Title: The genomic basis of evolutionary differentiation among honey bees

Running title: Honey bee evolution

Authors: Bertrand Fouks[1,2], Philipp Brand[3,4], Hung N. Nguyen[5], Jacob Herman[1], Francisco Camara[6], Daniel Ence[7,8], Darren E. Hagen[9], Katharina J. Hoff[10,11], Stefanie Nachweide[10], Lars Romoth[10], Kimberly K. O. Walden[12], Roderic Guigo[6,13], Mario Stanke[10,11], Giuseppe Narzisi[14], Mark Yandell[8,15], Hugh M. Robertson[12], Nikolaus Koeniger[16], Panuwan Chantawannakul[17], Michael C. Schatz[18], Kim C. Worley[19], Gene E. Robinson[12,20,21], Christine G. Elsik[5,22,23], Olav Rueppell[1,24,*]

Affiliations:

[1] Department of Biology, University of North Carolina Greensboro, USA

[2] Institute for Evolution and Biodiversity, Molecular Evolution and Bioinformatics, Westfälische Wilhelms-Universität, Germany

[3] Department of Evolution and Ecology, Center for Population Biology, University of California Davis, USA

[4] Laboratory of Neurophysiology and Behavior, The Rockefeller University, USA

[5] MU Institute for Data Science & Informatics, University of Missouri, USA

[6] Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Spain

[7] School of Forest Resources and Conservation, University of Florida, USA

[8] Department of Human Genetics, University of Utah, USA

[9] Department of Animal and Food Sciences, Oklahoma State University, USA

[10] University of Greifswald, Institute for Mathematics and Computer Science, Bioinformatics Group, Germany

[11] University of Greifswald, Center for Functional Genomics of Microbes, Germany

[12] Department of Entomology, University of Illinois at Urbana-Champaign, USA

[13] Universitat Pompeu Fabra (UPF), Spain

[14] New York Genome Center, USA

[15] Utah Center for Genetic Discovery, University of Utah, USA

[16] Department of Behavioral Physiology and Sociobiology (Zoology II), University of Würzburg, Germany

[17] Environmental Science Research Center (ESRC) and Department of Biology, Faculty of Science, Chiang Mai University, Thailand

[18] Departments of Computer Science and Biology, Johns Hopkins University, USA

[19] Department of Molecular and Human Genetics, Human Genome Sequencing Center, Baylor College of Medicine, USA

[20] Carl R. Woese Institute for Genomic Biology, University of Illinois Urbana-Champaign, USA

[21] Neuroscience Program, University of Illinois at Urbana-Champaign, USA

[22] Division of Animal Sciences, University of Missouri, USA

[23] Division of Plant Sciences, University of Missouri, USA

[24] Department of Biological Sciences, University of Alberta, Canada

Keywords: *Apis*, piRNA pathway, evolution of dance language, adaptation to migration, eusociality, transposable elements, lineage-specific genes, chemosensory genes, positive selection

* Corresponding Author

Phone: (+1) 336-202-2349; FAX: (+1) 780-492-9234; Email: olav@ualberta.ca

CW 405 Biological Sciences Building, Edmonton, AB, T6G 2E9, Canada

Abstract

In contrast to the western honey bee, *Apis mellifera*, other honey bee species have been largely neglected despite their importance and diversity. The genetic basis of the evolutionary diversification of honey bees remains largely unknown. Here, we provide a genome-wide comparison of three honey bee species each representing one of the three subgenera of honey bees, namely the dwarf (*Apis florea*), giant (*A. dorsata*) and cavity-nesting (*A. mellifera*) honey bees with bumblebees as outgroup. Our analyses resolve the phylogeny of honey bees with the dwarf honey bees diverging first. We find that evolution of increased eusocial complexity in *Apis* proceeds via increases in the complexity of gene regulation, which is in agreement with previous studies. However, this process seems to be related to pathways other than transcriptional control. Positive selection patterns across *Apis* reveal a trade-off between maintaining genome stability and generating genetic diversity, with a rapidly evolving piRNA pathway leading to genomes depleted of transposable elements, and a rapidly evolving DNA repair pathway associated with high recombination rates in all *Apis* species. Diversification within *Apis* is accompanied by positive selection in several genes whose putative functions present candidate mechanisms for lineage-specific adaptations, such as migration, immunity, and nesting behavior.

Introduction

How genomes diverge to give rise to organismal diversity remains one of the most fundamental questions in biology. Comparative functional genomics has drastically expanded our knowledge on the relative contributions of genetic novelty and co-option (Jasper et al. 2015; Warner et al. 2019), structural and regulatory innovation (Deplancke et al. 2016), as well as cis- and trans-regulation of gene expression (Green et al. 2019) to phenotypic diversification. As a consequence, the genotype-phenotype map is being elucidated at ever-increasing detail (Zhou et al. 2020). In addition to broad-scale macroevolutionary studies, taxon-specific comparative genomics is generating novel insights, particularly with respect to structural genome evolution (Figueiró et al. 2017; Chavez et al. 2019; Sun et al. 2021).

The evolution of complex insect societies represents one of the major evolutionary transitions (Maynard Smith and Szathmáry 1995). Genomic signatures of this transition share few commonalities across taxa, except for an increase in gene regulatory capacity (Gadau et al. 2012; Simola et al. 2013; Terrapon et al. 2014; Kapheim et al. 2015; Harpur et al. 2017; Harrison et al. 2018). In contrast to the major focus on studying the genomic bases of the origin of sociality and associated traits, the maintenance and diversification of social traits has received limited attention (Simola et al. 2013; Jasper et al. 2015; Araujo and Arias 2021; Sun et al. 2021).

Here, we use a comparative, lineage-specific approach to identify genetic loci associated with evolutionary adaptations underlying the organization of complex insect societies in the eusocial honey bee genus *Apis*. Due to its scientific and practical importance, the Western honey bee *Apis mellifera* (L.) was among the first metazoans with a completed genome project (Weinstock et al. 2006). It has since served as a model for genomic studies of adaptation (Wallberg et al. 2014), invasion (Calfee et al. 2020), and social traits, such as caste differentiation (Chen et al. 2012), division of labor (Smith et al. 2008), and other social behaviors (Zayed and Robinson 2012).

In addition to the cavity-nesting *A. mellifera* and closely related species, the genus *Apis* contains two other lineages, the dwarf honey bees and giant honey bees (Raffiudin and Crozier 2007). Although their evolutionary origins are not clear (Kotthoff et al. 2013), all species share a social lifestyle in complex societies with thousands of workers and a single, polyandrous queen and nest in vertical wax comb to store food and raise brood (Oldroyd and Wongsiri 2006). However, the three subgenera exhibit pronounced differences in body size, colony size, mating behavior, caste divergence, nesting habits, thermoregulatory ability, recruitment dances, and defensive and migratory behaviors (Dyer and Seeley 1991; Oldroyd and Wongsiri 2006; Koeniger et al. 2010; Hepburn and Radloff 2011; Rueppell et al. 2011b).

The genetic architecture underlying the diversification of the *Apis* lineages remains largely unknown. Intra-specific studies have addressed the genetic basis of some key social traits, such as worker ovary size and caste differentiation (Cardoen et al. 2011; Graham et al. 2011; Chen et al. 2012), dance language (Johnson et al. 2002), and defensive behavior (Hunt et al. 2007; Alaux et al. 2009) in *A. mellifera*. However, it is unclear to what extent the identified genetic mechanisms involved in intra-specific variation can explain the inter-specific differentiation among *Apis* species (Dieckmann et al. 2004). Broad comparisons in *Apis* (Sarma et al. 2007, 2009) have been hampered by the lack of available genomic resources in species other than *A. mellifera* (Weinstock et al. 2006; Elsik et al. 2014) and the closely related *A. cerana* (Park et al. 2015), although the genome of *A. dorsata* has recently also been published (Oppenheim et al. 2020) and targeted analyses have helped to resolve particular gene families (Helbing et al. 2017).

Here, we present a comprehensive analysis of the molecular evolution of protein-coding genes across *Apis* based on homologous gene sets derived from genomes of all three major honey bee lineages. At the genome level, we reconstruct the phylogenetic relationships among the *Apis* lineages and identify key targets of positive selection associated with social complexity, ecological specialization, and chemosensation, elucidating the genomic basis of evolutionary diversification within honey bees.

Results

Honey bee genomes and phylogenetic inference

We identified all single-copy orthologs between the western honey bee *Apis mellifera*, the dwarf honey bee *A. florea*, and the giant honey bee *A. dorsata,* with bumblebees as outgroup. Our analysis included the published genomes of *A. mellifera* (Elsik et al. 2014) and *Bombus impatiens* and *B. terrestris* (Sadd et al. 2015). In addition, we sequenced, assembled, and annotated the genomes of *A. florea* and *A. dorsata*. This produced two high-quality genome assemblies of similar length and GC content (*A. dorsata*: 230 Mb, N50: 732kb, GC: 32.5%; *A. florea*: 229 Mb, N50: 2.86Mb, GC: 34.9%) but different contiguity (*A. dorsata*: size of scaffolds: 200 bp - 3.6 Mb, total count: 4040; *A. florea*: size of scaffolds: 500 bp - 9.6 Mb, total count: 6983), likely explained by differences in repetitive sequences (*A. dorsata*: 17.5%, 40.4 Mb; *A. florea*: 14.3%, 32.9 Mb). Even though a newer assembly for *A. mellifera* has been published since our analysis (Wallberg et al. 2019) and our sequencing and assembly strategies for *A. florea* and *A. dorsata* have been replaced by more modern approaches (Phillippy 2017), the generated datasets proved to be informative and appropriate for our subsequent analyses: A high level of gene completeness (*A. dorsata*: 93.7%, *A. florea*: 91.9%) was confirmed by a BUSCO analysis (Simão et al. 2015) with the hymenoptera lineage dataset.

The gene sets for comparison across species (see Methods) were of similar size among all bees (Figure 1). A total of 3,858 genes were present in only a single species (2,130 in *A. florea*, 584 in *A. dorsata*, and 1,144 in *A. mellifera*) and thus were categorized as lineage-specific. Among the 1,506 genes identified as homologs in only two species 570 were shared between *A. mellifera* and *A. dorsata* (570), more than either species with *A. florea* (386 and 550, respectively). 15,182 genes were shared among all species with 9,310 belonging to single-copy ortholog groups (Figure 1). The concatenated single-copy orthologs resulted in an alignment of 4,680,591 amino-acids, which we used to resolve the relationships among the three honey bee lineages. We recovered a highly supported phylogeny of *Apis* with the dwarf honey bees as outgroup to the other two lineages (Figure 1), agreeing with previous work (Raffiudin and Crozier 2007).

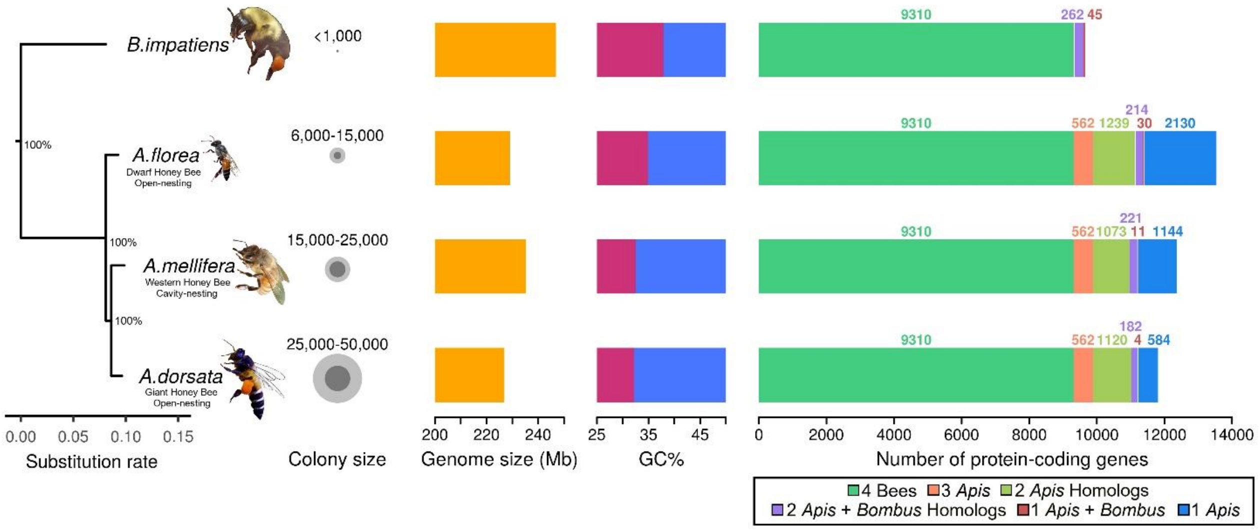

Figure 1 - Phylogenetic, genomic and gene content comparisons of 3 honey bee species. From left to right: Maximum likelihood phylogeny built from 9310 concatenated single-copy orthologous proteins from sequenced honeybees and bumblebee outgroup indicated that *A. florea* diverged first from the most recent common ancestor of honey bees (all nodes 100% bootstrap supported). *A. florea* represents the dwarf honey bees, while *A. mellifera* and *A. dorsata* represent the cavity nesters and the giant honey bees, respectively. Tree visualization was performed using ggtree (Yu 2020). Circles represent colony size ranges with dark grey indicating the lowest and light grey the highest colony size, the yellow bars depict the genome size of each species, and the red/blue bars correspond to the average GC content of the genome of each species. Average genome GC content decreases with increasing colony size. The rightmost horizontal bar plots show total gene counts for each species partitioned according to their orthology profiles. *A. florea* possessed the greatest number of lineage-specific genes followed by *A. mellifera*.

Genome-wide patterns of positive selection

To identify positive selection that acted on protein coding genes during the evolution of honey bees, we used the adaptive branch-site random effects likelihood (aBRSEL) method in Hyphy (Kosakovsky Pond et al. 2019; Smith et al. 2015) on 8,115 single-copy orthogroups (see Methods). We identified 149 single-copy orthogroups (1.85%) with signals of positive selection in at least one of the four branches at a 10% false discovery rate (FDR). Patterns of positive selection were equally distributed among the three honey bee species lineages with a proportion of 0.49-0.60% of all orthogroups tested (Supplemental Tables S1, S2). The basal *Apis* branch, however, was under positive selection in only 0.27% of orthogroups, representing a significantly lower proportion in comparison to the three species branches (Chi-squared test: $\chi^2 = 10.48$, df = 3, p = 0.0149). This result was not due to reduced power associated with short branches (Anisimova and Yang 2007) because the *Apis* branch had an overall increased branch length (Mean branch length (± standard error) of *Apis*: 0.37 ± 0.02, *A. mellifera*: 0.06 ± 0.0005, *A. florea*: 0.05 ± 0.0004, *A. dorsata*: 0.04 ± 0.0003; Kruskal-Wallis test: $\chi^2 = 3280$, df = 3, $p < 2.2 \times 10^{-16}$) and orthogroup test scores were positively correlated with the length of the tested branches (log-likelihood ratio; Spearman's correlation $\rho=0.20$, $p < 2.2 \times 10^{-16}$).

Next, we categorized each orthogroup by its homology to genes with known function in *A. mellifera,* in order to test whether the identified patterns of positive selection correlated with known functions. 6,719 of the 8,115 orthogroups (82.8%) included in the analysis could be categorized this way, while the function of 1,396 (17.2%) remained unknown. The proportion of genes with known (83.1%) and unknown (16.9%) function under positive selection did not differ from the overall distribution (Chi square test: $\chi^2 < 0.01$, df = 1, p = 1). However, genes with unknown function had a significantly higher median evolutionary rate ratio ($d_N/d_{S(\text{known function})} = 0.077$, $d_N/d_{S(\text{unknown function})} = 0.157$; Wilcoxon Rank Sum test: $W = 5.4 \times 10^7$, $p < 2.2 \times 10^{-16}$) compared to those with a known function. While this result is not surprising because genes with higher divergence rates are more difficult to annotate based on homology to genes of known function, it does emphasize the significance of studying genes of unknown function.

Most of the significant gene families were found to be positively selected in a single branch, although the following five were found to be positively selected in two branches: *muscle myosin heavy chain*, which is involved in muscle contraction (Holmes 2004; Odronitz and Kollmar 2008), was under positive selection in both *A. dorsata* and *A. florea*; *four and a half LIM domains protein 2*, involved in heart physiology and muscle formation (Johannessen et al. 2006), was under positive selection in both *A. dorsata* and *mellifera*; *serine-rich adhesin for platelets*, which plays a role in cell adhesion (Sanchez et al. 2010), was positively selected in the *Apis* branch and in *A. florea*; and *alpha-glucosidase 2* (*AmGCS2α*), which is involved in glucose metabolism, and one additional orthogroup of unknown function were positively selected in both the *Apis* branch and *A. mellifera*. In the three species branches, as well as the ancestral *Apis* branch, several positively selected genes were identified with a function in the regulation of gene expression, cell signaling, and neural processes, as well as with an association with resistance against pathogens and xenobiotics (Supplemental Tables S1, S2).

Tests of functional category enrichment

To identify whether positive selection across the honey bee species quantitatively relates to particular functions, we classified genes based on their Gene Ontology (GO) annotation from *A. mellifera* orthologs. Using SUMSTAT (Roux et al. 2014) with the topGO R package (Alexa et al. 2006) to test for gene set enrichment, we identified 51 significant functional categories, of which 45 were enriched and six depleted in genes under positive selection at 20% FDR. Most functional categories enriched with positively selected genes were unique for each branch, with the exception of "ATP-dependent microtubule motor activity", which was shared among the three *Apis* species and "mitochondrial translation-related functions", which was enriched in all branches but *A. florea* (Figure 2). In addition, *A. dorsata* and *A. mellifera* shared similar functional categories involved in cellular ion exchange (Supplemental Table S3). GO terms depleted of positively selected genes were mostly found in the Apis branch and were linked to the regulation of transcription (Figure 3).

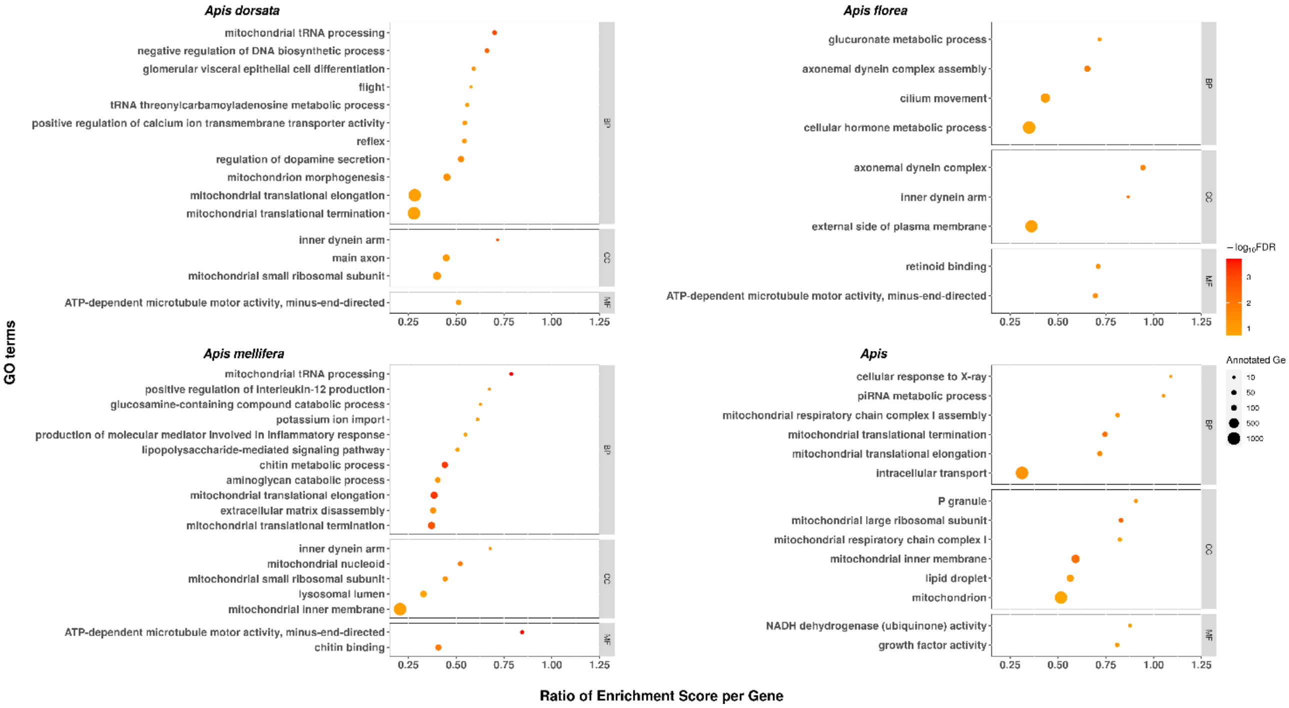

Figure 2 – Functional categories enriched with genes under positive selection in each honey bee species and their most recent common ancestor. GO terms enriched in positively-selected genes are depicted as spheres representing the number of annotated genes (sphere size) and the -$\log_{10}$ of their FDR (color intensity). GO enrichment scores, normalized by the number of annotated genes, are indicated by the x-axis. Most enriched GO terms with positively selected genes can be interpreted as adaptations to long

distance migration and increased colony size in *A. dorsata*, colony defense in *A. florea*, immunity in *A. mellifera*, and TE silencing and high recombination rates in the basal *Apis* lineage. BP = Biological Process, CC = Cellular Component, MF = Molecular Function.

The *Apis* branch revealed 14 enriched GO categories including the "piRNA metabolic process" and "cellular response to X-ray". The former could relate to the particularly low TE content of honey bees (Petersen et al. 2019) because piRNAs silence transposable elements (Ernst et al. 2017), while the latter might explain the honey bees' high genomic recombination rates (Rueppell et al. 2016) due to its link to DNA double strand breaks (DSB) that are required to initiate recombination (Aguilera and Gómez-González 2008). GO categories enriched in *A. florea* included "hormone and glucuronate metabolism", and "retinal proteins". The GO categories "glomerular visceral epithelial cell differentiation", "dopamine metabolism", "flight", and "negative regulation of DNA biosynthesis" were enriched for positive selection in *A. dorsata*. The *A. mellifera* branch was enriched in "chitin metabolism" and "inflammatory response".

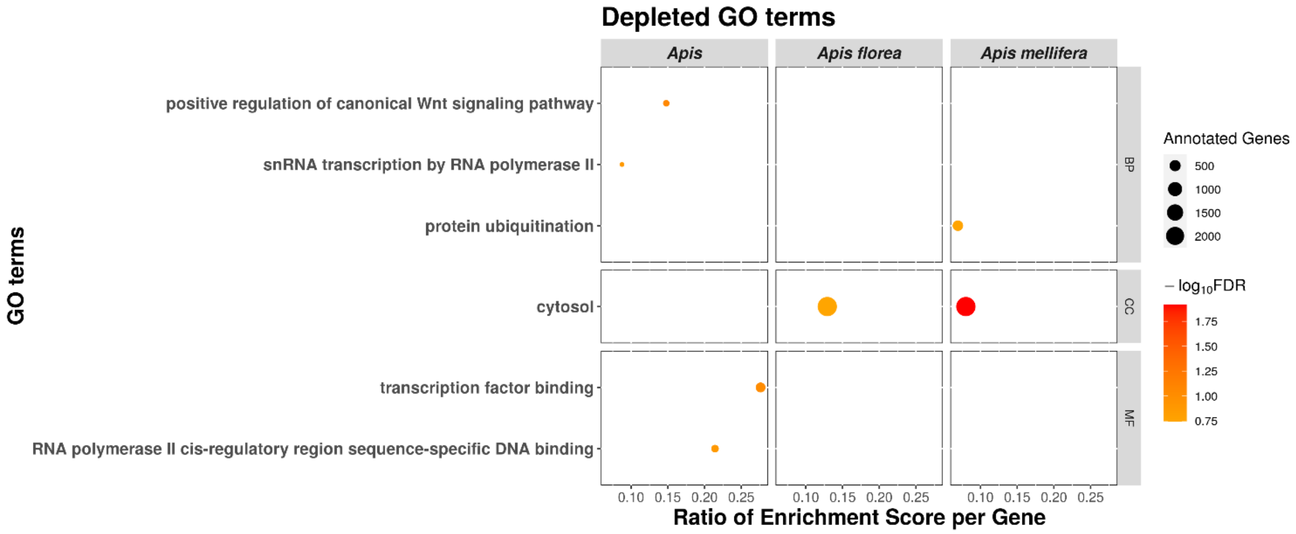

Figure 3 - Functional categories depleted of genes under positive selection in each honey bee species and their most recent common ancestor. Spheres indicate GO terms depleted of positively-selected genes, where size represents the number of annotated genes and color intensity the significance (-log$_{10}$ of their FDR). The x-axis represents the normalized GO enrichment score divided by the number of annotated genes. Most of the GO terms depleted in genes under positive selection are found in the basal *Apis* branch and relate to transcription functions. No depleted GO term was found in *A. dorsata*. BP = Biological Process, CC = Cellular Component, MF = Molecular Function.

Overlap analyses

A comparison of genes we identified as positively selected with published lists of genes of functional significance in *Apis* identified numerous overlapping genes (Supplemental Table S4) but did not reveal any quantitatively significant overlap. None of our four lists (*Apis* branch, *A. florea* branch, *A. dorsata* branch, and *A. mellifera* branch) exhibited significantly more overlap than expected by chance with inter-specific differences in brain gene expression (Sarma et al. 2007). There was also no significant overlap with functional gene lists identified by intra-specific studies, such as selected genes within *A. mellifera* (Wallberg et al. 2014) genes involved in *A. mellifera* caste determination (Chen et al. 2012), worker reproduction (Cardoen et al. 2011), worker behavioral ontogeny (Whitfield et al. 2006; Khamis et al. 2015), and queen-worker brain differences (Grozinger et al. 2007). The largest overlap ($p = 0.0012$) was found between genes selected in the *A. mellifera* branch and genes in the midgut that were up-regulated in *A. mellifera* foragers compared to nurses (Jasper et al. 2015) but correcting for the 72 independent comparisons made to this particular data set alone rendered the overlap non-significant.

QTL	**Branch with sign of selection**	**RefSeq ID**	**Gene Description**	***Apis mellifera* homolog**	**Putative function**
pln1	*A. dorsata*	102675389	forkhead box protein P1-like	AMEL3B67976-RH	Versatile transcription factor
pln4	*A. dorsata*	102679494	arrestin domain-containing protein 17-like	AMEL3B68030-RA	Unknown
pln4	*A. dorsata*	102674786	intersectin-1-like	AMEL3B68033-RB	Neuronal endocytosis
wos1	*A. dorsata*	102679612	dynamin	AMEL3B62415-RB	Membrane fissioning in the nervous system
wos2	*A. mellifera*	102653640	glutamate receptor 1	AMEL3B61681-RB	Neurotransmission
wos2	*A. dorsata*	102677058	deubiquitinase DESI2	AMEL3B61581-RA	Deubiquitination
wos2	*A. florea*	100867905	uncharacterized LOC100867905	AMEL3B61701-RA	Unknown
wos2	*A. florea*	100863251	high affinity cAMP-specific and IBMX-insensitive 3',5'-cyclic phosphodiesterase 8	AMEL3B61641-RA	Intracellular signaling
wos3	*A. mellifera*	726989	E3 ubiquitin-protein ligase listerin	AMEL3B62585-RA	Neurodegeneration

Table 1 – Overlap of positively selected genes with genes present in QTL studies.

The positively selected genes were also compared to positional candidates in the confidence intervals of published intra-specific quantitative trait loci for the pollen hoarding syndrome, specifically foraging behavior (pln1-4) and ovary size (wos1-5) (Hunt et al. 2007; Graham et al. 2011; Rueppell et al. 2011a). Nine positively selected genes were located in these genome regions. Five of these genes showed evidence of selection in the *A. dorsata* branch and none in the *Apis* branch. Known functions of the genes were diverse with a bias towards functions in the nervous system (Table 1).

Lineage-specific genes

Lineage-specific genes have received increased attention, due to their potential role in lineage- or species-specific trait evolution (Simola et al. 2013; Jasper et al. 2015). To understand the role of lineage-specific genes in the diversification of honey bees, we performed a gene-set-enrichment analysis by comparing GO term annotations of the lineage-specific genes (Figure 1) to our orthogroups. The majority of lineage-specific genes (1,994 in *A. florea* (92.2%)*,* 560 in *A. dorsata* (95.2%), and 1,218 in *A. mellifera* (91.5%)) could not be categorized into a functional group nor into previously characterized protein families (Supplemental Table S5). Accordingly, the GO analysis revealed only a few enriched terms for *A. florea* at 20% FDR, including “carbohydrate metabolic process”, “hydrolase activity, hydrolyzing O-glycosyl compounds”, and “DNA integration” (Supplemental Table S5). Although not significantly enriched in the GO term analysis, the *A. dorsata* genome contained two lineage-specific genes related to vision, *gelsolin-like* and *calphotin-like* and the *A. mellifera* genome also revealed several lineage-specific genes of interest (Supplemental Table S5).

Chemosensory gene evolution

Chemosensory diversification is important for insect evolution (McBride et al. 2014; Brand et al. 2020) but automated annotation of chemosensory genes remains problematic. Thus, we manually annotated and analyzed five chemosensory gene families involved in olfaction and gustation: odorant binding proteins (OBPs), chemosensory proteins (CSPs), odorant receptors (ORs), gustatory receptors (GRs) and ionotropic receptors (IRs) (Sánchez-Gracia et al. 2009; Croset et al. 2010).

The number of chemosensory genes in *A. dorsata* and *A. florea* (Supplemental Table S6) was similar to the previously described gene sets in *A. mellifera* for all chemosensory gene families (Brand and Ramírez 2017; Karpe et al. 2016; Robertson and Wanner 2006), with a large number of 1:1:1 orthologous genes between the three species (from 66% in ORs to 100% in CSPs and IRs). Additionally, we found conservation of genes, such as the 9-ODA receptor gene *OR11*, across species. While we did not detect any variation in CSPs and IRs across the honey bees, OBPs, ORs, and GRs varied in the number of genes, revealing gains and losses (Figure 4, Supplemental Figs. S1, S2). The most variable clades in all three of these gene families, previously identified as specific to honey bees in comparison to other corbiculate bees (Brand and Ramírez 2017), were similar in numbers for all three species analyzed but revealed complex phylogenetic relationships, including the OR 9-exon subfamily.

In addition to these patterns shared among gene families, we found that the number of GRs in the newly annotated *A. florea* and *A. dorsata* genomes differed substantially from *A. mellifera*. Previous annotations of the *A. mellifera* genome reported a total of 15 GR genes including 11 functional and 4 pseudogenized copies (Robertson and Wanner 2006; Smith et al. 2011). In addition to single copies for each of the functional GRs known from *A. mellifera*, we identified 19 and 15 GRs in *A. dorsata* and *A. florea*, respectively (Figure 4). Of these, 8 and 2 were likely pseudogenes, respectively, and all of these GRs formed a monophyletic clade with the three previously described X, Y, and Z *A. mellifera* pseudogenes (Figure 4). Several of the XYZ-homologous GRs showed 1:1 homology between *A. dorsata* and *A. florea*, as well as the *A. mellifera* pseudogenes. A reannotation of the *A. mellifera* GR gene family including the previously reported >50 fragmented GR pseudogenes (Robertson and Wanner 2006), reconstructed all known functional GRs and 88 additional sequences with homology to the X, Y, and Z GR pseudogenes. Six of 11 GRs with a length of at least 300 amino acids contained premature stop codons, while the other 5 represent new, potentially functional GRs.

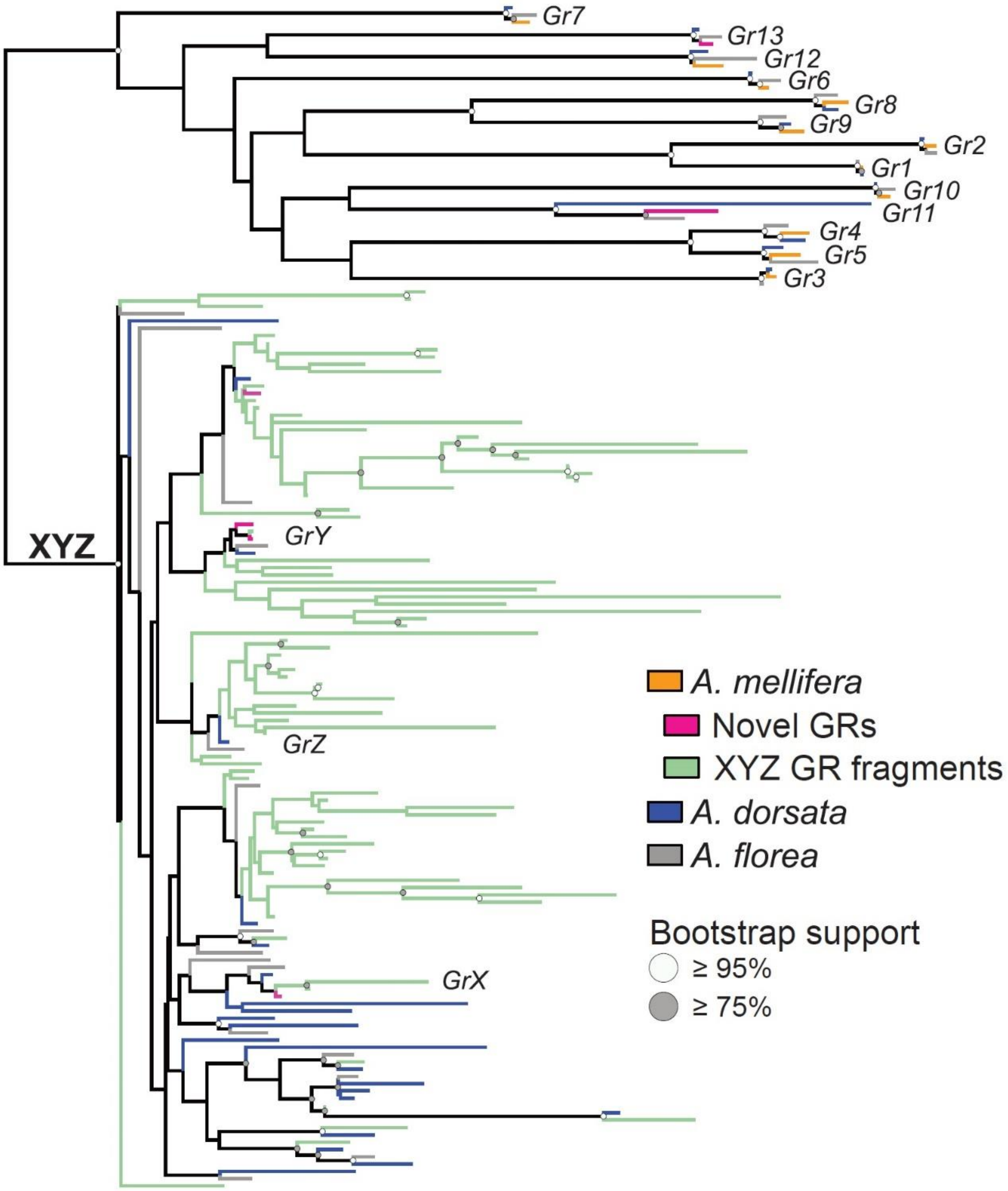

Figure 4 – Gustatory receptor (GR) gene family phylogeny including newly annotated genes of three honey bee species. The Maximum Likelihood tree contained two clades, one including a single ortholog of all putatively functional GRs previously described in *A. mellifera* (highlighted in orange) in each species (blue: *A. dorsata*, grey: *A. florea*), and the XYZ clade (supported with 99% bootstrap support) previously thought to be entirely pseudogenized (Robertson and Wanner 2006, Sadd et al. 2015). Five newly identified full-length GRs for *A. mellifera* are highlighted in pink, some of which are among the newly identified XYZ GRs (4 in *A. mellifera*, 15 in *A. florea*, and 19 in *A. dorsata*). All GR groupings outside the XYZ clade have high bootstrap support (see Supplemental Figure S2 for exact support values), highlighting

the conservation of GR gene number in this group across *Apis*. In addition to >50 small fragments with homology to GRs (light green, only *A. mellifera* fragments shown), we newly identified a number of full-length genes in the XYZ clade, all of which are supported by gene expression data in *A. mellifera*. The fragments are included here for to represent all of our results, although the GR phylogeny is much clearer without them (Supplemental Figure S2). With 16 to 26 putatively functional GRs per species, honey bees are similar to other corbiculate bees (Brand and Ramirez 2017), suggesting that the sense of taste in honey bees is more sophisticated than previously thought.

To validate potential functionality of the newly described GRs, we visualized gene models along with RNA-seq tracks in the *A. mellifera* Apollo browser (Dunn et al. 2019) available at the Hymenoptera Genome Database (Elsik et al. 2016). Four of the GR gene models were supported by RNA-seq reads spanning predicted exon-intron boundaries, indicating they are actively transcribed and thus functional receptors. The only novel full-length GR without expression support was highly similar to *GR13*, which was also present in the genomes of *A. dorsata* and *A. florea* and has known orthologs in several other corbiculate bees (Brand and Ramírez 2017), suggesting it is a conserved functional GR as well. Several of the smaller fragments were also supported by expression data, suggesting that they might be part of coding genes that are not well assembled. Indeed, all but one of the newly identified GR sequences were located on small scaffolds not assigned to linkage groups (‘Un’-scaffolds) and gene models were often truncated at the end of a scaffold. Accordingly, it is likely that the additional 5 GRs we identified for *A. mellifera* are an underestimation of the real number of honey bee-specific GRs in the XYZ-subfamily (Brand and Ramírez 2017).

Discussion

Fine-scale comparative genomic analyses lead to a better understanding of the molecular basis of species diversification and increased resolution of genomic feature evolution. Our genome-wide analysis reveals increased positive selection pressure during the diversification of the three honey bee lineages after the divergence of *Apis* from its most recent common ancestor with *Bombus*. Our results parallel previous analyses that indicate accelerated evolution during the diversification of species within a family (Nevado et al. 2016; Tollis et al. 2018; Vianna et al. 2020), suggesting a common evolutionary pattern. We also find evidence for selection for sequence changes in existing protein coding regions and evolutionary turn-over of genes, similar to a genomic study of the radiation of closely related bumble bees (Sun et al. 2021). These two sources of evolutionary change may be important in bee social evolution in addition to regulatory diversification (Kapheim et al. 2015). Practically, rapid evolutionary divergence may not be easy to distinguish from evolution of novel genes, unless sufficient similarity remains to distinguish orthologs from paralogs as in our manual *Apis* chemoreceptor analyses. We believe that our extensive search for taxonomically restricted genes resulted in unrealistically high estimates of novel genes because the majority of these genes have only support from one prediction method. However, the findings suggest the existence of at least some additional species-specific genes within *Apis* that deserve further study.

We did not identify significant overlap between the genes found to be positively selected among species and genes that determine intra-specific variation in key traits of honey bees, which we predicted based on the hypothesis that phenotypic plasticity is a main driver of *Apis* diversification (West-Eberhard 2003; Kapheim et al. 2020). In contrast to the stark phenotypic differences of honey bees to their closest contemporary relatives, relatively few genes were identified as positively selected in the shared evolution of all honey bees (basal *Apis* branch) compared to the number of positively selected genes detected across branches within *Apis* (species branches). Although we lack a comprehensive explanation for the relatively low number of positively selected genes, it is plausible that evolution at this stage was more strongly driven by gene regulatory changes (Kapheim et al. 2015) or the appearance of *Apis*-specific genes.

In additional to the computational prediction of additional genes, our manual analysis corrected previous results of low numbers of GR genes in honey bees (11 GRs, Robertson and Wanner 2006): We were able to identify 22, 26, and 16 complete GR genes in *A. dorsata*, *A. florea*, and *A. mellifera*, respectively, aided by an updated genome assembly for *A. mellifera* (Elsik et al. 2014). This increase of full-length GRs in *A. mellifera* by almost 50% is presumably still an underestimate due to low quality sequence assembly of the respective parts of the genome. Thus, the sense of taste in honey bees may be more sophisticated than previously thought (Wright et al. 2010). Furthermore, the XYZ-subfamily, which is only found in *Apis* (although one instance has been reported from *Bombus terrestris* (Sadd et al. 2015)), revealed complex evolutionary dynamics suggesting an evolutionary history of gustatory functions specific to honey bees.

Together, this makes the XYZ-subfamily an interesting target to understand the evolution of chemosensory capabilities in honey bees.

The evolution of *Apis* supports previous studies on the molecular basis of increased social complexity

The rise of eusociality in insects has been linked with an increased capacity of gene regulation and the rapid evolution of chemoreceptors, despite the small number of fast-evolving genes shared among eusocial insects (Woodard et al. 2007; Simola et al. 2013; Terrapon et al. 2014; Kapheim et al. 2015, 2020; Harrison et al. 2018).

While our analyses support the importance of chemosensation, we found that the divergence of the *Apis* ancestor from the most recent common ancestor with *Bombus* was accompanied by a depletion of positively selected genes from functional categories related to transcription, such as “transcription factor binding”. The major evolutionary transition to eusociality was not captured in our contrast between *Bombus* and *Apis* and our results may thus reflect a subsequent conservation of gene regulatory mechanisms that consolidate and stabilize the progress of a rapid transition to sociality. Subsequent gene regulatory changes in the evolution of *Apis* may have been achieved by more specific mechanisms: genes involved in growth factor activity, a major pathway of the regulation of gene expression, were fast evolving in the ancestor of all *Apis* species. The rapid evolution of piRNA metabolism in honey bees might also be linked to the regulation of gene expression in *Apis*, as it regulates gene expression and epigenetic effects in *Drosophila* (Weick and Miska 2014; Glastad et al. 2018) and piRNAs target regions anti-sense of protein-coding genes in honey bees, suggesting that they could control transcription (Wang et al. 2017).

Chemosensory gene evolution has been hypothesized to be important during the evolution of eusociality (Harrison et al. 2018). The 9-exon OR gene family has been hypothesized to be important in social communication in Hymenoptera, due to a role of 9-exon ORs in the detection of CHCs in ants (Smith et al. 2011; McKenzie et al. 2016; Slone et al. 2017; Pask et al. 2017). Our results demonstrate that the OR 9-exon subfamily evolves rapidly between the three *Apis* species, which occurs also more widely (Sadd et al. 2015; Brand and Ramírez 2017). In contrast, sex pheromone receptor genes (*OR11*, *OR10*, *OR18*, and *OR170*) were highly conserved. Moreover, we found that the expansion of OBPs is not specific to *A. mellifera* (Brand and

Ramírez 2017) but most likely occurred in the common ancestor of *Apis* species, pointing to a role in chemosensory behaviors unique to honey bees.

Apis evolution reveals an evolutionary trade-off between genome stability and variability

While genome stability is vital for organisms and crucial for maintenance of optimally adapted phenotypes, it restricts genetic diversity, which is essential for evolutionary and physiological processes, particularly in eusocial insects (Mattila and Seeley 2007; Seeley and Tarpy 2007; Kent et al. 2012). The resulting trade-off between genome stability and diversity was reflected in our findings that TE silencing and DSB repair pathways in the *Apis* lineage were positively selected. The honey bee genomes are depleted of TEs (Elsik et al. 2014; Park et al. 2015) and we found that the regulation of one of the major mechanisms to prevent TE spread within a genome, piRNAs (Brennecke et al. 2007; Ernst et al. 2017), was positively selected in *Apis*. The enrichment of the piRNA regulatory pathway, as well as the GO term “P granule cellular component” (Lim and Kai 2007), among positively selected genes in the *Apis* lineage suggests that positive selection can act on piRNAs over evolutionary time to limit the spread of TEs despite consistently high rates of recombination (Rueppell et al. 2016).

The high recombination rates of all *Apis* species studied so far, ranging from 20 to 25 cM/Mb (Hunt and Page 1995; Meznar et al. 2010; Ross et al. 2015; Rueppell et al. 2016), may increase genetic diversity and facilitate evolutionary novelties (Kent et al. 2012). The enrichment of rapidly evolving genes associated with the cellular response to X-rays in the *Apis* ancestor indicates a corresponding adaptation to double strand breaks (DSBs) of DNA (Rothkamm and Löbrich 2003). It is unclear whether this selective signature should be interpreted as a cause or consequence of the high recombination rates but mutations in genes involved in DSB repair can lead to higher homologous recombination rates (Aguilera and Gómez-González 2008). The accelerated molecular evolution of DSB repair genes may thus have enabled the high meiotic recombination rates of honey bees, with potential effects on genome evolution and diversity (Kent et al. 2012).

The continuous oogenesis of Hymenoptera (Büning 1994) can exacerbate the accumulation of mutations during later-life meiosis (Bromham and Leys 2005; Thomas et al. 2010), particularly in females that produce numerous offspring. The resulting mutational load is particularly severe in mitochondria (Neiman and

Taylor 2009). Nuclear genomes can co-evolve to compensate the loss of mitochondrial function via the accumulation of deleterious mutations (Hill 2020), resuling in increased evolutionary rates of mitochondrion-destined nuclear genes (Li et al. 2017). Correspondingly, we found positive selection of nuclear genes involved in the mitochondrial translation elongation and termination pathway in the *Apis* lineage and in the *A. mellifera* and *A. dorsata* branches, the two species with the largest colony sizes, suggesting selection for increased efficiency and accuracy of mitochondrial translation (Schneider 2011) in the face of increased mutations with colony size increases. This hypothesis is also compatible with the strong positive selection targeting the negative regulation of DNA biosynthesis and the tRNA threonylcarbamoyladenosine metabolism essential for accurate translation (Yarian et al. 2002) in *A. dorsata*, the honey bee species with the greatest colony size (Oldroyd and Wongsiri 2006). Hence, the molecular evolution of honey bee genomes suggests an evolutionary trade-off between maintaining genome integrity and generating genetic diversity.

Fine-scale comparative genomics reveals candidates for the evolution of key phenotypic traits

Accordingly with fundamental differences in body size and queen-worker caste divergence among the three *Apis* lineages (Wongsiri and Oldroyd 2006; Rueppell et al. 2011b), we found several positively selected genes predicted to belong to gene families involved in growth and reproductive processes: a *G-protein-coupled receptor* with similarities to the life-history regulator *methuselah* (Delanoue et al. 2016) and the ovary determinant *tudor* (Xie et al, 2019) in the basal *Apis* branch, *pde8* involved in ERK-signaling that has multiple life-history coordinating roles (Brown et al. 2013) in the *A. florea* branch, and the putative growth effectors *short neuropeptide F receptor* (Lee et al. 2008), *farnesol-dehydrogenase* (Mayoral et al. 2009), and *cdk2* (Vidwans and Su 2001) in the giant honey bee lineage.

The evolutionary diversification of nesting behavior into cavity-nesting in *A. mellifera* and related species versus open-nesting in the other lineages has been highly controversial for decades and has direct ramifications for understanding the evolution of the honey bee dance language (Koeniger 1976; Oldroyd and Wongsiri 2006; Raffiudin and Crozier 2007; Koeniger et al. 2011). Our analysis cannot resolve this controversy but provides some support for a transition from cavity-nesting to open-nesting within *Apis*: While no genes or GO terms that could be interpreted as adaptions to open-nesting were found to evolve under positive selection in the ancestral *Apis* branch, in *A. florea*, which accurately controls nest temperature despite its open-nesting habit (Oldroyd and Wongsiri 2006), lineage-specific genes were associated with carbohydrate metabolism, a pathway associated with thermoregulation in bees (Woodard et al. 2011).

While all honey bees migrate, only giant honey bees seasonally migrate over long distances, up to 100-200km in *Apis dorsata* (Oldroyd and Wongsiri 2006). Correspondingly, we found potential molecular signatures of adaptions to long-distance migration in the *A. dorsata* lineage: Positive selection in genes linked to “flight” along with large musculature and body size (Dulta and Verma 1987), involved in “mitochondrial morphogenesis” that may affect energy metabolism during migration (Sogl et al. 2000; Li et al. 2018), associated with the renal system (i.e. “glomerular visceral epithelial cell differentiation”) allowing water conservation during migration (Wigglesworth 1932), and “regulation of dopamine secretion”, a pathway involved in migration in locusts (Ma et al. 2011). The adaptation to night foraging in *A. dorsata* enables them to detect objects at lower light intensity than expected by their ommatidium structure (Warrant et al. 1996). This might be explained by 2 *A. dorsata*-specific genes, homologs of genes involved in phototaxis, *gelsolin-like* (Stocker et al. 1999), and vision, *calphotin-like* (Yang and Ballinger 1994). An enhanced floral scent detection in *A. dorsata* may also be beneficial for night foraging, which is suggested by the lineage-specific duplications and pseudogenization events of *OR151* and *OR152*, important for detection of floral compounds (Claudianos et al. 2014).

The *A. mellifera* branch is mainly associated with positive selection on genes involved in chitin metabolic processes, as previously found to be enriched in positively selected genes in *A. mellifera* and bumble bees (Harpur et al. 2014; Sun et al. 2021). They mostly relate to caste differentiation (Santos and Hartfelder 2015; Malka et al. 2014; Li et al. 2012) and immunity (Oddie et al. 2018; Harpur and Zayed 2013), which may be caused by pathogen pressure in the relative stable and long-lasting nests of cavity-nesting species.

Focusing on the main lineages of the unique honey bee genus, our study identifies positively selected genes that warrant further study. Of particular interests are selected genes with putative molecular functions that may link them to key adaptations and the diversification among *Apis* species. Even though the genus *Apis* is small and contains only the three subgeneric lineages included in this study, sequencing other *Apis* species to increase phylogenetic depth may further refine our conclusions about *Apis* evolution and enhance our understanding of genome evolution in dwarf, giant, and cavity-nesting honey bees. Overall, our results provide an evolutionary scenario of an *Apis* ancestor adapted to building a vertical comb, likely in cavities, that allowed for increased colony size.

Methods

Specimen collection

Haploid drones collected from a single colony per species were used for *A. florea* and *A. dorsata* genome sequencing. The samples of *A. florea* were collected in 2009 from Chiang Mai, Thailand. The samples of *A. dorsata* were collected in the vicinity of the Agricultural Research Station Tenom (Sabah, Malaysia: 5.4° N/115.6° E) in March 2007. Samples were preserved in RNAlater™ and subsequently frozen until total DNA extraction from single individuals.

Genome sequencing and assembly

Two types of WGS libraries, a fragment library and mate-pair libraries with 8 kb inserts, were used to generate the *Apis florea* genome sequencing data using 454 Titanium technology. The Aflo_1.0 genome assembly was generated by assembling WGS reads using Newbler (2.3-PreRelease-10/19/2009) (Margulies et al. 2005). Reads from each Newbler scaffold were grouped, along with any missing mate-pairs, and reassembled using PHRAP (Bastide and McCombie 2007) in an attempt to close the gaps within Newbler scaffolds.

For *A. dorsata*, four libraries were sequenced on an Illumina GA platform for the assembly: (1) 2 × 125bp paired-end reads from a 500bp library; (2) 2 × 125bp mate-pairs from a 1.2kbp library; (3) 2 × 125bp mate-pairs from a 3kbp library, and (4) 2 × 36bp mate-pairs from a 5kbp library. The sequencing reads from all four libraries were first error corrected and trimmed using Quake v0.2.0 (Kelley et al. 2010). Error corrected reads were then assembled using SOAPdenovo v1.0.5 (Li et al. 2010) (Supplemental Methods).

Completeness of the two assemblies was assessed by identifying Benchmarking Universal Single-Copy Orthologs (BUSCOs) using the BUSCO v5beta pipeline in genome mode (Simão et al. 2015). For this analysis, we identified single-copy orthologs based on the hymenoptera_db10.

Genome annotation

To avoid artifacts stemming from different annotation methods (Supplemental Methods) a combined gene set was created for each species, by adding non-overlapping genes from different annotation pipelines to a fundamental NCBI RefSeq annotation in the following orders: *A. dorsata*, RefSeq → EVM (Haas et al. 2008) → MAKER (Holt and Yandell 2011) → AUGUSTUS -CGP (Stanke et al. 2008; König et al. 2016; Nachtweide and Stanke 2019); *A. florea*, RefSeq → EVM → AUGUSTUS -CGP → BGI (Kapheim et al. 2015); *A. mellifera*, RefSeq → OGS (Elsik et al. 2014) → AUGUSTUS -CGP. Accuracy of all gene prediction methods were assessed (Supplemental Tables S7, S8) and combined in EVM with different weights (Supplemental Tables S9, S10) based on different sources (Supplemental Tables 11, 12), resulting in 12,172 genes for *A. dorsata* (Supplemental Table S13) and 14,393 for *A. florea* (Supplemental Table S14).

Exonerate protein2genome (Slater and Birney 2005) was used to align protein sequences from each species to the genome assemblies of the other two species (*A. mellifera*: BioProject PRJNA10625 and *Bombus impatiens*: BioProject PRJNA61101 and *B. terrestris* BioProject PRJNA45869). For each species, a new gene model was created wherever there was a protein alignment that did not overlap with an existing gene model. At each new gene locus with more than one alternate species alignment, the alignment with the best score was used to generate a single protein-coding gene model, correcting any artifactual frameshifts in protein and coding sequences. The protein homolog-based gene models were added to the combined gene sets to create the final gene sets, deemed “comparative gene sets”, used in this study. Although some of the protein homolog-based predictions were not of sufficient quality for evolutionary analysis, including them in the comparative gene sets allowed us to determine more realistic numbers of species-specific genes.

Gene set annotation

We used InterProScan (Zdobnov and Apweiler 2001) to compare protein sequences to InterPro (Finn et al. 2017) protein domain and other motif databases (Supplemental Methods). InterProScan assigns Gene Ontology (GO) (Ashburner et al. 2000) terms and pathway ids from KEGG (Chen et al. 2012), MetaCyc (Caspi et al. 2018) and Reactome (Fabregat et al. 2018) based on protein domain content. We used FASTA (Pearson and Lipman 1988) with an E-value threshold of 1×10^{-6} to compute reciprocal alignments between *Apis* comparative proteins and a *Drosophila melanogaster* protein set consisting of the longest protein isoform of each gene (annotation version r6.14). We identified reciprocal best hits (RBH) and transferred GO, KEGG, PANTHER and REACTOME annotations from the *D. melanogaster* protein to the *Apis* protein for each RBH pair, using the annotation files available at FlyBase (Gramates et al. 2017). Finally, we obtained gene descriptions from NCBI for the RefSeq (O'Leary et al. 2016) gene annotations.

Ortholog prediction

We created ortholog groups containing one gene from the two newly annotated genomes of *Apis dorsata* and *A. florea* and the existing *A. mellifera* genome (Amel_4.5, under BioProject PRJNA10625). Protein sequences from the three comparative gene sets were combined into one file that was used in an all by all protein comparison with FASTA (Pearson and Lipman 1988) using an E-value threshold 0.001 to identify single-copy orthologs (Supplemental Methods). This process resulted in 15,182 families of *Apis* orthologs. Of those, 5310 families were flagged because a translational discrepancy in the NCBI GFF or a frameshift/gap in the Exonerate alignment were indicated. After creating the families of *Apis* orthologs, a *Bombus* protein to serve as an outgroup was identified for each family (Supplemental Methods). 9310 *Apis* ortholog families were assigned a *Bombus* protein.

Multiple sequence alignment

For each ortholog family, the longest protein isoforms for each species were used in multiple sequence alignment with PRANK (v.150803) (Löytynoja and Goldman 2008) and unreliably aligned residues were masked with GUIDANCE (v2.02) (Penn et al. 2010). A custom Python script (Supplemental Code) was then used to replace protein sequences with coding sequences in the multiple alignments, resulting in 8115 gene families after filtering (Supplemental Methods). The mean length of filtered alignment was 1621 nucleotides (median = 1233 nucleotides), ranging from 303 to 22830 nucleotides.

Phylogeny

Gene family phylogenies were built using RAxML (v7.2.9) (Stamatakis 2006) from the amino acid sequences (9310 *Apis* ortholog families). For each ortholog family, ModelGenerator was used to select the best amino-acid matrix and substitution model (Keane et al. 2006). The species phylogeny was built from a concatenation of all amino-acid alignments with *B. impatiens* or *B. terrestris* data (9275), using RaxML with an estimated amino-acid matrix based on our data (GTR) and the CAT model (Rokas 2011).

Branch-site test for positive selection

The adaptive branch-site random effects model [aBSREL, (Smith et al. 2015)] from Hyphy software package (Kosakovsky Pond et al. 2019) was used to detect positive selection experienced by a gene family in a subset of sites in a specific branch of its phylogenetic tree. Due to our low phylogenetic depth, test for positive selection was run only on the *Apis*, *A. mellifera*, *dorsata* and *florea* branches (all “leaves”). To account for multiple testing (Anisimova and Yang 2007), p-values from the successive 32460 tests were corrected using the False Discovery Rate (FDR) (Benjamini and Hochberg 1995). Due to our stringent alignment filtering and the multiple testing correction as one series, we set our significant threshold at 10%. We visually checked alignments of positive results and excluded GC-biased gene conversion because our ω estimates were negatively correlated with GC content (Spearman’s Correlation: S = 6.7e12, rho= -0.17, $P < 2.2 \times 10^{-16}$).

Overlap analysis

Our lists of selected genes were compared to multiple other studies. The only other available inter-specific study (Sarma et al. 2009) and the following intra-specific studies that have identified gene sets of functional significance for the observed inter-specific differences within *Apis* were selected: Genes involved in caste determination (Chen et al. 2012), reproductive phenotypes (Grozinger et al. 2007; Cardoen et al. 2011), and genes involved in local adaptation (Wallberg et al. 2014). In addition, overlap to quantitative trait loci for ovary size (Rueppell et al. 2011a; Graham et al. 2011) and social behavior (Hunt et al. 2007; Rueppell 2009) was evaluated.

Tests of functional category enrichment

Gene Ontology (GO) (Ashburner et al. 2000) annotations for our gene families were taken from *A. mellifera*, annotated with GO terms as described above. To identify functional biases, the package topGO version 2.4 (Alexa et al. 2006) of Bioconductor (Gentleman et al. 2004) was used with the full data-set (before filtering) of genes containing a GO annotation as reference. Functional biases were detected using Fisher's exact test with the 'elim' algorithm of topGO and selected based on FDR<20% (Supplemental Methods). Gene Ontology categories mapped to less than 10 genes were discarded. To identify functional categories enriched with genes under positive selection, the SUMSTAT test was used (Supplemental Methods). We performed bidirectional tests to account for enrichment and depletion for positively selected genes in a gene set. Gene Ontology categories mapped to less than 10 genes were discarded.

Lineage Specific Genes

We identified genes specific to one or two *Apis* genomes using outputs of the all-by-all FASTA protein comparison and Exonerate protein2genome alignments described above. If all protein isoforms encoded by a particular gene were missing protein or Exonerate alignments to another species, that gene was considered missing in the other species. We excluded genes due to bacterial contamination (Supplemental Methods). To investigate whether lineage specific genes of each *Apis* species are associated with features of their biology, their GO annotations were compared to the ortholog families dataset using Fisher's exact test with the 'elim' algorithm of topGO.. Gene Ontology categories mapped to less than 10 genes were discarded.

Chemosensory gene family analysis

Annotation and selection analysis of chemosensory gene families followed Brand and Ramírez (2017). In brief, high-quality annotations for *A. mellifera* were used to annotate odorant receptors (Robertson and Wanner 2006), odorant binding proteins (Forêt and Maleszka 2006), chemosensory genes (Forêt et al. 2007), gustatory receptors (Robertson and Wanner 2006), and ionotropic receptors (Croset et al. 2010) using exonerate (Slater and Birney 2005) coupled with manual curation and, if necessary, correction of gene models for *A. dorsata* and *A. florea*. In addition, we re-annotated the OR and GR gene families in *A. mellifera* (Robertson and Wanner 2006), and the OR gene family for *A. florea* (Karpe et al. 2016). The resulting gene models were aligned with MAFFT (Katoh and Standley 2013) and used to reconstruct gene family-specific gene trees with RaxML (Stamatakis 2006) using 20 independent ML searches and 100 bootstrap replicates. Selection analyses were performed with the aBSREL algorithm in HYPHY. ORs were divided into subfamilies as defined in Brand and Ramirez 2017, while all other gene families were analyzed as a whole. P values for each independent aBSREL run were corrected for multiple testing using a FDR of 5%.

Data Access

The biological data, sequencing data, assembled genome sequences, and annotations generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession numbers PRJNA174631 (*A. dorsata*) and PRJNA45871 (*A. florea*).

Acknowledgements

We thank Salim Tingek for hosting OR and NK at the Agricultural Research Station, Tenom, Malaysia. Research reported by OR in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number R15GM102753, the National Institute on Aging under award number R21AG046837, and the UNCG Florence Schaeffer Endowment. CGE was supported by the National Science Foundation under award # IIA-1355406. Further support was provided to GER by the Illinois Sociogenomics Initiative. Funding for *A. florea* genome sequencing and assembly and was provided by NHGRI grant U54 HG003273 to R.A. Gibbs. MCS acknowledges the US National Science

Foundation award number DBI-1627442 and PC acknowledges the support of the Chiang Mai University fund. Other parts of this work were supported by the INB ("Instituto Nacional de Bioinformatica")-Elixir Spain Project PT13/0001/0021 (ISCIII -FEDER) and the Spanish Ministry of Economy and Competitiveness, under 'Centro de Excelencia Severo Ochoa 2013-2017', SEV-2012-0208."

Author Contributions

BF processed the data, performed the main analyses, interpreted the results, and wrote the first draft of the overall manuscript. PB manually annotated all chemosensory genes with assistance from HNN, interpreted the results, and wrote the corresponding manuscript parts. HNN assisted with the *A. dorsata* genome assembly and identifying QTL overlapping genes, JJH helped write the manuscript, and DE provided MAKER gene models for *A. dorsata* under the supervision of MY. KJH performed the BUSCO analyses and worked together with SN and LR under the leadership of MS to perform gene prediction in the three *Apis* species with AUGUSTUS-CGP. DEH generated RNA-seq alignments, transcript assemblies and intron hints for input to gene prediction. KKOW, HMR and GER were responsible for the main part of the *A. dorsata* sequencing. FC annotated the *A. dorsata* and *A. florea* genomes with EVM under the supervision of RG. NK was responsible for initiating the project and facilitating field collections. PC was hosting the collection of *A. florea* samples. GN performed the *A. dorsata* assembly under the leadership of MCS. KCW was responsible for the *A. florea* genome sequencing, assembly and primary annotation. CGE coordinated the overall project and particularly all gene annotation efforts, performed repeat masking of genomes, generated the final comparative gene sets, annotated proteins using InterPro, searched genomes and proteins for bacterial contaminants, generated the main datasets of orthologs and lineage-specific genes, and participated in the analyses and results interpretation. OR designed and coordinated the overall project, provided the *A. dorsata* samples, secured funds for the project, performed the gene overlap analysis, and helped write the manuscript. All authors read the manuscript and provided feedback to improve the final version.

Disclosure Declaration

The authors have no conflicts of interests to declare.

References

Aguilera A, Gómez-González B. 2008. Genome instability: A mechanistic view of its causes and consequences. *Nat Rev Genet* **9**: 204–217.

Alaux C, Sinha S, Hasadsri L, Hunt GJ, Guzmán-Novoa E, DeGrandi-Hoffman G, Uribe-Rubio JL, Southey BR, Rodriguez-Zas S, Robinson GE. 2009. Honey bee aggression supports a link between gene regulation and behavioral evolution. *Proc Natl Acad Sci U S A* **106**: 15400–15405.

Alexa A, Rahnenführer J, Lengauer T. 2006. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* **22**: 1600–1607.

Anisimova M, Yang Z. 2007. Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. *Mol Biol Evol* **24**: 1219–1228.

Araujo N de S, Arias MC. 2021. Gene expression and epigenetics reveal species-specific mechanisms acting upon common molecular pathways in the evolution of task division in bees. *Scientific Reports* **11**: 3654.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. Gene ontology: Tool for the identification of biology. *Nat Genet* **25**: 25–29.

Aufauvre J, Misme-Aucouturier B, Viguès B, Texier C, Delbac F, Blot N. 2014. Transcriptome analyses of the honeybee response to *Nosema ceranae* and insecticides. *PLoS One* **9**.

Bastide M, McCombie WR. 2007. Assembling Genomic DNA Sequences with PHRAP. *Curr Protoc Bioinforma* **17**: 1–15.

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Ser B (Methodol)* **57**: 289–300.

Brand P, Hinojosa-Díaz IA, Ayala R, Daigle M, Yurrita Obiols CL, Eltz T, Ramírez SR. 2020. The evolution of sexual signaling is linked to odorant receptor tuning in perfume-collecting orchid bees. *Nat Commun* **11**: 244.

Brand P, Ramírez SR. 2017. The evolutionary dynamics of the odorant receptor gene family in corbiculate bees. *Genome Biol Evol* **9**: 2023–2036.

Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, Hannon GJ. 2007. Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell* **128**: 1089–1103.

Bromham L, Leys R. 2005. Sociality and the rate of molecular evolution. *Mol Biol Evol* **22**: 1393–1402.

Brown KM, Day JP, Huston E, Zimmermann B, Hampel K, Christian F, Romano D, Terhzaz S, Lee LC, Willis MJ, Morton DB, Beavo JA, Shimizu-Albergine M, Davies SA, Kolch W, Houslay MD, Baillie GS. 2013. Phosphodiesterase-8A binds to and regulates Raf-1 kinase. *Proc Natl Acad Sci U S A* **110**: E1533--E1542.

Büning J. 1994. *The Insect Ovary: Ultrastructure, Previtellogenic Growth and Evolution*. Chapman & Hall, London.

Calfee E, Agra MN, Palacio MA, Ramírez SR, Coop G. 2020. Selection and hybridization shaped the Africanized honey bee invasion of the Americas. *PLoS Genet* **16**: e1009038.

Cardoen D, Wenseleers T, Ernst UR, Danneels EL, Laget D, De Graaf DC, Schoofs L, Verleyen P. 2011. Genome-wide analysis of alternative reproductive phenotypes in honeybee workers. *Mol Ecol* **20**: 4070–4084.

Caspi R, Billington R, Fulcher CA, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Midford PE, Ong Q, Ong WK, et al. 2018. The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Res* **46**: D633–D639.

Chavez DE, Gronau I, Hains T, Kliver S, Koepfli KP, Wayne RK. 2019. Comparative genomics provides new insights into the remarkable adaptations of the African wild dog (*Lycaon pictus*). *Sci Rep* **9**: 1–14.

Chen X, Hu Y, Zheng H, Cao L, Niu D, Yu D, Sun Y, Hu S, Hu F. 2012. Transcriptome comparison between honey bee queen- and worker-destined larvae. *Insect Biochem Mol Biol* **42**: 665–673.

Claudianos C, Lim J, Young M, Yan S, Cristino AS, Newcomb RD, Gunasekaran N, Reinhard J. 2014. Odor memories regulate olfactory receptor expression in the sensory periphery. *Eur J Neurosci* **39**: 1642–1654.

Croset V, Rytz R, Cummins SF, Budd A, Brawand D, Kaessmann H, Gibson TJ, Benton R. 2010. Ancient protostome origin of chemosensory ionotropic glutamate receptors and the evolution of insect taste and olfaction. *PLoS Genet* **6**: e1001064.

Delanoue R, Meschi E, Agrawal N, Mauri A, Tsatskis Y, McNeill H, Leopold P. 2016 *Drosophila* insulin release is triggered by adipose Stunted ligand to brain Methuselah receptor. *Science* **353**: 1553-1556.

Deplancke B, Alpern D, Gardeux V. 2016. The genetics of transcription factor DNA binding variation. *Cell* **166**: 538–554.

Dieckmann U, Doebeli M, Metz JA, Tautz D. 2004. *Adaptive speciation*. Cambridge University Press.

Dulta P, Verma L. 1987. Comparative biometric studies on flight muscles of honeybees in the genus *Apis*. *J Apic Res* **26**: 205–209.

Dunn NA, Unni DR, Diesh C, Munoz-Torres M, Harris NL, Yao E, Rasche H, Holmes IH, Elsik CG, Lewis SE. 2019. Apollo: Democratizing genome annotation. *PLoS Comput Biol* **15**: 1–14.

Dyer FC., Seeley TD. 1991. Nesting behavior and the evolution of worker tempo in four honey bee species. *Ecology* **72**: 156–170.

Elsik CG, Tayal A, Diesh CM, Unni DR, Emery ML, Nguyen HN, Hagen DE. 2016. Hymenoptera Genome Database: Integrating genome annotations in HymenopteraMine. *Nucleic Acids Res* **44**: D793–D800.

Elsik CG, Worley KC, Bennett AK, Beye M, Camara F, Childers CP, Graaf DC De, Debyser G, Deng J, Devreese B, et al. 2014. Finding the missing honey bee genes: lessons learned from a genome upgrade. *BMC Genomics* **15**: 86.

Ernst C, Odom DT, Kutter C. 2017. The emergence of piRNAs against transposon invasion to preserve mammalian genome integrity. *Nat Commun* **8**: 1–9.

Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, Haw R, Jassal B, Korninger F, May B, et al. 2018. The Reactome Pathway Knowledgebase. *Nucleic Acids Res* **46**: D649–D655.

Figueiró H V., Li G, Trindade FJ, Assis J, Pais F, Fernandes G, Santos SHD, Hughes GM, Komissarov A, Antunes A, et al. 2017. Genome-wide signatures of complex introgression and adaptive evolution in the big cats. *Sci Adv* **3**: 1–14.

Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, Chang HY, Dosztanyi Z, El-Gebali S, Fraser M, et al. 2017. InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res* **45**: D190–D199.

Forêt S, Maleszka R. 2006. Function and evolution of a gene family encoding odorant binding-like proteins in a social insect, the honey bee (*Apis mellifera*). *Genome Res* **16**: 1404–1413.

Forêt S, Wanner KW, Maleszka R. 2007. Chemosensory proteins in the honey bee: Insights from the annotated genome, comparative analyses and expressional profiling. *Insect Biochem Mol Biol* **37**: 19–28.

Gadau J, Helmkampf M, Nygaard S, Roux J, Simola DF, Smith CDR, Suen G, Wurm Y, Smith CDR. 2012. The genomic impact of 100 million years of social evolution in seven ant species. *Trends Genet* **28**: 14–21.

Gentleman R, Carey V, Bates D, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al. 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**: R80.

Glastad KM, Hunt BG, Goodisman MAD. 2018. Epigenetics in insects: Genome regulation and the generation of phenotypic diversity. *Annu Rev Entomol* **64**: 185–203.

Graham AM, Munday MD, Kaftanoglu O, Page RE, Amdam G V, Rueppell O. 2011. Support for the reproductive ground plan hypothesis of social evolution and major QTL for ovary traits of Africanized worker honey bees (*Apis mellifera* L.). *BMC Evol Biol* **11**: 95.

Gramates LS, Marygold SJ, Dos Santos G, Urbano JM, Antonazzo G, Matthews BB, Rey AJ, Tabone CJ, Crosby MA, Emmert DB, et al. 2017. FlyBase at 25: Looking to the future. *Nucleic Acids Res* **45**: D663–D671.

Green L, Battlay P, Fournier-Level A, Good RT, Robin C. 2019. Cis- And trans-acting variants contribute to survivorship in a naïve *Drosophila melanogaster* population exposed to ryanoid insecticides. *Proc Natl Acad Sci U S A* **116**: 10424–10429.

Grozinger CM, Fan Y, Hoover SER, Winston ML. 2007. Genome-wide analysis reveals differences in brain gene expression patterns associated with caste and reproductive status in honey bees (*Apis mellifera*). *Mol Ecol* **16**: 4837–4848.

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J et. al. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biology **9**: R7.

Harpur BA, Zayed A. 2013. Accelerated Evolution of Innate Immunity Proteins in Social Insects: Adaptive Evolution or Relaxed Constraint? *Mol Biol Evol* **30**: 1665–1674.

Harpur BA, Dey A, Albert JR, Patel S, Hines HM, Hasselmann M, Packer L, Zayed A. 2017. Queens and workers contribute differently to adaptive evolution in bumble bees and honey bees. *Genome Biol Evol* **9**: 2395–2402.

Harpur BA, Kent CF, Molodtsova D, Lebon JMDD, Alqarni AS, Owayss AA, Zayed A. 2014. Population genomics of the honey bee reveals strong signatures of positive selection on worker traits. *Proc Natl Acad Sci U S A* **111**: 2614–9.

Harrison MC, Jongepier E, Robertson HM, Arning N, Bitard-Feildel T, Chao H, Childers CP, Dinh H, Doddapaneni H, Dugan S, et al. 2018. Hemimetabolous genomes reveal molecular basis of termite eusociality. *Nat Ecol Evol* **2**: 557–566.

Helbing S, Michael H, Lattorff G, Moritz RFA, Buttstedt A. 2017. Comparative analyses of the major royal jelly protein gene cluster in three *Apis* species with long amplicon sequencing. *DNA Research* **24** (3): 279-287.

Hepburn HR, Radloff SE. 2011. *Honeybees of Asia*. Springer-Verlag, Berlin Heidelberg.

Hill GE. 2020. Mitonuclear Compensatory Coevolution. *Trends in Genetics* **36**: 403-414.

Holmes KC. 2004. Myosin, muscle and motility: Introduction. *Philos Trans R Soc B Biol Sci* **359**: 1813–1818.

Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**(1): 491.

Hunt GJ, Amdam G V., Schlipalius D, Emore C, Sardesai N, Williams CE, Rueppell O, Guzmán-Novoa E, Arechavaleta-Velasco M, Chandra S, et al. 2007. Behavioral genomics of honeybee foraging and nest defense. *Naturwissenschaften* **94**: 247–267.

Hunt GJ, Page RE. 1995. Linkage map of the honey bee, *Apis mellifera*, based on RAPD Markers. *Genetics* **139**: 1371–1382.

Jasper WC, Linksvayer TA, Atallah J, Friedman D, Chiu JC, Johnson BR. 2015. Large-scale coding sequence change underlies the evolution of postdevelopmental novelty in honey bees. *Mol Biol Evol* **32**: 334–346.

Johannessen M, Møler S, Hansen T, Moens U, Van Ghelue M. 2006. The multifunctional roles of the four-and-a-half-LIM only protein FHL2. *Cell Mol Life Sci* **63**: 268–284.

Johnson RN, Oldroyd BP, Barron AB, Crozier RH. 2002. Genetic Control of the Honey Bee (*Apis mellifera*) Dance language: segregating dance forms in a backcrossed colony. *J Hered* **93**: 170–173.

Kapheim KM, Jones BM, Pan H, Li C, Harpur BA, Kent CF, Zayed A, Ioannidis P, Waterhouse RM, Kingwell C, et al. 2020. Developmental plasticity shapes social traits and selection in a facultatively eusocial bee. *Proc Natl Acad Sci U S A* **117**: 13615–13625.

Kapheim KM, Pan H, Li C, Salzberg SL, Puiu D, Magoc T, Robertson HM, Hudson ME, Venkat A, Fischman BJ, et al. 2015. Genomic signatures of evolutionary transitions from solitary to group living. *Science* **348**: 1139–1143.

Karpe SD, Jain R, Brockmann A, 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.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol* **30**: 772–780.

Keane TM, Creevey CJ, Pentony MM, Naughton TJ, McInerney JO. 2006. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evol Biol* **6**: 1–17.

Kelley DR, Schatz MC, Salzberg SL. 2010. Quake: Quality-aware detection and correction of sequencing errors. *Genome Biol* **11**.

Kent CF, Minaei S, Harpur B a., Zayed A. 2012. Recombination is associated with the evolution of genome structure and worker behavior in honey bees. *Proc Natl Acad Sci U S A* **109**: 18012–18017.

Khamis AM, Hamilton AR, Medvedeva YA, Alam T, Alam I, Essack M, Umylny B, Jankovic BR, Naeger NL, Suzuki M, et al. 2015. Insights into the transcriptional architecture of behavioral plasticity in the honey bee *Apis mellifera*. *Sci Rep* **5**: 1–19.

König S, Romoth L, Gerischer L, Stanke M. 2016. Simultaneous gene finding in multiple genomes. 2016. *BMC Bioinformatics* **32**(22): 3388-3395

Koeniger N. 1976. Neue Aspekte der Phylogenie innerhalb der Gattung *Apis*. *Apidologie*: **7**: 357-366.

Koeniger N, Koeniger G, Smith D. 2011. Phylogeny of the Genus Apis. In: Honeybees of Asia. Eds Hepburn R and Radloff S. Springer, Heidelberg.

Koeniger N, Koeniger G, Tingek S. 2010. *Honey bees of Borneo: Exploring the Centre of Apis diversity*. Natural History Publications (Borneo), Kota Kinabalu.

Kosakovsky Pond SL, Poon AFY, Velazquez R, Weaver S, Hepler NL, Murrell B, Shank SD, Magalis BR, Bouvier D, Nekrutenko A, et al. 2019. HyPhy 2.5—A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies. *Mol Biol Evol* 1–5.

Kotthoff U, Wappler T, Engel MS. 2013. Greater past disparity and diversity hints at ancient migrations of European honey bee lineages into Africa and Asia. *J Biogeogr* **40**: 1832–1838.

Lee KS, Kwon OY, Lee JH, Kwon K, Min KJ, Jung SA, Kim AK, You KH, Tatar M, Yu K. 2008. *Drosophila* short neuropeptide F signalling regulates growth by ERK-mediated insulin signalling. *Nat. Cell Biol.* **10**: 468--475.

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, et al. 2010. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* **20**: 265–272.

Li Z, Liu F, Li W, Zhang S, Niu D, Xu H, Hong Q, Chen S, Su S. 2012. Differential transcriptome profiles of heads from foragers: Comparison between *Apis mellifera ligustica* and *Apis cerana cerana*. *Apidologie* **43**: 487–500.

Li X-D, Jiang G-F, Yan, L-Y, Li R, Mu Y, Deng W-A. 2018. Positive selection drove the adaptation of mitochondrial genes to the demands of flight and high-altitude environments in grasshoppers. *Frontiers in Genetics* **9**: 1-12.

Li Y, Zhang R, Liu, S, Donath A, Peters RS, Ware J, Misof B, Niehuis O, Pfrender ME, Zhou X. 2017. The molecular evolutionary dynamics of oxidative phosphorylation (OXPHOS) genes in Hymenoptera. *BMC Evol Biol* **17**: 269.

Lim AK, Kai T. 2007. Unique germ-line organelle, nuage, functions to repress selfish genetic elements in *Drosophila melanogaster*. *Proc Natl Acad Sci U S A* **104**: 20143–20143.

Löytynoja A, Goldman N. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* **320**: 1632–1635.

Ma Z, Guo W, Guo X, Wang X, Kang L. 2011. Modulation of behavioral phase changes of the migratory locust by the catecholamine metabolic pathway. *Proc Natl Acad Sci U S A* **108**: 3882–3887.

Malka O, Niño EL, Grozinger CM, Hefetz A. 2014. Genomic analysis of the interactions between social environment and social communication systems in honey bees (*Apis mellifera*). *Insect Biochem Mol Biol* **47**: 36–45.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376–380.

Mattila HR, Seeley TD. 2007. Genetic diversity in honey bee colonies enhances productivity and fitness. *Science* **317**: 362–364.

Maynard Smith J, Szathmáry E. 1995. *The Major Transitions in Evolution*. Oxford University Press. Oxford, UK.

Mayoral JG, Nouzova M, Navare A, Noriega FG. 2009. NADP-dependent farnesol dehydrogenase, a corpora allata enzyme involved in juvenile hormone synthesis. *Proc Natl Acad Sci U S A* **106**: 21091-21096.

McBride CS, Baier F, Omondi AB, Spitzer SA, Lutomiah J, Sang R, Ignell R, Vosshall LB. 2014. Evolution of mosquito preference for humans linked to an odorant receptor. *Nature* **515**: 222–227.

McKenzie SK, Fetter-Pruneda I, Ruta V, Kronauer DJCC. 2016. Transcriptomics and neuroanatomy of the clonal raider ant implicate an expanded clade of odorant receptors in chemical communication. *Proc Natl Acad Sci U S A* **113**: 14091–14096.

Meznar ER, Gadau J, Koeniger N, Rueppell O. 2010. Comparative linkage mapping suggests a high recombination rate in all honeybees. *J Hered* **101**: 118–126.

Nachtweide S, Stanke M. 2019. Multi-genome annotation with AUGUSTUS. *Methods Mol Biol* **1962**: 139-160.

Neiman M, Taylor DR. 2009. The causes of mutation accumulation in mitochondrial genomes. *Proc R Soc B* **276**: 1201‐1209.

Nevado B, Atchison GW, Hughes CE, Filatov DA. 2016. Widespread adaptive evolution during repeated evolutionary radiations in New World lupins. *Nat Comm* **7**: 12384.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al. 2016. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**: D733–D745.

Oddie M, Büchler R, Dahle B, Kovacic M, Le Conte Y, Locke B, De Miranda JR, Mondet F, Neumann P. 2018. Rapid parallel evolution overcomes global honey bee parasite. *Sci Rep* **8**: 1–9.

Odronitz F, Kollmar M. 2008. Comparative genomic analysis of the arthropod muscle myosin heavy chain genes allows ancestral gene reconstruction and reveals a new type of "partially" processed pseudogene. *BMC Mol Biol* **9**: 1–17.

Oldroyd BP, Wongsiri S. 2006. *Asian Honey Bees: Biology, Conservation, and Human Interactions*. Harvard Univerity Press.

Oppenheim S, Cao X., Rueppell O., Chantawannakul P., Krongdang S., Phokasem P., DeSalle R., Goodwin S., Xing J., Rosenfeld O. (2020) Whole genome sequencing and assembly of the Asian Honey Bee *Apis dorsata*. *Genome Biology and Evolution* **12**: 3677-3683.

Park D, Jung WW, Choi BS, Jayakodi M, Lee J, Lim J, Yu Y, Choi YS, Lee ML, Park Y, et al. 2015. Uncovering the novel characteristics of Asian honey bee, *Apis cerana*, by whole genome sequencing. *BMC Genomics* **16**: 1–16.

Pask GM, Slone JD, Millar JG, Das P, Moreira JA, Zhou X, Bello J, Berger SL, Bonasio R, Desplan C, et al. 2017. Specialized odorant receptors in social insects that detect cuticular hydrocarbon cues and candidate pheromones. *Nat Commun* **8**: 1–10.

Pearson WR, Lipman DJ. 1988. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A* **85**: 2444–2448.

Penn O, Privman E, Landan G, Graur D, Pupko T. 2010. An alignment confidence score capturing robustness to guide tree uncertainty. *Mol Biol Evol* **27**: 1759–1767.

Petersen M, Armisén D, Gibbs RA, Hering L, Khila A, Mayer G, Richards S, Niehuis O, Misof B. 2019. Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects. *BMC Evol Biol* **19**(1): 11.

Phillippy AM. 2017. New advances in sequence assembly. *Genome Res* **27**: xi–xiii.

Raffiudin R, Crozier RH. 2007. Phylogenetic analysis of honey bee behavioral evolution. *Mol Phylogenet Evol* **43**: 543–552.

Robertson HM, Wanner KW. 2006. The chemoreceptor superfamily in the honey bee, *Apis mellifera*: Expansion of the odorant, but not gustatory, receptor family. *Genome Res* **16**: 1395–1403.

Rokas A. 2011. Phylogenetic analysis of protein sequence data using the Randomized Axelerated Maximum Likelihood (RAXML) Program. *Curr Protoc Mol Biol* 96:19.11.1-19.11.14

Ross CR, DeFelice DS, Hunt GJ, Ihle KE, Amdam GV, Rueppell O. 2015. Genomic correlates of recombination rate and its variability across eight recombination maps in the western honey bee (*Apis mellifera* L.). *BMC Genomics* **16**:107.

Rothkamm K, Löbrich M. 2003. Evidence for a lack of DNA double-strand break repair in human cells exposed to very low x-ray doses. *Proc Natl Acad Sci U S A* **100**: 5057–5062.

Roux J, Privman E, Moretti S, Daub JT, Robinson-Rechavi M, Keller L. 2014. Patterns of positive selection in seven ant genomes. *Mol Biol Evol* **31**: 1661–1685.

Rueppell O. 2009. Characterization of quantitative trait loci for the age of first foraging in honey bee workers. *Behav Genet* **39**: 541–553.

Rueppell O, Kuster R, Miller K, Fouks B, Correa SR, Collazo J, Phaincharoen M, Tingek S, Koeniger N. 2016. A new metazoan recombination rate record and consistently high recombination rates in the honey bee genus *Apis* accompanied by frequent inversions but not translocations. *Genome Biol Evol* **8**: 3653–3660.

Rueppell O, Metheny JD, Linksvayer T, Fondrk MK, Page RE, Amdam G V. 2011a. Genetic architecture of ovary size and asymmetry in European honeybee workers. *Heredity* **106**: 894–903.

Rueppell O, Phaincharoen M, Kuster R, Tingek S. 2011b. Cross-species correlation between queen mating numbers and worker ovary sizes suggests kin conflict may influence ovary size evolution in honeybees. *Naturwissenschaften* **98**: 795–799.

Sadd BM, Barribeau SM, Bloch G, de Graaf DC, Dearden P, Elsik CG, Gadau J, Grimmelikhuijzen CJP, Hasselmann M, Lozier JD, et al. 2015. The genomes of two key bumblebee species with primitive eusocial organization. *Genome Biol* **16**: 76.

Sánchez-Gracia A, Vieira FG, Rozas J. 2009. Molecular evolution of the major chemosensory gene families in insects. *Heredity* **103**: 208–216.

Sanchez CJ, Shivshankar P, Stol K, Trakhtenbroit S, Sullam PM, Sauer K, Hermans PWM, Orihuela CJ. 2010. The pneumococcal serine-rich repeat protein is an intraspecies bacterial adhesin that promotes bacterial aggregation in vivo and in biofilms. *PLoS Pathog* **6**: 33–34.

Santos CG, Hartfelder K. 2015. Insights into the dynamics of hind leg development in honey bee (*Apis mellifera* L.) queen and worker larvae - A morphology/differential gene expression analysis. *Genet Mol Biol* **38**: 263–277.

Sarma M Sen, Rodriguez-Zas SL, Hong F, Zhong S, Robinson GE. 2009. Transcriptomic profiling of central nervous system regions in three species of honey bee during dance communication behavior. *PLoS One* **4**: e6408.

Sarma M Sen, Whitfield CW, Robinson GE. 2007. Species differences in brain gene expression profiles associated with adult behavioral maturation in honey bees. *BMC Genomics* **8**: 202–216.

Schneider A. 2011. Mitochondrial tRNA import and its consequences for mitochondrial translation. *Annu Rev Biochem* **80**: 1033–1053.

Seeley TD, Tarpy DR. 2007. Queen promiscuity lowers disease within honeybee colonies. *Proc R Soc B Biol Sci* **274**: 67–72.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, *Bioinformatics* **31:**3210–3212,

Simola DF, Wissler L, Donahue G, Waterhouse RM, Helmkampf M, Roux J, Nygaard S, Glastad KM, Hagen DE, Viljakainen L, et al. 2013. Social insect genomes exhibit dramatic evolution in gene composition and regulation while preserving regulatory features linked to sociality. *Genome Res* **23**: 1235–1247.

Slater GSC, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**: 1–11.

Slone JD, Pask GM, Ferguson ST, Millar JG, Berger SL, Reinberg D, Liebig J, Ray A, Zwiebel LJ. 2017. Functional characterization of odorant receptors in the ponerine ant, *Harpegnathos saltator*. *Proc Natl Acad Sci U S A* **114**: 8586–8591.

Smith CD, Zimin A, Holt C, Abouheif E, Benton R, Cash E, Croset V, Currie CR, Elhaik E, Elsik CG, et al. 2011. Draft genome of the globally widespread and invasive Argentine ant (*Linepithema humile*). *Proc Natl Acad Sci U S A* **108**: 5673–5678.

Smith CR, Toth AL, Suarez A V., Robinson GE. 2008. Genetic and genomic analyses of the division of labour in insect societies. *Nat Rev Genet* **9**: 735–748.

Smith MD, Wertheim JO, Weaver S, Murrell B, Scheffler K, Kosakovsky Pond SL. 2015. Less is more: An adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol Biol Evol* **32**: 1342–1353.

Sogl B, Gellissen G, Wiesner RJ. 2000. Biogenesis of giant mitochondria during insect flight muscle development in the locust, *Locusta migratoria* (L.): Transcription, translation and copy number of mitochondrial DNA. *European Journal of Biochemistry* **267**: 11-17.

Stamatakis A. 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**: 2688–2690.

Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**: 637–644.

Stocker S, Hiery M, Marriott G. 1999. Phototactic migration of *Dictyostelium* cells is linked to a new type of gelsolin-related protein. *Mol Biol Cell* **10**: 161–178.

Sun C, Huang J, Wang Y, Zhao X, Su L, Thomas WC, Zhao M, Zhang X, Jungreis I, Kellis M, et al. 2021. Genus-wide characterization of bumblebee genomes reveals variation associated with key ecological and behavioral traits of pollinators. *Mol Biol Evol* **38**: 486–501.

Terrapon N, Li C, Robertson HM, Ji L, Meng X, Booth W, Chen Z, Childers CP, Glastad KM, Gokhale K, et al. 2014. Molecular traces of alternative social organization in a termite genome. *Nat Commun* **5**: 3636.

Thomas JA, Welch JJ, Lanfear R, Bromham L. 2010. A generation time effect on the rate of molecular evolution in invertebrates. *Mol Biol Evol* **27**: 1173–1180.

Tollis M, Hutchins ED, Stapley J, Rupp SM, Eckalbar WL, Maayan I, Lasku E, Infante CR, Dennis SR, Robertson JA, et al. 2018. Comparative genomics reveals accelerated evolution in conserved pathways during the diversification of anole lizards. *Genome Biol Evol* **10**: 489-506.

Vianna JA, Fernandes FAN, Frugone MJ, Figuero HV, Pertierra LR, Noll D, Bi K, Wang-Claypool CY, Lowther A, et al. 2020. Genome-wide analyses reveal drivers of penguin diversification. *Proc Natl Acad Sci U S A* **117**: 22303-22310.

Vidwans SJ, Su TT. 2001. Cycling through development in *Drosophila* and other metazoa. *Nature Cell Biology* **3**: E35-39.

Wallberg A, Han F, Wellhagen G, Dahle B, Kawata M, Haddad N, Simões ZLP, Allsopp MH, Kandemir I, De La Rúa P, et al. 2014. A worldwide survey of genome sequence variation provides insight into the evolutionary history of the honeybee *Apis mellifera*. *Nat Genet* **46**: 1081–1088.

Wallberg A, Bunikis I, Pettersson OV, Childers AK, Evans JD, Mikheyev AS, Robertson HM, Robinson GE, Webster MT. 2019. A hybrid de novo genome assembly of the honeybee, *Apis mellifera*, with chromosome-length scaffolds. *BMC Genomics* **20**: 275.

Wang W, Ashby R, Ying H, Maleszka R, Forêt S. 2017. Contrasting sex-and caste-dependent piRNA Profiles in the transposon depleted haplodiploid honeybee *Apis mellifera*. *Genome Biol Evol* **9**: 1341–1356.

Warner MR, Qiu L, Holmes MJ, Mikheyev AS, Linksvayer TA. 2019. Convergent eusocial evolution is based on a shared reproductive groundplan plus lineage-specific plastic genes. *Nat Commun* **10**: 1–11.

Warrant E, Porombka T, Kirchner WH. 1996. Neural image enhancement allows honeybees to see at night. *Proc R Soc B Biol Sci* **263**: 1521–1526.

Weick E-M, Miska EA. 2014. piRNAs: from biogenesis to function. *Development* **141**: 3458–3471.

Weiner J. 2015. tagcloud. R package version, 0.6. https://cran.r-project.org/web/packages/tagcloud/index.html

Weinstock GM, Robinson GE, Gibbs RA, Weinstock GM, Weinstock GM, Robinson GE, Worley KC, Evans JD, Maleszka R, Robertson HM, et al. 2006. Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature* **443**: 931–949.

West-Eberhard MJ. 2003. *Developmental Plasticity and Evolution*. Oxford University Press, Oxford, UK.

Whitfield CW, Ben-Shahar Y, Brillet C, Leoncini I, Crauser D, LeConte Y, Rodriguez-Zas S, Robinson GE. 2006. Genomic dissection of behavioral maturation in the honey bee. *Proc Natl Acad Sci U S A* **103**: 16068–16075.

Wigglesworth VB. 1932. Memoirs: On the Function of the so-called “Rectal Glands” of Insects. *Q J Microsc Sci* **s2-75**: 131–150.

Wilfert L, Gadau J, Schmid-Hempel P. 2007. Variation in genomic recombination rates among animal taxa and the case of social insects. *Heredity* **98**: 189–197.

Woodard C, Alcorta E, Carlson J. 2007. The rdgB gene of *Drosophila*: A link between vision and olfaction. *J Neurogenet* **21**: 291–305.

Woodard SH, Fischman BJ, Venkat A, Hudson ME, Varala K, Cameron SA, Clark AG, Robinson GE. 2011. Genes involved in convergent evolution of eusociality in bees. *Proc Natl Acad Sci U S A* **108**: 7472–7477.

Wright GA, Mustard JA, Simcock NK, Ross-Taylor AAR, McNicholas LD, Popescu A, Marion-Poll F. 2010. Parallel reinforcement pathways for conditioned food aversions in the honeybee. *Curr Biol* **20**: 2234–2240.

Xie Y-F, Shang F, Ding B-Y, Wu Y-B, Niu J-Z, Wei D, Dou W, Christiaens O, Smagghe G, Wang J-J. 2019. Tudor knockdown disrupts ovary development in *Bactrocera dorsalis*. *Insect Molecular Biology* **28**: 136-144.

Yang Y, Ballinger D. 1994. Mutations in calphotin, the gene encoding a *Drosophila* photoreceptor cell-specific calcium-binding protein, reveal roles in cellular morphogenesis and survival. *Genetics* **138**: 413–421.

Yarian C, Townsend H, Czestkowski W, Sochacka E, Malkiewicz AJ, Guenther R, Miskiewicz A, Agris PF. 2002. Accurate translation of the genetic code depends on tRNA modified nucleosides. *J Biol Chem* **277**: 16391–16395.

Yu G. 2020. Using ggtree to visualize data on tree-like structures. *Curr Protoc Bioinfo*, **69:** e96.

Zayed A, Robinson GE. 2012. Understanding the relationship between brain gene expression and social behavior: lessons from the honey bee. *Annu Rev Genet* **46**: 591–615.

Zdobnov EM, Apweiler R. 2001. InterProScan - An integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**: 847–848.

Zhou S, Morgante F, Geisz MS, Ma J, Anholt RRH, Mackay TFC. 2020. Systems genetics of the *Drosophila* metabolome. *Genome Res* **30**: 392–405.

The genomic basis of evolutionary differentiation among honey bees

Bertrand Fouks, Philipp Brand, Hung N Nguyen, et al.

Genome Res. published online May 4, 2021
Access the most recent version at doi:10.1101/gr.272310.120