolism, immunity, and/or development, such as C-type lectins, serine proteases, and dorsal,[46] as well as ion and sugar transportations. Other groups are involved in detoxification, such as cytochrome P450, GST-C, and carboxylases.[47] Indeed, bees have evolved a reduced repertoire of immunity and detoxification genes, likely due to the evolution of social behavior and their life in an overprotective and clean shelter, i.e., the nest.[49,50] Cytochrome P450 genes are more expressed in foraging workers than in the castes that remain in the nest (i.e., the queen and nurse workers).[51] The reduction of peptidases in both the honey bee and the bee louse fly could also be due to the low protein content of some of their food, i.e., nectar and honey. We also noted an underrepresentation of chitin-binding domain proteins and chitinases in the bee louse fly and the honey bee. Cuticles could act as barriers against environmental toxins, which may not be highly encountered in the nest. Remarkably, *B. coeca* is unique among cyclorrhaphan dipterans as its pupa, similar to the honey bee's,[52] is contained in the unmodified cuticle of the third instar larva and no sclerotized puparium is formed.[6,7] Whereas assembly and annotation errors can bias general estimates of gene losses, they should not specifically target the gene families that are ecologically relevant to both the host and the inquiline.

Honey and wax feeding drove the loss of almost all bitter-tasting GRs

The two most significantly evolving gene families in the bee louse fly, i.e., GRs and IR-DIVs, allow the detection of soluble cues (Table 1). There are 60 GRs in *D. melanogaster*, of which nine and 49 receptors respond primarily to sweet and bitter tastes, respectively, and two receptors respond to carbon dioxide (CO_2).[53] The three categories clustered into 35 orthogroups (Figure 3A), whose phylogenetic analysis indicates that the ancestral drosophilid repertoire consisted of 6 sweet, 26 bitter, and 2 CO_2 GRs assuming functional conservation of gustatory categories (Figure 3A). We identified 11 GRs in the bee louse fly with no duplications, using InsectOR[54] and manual curation. These GRs could be classified according to their *D. melanogaster* orthologs into 3 sweet, 6 bitter, and 2 CO_2. That means that the *D. melanogaster* lineage disproportionally evolved more bitter receptors from the ancestral repertoire, whereas *B. coeca* disproportionally lost bitter receptors (Figure 3A). InsectOR inferred the number of GRs in the steganine species *L. varia* and *P. variegata* to be 21 and 26, respectively, further confirming that *B. coeca* has lost a significant portion of the ancestral GR repertoire (Figure S4). Honey bees have only 11 GRs, of which 7 are orthologous to sweet *Drosophila* GRs.[55] This is likely due to the bees' strong diet reliance on sweet floral nectars and honey.[56] The loss of *B. coeca* bitter GRs and its retention of 2 ancestral sweet receptors is a strong convergence with its host.

Ionotropic receptors are another major class of chemoreceptors. They are divided into antennal IRs, which are conserved across insects and are most likely involved in olfaction, and divergent IRs (IR-DIVs), which evolve rapidly and are mostly involved in the taste perception of carboxylic and amino acids. Only IR-DIVs showed a significant loss in *B. coeca* (Table 1). However, our knowledge about the function of the 42

Please cite this article in press as: Bastide et al., The genome of the blind bee louse fly reveals deep convergences with its social host and illuminates *Drosophila* origins, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.01.034

Table 1. Rapidly evolving gene groups in *Braula coeca* show parallel reduction with bees

		Evolution in *Braula*		
Gene group	Function	Change	FDR p value	Evolution in bees
Gustatory receptors	chemosensory	reduced	1.7×10^{-4}	reduced in bees
Divergent ionotropic receptors	chemosensory	reduced	1.7×10^{-4}	reduced in Hymenoptera
S1A non-peptidase homologs	immunity, morphogenesis	reduced	1.9×10^{-4}	reduced in bees (particularly in *A. mellifera*)
Cytochrome P450-CYP3 clan	detoxification	reduced	9.8×10^{-4}	reduced in *Apis* (particularly in *A. mellifera*)
Odorant receptors	chemosensory	reduced	9.9×10^{-4}	reduced in the Apidae
Odorant binding proteins	chemosensory	reduced	0.0012	reduced in bees
Cytosolic glutathione S-transferases	detoxification	reduced	0.0034	reduced in *Apis*
Ecdysteroid kinase-like	detoxification	reduced	0.0034	reduced in bees
C-type lectin-like	immunity	reduced	0.0072	reduced in Hymenoptera
SLC22 family of organic ions transporters	development, detoxification	reduced	0.0084	reduced in *Apis*
SLC2 family of hexose sugar transporters	metabolism	reduced	0.0144	–
S1A serine proteases-chymotrypsin-like	metabolism, immunity, morphogenesis	reduced	0.0144	reduced in bees
Chitin-binding domain-containing proteins	morphogenesis, immunity	reduced	0.0233	reduced in bees
Carboxylesterases	detoxification	reduced	0.0450	reduced in long-tongued bees (but not in *A. mellifera*)
S1A serine proteases-trypsin-like	metabolism, immunity, morphogenesis	reduced	0.0450	reduced in bees
Cytochrome P450-CYP4 clan	detoxification	reduced	0.0478	reduced in bees
Dorsal group	morphogenesis, immunity	reduced	0.0478	reduced in Hymenoptera

Gene groups were defined according to *D. melanogaster* genes clustered with orthologous sequences from 42 dipteran and hymenopteran genomes by OrthoFinder. Putative functions of each group are given following FlyBase definitions and references therein. Evolutionary rate was estimated by CAFE5, with p values corrected for multiple testing using false discovery rate (FDR) analysis. See also Figure S3.

D. melanogaster IR-DIVs is still limited.[57] We inferred the ancestral IR-DIV drosophilid repertoire to contain 29 receptors, of which only nine were retained in *B. coeca*. Remarkably, whereas we found almost no direct orthologs between Diptera and Hymenoptera for IR-DIVs (Figure S3), bees are known to have few IRs in general,[58] pointing to another possible taste convergence between the bee louse fly and its host.

One-fifth of ancestral ORs were lost, including one receptor that is involved in anti-ovarian response in *Drosophila melanogaster*

ORs are essential to detect volatile chemical cues from the environment. This family has expanded in the honey bee to reach 170.[59] However, only nine of the honey bee genes have orthologs with *D. melanogaster*, and phylogenetic analysis indicates that this common OR repertoire has been gradually reduced during the evolution of *Apis* (Table 1; Figure S3). The 60 ORs of *D. melanogaster* are clustered within 16 orthogroups (Figure 3B). We inferred the ancestral drosophilid OR repertoire to contain 44 ORs, with at least one representative for each orthogroup (Figure 3B). We identified in *B. coeca*, following InsectOR[54] and manual curation, 35 ORs in addition to *Orco*, i.e., one-fifth of the ancestral repertoire was lost. The number of ORs was 50 and 51 in the two closely related steganine species, *L. varia* and *P. varia*, respectively (Figure S4). *Braula* ORs were direct orthologs to 18 genes in *D. melanogaster* (Figure 3B). Judging from the response of these orthologs to different volatiles in *D. melanogaster*, as curated in the Database of Odorant Responses (DoOR),[60] and assuming the potential conservation of function, the retained bee louse fly ORs may respond to compounds produced by honey bee workers in a defense context (e.g., 1-hexanol, farnesol, and 2-heptanone)[61] and/or to compounds of floral, pollen, and nectar aromas, such as acetophenone and benzaldehyde, a major volatile of honey.[62,63] Two cases of tetraplications were observed. One case involved three recent duplications of genes orthologous to DmOr67b, a gene that is highly responsive in *D. melanogaster* to both acetophenone and 1-hexanol. The second case involved three successive duplications of a gene orthologous to DmOr74a, which responds in *D. melanogaster* larvae to 1-nonanol and 1-heptanol, the latter being a major brood volatile,[64] and 1-hexanol, a component of the alarm pheromone.[65] Of these three duplications, two were unique to *B. coeca* compared with its closely related steganine species (Figure S4). Low concentrations of isopentyl acetate, the main component of the alarm pheromone released by unstressed workers at the nest entrances, attract the parasitic nest beetle *Aethina tumida*,[66] suggesting that the detection of the host odors could be a common strategy among phylogenetically distant inquilines and parasites of social insects.

Whereas major molecular convergences could exist between the inquiline and its social host, divergent strategies to adapt to the eusocial lifestyle requirements are still needed. In honey

Please cite this article in press as: Bastide et al., The genome of the blind bee louse fly reveals deep convergences with its social host and illuminates *Drosophila* origins, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.01.034

CellPress

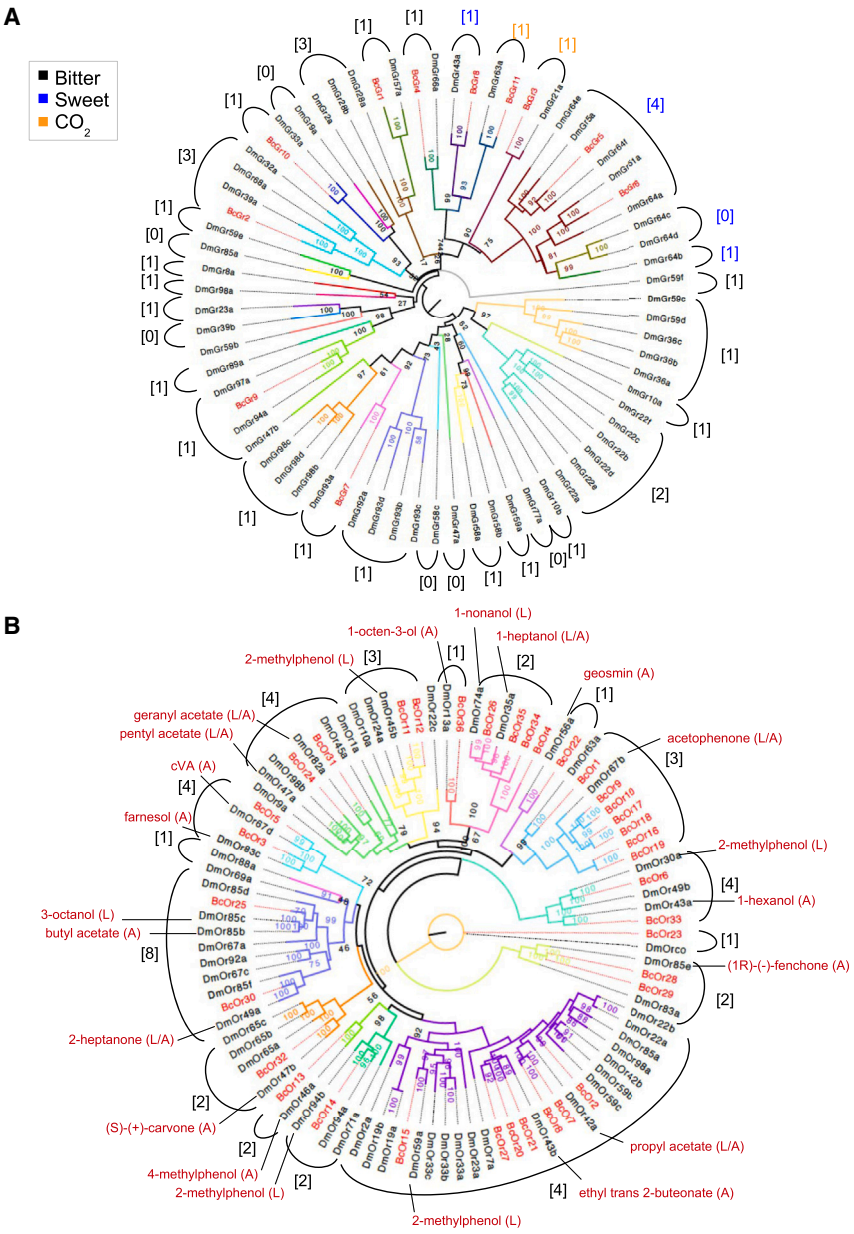


Figure 3. Evolution of chemosensory receptor gene families in *Braula coeca* and *Drosophila melanogaster*

(A) Maximum-likelihood phylogeny of gustatory receptors (GRs), with main taste categories color code given in a frame.

(B) Maximum-likelihood phylogeny of odorant receptors (ORs), with main ligands for each *D. melanogaster* receptor given in dark red. L, larva; A, adult expression.

For (A) and (B), ultra-fast bootstrap values are given above nodes. Branches are colored according to orthogroups defined by OrthoFinder for 42 dipteran and hymenopteran species. Numbers in broken brackets before each orthogroup reflect the presumed ancestral gene content inferred by phytools. See also Figure S4.

bees, colony cohesion is driven by the volatile queen's mandibular pheromone (QMP), which "sterilizes" the bee workers.[67] This pheromone elicits an anti-ovarian response in other insects, including *D. melanogaster*.[28] An RNA interference (RNAi) screen identified DmOr49b, DmOr56a, and DmOr98a to be potentially involved in the detection of the QMP compounds and the suppression of fecundity.[28,68] A *sine qua non* condition for a drosophilid to reproduce in a bee nest would, therefore, be to lose those receptors or to modify their response or effect. We found that the bee louse fly does not have an ortholog for DmOr98a, a receptor specific to the genus *Drosophila* (Figure S4). The bee louse fly has a pseudogene, orthologous to DmOr49b, that InsectOR identified. Orthologs of this *D. melanogaster* receptor are present and complete in all dipteran species, including *L. varia* and *P. variegata* (Figure S4). The bee louse fly had a receptor that we called BcOr22, which was orthologous to DmOr56a (Figure 3B). This last receptor is narrowly tuned in many *Drosophila* species to a single component, the mold volatile geosmin, whose perception also inhibits oviposition in *D. melanogaster*,[69] pointing to a possible conserved role in reproduction. Therefore, further functional analyses of the response of candidate ORs to various QMP compounds are required in both *D. melanogaster* and *B. coeca* to understand how modifications of these genes in *B. coeca* might have facilitated the evolution of the bee louse fly's inquilinism.

Blindness and life in a dark nest were accompanied by the loss of multiple rhodopsins

The species Latin name of the bee louse fly refers to the assumption that it was blind due to the reduction of the eye size and the loss of the ocelli. In agreement with reduced vision in the bee louse fly, we found only two out of the seven rhodopsin genes, which are responsible for colored vision and positive phototaxis in *D. melanogaster* and which were all present in the ancestral drosophilid repertoire. *D. melanogaster* orthologs of the *Rh1* and *Rh6* genes are expressed in the ommatidia and are sensitive to light.[70] The role of these opsins in light detection, despite the absence of ommatidia in the bee louse fly, is unclear. Remarkably, *Rh1* and *Rh6* are structurally required in mechanosensory bristles to control larval locomotion.[71] They also detect temperature.[72] Therefore, the retention of these rhodopsins in the bee louse fly could mainly be due to their unconventional functions. On the other hand, the rhodopsin *Rh2*, which is exclusively expressed in the ocelli and used for horizon detection in *D. melanogaster*,[73] is among those lost in the bee louse fly, in agreement with the loss of the ocelli. Regression of the visual system and its underlying opsin genes is common in animals inhabiting dark environments,

Please cite this article in press as: Bastide et al., The genome of the blind bee louse fly reveals deep convergences with its social host and illuminates *Drosophila* origins, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.01.034

such as fossorial mammals[74] and cavefishes,[75] representing a major example of deep convergences.

Apterism was not accompanied by the loss of major wing development genes

Small size, loss of wings, and the evolution of strongly clinging legs are all morphological changes that could prevent the honey bees from getting rid of bee lice.[76] All these potential adaptations are convergent with ectoparasitic true lice and, for some, such as apterism, represent major recurrent changes that have responded to distinct pressures throughout the history of insects.[77] We found intact most of the main wing development genes whose mutations severely reduce the wing in *D. melanogaster*, such as *wingless*, *apterous*, or *vestigial*. This means that the major morphological changes more likely resulted from regulatory changes of these core genes or modifications of other genes. Future developmental studies, specifically comparing the expression of wing and leg morphogenic genes between the bee louse fly and *D. melanogaster*, will definitively help shed light on the transcriptomic shifts underlying the major morphological changes of the bee louse fly.

Conclusions

That the enigmatic bee louse fly is indeed a drosophilid, a lineage within the most-investigated insect family, with more than 150 fully sequenced genomes, is undoubtedly one of the most exciting discoveries in dipteran phylogeny. How could a fly with an ancestral drosophilid genome become ecologically adapted to bees and morphologically similar to lice? Our results show that a mosaic of deep convergences at the genomic level underlies the relatively recent and dramatic changes of the bee louse fly to nest inquilinism. This mosaicism involved deep convergences with the host, mostly in genes likely involved in immunity, detoxification, and chemical perception, as well as convergences with general features of fossorial animals in the visual systems. Future developmental studies may elucidate whether general morphologies, such as apterism and leg modifications, could also be shared between *Braula* and other ectoparasites. Due to its genetic relatedness to *Drosophila* and ecological association to *Apis*, two major laboratory models, the new genomic resources presented here can help establish the bee louse fly as a promising model to address questions related to deep convergences that are difficult to approach in multiple highly specializing animals.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Sample collection and genomic library preparation
- METHOD DETAILS
  - Genome sequencing and assembly
  - Estimation of genome size and endosymbionts detection
  - Genome annotation
  - Phylogenomic analysis of the Ephydroidea
  - Reconstruction of the ancestral ecological niches
  - Phylogenomic analysis of Diptera and Hymenoptera
  - Genome size and gene content evolution
  - Transposons annotation and detection of Horizontal Transposon Transfer (HTT)
  - Gene family evolution
  - Chemosensory superfamilies evolution
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Genome size and protein-coding gene content analyses
  - Likelihood Ratio Test (LTR) comparison of CAFE5 models
  - False Discovery Rate (FDR) estimation of the p-values of the best CAFE5 model

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cub.2024.01.034.

ACKNOWLEDGMENTS

We are very grateful to Marcus Stensmyr for insightful comments on an early draft of the manuscript, the *Association Conservatoire de l'Abeille Noire Bretonne* (ACANB) for help collecting *B. coeca* flies, Etienne Minaud for photographic images of *B. coeca*, Héloïse Muller for help with the genome annotation pipeline, and Christian Cheminade and Michael Lang for help translating early literature on *Braula*. *Braula* genome sequencing was funded by a grant from Université Paris-Saclay (ADAPAR) to H.B.

AUTHOR CONTRIBUTIONS

Conceptualization, H.B. and A.Y.; investigation, H.B., H.L., N.D., D.O., C.P., J.F., C.G., and A.Y.; resources, H.L. and L.G.; writing – original draft, H.B., H.L., D.O., J.F., C.G., and A.Y.; writing – reviewing & editing, H.B., J.C., C.G., F.M.-P., F.R., J.-C.S., and A.Y.; visualization, H.B., J.F., C.G., and A.Y.; supervision, H.B.; funding acquisition, H.B.

DECLARATION OF INTERESTS

The authors declare no competing interests.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work, the authors used Grammarly (Grammarly Inc.) in order to improve language and readability. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Received: January 22, 2023
Revised: October 22, 2023
Accepted: January 12, 2024
Published: February 2, 2024

REFERENCES

1. Cini, A., Sumner, S., and Cervo, R. (2019). Inquiline social parasites as tools to unlock the secrets of insect sociality. Philos. Trans. R. Soc. Lond. B Biol. Sci. *374*, 20180193.

2. Rabeling, C., Schultz, T.R., Pierce, N.E., and Bacci, M. (2014). A Social Parasite Evolved Reproductive Isolation from Its Fungus-Growing Ant Host in Sympatry. Curr. Biol. *24*, 2047–2052.
3. Borowiec, M.L., Cover, S.P., and Rabeling, C. (2021). The evolution of social parasitism in Formica ants revealed by a global phylogeny. Proc. Natl. Acad. Sci. USA *118*, e2026029118.
4. Dejean, A., Orivel, J., Azémar, F., Hérault, B., and Corbara, B. (2016). A cuckoo-like parasitic moth leads African weaver ant colonies to their ruin. Sci. Rep. *6*, 23778.
5. Maruyama, M., and Parker, J. (2017). Deep-Time Convergence in Rove Beetle Symbionts of Army Ants. Curr. Biol. *27*, 920–926.
6. Skaife, S.H. (1922). On *Braula Coeca*, Nitzsch, a Dipterous parasite of the honey bee. Trans. R. Soc. S. Afr. *10*, 41–48.
7. Imms, A.D. (1942). On Braula coeca Nitsch and its affinities. Parasitology *34*, 88–100.
8. Bayless, K.M., Trautwein, M.D., Meusemann, K., Shin, S., Petersen, M., Donath, A., Podsiadlowski, L., Mayer, C., Niehuis, O., Peters, R.S., et al. (2021). Beyond Drosophila: resolving the rapid radiation of schizophoran flies with phylotranscriptomics. BMC Biol. *19*, 23.
9. Winkler, I.S., Kirk-Spriggs, A.H., Bayless, K.M., Soghigian, J., Meier, R., Pape, T., Yeates, D.K., Carvalho, A.B., Copeland, R.S., and Wiegmann, B.M. (2022). Phylogenetic resolution of the fly superfamily Ephydroidea–Molecular systematics of the enigmatic and diverse relatives of Drosophilidae. PLoS One *17*, e0274292.
10. Avalos, J., Rosero, H., Maldonado, G., and Reynaldi, F.J. (2019). Honey bee louse (Braula schmitzi) as a honey bee virus vector? J. Apic. Res. *58*, 427–429.
11. de Réaumur, R.-A.F. (1740). Mémoires pour servir à l'histoire des insectes*5* (de l'Imprimerie royale).
12. Nitzsch, C. L. (1818). Die Familien und Gattungen der Thierinsekten (insecta epizoica); als Prodromus einer Naturgeschichte derselben. Magazin der Entomologie *3*, 261–316.
13. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics *31*, 3210–3212.
14. Lewin, H.A., Robinson, G.E., Kress, W.J., Baker, W.J., Coddington, J., Crandall, K.A., Durbin, R., Edwards, S.V., Forest, F., Gilbert, M.T.P., et al. (2018). Earth BioGenome Project: Sequencing life for the future of life. Proc. Natl. Acad. Sci. USA *115*, 4325–4333.
15. Rhie, A., Walenz, B.P., Koren, S., and Phillippy, A.M. (2020). Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. *21*, 245.
16. Koren, S., Phillippy, A.M., Simpson, J.T., Loman, N.J., and Loose, M. (2019). Reply to 'Errors in long-read assemblies can critically affect protein prediction'. Nat. Biotechnol. *37*, 127–128.
17. Yassin, A. (2013). Phylogenetic classification of the Drosophilidae Rondani (Diptera): the role of morphology in the postgenomic era. Syst. Entomol. *38*, 349–364.
18. Rondani, C. (1856). Dipterologiae Italicae prodromus: Genera Italica Ordinis Dipterorum (Stocchi).
19. Egger, S. (1853). Himmlische Waffenrüstung für die Jugend bestehend aus den heiligen Sakramenten der Busse, des Altars und der Firmung: Ein praktischer Unterricht (Schmid).
20. Throckmorton, L. H. (1975). The phylogeny, ecology, and geography of *Drosophila*. In Handbook of Genetics (Plenum Press).
21. Keiper, J.B., Walton, W.E., and Foote, B.A. (2002). Biology and ecology of higher Diptera from freshwater wetlands. Annu. Rev. Entomol. *47*, 207–232.
22. Pollock, J.N. (2002). Observations on the biology and anatomy of Curtonotidae (Diptera: Schizophora). J. Nat. Hist. *36*, 1725–1745.
23. Markow, T.A., and O'Grady, P.M. (2005). Evolutionary genetics of reproductive behavior in *Drosophila*: connecting the dots. Annu. Rev. Genet. *39*, 263–291.
24. Ashburner, M. (1981). Entomophagous and other bizarre *Drosophilidae*. In The genetics and biology of Drosophila (Academic Press).
25. Duchêne, D., Duchêne, S., and Ho, S.Y.W. (2015). Tree imbalance causes a bias in phylogenetic estimation of evolutionary timescales using heterochronous sequences. Mol. Ecol. Resour. *15*, 785–794.
26. Shell, W.A., Steffen, M.A., Pare, H.K., Seetharam, A.S., Severin, A.J., Toth, A.L., and Rehan, S.M. (2021). Sociality sculpts similar patterns of molecular evolution in two independently evolved lineages of eusocial bees. Commun. Biol. *4*, 253.
27. Noll, F.B. (2002). Behavioral phylogeny of corbiculate Apidae (Hymenoptera; Apinae), with special reference to social behavior. Cladistics *18*, 137–153.
28. Galang, K.C., Croft, J.R., Thompson, G.J., and Percival-Smith, A. (2019). Analysis of the Drosophila melanogaster anti-ovarian response to honey bee queen mandibular pheromone. Insect Mol. Biol. *28*, 99–111.
29. Grimaldi, D., and Underwood, B.A. (1986). Megabraula, a new genus for two new species of Braulidae (Diptera), and a discussion of braulid evolution. Syst. Entomol. *11*, 427–438.
30. Dobson, J.R. (1999). A "bee-louse" Braula schmitzi örösi-pál (Diptera: Braulidae) new to the British Isles, and the status of Braula spp. in England and Wales. British Journal of Entomology & Natural History *11*, 139–148.
31. Kelley, J.L., Peyton, J.T., Fiston-Lavier, A.S., Teets, N.M., Yee, M.C., Johnston, J.S., Bustamante, C.D., Lee, R.E., and Denlinger, D.L. (2014). Compact genome of the Antarctic midge is likely an adaptation to an extreme environment. Nat. Commun. *5*, 4611.
32. Kim, B.Y., Wang, J.R., Miller, D.E., Barmina, O., Delaney, E., Thompson, A., Comeault, A.A., Peede, D., D'Agostino, E.R.R., Pelaez, J., et al. (2021). Highly contiguous assemblies of 101 drosophilid genomes. eLife *10*, e66405.
33. Cantarel, B.L., Korf, I., Robb, S.M.C., Parra, G., Ross, E., Moore, B., Holt, C., Sánchez Alvarado, A.S., and Yandell, M. (2008). MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. *18*, 188–196.
34. Korf, I. (2004). Gene finding in novel genomes. BMC Bioinformatics *5*, 59.
35. König, S., Romoth, L.W., Gerischer, L., and Stanke, M. (2016). Simultaneous gene finding in multiple genomes. Bioinformatics *32*, 3388–3395.
36. van Breugel, F., and Dickinson, M.H. (2017). Superhydrophobic diving flies (*Ephydra hians*) and the hypersaline waters of Mono Lake. Proc. Natl. Acad. Sci. USA *114*, 13483–13488.
37. Flynn, J.M., Hubley, R., Goubert, C., Rosen, J., Clark, A.G., Feschotte, C., and Smit, A.F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. Proc. Natl. Acad. Sci. USA *117*, 9451–9457.
38. Smit, A.F.A., Hubley, R., and Green, P. (2015). RepeatMasker Open-4.0. 2013–2015. http://www.repeatmasker.org/.
39. Mérel, V., Boulesteix, M., Fablet, M., and Vieira, C. (2020). Transposable elements in *Drosophila* Mobile. Mob. DNA *11*, 23.
40. Gilbert, C., Peccoud, J., and Cordaux, R. (2021). Transposable elements and the evolution of insects. Annu. Rev. Entomol. *66*, 355–372.
41. Venner, S., Miele, V., Terzian, C., Biémont, C., Daubin, V., Feschotte, C., and Pontier, D. (2017). Ecological networks to unravel the routes to horizontal transposon transfers. PLoS Biol. *15*, e2001536.
42. Barry, E.G., Witherspoon, D.J., and Lampe, D.J. (2004). A Bacterial Genetic Screen Identifies Functional Coding Sequences of the Insect *mariner* Transposable Element *Famar1* Amplified From the Genome of the Earwig, *Forficula auricularia*. Genetics *166*, 823–833.
43. Carr, S.M. (2023). Multiple mitogenomes indicate things fall apart with Out of Africa or Asia hypotheses for the phylogeographic evolution of honey bees (*Apis mellifera*). Sci. Rep. *13*, 9386.

44. Gilbert, C., Schaack, S., Pace, J.K., Brindley, P.J., and Feschotte, C. (2010). A role for host–parasite interactions in the horizontal transfer of transposons across phyla. Nature *464*, 1347–1350.
45. Gilbert, C., and Maumus, F. (2022). Multiple Horizontal Acquisitions of Plant Genes in the Whitefly Bemisia tabaci. Genome Biol. Evol. *14*, evac141.
46. Emms, D.M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. *20*, 238.
47. Mendes, F.K., Vanderpool, D., Fulton, B., and Hahn, M.W. (2021). CAFE 5 models variation in evolutionary rates among gene families. Bioinformatics *36*, 5516–5518.
48. Thurmond, J., Goodman, J.L., Strelets, V.B., Attrill, H., Gramates, L.S., Marygold, S.J., Matthews, B.B., Millburn, G., Antonazzo, G., Trovisco, V., et al. (2019). FlyBase 2.0: the next generation. Nucleic Acids Res. *47*, D759–D765.
49. Evans, J.D., Aronstein, K., Chen, Y.P., Hetru, C., Imler, J.L., Jiang, H., Kanost, M., Thompson, G.J., Zou, Z., and Hultmark, D. (2006). Immune pathways and defence mechanisms in honey bees *Apis mellifera*. Insect Mol. Biol. *15*, 645–656.
50. Berenbaum, M.R., and Johnson, R.M. (2015). Xenobiotic detoxification pathways in honey bees. Curr. Opin. Insect Sci. *10*, 51–58.
51. Chan, Q.W.T., Chan, M.Y., Logan, M., Fang, Y., Higo, H., and Foster, L.J. (2013). Honey bee protein atlas at organ-level resolution. Genome Res. *23*, 1951–1960.
52. Winston, M.L. (1987). The Biology of the Honey Bee (Harvard University Press).
53. Weiss, L.A., Dahanukar, A., Kwon, J.Y., Banerjee, D., and Carlson, J.R. (2011). The Molecular and Cellular Basis of Bitter Taste in Drosophila. Neuron *69*, 258–272.
54. Karpe, S.D., Tiwari, V., and Ramanathan, S. (2021). InsectOR—Webserver for sensitive identification of insect olfactory receptor genes from non-model genomes. PLoS One *16*, e0245324.
55. Sadd, B.M., Barribeau, S.M., Bloch, G., de Graaf, D.C., Dearden, P., Elsik, C.G., Gadau, J., Grimmelikhuijzen, C.J., Hasselmann, M., Lozier, J.D., et al. (2015). The genomes of two key bumblebee species with primitive eusocial organization. Genome Biol. *16*, 76.
56. Robertson, H.M., and Wanner, K.W. (2006). The chemoreceptor superfamily in the honey bee, *Apis mellifera*: expansion of the odorant, but not gustatory, receptor family. Genome Res. *16*, 1395–1403.
57. Ni, L. (2020). The structure and function of ionotropic receptors in *Drosophila*. Front. Mol. Neurosci. *13*, 638839.
58. Park, D., Jung, J.W., Choi, B.S., Jayakodi, M., Lee, J., Lim, J., Yu, Y., Choi, Y.S., Lee, M.L., Park, Y., et al. (2015). Uncovering the novel characteristics of Asian honey bee, *Apis cerana*, by whole genome sequencing. BMC Genomics *16*, 1.
59. Karpe, S.D., Jain, R., Brockmann, A., and Sowdhamini, R. (2016). Identification of Complete Repertoire of Apis florea Odorant Receptors Reveals Complex Orthologous Relationships with Apis mellifera. Genome Biol. Evol. *8*, 2879–2895.
60. Münch, D., and Galizia, C.G. (2016). DoOR 2.0 - Comprehensive mapping of *Drosophila melanogaster* odorant responses. Sci. Rep. *6*, 21841.
61. Wang, Z., and Tan, K. (2019). Honey Bee Alarm Pheromone Mediates Communication in Plant–Pollinator–Predator Interactions. Insects *10*, 366.
62. Machado, A.M., Miguel, M.G., Vilas-Boas, M., and Figueiredo, A.C. (2020). Honey Volatiles as a Fingerprint for Botanical Origin—A Review on their Occurrence on Monofloral Honeys. Molecules *25*, 374.
63. Starowicz, M., Hanus, P., Lamparski, G., and Sawicki, T. (2021). Characterizing the Volatile and Sensory Profiles, and Sugar Content of Beeswax, Beebread, Bee Pollen, and Honey. Molecules *26*, 3410.
64. Noël, A., Dumas, C., Rottier, E., Beslay, D., Costagliola, G., Ginies, C., Nicolè, F., Rau, A., Le Conte, Y.L., and Mondet, F. (2023). Detailed chemical analysis of honey bee (*Apis mellifera*) worker brood volatile profile from egg to emergence. PLoS One *18*, e0282120.
65. Collins, A.M., and Blum, M.S. (1983). Alarm responses caused by newly identified compounds derived from the honeybee sting. J. Chem. Ecol. *9*, 57–65.
66. Torto, B., Boucias, D.G., Arbogast, R.T., Tumlinson, J.H., and Teal, P.E.A. (2007). Multitrophic interaction facilitates parasite–host relationship between an invasive beetle and the honey bee. Proc. Natl. Acad. Sci. USA *104*, 8374–8378.
67. Van Oystaeyen, A., Oliveira, R.C., Holman, L., van Zweden, J.S., Romero, C., Oi, C.A., d'Ettorre, P., Khalesi, M., Billen, J., Wäckers, F., et al. (2014). Conserved Class of Queen Pheromones Stops Social Insect Workers from Reproducing. Science *343*, 287–290.
68. Camiletti, A.L., Percival-Smith, A., Croft, J.R., and Thompson, G.J. (2016). A novel screen for genes associated with pheromone-induced sterility. Sci. Rep. *6*, 36041.
69. Stensmyr, M.C., Dweck, H.K.M., Farhan, A., Ibba, I., Strutz, A., Mukunda, L., Linz, J., Grabe, V., Steck, K., Lavista-Llanos, S., et al. (2012). A Conserved Dedicated Olfactory Circuit for Detecting Harmful Microbes in Drosophila. Cell *151*, 1345–1357.
70. Senthilan, P.R., and Helfrich-Förster, C. (2016). Rhodopsin 7–The unusual Rhodopsin in Drosophila. PeerJ *4*, e2427.
71. Zanini, D., Giraldo, D., Warren, B., Katana, R., Andrés, M., Reddy, S., Pauls, S., Schwedhelm-Domeyer, N., Geurten, B.R.H., and Göpfert, M.C. (2018). Proprioceptive Opsin Functions in Drosophila Larval Locomotion. Neuron *98*, 67–74.e4.
72. Leung, N.Y., and Montell, C. (2017). Unconventional Roles of Opsins. Annu. Rev. Cell Dev. Biol. *33*, 241–264.
73. Mishra, A.K., Fritsch, C., Voutev, R., Mann, R.S., and Sprecher, S.G. (2021). Homothorax controls a binary Rhodopsin switch in Drosophila ocelli. PLoS Genet. *17*, e1009460.
74. Partha, R., Chauhan, B.K., Ferreira, Z., Robinson, J.D., Lathrop, K., Nischal, K.K., Chikina, M., and Clark, N.L. (2017). Subterranean mammals show convergent regression in ocular genes and enhancers, along with adaptation to tunneling. eLife *6*, e25884.
75. Policarpo, M., Fumey, J., Lafargeas, P., Naquin, D., Thermes, C., Naville, M., Dechaud, C., Volff, J.N., Cabau, C., Klopp, C., et al. (2021). Contrasting Gene Decay in Subterranean Vertebrates: Insights from Cavefishes and Fossorial Mammals. Mol. Biol. Evol. *38*, 589–605.
76. Büscher, T.H., Petersen, D.S., Bijma, N.N., Bäumler, F., Pirk, C.W.W., Büsse, S., Heepe, L., and Gorb, S.N. (2022). The exceptional attachment ability of the ectoparasitic bee louse Braula coeca (Diptera, Braulidae) on the honeybee. Physiol. Entomol. *47*, 83–95.
77. Roff, D.A. (1990). The Evolution of Flightlessness in Insects. Ecol. Monogr. *60*, 389–421.
78. Leger, A., and Leonardi, T. (2019). pycoQC, interactive quality control for Oxford Nanopore Sequencing. JOSS *4*, 1236.
79. Zimin, A.V., Puiu, D., Luo, M.C., Zhu, T., Koren, S., Marçais, G., Yorke, J.A., Dvořák, J., and Salzberg, S.L. (2017). Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. Genome Res. *27*, 787–792.
80. Kokot, M., Długosz, M., and Deorowicz, S. (2017). KMC 3: counting and manipulating k-mer statistics. Bioinformatics *33*, 2759–2761.
81. Ranallo-Benavidez, T.R., Jaron, K.S., and Schatz, M.C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat. Commun. *11*, 1432.
82. Laetsch, D.R., and Blaxter, M.L. (2017). BlobTools: Interrogation of genome assemblies. F1000Res. *6*, 1287.
83. Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods *12*, 59–60.
84. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100.
85. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421.

86. Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat. Biotechnol. *29*, 644–652.

87. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780.

88. Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol. Biol. Evol. *37*, 1530–1534.

89. Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., and Jermiin, L.S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods *14*, 587–589.

90. Pagel, M., and Meade, A. (2006). Bayesian analysis of correlated evolution of discrete characters by reversible-jump markov chain monte carlo. Am. Nat. *167*, 808–825.

91. Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. *24*, 1586–1591.

92. Revell, L.J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol. Evol. *3*, 217–223.

93. Gilleland, E., and Katz, R.W. (2016). extRemes 2.0: An Extreme Value Analysis Package in R. Journal. J. Stat. Soft. *72*, 1–39.

94. Slater, G.S.C., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics *6*, 31.

95. Murray, M.H., and Blume, J.D. (2021). FDRestimation: Flexible False Discovery Rate Computation in R. F1000Res *10*, 441.

96. Miller, D.E., Staber, C., Zeitlinger, J., and Hawley, R.S. (2018). Highly Contiguous Genome Assemblies of 15 Drosophila Species Generated Using Nanopore Sequencing. G3 (Bethesda) *8*, 3131–3141.

97. Lu, H., Giordano, F., and Ning, Z. (2016). Oxford Nanopore MinION Sequencing and Genome Assembly. Genomics Proteomics Bioinformatics *14*, 265–279.

98. Muller, H., Ogereau, D., Da Lage, J.L., Capdevielle, C., Pollet, N., Fortuna, T., Jeannette, R., Kaiser, L., and Gilbert, C. (2021). Draft nuclear genome and complete mitogenome of the Mediterranean corn borer, Sesamia nonagrioides, a major pest of maize. G3 (Bethesda) *11*, jkab155.

99. Boekel, J., Chilton, J.M., Cooke, I.R., Horvatovich, P.L., Jagtap, P.D., Käll, L., Lehtiö, J., Lukasse, P., Moerland, P.D., and Griffin, T.J. (2015). Multi-omic data analysis using Galaxy. Nat. Biotechnol. *33*, 137–139.

100. Hiltemann, S., Rasche, H., Gladman, S., Hotz, H.R., Larivière, D., Blankenberg, D., Jagtap, P.D., Wollmann, T., Bretaudeau, A., Goué, N., et al. (2023). Galaxy Training: A powerful framework for teaching! PLoS Comput. Biol. *19*, e1010752.

101. Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., and Vinh, L.S. (2018). UFBoot2: improving the ultrafast bootstrap approximation. Mol. Biol. Evol. *35*, 518–522.

102. Okada, T. (1962). Bleeding sap preference of the Drosophilid flies. Jpn. J. Appl. Entomol. Zool. *6*, 216–229.

103. Misof, B., Liu, S., Meusemann, K., Peters, R.S., Donath, A., Mayer, C., Frandsen, P.B., Ware, J., Flouri, T., Beutel, R.G., et al. (2014). Phylogenomics resolves the timing and pattern of insect evolution. Science *346*, 763–767.

104. Suvorov, A., Kim, B.Y., Wang, J., Armstrong, E.E., Peede, D., D'Agostino, E.R.R., Price, D.K., Waddell, P.J., Lang, M., Courtier-Orgogozo, V., et al. (2022). Widespread introgression across a phylogeny of 155 *Drosophila* genomes. Curr. Biol. *32*, 111–123.e5.

105. Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. PeerJ *4*, e2584.

106. Lampe, D.J., Witherspoon, D.J., Soto-Adames, F.N., and Robertson, H.M. (2003). Recent horizontal transfer of mellifera subfamily mariner transposons into insect lineages representing four different orders shows that selection acts only during horizontal transfer. Mol. Biol. Evol. *20*, 554–562.

107. Zhang, H.H., Peccoud, J., Xu, M.R., Zhang, X.G., and Gilbert, C. (2020). Horizontal transfer and evolution of transposable elements in vertebrates. Nat. Commun. *11*, 1362.

108. Edgar, R.C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics *5*, 113.

109. Montell, C. (2009). A taste of the *Drosophila* gustatory receptors. Curr. Opin. Neurobiol. *19*, 345–353.

Please cite this article in press as: Bastide et al., The genome of the blind bee louse fly reveals deep convergences with its social host and illuminates *Drosophila* origins, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.01.034

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
15 *B. coeca* individuals from the Island of Ouessant, France.	*Association Conservatoire de l'Abeille Noire Bretonne* (A.C.A.N.B.)	N/A
Chemicals, peptides, and recombinant proteins		
Nucleobond AXG20 kit	Macherey-Nagel	740544
buffer set IV	Macherey-Nagel	740604
SRE XS kit	Circulomics	SKU 102-208-200
Ligation Sequencing kit	Oxford Nanopore Technology	SQK-LSK110
R9.4.1 flow cell	Oxford Nanopore Technology	FLO-Min106
Illumina paired-end sequencing	Novogene Company Limited	N/A
Deposited data		
Braula genome sequence	This paper	NCBI: PRJNA1000103
Commands of all programs	This paper	https://doi.org/10.5281/zenodo.10433104
Genome assemblies and all associated data	This paper	https://doi.org/10.6084/m9.figshare.c.6997056.v1
Software and algorithms		
Guppy v5.0.11		https://nanoporetech.com
PycoQC	Leger and Leonardi [78]	https://github.com/a-slide/pycoQC
FastQC		http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
MaSuRCA v4.0.3	Zimin et al.[79]	https://genome.umd.edu/masurca.html
BUSCO v5.0.0	Simão et al.[13]	http://busco.ezlab.org
Merqury	Rhie et al.[15]	https://github.com/marbl/merqury
KMC 3	Kokot et al.[80]	https://github.com/refresh-bio/KMC
GenomeScope v2.0	Ranallo-Benavidez et al.[81]	https://github.com/tbenavi1/genomescope2.0
Smudgeplot	Ranallo-Benavidez et al.[81]	https://github.com/KamilSJaron/smudgeplot
Blobtools	Laetsch and Blaxter[82]	https://github.com/DRL/blobtools
Diamond	Buchfink et al.[83]	https://github.com/bbuchfink/diamond
Minimap2	Li[84]	https://github.com/lh3/minimap2
Maker v2.31.10	Cantarel et al.[33]	https://www.yandell-lab.org/software/maker.html
SNAP v.2006-07-28	Korf[34]	https://github.com/KorfLab/SNAP
Augustus v.3.3.3	König et al.[35]	http://bioinf.uni-greifswald.de/augustus/
RepeatModeler v2.0.1	Flynn et al.[37]	https://github.com/Dfam-consortium/RepeatModeler http://www.repeatmasker.org/RepeatModeler/
RepeatMasker v4.0.9	Smit et al.[38]	https://www.repeatmasker.org/RepeatMasker/
BLAST 2.9.0+	Camacho et al.[85]	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST
Trinity	Grabherr et al.[86]	https://github.com/trinityrnaseq/trinityrnaseq/releases
Supermatrix.pl	This study	https://github.com/AmirYassinLab/Supermatrix
MAFFT	Katoh and Standley[87]	https://mafft.cbrc.jp/alignment/software/
IqTREE 2	Minh et al.[88]	http://www.iqtree.org/
ModelFinder	Kalyaanamoorthy et al.[89]	Implemented in IqTREE 2 http://www.iqtree.org/
BayesTraits v.4	Pagel and Meade[90]	http://www.evolution.reading.ac.uk/BayesTraitsV4.1.1/BayesTraitsV4.1.1.html
OrthoFinder	Emms and Kelly[46]	https://github.com/davidemms/OrthoFinder

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
MCMCTree	Yang[91]	http://abacus.gene.ucl.ac.uk/software/paml.html
Phytools	Revell[92]	https://cran.r-project.org/package=phytools
CAFE v. 5	Mendes et al.[47]	https://github.com/hahnlab/CAFE5/releases
OG2GG.pl	This study	https://github.com/AmirYassinLab/OG2GG
extRemes	Gilleland[93]	https://cran.r-project.org/package=extRemes
FDRestimation	Murray and Blume[94]	https://cran.r-project.org/package=FDRestimation
Exonerate ver. 2.2	Slater and Birney[95]	https://www.ebi.ac.uk/about/vertebrate-genomics/software/exonerate
InsectOR	Karpe et al.[54]	https://github.com/sdk15/insectOR

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Héloïse Bastide (heloise.bastide@universite-paris-saclay.fr).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Raw sequence data are deposited on NCBI Sequence Read Archive (SRA). Bioproject accession number is listed in the key resources table. Genome assemblies and all data associated to this study including translation of early taxonomic literature are deposited in Figshare. DOI is listed in the key resources table.
- All original code and commands for all programs have been deposited at Github depository. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Sample collection and genomic library preparation

Samples of *Braula coeca* were collected from honey bee colonies on the Island of Ouessant in France and kindly provided to us by the *Association Conservatoire de l'Abeille Noire Bretonne* (A.C.A.N.B.). Genomic DNA was extracted from 15 unsexed individuals conserved in alcohol using the Nucleobond AXG20 kit and buffer set IV from Macherey-Nagel (ref. 740544 and 740604, https://www.mn-net.com, Düren, Germany).

METHOD DETAILS

Genome sequencing and assembly

We used a hybrid approach to assemble a draft genome of *B. coeca* using both long-read Oxford Nanopore Technology (ONT) and short-read Illumina sequencing.[96] Before nanopore sequencing, a size selection was conducted on the DNA using the SRE XS kit from Circulomics (https://www.circulomics.com/, Baltimore, Maryland, USA). The Ligation Sequencing kit SQK-LSK110 from ONT (https://nanoporetech.com/)[97] was then used to prepare the samples for nanopore sequencing following the manufacturer's protocol. The library was loaded and sequenced on an R9.4.1 flow cell (ref. FLO-Min106) for sequencing. Raw data were basecalled using Guppy v5.0.11 and the "sup" algorithm. The ONT raw data size was 4.4 Gb in 1,399,323 reads (mean read length 3,146 kb, longest read of 123.3 Mb), with an N50 of 4,677 kb. Phred scores ranged from 8 to 18, with a median of 13, as assessed by PycoQC.[78] Illumina paired-end sequencing was performed by Novogene Company Limited (https://en.novogene.com, Cambridge, UK) on the same DNA sample. The Illumina sequencing produced 119,719,537 paired 150 bp reads. Phred scores averaged 36 per read as analyzed by FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). We used MaSuRCA v4.0.3[79] to produce the hybrid assembly of our genome using the Cabog assembler. We obtained a final assembly size of 309,35Mb in 2477 contigs, with a N50 of 347,227 bp. The completeness of the assembly was estimated to 95.8% with BUSCO v5.0.0 on the *diptera_odb10* dataset (C:95.8%[S:94.5%,D:1.3%],F:0.7%,M:3.5%,n:3285), and to 93.6% using Merqury.

Please cite this article in press as: Bastide et al., The genome of the blind bee louse fly reveals deep convergences with its social host and illuminates *Drosophila* origins, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.01.034

Estimation of genome size and endosymbionts detection

K-mers frequencies within short-read data were obtained with KMC 3.[80] Genome size and ploidy were inferred using GenomeScope v2.0 with k-mer size = 21 and Smudgeplot.[81] Contig taxonomy was performed using Blobtools[82] with Diamond as search engine[83] against the UniProt database using a local copy of the NCBI TaxID file for the taxonomic assignation of the best hit. Minimap2[84] was used for read mapping (Figure S1).

Genome annotation

The *B. coeca* genome was annotated using Maker v2.31.10,[33] following Muller et al.'s[98] protocol, wherein multiple rounds of Maker supported by the training of the SNAP v.2006-07-28[34] and Augustus v.3.3.3[35] gene finding and prediction tools, were conducted. RepeatModeler v2.0.1 was first used to identify the repeat-enriched regions that were masked by RepeatMasker v4.0.9 as implemented in Maker. Proteomes of five *Drosophila* species, namely *D. innubila*, *D. albomicans*, *D. bipectinata*, *D. melanogaster*, and *D. virilis* were obtained from NCBI and used to guide the annotation. Protein-Protein BLAST 2.9.0+[85] (-evalue 1e-6 -max_hsps 1 -max_target_seqs 1) was then used to assess putative protein functions in *B. coeca* by comparing the protein sequences given by Maker to the protein sequences from the annotated genome of *D. melanogaster*. The completeness of genome annotation was assessed using BUSCO at each round and the round with the highest score was retained.

Phylogenomic analysis of the Ephydroidea

Besides our *B. coeca* assembly, we obtained from NCBI repository genome assemblies for 12 species, transcriptome shotgun assemblies (TSA) for four species, and sequence read runs (SRR) for three species (Table S1). Paired-end DNA raw data of two species, namely *Rhinoleucophenga* cf. *bivisualis* and *Cacoxenus indagator* were assembled using MaSuRCA with default parameters. The transcriptome of *Acletoxenus* sp. was assembled using Trinity software package[86] on the Galaxy Europe website[99] following standard protocol.[100] BUSCO v.5.0 was used to assess the completeness of those assemblies and to extract single-copy BUSCO genes for all species. Protein sequences of 3,100 single and complete BUSCO genes were aligned using MAFFT[87] and concatenated into a single supermatrix (2,557,349 amino acids). A maximum-likelihood (ML) phylogeny was then inferred for the supermatrix using IqTREE 2[88] with 1,000 ultrafast bootstrap iterations[101] and using the JTT+R substitution model inferred by ModelFinder[89] implemented by IqTree.

Reconstruction of the ancestral ecological niches

For each of the 20 ephydroid species we obtained a predominant ecological niche from the taxonomic literature.[20,23,24,102] Eight predominant niches were coded as discrete traits, and the Multistate program of the BayesTraits v.4 package[90] was used under the Reverse Jump MCMC model with 1,010,000 chain iterations and a burnin sample of 10,000.

Phylogenomic analysis of Diptera and Hymenoptera

The second phylogenomic analysis involved, besides *B. coeca*, 25 hymenopteran and 16 dipteran species for which an assembly can be downloaded from the NCBI Genome repository (Table S1). Protein sequences for all species but three, namely *Leucophenga varia*, *Phortica variegata*, and *Ephydra gracilis*, were obtained from NCBI. For these three species, we used the same four-round annotation procedure that we used for *B. coeca* to identify protein-coding genes and translate their sequences. We used BUSCO to assess the completeness of all annotated and downloaded genomes and their corresponding assemblies. OrthoFinder[46] was used to generate protein sequences of protein-coding-genes of the 42 species and to cluster these sequences into orthogroups. Only the longest isoform (*i.e.* the primary transcript) was used for genes with multiple isoforms. 79 orthogroups contained a single copy ortholog from each species, and their protein sequences were aligned using MAFFT and concatenated into a single supermatrix (63,192 amino acids). A maximum-likelihood (ML) phylogeny was then inferred for the supermatrix using IqTREE 2[88] with 1,000 ultrafast bootstrap iterations[101] and the JTT+R substitution model inferred by ModelFinder[89] implemented by IqTree.

MCMCTree[91] was used to date the inferred ML trees based on recently published fossil-calibrated phylogenies. First, two time points were obtained for the 42-species phylogeny. These included the divergence between ants and bees between 90-120 myr ago[103] and between *Scaptodrosophila* and *Drosophila* between 50-56 myr ago,[104] with a maximum root age for the ancestor of Hymenoptera and Diptera at 344 myr ago.[103]

Genome size and gene content evolution

Genome size and gene content (number of OrthoFinder generated protein-coding-genes after retaining the longest isoform for genes with multiple transcripts) inferred for each of the 42 dipteran and hymenopteran genomes were mapped on the phylogenetic tree, and values at the ancestral nodes were inferred and visualized using the fastAnc command in the R package Phytools v0.2.2.[92]

Transposons annotation and detection of Horizontal Transposon Transfer (HTT)

Transposons were identified in the 42 dipteran and hymenopteran genomes following a two-step protocol. First, we used RepeatModeler v2.0.1[37] with default parameters to generate a *de novo* library of repetitive regions. RepeatMasker v 4.0.9[37] was then run with the newly generated library and the options -a (create a.align output file) and -s (slow search; more sensitive) to create a summary of the families of transposable elements found in each genome along with the percentage of the genome they represent. To detect possible HTT between *Braula coeca* and *Apis mellifera*, we used the *B. coeca* whole genome as query to perform a blastn

Please cite this article in press as: Bastide et al., The genome of the blind bee louse fly reveals deep convergences with its social host and illuminates *Drosophila* origins, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.01.034

similarity search against the whole *A. mellifera* genome (all default options, including "-task megablast"). All *B. coeca* genome regions longer than 299 bp and aligning to *A. mellifera* with an e-value lower than 0.0001 were extracted and clustered at 80% nucleotide identity threshold with vsearch.[105] The consensus sequence of each of the 50 resulting clusters were used as queries to perform blastx searches on the non-redundant protein database of NCBI using Diamond.[83] A total of eight consensus sequences had best hits to the *Famar1* element previously described in the earwig *Forficula auricularia*, known to be also present in *A. mellifera* as a result of horizontal transfer.[42,106] To verify that the *Famar1*-like element from *B. coeca* has indeed been involved in HTT, we compared the *Famar1*-like synonymous distance (dS) to a distribution of dS expected under vertical transmission since the last common ancestor of *B. coeca* and *A. mellifera* following the approach developed in Zhang et al.[107] This approach assumes that in case of HTT, TE dS should be much lower than dS expected under vertical transmission. Briefly, we calculated the dS over the transposase open reading frame between one copy of the *Famar1*-like element extracted from *C. coeca* and another copy of this element from the *A. mellifera* genome. We then compared this distance to the distribution of dS calculated over 2,179 alignments between single copy BUSCO genes that produce best reciprocal hits in blastp similarity searches.[107] We found that the *Famar1*-like dS (=0.12) fall below the 0.5% quantile (=1.76) of the distribution of dS calculated for orthologous genes (Figure S2), confirming that the element has been acquired through HTT in *B. coeca* and *A. mellifera*. To assess whether the tight ecological interactions existing between *B. coeca* and *A. mellifera* might have favored direct transfer of this element between the two species, we assessed how closely related are *B. coeca Famar1*-like copies to those from *A. mellifera*. We first screened for the presence of this element in other animal genomes. We used the *Famar1* sequence[42] as query to perform online blastn similarity searches (all default options, including "-task megablast") on a total of 8,180 animal genomes belonging to 11 insect orders as well as to Annelida, Chelicerata, Chiroptera, Cnidaria, Myriapoda, Nematoda, Platyhelminthes and Teleostei. We found full length copies showing >79% of nucleotide identity to this element in a total of 37 species. We aligned up to ten copies from each genome the most similar to *Famar1* using Muscle.[108] We then reconstructed a maximum-likelihood phylogeny of these copies using IqTree after nucleotide model detection using ModelFinder. Node support was quantified using ultrafast bootstrap as implemented in IqTree.

Gene family evolution

We used CAFE v. 5[47] to model and infer gene family evolution. We conducted CAFE5 using an error model on the 19,011 orthgroups generated by OrthoFinder and using the time-calibrated phylogenetic tree of the 42 dipteran and hymenopteran species. The analysis showed that *B. coeca* has gained 439 while losing 1,517 protein-coding genes in agreement with the low gene content of this species. To gain further functional insights on the ecological or biological relevance of the evolving genes, we grouped OrthoFinder orthogroups into functional gene groups using the customized perl script OG2GG.pl (https://github.com/AmirYassinLab/OG2GG) leveraging the proximity of *B. coeca* to *D. melanogaster*. The script assigns each *D. melanogaster* ortholog to its largest functional gene groups in Flybase[48] ("gene_group_data_fb_2023_02.tsv") and then assigns each orthogroup to the largest gene group of its constituent genes. *D. melanogaster* has 13,545 protein-coding genes that were clustered into 10,497 orthogroups. However, 8,202 *D. melanogaster* genes are assigned to at least one of 10,670 functional gene groups in the FlyBase database, of which some concern RNA genes that, by definition, are not analyzed by OrthoFinder. Because of the hierarchical nature of the functional gene groups annotation in FlyBase as well as to the pleiotropy of certain genes, each *D. melanogaster* gene was assigned to its largest group, *i.e.* the group with biggest number of genes. Consequently, 5,733 protein-coding genes were assigned to an orthogroup and a gene group. Because some orthogroups can have multiple genes with some assigned to different gene groups, each orthogroup was assigned to the largest gene group of its constituent genes. Orthogroups were then clustered according to their assigned functional groups, e.g., the odorant receptors family contained 16 orthgroups (and 60 *D. melanogaster* genes). Because some of the genes found in the orthogroups based on their sequence similarity have no functional annotation in Flybase, the total number of *D. melanogaster* protein-coding genes to be grouped into gene groups was 7,820 genes (and 6,317 protein-coding genes for *B. coeca*).

CAFE5 was then run on the gene groups' gene counts using four different birth rate models in an increasing order (lambda = 1, 2, 3 and 4) and the error model to correct for possible assembly and annotation errors. For each model we run four iterations. The likelihood of only the two simplest models, *i.e.* one- and two-lambda models, converged across the four iterations. Likelihood ratio test using the lr.test function of the extRemes R package[93] showed that the two-lambda better fit our data. This model imposed a different rate for only *B. coeca* compared to the rest of the tree and it was chosen for four-iterations of subsequent analyses using the estimated error rate. Multiple testing corrections were conducted using the False Discovery Rate (FDR) analysis of the FDRestimation[95] package implemented in R.

Chemosensory superfamilies evolution

To curate *B. coeca* gustatory receptors (GR) and odorant receptors (OR) genes, we queried *D. melanogaster* GRs and ORs protein sequences on *B. coeca*, *L. varia* and *P. variegata* assemblies using Exonerate ver. 2.2[94] with option –maxintron 2000 and -p pam250. The output, along with the assembly, were fed to InsectOR[54] with option 7tm_7 and 7tm_6 activated for GR and OR analyses, respectively. From the output files, we extracted 300-500 amino acids-long complete sequences with 7tm_7 or 7tm_6 motif detected and with start codon present and no internal stop codon, *i.e.* pseudogenes excluded. Protein sequences were aligned using MAFFT and a maximum-likelihood phylogenetic tree for each family using IqTREE 2 with the same options as the phylogenomic analysis. The literature was reviewed to classify GRs into bitter, sweet, and CO_2 categories[109] and identify volatile ligands eliciting the strongest response in odorant neurons in *D. melanogaster*.[60] Because CAFE5 inferred ancestral counts for orthogroups with significant

Please cite this article in press as: Bastide et al., The genome of the blind bee louse fly reveals deep convergences with its social host and illuminates *Drosophila* origins, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.01.034

deviation only, we estimated and visualized ancestral counts for each orthogroups of these two families using FastAnc command in the R package Phytools.

QUANTIFICATION AND STATISTICAL ANALYSIS

Genome size and protein-coding gene content analyses

We compared the estimates of genome size and protein-coding gene content of *B. coeca* to the distributions of these values for the 10 drosophilid genomes (data given in Table S1) using one-sample Student's *t* test and after testing for normality using Shapiro-Wilk test as implemented in R.

Likelihood Ratio Test (LTR) comparison of CAFE5 models

We compared the likelihood of the two simplest CAFE5 models, *i.e.* one- and two-lambda models which were the only ones to converge across the four iterations, using the lr.test function of the extRemes R package.[93] For each model, the likelihood estimates were averaged across the four iterations.

False Discovery Rate (FDR) estimation of the p-values of the best CAFE5 model

The two-lambda model had the best likelihood and was therefore subsequently run for four iterations. *p-value*s inferred for each gene group in the branch leading to *B. coeca* in the iteration with the best likelihood were extracted, and multiple testing corrections were conducted using the False Discovery Rate (FDR) analysis of the FDRestimation[95] package implemented in R.