Ecological correlates of gene family size in a pine-feeding sawfly genome and across Hymenoptera

Kim L. Vertacnik[1,2], Danielle K. Herrig[1], R. Keating Godfrey[3], Tom Hill[4,5], Scott M. Geib[6], Robert L. Unckless[4], David R. Nelson[7], and Catherine R. Linnen[1]

[1]Department of Biology, University of Kentucky, Lexington, KY 40506, USA
[2]Columbia River Inter-Tribal Fish Commission, Hagerman, ID 83332, USA (current address)
[3]Department of Neuroscience, University of Arizona, Tucson, AZ 85721, USA
[4]Department of Molecular Biosciences, University of Kansas, Lawrence, Kansas 66045, USA
[5]NIH Frederick National Laboratory for Cancer Research, Frederick, MD 21702, USA
[6]Tropical Crop and Commodity Protection Research Unit, United States Department of Agriculture: Agriculture Research Service Pacific Basin Agricultural Research Center, Hilo, Hawaii 96720, USA
[7]Department of Microbiology, Immunology and Biochemistry, University of Tennessee Health Science Center, Memphis, TN 38163, USA

Correspondence: kvertacnik@critfc.org

Keywords: evolutionary predictability, adaptation, chemosensation, detoxification, immunity, Hymenoptera, draft genome

Abstract

A central goal in evolutionary biology is to determine the predictability of adaptive genetic changes. Despite many documented cases of convergent evolution at individual loci, little is known about the repeatability of gene family expansions and contractions. To address this void, we examined gene family evolution in the redheaded pine sawfly *Neodiprion lecontei*, a non-eusocial hymenopteran and exemplar of a pine-specialized lineage evolved from angiosperm-feeding ancestors. After assembling and annotating a draft genome, we manually annotated multiple gene families with chemosensory, detoxification, or immunity functions and characterized their genomic distributions and evolutionary history. Our results suggest that expansions of bitter gustatory receptor (GR), clan 3 cytochrome P450 (CYP3), and antimicrobial peptide (AMP) subfamilies may have contributed to pine adaptation. By contrast, there was no evidence of recent gene family contraction via pseudogenization. Next, we compared the number of genes in these same families across insect taxa that vary in diet, dietary specialization, and social behavior. In Hymenoptera, herbivory was associated with small GR and olfactory receptor (OR) families, eusociality was associated with large OR and small AMP families, and—unlike investigations in more closely related taxa—ecological specialization was not related to gene family size. Overall, our results suggest that gene families that mediate ecological interactions may expand and contract predictably in response to particular selection pressures, however, the ecological drivers and temporal pace of gene gain and loss likely varies considerably across gene families.

Introduction

Changes in gene family size are a potentially important source of evolutionary innovation. When gene families grow via duplication, for example, reduced functional constraints may facilitate the development of phenotypic novelty (Ohno 1970; Demuth and Hahn 2009). Reductions in gene family size can also enable novel traits. For example, the colonization of highly specialized niches like oligotrophic caves (Protas et al. 2006; Gross et al. 2009; Yang et al. 2016) and toxic host plants (Matsuo et al. 2007; McBride 2007; Good et al. 2014) is linked to rampant pseudogenization. Together, these observations suggest that gene families predictably expand or contract in response to specific selection pressures. Yet compared to the rich and growing literature on genetic convergence at individual loci (Martin and Orgogozo 2013), the repeatability and predictability of gene family evolution remains understudied.

The evolution of many gene families, defined here as groups of genes that share sequence and functional similarity from common ancestry (Dayhoff 1976; Demuth and Hahn 2009), is consistent with a birth-death model where genes arise via duplication (gene gain) and are lost via deletion or pseudogenization (gene loss) (Hughes and Nei 1992; Nei and Rooney 2005). When frequency rates of

55 duplication and deletion evolve primarily through genetic drift, over time gene family sizes contract and
56 expand via a process dubbed genomic drift (Nei 2007; Nozawa et al. 2007). Overall, the stochastic birth-
57 death process of genomic drift (which differs from Nei's conceptual birth-death model of gene family
58 evolution (Hahn et al. 2005)) sufficiently explains most gene family size distributions within genomes
59 (Karev et al. 2002) and between species (Hahn et al. 2007).
60 But during an ecological shift, natural selection can influence birth-death dynamics by promoting
61 the expansion or contraction of specific gene families. Thus, taxa adapted to a novel niche may have
62 genomic evidence of selective maintenance for gene duplications or deletions. For example, if selection
63 favors gene gain, novel gene duplicates will tend to persist in the population and form subfamilies of
64 recently diverged paralogs. If the mutational mechanism that generates new duplicate genes is unequal
65 crossing over during meiosis, these recently diverged paralogs will be arranged in tandem arrays across
66 the genome (Zhang 2003). Moreover, if duplicates experience positive selection for novel functions, they
67 can have elevated amino acid substitution rates. Conversely, some genetic functions may become obsolete
68 or even deleterious in the novel habitat. In this case, positive or relaxed purifying selection will cause
69 some gene families to accumulate loss-of-function mutations at an accelerated rate.
70 After an ecological shift, impacted gene families will eventually reach a new equilibrium state
71 where gene number evolves primarily through negative selection and genomic drift. Likewise, tandem
72 array lengths will reflect local recombination rates (Akhunov et al. 2003; Zhang and Gaut 2003; Rizzon et
73 al. 2006; Thomas 2006) and pseudogenes will fade into the genomic background (Petrov et al. 1996;
74 Petrov and Hartl 1997, 1998). Thus, within-genome signatures of adaptive changes in gene family size are
75 likely ephemeral and best detected in lineages that *recently* shifted to a novel niche. Plus, if selection
76 consistently favors the expansion or contraction of specific gene families in specific environments,
77 among-taxon correlations between gene family size and ecology should be maintained. Currently, the
78 extent to which different taxa converge at the level of gene family size changes is largely unknown.
79 Arguably, the genes most likely to expand and contract convergently in response to similar
80 selection pressures are those that mediate organismal interactions with their biotic and abiotic
81 environments. These "environmentally responsive genes" include chemosensory (e.g., olfactory and
82 gustatory receptors), detoxifying (e.g., cytochrome P450), and immunity (e.g., immunoglobulin and
83 MHC) genes. To cope with constantly changing pressures, environmentally responsive genes tend to be
84 characterized by elevated sequence diversity, duplication rates, substitution rates, and genomic clustering,
85 as well as tissue- or temporal-specific expression (Berenbaum 2002) and limited pleiotropy (Arguello et
86 al. 2016). Importantly, causal links between changes in environmentally responsive genes and adaptation
87 to novel niches have been established for multiple taxa (Després et al. 2007; Matsuo et al. 2007; Dobler et
88 al. 2012; Zhen et al. 2012; Sezutsu et al. 2013).

With exceptionally diverse ecologies and an ever-increasing availability of annotated genomes (Consortium 2013; Poelchau et al. 2015), insects are a powerful system for investigating the predictability of size changes in environmentally responsive gene families. To date, at least two ecological transitions are hypothesized to have a predictable impact on gene family size in insect lineages. In plant-feeding insects, the evolution of increased dietary specialization (i.e., smaller diet breadth) is associated with reduced chemosensory and detoxifying gene family sizes (McBride 2007; McBride and Arguello 2007; Good et al. 2014; Goldman-Huertas et al. 2015; Calla et al. 2017; Comeault et al. 2017) but see (Gardiner et al. 2008). In hymenopteran insects, eusociality is associated with expansions of the olfactory-receptor family and contractions of the gustatory-receptor family (Robertson and Wanner 2006; Zhou et al. 2015; McKenzie et al. 2016; Brand and Ramírez 2017) but see (Fischman et al. 2011; Johnson et al. 2018). Most of these studies, however, consider a single ecological characteristic or gene family (but see Robertson and Wanner 2006) which is problematic since changes in social behavior may often be accompanied by changes in other ecological characteristics and vice versa (Faulkes et al. 1997; Duffy and Macdonald 2010; Ross et al. 2013). A better understanding of ecology and gene family size relationships requires simultaneous consideration of multiple ecological characteristics and diverse gene families.

Here, we characterize multiple environmentally responsive gene families in the genome of the redheaded pine sawfly, *Neodiprion lecontei* (Order: Hymenoptera; Family: Diprionidae). This species provides an opportunity to examine both within-genome signatures of adaptive gene family contractions and expansions, and among-lineage correlations between ecology and gene family size. First, for within-genome signatures, *N. lecontei* is an exemplar of an herbivorous hymenopteran lineage (Diprionidae) that underwent a drastic host shift: sometime within the last 60 million years, this lineage transitioned from angiosperms to coniferous host plants in the family Pinaceae (Boevé et al. 2013; Peters et al. 2017). To defend against herbivores and pathogens, Pinaceae produce viscous oleoresin secretions that are sticky and have unique antimicrobial properties (Trapp and Croteau 2001; Gershenzon and Dudareva 2007). To manage these toxic and extraordinarily sticky resins, *N. lecontei* and related diprionids evolved specialized feeding and egg-laying traits (Figure 1). Beyond these traits, we hypothesize that pine specialization likely resulted in pronounced changes to the selection pressures acting on multiple gene families, especially those involved in chemosensation, detoxification, and immune function. Second, with respect to among-lineage correlations between ecology and gene-family size, *N. lecontei* is an herbivorous, non-eusocial insect from the Eusymphyta, a massively understudied hymenopteran clade (Peters et al. 2017). Although many assembled and annotated hymenopteran genomes are currently available, almost all have come from apocritans (bees, wasps, and ants, but see (Robertson et al. 2018)). Thus, *N. lecontei* increases the ecological, behavioral, and taxonomic diversity of hymenopteran genomes for evaluating ecological correlates of gene family size among taxa.

To evaluate the predictability of gene family evolution, we assembled a draft genome for *N. lecontei* and manually annotated genes for five environmentally responsive gene families: olfactory receptor (OR), gustatory receptor (GR), odorant binding protein (OBP), cytochrome P450 (CYP), and antimicrobial peptide (AMP). For gene families that underwent a size change related to pine adaptation, we expected one or more of the following patterns: (1) clusters of recently diverged paralogs in gene-family trees, (2) a high proportion of genes in tandem arrays, (3) signatures of positive selection among paralogs, and (4) elevated rates of pseudogenization. Then, for the same five gene families, we asked whether gene-family size correlated with ecology among distantly related insect taxa. To do so, we compiled published gene annotations and ecological variables (diet type, degree of ecological specialization, presence/absence of eusociality) for hymenopteran taxa. Together, these analyses identify possible candidate gene families underlying pine specialization and reveal that relationships between gene family size and ecology differ among environmentally responsive gene families.

Results

Genome assembly and annotation

Sequencing and assembly

We sequenced one mate-pair and two small-insert Illumina libraries made from haploid male siblings (see Methods). After read processing, we retained 268 billion PE100 reads with a combined read depth of 112x (Table S1). ALLPATHS-LG (v47417) (Gnerre et al. 2011) produced a 239-Mbp assembly consisting of 4523 scaffolds, with a scaffold N50 of 243 kbp (Table S2). Prior studies identified seven chromosomes in *N. lecontei* (Smith 1941; Maxwell 1958; Sohi and Ennis 1981; Linnen et al. 2018). With an estimated genome size (1C) of 331 ±9.6 Mbp, our assembly captured 72% of the genome. Overall, these metrics are comparable to other hymenopteran assemblies (Table S2).

To measure assembly completeness and artificial sequence duplication, we used CEGMA (Parra et al. 2007) and BUSCO (Simão et al. 2015). Both search the assembly for a set of single-copy, conserved genes, however, the CEGMA software has been deprecated (http://korflab.ucdavis.edu/Datasets/cegma). Of the 248 CEGMA core eukaryotic genes, 90% aligned as complete, single copies and 8% aligned complete but duplicated. For BUSCO, we used the OrthoDB arthropod dataset, and out of 2675 groups 77% were complete, single copies and 3% were complete but duplicated. These metrics indicate the presence of artificial duplicate sequences, but otherwise the assembly was reasonably complete and suitable for annotation.

About 15.8% of the assembly consisted of repetitive elements, including 122 unknown transposable elements that were mostly unique to *N. lecontei* (Table S3), and 212 other families of transposable elements and simple repeats. This 15.8% corresponds to 11.4% of the actual 331-Mb

genome, of which we predict 27.6% is repetitive, suggesting that ~16.1% of the missing ~28% of the genome is repetitive content (Table S3). For de novo gene prediction, we included the *N. lecontei* transcriptome and protein-coding genes from *Atta cephalotes* (OGSv1.2), *Acromyrmex echinatior* (OGSv3.8), *Apis melifera* (OGSv3.2), *Athalia rosae* (OGSv1.0), and *Nasonia vitripennis* (OGSv1.0) to guide annotation. The official gene set (OGSv1) had 12,980 gene models while the transcriptome had an average of 26,000 transcripts per tissue (Table S4).

Olfactory receptor

The OR gene family had 56 genes total, including the co-receptor *Orco*; one gene contained stop codons, three were partial annotations, and 52 genes were intact (Table 1). In *D. melanogaster* most olfactory sensory neurons (OSNs) express a single OR (along with the coreceptor, *Orco*), and OSNs expressing a particular OR converge on a single glomerulus in the antennal lobe (Gao et al. 2000; Vosshall et al. 2000; Couto et al. 2005) but see (Fishilevich and Vosshall 2005). This anatomy results in a general one-to-one correspondence between the number of ORs and the number of glomeruli, a correspondence also observed in the hymenopteran European honey bee (*Apis mellifera*, (Robertson and Wanner 2006)). Based on these studies and examination of the antennal lobes of adult male and adult female *N. lecontei*, we expected to find a minimum of 49 functional ORs (Table S5, Figure 2). The close correspondence between our gene annotations and glomeruli counts suggests that we have located all or most *N. lecontei* OR genes.

59% of ORs were in genomic clusters of two or more genes (Figure 3), a low proportion compared to many other hymenopteran OR families (Robertson and Wanner 2006; Zhou et al. 2015; McKenzie et al. 2016; Brand and Ramírez 2017). A phylogenetic analysis of OR protein sequences from *Neodiprion,* six other hymenopterans, and *D. melanogaster* identified three *Neodiprion*-specific clades with at least five genes (Figure S1a). These same three clades were also recovered in a phylogenetic analysis of *Neodiprion* OR cDNA sequences (Figure S1b). For each *Neodiprion*-specific OR clade (and *Neodiprion*-specific clades in other gene families, see below), we used the *Neodiprion* cDNA tree, the codeml program in the PAML package (Yang 2007), and likelihood-ratio tests to ask: (1) whether the ratio of non-synonymous to synonymous substitution rates (dN/dS or ω) for the focal OR clade differed from the rest of the *Neodiprion* OR gene family and, if so, whether they exhibited evidence of positive selection (ω>1) (branch tests); and (2) whether ω differed among sites across members of *Neodiprion*-specific clades and, if so, which sites exhibited evidence of positive selection (site tests). For only one OR clade (OR clade 1) did we detect evidence of branch-specific positive selection (i.e., rejection of both 1-ratio and fixed-ω models), but this clade lacked evidence of site-specific positive selection (Table 2).

Gustatory receptor

190 The GR gene family had 44 genes total; two genes contained stop codons, two were partial
191 annotations (one annotation was both partial and pseudogenized), and 41 were intact (Table 1). 76% of
192 the GRs that could be placed on chromosomes were in genomic clusters (Figure 3) with three *Neodiprion*-
193 specific clades of at least five genes (Figures S2a and S2b). Only one clade (GR clade 3) had evidence of
194 branch-specific positive selection (Table 2). This clade also had evidence of positive selection at some
195 amino acid positions among paralogs (Table 2; sites with evidence of positive selection include: 77E,
196 79S, 146N, 275S, 301S). Notably, GR Clade 3 is an expansion of six paralogs orthologous to *DmGR66a*,
197 a bitter receptor specifically for caffeine (Moon et al. 2006). However, *N. lecontei* orthologs were not
198 found for *DmGR93a* (Lee et al. 2009) and *DmGr33a* (Moon et al. 2009), coreceptors possibly required
199 for caffeine detection. Together, these data suggest that caffeine-like GR receptors have been coopted for
200 novel functions in *N. lecontei*.
201 The GR family also had orthologs for sugar receptors *DmGR5a* (trehalose) (Dahanukar et al.
202 2001), *DmGR43a* (fructose) (Miyamoto et al. 2012), and *DmGR64a-f* (multiple sugars) (Slone et al.
203 2007) as well as carbon dioxide receptors *DmGR21a* and *DmGR63a* (Jones et al. 2007) (Figure S2a).
204 Orthologs to these carbon dioxide receptors have not been found in Apocrita but seem to be preserved in
205 Symphyta, like *N. lecontei* (Robertson and Kent 2009; Robertson et al. 2018).
206 *Odorant binding protein*
207 The OBP gene family had 13 genes total; none were pseudogenized or partial annotations (Table
208 1). In this family, 38% of genes were in genomic clusters, including a cluster of five genes on
209 chromosome 6 (Figure 3). *Neodiprion*-specific OBP clades were not found, even for the chromosome 6
210 cluster. We note, however, that the OBP phylogenies had low bootstrap support (Figure S3a,b), making it
211 difficult to infer relationships among paralogs.
212 *Cytochrome P450*
213 The CYP gene family had 107 genes total; twelve genes contained stop codons, two were partial
214 annotations, and 93 were intact (Table 1). In insects, CYPs belong to four major clades, which are
215 referred to as clans (Feyereisen 2012). When we split the CYP gene family by clan, the CYP2 clan had
216 nine intact genes; the CYP3 clan had 47 intact genes and eight pseudogenes; the CYP4 clan had 27 intact
217 genes, four pseudogenes, and two partial genes; and the mitochondrial CYP clan had 10 intact genes
218 (Table 1). Across all CYPs, 66% were in genomic clusters (Figure 3). Looking at the four major clans
219 separately, the percentage of clustered genes were: 33% for CYP2, 81% for CYP3, 55% for CYP4, and
220 50% for mitochondrial CYP.
221 The CYP gene family had five *Neodiprion*-specific clades with at least five genes (Figure S4a,b),
222 four of which were in the CYP3 clan. Of these, two clades that were both within the CYP3 clan (CYP
223 clades 3 and 5) had evidence of branch-specific, but not site-specific, positive selection (Table 2). CYP

224 clade 3 contained members of the CYP6 subfamily, and the CYP clade 5 contained members of the
225 CYP336 subfamily. Several studies to date suggest that members of the CYP3 clan—and the CYP6
226 subfamily in particular—play an important role in detoxifying pesticides and host-plant allelochemicals
227 (Feyereisen 2012).
228 Orthologs were found for all the Halloween genes (which include genes from both the CYP2 and
229 mitochondrial CYP clans) of the 20-hydroxy ecdysone biosynthesis pathway: *CYP302A1* (disembodied),
230 *CYP306A1* (phantom), *CYP307A2* (spookier), *CYP307B1* (spookiest), *CYP314A1* (shade), *CYP315A1*
231 (shadow), and *CYP18A1* which turns over 20-hydroxy ecdysone (Rewitz et al. 2007; Feyereisen 2011;
232 Guittard et al. 2011; Qu et al. 2015). The juvenile hormone biosynthesis gene *CYP15A1* was present as
233 well (Helvig et al. 2004). Finally, *N. lecontei* had orthologs for the two CYP4G enzymes that synthesize
234 the cuticular hydrocarbons used as external waterproof coating (Qiu et al. 2012).
235 *Immunity*
236 Antimicrobial peptides (AMPs) are expressed upon infection to kill or inhibit microbes. Based on
237 hymenopteran sequences, the *N. lecontei* AMP gene family had 21 genes (Table 1; Table S6), including
238 single copies of *Hymenoptaecin*, *Abaecin*, and *Tachystatin*, but no clear *Defensin* ortholog. Over 18
239 *Hisnavicin* genes were identified, including a *Neodiprion*-specific expansion of eight histidine-rich
240 paralogs orthologous to *Hisnavicin-4*, which has been characterized as a larval cuticle protein and AMP,
241 but not functionally tested (Tian et al. 2010). The *N. lecontei* Hisnavicins had a conserved 62 amino acid
242 motif that appeared up to 19 times in a single protein; the purpose of this amplification is unknown. 95%
243 of the AMPs were in genomic clusters (Figure 3). Due to low bootstrap support on many of the branches
244 in our Hisnavicin protein tree, we could not identify unambiguous *Neodiprion*-specific clades (Figure
245 S5a). However, our *Neodiprion* cDNA tree (Figure S5b) did reveal strong support for the monophyly of a
246 cluster of 15 *Hisnavicins* on linkage group 5 (Figure 3), and this cluster had some evidence of positive
247 selection (Table 2).
248 Outside of the AMP family, most immune pathways had direct orthologs between *N. lecontei* and
249 *D. melanogaster* (Figure S6, Table S7). The basic viral siRNA response pathway was completely
250 conserved between species. The immune deficiency (IMD) pathway was missing an ortholog for the
251 peptidoglycan recognition receptor *PGRP-LC*, but it is likely that another *PGRP* replaced *PGRP-LC* in *N.*
252 *lecontei*; assigning PGRP orthology was also difficult in ants (Gupta et al. 2015). Also missing is the
253 *Drosophila* mitogen activated protein kinase kinase kinase, TGF-β activated kinase 1 (*Tak1*), but *N.*
254 *lecontei* had a similar TGF-β activated kinase that is a close ortholog to several *Tak1-like D.*
255 *melanogaster* proteins possibly involved in immune deficiency signaling. The encapsulation/melanization
256 pathway was missing one of the two *Drosophila* GTPases (*Rak2*). The *N. lecontei Rak1* ortholog may be
257 playing both roles, but again this is likely due to the difficulty of assigning one-to-one orthologs. The

258 Duox pathway was missing the top G-protein coupled receptor, but this is unknown in *D. melanogaster*
259 and unidentified in other Hymenoptera (Evans et al. 2006). Interestingly, *N. lecontei* had two copies of
260 Dual Oxidase (*Duox*), which regulates commensal gut microbiota and infectious microbes (Ha et al. 2005;
261 Lee et al. 2015); *Apis mellifera* had one copy. Finally, the Toll pathway *NF-kappaB* transcription factor,
262 *Dorsal-related immunity factor (Dif)* does not have a one-to-one ortholog in *N. lecontei*, but two copies of
263 its paralog, *Dorsal*, were present.

264 **Within-genome signatures of adaptive expansions and contractions**

265 *Evidence of selection in* Neodiprion-*specific gene family clades*

266 Massive gene family expansions with dozens of genes were not found in *N. lecontei* (in contrast
267 to (Smadja et al. 2009; Zhou et al. 2015)). Instead, the largest *Neodiprion*-specific clade had 22 genes
268 (CYP gene family) and the rest had fewer than 10 genes. Nevertheless, we did identify 11 *Neodiprion*-
269 specific clades containing at least 5 closely related paralogs and a monophyletic clade of 15 AMPs with
270 ambiguous ancestry (Table 1). Of these 12 clades, four had significant branch positive selection (OR
271 clade 1, GR clade 3, CYP clade 3, and CYP clade 5) (Table 2). Of these four clades, only one also had
272 significant site-specific positive selection (GR clade 3) (Table 2).

273 *Clustering*

274 Our five focal gene families varied in the proportion of genes that were found in clusters of two
275 or more genes (Fisher's exact test, $P = 0.002$; Table 1). Post-hoc tests revealed that much of this variation
276 was due to differences between the highly clustered AMP family and all other families except GR (AMP
277 vs. OR: $P = 0.0091$; AMP vs. OBP: $P = 0.0053$; AMP vs. CYP: $P = 0.024$, AMP vs. GR: $P = 0.12$; all p-
278 values are FDR-corrected). The only other difference in clustering that we detected was between the GR
279 and singleton-heavy OBP families (FDR-corrected $P = 0.045$).

280 Differences in clustering were even more pronounced when we separated the CYP family by clan
281 (Fisher's exact test, $P < 0.0001$; Table 1). In addition to the pairwise differences described above, we also
282 found that the proportion of CYP3 genes found in clusters differed significantly from ORs, CYP2s, and
283 CYP4s (all FDR-corrected $P < 0.05$), but not AMPs, GRs, and mitochondrial CYPs. Additionally, AMP
284 clustering differed from CYP2, CYP4, and mitochondrial CYP, while GR differed from CYP2 (all FDR-
285 corrected $P < 0.05$). Together, these analyses identified AMP, GR, and CYP3 as having an unusually high
286 proportion of genes found in clusters compared to other environmentally responsive gene families.

287 *Pseudogenization*

288 Overall, we found very few pseudogenes, and the proportion of pseudogenized genes did not
289 differ significantly among gene families (Fisher's exact test, $P = 0.12$; Table 1). The chemoreceptors had
290 one pseudogene each while CYP had 12, which is about 10% of the family, but this was also the largest
291 gene family. Although CYP3 had more pseudogenes than other CYP clans, the proportion of

292 pseudogenized genes still did not differ when we compared CYP clans (Fisher's exact test, $P = 0.10$).
293 Given these low rates of pseudogenization, it is unlikely that *N. lecontei* gene families underwent
294 substantial, recent contractions.

295 **Ecological correlates of gene-family size across insects**

296 We first examined broad-scale variation in the sizes of our five focal gene families and four CYP
297 clans among different insect orders (Figure S7). Not surprisingly, sample sizes were highly variable
298 across gene families and insect orders. Despite this variation, we observed some intriguing differences
299 among gene families and taxa. We detected significant differences in gene family size among orders for
300 OR (Kruskal-Wallas chisq = 48.2, df = 12, $P < 1 \times 10^{-5}$), GR (K-W chisq = 25.5, df = 9, $P = 0.0025$),
301 and OBP (K-W chisq = 37.6, df = 9, $P < 1 \times 10^{-4}$), but not CYP (K-W chisq = 10.3, df = 7, $P = 0.17$) or
302 AMP (K-W chisq = 7.93, df = 5, $P = 0.16$). We note, however, that the AMP sample size was
303 considerably smaller than the other gene families. When we looked at CYP clans individually, we found
304 differences among orders for CYP4 (K-W chisq = 19.0, df = 7, $P = 0.0083$) and mitochondrial CYP (K-W
305 chisq = 16.3, df =7, $P = 0.022$), but not CYP2 (K-W chisq = 9.19, df = 7, $P = 0.24$) or CYP3 (K-W chisq
306 = 8.76, df = 7, $P = 0.27$).

307 For the OR family, among-group differences in gene number were mostly attributable to an
308 unusually large number of OR genes in Hymenoptera (significant post-hoc tests include Diptera vs.
309 Hymenoptera: $P = 0.0018$; Hemiptera vs. Hymenoptera: $P = 0.00014$; and Odonata vs. Hymenoptera: $P =$
310 0.011; all p-values are FDR-corrected). By contrast, the size of the OBP family was larger in Diptera than
311 other orders (significant post-hoc tests include Diptera vs. Hymenoptera: $P = 0.00037$; Diptera vs.
312 Hemiptera: $P = 0.00092$; all p-values are FDR-corrected). Although none of the post-hoc tests were
313 significant for GR family size, the Blattodea appear to have more GRs on average than other insect orders
314 (Figure S7). For CYP clans, posthoc tests revealed that hymenopterans have fewer CYP4s than dipterans
315 (FDR-corrected $P = 0.010$) and fewer mitochondrial CYPs than both dipterans and lepidopterans (FDR-
316 corrected $P = 0.024$ and 0.023, respectively).

317 We next examined how gene family size correlated with ecology within the hymenopteran clade
318 (Figures 4 and 5). Once again, we observed differences among gene families. We found that the number
319 of ORs differed significantly among hymenopteran species that differed in diet (Kruskal-Wallas chisq =
320 15.8, df = 3, $P = 0.0012$) and sociality (Wilcoxon rank-sum test W = 115; $P = 0.00094$). For diet, we
321 found that herbivores had fewer ORs than all other diet types (fungivores vs. herbivores: $P = 0.015$;
322 omnivores vs. herbivores: $P = 0.015$; insectivores vs. herbivores: $P = 0.048$; all p-values are FDR-
323 corrected). We observed an even more striking difference between eusocial and non-eusocial
324 hymenopterans, with the former having larger OR families, on average. By contrast, GR family size was
325 related to diet (Kruskal-Wallas chisq = 11.8, df = 3, $P = 0.0082$), but not sociality (Wilcoxon rank sum

test W = 30; $P = 0.65$). And CYP family size was related to sociality (W = 2; $P = 0.045$), but not diet ($P = 0.38$). Finally, specialists and generalists did not differ significantly in gene family size in any of the gene families and ecology was unrelated to gene family size for OBP and CYP (total CYP number and individual CYP clans). Although these analyses have several limitations (see discussion), these results are consistent with the hypothesis that environmentally responsive gene families may contract or expand predictably in response to particular selection pressures.

Discussion

The predictability of gene family expansion or contraction in response to specific selection pressures is still an open question. Here, we evaluated genomic signatures of adaptive gene family size changes in five environmentally responsive gene families within the *N. lecontei* draft genome, a hymenopteran exemplar of a pine-specialized lineage. Although we saw minimal evidence of recent gene loss via pseudogenization, at least three gene families (AMP, GR, and CYP3) had genomic distributions consistent with the selective maintenance of novel gene duplicates, and two of these families also had evidence of positive selection within *Neodiprion*-specific clades (GR and CYP3). Next, we examined these same gene families in other hymenopterans to see if family size correlated with diet, ecological specialization, or eusocial behavior. Among Hymenoptera, we found that OR family size was correlated with eusociality and diet type, but not dietary specialization; GR family size was correlated with diet type; and AMP family size was associated with eusociality. These results suggest that ecology can have a predictable impact on gene family size and that different selection pressures impact different gene families. Below, we discuss both the implications and limitations of our analyses and suggest priorities for future comparative work on gene family size evolution.

Within-genome signatures of gene-family size change

During a niche shift, new selective pressures can leave footprints in the genomes of evolving lineages; such signatures of positive selection are well described for individual loci (Nielsen et al. 2005; Vitti et al. 2013). Similarly, strong selection for increases or decreases in the size of a particular gene family should also leave characteristic genomic footprints. We argue that these footprints include monophyletic groups of closely related paralogs in gene-family trees (from the selective maintenance of novel duplicates), genomic clustering (when novel genes arise via unequal crossing over), evidence of positive selection among paralogs (given selection for sub- or neofunctionalization), and high rates of pseudogenization (from the selective maintenance of loss-of-function mutations). Of the environmentally responsive gene families we evaluated, none exhibited patterns consistent with selection for a decrease in gene family size. By contrast, at least three families had characteristics consistent with selection for an increase in gene family size. Two of these families, GR and CYP3, were highly clustered in the genome

360 and exhibited evidence of positive selection, making these especially promising candidates for expansions
361 related to a novel coniferous host. Additionally, although the AMP family lacked evidence of positive
362 selection, its unusually clustered distribution in the *Neodiprion* genome could be related to selection for
363 increased dosage of a conserved protein function (Perry et al. 2007). Below we discuss the functions of
364 these three candidate families in more detail.

365 Shifts to pine feeding likely involved changes in the detection of and response to pine-specific
366 cues. Intriguingly, the one GR clade with evidence of positive selection—GR clade 3—is an expansion of
367 six paralogs (one is pseudogenized) orthologous to *DmGR66a*, a bitter receptor specifically for caffeine
368 (Moon et al. 2006). However, orthologs were not found for *DmGR93a* (Lee et al. 2009) and *DmGr33a*
369 (Moon et al. 2009), coreceptors possibly required for caffeine detection. Nevertheless, honeybees, which
370 also lack clear orthologs to these putative coreceptors (Wanner and Robertson 2008), can detect and even
371 prefer low concentrations of caffeine and nicotine (Singaravelan et al. 2005, but see de Brito Sanchez
372 2011). Although pines do not contain caffeine, they do synthesize alkaloids that could confer some
373 bitterness (Mumm and Hilker 2006). Thus, despite lacking caffeine coreceptor orthologs, members of GR
374 clade 3 may still be involved in the detection of pine-specific bitter compounds. Duplications of putative
375 bitter GRs are documented in other host-specialized insects, such as *Heliconius, Danaus,* and *Bombyx*
376 butterflies (Wanner and Robertson 2008; Briscoe et al. 2013). Our sawfly-specific GR expansion, coupled
377 with the finding that GR family size is associated with diet (see below), lends support to the hypothesis
378 that expansions of GR bitter receptors repeatedly contribute to changes in oviposition and feeding
379 behaviors in plant-feeding insects.

380 Because pines contain toxic components like terpenoids and phenolics, detoxifying gene families
381 are also promising candidates for pine adaptation. The mountain pine beetle (*Dendroctonus ponderosae*),
382 feeds on pine bark and wood and has gene "blooms" (species-specific gene gains) in the CYP3 and CYP4
383 clans (Keeling et al. 2013). Similarly, in *N. lecontei*, the CYP family had five blooms (Figure S4a): four
384 CYP3 and one CYP4. CYP3 blooms are also found in wood-feeding insects that do not use pine, such as
385 the emerald ash borer (*Agrilus planipennis*) (David Nelson, unpublished data) and the Asian longhorned
386 beetle (*Anoplophora glabripennis*) (McKenna et al. 2016). Notably, *N. lecontei* larvae frequently ingest
387 pine bark in addition to pine needles (Wilson 1992), suggesting that CYP3 may expand predictably in
388 wood feeders. Additionally, one of the two *Neodiprion*-specific CYP3 clades with evidence of positive
389 selection (Table 3) belongs to the CYP6 subfamily, which is linked to host plant adaptation in several
390 insect taxa (Li et al. 2003; Li et al. 2007; Feyereisen 2012; Mittapelly et al. 2019).

391 Because pine resin has antimicrobial (Himejima et al. 1992; Cowan 1999; Gershenzon and
392 Dudareva 2007) and fungicidal properties (Grayer and Harborne 1994), we hypothesized that *N. lecontei*
393 co-opted these compounds for its own defense, leading to relaxed selection on genes involved in

394 immunity and a reduced innate immune response. In other Hymenoptera, honeybees (*Apis mellifera*)
395 exposed to plant resin have reduced expression of immune-related genes (Simone et al. 2009) and wood
396 ants (*Formica paralugubris*) that use conifer resin as building material have slightly reduced inducible
397 immune system activity and nests with lower bacterial and fungal loads (Castella et al. 2008). In Diptera,
398 AMP loss is associated with herbivorous lineages that live within host tissue, a more sterile habitat than is
399 experienced by most dipterans (Hanson et al. 2019). Unexpectedly, we found a large species-specific
400 clade of *Hisnavicin*-like AMPs in *Neodiprion*. Although additional data are needed to confirm that
401 *Hisnavicin* orthologs act as AMPs in *N. lecontei*, one possible explanation for this putatively adaptive
402 expansion that lacked an accompanying change in non-synonymous substitution rate is that having large
403 numbers of *Hisnavicin*-like AMPs confers protection against pathogens unique to pine trees. That said,
404 our data do not rule out adaptive AMP loss. For example, *N. lecontei* lacks a clear *Defensin* ortholog, a
405 gene present in all dipterans tested to date (Hanson et al. 2019).

406 *Limitations of within-genome analyses*

407 One benefit to studying adaptive expansions/contractions within a single taxon is that gene
408 families have likely experienced similar demographic histories, which can also impact gene birth and
409 death rates. That said, each of our within-genome signatures of selection has limitations that should be
410 revisited with additional data. First, our analysis of genomic clustering does not account for local
411 recombination rate variation, which correlates with tandem array size in several taxa (Gaut et al. 2007). A
412 fine-scale recombination rate map, coupled with clustering analyses for many additional gene families,
413 would more rigorously test the extent to which individual gene family clustering deviates from the
414 genome-wide relationship between recombination rate and tandem array size.

415 Second, a lack of comparable data from other Eusymphyta meant that our gene family
416 phylogenies lacked orthologues from closely related sawfly taxa. Thus, the "*Neodiprion*-specific" clades
417 may not be unique to pine-feeding sawflies. If these paralogs were present prior to the shift to pine hosts,
418 this would not support a scenario in which new duplicates were selectively maintained in the novel niche.
419 Signatures of positive selection may still be related to pine adaptation but would indicate selection on
420 preexisting loci rather than selection favoring gene family expansion.

421 Third, signatures of adaptive gene family expansions and contractions may be ephemeral, and the
422 shift to pine use could have occurred too long ago to detect these signatures in *N. lecontei*. For example,
423 in *Drosophila*, pseudogenes have an estimated half-life of ~14.3 million years (Petrov et al. 1996; Petrov
424 and Hartl 1997, 1998). If the rate of gene decay is similar in *Neodiprion*, then pseudogenes that formed
425 after a shift to pine (up to 60 mya) may no longer be detectable in the genomes of extant sawflies.
426 Likewise, gene clustering patterns are likely to change over time from chromosomal rearrangements and
427 additional gene duplications and deletions. To investigate how the number and position of genes in these

428 focal families has changed over time, high quality gene annotations for diprionids and many additional
429 sawfly outgroups are needed. Fortunately, even if footprints of recent gene family size changes are too
430 ephemeral to be detected in most taxa, consistent relationships with ecology should still be detectable
431 given sufficient sampling of taxa differing in ecological traits of interest.

432 **Ecological correlates of gene family size among hymenopteran taxa**

433 The largest insect OR gene families are in eusocial Hymenoptera, leading to the hypothesis that
434 OR family size expansions were favored in these lineages because they facilitate complex chemical
435 communication (Robertson and Wanner 2006; LeBoeuf et al. 2013; Zhou et al. 2015). To date, evidence
436 in support of this hypothesis has been mixed (e.g., (Roux et al. 2014; Brand and Ramírez 2017).
437 Consistent with the OR-eusociality hypothesis, we found that, on average, eusocial hymenopterans had
438 larger OR families than non-eusocial hymenopterans. However, it is likely that eusocial taxa differ from
439 non-eusocial taxa in many other aspects of their ecology that should also impact OR evolution. Indeed,
440 we found that herbivorous hymenopterans tended to have fewer OR genes than non-herbivores.

441 Whereas all eusocial hymenopterans had relatively large OR families, some eusocial
442 hymenopterans had relatively small GR families (Figures 4, 5; (Zhou et al. 2015)). To explain the
443 strikingly small set of GR genes in honeybee, Wanner and Robinson (2006) proposed that a stable hive
444 environment and a mutualistic relationship with flowering plants resulted in a lack of selection for GR
445 expansions. Intriguingly, our data indicate that among hymenopterans, GR family size is associated with
446 diet, but not eusociality. Like the ORs, GR gene family size tends to be smaller in herbivores than in non-
447 herbivorous taxa, regardless of social behavior. The directionality of this change, however, is unclear: do
448 shifts to plant diets favor reductions in GR families, do shifts to non-plant diets favor GR expansions, or
449 is it both? Answering this question will require characterizing GR families across many independent
450 transitions to and from herbivory, as well as polarizing directions of change (i.e., distinguishing GR gains
451 from GR losses). Fortunately, there are many such diet transitions across diverse clades of insects (Wiens
452 et al. 2015).

453 Unlike sociality and diet, ecological specialization was not associated with gene family size in
454 any of the five gene families we evaluated. This result was unexpected because specialization-associated
455 reductions in gene family size are documented in diverse taxa and multiple gene families, including the
456 families examined here (McBride 2007; Smadja et al. 2009; Cao et al. 2014; Goldman-Huertas et al.
457 2015; Suzuki et al. 2018). One explanation for the lack of association between gene family sizes and
458 specialization in our data is that our "generalist" and "specialist" categories are not meaningful across
459 diverse diets (Forister et al. 2012). Additionally, within a particular diet, the degree of specialization may
460 be highly labile, with rapid fluctuations that are not captured in our broad, order-wide comparison.
461 Indeed, previous studies that reported correlations between gene family size and ecological specialization

462 focused on closely related species. Thus, to fully understand how changes in ecology shape gene family
463 evolution, it will be necessary to evaluate ecological correlates of gene family size at multiple time scales
464 of taxonomic divergence.

465 Compared to ORs and GRs, our other focal gene families had far less manual annotation data
466 available for analysis. This may explain, in part, why we did not detect strong ecological correlates for the
467 other gene families. It is also possible that by focusing on the sizes of entire gene families, we missed
468 relevant signals in particular subfamilies (Hahn et al. 2007). For example, as noted above, expansions of
469 CYP3 and CYP4 subfamilies are associated with wood-feeding insects and CYP3 clan subfamilies were
470 also linked to detoxification in honey bee (Berenbaum and Johnson 2015; Johnson et al. 2018). However,
471 we did not detect any correlations between ecology and CYP clan sizes. Despite these limitations, we did
472 uncover hints that AMP gene family size may be larger in non-eusocial lineages. If eusocial taxa tend to
473 inhabit more sterile environments (nests and hives) than non-eusocial taxa, this finding is consistent with
474 associations between habitat and AMP loss reported in dipterans (Hanson et al. 2019). Given that AMPs
475 were also implicated in our within-genome analysis, immune-related genes are especially promising
476 candidates for future manual annotation projects.

477 *Limitations of among-taxon analyses*

478 Comparative analysis is a powerful approach for evaluating the repeatability and predictability of
479 evolutionary outcomes. Although our comparison of candidate gene family sizes among ecologically
480 diverse hymenopterans hints at intriguing relationships between ecology and gene family size, it also had
481 several limitations that should be revisited in future work. First, because several taxa in our manual
482 annotation dataset are missing from published hymenopteran phylogenies (Peters et al. 2017), we were
483 unable to correct for phylogenetic non-independence and polarize gene gain/loss (e.g., as in (Hahn et al.
484 2005; Han et al. 2013) without losing unacceptable amounts of data. Without accounting for similarity in
485 ecology and gene family size due to recent common ancestry, our Type I error rate is likely inflated and
486 p-values should be interpreted with caution. Nevertheless, variation in patterns of association among
487 ecological traits and gene families suggest that phylogeny and ecology are, to some extent, decoupled.

488 The gene annotation and ecological datasets also had limitations. For example, across studies that
489 included manual annotations, we observed a lack of consistency in the methods and criteria for manually
490 curated gene family datasets. The most problematic inconsistency was in the criteria for delineating intact,
491 partial, and pseudogenized gene annotations. “Intact” could mean an exon-by-exon check against closely
492 related orthologs, a minimum amino acid length, or merely the presence of an expected domain.
493 Meanwhile, in reference publications, the number of pseudogenized and partial annotations were not
494 always reported or were conflated. This is in addition to variation in the methods used to search for genes.
495 Inconsistency in annotation methods and criteria across studies may introduce taxon-specific biases

496 unrelated to ecology. Regarding ecology, categorizations are somewhat subjective. For example, this
497 study and Rane et al. (2016) classified bees as generalists since they collect nectar and pollen from
498 multiple plant families (we defined specialization as the use of a single taxonomic family). But Johnson et
499 al. (2018) classified bees as specialists as their diet consists of only nectar and pollen.
500 Finally, our attempts to correlate the size of different gene families with ecology suffered from
501 sampling biases in which species had genome assemblies and which gene families were manually
502 annotated. Species skewed heavily towards *Drosophila* and apocritan Hymenoptera, and annotations
503 toward the OR and CYP families (Table S8). To evaluate ecological correlates of gene family expansions
504 and contractions, it is essential to expand both the taxonomic breadth and depth of annotation sampling.
505 Taxa that capture independent ecological transitions (e.g., between herbivory and other diets) would be
506 especially useful, as would replicated groups of closely related species that vary in ecological axes of
507 interest (e.g., specialization or social behavior). By systematically sampling different ranges of divergence
508 times, we can evaluate the extent to which the tempo of gene family size change varies across different
509 gene families. To do so, however, will require high quality, manually curated datasets produced using
510 consistent methods and standards for many different environmentally responsive gene families.
511

512 ## Conclusions

513 Gene families that mediate ecological interactions may predictably expand and contract in
514 response to changing selection pressures. These adaptive changes in gene family size should leave
515 detectable genomic footprints in recent niche colonists and across taxa with convergent niche shifts.
516 Consistent with these predictions, (1) our analysis of gene family evolution in a derived pine feeder
517 suggests that expansions of GRs, CYP3s, and AMPs may have accompanied pine adaptation, and (2) our
518 comparison among ecologically diverse hymenopterans links two of these families to variation in diet
519 (GR) and eusociality (AMP). In the order Hymenoptera, the OR gene family was associated with ecology
520 (eusociality), however, the size of all five candidate gene families was not linked to other ecological axes
521 of variation (specialization/generalization); they were in other comparisons of closely related species
522 (McBride 2007)). Together, these results suggest that the size changes of environmentally responsive
523 gene families vary in both temporal dynamics (shallow vs. deep divergence times) and in ecological
524 drivers. Teasing apart these relationships will require high quality annotation data across diverse gene
525 families, ecologies, and divergence times. For hymenopterans, increased effort in understudied
526 symphytan, parasitoid, and herbivorous taxa would be especially useful for disentangling different axes of
527 ecological variation contributing to gene family size change.
528

529 # Materials and methods

530 **Biological material**

531 To minimize the confounding effects heterozygosity has on genome assembly, we sequenced
532 haploid siblings. Like all Hymenoptera, sawflies have haplodiploid sex determination in which males
533 (haploid genomes) emerge from unfertilized eggs and females (diploid genomes) from fertilized eggs. A
534 virgin female will bear a clutch of all-male offspring with haploid recombinants of the maternal genome.
535 But the individual genomes are not identical, so an assembly derived from a single clutch is akin to a
536 diploid assembly made from a single individual.

537 All insects were reared in custom, climate-controlled environmental chambers (18:6 light cycle,
538 22˚C, 70% RH) on jack pine (*Pinus banksiana*) foliage. Our laboratory line of *N. lecontei* was established
539 from multiple larval colonies collected from a mugo pine (*P. mugo*) in Lexington, Kentucky, USA
540 (37°59'01.6"N 84°30'38.8"W; population ID: RB017). For the transcriptome, adults and larvae were
541 collected from the first lab-reared generation; both were stored at -80˚C. For the genome assembly, the
542 founding population was propagated in the lab for two generations, followed by brother-sister matings for
543 an additional two generations. At this point, a single, virgin, adult female (I2G2-V, 4th generation in the
544 lab) was allowed to lay unfertilized eggs onto jack pine seedlings. The offspring (haploid male brothers
545 from an inbred mother) were reared until the eonymph (prepupal) life stage, at which point they were
546 isolated without food for 24 hours prior to preservation in absolute ethanol at -20˚C. Although eonymphs
547 are non-feeding, they were starved to ensure the gut contents were completely voided.

548 **Sample preparation and sequencing**

549 *Genomic DNA*

550 Whole eonymph bodies were individually frozen inside microcentrifuge tubes with liquid
551 nitrogen and ground with pestles made from 1-mL micropipette tips; the resulting powder was incubated
552 in CTAB buffer supplemented with proteinase K and RNase A. After PCI extraction and ethanol
553 precipitation, the precipitate was dried overnight before being resuspended in TE buffer. DNA integrity
554 was assessed with 0.7% agarose gel, purity was measured with the 260/280 ratio, and concentration was
555 measured with a Quant-iT dsDNA High-Sensitivity fluorescence assay (Thermo Fisher Scientific).

556 The HudsonAlpha Genomic Services Lab (Huntsville, AL, USA) prepared and sequenced the
557 DNA libraries. Two small-insert, barcoded libraries with average fragment sizes of 337 bp and 864 bp
558 were made from a single individual. A 4.6-kbp mate-pair, barcoded library was made from 25 pooled
559 individuals. All individuals were brothers from the same I2G2-V mother. The libraries were sequenced on
560 Illumina HiSeq 2000 with paired-end, 100 bp (PE100) reads: the small-insert libraries each had ¼ of a
561 flow cell lane and the mate-pair library had an entire lane.

562 *mRNA*

563 The RNeasy Mini extraction kit (Qiagen) was used to collect total RNA from adult female body,
564 adult female head, adult male body, adult male head, eonymph body, feeding larval body, and feeding
565 larval head. RNA from eonymph head was extracted but not sequenced due to insufficient yield. Each
566 tissue was represented with one replicate that had equal RNA contributions from eight individuals, except
567 for eonymph body which was comprised of three individuals. RNA integrity and concentration were
568 measured with a 2100 Bioanalyzer (Agilent).
569 The HudsonAlpha Genomic Services Lab (Huntsville, AL, USA) handled library preparation and
570 sequencing. Non-stranded, barcoded libraries were made for each of the seven tissue samples; on average,
571 mRNA was sheared to 200 bp. The libraries were combined and sequenced on an entire flow cell of
572 Illumina MiSeq with PE250 reads in addition to one lane of Illumina HiSeq 2000 with PE50 reads.
573 **Read processing and assembly**
574 *De novo genome assembly*
575 Sequencing reads were chastity-filtered and adaptor-trimmed with fastq-mcf (ea-utils v1.04.803)
576 (Aronesty 2011), and quality-filtered with fastq_quality_filter (FASTX Toolkit v0.0.13.2) (Gordon and
577 Hannon 2019). The 337-bp small-insert reads and the 4.6-kbp mate-pair reads were quality-filtered to
578 retain reads where at least 80% of the bases had a quality score of 20 or higher (parameters: `-q 20 -p`
579 `80`). Due to sequencing quality, the 864-bp small-insert reads were filtered to retain reads where at least
580 70% of the bases had a quality score of 20 or higher (R1) or 60% (R2) (parameters: `-q 20 -p`
581 `60/70`). In situations where only one end of the paired-end reads passed filtering, the passed reads were
582 kept and treated as single-end data. Kmer counting was used to measure read depth before and after
583 filtering (Jellyfish v1.1.11) (Marçais and Kingsford 2011). Finally, reads were screened for sequencing
584 contamination by mapping the reads (BWA v0.7.12-r1039) (Li and Durbin 2009) to reference genomes
585 for *Escherichia coli* (K12 substr. DH10B uid58979), human (v37), loblolly pine (*Pinus taeda*, v0.8), and
586 *Wolbachia* (endosymbiont of Dmel uid57851).
587 The genome was assembled with ALLPATHS-LG (v47417) (Gnerre et al. 2011) using default
588 settings, including a minimum scaffold size of 1000 bp. The error-correction module was run on the reads
589 prior to assembly. After assembly, GapFiller (v1.11) (Boetzer and Pirovano 2012) was used to help close
590 intra-scaffold gaps. Spurious scaffolds were identified with SOAP.coverage (v2.7.7) (Li et al. 2009):
591 reads were mapped to the assembly scaffolds and scaffolds with a read depth < 15 and nucleotide
592 percentage < 40 were removed. The completeness of the final assembly was measured with CEGMA
593 (v2.5) (Parra et al. 2007) and BUSCO (v1.22) (Simão et al. 2015) benchmarks. BUSCO was run with the
594 arthropoda-25oct16 database (parameters: `--long`).
595 *De novo transcriptome assembly*

596 For both the PE250 MiSeq and the PE50 HiSeq reads, fastq-mcf (ea-utils v.1.04.803) (Aronesty
597 2011) was used for chastity filtering and Trimmomatic (v0.32) (Bolger et al. 2014) was used to adaptor
598 clip, trim, and quality-filter. The PE250 MiSeq reads were processed with the Trimmomatic parameters
599 `ILLUMINACLIP: 2:15:5, HEADCROP: 10, CROP: 60, MINLEN: 60, AVGQUAL: 25`
600 whereas the PE50 HiSeq reads were processed with `ILLUMINACLIP: 2:15:5, HEADCROP: 15,`
601 `MINLEN: 35, AVGQUAL: 25`. Because the mRNA libraries had an average insert size of 200 bp, the
602 MiSeq reads required extensive adaptor trimming. Reads were screened for contamination as described in
603 *De novo genome assembly*.
604 For each tissue, transcriptomes were assembled with Trinity (r2013_08_14) (Grabherr et al. 2011;
605 Haas et al. 2013) using default settings and the `--jaccard_clip` option. Spurious sequences were
606 identified by mapping the sequencing reads to the assembled transcripts with RSEM (v1.2.18) (Li and
607 Dewey 2011); transcripts with either FPKM or TPM values < 1 were removed. After filtering, the
608 transcriptomes were combined, and duplicate sequences were removed.
609 **Genome size estimation**
610 Flow cytometry was described in (Harper et al. 2016). For this analysis, we used adult
611 males and females from a lab line of *N. lecontei* established from a colony collected in Auburn,
612 GA (33°59'22.4" N, 83°47'44.6"W; population ID: RB027). Briefly, cell nuclei were collected
613 from the heads of 7 individuals (4 female, 3 male) and stained with propidium iodide. Mean
614 fluorescence for each sample was measured with a BD FACSCalibur flow cytometer (BD
615 Biosciences) and compared to two external standards: *Drosophila melanogaster* (adult female
616 heads, 1C = 175 Mbp) and *Gallus gallus domesticus* (CEN singlets from BioSure, Grass Valley,
617 CA, 1C = 1222.5 Mbp). To correct for ploidy differences between haploid males and diploid
618 standards, we multiplied the *N. lecontei* male estimates by 2. To obtain a single size estimate for
619 each N. lecontei sample, we averaged values obtained for the two standards.
620 **Repeat annotation**
621 The N. lecontei genome assembly was masked with a custom repeat library. A lineage-
622 specific de novo repeat library was made with RepeatModeler (v1.0.7) (Smit and Hubley 2008-
623 2015) and combined with the hymenopteran repetitive element database (Nov. 2013) from
624 Repbase (Jurka et al. 2005). The custom library was used by RepeatMasker (v4.0.3) (Smit et al.
625 2013-2015) (parameters: `-cutoff 250 -s -pa 15 -gc 40 -a -poly`) to identify and
626 mask repetitive elements in the genome, including low-complexity DNA and simple repeats.

627 Transposable element (TE) family consensus sequences were identified by rerunning
628 RepeatModeler (Smit and Hubley 2008-2015) on the genome assembly using the "ncbi" search engine.
629 The resulting sequences were provided to RepeatMasker (Smit et al. 2013-2015) as a custom library to
630 locate associated TE copies in the genome (parameters: `-gc 40 -cutoff 250 -gff -gccalc -`
631 `norna -nolow -no_is -poly`). TE families with at least 10 fragments longer than 100 bp were
632 extracted for further analysis.
633 The sequencing reads were mapped to a concatenation of the masked genome and the consensus
634 TE sequences (BWA MEM (parameters: `-M`) (Li and Durbin 2009)). Families that had at least 1x the
635 median coverage to the reference genome for at least 80% of their sequence (to support at least one full
636 insertion found by RepeatModeler) and at least 2x the maximum coverage of the reference genome (to
637 support multiple insertions of the family) were extracted with genomeCoverageBed (BEDtools (Quinlan
638 and Hall 2010)). We attempted to identify the consensus sequences with BLASTN and BLASTX
639 (Altschul et al. 1990) searches against a database of repeat elements, but the only hits were to the lineage-
640 specific elements identified by RepeatModeler. Sequences were also filtered for BLAST hits to rRNA or
641 mitochondrial sequences.
642 We also used dnaPipeTE (Goubert et al. 2015) to identify what proportion of our short reads was
643 composed of repetitive content, we used a random subset of reads corresponding to 1-fold coverage of the
644 genome (331Mb) and took the total for three separate random samplings of reads (parameters: genome
645 size = 331000000 genome coverage = 1 samples number = 3). We then compared this annotation to the
646 RepeatModeler annotation.
647 **Gene and functional annotation**
648 *Automated gene annotation*
649 RNA-Seq data for *N. lecontei* was used to generate training models for gene prediction along with
650 utilization of peptide sequences from other species. PASA (r20130425beta) was used to build a
651 comprehensive transcriptome set from Trinity assembled transcripts along with RNA-Seq read mapping
652 predictions generated from the Tuxedo pipeline. To improve annotation quality, in addition to this *N.*
653 *lecontei* transcriptome, annotated proteins from *Atta cephalotes* (OGSv1.2), *Acromyrmex echinatior*
654 (OGSv3.8), *Apis melifera* (OGSv3.2), *Athalia rosae* (OGSv1.0), and *Nasonia vitripennis* (OGSv1.0) were
655 provided to Maker (2.09) (Cantarel et al. 2008) as evidence for structural gene prediction. Prior to
656 annotation, the genome was masked using a custom repeat database built using RepeatModeler (v1.0.8)
657 and the annotation was run using the *ab initio* gene predictors Augusts, Genemark-ES and snap in
658 addition to the evidence provided. The functions of the predicted protein-coding genes were putatively
659 established with BLASTP alignments (Altschul et al. 1990) to the Swiss-Prot database (accessed 20 Apr
660 12) (Apweiler et al. 2004). In cases of multiple matches, the top-ranked alignment was assigned to the

661 gene annotation. Protein motifs and functional domains within the annotations were also identified with
662 an InterProScan (v5.3.46.0) (Jones et al. 2014) search against the InterPro database with gene ontology
663 and IPR lookup (Finn et al. 2016). For the official gene set (OGS), the Maker annotations were filtered by
664 hits to the reference databases and/or a minimum eAED score of 0.1. A second set of gene annotations
665 was generated with the NCBI GNOMON pipeline (annotation release 100 on Nlec1.0 assembly,
666 GCF_001263575.1) (Souvorov et al. 2010).
667 As the genome was annotated prior to submission to NCBI, we encountered a problem when the
668 NCBI contamination software flagged vector/adaptor sequences for removal; this would disrupt the
669 coordinates provided by Maker. We used a modified version of GAG (Hall et al. 2014) that could accept
670 the flagged coordinates from NCBI to edit the assembly and update annotation coordinates accordingly.
671 *Chemoreceptor genes*
672 The olfactory (OR) and gustatory (GR) receptor genes were manually curated following
673 Robertson et al. (2003, 2006). Amino acid sequences of manually curated chemoreceptor genes from *Apis*
674 *mellifera* (Robertson and Wanner 2006; Smith et al. 2011), *Bombus terrestris* (Sadd et al. 2015) and
675 *Cephus cinctus* (Robertson et al. 2018), *Drosophila melanogaster* (Flybase release FB2017_04), and
676 *Nasonia vitripennis* (Robertson et al. 2010) were used as queries in TBLASTN (v2.2.19) (Altschul et al.
677 1990) searches against the *N. lecontei* draft genome (parameters: `-e 100000 -F F`). Gene models
678 were manually built in TextWrangler (v5.5) (Bare Bones Software), using protein alignment to identify
679 exons and refine the gene structures; alignments were visualized with Clustal X (v2.1) (Larkin et al.
680 2007). The Neural Network Splice Predictor program from the Berkeley *Drosophila* Genome Project was
681 used to help identify intron splice sites (http://www.fruitfly.org/seq_tools/splice.html). New gene models
682 were added to TBLASTN searches and this process continued iteratively until new chemoreceptors were
683 no longer found. The gene models were checked against RNAseq reads from tissue-specific
684 transcriptomes (adult antennae, mouthparts, heads, legs, genitalia, and larval heads (Herrig et al. 2019))
685 and against orthologs in the *N. pinetum* draft genome assembly (NCBI accession GCA_004916985.1).
686 *Odorant binding proteins*
687 Custom scripts were used to identify Maker gene annotations (see *Automated gene annotation*)
688 that contained the classic/6C, Plus-C, Minus-C, or atypical odorant binding protein (OBP) motif (Xu et al.
689 2009). These as well as OBPs from *Apis mellifera* and *Nasonia vitripennis* were used as queries for
690 TBLASTN searches against the *N. lecontei* genome; searches did not yield any new OBPs. All genomic
691 regions identified as potential OBPs were manually curated as described for chemoreceptor genes. After
692 manual annotation, duplicate annotations or genes that lacked OBP motifs were removed.
693 *Cytochrome P450 genes*
694 A broad set of 52 insect CYP genes (covering the diversity of insect CYP families) were

695 searched against the *N. lecontei* genome assembly (E-value cutoff 1e3). Scaffolds with hits were then
696 searched against 8782 known insect CYPs. The top 10 hits were returned (later increased to 15 to recover
697 more sequences) and filtered for duplicates. An alternative search of the NCBI GNOMON predictions
698 ("Neodiprion lecontei[orgn] AND P450 NOT reductase") was also performed and new sequences were
699 added to the dataset. This approach found all the loci identified by the initial search, indicating that the
700 GNOMON annotation tool was able to comprehensively search for CYP sequences. Finally, the candidate
701 *N. lecontei* CYP sequences were manually curated based on comparison to the best BLAST hits.

702 *Immune-related genes*

703 Because of the relative completeness of its immune annotation, *Drosophila melanogaster*
704 immunity genes were used to guide annotation. Reference immune genes from *D. melanogaster* tagged
705 with the gene ontology term "GO:0002376 – Immune system process" were compiled from Flybase
706 (release 6.13). Orthology with *N. lecontei* proteins was assigned initially with reciprocal BLASTP
707 (Altschul et al. 1990) searches (E-value cutoff 1e-10). Reference *D. melanogaster* genes without obvious
708 one-to-one orthologs in *N. lecontei* were examined individually to determine whether closely related
709 paralogs in one or both species interfered with the inference of orthology. If not, they were searched
710 against the *N. lecontei* genome assembly using TBLASTN (Altschul et al. 1990) in an attempt to identify
711 unannotated orthologs.

712 Since antimicrobial peptides (AMP) are unlikely to be conserved between *D. melanogaster* and
713 *N. lecontei*, AMPs from three representative hymenopterans *Apis mellifera* (Danihlík et al. 2015),
714 *Nasonia vitripennis* (Tian et al. 2010), and *Camponotus floridanus* (Ratzka et al. 2012; Zhang and Zhu
715 2012; Gupta et al. 2015) were used for BLAST queries. Furthermore, since AMP copy number is fast
716 evolving, we attempted to find all the *N. lecontei* orthologs of each hymenopteran AMP instead of
717 focusing on one-to-one orthology. Once again, BLASTP searches were performed against the annotated
718 proteins and TBLASTN searches were performed against the assembled genome; the TBLASTN search
719 did not reveal additional AMPs. Putative *N. lecontei* orthologs were reciprocally blasted against the
720 appropriate hymenopteran proteome to assure that the best hits were indeed AMPs.

721 Amino acid and cDNA sequences for all manual annotated genes are available in File S1.

722 **Glomeruli counts**

723 *Antennal lobe histology*

724 Whole heads of adult *N. lecontei* of both sexes were fixed in 2% paraformaldehyde, 2%
725 glutaraldehyde in PBS for 5 days. Heads were rinsed for 40 minutes three times and the brains dissected
726 out in cold PBS. Following blocking with goat serum, brains were permeabilized with 1% Triton X-100
727 in PBS (Electron Microscopy Supply, Fort Washington, PA; PBS-TX), rinsed with 0.1% PBS-TX, and
728 incubated on a shaker at 25°C for three nights in primary antibody (1:500 in 2% goat serum in 0.2% PBS-

729 TX). Monoclonal *Drosophila* synapsin I antibody (SYNORF1, AB_2315426) from the Developmental
730 Studies Hybridoma Bank (catalog 3C11) was used to label synapsin. Subsequently, brains were washed
731 in 0.1% PBS-TX and incubated for two nights in Alexa Fluor 568 (ThermoFisher) goat anti-mouse
732 secondary antibody (1:100 in PBS) in the dark at room temperature on a shaker. After secondary
733 incubation, brains were rinsed with distilled water, dehydrated in increasing concentrations of ethanol,
734 and mounted in custom-made aluminum well slides. Brains were cleared by removing ethanol and
735 replacing it with methyl salicylate. Brains were imaged on an inverted Zeiss 880 Laser Scanning
736 Confocal Microscope with a 20X plan-Apochromat 20x 0.8 aperture objective and optically sectioned in
737 the horizontal plane at 3-micron intervals.

738 *Glomeruli segmentation*

739 Whole-brain images of one female and one male were manually segmented using the TrakEM2
740 software package in ImageJ (Cardona et al. 2012; Schindelin et al. 2012). Individual glomeruli were
741 traced in both brain hemispheres. Glomeruli near the center of the antennal lobe can be difficult to
742 distinguish, meaning counts are biased toward fewer glomeruli and the largest number of glomeruli
743 confidently detected represents a minimum of the number of expected glomeruli. Male *Neodiprion* have a
744 collection of smaller synaptic clusters in their antennal lobe (Dacks and Nighorn 2011), but the functional
745 significance of this anatomy is not known. There are more than 50 of these smaller synaptic clusters and
746 we suspect they do not represent the traditional one-to-one OR-to-glomerulus organization. Therefore,
747 these structures were not included in counts. Male glomeruli number may be lower if particular OSNs
748 contribute to these clusters instead of forming traditional glomeruli.

749 **Within-genome signatures of adaptive expansions and contractions**

750 *Clustering and pseudogene analyses*

751 To evaluate the extent to which members of our five focal gene families were located in tandem
752 arrays, we placed our annotated genes on a linkage-map anchored version of the *N. lecontei* genome
753 assembly described in Linnen et al. 2018. We considered genes to be clustered if they were located within
754 a genomic region of 20(n - 1) kilobases, where n is the number of genes in the cluster under
755 consideration. This criterion was chosen based on average gene densities in *Nasonia* (Niehuis et al. 2010)
756 and clustering criteria described *Drosophila* (Vieira et al. 2007). For scaffolds that could not be placed on
757 linkage groups, we evaluated clustering only if genes were more than 20 kb from either scaffold end.

758 To evaluate whether the five focal gene families differed in (1) the proportion of genes found in
759 clusters of two or more or (2) the proportion of pseudogenized genes, we performed Fisher's exact tests in
760 R v3.5.0 ("fisher.test" function) (R-Core-Team 2018). For significant Fisher's exact tests, we performed
761 additional posthoc tests using the "fisher.multcomp" function (from R package RVAideMemoire v. 0.9-
762 72) with FDR correction (Benjamini-Hochberg method) for multiple comparisons.

763 *Identification of* Neodiprion-*specific clades and tests of positive selection*

764 First, we identified clades unique to *N. lecontei*. For each gene family, a multi-species, amino
765 acid phylogeny was constructed with manually curated annotations from *N. lecontei*, select Hymenoptera,
766 and *D. melanogaster*. Sequences were size filtered (350≥ for GR, OR, CYP; 100≥ for histnavicin and
767 OBP), but pseudogenes and partial annotations that met the length requirement were retained. MAFFT
768 alignments (v7.305b) (Katoh et al. 2002) (parameters: `--maxiterate 1000 -localpair`) were
769 visually inspected to remove sequences with large alignment gaps, and sites with more than 20% gaps
770 were removed with trimAl (v1.4.rev15 build[2013-12-17]) (Capella-Gutiérrez et al. 2009) (parameters:
771 `-gapthreshold 0.8`). Maximum likelihood phylogenies were made in RAxML (v8.2.4) (Stamatakis
772 2014) (parameters: `-f a -x 12345 -p 12345 -# autoMRE`) using protein substitution models
773 chosen from ProtTest3 (v3.4.2) (Abascal et al. 2010; Darriba et al. 2011).

774 *Neodiprion*-specific clades were defined as those with at least five *N. lecontei* genes (not
775 including partial and pseudogenes) and a bootstrap score ≥70 (Engsontia et al. 2015). Second, the clades
776 were confirmed with cDNA phylogenies for each *N. lecontei* gene family. Amino acid sequences were
777 aligned as above, however, after alignment TranslatorX (Abascal et al. 2010) was used to map cDNA
778 sequences to the amino acid alignment. After trimming, the cDNA alignments were passed to RAxML to
779 construct maximum likelihood gene family trees with the nucleotide substitution model `-m GTRGAMMA`.

780 Site tests were conducted with codeml (part of the PAML package (PAML v4.9e) (Yang 2007))
781 using the cDNA phylogenies and sequences as inputs. For each *Neodiprion*-specific clade, the gene
782 family cDNA phylogeny was pruned to remove all branches except for that clade. Codeml models M7,
783 M8, and M8a were fitted to the cDNA sequence and phylogeny data. Likelihood-ratio tests were
784 performed for the nested models M7-M8 (null model M7 that equally distributes amino acid sites across
785 10 classes of ω parameter values (p, q) against alternative model M8 that has an 11th class for positively
786 selected sites) and M8-M8a (null model M8a that has 11 classes and does not allow positive selection
787 against alternative model M8). Bonferroni correction was applied to the likelihood-ratio test probability
788 values; each value was multiplied by two since two tests that used M8 as the alternative model were
789 performed on each clade. For clades with significant likelihood-ratio tests, sites under selection were
790 identified by looking at the Bayes Empirical Bayes analysis within the alternative models.

791 For branch tests, the cDNA phylogenies for each *N. lecontei* gene family were used to compare
792 the lineage-specific clade to the rest of the gene family. To determine if the foreground branch dN/dS
793 (i.e., the branch with the species-specific expansion) was significantly different from the background (i.e.,
794 the rest of the gene family), in codeml we ran a two-ratio model (Model=2, fix_omega=0) and a one-ratio
795 model (Model=0, fix_omega=0) for that clade and performed a likelihood-ratio test comparing the two

models. To determine if the foreground branch is under positive selection (dN/dS>1), we performed a likelihood-ratio test comparing the two-ratio model to a neutral model (fix_omega=1).

Ecological correlates of gene family size among insects

All the insect genome assembly projects we could find (published and unpublished) were searched for manually curated OR, GR, OBP, CYP, and AMP gene annotations. If fasta sequence files were available, the number of intact, partial, and pseudogenized genes was determined by gene names (e.g., labels with "pse" or "partial") and compared to values reported in the publication. Otherwise, we relied on reported values. If gene family size was reported but not broken down into intact, partial, and pseudogenized, and sequence files were unavailable, we assumed that the reported number referred to intact genes. Splice variants were not included in the gene count. It is important to note that different authors likely used different criteria for these categories.

Only putatively functional (intact) gene were used in gene family size comparisons. Species were classified according to taxonomic order, diet type, dietary specialization, and sociality. An order needed at least two species to be included. Specialization was defined as the use of a single taxonomic family and only referred to the realized diet niche, ignoring reports of feeding under laboratory conditions. If a species had a preferred host or both specialist and generalist life stages, it was classified as specialist. Comparisons were made in R (v3.5.0) where species were grouped by the different classifications.

Because both gene family size and ecology are likely to correlate with phylogeny, the ideal approach to identifying ecological correlates of gene family evolution is to use statistical methods that account for phylogenetic relationships (Hahn et al. 2005; De Bie et al. 2006; Han et al. 2013). Unfortunately, a lack of overlap between species with manual annotations for our focal gene families and species included in published hymenopteran genomes precluded us from such an analysis without a substantial loss of sample size. Therefore, as a first step to evaluating ecological correlates of gene family size, we used non-parametric tests to determine whether gene family size differed among taxa. For sociality and specialization, we used two-tailed Wilcoxon rank-sum tests ("wilcox.exact" function in the exactRankTests v0.8-30 package). For taxonomic order and diet, both of which have more than two categories, we used Kruskal-Wallis tests ("kruskal.test" function) followed by Dunn's post-hoc tests of multiple comparisons ("dunnTest" function in the FSA v0.8.23 package).

Acknowledgements

We thank Linnen lab members for insect collection, insect reading, and reading earlier manuscripts; Jeramiah Smith, Erin Scully, and Romain Studer for advice. We are especially grateful to Hugh Robertson for his guidance on manual chemoreceptor gene annotation. This work was supported by the University of Kentucky Center for Computational Sciences and the Lipscomb High Performance

830 Computing Cluster, the United States Department of Agriculture National Institute of Food and
831 Agriculture (2016-67014-2475; CRL), the Kentucky Science and Engineering Foundation (KSEF-3492-
832 RDE-019; CRL), and the University of Kentucky (Lyman T. Johnson Fellowship; KV).
833

834 Data availability

835 The genome assembly, official gene set (OGS), and transcriptome described in this paper (v1
836 versions) can be found at https://i5k.nal.usda.gov/neodiprion-lecontei
837 On GenBank (NCBI), the genome assembly is labeled whole genome shotgun sequencing project
838 accession PRJNA28045 and the genomic sequencing reads are RefSeq accession PRJNA312506. The
839 transcriptome is transcriptome shotgun assembly accession GEDM00000000; this is a combined
840 transcriptome of all seven tissue types. The mRNA sequencing reads for each tissue type was submitted
841 separately under BioSample and short read archive accessions SAMN04302192 (adult female head),
842 SAMN04302193 (adult female body), SAMN04302194 (adult male head), SAMN04302195 (adult male
843 body), SAMN04302196 (feeding larval head), SAMN04302197 (feeding larval body), and
844 SAMN04302198 (eonymph body). The predicted gene annotations on NCBI are from Gnomon, the NCBI
845 annotation pipeline, and were not described in this paper. Finally, the clustering analysis was based on a
846 linkage-map anchored version of the genome assembly described in Linnen et al. 2018. This anchored
847 assembly is denoted as v1.1 in NCBI and the *N. lecontei* i5k Workspace@NAL (USDA).
848

849 References

850 Abascal F, Zardoya R, Telford MJ. 2010. TranslatorX: multiple alignment of nucleotide sequences guided
851 by amino acid translations. Nucleic Acids Res 38(suppl 2):W7-W13.
852 Akhunov ED, Goodyear AW, Geng S, Qi L-L, Echalier B, Gill BS, Gustafson JP, et al. 2003. The
853 organization and rate of evolution of wheat genomes are correlated with recombination rates along
854 chromosome arms. Genome Res 13(5):753-763.
855 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol
856 Biol 215(3):403-410.
857 Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, et al. 2004. UniProt:
858 the universal protein knowledgebase. Nucleic Acids Res 32(suppl 1):D115-D119.
859 Arguello JR, Cardoso-Moreira M, Grenier JK, Gottipati S, Clark AG, Benton R. 2016. Extensive local
860 adaptation within the chemosensory system following Drosophila melanogaster's global expansion.
861 Nat Commun 7(1):1-12.
862 Aronesty E. 2011. ea-utils : "Command-line tools for processing biological sequencing data".

863 Berenbaum MR. 2002. Postgenomic chemical ecology: from genetic code to ecological interactions. J
864 Chem Ecol 28(5):873-896.
865 Berenbaum MR, Johnson RM. 2015. Xenobiotic detoxification pathways in honey bees. Curr Opin Insect
866 Sci 10:51-58.
867 Boetzer M, Pirovano W. 2012. Toward almost closed genomes with GapFiller. Genome Biol 13(6):R56.
868 Boevé J-L, Blank SM, Meijer G, Nyman T. 2013. Invertebrate and avian predators as drivers of chemical
869 defensive strategies in tenthredinid sawflies. BMC Evol Biol 13(1):198.
870 Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data.
871 Bioinformatics 30(15):2114-2120.
872 Brand P, Ramírez SR. 2017. The evolutionary dynamics of the odorant receptor gene family in
873 corbiculate bees. Genome Biol Evol 9(8):2023-2036.
874 Briscoe AD, Macias-Munoz A, Kozak KM, Walters JR, Yuan F, Jamie GA, Martin SH, et al. 2013.
875 Female behaviour drives expression and evolution of gustatory receptors in butterflies. PLoS Genet
876 9(7):e1003620.
877 Calla B, Noble K, Johnson RM, Walden KKO, Schuler MA, Robertson HM, Berenbaum MR. 2017.
878 Cytochrome P450 diversification and hostplant utilization patterns in specialist and generalist moths:
879 Birth, death and adaptation. Mol Ecol 26(21):6021-6035.
880 Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, Holt C, Alvarado AS, Yandell M. 2008.
881 MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes.
882 Genome Res 18(1):188-196.
883 Cao D, Liu Y, Walker WB, Li J, Wang G. 2014. Molecular characterization of the Aphis gossypii
884 olfactory receptor gene families. PLoS One 9(6):e101187.
885 Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment
886 trimming in large-scale phylogenetic analyses. Bioinformatics 25(15):1972-1973.
887 Cardona A, Saalfeld S, Schindelin J, Arganda-Carreras I, Preibisch S, Longair M, Tomancak P,
888 Hartenstein V, Douglas RJ. 2012. TrakEM2 software for neural circuit reconstruction. PLoS One
889 7(6):e38011.
890 Castella G, Chapuisat M, Moret Y, Christe P. 2008. The presence of conifer resin decreases the use of the
891 immune system in wood ants. Ecological Entomology 33(3):408-412.
892 Comeault AA, Serrato-Capuchina A, Turissini DA, McLaughlin PJ, David JR, Matute DR. 2017. A
893 nonrandom subset of olfactory genes is associated with host preference in the fruit fly Drosophila
894 orena. Evolution Letters 1(2):73-85.
895 Consortium iK. 2013. The i5K Initiative: advancing arthropod genomics for knowledge, human health,
896 agriculture, and the environment. J Hered, 104(5):595-600.

897 Couto A, Alenius M, Dickson BJ. 2005. Molecular, anatomical, and functional organization of the
898 Drosophila olfactory system. Curr Biol 15(17):1535-1547.
899 Cowan MM. 1999. Plant products as antimicrobial agents. Clin Microbiol Rev 12(4):564-582.
900 Dacks AM, Nighorn AJ. 2011. The organization of the antennal lobe correlates not only with
901 phylogenetic relationship, but also life history: a basal hymenopteran as exemplar. Chem Senses
902 36(2):209-220.
903 Dahanukar A, Foster K, Carlson JR. 2001. A Gr receptor is required for response to the sugar trehalose in
904 taste neurons of Drosophila. Nat Neurosci 4(12):1182-1186.
905 Danihlík J, Aronstein K, Petřivalský M. 2015. Antimicrobial peptides: a key component of honey bee
906 innate immunity: Physiology, biochemistry, and chemical ecology. Journal of Apicultural Research
907 54(2):123-136.
908 Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein
909 evolution. Bioinformatics 27(8):1164-1165.
910 Dayhoff MO. 1976. The origin and evolution of protein superfamilies Federation Proceedings
911 35(10):2132-2138.
912 De Bie T, Cristianini N, Demuth JP, Hahn MW. 2006. CAFE: a computational tool for the study of gene
913 family evolution. Bioinformatics 22(10):1269-1271.
914 de Brito Sanchez MG. 2011. Taste perception in honey bees. Chem Senses 36(8): 675-692.
915 Demuth JP, Hahn MW. 2009. The life and death of gene families. Bioessays 31(1):29-39.
916 Després L, David J-P, Gallet C. 2007. The evolutionary ecology of insect resistance to plant chemicals.
917 Trends Ecol Evol 22(6):298-307.
918 Dobler S, Dalla S, Wagschal V, Agrawal AA. 2012. Community-wide convergent evolution in insect
919 adaptation to toxic cardenolides by substitutions in the Na, K-ATPase. Proc Natl Acad Sci USA
920 109(32):13040-13045.
921 Duffy JE, Macdonald KS. 2010. Kin structure, ecology and the evolution of social organization in shrimp:
922 a comparative analysis. Proc Royal Soc B 277(1681):575-584.
923 Engsontia P, Sangket U, Robertson HM, Satasook C. 2015. Diversification of the ant odorant receptor
924 gene family and positive selection on candidate cuticular hydrocarbon receptors. BMC Res Notes
925 8(1):380.
926 Evans JD, Aronstein K, Chen YP, Hetru C, Imler JL, Jiang H, Kanost M, Thompson GJ, Zou Z, Hultmark
927 D. 2006. Immune pathways and defence mechanisms in honey bees Apis mellifera. Insect Mol Biol
928 15(5):645-656.
929 Faulkes CG, Bennett NC, Bruford MW, O'Brien HP, Aguilar GH, Jarvis JU. 1997. Ecological constraints
930 drive social evolution in the African mole–rats. Proc Royal Soc B 264(1388):1619-1627.

931 Feyereisen R. 2011. Arthropod CYPomes illustrate the tempo and mode in P450 evolution. Biochimica et
932 Biophysica Acta (BBA)-Proteins and Proteomics 1814(1):19-28.
933 Feyereisen R. 2012. Insect CYP genes and P450 enzymes. In. Insect Mol Biol and Biochemistry:
934 Academic Press. p. 236-316.
935 Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, et al. 2016. The Pfam
936 protein families database: towards a more sustainable future. Nucleic Acids Res 44(D1):D279-D285.
937 Fischman BJ, Woodard SH, Robinson GE. 2011. Molecular evolutionary analyses of insect societies.
938 Proc Natl Acad Sci USA 108(suppl 2):10847-10854.
939 Fishilevich E, Vosshall LB. 2005. Genetic and functional subdivision of the Drosophila antennal lobe.
940 Curr Biol 15(17):1548-1553.
941 Forister ML, Dyer LA, Singer MS, Stireman III JO, Lill JT. 2012. Revisiting the evolution of ecological
942 specialization, with emphasis on insect–plant interactions. Ecology 93(5):981-991.
943 Gao Q, Yuan B, Chess A. 2000. Convergent projections of Drosophila olfactory neurons to specific
944 glomeruli in the antennal lobe. Nat Neurosci 3(8):780-785.
945 Gardiner A, Barker D, Butlin RK, Jordan WC, Ritchie MG. 2008. Drosophila chemoreceptor gene
946 evolution: selection, specialization and genome size. Mol Ecol 17(7):1648-1657.
947 Gaut BS, Wright SI, Rizzon C, Dvorak J, Anderson LK. 2007. Recombination: an underappreciated factor
948 in the evolution of plant genomes. Nat Rev Genet 8(1):77-84.
949 Gershenzon J, Dudareva N. 2007. The function of terpene natural products in the natural world. Nat
950 Chem Biol 3(7):408-414.
951 Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, et al. 2011. High-
952 quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl
953 Acad Sci USA 108(4):1513-1518.
954 Goldman-Huertas B, Mitchell RF, Lapoint RT, Faucher CP, Hildebrand JG, Whiteman NK. 2015.
955 Evolution of herbivory in Drosophilidae linked to loss of behaviors, antennal responses, odorant
956 receptors, and ancestral diet. Proc Natl Acad Sci USA 112(10):3026-3031.
957 Good RT, Gramzow L, Battlay P, Sztal T, Batterham P, Robin C. 2014. The molecular evolution of
958 cytochrome P450 genes within and between Drosophila species. Genome Biol Evol 6(5):1118-1134.
959 Gordon A, Hannon GJ. 2019. "Fastx-toolkit" FASTQ/A short-reads preprocessing tools (unpublished).
960 Goubert C, Modolo L, Vieira C, Claire ValienteMoro, Mavingui P, Boulesteix M. 2015. De novo
961 assembly and annotation of the Asian tiger mosquito (Aedes albopictus) repeatome with dnaPipeTE
962 from raw genomic reads and comparative analysis with the yellow fever mosquito (Aedes aegypti).
963 Genome Biol Evol 7(4):1192-1205.

964 Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, et al. 2011. Full-
965 length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol
966 29(7):644-652.
967 Grayer RJ, Harborne JB. 1994. A survey of antifungal compounds from higher plants, 1982–1993.
968 Phytochemistry 37(1):19-42.
969 Gross JB, Borowsky R, Tabin CJ. 2009. A novel role for Mc1r in the parallel evolution of depigmentation
970 in independent populations of the cavefish Astyanax mexicanus. PLoS Genet 5(1):e1000326.
971 Guittard E, Blais C, Maria A, Parvy J-P, Pasricha S, Lumb C, Lafont R, Daborn PJ, Dauphin-Villemant
972 C. 2011. CYP18A1, a key enzyme of Drosophila steroid hormone inactivation, is essential for
973 metamorphosis. Dev Bio l349(1):35-45.
974 Gupta SK, Kupper M, Ratzka C, Feldhaar H, Vilcinskas A, Gross R, Dandekar T, Förster F. 2015.
975 Scrutinizing the immune defence inventory of Camponotus floridanus applying total transcriptome
976 sequencing. BMC Genom 16(1):1-21.
977 Ha E-M, Oh C-T, Bae YS, Lee W-J. 2005. A direct role for dual oxidase in Drosophila gut immunity.
978 Science 310(5749):847-850.
979 Haas BJ, Papanicolaou A, Yassour M, Grabherr MG, Blood PD, Bowden J, Couger MB, et al. 2013. De
980 novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference
981 generation and analysis. Nat Protoc 8(8):1494-1512.
982 Hahn MW, de Bie T, Stajich JE, Nguyen C, Cristianini N. 2005. Estimating the tempo and mode of gene
983 family evolution from comparative genomic data. Genome Res 15(8):1153-1160.
984 Hahn MW, Han MV, Han S-G. 2007. Gene family evolution across 12 Drosophila genomes. PLoS Genet
985 3(11):e197.
986 Hall B, DeRego T, Geib SM. 2014. GAG: the genome annotation generator (version 1.0).
987 Han MV, Thomas GW, Lugo-Martinez J, Hahn MW. 2013. Estimating gene gain and loss rates in the
988 presence of error in genome assembly and annotation using CAFE 3. Mol Biol Evol 30(8):1987-1997.
989 Hanson MA, Lemaitre B, Unckless RL. 2019. Dynamic evolution of antimicrobial peptides underscores
990 trade-offs between immunity and ecological fitness. Front Immunol 10:2620.
991 Harper KE, Bagley RK, Thompson KL, Linnen CR. 2016. Complementary sex determination, inbreeding
992 depression and inbreeding avoidance in a gregarious sawfly. Heredity 117(5):326-335.
993 Helvig C, Koener JF, Unnithan GC, Feyereisen R. 2004. CYP15A1, the cytochrome P450 that catalyzes
994 epoxidation of methyl farnesoate to juvenile hormone III in cockroach corpora allata. Proc Natl Acad
995 Sci USA 101(12):4024-4029.
996 Herrig DK, Vertacnik KL, Linnen CR. 2019. Testing the adaptive decoupling hypothesis in a
997 hypermetamorphic and sexually dimorphic insect. bioRxiv 2019.12.20.882803.

998 Himejima M, Hobson KR, Otsuka T, Wood DL, Kubo I. 1992. Antimicrobial terpenes from oleoresin of
999 ponderosa pine tree Pinus ponderosa: A defense mechanism against microbial invasion. J Chem Ecol
1000 18(10):1809-1818.
1001 Hughes AL, Nei M. 1992. Maintenance of MHC polymorphism. Nature 335:402–403.
1002 Johnson RM, Harpur BA, Dogantzis KA, Amro Z, Berenbaum MR. 2018. Genomic footprint of evolution
1003 of eusociality in bees: floral food use and CYPome "blooms". Insectes Soc 65(3):445-454.
1004 Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, et al. 2014. InterProScan 5:
1005 genome-scale protein function classification. Bioinformatics 30(9):1236-1240.
1006 Jones WD, Cayirlioglu P, Kadow IG, Vosshall LB. 2007. Two chemosensory receptors together mediate
1007 carbon dioxide detection in Drosophila. Nature 445(7123):86-90.
1008 Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase Update, a
1009 database of eukaryotic repetitive elements. Cytogenet Genome Res 110(1-4):462-467.
1010 Karev GP, Wolf YI, Rzhetsky AY, Berezovskaya FS, Koonin EV. 2002. Birth and death of protein
1011 domains: a simple model of evolution explains power law behavior. BMC Evol Biol 2(1):18.
1012 Keeling CI, Yuen MMS, Liao NY, Docking TR, Chan SK, Taylor GA, Palmquist DL, et al. 2013. Draft
1013 genome of the mountain pine beetle, Dendroctonus ponderosae Hopkins, a major forest pest. Genome
1014 Biol 14(3):R27.
1015 Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, et al.
1016 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23(21):2947-2948.
1017 LeBoeuf AC, Benton R, Keller L. 2013. The molecular basis of social behavior: models, methods and
1018 advances. Curr Opin Neurobiol 23(1):3-10.
1019 Lee K-A, Kim B, Bhin J, Kim DH, You H, Kim E-K, Kim S-H, Ryu J-H, Hwang D, Lee W-J. 2015.
1020 Bacterial uracil modulates Drosophila DUOX-dependent gut immunity via Hedgehog-induced
1021 signaling endosomes. Cell Host Microbe 17(2):191-204.
1022 Lee Y, Moon SJ, Montell C. 2009. Multiple gustatory receptors required for the caffeine response in
1023 Drosophila. Proc Natl Acad Sci USA 106(11):4495-4500.
1024 Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a
1025 reference genome. BMC Bioinform 12(1):323.
1026 Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform.
1027 Bioinformatics 25(14):1754-1760.
1028 Li R, Yu C, Li Y, Lam T-W, Yiu S-M, Kristiansen K, Wang J. 2009. SOAP2: an improved ultrafast tool
1029 for short read alignment. Bioinformatics 25(15):1966-1967.

1030 Li W, Schuler MA, Berenbaum MR. 2003. Diversification of furanocoumarin-metabolizing cytochrome
1031 P450 monooxygenases in two papilionids: specificity and substrate encounter rate. Proc Natl Acad
1032 Sci USA 100(suppl 2):14593-14598.
1033 Li X, Schuler MA, Berenbaum MR. 2007. Molecular mechanisms of metabolic resistance to synthetic and
1034 natural xenobiotics. Annu Rev Entomol 52:231-253.
1035 Linnen CR, O'Quin CT, Shackleford T, Sears CR, Lindstedt C. 2018. Genetic basis of body color and
1036 spotting pattern in redheaded pine sawfly larvae (Neodiprion lecontei). Genetics 209(1):291-305.
1037 Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of
1038 k-mers. Bioinformatics 27(6):764-770.
1039 Martin A, Orgogozo V. 2013. The loci of repeated evolution: a catalog of genetic hotspots of phenotypic
1040 variation. Evolution 67(5):1235-1250.
1041 Matsuo T, Sugaya S, Yasukawa J, Aigaki T, Fuyama Y. 2007. Odorant-binding proteins OBP57d and
1042 OBP57e affect taste perception and host-plant preference in Drosophila sechellia. PLoS Biology
1043 5(5):e118.
1044 Maxwell DE editor. Proceedings of the 10th International Congress of Entomology. 1958 Montreal,
1045 Canada.
1046 McBride CS. 2007. Rapid evolution of smell and taste receptor genes during host specialization in
1047 Drosophila sechellia. Proc Natl Acad Sci USA 104(12):4996-5001.
1048 McBride CS, Arguello RJ. 2007. Five Drosophila genomes reveal nonneutral evolution and the signature
1049 of host specialization in the chemoreceptor superfamily. Genetics 177(3):1395-1416.
1050 McKenna DD, Scully ED, Pauchet Y, Hoover K, Kirsch R, Geib SM, Mitchell RF, et al. 2016. Genome
1051 of the Asian longhorned beetle (Anoplophora glabripennis), a globally significant invasive species,
1052 reveals key functional and evolutionary innovations at the beetle–plant interface. Genome Biol
1053 17:227
1054 McKenzie SK, Fetter-Pruneda I, Ruta V, Kronauer DJC. 2016. Transcriptomics and neuroanatomy of the
1055 clonal raider ant implicate an expanded clade of odorant receptors in chemical communication. Proc
1056 Natl Acad Sci USA 113(49):14091-14096.
1057 Mittapelly P, Bansal R, Michel A. 2019. Differential expression of cytochrome P450 CYP6 genes in the
1058 brown marmorated stink bug, Halyomorpha halys (Hemiptera: Pentatomidae). J Econ Entomol
1059 112(3):1403-1410.
1060 Miyamoto T, Slone J, Song X, Amrein H. 2012. A fructose receptor functions as a nutrient sensor in the
1061 Drosophila brain. Cell 151(5):1113-1125.
1062 Moon SJ, Köttgen M, Jiao Y, Xu H, Montell C. 2006. A taste receptor required for the caffeine response
1063 in vivo. Curr Biol 16(18):1812-1817.

1064 Moon SJ, Lee Y, Jiao Y, Montell C. 2009. A Drosophila gustatory receptor essential for aversive taste
1065 and inhibiting male-to-male courtship. Curr Biol 19(19):1623-1627.
1066 Mumm R, Hilker M. 2006. Direct and indirect chemical defence of pine against folivorous insects. Trends
1067 Plant Sci 11(7):351-358.
1068 Nei M. 2007. The new mutation theory of phenotypic evolution. Proc Natl Acad Sci USA 104(30):12235-
1069 12242.
1070 Nei M, Rooney AP. 2005. Concerted and birth-and-death evolution of multigene families. Annu Rev
1071 Genet 39:121-152.
1072 Niehuis O, Gibson JD, Rosenberg MS, Pannebakker BA, Koevoets T, Judson AK, Desjardins CA, et al.
1073 2010. Recombination and its impact on the genome of the haplodiploid parasitoid wasp Nasonia.
1074 PLoS One 5(1):e8597.
1075 Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. 2005. Genomic scans for
1076 selective sweeps using SNP data. Genome Res 15(11):1566-1575.
1077 Nozawa M, Kawahara Y, Nei M. 2007. Genomic drift and copy number variation of sensory receptor
1078 genes in humans. Proc Natl Acad Sci USA 104(51):20421-20426.
1079 Ohno S. 1970. The enormous diversity in genome sizes of fish as a reflection of nature's extensive
1080 experiments with gene duplication. Trans Am Fish Soc 99(1):120-130.
1081 Parra G, Bradnam K, Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic
1082 genomes. Bioinformatics 23(9):1061-1067.
1083 Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, et al. 2007. Diet and the
1084 evolution of human amylase gene copy number variation. Nat Genet 39(10):1256-1260.
1085 Peters RS, Krogmann L, Mayer C, Donath A, Gunkel S, Meusemann K, Kozlov A, et al. 2017.
1086 Evolutionary history of the Hymenoptera. Curr Biol 27(7):1013-1018.
1087 Petrov DA, Hartl DL. 1997. Trash DNA is what gets thrown away: high rate of DNA loss in Drosophila.
1088 Gene 205(1-2):279-289.
1089 Petrov DA, Hartl DL. 1998. High rate of DNA loss in the Drosophila melanogaster and Drosophila virilis
1090 species groups. Mol Biol Evol 15(3):293-302.
1091 Petrov DA, Lozovskaya ER, Hartl DL. 1996. High intrinsic rate of DNA loss in Drosophila. Nature
1092 384(6607):346-349.
1093 Poelchau M, Childers C, Moore G, Tsavatapalli V, Evans J, Lee C-Y, Lin H, Lin J-W, Hacket K. 2015.
1094 The i5k Workspace@ NAL—enabling genomic data access, visualization and curation of arthropod
1095 genomes. Nucleic Acids Res 43(D1):D714-D719.

1096 Protas ME, Hersey C, Kochanek D, Zhou Y, Wilkens H, Jeffery WR, Zon LI, Borowsky R, Tabin CJ.
1097 2006. Genetic analysis of cavefish reveals molecular convergence in the evolution of albinism. Nat
1098 Genet 38(1):107-111.
1099 Qiu Y, Tittiger C, Wicker-Thomas C, Le Goff G, Young S, Wajnberg E, Fricaux T, Taquet N, Blomquist
1100 GJ, Feyereisen R. 2012. An insect-specific P450 oxidative decarbonylase for cuticular hydrocarbon
1101 biosynthesis. Proc Natl Acad Sci USA 109(37):14858-14863.
1102 Qu Z, Kenny NJ, Lam HM, Chan TF, Chu KH, Bendena WG, Tobe SS, Hui JHL. 2015. How did
1103 arthropod sesquiterpenoids and ecdysteroids arise? Comparison of hormonal pathway genes in
1104 noninsect arthropod genomes. Genome Biol Evol 7(7):1951-1959.
1105 Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features.
1106 Bioinformatics 26(6):841-842.
1107 R-Core-Team. 2018. R: A language and environment for statistical computing. Vienna, Austria: R
1108 Foundation for Statistical Computing.
1109 Ratzka C, Förster F, Liang C, Kupper M, Dandekar T, Feldhaar H, Gross R. 2012. Molecular
1110 characterization of antimicrobial peptide genes of the carpenter ant Camponotus floridanus. PLoS
1111 One 7(8):e43036.
1112 Rewitz KF, O'Connor MB, Gilbert LI. 2007. Molecular evolution of the insect Halloween family of
1113 cytochrome P450s: phylogeny, gene organization and functional conservation. Insect Biochem Mol
1114 Biol 37(8):741-753.
1115 Rizzon C, Ponger L, Gaut BS. 2006. Striking similarities in the genomic distribution of tandemly arrayed
1116 genes in Arabidopsis and rice. PLoS Comput Biol 2(9):e115.
1117 Robertson HM, Gadau J, Wanner KW. 2010. The insect chemoreceptor superfamily of the parasitoid
1118 jewel wasp Nasonia vitripennis. Insect Mol Biol 19:121-136.
1119 Robertson HM, Kent LB. 2009. Evolution of the gene lineage encoding the carbon dioxide receptor in
1120 insects. J Insect Sci 9(1):19.
1121 Robertson HM, Wanner KW. 2006. The chemoreceptor superfamily in the honey bee, Apis mellifera:
1122 expansion of the odorant, but not gustatory, receptor family. Genome Res 16(11):1395-1403.
1123 Robertson HM, Waterhouse RM, Walden KKO, Ruzzante L, Reijnders MJMF, Coates BS, Legeai F, et
1124 al. 2018. Genome sequence of the wheat stem sawfly, Cephus cinctus, representing an early-
1125 branching lineage of the Hymenoptera, illuminates evolution of hymenopteran chemoreceptors.
1126 Genome Biol Evol 10(11):2997-3011.
1127 Ross L, Gardner A, Hardy N, West SA. 2013. Ecology, not the genetics of sex determination, determines
1128 who helps in eusocial populations. Curr Biol 23(23):2383-2387.

1129 Roux J, Privman E, Moretti S, Daub JT, Robinson-Rechavi M, Keller L. 2014. Patterns of positive
1130 selection in seven ant genomes. Mol Biol Evol 31(7):1661-1685.
1131 Sadd BM, Barribeau SM, Bloch G, de Graaf DC, Dearden P, Elsik CG, Gadau J, et al. 2015. The
1132 genomes of two key bumblebee species with primitive eusocial organization. Genome Biol 16(1):1-
1133 32.
1134 Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, Preibisch S, et al. 2012. Fiji:
1135 an open-source platform for biological-image analysis. Nature Methods 9(7):676-682.
1136 Sezutsu H, Le Goff G, Feyereisen R. 2013. Origins of P450 diversity. Philos Trans R Soc Lond B Biol
1137 Sci 368:20120428.
1138 Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing
1139 genome assembly and annotation completeness with single-copy orthologs. Bioinformatics
1140 31(19):3210-3212.
1141 Simone M, Evans JD, Spivak M. 2009. Resin collection and social immunity in honey bees. Evolution
1142 63(11):3016-3022.
1143 Singaravelan N, Nee'man G, Inbar M, Izhaki I. 2005. Feeding responses of free-flying honeybees to
1144 secondary compounds mimicking floral nectars. J Chem Ecol 31(12):2791-2804.
1145 Slone J, Daniels J, Amrein H. 2007. Sugar receptors in Drosophila. Curr Biol 17(20):1809-1816.
1146 Smadja C, Shi P, Butlin RK, Robertson HM. 2009. Large gene family expansions and adaptive evolution
1147 for odorant and gustatory receptors in the pea aphid, Acyrthosiphon pisum. Mol Biol Evol
1148 26(9):2073-2086.
1149 Smit AFA, Hubley R. 2008-2015. RepeatModeler Open-1.0.
1150 Smit AFA, Hubley R, Green P. 2013-2015. RepeatMasker Open-4.0.
1151 Smith CR, Smith CD, Robertson HM, Helmkampf M, Zimin A, Yandell M, Holt C, et al. 2011. Draft
1152 genome of the red harvester ant Pogonomyrmex barbatus. Proc Natl Acad Sci USA 108(14):5667-
1153 5672.
1154 Smith SG. 1941. A new form of spruce sawfly identified by means of its cytology and parthenogenesis. J
1155 Agric Sci 21(5):245-305.
1156 Sohi SS, Ennis TJ. 1981. Chromosomal characterization of cell lines of Neodiprion lecontei
1157 (Hymenoptera: Diprionidae). Proc Entomol Soc Ont 112:45–48.
1158 Souvorov A, Kapustin Y, Kiryutin B, Chetvernin V, Tatusova T, Lipman DJ. 2010. Gnomon–NCBI
1159 eukaryotic gene prediction tool. National Center for Biotechnology Information.
1160 Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
1161 phylogenies. Bioinformatics 30(9):1312-1313.

1162 Suzuki HC, Ozaki K, Makino T, Uchiyama H, Yajima S, Kawata M. 2018. Evolution of gustatory
1163 receptor gene family provides insights into adaptation to diverse host plants in nymphalid butterflies.
1164 Genome Biol Evol 10(6):1351-1362.
1165 Thomas JH. 2006. Analysis of homologous gene clusters in Caenorhabditis elegans reveals striking
1166 regional cluster domains. Genetics 172(1):127-143.
1167 Tian C, Gao B, Fang Q, Ye G, Zhu S. 2010. Antimicrobial peptide-like genes in Nasonia vitripennis: a
1168 genomic perspective. BMC Genom 11(1):187.
1169 Trapp S, Croteau R. 2001. Defensive resin biosynthesis in conifers. Annu Rev Plant Biol 52(1):689-724.
1170 Vieira FG, Sánchez-Gracia A, Rozas J. 2007. Comparative genomic analysis of the odorant-binding
1171 protein family in 12 Drosophila genomes: purifying selection and birth-and-death evolution. Genome
1172 Biol 8(11):R235.
1173 Vitti JJ, Grossman SR, Sabeti PC. 2013. Detecting natural selection in genomic data. Annu Rev Genet
1174 4797-120.
1175 Vosshall LB, Wong AM, Axel R. 2000. An olfactory sensory map in the fly brain. Cell 102(2):147-159.
1176 Wanner KW, Robertson HM. 2008. The gustatory receptor family in the silkworm moth Bombyx mori is
1177 characterized by a large expansion of a single lineage of putative bitter receptors. Insect Mol Biol
1178 17(6):621-629.
1179 Wiens JJ, Lapoint RT, Whiteman NK. 2015. Herbivory increases diversification across insect clades. Nat
1180 Commun 6(1):1-7.
1181 Wilson LF, Wilkinson RC, Averill RC. 1992. Redheaded pine sawfly: its ecology and management. US
1182 Department of Agriculture, Forest Service, editor. Agriculture Handbook No. 694. Washington, DC.
1183 Xu Y-L, He P, Zhang L, Fang S-Q, Dong S-L, Zhang Y-J, Li F. 2009. Large-scale identification of
1184 odorant-binding proteins and chemosensory proteins from expressed sequence tags in insects. BMC
1185 Genom 10(1):632.
1186 Yang J, Chen X, Bai J, Fang D, Qiu Y, Jiang W, Yuan H, et al. 2016. The Sinocyclocheilus cavefish
1187 genome provides insights into cave adaptation. BMC Biol 14(1):1-13.
1188 Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24(8):1586-1591.
1189 Zhang J. 2003. Evolution by gene duplication: an update. Trends Ecol Evol 18(6):292-298.
1190 Zhang L, Gaut BS. 2003. Does recombination shape the distribution and evolution of tandemly arrayed
1191 genes (TAGs) in the Arabidopsis thaliana genome? Genome Res 13(12):2533-2540.
1192 Zhang Z, Zhu S. 2012. Comparative genomics analysis of five families of antimicrobial peptide-like
1193 genes in seven ant species. Dev Comp Immunol 38(2):262-274.
1194 Zhen Y, Aardema ML, Medina EM, Schumer M, Andolfatto P. 2012. Parallel molecular evolution in an
1195 herbivore community. Science 337(6102):1634-1637.

1196 Zhou X, Rokas A, Berger SL, Liebig J, Ray A, Zwiebel LJ. 2015. Chemoreceptor evolution in
1197 hymenoptera and its implications for the evolution of eusociality. Genome Biol Evol 7(8):2407-2416.
1198

1199 **Figure Legends**

1200 **Figure 1. Like other diprionids, *N. lecontei* has multiple morphological and behavioral adaptations**
1201 **to *Pinus* foliage. A.** An egg-laying *N. lecontei* female demonstrating several adaptations for dealing with
1202 thick, resinous pine needles, including: a robust saw-like ovipositor (visible within the needle), a tendency
1203 to lay many closely spaced eggs per needles, and a tendency to cut resin-draining slits on egg-bearing
1204 needles (circled). **B.** Prior to hatching, *N. lecontei* eggs absorb water from the host, causing the eggs to
1205 swell and the pockets to open. Throughout development, embryos are in close contact with living host
1206 tissue. **C.** Early-instar larvae have skeletonizing feeding behavior in which only the outer needle tissue is
1207 consumed, leaving the resinous interior intact. This strategy prevents small larvae from being
1208 overwhelmed by sticky resin. **D.** Mid- and late-instar larvae consume the entire pine needle. Larvae
1209 sequester pine resin in specialized pouches for use in self-defense (All photos by R.K. Bagley).
1210

1211 **Figure 2. Optical sections through the antennal lobes of adult female (left) and male (right) *N.***
1212 ***lecontei*.** White arrows indicate regions of male-specific synaptic clusters. Scale bars = 500 µm.
1213

1214 **Figure 3. Position of genes belonging to five environmentally responsive gene families along seven**
1215 ***N. lecontei* linkage groups.** Linkage groups (LG) are drawn to scale and ordered as in the linkage-group
1216 anchored assembly described in Linnen et al. 2018 (GenBank accession numbers are as follows: LG1 =
1217 CM009916.1; LG2 = CM009917.1; LG3 = CM009918.1; LG4 = CM009919.1; LG5 = CM009920.1;
1218 LG6 = CM009921.1; LG7 = CM009922.1). Gene family abbreviations: OR (olfactory receptor), GR
1219 (gustatory receptor), OBP (odorant binding protein), CYP (cytochrome P450), AMP (antimicrobial
1220 protein). Each gene family is represented by a different color. Horizontal lines indicate the approximate
1221 locations of genes within LG; diagonal lines that connect to horizontal lines are used to highlight groups
1222 of genes that met our clustering criteria. Genes that were found on scaffolds that have not been placed on
1223 linkage groups are indicated on the bottom left, with abbreviated scaffold names given in parentheses
1224 (e.g., S-210 = scaffold_210 = LGIB01000210.1 in the assemblies available on NCBI).
1225

1226 **Figure 4. Number of intact genes in hymenopteran genomes for each of five environmentally**
1227 **responsive gene families.** Phylogenetic relationships are as in Moreau et al. (2006); Hedtke et al. (2013);
1228 Roux et al. (2014); Brand et al. (2017); Branstetter et al. (2017); Peters et al. (2017). Branch lengths are
1229 arbitrary. Gene family abbreviations are as in Figure 3.

1230

1231 **Figure 5. Ecological correlates of gene family size in Hymenoptera.** Each point represents the number
1232 of intact genes for a hymenopteran species for which both manually curated gene annotations and
1233 ecological data are available. Asterisks indicate that gene number varies significantly among the
1234 ecological categories under consideration; for significant categories with >2 groups, letters indicate
1235 significance in post-hoc tests (groups that do not share a letter are significantly different). Gene number
1236 and ecological data for all taxa are provided in Table S8.

1237 **Table 1. Summary of within-genome signatures of adaptive expansions and contractions**
1238 **for five environmentally responsive gene families.**

1239

	Gene family size					Genomic Clustering		Molecular evolution		
Gene family*	Intact genes	Partial	Pseudo	Total genes	Prop. pseudo	Prop. in clusters†	Largest cluster	*Neodiprion*-specific clades‡	Significant branch tests§	Significant site tests**
OR	52	3	1	56	0.02	0.59	8	3	1	0
GR	41	2	2	44††	0.05	0.76	10	3	1	1
OBP	13	0	0	13	0	0.38	3	0	n/a	n/a
CYP (all)	93	2	12	107	0.11	0.66	16	5	2	0
CYP2 clan	9	0	0	9	0	0.33	2	0	0	0
CYP3 clan	47	0	8	55	0.15	0.81	16	4	2	0
CYP4 clan	27	2	4	33	0.12	0.55	3	1	0	0
mito CYP clan	10	0	0	10	0	0.50	3	0	0	0
AMP	21	0	0	21	0	0.95	15	?‡‡	0	0

1240

* Abbreviations: OR = olfactory receptor genes; GR = gustatory receptor genes; OBP = odorant binding protein genes; CYP = cytochrome P450 genes ("clans" refer to four major clades of CYPs present in insects); AMP = antimicrobial peptide genes.
† Calculated as: (number of genes in clusters of 2 or more)/(genes for which clustering could be evaluated).
‡ Defined as monophyletic clusters of 5 or more *Neodiprion* paralogs with a bootstrap support ≥ 70% in an amino acid phylogeny constructed with gene annotations from *Neodiprion*, select Hymenoptera, and *Drosophila melanogaster.*
§ To be counted, clades had to reject both 1-ratio and fixed-ratio models in dN/dS branch tests (see Table 2).
** To be counted, clades had to reject both M7 and M8a models in dN/dS site tests (see Table 2).
†† One gene was both a partial annotation and a pseudogene.
‡‡ Low bootstrap support precluded the identification of *Neodiprion*-specific clades.

Table 2. Likelihood-ratio tests (LRTs) of positive selection on *Neodiprion*-specific clades (branch models) and on amino acid sites within these clades (site models).

Clade Names*	n†	Model comparison‡	LRT statistic§	df	P-value**
Olfactory Receptor					
Clade 1	6	M8 vs M7	2.932	2	0.231
		M8 vs M8a	0.748	1	0.387
		2 ratio vs 1 ratio	5.408	1	**0.020**
		2 ratio vs neutral	7.800	1	**0.005**
Clade 2	5	M8 vs M7	3.941	2	0.139
		M8 vs M8a	1.525	1	0.217
		2 ratio vs 1 ratio	1.426	1	0.232
		2 ratio vs neutral	0.050	1	0.822
Clade 3	5	M8 vs M7	0	2	1
		M8 vs M8a	0	1	1
		2 ratio vs 1 ratio	2.395	1	0.122
		2 ratio vs neutral	6.371	1	**0.012**
Gustatory Receptor					
Clade 1	7	M8 vs M7	0.809	2	0.667
		M8 vs M8a	0.379	1	0.538
		2 ratio vs 1 ratio	0.003	1	0.954
		2 ratio vs neutral	0.645	1	0.422
Clade 2	8	M8 vs M7	6.049	2	**0.049**
		M8 vs M8a	2.654	1	0.103
		2 ratio vs 1 ratio	0.003	1	0.959
		2 ratio vs neutral	0.781	1	0.377
Clade 3	5	M8 vs M7	39.328	2	**2.884e-09**
		M8 vs M8a	35.167	1	**3.026e-09**
		2 ratio vs 1 ratio	14.789	1	**1.202e-04**
		2 ratio vs 2 ratio neutral	27.810	1	**1.338e-07**
Cytochrome P450					
Clade 1 (CYP4 clan)	8	M8 vs M7	0.615	2	0.735
		M8 vs M8a	0.866	1	0.352
		2 ratio vs 1 ratio	0.658	1	0.417
		2 ratio vs 2 ratio neutral	0.089	1	0.766
Clade 2 (CYP3 clan)	19	M8 vs M7	0	2	1
		M8 vs M8a	0	1	1
		2 ratio vs 1 ratio	2.077	1	0.149

* Clade names are as in Figures S1a,S1b,S2a,S2b,S3a,S3b,S4a,S4b, and S5a.
† Putatively functional genes. Pseudogenes and partial annotations were excluded from analysis.
‡ Site models unshaded; neutral M7 and M8a do not allow for positive selection. Branch models shaded; 1 ratio estimates a single ω value for all branches, 2 ratio estimates a separate ω value for the foreground branch, 2 ratio neutral fixes ω=1 for all branches.
§ Likelihood ratio test statistic, calculated twice the difference in model log likelihoods.
** Bolded values are significant at critical value 0.05.

		2 ratio vs 2 ratio neutral	0.076	1	0.783
Clade 3 (CYP3 clan)	6	M8 vs M7	7.152	2	**0.028**
		M8 vs M8a	0.649	1	0.421
		2 ratio vs 1 ratio	6.325	1	**0.012**
		2 ratio vs 2 ratio neutral	14.261	1	**1.59 e-04**
Clade 4 (CYP3 clan)	6	M8 vs M7	0	2	1
		M8 vs M8a	0.151	1	0.697
		2 ratio vs 1 ratio	0.002	1	0.964
		2 ratio vs 2 ratio neutral	0.936	1	0.333
Clade 5 (CYP3 clan)	5	M8 vs M7	0	2	1
		M8 vs M8a	0	1	1
		2 ratio vs 1 ratio	5.327	1	**0.021**
		2 ratio vs 2 ratio neutral	12.286	1	**4.56 e-04**
Hisnavicin (Antimicrobial Peptide)					
Clade 1[††]	15	M8 vs M7	2.388	2	0.665
		M8 vs M8a	0	1	1
		2 ratio vs 1 ratio	7.908	1	**0.010**
		2 ratio vs 2 ratio neutral	0.999	1	0.635

[††] Although this clade did not meet the bootstrap criteria for species-specific clades (>70), it was included in this analysis because it contained almost all *N. lecontei* hisnavicin paralogs.

Figure 1. Like other diprionids, *N. lecontei* has multiple morphological and behavioral adaptations to *Pinus* foliage. A. An egg-laying *N. lecontei* female demonstrating several adaptations for dealing with thick, resinous pine needles, including: a robust saw-like ovipositor (visible within the needle), a tendency to lay many closely spaced eggs per needles, and a tendency to cut resin-draining slits on egg-bearing needles (circled). **B.** Prior to hatching, *N. lecontei* eggs absorb water from the host, causing the eggs to swell and the pockets to open. Throughout development, embryos are in close contact with living host tissue. **C.** Early-instar larvae have skeletonizing feeding behavior in which only the outer needle tissue is consumed, leaving the resinous interior intact. This strategy prevents small larvae from being overwhelmed by sticky resin. **D.** Mid- and late-instar larvae consume the entire pine needle. Larvae sequester pine resin in specialized pouches for use in self-defense (All photos by R.K. Bagley).

Figure 2. Optical sections through the antennal lobes of adult female (left) and male (right) ***N. lecontei*****.** White arrows indicate regions of male-specific synaptic clusters. Scale bars = 500 µm.

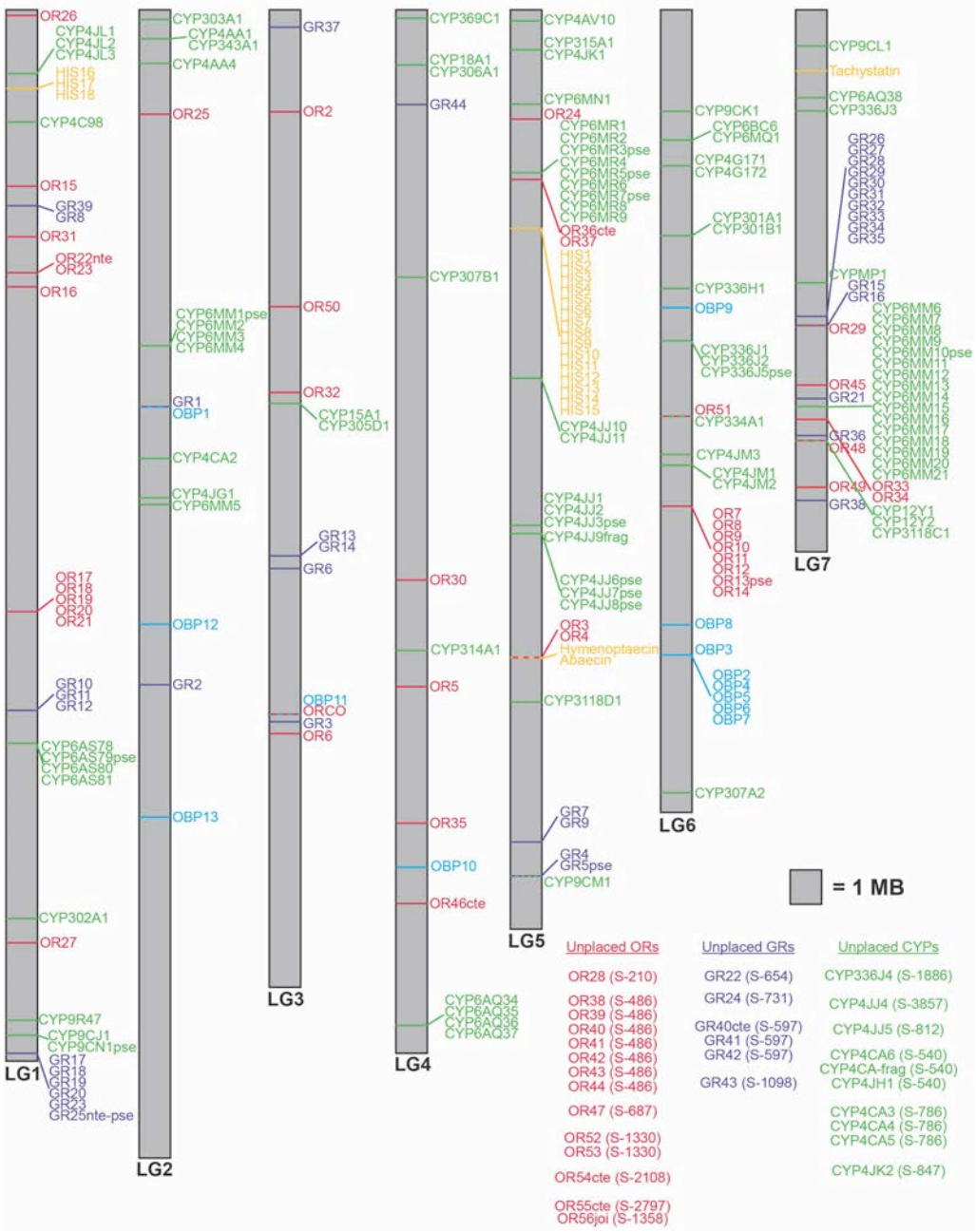
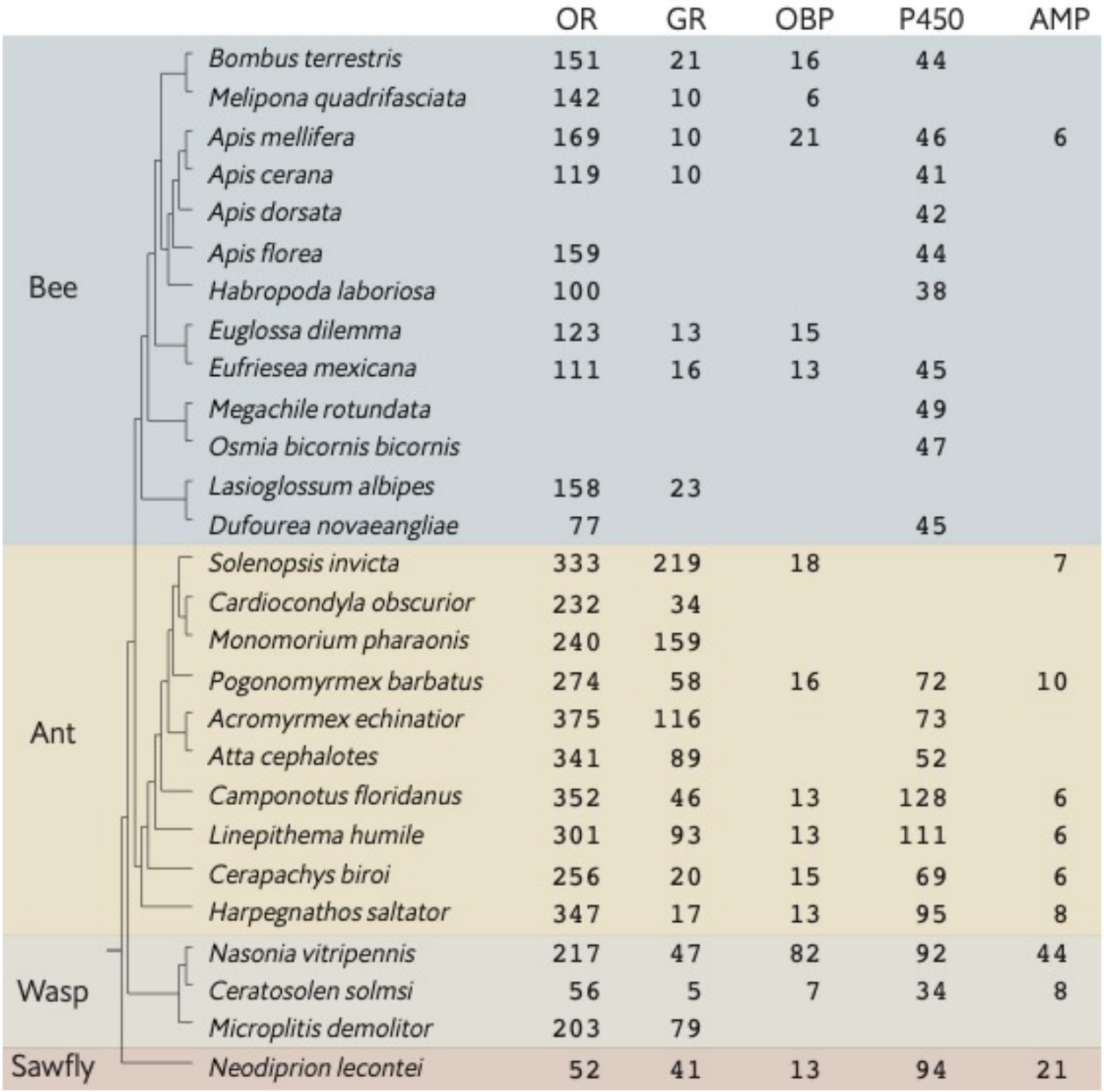


Figure 3. Position of genes belonging to five environmentally responsive gene families along seven *N. lecontei* linkage groups. Linkage groups (LG) are drawn to scale and ordered as in the linkage-group anchored assembly described in Linnen et al. 2018 (GenBank accession numbers are as follows: LG1 = CM009916.1; LG2 = CM009917.1; LG3 = CM009918.1; LG4 = CM009919.1; LG5 = CM009920.1; LG6 = CM009921.1; LG7 = CM009922.1). Gene family abbreviations: OR (olfactory receptor), GR (gustatory receptor), OBP (odorant binding protein), CYP (cytochrome P450), AMP (antimicrobial protein). Each gene family is represented by a different color. Horizontal lines indicate the approximate locations of genes within LG; diagonal lines that connect to horizontal lines are used to highlight groups of genes that met our clustering criteria. Genes that were found on scaffolds that have not been placed on linkage groups are indicated on the bottom left, with abbreviated scaffold names given in parentheses (e.g., S-210 = scaffold_210 = LGIB01000210.1 in the assemblies available on NCBI).



		OR	GR	OBP	P450	AMP
Bee	*Bombus terrestris*	151	21	16	44	
	Melipona quadrifasciata	142	10	6		
	Apis mellifera	169	10	21	46	6
	Apis cerana	119	10		41	
	Apis dorsata				42	
	Apis florea	159			44	
	Habropoda laboriosa	100			38	
	Euglossa dilemma	123	13	15		
	Eufriesea mexicana	111	16	13	45	
	Megachile rotundata				49	
	Osmia bicornis bicornis				47	
	Lasioglossum albipes	158	23			
	Dufourea novaeangliae	77			45	
Ant	*Solenopsis invicta*	333	219	18		7
	Cardiocondyla obscurior	232	34			
	Monomorium pharaonis	240	159			
	Pogonomyrmex barbatus	274	58	16	72	10
	Acromyrmex echinatior	375	116		73	
	Atta cephalotes	341	89		52	
	Camponotus floridanus	352	46	13	128	6
	Linepithema humile	301	93	13	111	6
	Cerapachys biroi	256	20	15	69	6
	Harpegnathos saltator	347	17	13	95	8
Wasp	*Nasonia vitripennis*	217	47	82	92	44
	Ceratosolen solmsi	56	5	7	34	8
	Microplitis demolitor	203	79			
Sawfly	*Neodiprion lecontei*	52	41	13	94	21

Figure 4. Number of intact genes in hymenopteran genomes for each of five environmentally responsive gene families. Phylogenetic relationships are as in Moreau et al. (2006); Hedtke et al. (2013); Roux et al. (2014); Brand et al. (2017); Branstetter et al. (2017); Peters et al. (2017). Branch lengths are arbitrary. Gene family abbreviations are as in Figure 3.

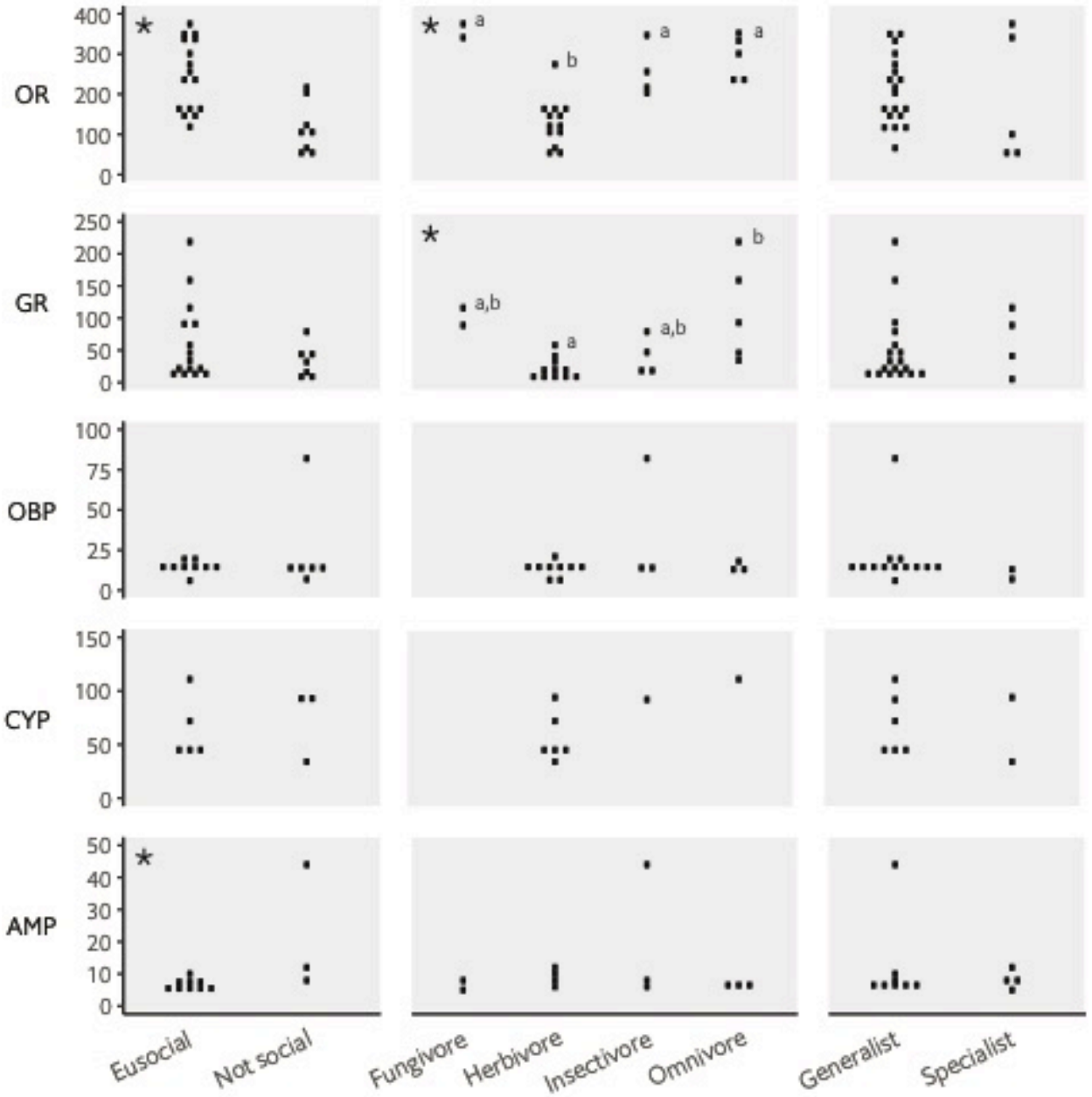


Figure 5. Ecological correlates of gene family size in Hymenoptera. Each point represents the number of intact genes for a hymenopteran species for which both manually curated gene annotations and ecological data are available. Asterisks indicate that gene number varies significantly among the ecological categories under consideration; for significant categories with >2 groups, letters indicate significance in post-hoc tests (groups that do not share a letter are significantly different). Gene number and ecological data for all taxa are provided in Table S8.