The birth-and-death evolution of cytochrome P450 genes in bees

Kathy Darragh¹, David R. Nelson² & Santiago R. Ramírez¹

¹Department of Evolution and Ecology, University of California, Davis, CA, 95616

²Department of Molecular Sciences, University of Tennessee, Memphis, TN, 38163

ORCID Kathy Darragh 0000-0002-9215-014X

ORCID David R. Nelson 0000-0003-0583-5421

ORCID Santiago R. Ramírez 0000-0003-1306-1315

Key words: cytochrome P450, birth-and-death evolution, gene family evolution, orchid bees.

Abstract

The birth-and-death model of multigene family evolution describes how families can expand by duplication and contract by gene deletion and formation of pseudogenes. The phylogenetic stability of a gene is thought to be related to the degree of functional importance. However, it is unclear how much evolution of a gene in a gene family is driven by adaptive versus neutral processes. The cytochrome P450s are one of the most diverse and well-studied multigene families, involved in both physiological and xenobiotic functions. Bees have a high toxin exposure due to their diet of nectar and pollen, as well as the resin-collecting behavior exhibited by some bees. Here, we describe the P450s of the orchid bee *Euglossa dilemma*. Orchid bees are a neotropical clade in which males form perfume bouquets used in courtship displays by collecting a diverse set of volatile compounds, resulting in high chemical compound exposure. We conducted phylogenetic and selection analyses across ten bee species encompassing three bee families. We do not find a relationship between the ecology of a bee species and its P450 repertoire. Our analyses reveal that P450 clades can be classified into stable and unstable clades, and that genes involved in xenobiotic metabolism are more likely to belong to unstable clades. Furthermore, we find that unstable clades are under more dynamic evolutionary pressures, with signals of adaptive evolution detected, suggesting that both gene duplication and positive selection driving sequence divergence have played a role in the diversification of bee P450s. Our work highlights the complexity of multigene family evolution which does not always follow generalized predictions.

37 Significance statement

38 Gene family evolution is characterized by deletion and duplication, but it is unclear if this is
 39 driven by adaptive or neutral evolutionary processes. To investigate the dynamics of gene
 40 family evolution we analyze the P450s of ten bee species, a well-studied gene family
 41 involved in both physiological and detoxification roles. We find that genes involved in
 42 detoxification are more likely to belong to unstable clades, exhibit dynamic evolutionary
 43 pressures, and show signals of adaptive evolution. However, we do not find evidence for a
 44 relationship between resin collection or sociality, and the P450 repertoire of a bee species.
 45 Our findings highlight the complexity of multigene family evolution which does not always
 46 follow generalized predictions.

47

Introduction

Multigene families arise from the duplication of a common ancestral gene, resulting in groups of genes that share similar sequences, and often functions (Nei & Rooney 2005). To understand how these families evolve, early studies focused on concerted evolution in rRNA, whereby members of a gene family evolve together due to repeated unequal crossover events (Brown et al. 1972). More recently, an alternative model, the birth-and-death model, has been used to explain gene family evolution. In this model, genes evolve independently and expansion and contraction occurs through gene duplication, formation of pseudogenes, and gene deletion (Nei & Rooney 2005). The relative roles of neutral versus adaptive evolutionary forces within the framework of the birth-and-death model has proven more complex (Eirín-López et al. 2012).

Cytochrome P450s are one of the best-studied and most diverse multigene families (Nelson et al. 2013). They are enzymes which use molecular oxygen to change the structure of their substrates, a reaction which has an important role in both physiological endogenous processes, and also ecological and xenobiotic processes (Nelson et al. 2013). P450s are highly diverse and groups have been named based on sequence similarity, with four prominent groups described in insects (Nelson 1998; Tijet et al. 2001; Dermauw et al. 2020). Of these, the CYP3 group is the largest and most dynamic, exhibiting many lineage-specific duplications in insects, often linked to insecticide resistance and xenobiotic metabolism (Feyereisen 2006).

There appears to be a link between the function of P450s and their evolutionary dynamics. In vertebrates, genes involved in xenobiotic detoxification are more likely to be evolutionarily unstable, exhibiting duplications and deletions between species (Thomas 2007). In contrast, those involved in viability are more likely to be stable with a one-to-one

orthology between species (Thomas 2007). This pattern has also been found in *Drosophila* where P450s associated with a role in development are duplicated less than those implicated in detoxification or have unknown functions (*Drosophila* 12 Genomes Consortium 2007).

Whilst the link between evolutionary instability in P450s and xenobiotic function has been shown, this does not test whether selection is acting on these P450s. The expansion of P450 subfamilies involved in xenobiotic functions is often assumed to be due to environmental adaptation driven by natural selection. However, even if an initial duplication is selected for, further duplication which lead to large subfamilies could simply be due to the self-sustaining process whereby a duplicated gene has twice the likelihood of duplicating again (Feyereisen 2011). In contrast to this neutral viewpoint, at a molecular level, correlation between amino acid replacement and the number of duplications found for a P450 lineage in the *Drosophila* phylogeny is inconsistent with stochastic models (Good et al. 2014). Furthermore, CYP expansions have been linked to environmental factors, such as specialized diets in *Lepidoptera* (Calla et al. 2017), implicating adaptive evolutionary forces in P450 expansions. However, the observed pattern of many groups with few genes and few groups with many genes (power-law distribution) does not require an adaptive explanation (Dermauw et al. 2020). Birth-and-death models of gene family evolution are sufficient to explain the pattern, not requiring any further explanation based on the ecology or life-history of the species (Sezutsu et al. 2013).

An interesting group for the study of P450s and detoxification is bees. Whilst providing important ecosystem services as pollinators, bees are exposed to a wide range of toxins, both natural and synthetic (Johnson 2015). In particular, bees differ from most pollinators in the fact that they are specialized on consuming pollen and nectar during all life-stages, diet sources with high levels of potentially toxic flavonoids, and have been called “flavonoid specialists” due to their consistent exposure to these compounds (Johnson et al.

2018). Exposure to flavonoids is thought to be higher in perennial eusocial bees due to the concentration of flavonoids when nectar is converted into honey and pollen into beebread for storage (Johnson et al. 2018). Furthermore, resin collection, associated with the evolution of sociality but also found in some solitary species, also increases flavonoid exposure (Bankova et al. 1983).

Understanding the molecular basis of toxin sensitivity is important to protect bees, especially from the negative effects of pesticides (Berenbaum & Johnson 2015). The sequencing of the honeybee genome (*Apis mellifera*) revealed a lower diversity of detoxification genes than expected when compared with other insect species (Claudianos et al. 2006; The Honeybee Genome Sequencing Consortium 2006). While the beetle *Tribolium castaneum* and the mosquito *Anopheles gambiae* have over 100 P450s, the honey bee *Apis mellifera* has less than 50 (Ranson et al. 2002; Claudianos et al. 2006; Oakeshott et al. 2010; Zhu et al. 2013). This deficit of P450s is not only exclusive to *Apis mellifera*, but is also the case in other bee species, including the solitary leafcutter bees *Megachile rotundata* and *Osmia bicornis* (Beadle et al. 2019; Hayward et al. 2019). Furthermore, this reduction is not equal across all P450s. Mitochondrial and CYP2 P450s have mostly been maintained, with the genes present from these groups linked to physiological roles in other insects such as hormone biosynthesis (Claudianos et al. 2006). In contrast, the CYP4s are greatly reduced, especially those members thought to play a role in xenobiotic detoxification. The CYP3s are also greatly reduced in diversity, but widespread duplication in the CYP6AS family within the CYP3s means that these make up a large proportion of total bee P450S (Claudianos et al. 2006; Berenbaum & Johnson 2015). This may explain why, in spite of these reductions, honey bees are not more sensitive to insecticides than other insects (Hardstone & Scott 2010).

One particular group of bees, the orchid bees, exhibit unique natural history and adaptation that presents additional challenges for detoxification. Male bees collect chemical

compounds from different sources, such as orchid flowers and fungi, which they store in specialized leg pouches to concoct a ‘perfume’ bouquet (Dressler 1982). These bouquets are then released during courtship displays and are thought to play a key role in mate choice (Eltz et al. 2005; Pokorny et al. 2017). The chemical bouquets differ between species, but also between individuals of the same species and may provide information to the female regarding mate quality (Eltz et al. 1999; Zimmermann et al. 2009; Weber et al. 2016). Male orchid bees collect a wide range of chemical compounds, many of which are expected to be toxic to bees. Collection of these compounds, therefore, could act as an indicator of male quality, through their ability to handle and detoxify these compounds (Eltz et al. 1999; Arriaga-Osnaya et al. 2017). We may expect adaptation at the molecular level, perhaps in the form of an increased P450 repertoire, or increased expression levels, to be able to detoxify such a wide range of compounds.

Here, we study the molecular evolution of cytochrome P450s in bees and evaluate the hypothesis that orchid bees have an expanded P450 repertoire. We annotate P450s in the genome of *Euglossa dilemma* and combine these newly annotated P450s with those previously identified in nine other bee species: *Apis mellifera*, *Bombus terrestris*, *Dufourea novaeangliae*, *Eufriesea mexicana*, *Habropoda laboriosa*, *Lasioglossum albipes*, *Megachile rotundata*, *Melipona quadrifasciata*, and *Osmia bicornis*. These species are from three different families, and include *Eufresia mexicana*, another orchid bee. To determine the evolutionary history of these P450s we carry out phylogenetic analyses and classify clades as stable or unstable based on their history across these species. We measure evolutionary change and search for signals of adaptive evolution in these clades to identify the selection pressures acting on P450s in these bee species. We also investigate the patterns of correlation between the level of sociality of a bee species, whether or not it collects resin, and its P450

146 inventory. We aim to shed light both on the evolution of the P450 family and orchid bee
147 biology in relation to detoxification.

148

Results

Phylogenetic analysis of bee P450s

To compile a dataset of bee P450s we firstly combined previously annotated P450s for nine species: *Apis mellifera*, *Bombus terrestris*, *Dufourea novaeangliae*, *Eufriesea mexicana*, *Habropoda laboriosa*, *Lasioglossum albipes*, *Megachile rotundata*, *Melipona quadrifasciata*, and *Osmia bicornis* (Hayward et al. 2019; Beadle et al. 2019; Kapheim et al. 2015; Johnson et al. 2018) (Figure 1). To this set of P450s we then added 41 P450s which we identified and annotated in *Euglossa dilemma* with complete protein domains. We also identified 4 P450s with incomplete protein domains which may represent pseudogenes or may be due to poor assembly, sequencing errors, or sequencing gaps. We do not include these for the following analyses (incomplete_genes.fa).

To determine orthology of P450s we constructed a phylogeny combining the *Eg. dilemma* P450s from those previously identified in nine other bee species. The P450s group into the four expected clusters in insects, CYP2, CYP3, CYP4 and mitochondrial (Figure 2, Supplementary Figure 1). As expected for bees, the CYP4s are reduced across all species, and the CYP6AS family is expanded. In some clades the phylogeny of P450s does not reflect the species' phylogeny. This is likely due to errors in phylogenetic reconstruction in these particular P450 clades rather than complicated patterns of duplicated and deletion between species. No members of the CYP6AS clade containing CYP6AS10 from *B. terrestris* and *A. mellifera* are found in either *Eg. dilemma* or *Ef. mexicana*, suggesting this is a loss in orchid bees. Overall, *Eg. dilemma* has a comparable number and distribution of P450s as other previously studied bee species (Figure 1, Supplementary Figure 1).

Figure 1: Phylogeny of cytochrome P450s in bees. The phylogeny was constructed in IQ-TREE (model GTR+F+R4) using five nuclear genes at 10508 nucleotide sites. Bootstrap values ($n=1000$) are illustrated. For each species its resin collecting behavior, level of sociality, and number of P450s in each family is shown. *Melipona quadrifasciata* photo courtesy of Lucas Rubio, *Osmia bicornis* photo courtesy of Amelia Bassiti.

Is the number of P450s in the genome associated with species' ecology?

We found that total number of P450s in the genome of each species ranged from 40 in *H. laboriosa* to 59 in *D. novaeangliae*, with most of this variation from the number of CYP3s found in each species (Figure 1). We investigated two aspects of species' ecology: resin collection and level of sociality. We tested for relationships between P450 number and ecology while accounting for phylogeny. We did not find a relationship between sociality and the number of P450s (Supplementary Figure 2A), the number of CYP3s (Supplementary Figure 2B), or the number of CYP6AS genes (Supplementary Figure 2C). Furthermore, we did not find a difference between bees that exhibit substantial resin collection and those which do not in the number of P450s (Supplementary Figure 3A), the number of CYP3s (Supplementary Figure 3B), or the number of CYP6AS genes (Supplementary Figure 3C).

Evolutionary history of P450s in bees

To investigate the dynamics of gene family evolution and phylogenetic instability there are two main approaches: the reconciliation of the gene tree with the species tree to detect duplication and loss events, and the use of birth-death models fitted to a species tree (Yohe et al. 2019).

We used MiPhy to implements the first approach of species tree gene tree reconciliation to identify clades and estimate phylogenetic instability scores for each clade. Without the “merge singletons” option, MiPhy divided our data into 42 clades including two clades which only contained a sequence from *Habropoda laboriosa*. For the final analysis, we used the option “merge singletons” to divide our data into 40 clades ranging in their instability score from -0.1 to 25.05 (Supplementary Figure 4). The instability scores of the clades fell into two broad groups. Clades with low instability scores (-0.1 to 3.61) were classified as stable, and those with higher instability scores (9.12-25.05) were classified as

unstable. In total, 27.5% (11/40) clades were classified as stable. The distribution of stable and unstable clades was not equal amongst P450 families, with 52% of CYP3 clades found to be unstable, in comparison with only 0% of CYP2, CYP4 and mitochondrial P450 clades (Figure 2). CYP3s have a higher instability score than the other CYP groups (Supplementary Figure 5).

To apply the second approach to our dataset we use Computational Analysis of gene Family Evolution (CAFÉ), a program which uses birth-death to model gene gain and loss across a species tree, using the clades previously identified by the MiPhy analysis (Mendes et al. 2020). Clades which are fast evolving are identified by comparing models in which all clades evolve at the same rate with models in which different clades vary in their evolutionary rate. Using this approach, we identified one CYP6AS clade, CYP6AS1, as an outlier when compared to all other families, with respect to its evolutionary rate of family size change ($p=0.029$). This was the same clade which had the highest instability in the MiPhy analysis (Supplementary Figure 4). As expected, this family also had the strongest assignment to the gamma category of the highest median lambda (rate of evolutionary change), (posterior probability=1.0). Furthermore, whilst not detected as outliers relative to the entire dataset, seven additional clades were also inferred to be rapidly evolving in size, as indicated by their significant placement in the same fast evolving gamma category (Supplementary Table 1; posterior probability > 0.95). All of these were classed as unstable by the MiPhy analysis. Whilst all of the clades identified by CAFÉ were identified as unstable in the MiPhy analysis, this was not true in reverse, probably as CAFÉ uses gene counts and does not consider the gene tree, making it unable to distinguish between independent gene duplication events and inheritance of paralogs (Curran et al. 2018).

The exact pattern of duplications and deletions is unclear, but certainly some have happened on a more recent timescale. For example, CYP9DN1 is not found in *A. mellifera*,

228 *B. terrestris*, *D. novaeangliae*, or *Ef. mexicana*, suggesting this deletion has occurred at least
 229 twice as *D. novaeangliae* is not closely related to the other species. Of the 11 unstable clades,
 230 three belong to this category of requiring at least two deletion events to explain the
 231 phylogenetic pattern. Other clades could be explained by one deeper internal deletion event,
 232 but more species would be needed to determine this. For example, the clade containing *Apis*
 233 *mellifera* CYP6AS10 is missing in both *Ef. mexicana* and *Eg. dilemma*, which could be
 234 explained by a deletion in all orchid bees, or two separate deletion events, with further
 235 sampling necessary to distinguish the two. Duplication events are also a mix of lineage-
 236 specific and potentially deeper internal duplications. The genes CYP6AS131-135 and
 237 CYP6AS91-95 are expanded specifically in *O. bicornis* and *D. novaeangliae*, respectively,
 238 suggesting more recent duplication events. In contrast, CYP336 shows a pattern consistent
 239 with duplications deeper in the lineage leading to *O. bicornis* and *M. rotundata*
 240 (Supplementary figure 1).

Figure 2: Phylogeny of cytochrome P450s in bees. The phylogeny was constructed in IQ-TREE (model JTT+F+R10) using 481 amino acid sequences across ten bee species belonging to three different families. Clades are collapsed to show 40 clades. Bootstrap values (n=1000) are illustrated.

Do unstable clades of P450s exhibit increased evolutionary change?

The clades were found to differ in their phylogenetic instability phylogenetic instability and rate of size evolution. To further investigate the evolutionary dynamics of these clades we compared various evolutionary measures. Branch length can be used as a measure of evolutionary change in a phylogenetic tree. We found a correlation between instability and branch length, when gene number is corrected for, suggesting higher rates of evolution than for genes found in stable clades (Fig 3). Unstable clades also have a higher cumulative patristic distance, which includes the length of internal branches also (Supplementary Figure 6). One clade, CYP9FX, did not follow this trend, with a low instability score but high CBL and cumulative patristic distance. This is the smallest clade in our analyses with only three genes from *Dufourea novaeangliae* and *Lasioglossum albipes*.

Are there signals of adaptive evolution in bee P450s?

Branch-specific models

Although branch length can give an indication of the amount of divergence between two P450s, it does not provide information on the predominant type of selection that shaped their evolution. To explore both the amount and the type of evolutionary change happening in stable and unstable clades, we compared the rate of nonsynonymous to synonymous mutations (dN/dS) in different lineages. The simplest analysis carried out was a “one-ratio” model for each clade. In these models, each branch in a clade is assumed to have the same dN/dS value. All values obtained were below one when all members of a clade are forced to have the same dN/dS value, suggesting that in general purifying selection is acting. To test if unstable and stable clades are under different selection pressures, we compared their “one-ratio” dN/dS values. We found that all clades are under purifying selection with dN/dS values less than one. However, we detected a correlation between clade instability and dN/dS ratio,

with more unstable clades having a higher dN/dS ratio (Fig. 4A). Furthermore, we found a correlation between the cumulative branch length of a clade, the amount of evolutionary change, and the dN/dS value of the clade (Fig. 4B).

The “one-ratio” model is restrictive, only allowing a single dN/dS value per clade. To investigate which clades are under more dynamic selection pressures than the simple “one-ratio” model, we compared these models to models under which each lineage can vary in its dN/dS ratio, also called “free-ratio” models (Supplementary Table 2). As for the “one-ratio” models above, these “free-ratio” models were carried out by modelling each clade individually. The evolutionary history of unstable clades is more dynamic than stable clades. The more complex model has a better fit for more of the unstable clades (100%, 11/11) than the stable clades (69%, 20/29) (Fisher’s exact test, $p=0.0433$).

As well as testing selection pressures at a clade level we tested each branch in each clade individually for evidence of selection. We compared “one-ratio” models for each clade with models in which only one branch is allowed to vary in its dN/dS ratio (“two-ratio” models). In these models we can ask whether the branch of interest has a dN/dS value significantly different from the “background” dN/dS of the clade. Again, we found the same pattern in the resulting dN/dS values. Of all the branches with a dN/dS ratio significantly different from the background of the clade, dN/dS was higher for branches from unstable than stable clades (Supplementary Figure 7). Furthermore, all branches with a significant dN/dS greater than 1.1 (excluding branches where $dS < 0.01$ and $dS > 2$), indicating positive selection, were from the CYP3 clade. Of these, 70% (7/10) were from unstable CYP3 clades, and 70% (7/10) were from the CYP6AS family (Table 1). The three genes which were not found in unstable clades were all in the CYP6AS7 clade. Full list of tests can be found in Supplementary data 2.

Site-specific models

Branch-specific models assumes a consistent dN/dS ratio across all sites in a given gene. We also carried out site-specific tests which allow dN/dS to vary among sites. We found evidence of positively selected sites in a quarter (10/40) of all clades in the M7 and M8 model, including clades from CYP2, CYP4 and CYP3 (Supplementary Table 3). For four of these clades, CYP6AS1, CYP6AS7, CYP6AS8 and CYP336, positively selected branches were previously identified in the branch-specific test comparing “one-ratio” and “two-ratio” models. CYP6AS1 and CYP6336 are the two clades with highest instability as identified by MiPhy and CYP6AS1 was also identified as the sole outlier family in terms of evolutionary rate by CAFE. We identified positively selected sites at 90% posterior probability for eight of these clades: CYP306A1, CYP305D1 CYP4G202, CYP336, CYP6BD1, CYP6AS1, CYP6AS7, and CYP6AS8. We found no evidence of positively selected sites in the M1a-M2a comparison.

To determine where on the protein these sites were located, we took advantage of the AlphaFold2 algorithm (Jumper et al. 2021; Mirdita et al. 2021) to create protein structural models for a representative sequence in each clade. In general, the residues are found on surface regions of the proteins. The residues identified in CYP6AS1, CYP6AS7, CYP6AS8 and CYP6BD1 are all found in the same surface region of the protein (Supplementary Figure 8). The residues in CYP6AS1 and CYP6AS7 are both found between the F' and F helices, and those for CYP6AS8 and CYP6BD1 are both found between the G and G' helices. This region is not close to the active site, and may be involved in protein-protein interactions (Sevrioukova et al. 1999). The residues for the other four P450s are found in other regions of the structure (Supplementary Figure 9). The positively selected site identified in CYP306A1 is found at the C-terminal of the protein. Two residues were identified in CYP305D1, one at the end of the F helix and the other on a surface loop between the I and H helices. Two were also detected for CYP4G202, one in the turn between beta 1 and beta 2 near the N-terminal,

330 and the other on a large surface loop. Finally, the residue in CYP336 is located on a non-
331 helical, non-beta surface area of the protein. None of these residues are part of the active site
332 of the protein.

333

334

Table 1: Branches with dN/dS greater than one and significantly different from

background dN/dS of the clade in the “two-ratio” tests (adjusted p values <0.05). Branches

with $dS < 0.01$ and $dS > 2$ have been removed.

Name	dN/dS	CYP	CYP family	Genus
		clade		
GB40288	2.316	CYP3	CYP6AS1	<i>Apis mellifera</i>
LALB19295RA_B	1.822	CYP3	CYP6AS200	<i>Lasioglossum</i>
				<i>albipes</i>
MQUA24239RA	1.597	CYP3	CYP336A42	<i>Melipona</i>
				<i>quadrifasciata</i>
Internal branch	1.597	CYP3	CYP336	
MH500637	1.322	CYP3	CYP6AS127	<i>Osmia bicornis</i>
NW_015373885.1C	1.268	CYP3	CYP9FW6	<i>Dufourea</i>
				<i>novaeangliae</i>
Internal branch	1.243	CYP3	CYP6AS7	
MH500655	1.224	CYP3	CYP6AS131	<i>Osmia bicornis</i>
MH500653	1.206	CYP3	CYP6AS132	<i>Osmia bicornis</i>
XM_017902249.1	1.137	CYP3	CYP6AS87	<i>Eufriesea</i>
				<i>mexicana</i>

Figure 4. Relationship between the dN/dS of a clade, clade instability, and normalized CBL. A) Correlation between the instability of a clade and its dN/dS ratio in “one-ratio” models (Spearman’s Correlation, $\rho=0.567$, $p=0.00014$). B) Correlation between Normalized CBL (average branch length per gene in a clade) and dN/dS of the clade in a “one-ratio” model (Spearman’s Correlation, $\rho=0.754$, $p=1.889e-08$). CBL, cumulative branch length.

Discussion

Whilst the role of duplication and deletion in the evolution of multigene families is well-established, controversy remains surrounding whether this evolution is primarily adaptive or neutral. In this study, we demonstrate that genes known to be involved in xenobiotic metabolism (CYP3 family) are more likely to be unstable in the form of more gene duplication and deletion. Furthermore, we find that not only are these genes more likely to be unstable but also under more dynamic evolutionary pressures, and exhibit signals of adaptive evolution. This suggests that both gene duplication and positive selection driving sequence divergence contribute to the diversification of P450s. We do not find evidence for a correlation between the P450 repertoire of a bee species and its biology. Furthermore, our hypothesis that orchid bees would have an expansion of P450s due to their perfume collection behavior was also not supported.

The patterns we identify here are very similar to those previously identified in *Drosophila*. P450s with developmental functions are more likely to be found in stable clades, and also more likely to be under purifying selection (*Drosophila* 12 Genomes Consortium 2007; Good et al. 2014). These patterns, however, are not restricted to P450s, and have also been identified in the fatty acyl-coA synthase multigene family in *Drosophila*, which are involved in both essential physiological and chemical communication functions (Finet et al. 2019). While no single evolutionary model will be able to explain multigene family evolution, we can understand some common underlying principles using the framework of the birth-and-death model (Eirín-López et al. 2012).

Another common pattern found across multigene families is lineage-specific expansions, which are often associated with the evolution of novelty (Lespinet et al. 2002). In the CYP3 clade of P450s in bees, we find extensive evidence of duplications, particularly in

the CYP6AS subfamily. It has been suggested that this expansion is related to specialization on a plant-based diet and the ability to metabolize flavonoids, in comparison to solitary carnivorous ancestors such as solitary wasps (Johnson et al. 2018). As previously described by Johnson and colleagues (Johnson et al. 2018), we find that not only are CYP6AS subfamilies expanded by duplication, but also show evidence of positive selection, with the majority of positively selected genes detected belonging to this family. CYP6AS3 and CYP6AS4 from *Apis mellifera* were found to be under positive selection by Johnson and colleagues. Here we find CYP6AS1 to be under positive selection in the branch-specific models and evidence for positively selected sites in the clade containing CYP6AS1 and CYP6AS3 in the site-specific models. CYP6AS1, CYP6AS3, CYP6AS4, and CYP6AS10, have all been linked to quercetin metabolism, a plant flavonoid found in honey (Mao et al. 2009). Other CYP6AS genes, as well as CYP9R1, CYP9P1, CYP9S1, and CYP9Q family genes are upregulated in response to quercetin treatment in *Apis mellifera* (Mao et al. 2017). Furthermore, CYP9Q4 in *Bombus* and CYP9Q3 in *Apis* metabolize the neonicotinoid thiacloprid (Manjon et al. 2018). All of these genes known to be involved in xenobiotic metabolism were identified in our study as members of unstable clades, again linking instability and gene duplication with xenobiotic function.

In contrast to these genes with known xenobiotic function, we find that genes predicted to have physiological roles are mostly found in stable clades under purifying selection. The mitochondrial clade genes CYP302A1, -314A1, AND -315A1, which are orthologs of the *Drosophila melanogaster* Halloween genes, are thought to be involved in ecdysteroid synthesis and are all found in stable clades in our analysis (Gilbert 2004; Rewitz et al. 2007; Feyereisen 2011). From the CYP2 clade, CYP18A1, thought to be involved in ecdysteroid inactivation, CYP15A1, a predicted juvenile hormone epoxidase, and CYP306A1, a CYP2 member orthologous to a *D. melanogaster* ecdysteroid 25-hydroxylase,

are also found in stable clades (Helvig et al. 2004; Niwa et al. 2004; Claudianos et al. 2006). However, CYP306A1, despite being a stable clade is also one of the clades with evidence for a positively selected residue. Multiple other stable clades also showed evidence of positive selection in site-specific models (CYP305D1, CYP4G202, CYP6AS7 and CYP6BD1). Furthermore, the dichotomy of physiological and xenobiotic genes is not so simple, as both can be found in phylogenetic proximity (Dermauw et al. 2020). Furthermore, members of the same subfamily may have divergent roles. For example, while the CYP6AS subfamily is known to play a role in xenobiotic detoxification, CYP6AS8 and CYP6AS11 of *Apis mellifera* are expressed in mandibular glands, potentially involved in fatty acid signal synthesis (Wu et al. 2017; Dermauw et al. 2020). This highlights the complexity of multigene family evolution which does not always follow generalized predictions.

If P450 function is linked to their evolutionary dynamics, we might expect species with different biology to differ in their P450s (Rane et al. 2019). Specifically, we hypothesized that *Euglossa dilemma* would exhibit an expanded repertoire of P450s when compared with other bees, due to its perfume-collecting behavior increasing exposure to different chemical compounds. In contrast, we found that *Eg. Dilemma* has a comparable set of P450s similar to other bees, with no significant expansions detected. This does not rule out an increased capacity for xenobiotic metabolism, which would need to be tested experimentally. Specialization could also occur at the sequence level rather than the number of genes; however, we did not find evidence for positive selection in *Eg. Dilemma* P450s. One alternative is that *Eg. Dilemma* exhibits increased gene diversity at another step in the detoxification pathway. In honeybees, not only are P450s reduced in comparison to other insects, but also glutathione-S-transferases (GSTs) and carboxyl/cholinesterases (CCEs) (Claudianos et al. 2006; Berenbaum & Johnson 2015). Both GSTs and CCEs are known to be involved in xenobiotic metabolism in insects, thus meriting further investigation in *Eg.*

Dilemma (Li et al. 2007). Another possibility is that orchid bees do not demonstrate higher abilities to metabolize xenobiotic compounds, but instead reduce the amount of compound that can enter the body, for example by decreasing cuticle penetration, as reported in *Anopheles gambiae* (Balabanidou et al. 2016).

We also might expect a reduction in P450 repertoire in species which have a potentially lower xenobiotic exposure due to a specialist lifestyle (Rane et al. 2019). *Drosophila sechellia*, a specialist island species only found within a narrow ecological niche exhibits an increased loss of P450 genes compared to other *Drosophila* species (Good et al. 2014). In our dataset, the two most specialized species are *D. novaeangliae*, a specialist, with only one known pollen source (Eickwort et al. 1986), and *H. laboriosa*, with only a few known pollen sources (Pascarella 2007). Interestingly, *D. novaeangliae* has the largest P450 repertoire of all species in our analysis and *H. laboriosa* has the fewest P450s. We therefore do not find a strong link between specialization in diet and P450 repertoire in our bee dataset. More generally, we did not find a relationship between P450s and the biology of different bee species, neither at the total number level or subfamily level. This is in contrast to the previously identified trend that degree of sociality correlates with CYP6AS subfamily size, and that resin-collecting bees have more CYP6AS members (Johnson et al. 2018). This is likely because of the inclusion of *Osmia bicornis* in this study, a solitary bee which does not collect resin and yet has a large P450 repertoire, and the inclusion of phylogeny in these analyses. Future studies with more species, in particular species such as specialists, will allow us to untangle the relationship between species' biology and P450s.

The P450 family in bees can provide us with insights into the evolution of multigene families. In addition to this evolutionary perspective, understanding detoxification in bees is crucial to native bee conservation and the management of agricultural pollination in the face of widespread pesticide use. Bees provide an important ecosystem service through

pollination, both of wild plants and economically important crop species (Klein et al. 2007; Vanbergen & Initiative 2013). Whilst most insecticide assessments are carried out in *Apis mellifera*, they do not always reflect the sensitivity of native bees, with variation between species (Arena & Sgolastra 2014; Franklin & Raine 2019). Toxicity assays require large numbers of bees, difficult to achieve with many native bee species. Understanding the genetic mechanisms underlying insecticide sensitivity may allow us to make predictions based on P450 repertoires and sequences in different species (López-Osorio & Wurm 2020) . A recent example of this comes from *Megachile rotundata*, a species which exhibits increased sensitivity to neonicotinoids, and also lacks the P450 enzymes known to be involved in neonicotinoid metabolism in other species (Hayward et al. 2019). A combination of comparative genomic studies, alongside toxicity assays and molecular and functional studies of metabolism and insecticide resistance will allow us to extend our understanding to other bee species.

Materials and Methods

Gene family annotation

First, we gathered previously annotated P450s for nine species: *Apis mellifera*, *Bombus terrestris*, *Dufourea novaeangliae*, *Eufriesea mexicana*, *Habropoda laboriosa*, *Lasioglossum albipes*, *Megachile rotundata*, *Melipona quadrifasciata*, and *Osmia bicornis* (Hayward et al. 2019; Beadle et al. 2019; Kapheim et al. 2015; Johnson et al. 2018). We combined all P450s described in previous publications and aligned them using MAFFT v7.453 (-maxiterate 1000, using L-INS-I algorithm) (Katoh et al. 2002, 2005; Katoh & Standley 2013). We then used the *bio3d* package in R to compare the sequences, removing duplicates to curate a final dataset (Grant et al. 2021). We searched either the National Center for Biotechnology Information (NCBI) or the Hymenoptera Genome Database to provide accession numbers for the included P450s. (Elsik et al. 2018; NCBI Resource Coordinators 2018). For genes not currently included in an annotation but previously manually annotated, we included the scaffold accession number where the sequence can be found (Supplementary data 1). We checked each sequence for a complete protein domain using the NCBI conserved domain search (Marchler-Bauer et al. 2015). In some cases the sequence from the database contained an incomplete domain which had been manually edited in the dataset provided by Johnson and colleagues (Johnson et al. 2018), and in these cases we used the manually curated sequence and noted this (Supplementary data 1).

We then annotated the P450 gene family in the *Euglossa dilemma* genome (Brand et al. 2017) using all previously annotated *Apis mellifera* protein sequences as a reference (Elsik et al. 2018). Firstly, we identified scaffolds in the *Eg. Dilemma* genome containing P450s using TBLASTN (Altschul et al. 1990), and then annotated proteins on these scaffolds using exonerate (Slater & Birney 2005). Gene models were extracted, translated to amino acid

sequences and checked by comparison to *Apis mellifera* proteins. Gene models were manually curated using IGV, checking for splice sites and start and stop codons (Robinson et al. 2011). To check for P450s missed by the initial search, the predicted *Eg. Dilemma* P450s, as well as previously described *Ef. mexicana* P450s, were used as a query for the initial step of TBLASTN followed by exonerate predictions (Altschul et al. 1990; Slater & Birney 2005). As a further check we used already published RNA-seq data of ovaries and brains of *Eg. Dilemma* (Bioproject PRJNA523381) to improve the annotations (Saleh & Ramírez 2019). We trimmed the reads using TrimGalore! (Martin 2011). We then mapped the genes (2pass) to the *Eg. dilemma* genome (Brand et al. 2017) with the manually edited annotation (Edil_v1.0_revised.gff) using STAR (Dobin et al. 2013). We used StringTie (Pertea et al. 2016) to assemble and merge the transcripts to form an annotation which we compared to the edited annotation in IGV (Robinson et al. 2011). To check that the predicted P450s contain complete functional protein domains we used the NCBI conserved domain search (Marchler-Bauer et al. 2015). One gene, Edil_14289, is found at the end of a scaffold in a poorly sequenced area. The exon from the unsequenced area is found at the end of another scaffold and we combined these to form a complete gene in our analyses. All P450s identified were included in a revised version of the current *Eg. dilemma* annotation (Edil_v1.0_revised.gff) and included in Supplementary data 1.

Phylogenetic analyses

All 481 bee P450 sequences were combined with CYP51 from *Mus musculus* as an outgroup. The sequences were aligned using MAFFT v7.453 (-maxiterate 1000, using L-INS-I algorithm) (Katoh et al. 2002, 2005; Katoh & Standley 2013). Phylogenetic trees were constructed using IQ-TREE with the ModelFinder function to determine the best-fit model (Nguyen et al. 2015; Kalyaanamoorthy et al. 2017; Hoang et al. 2018). The Newick phylogeny was plotted using MEGA X, and the packages *ape*, *evobiR*, and *geiger* in R

version 3.6.2 (Pennell et al. 2014; Blackmon & Adams 2015; Kumar et al. 2018; Paradis & Schliep 2018; R Core Team 2019; Stecher et al. 2020).

To test for correlation between P450 repertoire and bee biology we carried out phylogenetic comparative analyses. The dataset used to produce the species-level phylogeny was composed of five nuclear genes: wingless, arginine kinase, opsin, Nak, Efl α (F2 copy) (Danforth et al. 2011). We identified copies of the five genes in *Apis mellifera* from NCBI (NCBI Resource Coordinators 2018). We then used TBLASTN to search the nucleotide datasets on NCBI for the sequences in *Bombus terrestris*, *Dufourea novaeangliae*, *Eufriesea mexicana*, *Habropoda laboriosa*, *Megachile rotundata*, and *Osmia bicornis* (Altschul et al. 1990). For *Euglossa dilemma*, *Lasioglossum albipes* and *Melipona quadrifasciata*, we used TBLASTN and exonerate to find and annotate the proteins in the genome (Altschul et al. 1990; Slater & Birney 2005; Kapheim et al. 2015; Kocher et al. 2013; Brand et al. 2017). The sequences were aligned using MAFFT v7.453 (-maxiterate 1000, using L-INS-I algorithm) (Katoh et al. 2002, 2005; Katoh & Standley 2013). Alignments were trimmed using trimAl (-gt 0.6, sites only included where sequence from 6/10 bee species present) (Capella-Gutierrez et al. 2009), and trimmed alignments were then concatenated. We constructed phylogenetic trees constructed using IQ-TREE with the ModelFinder function to determine the best-fit model (Nguyen et al. 2015; Kalyanamoorthy et al. 2017; Hoang et al. 2018), and constrained the relationship of Corbiculate bees according to a previous phylogeny (Romiguier et al. 2016). We rooted the tree using the midpoint.root function in the *phytools* package in R (Revell 2012).

Bees were determined as resin collecting or non-resin collecting following Johnson and colleagues (Johnson et al. 2018). We designated level of sociality as described previously (1= ancestrally solitary, 2= facultative basic eusociality, 3= obligate basic sociality, 4= obligate complex eusociality) (Kapheim et al. 2015). We tested for correlation between number of

P450s, CYP3s, or CYP6AS genes and the level of sociality by using a Spearman's rank correlation on phylogenetically independent contrasts (PICs). We calculated PICs using the function `pic` in the *ape* package (Paradis & Schliep 2018). We tested for a relationship between resin collection and P450 repertoire, using phylogenetic ANOVA, with the function `aov.phylo` from the package *geiger* (Pennell et al. 2014).

Clade stability and branch length analyses

We used MiPhy to identify clades and assess their stability (Curran et al. 2018). MiPhy requires a rooted tree as an input, however, our root did not belong to any of the species present in the species tree. Therefore, we removed the *Mus musculus* CYP15 from the dataset, aligned the sequences using MAFFT v7.453 (-maxiterate 1000, using L-INS-I algorithm) (Kato et al. 2002, 2005; Kato & Standley 2013), and constructed a phylogenetic tree using IQ-TREE with the ModelFinder function to determine the best-fit model (Nguyen et al. 2015; Kalyaanamoorthy et al. 2017; Hoang et al. 2018). We rooted the tree using the `midpoint.root` function in the *phytools* package in R (Revell 2012). We used default parameter settings along with the option "merge singletons" to ensure that all clades had more than one gene. Clades were named using nomenclature from *A. mellifera*, for the CYP6AS subfamily which contained multiple clades, we named the clades using the lowest numbered group member from *A. mellifera*.

To compare instability between the four CYP groups, we used a Welch one-way ANOVA test, followed by a post-hoc Games-Howell test using the package *rstatix* (Kassambara 2021). These tests do not assume homogeneity of variance.

We then applied CAFE to our dataset to analyze the evolutionary rates of different clades. This program uses a birth-death process to model gene gain and loss across a species tree to identify fast-evolving clades. The input files required are a species tree and gene

counts for each species in each clade, in this case the clades identified by MiPhy. We used the python script provided in the CAFE tutorial with the program r8s to convert the species tree constructed as described above into an ultrametric tree, as required for CAFE input (Sanderson 2003). We calibrated the tree using a previously published time-calibrated phylogeny (Cardinal et al. 2018). We found that three gamma rate categories ($k=3$) was the best fit for our data and showed convergence between runs.

For each clade we determined the cumulative branch length (CBL) by adding the terminal branches for each gene within a clade, obtained from the IQ-TREE output. As each clade has a different number of genes, we normalized the CBL by dividing the CBL by the number of genes in the clade, following (Finet et al. 2019). We also calculated the cumulative patristic distance per clade by summing all branch lengths, including internal branches within the clade. Again, we normalized the cumulative patristic distance by dividing by the number of genes in the clade. We compared both CBL and cumulative patristic distance between stable and unstable clades using a *t*-test. We tested for correlation between both CBL and cumulative patristic distance and the instability of a clade using Spearman's rank correlation.

Selection analyses

We split the dataset into the previously identified clades and carried out analysis individually on each clade. To construct a codon alignment for selection analyses we first aligned amino acid sequences using MAFFT v7.453 (-maxiterate 1000, using L-INS-I algorithm) (Kato et al. 2002, 2005; Kato & Standley 2013). We then used PAL2NAL to align the corresponding nucleotide sequences to the amino acid alignment (Suyama et al. 2006). Phylogenetic trees were constructed using IQ-TREE (Nguyen et al. 2015; Kalyaanamoorthy et al. 2017; Hoang et al. 2018).

Gene conversion

As gene conversion can lead to false positive results when testing for positive selection with PAML, we first tested our dataset for evidence of gene conversion (Casola & Hahn 2009). We split the dataset by species and aligned the sequences from each species separately using MAFFT v7.453 (-maxiterate 1000, using L-INS-I algorithm) (Katoh et al. 2002, 2005; Katoh & Standley 2013). We tested for gene conversion using GENECONV in RDP5 (Martin et al. 2021; Padidam et al. 1999). We tested the sequences as linear sequences and found no evidence for gene conversion in our dataset.

Branch-specific models

To compare selection pressure on clades which have undergone expansions with those which have not we used codon substitution models implemented in Phylogenetic Analysis by Maximum Likelihood (PAML) (Yang 2007). Firstly, we carried out “one-ratio” models. For these, we assumed that dN/dS (the ω ratio of nonsynonymous to synonymous substitutions) has one value across the whole clade. We compared the dN/dS value between stable and unstable clades using a *t*-test, and tested for a correlation between the dN/dS value of a clade and its normalized CBL using Spearman’s rank correlation.

We then carried out “free-ratio” models, in which we allowed ω to vary between lineages. We compared the two models (“one-ratio” and “free-ratio”) for each clade using a likelihood ratio test (LRT). To correct for multiple-testing, we used the p.adjust function in R, with false detection rate (fdr) correction, which controls for the proportion of false positives. To test if the unstable clades were more likely to be explained by a “free-ratio” model than stable clades we used a Fisher’s exact test.

Furthermore, to compare the selection pressures on individual genes or groups of genes within the clades, we again compared two models. The first assumed a single ω for the clade (“one-ratio”, as above), the second assumes two ω ratios, one for the lineage of interest, or group of interest, and the second for the rest of the tree. Again, we compared the two models using a likelihood ratio test (LRT), correcting for multiple-testing. To automate the process of testing each branch we used ETE3 framework to implement codeml (--leaves—internals --codeml_param CodonFreq,3 Nssites,0 fix_omega,0 omega,1 fix_kappa,0 kappa,0 cleandata,1 fix_blength,1) (Huerta-Cepas et al. 2016).

Site-specific models

To test for positively selected sites, we carried out two model comparisons: M1a vs. M2a, and M7 vs. M8. Model M1a is a nearly neutral model with two classes of sites, one evolving neutrally and the other under purifying selection. M2a is the same as M1a but with an additional site class for positive selection. Model M7 allows for a beta distribution of dN/dS across sites, while M8 has a beta distribution plus an additional site class with dN/dS>1 for positive selection. As above, we compared the two sets of models using likelihood ratio tests (LRT) and corrected for multiple-testing. We identified sites under positive selection using the Bayes Empirical Bayesian (BEB) in the PAML output (Yang et al. 2005). We considered sites to be under positive selection when the BEB posterior probability was greater than 0.9.

To determine where on the proteins these positively selected residues are located, we created protein models using ColabFold (Mirdita et al. 2021). ColabFold predicts protein structure based on AlphaFold2 (Jumper et al. 2021). We used the first sequence from each clade as a representative. We then visualized the resulting models using PyMOL (Schrödinger 2021).

627 We made figures using the packages *ggplot2* and *cowplot* in R version 3.6.2
628 (Wickham 2009, 2; R Core Team 2019).

Data Availability Statement

Data and R scripts used for analysis are available from Open Science Framework

(https://osf.io/9tdqu/?view_only=f7514c174fdb4574b77ed37ac1983f24).

Acknowledgements

Santiago R. Ramírez was funded by the David and Lucile Packard Foundation. We thank members of the Ramírez lab for their feedback. We also thank Philipp Brand for his helpful comments and advice. We are grateful to Dave Curran who added the option to merge singletons to MiPhy at our request, and also Fabio Henrique Kuriki Mendes who helped with the interpretation of the CAFE results. We thank Lucas Rubio and Amelia Bassiti for allowing us to use their photos of *Melipona quadrifasciata* and *Osmia bicornia*, respectively.

References

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403–410. doi: 10.1016/S0022-2836(05)80360-2.
- Arena M, Sgolastra F. 2014. A meta-analysis comparing the sensitivity of bees to pesticides. *Ecotoxicology.* 23:324–334. doi: 10.1007/s10646-014-1190-1.
- Arriaga-Osnaya BJ et al. 2017. Are body size and volatile blends honest signals in orchid bees? *Ecol. Evol.* 7:3037–3045. doi: 10.1002/ece3.2903.
- Balabanidou V et al. 2016. Cytochrome P450 associated with insecticide resistance catalyzes cuticular hydrocarbon production in *Anopheles gambiae*. *Proc. Natl. Acad. Sci. U. S. A.* 113:9268–9273. doi: 10.1073/pnas.1608295113.
- Bankova VS, Popov SS, Marekov NL. 1983. A Study on Flavonoids of Propolis. *J. Nat. Prod.* 46:471–474. doi: 10.1021/np50028a007.
- Beadle K et al. 2019. Genomic insights into neonicotinoid sensitivity in the solitary bee *Osmia bicornis* Gojobori, T, editor. *PLOS Genet.* 15:e1007903. doi: 10.1371/journal.pgen.1007903.
- Berenbaum MR, Johnson RM. 2015. Xenobiotic detoxification pathways in honey bees. *Curr. Opin. Insect Sci.* 10:51–58. doi: 10.1016/j.cois.2015.03.005.
- Blackmon H, Adams RA. 2015. EvobiR: Tools for comparative analyses and teaching evolutionary biology. doi: 10.5281/zenodo.30938.
- Brand P et al. 2017. The Nuclear and Mitochondrial Genomes of the Facultatively Eusocial Orchid Bee *Euglossa dilemma*. *G3 Genes Genomes Genet.* 7:2891–2898. doi: 10.1534/g3.117.043687.
- Brown DD, Wensink PC, Jordan E. 1972. A comparison of the ribosomal DNA's of *Xenopus laevis* and *Xenopus mulleri*: the evolution of tandem genes. *J. Mol. Biol.* 63:57–73. doi: 10.1016/0022-2836(72)90521-9.
- Calla B et al. 2017. Cytochrome P450 diversification and hostplant utilization patterns in specialist and generalist moths: Birth, death and adaptation. *Mol. Ecol.* 26:6021–6035. doi: 10.1111/mec.14348.
- Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics.* 25:1972–1973. doi: 10.1093/bioinformatics/btp348.
- Cardinal S, Buchmann SL, Russell AL. 2018. The evolution of floral sonication, a pollen foraging behavior used by bees (Anthophila). *Evolution.* 72:590–600. doi: <https://doi.org/10.1111/evo.13446>.
- Casola C, Hahn MW. 2009. Gene Conversion Among Paralogs Results in Moderate False Detection of Positive Selection Using Likelihood Methods. *J. Mol. Evol.* 68:679–687. doi: 10.1007/s00239-009-9241-6.

678 Claudianos C et al. 2006. A deficit of detoxification enzymes: pesticide sensitivity and
679 environmental response in the honeybee. *Insect Mol. Biol.* 15:615–636. doi: 10.1111/j.1365-
680 2583.2006.00672.x.

681 Curran DM, Gilleard JS, Wasmuth JD. 2018. MIPhy: identify and quantify rapidly evolving
682 members of large gene families. *PeerJ.* 6:e4873. doi: 10.7717/peerj.4873.

683 Danforth BN et al. 2011. Phylogeny and molecular systematics of bees (Hymenoptera:
684 Apoidea). <https://www.danforthlab.entomology.cornell.edu/research/resources/>.

685 Dermauw W, Van Leeuwen T, Feyereisen R. 2020. Diversity and evolution of the P450
686 family in arthropods. *Insect Biochem. Mol. Biol.* 127:103490. doi:
687 10.1016/j.ibmb.2020.103490.

688 Dobin A et al. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 29:15–21.
689 doi: 10.1093/bioinformatics/bts635.

690 Dressler RL. 1982. Biology of the Orchid Bees (Euglossini). *Annu. Rev. Ecol. Syst.* 13:373–
691 394. doi: 10.1146/annurev.es.13.110182.002105.

692 Drosophila 12 Genomes Consortium. 2007. Evolution of genes and genomes on the
693 Drosophila phylogeny. *Nature.* 450:203–218. doi: 10.1038/nature06341.

694 Eickwort GC, Kukuk PF, Wesley FR. 1986. The Nesting Biology of *Dufourea novaeangliae*
695 (Hymenoptera: Halictidae) and the Systematic Position of the Dufoureae Based on
696 Behavior and Development. *J. Kans. Entomol. Soc.* 59:103–120.

697 Eirín-López JM, Rebordinos L, Rooney AP, Rozas J. 2012. The birth-and-death evolution of
698 multigene families revisited. *Repetitive DNA.* 7:170–196. doi: 10.1159/000337119.

699 Elsik CG, Tayal A, Unni DR, Burns GW, Hagen DE. 2018. Hymenoptera Genome Database:
700 Using HymenopteraMine to Enhance Genomic Studies of Hymenopteran Insects. In:
701 Eukaryotic Genomic Databases: Methods and Protocols. Kollmar, M, editor. *Methods in*
702 *Molecular Biology* Springer: New York, NY pp. 513–556. doi: 10.1007/978-1-4939-7737-
703 6_17.

704 Eltz T, Sager A, Lunau K. 2005. Juggling with volatiles: exposure of perfumes by displaying
705 male orchid bees. *J. Comp. Physiol. A.* 191:575–581. doi: 10.1007/s00359-005-0603-2.

706 Eltz T, Whitten WM, Roubik DW, Linsenmair KE. 1999. Fragrance Collection, Storage, and
707 Accumulation by Individual Male Orchid Bees. *J. Chem. Ecol.* 25:157–176. doi:
708 10.1023/A:1020897302355.

709 Feyereisen R. 2011. Arthropod CYPomes illustrate the tempo and mode in P450 evolution.
710 *Biochim. Biophys. Acta BBA - Proteins Proteomics.* 1814:19–28. doi:
711 10.1016/j.bbapap.2010.06.012.

712 Feyereisen R. 2006. Evolution of insect P450. *Biochem. Soc. Trans.* 34:1252–1255. doi:
713 10.1042/BST0341252.

714 Finet C, Slavik K, Pu J, Carroll SB, Chung H. 2019. Birth-and-Death Evolution of the Fatty
715 Acyl-CoA Reductase (FAR) Gene Family and Diversification of Cuticular Hydrocarbon
716 Synthesis in *Drosophila*. *Genome Biol. Evol.* 11:1541–1551. doi: 10.1093/gbe/evz094.

717 Franklin EL, Raine NE. 2019. Moving beyond honeybee-centric pesticide risk assessments to
718 protect all pollinators. *Nat. Ecol. Evol.* 3:1373–1375. doi: 10.1038/s41559-019-0987-y.

719 Gilbert LI. 2004. Halloween genes encode P450 enzymes that mediate steroid hormone
720 biosynthesis in *Drosophila melanogaster*. *Mol. Cell. Endocrinol.* 215:1–10. doi:
721 10.1016/j.mce.2003.11.003.

722 Good RT et al. 2014. The Molecular Evolution of Cytochrome P450 Genes within and
723 between *Drosophila* Species. *Genome Biol. Evol.* 6:1118–1134. doi: 10.1093/gbe/evu083.

724 Grant BJ, Skjaerven L, Yao X-Q. 2021. The Bio3D packages for structural bioinformatics.
725 *Protein Sci. Publ. Protein Soc.* 30:20–30. doi: 10.1002/pro.3923.

726 Hardstone MC, Scott JG. 2010. Is *Apis mellifera* more sensitive to insecticides than other
727 insects? *Pest Manag. Sci.* 66:1171–1180. doi: 10.1002/ps.2001.

728 Hayward A et al. 2019. The leafcutter bee, *Megachile rotundata*, is more sensitive to N -
729 cyanoamidine neonicotinoid and butenolide insecticides than other managed bees. *Nat. Ecol.*
730 *Evol.* 3:1521–1524. doi: 10.1038/s41559-019-1011-2.

731 Helvig C, Koener JF, Unnithan GC, Feyereisen R. 2004. CYP15A1, the cytochrome P450
732 that catalyzes epoxidation of methyl farnesoate to juvenile hormone III in cockroach corpora
733 allata. *Proc. Natl. Acad. Sci. U. S. A.* 101:4024–4029. doi: 10.1073/pnas.0306980101.

734 Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving
735 the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* 35:518–522. doi:
736 10.1093/molbev/msx281.

737 Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: Reconstruction, Analysis, and Visualization
738 of Phylogenomic Data. *Mol. Biol. Evol.* 33:1635–1638. doi: 10.1093/molbev/msw046.

739 Johnson RM. 2015. Honey Bee Toxicology. *Annu. Rev. Entomol.* 60:415–434. doi:
740 10.1146/annurev-ento-011613-162005.

741 Johnson RM, Harpur BA, Dogantzis KA, Zayed A, Berenbaum MR. 2018. Genomic
742 footprint of evolution of eusociality in bees: floral food use and CYPome “blooms”. *Insectes*
743 *Sociaux.* 65:445–454. doi: 10.1007/s00040-018-0631-x.

744 Jumper J et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature.* 1–
745 7. doi: 10.1038/s41586-021-03819-2.

746 Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder:
747 fast model selection for accurate phylogenetic estimates. *Nat. Methods.* 14:587–589. doi:
748 10.1038/nmeth.4285.

749 Kapheim KM et al. 2015. Genomic signatures of evolutionary transitions from solitary to
750 group living. *Science.* 348:1139–1143. doi: 10.1126/science.aaa4788.

751 Kassambara A. 2021. *rstatix: Pipe-Friendly Framework for Basic Statistical Tests*. R
752 *package version 0.7.0*. <https://CRAN.R-project.org/package=rstatix>.

753 Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of
754 multiple sequence alignment. *Nucleic Acids Res.* 33:511–518. doi: 10.1093/nar/gki198.

755 Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple
756 sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30:3059–3066. doi:
757 10.1093/nar/gkf436.

758 Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7:
759 Improvements in Performance and Usability. *Mol. Biol. Evol.* 30:772–780. doi:
760 10.1093/molbev/mst010.

761 Klein A-M et al. 2007. Importance of pollinators in changing landscapes for world crops.
762 *Proc. R. Soc. B Biol. Sci.* 274:303–313. doi: 10.1098/rspb.2006.3721.

763 Kocher SD et al. 2013. The draft genome of a socially polymorphic halictid bee,
764 *Lasioglossum albipes*. *Genome Biol.* 14:R142. doi: 10.1186/gb-2013-14-12-r142.

765 Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: Molecular Evolutionary
766 Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* 35:1547–1549. doi:
767 10.1093/molbev/msy096.

768 Lespinet O, Wolf YI, Koonin EV, Aravind L. 2002. The Role of Lineage-Specific Gene
769 Family Expansion in the Evolution of Eukaryotes. *Genome Res.* 12:1048–1059. doi:
770 10.1101/gr.174302.

771 Li X, Schuler MA, Berenbaum MR. 2007. Molecular Mechanisms of Metabolic Resistance to
772 Synthetic and Natural Xenobiotics. *Annu. Rev. Entomol.* 52:231–253. doi:
773 10.1146/annurev.ento.51.110104.151104.

774 López-Osorio F, Wurm Y. 2020. Healthy Pollinators: Evaluating Pesticides with Molecular
775 Medicine Approaches. *Trends Ecol. Evol.* 35:380–383. doi: 10.1016/j.tree.2019.12.012.

776 Manjon C et al. 2018. Unravelling the Molecular Determinants of Bee Sensitivity to
777 Neonicotinoid Insecticides. *Curr. Biol.* 28:1137–1143.e5. doi: 10.1016/j.cub.2018.02.045.

778 Mao W et al. 2009. Quercetin-metabolizing CYP6AS enzymes of the pollinator *Apis*
779 *mellifera* (Hymenoptera: Apidae). *Comp. Biochem. Physiol. B Biochem. Mol. Biol.*
780 154:427–434. doi: 10.1016/j.cbpb.2009.08.008.

781 Mao W, Schuler MA, Berenbaum MR. 2017. Disruption of quercetin metabolism by
782 fungicide affects energy production in honey bees (*Apis mellifera*). *Proc. Natl. Acad. Sci.*
783 114:2538–2543. doi: 10.1073/pnas.1614864114.

784 Marchler-Bauer A et al. 2015. CDD: NCBI’s conserved domain database. *Nucleic Acids Res.*
785 43:D222–226. doi: 10.1093/nar/gku1221.

786 Martin DP et al. 2021. RDP5: a computer program for analyzing recombination in, and
787 removing signals of recombination from, nucleotide sequence datasets. *Virus Evol.* 7. doi:
788 10.1093/ve/veaa087.

789 Mendes FK, Vanderpool D, Fulton B, Hahn MW. 2020. CAFE 5 models variation in
790 evolutionary rates among gene families. *Bioinformatics*. 36:5516–5518. doi:
791 10.1093/bioinformatics/btaa1022.

792 Mirdita M, Ovchinnikov S, Steinegger M. 2021. *ColabFold - Making protein folding*
793 *accessible to all*. doi: 10.1101/2021.08.15.456425.

794 NCBI Resource Coordinators. 2018. Database resources of the National Center for
795 Biotechnology Information. *Nucleic Acids Res.* 46:D8–D13. doi: 10.1093/nar/gkx1095.

796 Nei M, Rooney AP. 2005. Concerted and birth-and-death evolution of multigene families.
797 *Annu. Rev. Genet.* 39:121–152. doi: 10.1146/annurev.genet.39.073003.112240.

798 Nelson DR. 1998. Metazoan cytochrome P450 evolution. *Comp. Biochem. Physiol. C*
799 *Pharmacol. Toxicol. Endocrinol.* 121:15–22. doi: 10.1016/S0742-8413(98)10027-0.

800 Nelson DR, Goldstone JV, Stegeman JJ. 2013. The cytochrome P450 genesis locus: the
801 origin and evolution of animal cytochrome P450s. *Philos. Trans. R. Soc. B Biol. Sci.*
802 368:20120474. doi: 10.1098/rstb.2012.0474.

803 Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A Fast and Effective
804 Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.*
805 32:268–274. doi: 10.1093/molbev/msu300.

806 Niwa R et al. 2004. CYP306A1, a cytochrome P450 enzyme, is essential for ecdysteroid
807 biosynthesis in the prothoracic glands of *Bombyx* and *Drosophila*. *J. Biol. Chem.* 279:35942–
808 35949. doi: 10.1074/jbc.M404514200.

809 Oakeshott JG et al. 2010. Metabolic enzymes associated with xenobiotic and chemosensory
810 responses in *Nasonia vitripennis*. *Insect Mol. Biol.* 19 Suppl 1:147–163. doi: 10.1111/j.1365-
811 2583.2009.00961.x.

812 Padidam M, Sawyer S, Fauquet CM. 1999. Possible Emergence of New Geminiviruses by
813 Frequent Recombination. *Virology*. 265:218–225. doi: 10.1006/viro.1999.0056.

814 Paradis E, Schliep K. 2018. ape 5.0: an environment for modern phylogenetics and
815 evolutionary analyses in R. *Bioinformatics*. doi: 10.1093/bioinformatics/bty633.

816 Pascarella J. 2007. Foraging patterns of the southeastern blueberry bee *Habropoda laboriosa*
817 (Apidae, Hymenoptera): Implications for understanding oligolecty. *J. Apic. Res. - J Apic.*
818 *RES.* 46:19–27. doi: 10.3896/IBRA.1.46.1.05.

819 Pennell MW et al. 2014. geiger v2.0: an expanded suite of methods for fitting
820 macroevolutionary models to phylogenetic trees. *Bioinformatics*. 30:2216–2218. doi:
821 10.1093/bioinformatics/btu181.

822 Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. 2016. Transcript-level expression
823 analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.*
824 11:1650–1667. doi: 10.1038/nprot.2016.095.

825 Pokorny T et al. 2017. Blown by the wind: the ecology of male courtship display behavior in
826 orchid bees. *Ecology*. 98:1140–1152. doi: 10.1002/ecy.1755.

827 R Core Team. 2019. *R: A language and environment for statistical computing*. R Foundation
828 for Statistical Computing: Vienna, Austria <http://www.R-project.org>.

829 Rane RV et al. 2019. Detoxifying enzyme complements and host use phenotypes in 160
830 insect species. *Curr. Opin. Insect Sci.* 31:131–138. doi: 10.1016/j.cois.2018.12.008.

831 Ranson H et al. 2002. Molecular analysis of multiple cytochrome P450 genes from the
832 malaria vector, *Anopheles gambiae*. *Insect Mol. Biol.* 11:409–418. doi: 10.1046/j.1365-
833 2583.2002.00350.x.

834 Revell LJ. 2012. phytools: an R package for phylogenetic comparative biology (and other
835 things). *Methods Ecol. Evol.* 3:217–223. doi: [https://doi.org/10.1111/j.2041-](https://doi.org/10.1111/j.2041-210X.2011.00169.x)
836 210X.2011.00169.x.

837 Rewitz KF, O'Connor MB, Gilbert LI. 2007. Molecular evolution of the insect Halloween
838 family of cytochrome P450s: phylogeny, gene organization and functional conservation.
839 *Insect Biochem. Mol. Biol.* 37:741–753. doi: 10.1016/j.ibmb.2007.02.012.

840 Robinson JT et al. 2011. Integrative genomics viewer. *Nat. Biotechnol.* 29:24–26. doi:
841 10.1038/nbt.1754.

842 Romiguier J et al. 2016. Phylogenomics Controlling for Base Compositional Bias Reveals a
843 Single Origin of Eusociality in Corbiculate Bees. *Mol. Biol. Evol.* 33:670–678. doi:
844 10.1093/molbev/msv258.

845 Saleh NW, Ramírez SR. 2019. Sociality emerges from solitary behaviours and reproductive
846 plasticity in the orchid bee *Euglossa dilemma*. *Proc. R. Soc. B Biol. Sci.* 286:20190588. doi:
847 10.1098/rspb.2019.0588.

848 Sanderson MJ. 2003. r8s: inferring absolute rates of molecular evolution and divergence
849 times in the absence of a molecular clock. *Bioinformatics.* 19:301–302. doi:
850 10.1093/bioinformatics/19.2.301.

851 Schrödinger L. 2021. *The PyMOL Molecular Graphics System*.

852 Sevrioukova IF, Li H, Zhang H, Peterson JA, Poulos TL. 1999. Structure of a cytochrome
853 P450–redox partner electron-transfer complex. *Proc. Natl. Acad. Sci.* 96:1863–1868. doi:
854 10.1073/pnas.96.5.1863.

855 Sezutsu H, Le Goff G, Feyereisen R. 2013. Origins of P450 diversity. *Philos. Trans. R. Soc.*
856 *B Biol. Sci.* 368:20120428. doi: 10.1098/rstb.2012.0428.

857 Slater GSC, Birney E. 2005. Automated generation of heuristics for biological sequence
858 comparison. *BMC Bioinformatics.* 6:31. doi: 10.1186/1471-2105-6-31.

859 Stecher G, Tamura K, Kumar S. 2020. Molecular Evolutionary Genetics Analysis (MEGA)
860 for macOS. *Mol. Biol. Evol.* 37:1237–1239. doi: 10.1093/molbev/msz312.

861 Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence
862 alignments into the corresponding codon alignments. *Nucleic Acids Res.* 34:W609–W612.
863 doi: 10.1093/nar/gkl315.

- 864 The Honeybee Genome Sequencing Consortium. 2006. Insights into social insects from the
865 genome of the honeybee *Apis mellifera*. Nature. 443:931–949. doi: 10.1038/nature05260.
- 866 Thomas JH. 2007. Rapid Birth–Death Evolution Specific to Xenobiotic Cytochrome P450
867 Genes in Vertebrates. PLOS Genet. 3:e67. doi: 10.1371/journal.pgen.0030067.
- 868 Tijet N, Helvig C, Feyereisen R. 2001. The cytochrome P450 gene superfamily in *Drosophila*
869 *melanogaster*: annotation, intron-exon organization and phylogeny. Gene. 262:189–198. doi:
870 10.1016/s0378-1119(00)00533-3.
- 871 Vanbergen AJ, Initiative the IP. 2013. Threats to an ecosystem service: pressures on
872 pollinators. Front. Ecol. Environ. 11:251–259. doi: 10.1890/120126.
- 873 Weber MG, Mitko L, Eltz T, Ramírez SR. 2016. Macroevolution of perfume signalling in
874 orchid bees. Ecol. Lett. 19:1314–1323. doi: 10.1111/ele.12667.
- 875 Wickham H. 2009. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York
876 <http://ggplot2.org>.
- 877 Wu Y et al. 2017. Comparative transcriptome analysis on the synthesis pathway of honey bee
878 (*Apis mellifera*) mandibular gland secretions. Sci. Rep. 7:4530. doi: 10.1038/s41598-017-
879 04879-z.
- 880 Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol.
881 24:1586–1591. doi: 10.1093/molbev/msm088.
- 882 Yang Z, Wong WSW, Nielsen R. 2005. Bayes Empirical Bayes Inference of Amino Acid
883 Sites Under Positive Selection. Mol. Biol. Evol. 22:1107–1118. doi: 10.1093/molbev/msi097.
- 884 Yohe LR, Liu L, Dávalos LM, Liberles DA. 2019. Protocols for the Molecular Evolutionary
885 Analysis of Membrane Protein Gene Duplicates. In: Computational Methods in Protein
886 Evolution. Sikosek, T, editor. Methods in Molecular Biology Springer: New York, NY pp.
887 49–62. doi: 10.1007/978-1-4939-8736-8_3.
- 888 Zhu F, Moural TW, Shah K, Palli SR. 2013. Integrated analysis of cytochrome P450 gene
889 superfamily in the red flour beetle, *Tribolium castaneum*. BMC Genomics. 14:174. doi:
890 10.1186/1471-2164-14-174.
- 891 Zimmermann Y, Ramírez SR, Eltz T. 2009. Chemical niche differentiation among sympatric
892 species of orchid bees. Ecology. 90:2994–3008. doi: 10.1890/08-1858.1.

893

894