The genome of the bee louse fly reveals deep convergences in the evolution of social inquilinism

Hélène Legout¹, David Ogereau¹, Julie Carcaud¹, Jonathan Filée¹, Lionel Garnery¹, Clément Gilbert¹, Fabrice Requier¹, Jean-Christophe Sandoz¹, Amir Yassin¹, Héloïse Bastide¹

¹Laboratoire Évolution, Génomes, Comportement et Écologie, CNRS, IRD, Université Paris-Saclay – Institut Diversité, Écologie et Évolution (IDEEV), 12 route 128, 91190 Gif-sur-Yvette, France.

Corresponding author: Héloïse Bastide (heloise.bastide@universite-paris-saclay.fr)
Abstract

The nests of social insects often harbor a rich fauna of intruders, known as inquilines. Close relatedness between the host and the inquiline prevails due to potential genetic predispositions but how phylogenetically distant inquilines adapt to their hosts remains unclear. Here, we analyzed the genome of the wingless and blind bee louse fly *Braula coeca*, an inquiline of the Western honey bee *Apis mellifera*. We found that unlike many parasites, such as the human louse, the bee louse genome did not show significant erosion or strict reliance on an endosymbiont, likely due to a relatively recent age of inquilinism. However, a strikingly parallel evolution in a set of gene families was observed between the honey bee and the bee louse. 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 hive and a major diet of honey, pollens, and beeswax.

Vision-related and odorant receptor genes also exhibited rapid losses. Only genes whose orthologs in the closely related *Drosophila melanogaster* respond to components of the honey bee alarm pheromones or floral aroma were retained, whereas the losses included orthologous receptors responsive to the anti-ovarian honey bee queen pheromone. These results establish a new model for the study of major morphological and neuroethological transitions and indicate that deep genetic convergences between phylogenetically distant organisms can underlie the evolution of social inquilinism.

Keywords: parasitism; morphological evolution; behavioral adaptations; gene family evolution; Drosophila.
Introduction

Inquilinism is a form of interspecific interactions wherein an organism, the inquiline, inhabits the "home" of another organism (Hegner 1926). Examples abound in nature (Moser 1964; Luczkovich et al. 1991; Sanver and Hawkins 2000; Kneitel and Miller 2002), and ancient cases are even present in the fossil record (Landman et al. 2014). The relationship between the inquiline and its host can range from commensalism to parasitism depending on the overlap in and availability of resources and space. Factors favoring the evolution of inquilinism greatly depend on the inquiline's capacity to hide its presence from the host. This is particularly true in the case of social inquilines, which constitute the most frequent case of inquilinism, wherein the efficient nest cleaning and care for the offspring performed by social organization are often fatal for the intruders. Therefore, phylogenetic relatedness with the host was suggested to play a major role in facilitating social inquilinism, since the inquiline would share some common genetic factors promoting camouflage, known as Emery's rule (Cini et al. 2019). Although recent phylogenetic studies have revised and sometimes rejected Emery's rule at the specific level (Huang and Dornhaus 2008; Lopez-Osorio et al. 2015; Romiguier et al. 2018; but see Savolainen and Vepsäläinen 2003; Degueldre et al. 2021), most cases of social inquilines remain related to the same genus (Jansen et al. 2010; Cardinal et al. 2010), tribe (Schrader et al. 2021), family (Ronquist 1994; Cardinal et al. 2010), or even order (Gilbert et al. 2012). The genetic basis underlying inquilines belonging to distinct orders or across wide phylogenetic distances is still less understood.

The hive of the Western honey bee Apis mellifera comprises several parasites and inquilines that are attracted by the hive's rich resources (honey, pollen, beeswax, and the bees themselves) and its clean and protective shelter (Winston 1987). These include several insects belonging to distinct orders such as wax moths, hive beetles, and endoparasitoid flies. None of these has endured as profound morphological changes that even its affiliation at the order-level was confounded, as the apterous and quasi-blind bee louse fly, Braula coeca (Figure 1A-C). The female lays eggs in honey (not brood) cells, and the hatched larvae eat pollen and wax, where they borrow tunnels in which they pupate without forming true puparia (Skaife 1922; Imms 1942). Following emergence, the adults attach to the body of worker bees, migrating from one individual to another until reaching the queen. There, they move to the head, stimulate regurgitation and imbibe from her mouth the honey and nectar (Skaife 1922; Imms 1942). The bee louse is considered an
inquiline kleptoparasite with potential negative effects on bee colony health due to galleries in bee combs and the facilitation of transmitting serious pathogenic viruses to the bees (Avalos et al. 2019).

The phylogenetic positioning within the Diptera of the family Braulidae, which contains the bee louse, has long been puzzling due to its modified morphology (Grimaldi and Underwood 1986). Interestingly, recent phylotranscriptomic and phylogenomic analyses show the bee louse to constitute a basal lineage within the Drosophilidae (Bayless et al. 2021; Winkler et al. 2022). This proximity to Drosophila melanogaster, the most investigated insect at the genetic, developmental, and neurobiological levels, makes the bee louse a unique model for the study of the genomic changes underlying major morphological and ecological shifts. We present here an annotated assembly of the bee louse B. coeca genome and compare the evolution of its genomic architecture and gene content with those of D. melanogaster (Adams et al. 2000), the honey bee A. mellifera (Weinstock et al. 2006), as well as to its homonym and morphologically-similar the ectoparasitic human louse Pediculus humanus (Kirkness et al. 2010). The comparisons revealed striking evidence of cross-order genomic parallelism and shed new light on the evolution of social inquilinism between phylogenetically distant organisms.

Results
Genome architecture
We sequenced whole genome from a pooled sample of 15 unsexed Braula coeca individuals, all collected from the same geographical location, the Island of Ouessant in France. We used a hybrid approach to assemble a draft genome using both long-read Oxford Nanopore Technology (ONT) and short-read Illumina sequencing (see Methods).

Size. The final assembly of the bee louse showed a size of 309 Mb, an N50 of 347227 bp and a GC content of 34.95%. BUSCO gave a score of 95.8% of the Dipteran conserved single-copy orthologs with 1.3% of duplicated genes. Analysis of heterozygous k-mers pairs distribution indicated that the genome is diploid (Supplementary Figure 1A) and genome size prediction using k-mers distribution spectra predict a genome size of 308 Mb, concordant with the assembly size (Supplementary Figure 1B). The bee louse genome is therefore larger than the genomes of the honey bee (227 Mb), Drosophila melanogaster (143 Mb), and the ectoparasitic human louse (108 Mb).
Endosymbionts. The taxonomic assignment of each contigs show that most of them (96%) match with arthropods indicating no (or very few) DNA contamination (Supplementary Figure 1C,D). Unlike in the human louse, no evidence for an obligate endosymbiont was detected in the bee louse. Imms (1942) discussed the possibility that the bee louse larvae have gut microbes that facilitate the digestion of the beeswax. We cannot rule out this hypothesis, but our results show that such microbiota, if present, likely do not persist in the adult stage.

Gene content. The annotation of the bee louse genome yielded 11,221 protein-coding genes. This number is higher than in A. mellifera and P. humanus where 9,935 and 10,773 protein-coding genes are found, respectively, but lower than that of D. melanogaster (13,968 protein-coding genes). The Annotation Edit Distance (AED), which measures the congruence between gene annotation and its supporting evidence was ≤ 0.5 for 96.7% of our gene models, indicating the near completeness of our annotation. Similarly, 79.79% of the corresponding proteins had a Pfam domain, which is another indication of annotation completeness since it varies between 57% and 75% in most eukaryotes.

Synteny. Orthologous genes were used to map B. coeca scaffolds to their corresponding Muller’s elements in D. melanogaster. The alignment showed strong consistency indicating persistent synteny between the two lineages (Figure 1D). However, given the short length of the scaffolds it was difficult to assess how much collinearity and rearrangement events took place since their divergence. There is no karyotypic map of B. coeca so it remains difficult at this stage to infer the chromosomal number from the assembly alone.

Mitochondrial genome. The mitochondrial genome consisted of a single scaffold of 19 kb (Figure 1E) unlike in the human louse, which has multiple fragments inside the mitochondria. Compared to conserved gene content and collinearity of the D. melanogaster and A. mellifera mitogenomes (Crozier and Crozier 1993), the mitochondrial gene content of the bee louse was incomplete, lacking the ND1 gene, included several duplications of the ND2 and ND4L genes that are truncated and/or display frameshifts and had a rearrangement between the COX1 and COX2 genes. The nuclear genomes contained 165 mitochondrial DNA insertions (NUMTs) distributed on 88 contigs and totaling 278 kb (~0.09% of the genome). The proportion of NUMTs in the bee louse approaches that of the honey bee (0.08%, (Behura 2007)) but exceeds that of the genus Drosophila (0.03%, (Rogers and Griffiths-Jones 2012).
Transposable elements. The large genome size of the bee louse despite its low gene content compared to *D. melanogaster* suggest an increase in repetitive sequences. RepeatModeler and RepeatMasker analyses indicated that nearly 41.41% of the *B. coeca* genome consist of such sequences, compared to 22.15% and 11.14% in *D. melanogaster* and *A. mellifera*, respectively (Supplementary Table 1). There is a far larger proportion of long interspersed nuclear elements (LINEs) retrotransposons in *B. coeca* (15.18%) compared to only 2.38% and none in *D. melanogaster* and *A. mellifera*, respectively. Long terminal repeat (LTR) elements on the other hand were fewer in *B. coeca* (0.57%) and *A. mellifera* (0.17%) than in *D. melanogaster* (6.99%).

Because host-parasite relationships have repeatedly been invoked as a factor that may favor horizontal transfer of transposable elements (TE) (Gilbert *et al.* 2010; Ortiz *et al.* 2015; Venner *et al.* 2017), we searched for evidence of such transfers between *B. coeca* and *A. mellifera* (Supplementary Text 1). We found one TE, *Famar1-like* element, previously described in the earwig *Forficula auricularia* (Barry *et al.* 2004a) that shows high similarity between *B. coeca* and *A. mellifera* but was absent in *D. melanogaster*, highly suggestive of acquisition through horizontal transfer (Supplementary Figure 2). However, phylogenetic analysis of multiple copies of this TE extracted from 37 widely divergent animal species (Supplementary Figure 3) shows that it was most likely acquired through independent transfer events in *B. coeca* and *A. mellifera*, and that the tight ecological connection between the two species are unlikely to have played a direct role in these transfers (Supplementary Text 1).

Gene family evolution

Families with excess losses. We identified gene families that underwent reduction or extension as compared to *D. melanogaster* using DAVID (Sherman *et al.* 2022). Underrepresented families (27 families with False Discovery Rate (FDR) < 0.05) showed a striking parallelism with the honey bee, with three of the most underrepresented families being similar (given in Honey bee Genome Sequencing Consortium 2006; Supplementary Table 2). These included InterPro-defined families such as Peptidase S1 (IPR001254), Ecdysteroid kinase-like (IPR004119) and Zinc finger AD-type (IPR012934). Peptidases play a major role in detoxification and digestion of protein-rich diets. Their reduction in both the honey bee and the bee louse could be due to the overprotectiveness of their mutual habitat, the hive, and/or the low protein content of
their food, nectar and honey. The Ecdysteroid kinase-like (EcKL) family is also suspected to include proteins involved in detoxification (Scanlan et al. 2020). We also noted a functional affinity between two underrepresented families involved in the formation of the exoskeleton: the Insect cuticle protein (IPR000618) in the honey bee and Chitin-binding domain (IPR002557) in the bee louse. Cuticles could act as barriers against environmental toxins, which may not be highly encountered in the hive. Remarkably, *B. coeca* is unique among Cyclorrhaphan Diptera as its pupa, similarly to the honey bee’s (Winston 1987), is contained in the unmodified cuticle of the third instar larva, and no sclerotized puparium is formed (Skaife 1922; Imms 1942).

Significant biological processes terms (FDR < 0.05) were mostly related to proteolysis and lipid metabolism (e.g., fatty-acyl-co-A biosynthesis), whereas cellular components terms were associated with the membranes and extracellular space in agreement with a biased loss of genes potentially involved with metabolism, detoxification and/or immunity (Supplementary Table 2).

**Families with excess gains.** We did not find a similar important overlap in overrepresented gene families between the bee louse and the honey bee (5 families with FDR < 0.05; Supplementary Table 3). Expanding gene families unique to the honey bee included 7TM odorant receptor (IPR004117), Ankyrin (IPR002110), Yellow/royal jelly protein (IPR003534) and LysR substrate-binding (IPR005119) (Weinstock et al. 2006). We found one family, Pleckstrin homology domain (IPR001849), that was overrepresented in both species. This domain is a part of several lipases that are known to be involved in wax ingestion. Indeed, larvae of the bee louse mine the beeswax (Imms 1942), whereas honey bee workers use their salivary lipases while chewing the wax to form the hive combs (Kurstjens et al. 1985).

Gene Ontology (GO) biological and cellular terms with significant enrichment at FDR < 0.05 associated with morphological (e.g., imaginal disc-derived leg morphogenesis, autophagy, and dorsal closure) and/or neurological (e.g., dendrite morphogenesis, axon guidance, sensory perception of sound, and neuromuscular junction) developments (Supplementary Table S3). The bee louse duplicated genes hence may play a role in the evolution of the particular morphologies that helped adaptation to the phoretic lifestyle.

**Vision.** The species Latin name refers to the early assumption that the bee louse was blind due to the reduction of the eye size and the loss of the ocelli. However, thin optic nerves connect the brain to the rudimentary eyes, which lack ommatidia and pigments.
(Müggenburg 1892), and the bee louse demonstrates negative phototaxis, indicating a certain degree of light perception (Kaschef 1959). In agreement with reduced vision in the bee louse, we found three out of the seven rhodopsin genes, which are responsible for colored vision and the positive phototaxis of *D. melanogaster*. Two of the three genes, *Rh1* and *Rh6*, are expressed in the ommatidia and are sensitive to light with long wavelengths (Senthilan and Helfrich-Förster 2016), whereas the third one, *Rh7*, is expressed in the brain and regulates light-dependent circadian entrainment (Ni et al. 2017). The role of these opsins in light detection despite the absence of ommatidia is unclear. Remarkably, *Rh1*, *Rh6* and *Rh7* are structurally required in mechanosensory bristles to control larval locomotion (Zanini et al. 2018), and *Rh1* and *Rh6* also detect temperature (Leung and Montell 2017), whereas *Rh1* and *Rh7* detect low concentrations of a bitter plant component (Leung et al. 2020). Therefore, the retention of these rhodopsins in the bee louse could mainly be due to their unconventional functions. On the other hand, the red-sensitive rhodopsin *Rh2*, which is exclusively expressed in the ocelli and used for horizon detection in *D. melanogaster* (Mishra et al. 2021), is among those which were lost in the bee louse, in agreement with the loss of the ocelli in the bee louse.

**Olfaction.** Odorant receptors (ORs) are essential to detect volatile chemical signals from the environment. In most *Drosophila* species, there are nearly 61 ORs, whereas in the honey bee this family has expanded to reach 160 (Robertson and Wanner 2006). We found 29 ORs in the bee louse in addition to the Orco ortholog. Of these four had no direct orthologs in *D. melanogaster*, but the remaining 25 genes were orthologous to 21 genes in *D. melanogaster* and 12 were duplicates specific to the bee louse lineage (Figure 2A; Supplementary Figure 3). Judging from the response of those ORs to different volatiles in *D. melanogaster* as curated in the DOOR database (Münch and Galizia 2016) and assuming potential conservation of function, we can divide the bee louse ORs into three categories. First, 11 ORs respond to different components of the honey bee workers alarm and mandibular gland pheromones, e.g., 2-heptanol, propyl acetate, 2-heptanone, 1-hexanol, butyl acetate, isopentyl acetate, etc. These included *Or13a* (x2 paralogues), *Or42a*, *Or42b/Or59b*, *Or43a*, *Or47b*, *Or49a/Or85f*, *Or67b* (x4), *Or74a*, *Or85b/Or85c*, and *Or85e* (x2). Second, 12 ORs respond to different floral, pollen and nectar aromas, such as acetophenones, ethylguaiacol, geranyl acetate or fenchone, but most importantly to benzaldehyde, a major volatile of honey (Machado et al. 2020; Starowicz et al. 2021). These included *Or30a*, *Or45b* (x2), *Or59a*, *Or67b* (x4), *Or82a*, *Or85e* (x2) and *Or94a*. Note
that the multiple *B. coeca* paralogs orthologous to *D. melanogaster* Or67b and Or85e genes belong to both categories. Third, one OR, Or67d, which is involved in sexual pheromone 11-cis-vaccenyl acetate perception (Ha and Smith 2006), is present in the bee louse. Three among the ORs that were lost in the bee louse compared to other drosophilids, Or49b, Or56a and Or98a, are responsible to *D. melanogaster* response to the anti-ovarian honey bee queen mandibular pheromones 9-oxo-2-decenoic acid (9ODA) and 10-hydroxy-2-decenoic acid (10HDA) (Galang et al. 2019). Their loss most likely protected the capacity of the bee louse to reproduce in the hive.

Taste. Gustatory receptors (GRs) allow detecting soluble chemical signals. There are nearly 68 GRs in *D. melanogaster*, that respond mostly to sweet, bitter and carbon dioxide (CO\(_2\)) tastes (Weiss et al. 2011). Unlike their expanded OR family, the honey bee has only 10 GRs, of which 7 are orthologous to sweet *Drosophila* GRs (Robertson and Wanner 2006). This is likely due to the bees' food reliance on sweet floral nectars and honey. We found 13 GRs in the bee louse, with no duplications (Figure 2B; Supplementary Figure 4). These GRs could be classified according to their *D. melanogaster* orthology into three categories. First, 6 GRs belong to the sweet class, namely Gr43a, Gr61a and Gr64a,b,e,f which usually respond to sucrose, maltose and fructose and other major honey sugars. Second, 5 GRs belong to the bitter class, namely Gr33a, Gr57a, Gr66a, Gr93a and Gr94a, suggesting that nearly 50 mostly bitter tasting drosophilid GRs were lost in the bee louse. Note that in *D. melanogaster*, Gr33a and Gr66a, are potentially involved in sexual pheromones detection (Lacaille et al. 2007; Moon et al. 2009). Third, 2 GRs, Gr21a and Gr63a, which detects CO\(_2\) odor in *D. melanogaster* (Jones et al. 2007) are present in the bee louse. Those GRs are absent in the honey bee despite its ability to perceive CO\(_2\) (Stange and Diesendorf 1973). High CO\(_2\) concentrations are probably characteristic of largely populated hives and induce fanning response in bees (Seeley 1974). The quasi-blind bee louse may therefore use CO\(_2\) concentrations to detect the bees in the dark hive environment.

Discussion

That the enigmatic bee louse is indeed a drosophilid (Winkler et al. 2022), a lineage within the most investigated insect family with more than 100 fully sequenced genomes (Kim et al. 2021), is undoubtedly one of the most exciting discoveries in Dipteran phylogeny. How could a fly with a typical drosophilid genome become partly ecologically
like a bee and partly morphologically like a louse? Our cross-order comparisons of the bee
louse to its homonyms shed significant light on the genomic basis of these spectacular
convergences.

The genome of the human louse is among the smallest sequenced insect genomes
(Kelley et al. 2014). Loss of significant portions of genomic and gene contents is a
characteristic of obligate parasites specializing on specific hosts or inhabiting extreme
environments. Indeed, the gene content of the bee louse genome approached that of the
human louse, but the ~309 Mb-long genome of the bee louse is longer than that of most
drosophilid species, even being slightly longer from the largest genome in this family
(~304 Mb-long; Kim et al. 2021). Besides, the human louse has an obligatory bacterial
endosymbiont that may compensate the loss of its genes (Kirkness et al. 2010) but no
evidence for such associations is present in the bee louse. These differences may mainly
indicate the relative recency of the shift to inquilinism in the bee louse compared to the
230 million years (myr) of specialization in true lice (Misof et al. 2014). Megabraula, the
closest relative to the genus Braula, is an inquiline of the giant honey bee Apis laboriosa,
whose divergence from the Western honey bee A. mellifera is around 23 myr ago
(Grimaldi and Underwood 1986; Cardinal et al. 2010). The crown age of the Drosophilidae
and the divergence time between Apis and its closest pollen-basket (corbiculate) bees are
estimated at 66-70 myr ago (Cardinal et al. 2010; Suvorov et al. 2021). This indicates that
association between braulids and Apis has arisen between 70 to 20 myr ago.

The origin of social inquiline requires the evolution of multiple convergences
that can locate the host, deceive it to enter the colony, adapt to its social organization, and
remain undetected or not easily removed from the colony. These phenotypes, which are
often host-specific, evolve more easily when the inquiline and the host are
phylogenetically close (Cini et al. 2019). Despite the distance of the bee louse from honey
bees, gene family analysis provided strong evidence for convergent evolution mostly for
genes potentially involved in immunity, detoxification, metabolism and chemical
perception. Although Braula has lost nearly half of the typical drosophilids odorant
receptor repertoire, in contrast to the major expansion of this family in the honey bee, it
predominantly retained genes whose orthologs in D. melanogaster detect compounds of
the honey bee pheromones and/or floral aroma and honey odors. Low concentrations of
isopentyl acetate (IPA), the major component of the alarm pheromone, released by
unstressed workers at hive entries attract the parasitic hive beetle Aethina tumida (Torto
et al. 2007), suggesting that the detection of the host odors could be a common strategy among phylogenetically distant social inquilines and parasites. We also noted that the bee louse has multiple copies related to Or67b which responds to several components of bee alarm pheromones, such as 1-hexanol, 2-heptanone, 1-butanol and 3-methyl-1-butanol, as well as to benzaldehyde, the major honey volatile, in D. melanogaster (Münch and Galizia 2016), suggesting a possible dual ancestral function that might have facilitated association with the honey bees. The multiple copies related to Or67b are remarkable since the copy number expansion of this gene was associated with the evolution of herbivory and strong plant-association in the drosophilid genus Scaptomyza (Goldman-Huertas et al. 2015; Matsunaga et al. 2022). Therefore, the bee louse might have evolved from a flower-breeding or plant-associated lineage within the Drosophilidae, a lifestyle that has recurrently evolved in this family (Yassin 2013).

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. The loss of the three odorant receptors, Or56a, Or49b and Or98a, that respond in D. melanogaster to the honey bee queen's pheromones which “sterilize” the bee workers, mainly 9-oxo-2-decenoic acid (9ODA) and possibly 9-hydroxy-2-decenoic acid (9HDA) (Galang et al. 2019), is a notable example. The queen’s pheromones elicit anti-ovarian response in other insects including D. melanogaster mostly through the activation of the neurons bearing these receptors. Therefore, a sine qua non condition for reproducing in a beehive is to protect against the effects of those pheromones, with the loss of the responding receptors being a preliminary and effective strategy. However, this raises the question of how the bee lice recognize the queen, which they preferentially infest (Imms 1942). It is therefore possible that higher sensitivity to worker pheromones, as suggested by the repertoire of retained odorant receptors (see above), may help to mostly discriminate the workers hence facilitating the recognition of the queen, who has a rudimentary sting gland. Another possibility is that strong CO₂ emission by the court surrounding the queen may be an indicator of her location, as could be suggested from the retention of the two CO₂-smelling gustatory receptors, Gr21a and Gr63a, whose orthologs are absent in the honey bee. It also remains unclear how the bee lice sexually communicate in the hive given the predominance of chemical camouflage; the hydrocarbon profile of bee lice from different colonies mimicked that of the colony of origin (Martin and Bayfield 2014). The retention of some chemosensory genes whose
orthologs are involved in sexual pheromone detection in *D. melanogaster*, such as *Or67d*, *Gr33a* and *Gr66a* (Kurtovic *et al.* 2007; Lacaille *et al.* 2007; Moon *et al.* 2009) suggest that pheromonal communications may still be present in the bee louse but perhaps at low undetectable levels.

Small size, loss of wings and the evolution of strongly clinging legs are all morphological adaptations that could prevent the honey bees getting rid of the bee lice. All these 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 (Roff 1990). 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. The regression of visual systems and their underlying genes in the bee louse spending most of its life cycle in the bee hives is a common phenomenon in animals inhabiting dark environments, such as fossorial mammals (Partha *et al.* 2017) and cavefishes (Policarpo *et al.* 2021). With its genetic relatedness to *Drosophila* and ecological association to *Apis*, two major laboratory models, and with new genomic tools presented here, the bee louse *Braula coeca* is a promising model to address questions related to deep convergences that are still difficult to approach in multiple highly specializing animals.

**Materials and Methods**

**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](https://www.mn-net.com), Düren, Germany).

**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, as in Miller *et al.* (2019). Before nanopore sequencing, a size selection was conducted on the DNA
using the SRE XS from Circulomics (https://www.circulomics.com/, Baltimore, Maryland, USA). The SQK-LSK110 kit from Oxford Nanopore Technology (Lu et al. 2016; https://nanoporetech.com/) was then used to prepare the samples for nanopore sequencing following manufacturer’s protocol. The library was loaded and sequenced on a R9.4.1 flow cell (ref FLO-Min106) for sequencing. Raw data were basecalled using Guppy v5.0.11 and the “sup” algorithm. Illumina paired-end sequencing was performed by Novogene Company Limited (https://en.novogene.com, Cambridge, UK) on the same DNA library.

We used MaSuRCA v4.0.3 (Zimin et al. 2017) 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 347211 pb. The completeness of the assembly was estimated to 95,8% with Busco v5.0 on the diptera_odb10 dataset (C:95.8%[S:94.6%,D:1.2%],F:0.7%,M:3.5%,n:3285 ).

Estimation of genome size, endosymbionts detection and mitogenome assembly

K-mers frequencies within short-read data were obtained with KMC 3 (Kokot et al. 2017). Genome size and ploidy were inferred using GenomeScope v2.0 with k-mer size = 21 and Smudgeplot (Ranallo-Benavidez et al. 2020). Contig taxonomy was performed using Blobtools (Laetsch and Blaxter 2017) with Diamond as search engine (Buchfink et al. 2015) against the UniProt database using a local copy of the NCBI TaxID file for the taxonomic assignation of the best hit. Minimap2 (Li 2018) was used for read mapping.

Genome annotation

The B. coeca genome was annotated using Maker v2.31.10 (Cantarel et al. 2008), following the protocol given in Muller et al. (2021), wherein multiple rounds of Maker supported by the training of the SNAP v.2006-07-28 (Korf 2004) and Augustus v.3.3.3 (König et al. 2016) gene finding and prediction tools, were conducted. Transcriptome of B. coeca (NCBI accession no. SRR2046564; 1KITE Consortium; (Bayless et al. 2021)) and proteomes of five Drosophila species, namely D. innubila (Hill et al. 2019), D. albomicans (Mai et al. 2020), D. bipectinata (Kim et al. 2021), D. melanogaster (Adams et al. 2000) and D. virilis (Clark et al. 2007) were used to guide the annotation. Protein-Protein BLAST 2.9.0+ (Altschul et al. 1997) (-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 colocalization of *B. coeca* and their *D. melanogaster* orthologs on their respective contigs and chromosomal arms was analyzed to test for synteny.

**Transposable elements annotation and transfer**

Transposable elements were identified in each species following a two-step protocol. First, we used RepeatModeler v 2.0.1 (Flynn *et al.* 2020) with default parameters to generate a *de novo* library of repetitive regions. RepeatMasker v 4.0.9 (Flynn *et al.* 2020) 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 reference genome along with the percentage of the genome they represent. Horizontal transfer analyses protocols of the *Famar1*-like element are given in Supplementary Text 1.

**Gene family evolution**

The Database for Annotation, Visualization and Integrated Discovery DAVID 2021 (Sherman *et al.* 2022) was used to test for gene ontology (GO) terms enrichments among lists of *D. melanogaster* orthologs annotated on the *B. coeca* genome (see above). The program was either fed with the list of *D. melanogaster* genes that were absent or duplicated in the *B. coeca* genome to test for under- and overrepresented gene families, respectively, compared to *D. melanogaster* genome, as denoted by the False Discovery Rate (FDR) correction for multiple tests. Each analysis was conducted for biological processes, molecular function, and cytological components.

**Chemosensory superfamilies evolution**

Protein sequences of the odorant (ORs) and gustatory (GRs) receptors of *A. mellifera* and *D. melanogaster* were obtained from Robertson and Wanner (2006) to which we added the annotated ORs and GRs protein sequences of *B. coeca*. Sequences were aligned using Molecular Evolutionary Genetic Analysis (MEGA X) software package (Kumar *et al.* 2018), which was also used to infer a maximum-likelihood phylogenetic tree for each family. We used iTOL v4 (Letunic and Bork 2019) to visualize the trees. *B. coeca* protein sequences are given in Supplementary Datasets 2 and 3 for ORs and GRs, respectively.
Acknowledgments

We are very grateful to Dr. Frédéric Marion-Poll for insightful comments on an early draft of the manuscript and to the Association Conservatoire de l'Abeille Noire Bretonne (A.C.A.N.B.) for help collecting *B. coeca* flies. *Braula* genome sequencing was funded by a grant from Université Paris Saclay (ADAPAR) to HB.

Conflicts of interest

The authors declare no conflicts of interest.

References


assemblies of 101 drosophilid genomes, (G. Coop, P. J. Wittkopp, and T. B. Sackton, Eds.).


Müggenburg F. H., 1892 Der Rüssel der Diptera pupipara. Nicolai.


Stange G., and M. Diesendorf, 1973 The response of the honeybee antennal CO2-receptors to
N2O and Xe. J. Comp. Physiol. 86: 139–158. https://doi.org/10.1007/BF00702534

Starowicz M., P. Hanus, G. Lamparski, and T. Sawicki, 2021 Characterizing the Volatile and Sensory Profiles, and Sugar Content of Beeswax, Beebread, Bee Pollen, and Honey. Molecules 26: 3410. https://doi.org/10.3390/molecules26113410


and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. Genome Res. 27: 787–792. https://doi.org/10.1101/gr.213405.116
Legends of figures

**Figure 1** – The bee louse fly (*Braula coeca*) morphology and genome. A) Two adults attached to a honey bee worker preserved in alcohol. B) Dorsal view of an adult showing the loss of the wings, halteres and scutum, the reduction of the mesonotum and the robustness of the legs. C) Frontal view of an adult showing the reduction of the eyes and the loss of the ocelli. Scale bars = 1 mm in A and 0.5 mm in B and C. D) Orthologous genes colocalization in *D. melanogaster* Muller’s elements and *B. coeca* contigs demonstrating the predominant conservation of synteny. E) The mitochondrial genome of *B. coeca*. Blue and violet ribbons indicate the protein-encoding genes in sens and anti-sens respectively, red and yellow ribbons indicate the rRNA and tRNA genes in sens and anti-sens respectively. Names with asterix indicate the presence of inactivated paralogous genes.

**Figure 2** – Maximum-likelihood phylogenetic tree for A) odorant receptors (ORs) and B) gustatory receptors (GRs) of *A. mellifera* (green), *B. coeca* (red) and *D. melanogaster* (blue). Genes commented in the text are labelled. For full labels see Supplementary Figures 4 and 5. Sequences for *B. coeca* are given in Supplementary Datasets 2 and 3 for ORs and GRs, respectively. Sequences for *A. mellifera* and *D. melanogaster* are from Robertson and Wanner (2006). a.a. = amino acids.
List of supplementary materials

Supplementary Dataset 1 – Sequence of all *Famar* and *Famar1*-like copies used to reconstruct the phylogeny shown in Supplementary Figure 3. The sequence names contain the name of the species, the Genbank accession number of the contig from which the copy was extracted, the start and end position of the copy in the contig, as well as a final number that allows making a correspondence with leaves in the tree shown in Supplementary Figure 3.

Supplementary Dataset 2 – Sequence of odorant receptors (ORs) of *Braula coeca* used to reconstruct the phylogeny shown in Supplementary Figure 4.

Supplementary Dataset 3 – Sequence of gustatory receptors (GRs) of *Braula coeca* used to reconstruct the phylogeny shown in Supplementary Figure 5.

Supplementary Figure 1 – Genomic features of *Braula coeca*. A) Ploidy estimation using the coverage and the distribution of heterozygous k-mers pairs. B) Genome size estimation using the k-mer profile spectrum. C) Taxonomic assignation of the contigs according to their GC% and their coverages. D) Proportion of short-reads that mapped onto the genome assembly (left) and onto the different contigs according to their taxonomic assignments (right).

Supplementary Figure 2 – Comparison of Famar1-like synonymous distance and orthologous gene synonymous distances between *Braula coeca* and *Apis mellifera*. To calculate the distribution of gene or gene fragment synonymous distances (dS) between *B. coeca* and *A. mellifera* we selected best reciprocal blastp hits between single copy BUSCO genes retrieved from the two species and calculated dS for each of them using the approach described in Zhang et al. (2020). The red line indicates the 0.5% quantile of this distribution (=1.76). The distribution is bimodal, with genes having highly saturated dS values showing a peak centered on 9.99 and genes showing less saturated dS values showing another peak around 2.5. We verified that genes showing less saturated values correspond to highly genes that evolve under strong purifying selection and are thus highly conserved between Hymenoptera and Diptera. The *Famar1*-like dS (green line, =
0.12) was calculated over the transposase open reading from of one copy of the element extracted from the *A. mellifera* genome and another copy extracted from the *B. coeca* genome.

**Supplementary Figure 3** – Phylogeny of Famar1-like copies from 38 animal species. The ten Famar1-like copies showing the highest nucleotide identity to the *Famar* element initially described in the earwig (Barry *et al.* 2004b) were retrieved using online blastn (see Supplementary Text 1) and extracted from the genome of 37 animal species. Filled circles indicate bootstrap value higher than 70%, with the diameter of the circle proportional to each individual value. Name of the copies are composed of the name of the species from which they were extracted and a unique number that allows making a correspondence with sequences provided in Supplementary Dataset 1.

**Supplementary Figure 4** – Labelled maximum-likelihood phylogeny of odorant receptors (ORS) of *A. mellifera* (green), *B. coeca* (blue), and *D. melanogaster* (red) show in Figure 2A.

**Supplementary Figure 5** – Labelled maximum-likelihood phylogeny of gustatory receptors (GRs) of *A. mellifera* (green), *B. coeca* (blue), and *D. melanogaster* (red) show in Figure 2B.

**Supplementary Table 1** – Proportions of transposable elements in the genomes of *B. coeca*, *D. melanogaster* and *A. mellifera*.

**Supplementary Table 2** – Gene Ontology (GO) enrichment for *D. melanogaster* genes that are absent from *B. coeca* genome as inferred using DAVID 2021.

**Supplementary Table 3** – Gene Ontology (GO) enrichment for *D. melanogaster* genes that have additional copies in *B. coeca* genome as inferred using DAVID 2021.

**Supplementary Text 1** – Analysis of horizontal transfer of transposable elements between *B. coeca* and *A. mellifera*.
Figure 1: Muller's element (D. melanogaster)