Ecological correlates of gene family size: the draft genome of the redheaded pine sawfly *Neodiprion lecontei*

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

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 large GR and small olfactory receptor (OR) families, eusociality was associated with large OR and small AMP families, and—unlike investigations among 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 give rise to novel traits. For example, rampant pseudogenization in particular gene families is linked to 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). 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 in which genes arise via duplication (gene gain) and are lost via deletion or pseudogenization (gene loss) (Hughes and Nei 1992; Nei and Rooney 2005). When the frequencies of gene duplications and deletions evolve primarily via genetic drift, gene families contract and expand over time via a process dubbed genomic drift (Nei 2007; Nozawa et al. 2007). Overall, this stochastic birth-death process (which differs from Nei's conceptual birth-death model of gene family evolution (Hahn et al. 2005)) sufficiently explains most gene family size distributions within genomes (Karev et al. 2002) and between species (Hahn et al. 2007).

But during an ecological shift, natural selection can influence birth-death dynamics by promoting the expansion or contraction of specific gene families. For lineages that adapt to a novel niche,

71 their genomes may exhibit evidence of selective maintenance of duplications or deletions in
72 certain gene families. For example, if selection favors gene gain, novel gene duplicates will tend
73 to persist in the population, resulting in subfamilies of recently diverged paralogs. If the
74 mutational mechanism that generates new duplicate genes is unequal crossing over during
75 meiosis, these recently diverged paralogs will be found in tandem arrays across the genome
76 (Zhang 2003). Moreover, if gene duplicates experience positive selection for novel functions,
77 this can result in elevated amino acid substitution rates among paralogs. Conversely, some
78 genetic functions may become obsolete or even deleterious in the novel habitat. When this
79 occurs, positive or relaxed purifying selection will cause some gene families to accumulate loss-
80 of-function mutations at an accelerated rate.
81
82 Sometime after an ecological shift, impacted gene families will reach a new equilibrium state in
83 which gene number evolves primarily via negative selection and genomic drift. Likewise,
84 pseudogenes will fade into the genomic background (Petrov et al. 1996; Petrov and Hartl 1997,
85 1998) and tandem array lengths will reflect local recombination rates (Akhunov et al. 2003;
86 Zhang and Gaut 2003; Rizzon et al. 2006; Thomas 2006). Thus, within-genome signatures of
87 adaptive changes in gene family size are likely ephemeral and best detected in lineages that
88 *recently* shifted to a novel niche. Nevertheless, if natural selection consistently favors the
89 expansion or contraction of specific gene families in specific environments, among-taxon
90 correlations between gene family size and ecology will be maintained. Currently, the extent to
91 which different taxa exhibit convergence at the level of gene family size changes is largely
92 unknown.
93
94 Arguably, the genes most likely to expand and contract convergently in response to similar
95 selection pressures are those that mediate organismal interactions with their biotic and abiotic
96 environments. These "environmentally responsive genes" include chemosensory (e.g., olfactory
97 and gustatory receptors), detoxifying (e.g., cytochrome P450), and immunity (e.g., MHC and
98 immunoglobulin) genes. To cope with constantly changing pressures, environmentally
99 responsive genes tend to be characterized by elevated sequence diversity, duplication rates,
100 substitution rates, and genomic clustering, as well as tissue- or temporal-specific expression
101 (Berenbaum 2002) and limited pleiotropy (Arguello et al. 2016). Importantly, causal links
102 between changes in environmentally responsive genes and adaptation to novel niches have been
103 established for multiple taxa (Després et al. 2007; Matsuo et al. 2007; Dobler et al. 2012; Zhen et
104 al. 2012; Sezutsu et al. 2013).

105 With their exceptionally diverse ecologies and ever-increasing availability of annotated genomes
106 (Consortium 2013; Poelchau et al. 2015), insects are a powerful system for investigating the
107 predictability of size changes in environmentally responsive gene families. To date, at least two
108 ecological transitions are hypothesized to have a predictable impact on gene family size in insect
109 lineages. In plant-feeding insects, the evolution of increased dietary specialization (i.e., smaller
110 diet breadth) is associated with reduced chemosensory and detoxifying gene family sizes
111 (McBride 2007; McBride and Arguello 2007; Good et al. 2014; Goldman-Huertas et al. 2015;
112 Calla et al. 2017; Comeault et al. 2017) but see (Gardiner et al. 2008). In hymenopteran insects,
113 eusociality is associated with expansions of the olfactory-receptor family and contractions of the
114 gustatory-receptor family (Robertson and Wanner 2006; Zhou et al. 2015; McKenzie et al. 2016;
115 Brand and Ramírez 2017) but see (Fischman et al. 2011; Johnson et al. 2018). Most of these
116 studies, however, considered a single ecological characteristic or gene family (but see Robertson

117 and Wanner 2006) which is potentially problematic in that changes in social behavior may often
118 be accompanied by changes in other ecological characteristics and vice versa (Faulkes et al.
119 1997; Duffy and Macdonald 2010; Ross et al. 2013). A better understanding of ecology and gene
120 family size relationships will require simultaneous consideration of multiple ecological
121 characteristics and diverse gene families.

122 Here, we characterize multiple environmentally responsive gene families in the genome of the
123 redheaded pine sawfly, *Neodiprion lecontei* (Order: Hymenoptera; Family: Diprionidae). This
124 species provides an opportunity to examine both within-genome signatures of adaptive gene
125 family contractions and expansions, and among-lineage correlations between ecology and gene
126 family size. First, for within-genome signatures, *N. lecontei* is an exemplar of an herbivorous
127 hymenopteran lineage (Diprionidae) that underwent a drastic host shift: sometime within the last
128 60 million years, this lineage transitioned from angiosperms to coniferous host plants in the
129 family Pinaceae (Boevé et al. 2013; Peters et al. 2017). To defend against herbivores and
130 pathogens, Pinaceae produce viscous oleoresin secretions that are sticky and have unique
131 antimicrobial properties (Trapp and Croteau 2001; Gershenzon and Dudareva 2007). To manage
132 these toxic and extraordinarily sticky resins, *N. lecontei* and related diprionids evolved
133 specialized feeding and egg-laying traits (Figure 1). Beyond these traits, we hypothesize that
134 pine specialization likely resulted in pronounced changes to the selection pressures acting on
135 multiple gene families, especially those involved in chemosensation, detoxification, and immune
136 function. Second, with respect to among-lineage correlations between ecology and gene-family
137 size, *N. lecontei* is an herbivorous, non-eusocial insect from the Eusymphyta, a massively
138 understudied hymenopteran clade (Peters et al. 2017). Although many assembled and annotated
139 hymenopteran genomes are currently available, almost all have come from apocritans (bees,
140 wasps, and ants, but see (Robertson et al. 2018)). Thus, *N. lecontei* increases the ecological,
141 behavioral, and taxonomic diversity of hymenopteran genomes for evaluating ecological
142 correlates of gene family size among taxa.
143
144 To evaluate the predictability of gene family evolution, we assembled a draft genome for *N.*
145 *lecontei* and manually annotated genes for five environmentally responsive gene families:
146 olfactory receptor (OR), gustatory receptor (GR), odorant binding protein (OBP), cytochrome
147 P450 (CYP), and antimicrobial peptide (AMP). For gene families that underwent a size change
148 related to pine adaptation, we expected one or more of the following patterns: (1) clusters of
149 recently diverged paralogs in gene-family trees, (2) a high proportion of genes in tandem arrays,
150 (3) signatures of positive selection among paralogs, and (4) elevated rates of pseudogenization.
151 Then, for the same five gene families, we asked whether gene-family size correlated with
152 ecology among distantly related insect taxa. To do so, we compiled published gene annotations
153 and ecological variables (diet type, degree of ecological specialization, presence/absence of
154 eusociality) for hymenopteran taxa. Together, these analyses identify possible candidate gene
155 families underlying pine specialization and reveal that relationships between gene family size
156 and ecology differ among environmentally responsive gene families.
157
158
159 **Results**
160 **Genome assembly and annotation**
161 *Sequencing and assembly*

162 We sequenced one mate-pair and two small-insert Illumina libraries made from haploid male
163 siblings (see Methods). After read processing, we retained 268 billion PE100 reads with a
164 combined read depth of 112x (Table S1). ALLPATHS-LG (v47417) (Gnerre et al. 2011)
165 produced a 239-Mbp assembly consisting of 4523 scaffolds, with a scaffold N50 of 243 kbp
166 (Table S2). Prior studies identified seven chromosomes in *N. lecontei* (Smith 1941; Maxwell
167 1958; Sohi and Ennis 1981; Linnen et al. 2018). With an estimated genome size (1C) of 331 ±9.6
168 Mbp, our assembly captured 72% of the genome. Overall, these metrics are comparable to other
169 hymenopteran assemblies (Table S2).
170
171 To measure assembly completeness and artificial sequence duplication, we used CEGMA (Parra
172 et al. 2007) and BUSCO (Simão et al. 2015). Both search the assembly for a set of single-copy,
173 conserved genes, however, the CEGMA software has been deprecated
174 (http://korflab.ucdavis.edu/Datasets/cegma). Of the 248 CEGMA core eukaryotic genes, 90%
175 aligned as complete, single copies and 8% aligned complete but duplicated. For BUSCO, we
176 used the OrthoDB arthropod dataset, and out of 2675 groups 77% were complete, single copies
177 and 3% were complete but duplicated. These metrics indicate the presence of artificial duplicate
178 sequences, but otherwise the assembly was reasonably complete and suitable for annotation.
179
180 About 15.8% of the assembly consisted of repetitive elements, including 122 unknown
181 transposable elements that were mostly unique to *N. lecontei* (Table S3), and 212 other families
182 of transposable elements and simple repeats. This 15.8% corresponds to 11.4% of the actual 331-
183 Mb genome, of which we predict 27.6% is repetitive, suggesting that ~16.1% of the missing
184 ~28% of the genome is repetitive content (Table S3). For de novo gene prediction, we
185 included the *N. lecontei* transcriptome and protein-coding genes from *Atta cephalotes*
186 (OGSv1.2), *Acromyrmex echinatior* (OGSv3.8), *Apis melifera* (OGSv3.2), *Athalia rosae*
187 (OGSv1.0), and *Nasonia vitripennis* (OGSv1.0) to guide annotation. The official gene set
188 (OGSv1) had 12,980 gene models while the transcriptome had an average of 26,000 transcripts
189 per tissue (Table S4).
190
191 *Olfactory receptor*
192 The OR gene family had 56 genes total, including the co-receptor *Orco*; one gene contained stop
193 codons, three were partial annotations, and 52 genes were intact (Table 1). In *D. melanogaster*
194 most olfactory sensory neurons (OSNs) express a single OR (along with the coreceptor, *Orco*),
195 and OSNs expressing a particular OR converge on a single glomerulus in the antennal lobe (Gao
196 et al. 2000; Vosshall et al. 2000; Couto et al. 2005) but see (Fishilevich and Vosshall 2005). This
197 anatomy results in a general one-to-one correspondence between the number of ORs and the
198 number of glomeruli, a correspondence also observed in the hymenopteran European honey bee
199 (*Apis mellifera*, (Robertson and Wanner 2006)). Based on these studies and examination of the
200 antennal lobes of adult male and adult female *N. lecontei*, we expected to find a minimum of 49
201 functional ORs (Table S5, Figure 2). The close correspondence between our gene annotations
202 and glomeruli counts suggests that we have located all or most *N. lecontei* OR genes.
203
204 59% of ORs were in genomic clusters of two or more genes (Figure 3), a low proportion
205 compared to many other hymenopteran OR families (Robertson and Wanner 2006; Zhou et al.
206 2015; McKenzie et al. 2016; Brand and Ramírez 2017). A phylogenetic analysis of OR protein
207 sequences from *Neodiprion,* six other hymenopterans, and *D. melanogaster* identified three

208 *Neodiprion*-specific clades with at least five genes (Figure S1a). These same three clades were
209 also recovered in a phylogenetic analysis of *Neodiprion* OR cDNA sequences (Figure S1b). For
210 each *Neodiprion*-specific OR clade (and *Neodiprion*-specific clades in other gene families, see
211 below), we used the *Neodiprion* cDNA tree, the codeml program in the PAML package (Yang
212 2007), and likelihood-ratio tests to ask: (1) whether the ratio of non-synonymous to synonymous
213 substitution rates (dN/dS or ω) for the focal OR clade differed from the rest of the *Neodiprion*
214 OR gene family and, if so, whether they exhibited evidence of positive selection (ω>1) (branch
215 tests); and (2) whether ω differed among sites across members of *Neodiprion*-specific clades
216 and, if so, which sites exhibited evidence of positive selection (site tests). For only one OR clade
217 (OR clade 1) did we detect evidence of branch-specific positive selection (i.e., rejection of both
218 1-ratio and fixed-ω models), but this clade lacked evidence of site-specific positive selection
219 (Table 2).
220
221 *Gustatory receptor*
222 The GR gene family had 44 genes total; two genes contained stop codons, two were partial
223 annotations (one annotation was both partial and pseudogenized), and 41 were intact (Table 1).
224 76% of the GRs that could be placed on chromosomes were in genomic clusters (Figure 3) with
225 three *Neodiprion*-specific clades of at least five genes (Figures S2a and S2b). Only one clade
226 (GR clade 3) had evidence of branch-specific positive selection (Table 2). This clade also had
227 evidence of positive selection at some amino acid positions among paralogs (Table 2; sites with
228 evidence of positive selection include: 77E, 79S, 146N, 275S, 301S). Notably, GR Clade 3 is an
229 expansion of six paralogs orthologous to *DmGR66a*, a bitter receptor specifically for caffeine
230 (Moon et al. 2006). However, *N. lecontei* orthologs were not found for *DmGR93a* (Lee et al.
231 2009) and *DmGr33a* (Moon et al. 2009), coreceptors possibly required for caffeine detection.
232 Together, these data suggest that caffeine-like GR receptors have been coopted for novel
233 functions in *N. lecontei*.
234
235 The GR family also had orthologs for sugar receptors *DmGR5a* (trehalose) (Dahanukar et al.
236 2001), *DmGR43a* (fructose) (Miyamoto et al. 2012), and *DmGR64a-f* (multiple sugars) (Slone et
237 al. 2007) as well as carbon dioxide receptors *DmGR21a* and *DmGR63a* (Jones et al. 2007)
238 (Figure S2a). Orthologs to these carbon dioxide receptors have not been found in Apocrita but
239 seem to be preserved in Symphyta, like *N. lecontei* (Robertson and Kent 2009; Robertson et al.
240 2018).
241
242 *Odorant binding protein*
243 The OBP gene family had 13 genes total; none were pseudogenized or partial annotations (Table
244 1). In this family, 38% of genes were in genomic clusters, including a cluster of five genes on
245 chromosome 6 (Figure 3). *Neodiprion*-specific OBP clades were not found, even for the
246 chromosome 6 cluster. We note, however, that the OBP phylogenies had low bootstrap support
247 (Figure S3a,b), making it difficult to infer relationships among paralogs.
248
249 *Cytochrome P450*
250 The CYP gene family had 107 genes total; twelve genes contained stop codons, two were partial
251 annotations, and 93 were intact (Table 1). In insects, CYPs belong to four major clades, which
252 are referred to as clans (Feyereisen 2012). When we split the CYP gene family by clan, the
253 CYP2 clan had nine intact genes; the CYP3 clan had 47 intact genes and eight pseudogenes; the

254 CYP4 clan had 27 intact genes, four pseudogenes, and two partial genes; and the mitochondrial
255 CYP clan had 10 intact genes (Table 1). Across all CYPs, 66% were in genomic clusters (Figure
256 3). Looking at the four major clans separately, the percentage of clustered genes were: 33% for
257 CYP2, 81% for CYP3, 55% for CYP4, and 50% for mitochondrial CYP.
258
259 The CYP gene family had five *Neodiprion*-specific clades with at least five genes (Figure S4a,b),
260 four of which were in the CYP3 clan. Of these, two clades that were both within the CYP3 clan
261 (CYP clades 3 and 5) had evidence of branch-specific, but not site-specific, positive selection
262 (Table 2). CYP clade 3 contained members of the CYP6 subfamily, and the CYP clade 5
263 contained members of the CYP336 subfamily. Several studies to date suggest that members of
264 the CYP3 clan—and the CYP6 subfamily in particular—play an important role in detoxifying
265 pesticides and host-plant allelochemicals (Feyereisen 2012).
266
267 Orthologs were found for all the Halloween genes (which include genes from both the CYP2 and
268 mitochondrial CYP clans) of the 20-hydroxy ecdysone biosynthesis pathway: *CYP302A1*
269 (disembodied), *CYP306A1* (phantom), *CYP307A2* (spookier), *CYP307B1* (spookiest),
270 *CYP314A1* (shade), *CYP315A1* (shadow), and *CYP18A1* which turns over 20-hydroxy ecdysone
271 (Rewitz et al. 2007; Feyereisen 2011; Guittard et al. 2011; Qu et al. 2015). The juvenile hormone
272 biosynthesis gene *CYP15A1* was present as well (Helvig et al. 2004). Finally, *N. lecontei* had
273 orthologs for the two CYP4G enzymes that synthesize the cuticular hydrocarbons used as
274 external waterproof coating (Qiu et al. 2012).
275
276 *Immunity*
277 Antimicrobial peptides (AMPs) are expressed upon infection to kill or inhibit microbes. Based
278 on hymenopteran sequences, the *N. lecontei* AMP gene family had 21 genes (Table 1; Table S6),
279 including single copies of *Hymenoptaecin*, *Abaecin*, and *Tachystatin*, but no clear *Defensin*
280 ortholog. Over 18 *Hisnavicin* genes were identified, including a *Neodiprion*-specific expansion
281 of eight histidine-rich paralogs orthologous to *Hisnavicin-4*, which has been characterized as a
282 larval cuticle protein and AMP, but not functionally tested (Tian et al. 2010). The *N. lecontei*
283 Hisnavicins had a conserved 62 amino acid motif that appeared up to 19 times in a single
284 protein; the purpose of this amplification is unknown. 95% of the AMPs were in genomic
285 clusters (Figure 3). Due to low bootstrap support on many of the branches in our Hisnavicin
286 protein tree, we could not identify unambiguous *Neodiprion*-specific clades (Figure S5a).
287 However, our *Neodiprion* cDNA tree (Figure S5b) did reveal strong support for the monophyly
288 of a cluster of 15 *Hisnavicins* on linkage group 5 (Figure 3), and this cluster had some evidence
289 of positive selection (Table 2).
290
291 Outside of the AMP family, most immune pathways had direct orthologs between *N. lecontei*
292 and *D. melanogaster* (Figure S6, Table S7). The basic viral siRNA response pathway was
293 completely conserved between species. The immune deficiency (IMD) pathway was missing an
294 ortholog for the peptidoglycan recognition receptor *PGRP-LC*, but it is likely that another *PGRP*
295 replaced *PGRP-LC* in *N. lecontei*; assigning PGRP orthology was also difficult in ants (Gupta et
296 al. 2015). Also missing is the *Drosophila* mitogen activated protein kinase kinase kinase, TGF-β
297 activated kinase 1 (*Tak1*), but *N. lecontei* had a similar TGF-β activated kinase that is a close
298 ortholog to several *Tak1-like D. melanogaster* proteins possibly involved in immune deficiency
299 signaling. The encapsulation/melanization pathway was missing one of the two *Drosophila*

300 GTPases (*Rak2*). The *N. lecontei Rak1* ortholog may be playing both roles, but again this is
301 likely due to the difficulty of assigning one-to-one orthologs. The Duox pathway was missing the
302 top G-protein coupled receptor, but this is unknown in *D. melanogaster* and unidentified in other
303 Hymenoptera (Evans et al. 2006). Interestingly, *N. lecontei* had two copies of Dual Oxidase
304 (*Duox*), which regulates commensal gut microbiota and infectious microbes (Ha et al. 2005; Lee
305 et al. 2015); *Apis mellifera* had one copy. Finally, the Toll pathway *NF-kappaB* transcription
306 factor, *Dorsal-related immunity factor (Dif)* does not have a one-to-one ortholog in *N. lecontei*,
307 but two copies of its paralog, *Dorsal*, were present.
308
309 **Within-genome signatures of adaptive expansions and contractions**
310
311 *Evidence of selection in* Neodiprion-*specific gene family clades*
312 Massive gene family expansions with dozens of genes were not found in *N. lecontei* (in contrast
313 to (Smadja et al. 2009; Zhou et al. 2015)). Instead, the largest *Neodiprion*-specific clade had 22
314 genes (CYP gene family) and the rest had fewer than 10 genes. Nevertheless, we did identify 11
315 *Neodiprion*-specific clades containing at least 5 closely related paralogs and a monophyletic
316 clade of 15 AMPs with ambiguous ancestry (Table 1). Of these 12 clades, four had significant
317 branch positive selection (OR clade 1, GR clade 3, CYP clade 3, and CYP clade 5) (Table 2). Of
318 these four clades, only one also had significant site-specific positive selection (GR clade 3)
319 (Table 2).
320
321 *Clustering*
322 Our five focal gene families varied in the proportion of genes that were found in clusters of two
323 or more genes (Fisher's exact test, $P = 0.002$; Table 1). Post-hoc tests revealed that much of this
324 variation was due to differences between the highly clustered AMP family and all other families
325 except GR (AMP vs. OR: $P = 0.0091$; AMP vs. OBP: $P = 0.0053$; AMP vs. CYP: $P = 0.024$,
326 AMP vs. GR: $P = 0.12$; all p-values are FDR-corrected). The only other difference in clustering
327 that we detected was between the GR and singleton-heavy OBP families (FDR-corrected $P =$
328 0.045).
329
330 Differences in clustering were even more pronounced when we separated the CYP family by
331 clan (Fisher's exact test, $P < 0.0001$; Table 1). In addition to the pairwise differences described
332 above, we also found that the proportion of CYP3 genes found in clusters differed significantly
333 from ORs, CYP2s, and CYP4s (all FDR-corrected $P < 0.05$), but not AMPs, GRs, and
334 mitochondrial CYPs. Additionally, AMP clustering differed from CYP2, CYP4, and
335 mitochondrial CYP, while GR differed from CYP2 (all FDR-corrected $P < 0.05$). Together, these
336 analyses identified AMP, GR, and CYP3 as having an unusually high proportion of genes found
337 in clusters compared to other environmentally responsive gene families.
338
339 *Pseudogenization*
340 Overall, we found very few pseudogenes, and the proportion of pseudogenized genes did not
341 differ significantly among gene families (Fisher's exact test, $P = 0.12$; Table 1). The
342 chemoreceptors had one pseudogene each while CYP had 12, which is about 10% of the family,
343 but this was also the largest gene family. Although CYP3 had more pseudogenes than other CYP
344 clans, the proportion of pseudogenized genes still did not differ when we compared CYP clans

(Fisher's exact test, $P = 0.10$). Given these low rates of pseudogenization, it is unlikely that *N. lecontei* gene families underwent substantial, recent contractions.

Ecological correlates of gene-family size across insects

We first examined broad-scale variation in the sizes of our five focal gene families and four CYP clans among different insect orders (Figure S7). Not surprisingly, sample sizes were highly variable across gene families and insect orders. Despite this variation, we observed some intriguing differences among gene families and taxa. We detected significant differences in gene family size among orders for OR (Kruskal-Wallas chisq = 48.2, df = 12, $P < 1 \times 10^{-5}$), GR (K-W chisq = 25.5, df = 9, $P = 0.0025$), 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 AMP (K-W chisq = 7.93, df = 5, $P = 0.16$). We note, however, that the AMP sample size was considerably smaller than the other gene families. When we looked at CYP clans individually, we found differences among orders for CYP4 (K-W chisq = 19.0, df = 7, $P = 0.0083$) and mitochondrial CYP (K-W 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 = 8.76, df = 7, $P = 0.27$).

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

We next examined how gene family size correlated with ecology within the hymenopteran clade (Figures 4 and 5). Once again, we observed differences among gene families. We found that the number of ORs differed significantly among hymenopteran species that differed in diet (Kruskal-Wallas chisq = 15.8, df = 3, $P = 0.0012$) and sociality (Wilcoxon rank-sum test W = 115; $P = 0.00094$). For diet, we found that herbivores had fewer ORs than all other diet types (fungivores vs. herbivores: $P = 0.015$; omnivores vs. herbivores: $P = 0.015$; insectivores vs. herbivores: $P = 0.048$; all p-values are FDR-corrected). We observed an even more striking difference between eusocial and non-eusocial hymenopterans, with the former having larger OR families, on average. By contrast, GR family size was 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 remains 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 selective maintenance of novel gene duplicates, and two of these families also exhibited 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 both 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 (if there is 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 and exhibited evidence of positive selection, making these especially promising candidates for expansions related to a novel coniferous host. Additionally, although the AMP family lacked evidence of positive selection, its unusually clustered distribution in the *Neodiprion* genome could be related to selection for increased dosage of a conserved protein function (Perry et al. 2007). Below we discuss the functions of these three candidate families in more detail.

Shifts to pine feeding likely involved changes in the detection of and response to pine-specific cues. Intriguingly, the one GR clade with evidence of positive selection—GR clade 3—is an expansion of six paralogs (one is pseudogenized) orthologous to *DmGR66a*, a bitter receptor specifically for caffeine (Moon et al. 2006). However, orthologs were not found for *DmGR93a* (Lee et al. 2009) and *DmGr33a* (Moon et al. 2009), coreceptors possibly required for caffeine detection. Nevertheless, honeybees, which also lack clear orthologs to these putative coreceptors (Wanner and Robertson 2008), can detect and even prefer low concentrations of caffeine and nicotine (Singaravelan et al. 2005, but see de Brito Sanchez 2011). Although pines do not

contain caffeine, they do synthesize alkaloids that could confer some bitterness (Mumm and Hilker 2006). Thus, despite lacking caffeine coreceptor orthologs, members of GR clade 3 may still be involved in the detection of pine-specific bitter compounds. Duplications of putative bitter GRs are documented in other host-specialized insects, such as *Heliconius, Danaus,* and *Bombyx* butterflies (Wanner and Robertson 2008; Briscoe et al. 2013). Our sawfly-specific GR expansion, coupled with the finding that GR family size is associated with diet (see below), lends support to the hypothesis that expansions of GR bitter receptors repeatedly contribute to changes in oviposition and feeding behaviors in plant-feeding insects.

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

Because pine resin has antimicrobial (Himejima et al. 1992; Cowan 1999; Gershenzon and Dudareva 2007) and fungicidal properties (Grayer and Harborne 1994), we hypothesized that *N. lecontei* co-opted these compounds for its own defense, leading to relaxed selection on genes involved in immunity and a reduced innate immune response. In other Hymenoptera, honeybees (*Apis mellifera*) exposed to plant resin have reduced expression of immune-related genes (Simone et al. 2009) and wood ants (*Formica paralugubris*) that use conifer resin as building material have slightly reduced inducible immune system activity and nests with lower bacterial and fungal loads (Castella et al. 2008). In Diptera, AMP loss is associated with herbivorous lineages that live within host tissue, a more sterile habitat than is experienced by most dipterans (Hanson et al. 2019). Unexpectedly, we found a large species-specific clade of *Hisnavicin*-like AMPs in *Neodiprion*. Although additional data are needed to confirm that *Hisnavicin* orthologs act as AMPs in *N. lecontei*, one possible explanation for this putatively adaptive expansion that lacked an accompanying change in non-synonymous substitution rate is that having large numbers of *Hisnavicin*-like AMPs confers protection against pathogens unique to pine trees. That said, our data do not rule out adaptive AMP loss. For example, *N. lecontei* lacks a clear *Defensin* ortholog, a gene present in all dipterans tested to date (Hanson et al. 2019).

Limitations of within-genome analyses

One benefit to studying adaptive expansions/contractions within a single taxon is that gene families have likely experienced similar demographic histories, which can also impact gene birth and death rates. That said, each of our within-genome signatures of selection has limitations that should be revisited with additional data. First, our analysis of genomic clustering does not account for local recombination rate variation, which correlates with tandem array size in several taxa (Gaut et al. 2007). A fine-scale recombination rate map, coupled with clustering analyses

482 for many additional gene families, would more rigorously test the extent to which individual
483 gene family clustering deviates from the genome-wide relationship between recombination rate
484 and tandem array size.
485
486 Second, a lack of comparable data from other Eusymphyta meant that our gene family
487 phylogenies lacked orthologues from closely related sawfly taxa. Thus, the "*Neodiprion*-
488 specific" clades may not be unique to pine-feeding sawflies. If these paralogs were present prior
489 to the shift to pine hosts, this would not support a scenario in which new duplicates were
490 selectively maintained in the novel niche. Signatures of positive selection may still be related to
491 pine adaptation but would indicate selection on preexisting loci rather than selection favoring
492 gene family expansion.
493
494 Third, signatures of adaptive gene family expansions and contractions may be ephemeral, and
495 the shift to pine use could have occurred too long ago to detect these signatures in *N. lecontei*.
496 For example, in *Drosophila*, pseudogenes have an estimated half-life of ~14.3 million years
497 (Petrov et al. 1996; Petrov and Hartl 1997, 1998). If the rate of gene decay is similar in
498 *Neodiprion*, then pseudogenes that formed after a shift to pine (up to 60 mya) may no longer be
499 detectable in the genomes of extant sawflies. Likewise, gene clustering patterns are likely to
500 change over time from chromosomal rearrangements and additional gene duplications and
501 deletions. To investigate how the number and position of genes in these focal families has
502 changed over time, high quality gene annotations for diprionids and many additional sawfly
503 outgroups are needed. Fortunately, even if footprints of recent gene family size changes are too
504 ephemeral to be detected in most taxa, consistent relationships with ecology should still be
505 detectable given sufficient sampling of taxa differing in ecological traits of interest.
506
507 **Ecological correlates of gene family size among hymenopteran taxa**
508 The largest insect OR gene families are in eusocial Hymenoptera, leading to the hypothesis that
509 OR family size expansions were favored in these lineages because they facilitate complex
510 chemical communication (Robertson and Wanner 2006; LeBoeuf et al. 2013; Zhou et al. 2015).
511 To date, evidence in support of this hypothesis has been mixed (e.g., (Roux et al. 2014; Brand
512 and Ramírez 2017). Consistent with the OR-eusociality hypothesis, we found that, on average,
513 eusocial hymenopterans had larger OR families than non-eusocial hymenopterans. However, it is
514 likely that eusocial taxa differ from non-eusocial taxa in many other aspects of their ecology that
515 should also impact OR evolution. Indeed, we found that herbivorous hymenopterans tended to
516 have fewer OR genes than non-herbivores.
517
518 Whereas all eusocial hymenopterans had relatively large OR families, some eusocial
519 hymenopterans had relatively small GR families (Figures 4, 5; (Zhou et al. 2015)). To explain
520 the strikingly small set of GR genes in honeybee, Wanner and Robinson (2006) proposed that a
521 stable hive environment and a mutualistic relationship with flowering plants resulted in a lack of
522 selection for GR expansions. Intriguingly, our data indicate that among hymenopterans, GR
523 family size is associated with diet, but not eusociality. Like the ORs, GR gene family size tends
524 to be smaller in herbivores than in non-herbivorous taxa, regardless of social behavior. The
525 directionality of this change, however, is unclear: do shifts to plant diets favor reductions in GR
526 families, do shifts to non-plant diets favor GR expansions, or is it both? Answering this question
527 will require characterizing GR families across many independent transitions to and from

528 herbivory, as well as polarizing directions of change (i.e., distinguishing GR gains from GR
529 losses). Fortunately, there are many such diet transitions across diverse clades of insects (Wiens
530 et al. 2015).
531
532 Unlike sociality and diet, ecological specialization was not associated with gene family size in
533 any of the five gene families we evaluated. This result was unexpected because specialization-
534 associated reductions in gene family size are documented in diverse taxa and multiple gene
535 families, including the families examined here (McBride 2007; Smadja et al. 2009; Cao et al.
536 2014; Goldman-Huertas et al. 2015; Suzuki et al. 2018). One explanation for the lack of
537 association between gene family sizes and specialization in our data is that our "generalist" and
538 "specialist" categories are not meaningful across diverse diets (Forister et al. 2012).
539 Additionally, within a particular diet, the degree of specialization may be highly labile, with
540 rapid fluctuations that are not captured in our broad, order-wide comparison. Indeed, previous
541 studies that reported correlations between gene family size and ecological specialization focused
542 on closely related species. Thus, to fully understand how changes in ecology shape gene family
543 evolution, it will be necessary to evaluate ecological correlates of gene family size at multiple
544 time scales of taxonomic divergence.
545
546 Compared to ORs and GRs, our other focal gene families had far less manual annotation data
547 available for analysis. This may explain, in part, why we did not detect strong ecological
548 correlates for the other gene families. It is also possible that by focusing on the sizes of entire
549 gene families, we missed relevant signals in particular subfamilies (Hahn et al. 2007). For
550 example, as noted above, expansions of CYP3 and CYP4 subfamilies are associated with wood-
551 feeding insects and CYP3 clan subfamilies were also linked to detoxification in honey bee
552 (Berenbaum and Johnson 2015; Johnson et al. 2018). However, we did not detect any
553 correlations between ecology and CYP clan sizes. Despite these limitations, we did uncover hints
554 that AMP gene family size may be larger in non-eusocial lineages. If eusocial taxa tend to
555 inhabit more sterile environments (nests and hives) than non-eusocial taxa, this finding is
556 consistent with associations between habitat and AMP loss reported in dipterans (Hanson et al.
557 2019). Given that AMPs were also implicated in our within-genome analysis, immune-related
558 genes are especially promising candidates for future manual annotation projects.
559
560 *Limitations of among-taxon analyses*
561 Comparative analysis is a powerful approach for evaluating the repeatability and predictability of
562 evolutionary outcomes. Although our comparison of candidate gene family sizes among
563 ecologically diverse hymenopterans hints at intriguing relationships between ecology and gene
564 family size, it also had several limitations that should be revisited in future work. First, because
565 several taxa in our manual annotation dataset are missing from published hymenopteran
566 phylogenies (Peters et al. 2017), we were unable to correct for phylogenetic non-independence
567 and polarize gene gain/loss (e.g., as in (Hahn et al. 2005; Han et al. 2013) without losing
568 unacceptable amounts of data. Without accounting for similarity in ecology and gene family size
569 due to recent common ancestry, our Type I error rate is likely inflated and p-values should be
570 interpreted with caution. Nevertheless, variation in patterns of association among ecological
571 traits and gene families suggest that phylogeny and ecology are, to some extent, decoupled.
572

The gene annotation and ecological datasets also had limitations. For example, across studies that included manual annotations, we observed a lack of consistency in the methods and criteria for manually curated gene family datasets. The most problematic inconsistency was in the criteria for delineating intact, partial, and pseudogenized gene annotations. “Intact” could mean an exon-by-exon check against closely related orthologs, a minimum amino acid length, or merely the presence of an expected domain. Meanwhile, in reference publications, the number of pseudogenized and partial annotations were not always reported or were conflated. This is in addition to variation in the methods used to search for genes. Inconsistency in annotation methods and criteria across studies may introduce taxon-specific biases that have nothing to do with ecology. Regarding ecology, categorizations are somewhat subjective. For example, this study and Rane et al. (2016) classified bees as generalists since they collect nectar and pollen from multiple plant families (we defined specialization as the use of a single taxonomic family). But Johnson et al. (2018) classified bees as specialists as their diet consists of only nectar and pollen.

Finally, our attempts to correlate the size of different gene families with ecology suffered from sampling biases in which species had genome assemblies and which gene families were manually annotated. Species skewed heavily towards *Drosophila* and apocritan Hymenoptera, and annotations toward the OR and CYP families (Table S8). To evaluate ecological correlates of gene family expansions and contractions, it is essential to expand both the taxonomic breadth and depth of annotation sampling. Taxa that capture independent ecological transitions (e.g., between herbivory and other diets) would be especially useful, as would replicated groups of closely related species that vary in ecological axes of interest (e.g., specialization or social behavior). By systematically sampling different ranges of divergence times, we can evaluate the extent to which the tempo of gene family size change varies across different gene families. To do so, however, will require high quality, manually curated datasets produced using consistent methods and standards for many different environmentally responsive gene families.

Conclusions

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

619
620 # Materials and methods
621 **Biological material**
622 To minimize the confounding effects heterozygosity has on genome assembly, we sequenced
623 haploid siblings. Like all Hymenoptera, sawflies have haplodiploid sex determination in which
624 males (haploid genomes) emerge from unfertilized eggs and females (diploid genomes) from
625 fertilized eggs. A virgin female will bear a clutch of all-male offspring with haploid
626 recombinants of the maternal genome. But the individual genomes are not identical, so an
627 assembly derived from a single clutch is akin to a diploid assembly made from a single
628 individual.
629
630 All insects were reared in custom, climate-controlled environmental chambers (18:6 light cycle,
631 22˚C, 70% RH) on jack pine (*Pinus banksiana*) foliage. Our laboratory line of *N. lecontei* was
632 established from multiple larval colonies collected from a mugo pine (*P. mugo*) in Lexington,
633 Kentucky, USA (37°59'01.6"N 84°30'38.8"W; population ID: RB017). For the transcriptome,
634 adults and larvae were collected from the first lab-reared generation; both were stored at -80˚C.
635 For the genome assembly, the founding population was propagated in the lab for two
636 generations, followed by brother-sister matings for an additional two generations. At this point, a
637 single, virgin, adult female (I2G2-V, 4th generation in the lab) was allowed to lay unfertilized
638 eggs onto jack pine seedlings. The offspring (haploid male brothers from an inbred mother) were
639 reared until the eonymph (prepupal) life stage, at which point they were isolated without food for
640 24 hours prior to preservation in absolute ethanol at -20˚C. Although eonymphs are non-feeding,
641 they were starved to ensure the gut contents were completely voided.
642
643 **Sample preparation and sequencing**
644 ***Genomic DNA***
645 Whole eonymph bodies were individually frozen inside microcentrifuge tubes with liquid
646 nitrogen and ground with pestles made from 1-mL micropipette tips; the resulting powder was
647 incubated in CTAB buffer supplemented with proteinase K and RNase A. After PCI extraction
648 and ethanol precipitation, the precipitate was dried overnight before being resuspended in TE
649 buffer. DNA integrity was assessed with 0.7% agarose gel, purity was measured with the
650 260/280 ratio, and concentration was measured with a Quant-iT dsDNA High-Sensitivity
651 fluorescence assay (Thermo Fisher Scientific).
652
653 The HudsonAlpha Genomic Services Lab (Huntsville, AL, USA) prepared and sequenced the
654 DNA libraries. Two small-insert, barcoded libraries with average fragment sizes of 337 bp and
655 864 bp were made from a single individual. A 4.6-kbp mate-pair, barcoded library was made
656 from 25 pooled individuals. All individuals were brothers from the same I2G2-V mother. The
657 libraries were sequenced on Illumina HiSeq 2000 with paired-end, 100 bp (PE100) reads: the
658 small-insert libraries each had ¼ of a flow cell lane and the mate-pair library had an entire lane.
659
660 ***mRNA***
661 The RNeasy Mini extraction kit (Qiagen) was used to collect total RNA from adult female body,
662 adult female head, adult male body, adult male head, eonymph body, feeding larval body, and
663 feeding larval head. RNA from eonymph head was extracted but not sequenced due to
664 insufficient yield. Each tissue was represented with one replicate that had equal RNA

contributions from eight individuals, except for eonymph body which was comprised of three individuals. RNA integrity and concentration were measured with a 2100 Bioanalyzer (Agilent).

The HudsonAlpha Genomic Services Lab (Huntsville, AL, USA) handled library preparation and sequencing. Non-stranded, barcoded libraries were made for each of the seven tissue samples; on average, mRNA was sheared to 200 bp. The libraries were combined and sequenced on an entire flow cell of Illumina MiSeq with PE250 reads in addition to one lane of Illumina HiSeq 2000 with PE50 reads.

Read processing and assembly

De novo genome assembly

Sequencing reads were chastity-filtered and adaptor-trimmed with fastq-mcf (ea-utils v1.04.803) (Aronesty 2011), and quality-filtered with fastq_quality_filter (FASTX Toolkit v0.0.13.2) (Gordon and Hannon 2019). The 337-bp small-insert reads and the 4.6-kbp mate-pair reads were quality-filtered to retain reads where at least 80% of the bases had a quality score of 20 or higher (parameters: `-q 20 -p 80`). Due to sequencing quality, the 864-bp small-insert reads were filtered to retain reads where at least 70% of the bases had a quality score of 20 or higher (R1) or 60% (R2) (parameters: `-q 20 -p 60/70`). In situations where only one end of the paired-end reads passed filtering, the passed reads were kept and treated as single-end data. Kmer counting was used to measure read depth before and after filtering (Jellyfish v1.1.11) (Marçais and Kingsford 2011). Finally, reads were screened for sequencing contamination by mapping the reads (BWA v0.7.12-r1039) (Li and Durbin 2009) to reference genomes for *Escherichia coli* (K12 substr. DH10B uid58979), human (v37), loblolly pine (*Pinus taeda*, v0.8), and *Wolbachia* (endosymbiont of Dmel uid57851).

The genome was assembled with ALLPATHS-LG (v47417) (Gnerre et al. 2011) using default settings, including a minimum scaffold size of 1000 bp. The error-correction module was run on the reads prior to assembly. After assembly, GapFiller (v1.11) (Boetzer and Pirovano 2012) was used to help close intra-scaffold gaps. Spurious scaffolds were identified with SOAP.coverage (v2.7.7) (Li et al. 2009): reads were mapped to the assembly scaffolds and scaffolds with a read depth < 15 and nucleotide percentage < 40 were removed. The completeness of the final assembly was measured with CEGMA (v2.5) (Parra et al. 2007) and BUSCO (v1.22) (Simão et al. 2015) benchmarks. BUSCO was run with the arthropoda-25oct16 database (parameters: `--long`).

De novo transcriptome assembly

For both the PE250 MiSeq and the PE50 HiSeq reads, fastq-mcf (ea-utils v.1.04.803) (Aronesty 2011) was used for chastity filtering and Trimmomatic (v0.32) (Bolger et al. 2014) was used to adaptor clip, trim, and quality-filter. The PE250 MiSeq reads were processed with the Trimmomatic parameters `ILLUMINACLIP: 2:15:5, HEADCROP: 10, CROP: 60, MINLEN: 60, AVGQUAL: 25` whereas the PE50 HiSeq reads were processed with `ILLUMINACLIP: 2:15:5, HEADCROP: 15, MINLEN: 35, AVGQUAL: 25`. Because the mRNA libraries had an average insert size of 200 bp, the MiSeq reads required extensive adaptor trimming. Reads were screened for contamination as described in *De novo genome assembly*.

710 For each tissue, transcriptomes were assembled with Trinity (r2013_08_14) (Grabherr et al.
711 2011; Haas et al. 2013) using default settings and the `--jaccard_clip` option. Spurious sequences
712 were identified by mapping the sequencing reads to the assembled transcripts with RSEM
713 (v1.2.18) (Li and Dewey 2011); transcripts with either FPKM or TPM values < 1 were removed.
714 After filtering, the transcriptomes were combined, and duplicate sequences were removed.
715
716 **Genome size estimation**
717 Flow cytometry was performed as described in (Harper et al. 2016). For this analysis, we used
718 adult males and females from a lab line of *N. lecontei* established from a colony collected in
719 Auburn, GA (33°59'22.4" N, 83°47'44.6"W; population ID: RB027). Briefly, cell nuclei were
720 collected from the heads of 7 individuals (4 female, 3 male) and stained with propidium iodide.
721 Mean fluorescence for each sample was measured with a BD FACSCalibur flow cytometer (BD
722 Biosciences) and compared to two external standards: *Drosophila melanogaster* (adult female
723 heads, 1C = 175 Mbp) and *Gallus gallus domesticus* (CEN singlets from BioSure, Grass Valley,
724 CA, 1C = 1222.5 Mbp). To correct for ploidy differences between haploid males and diploid
725 standards, we multiplied the *N. lecontei* male estimates by 2. To obtain a single size estimate for
726 each *N. lecontei* sample, we averaged values obtained for the two standards.

727 **Repeat annotation**
728 The *N. lecontei* genome assembly was masked with a custom repeat library. A lineage-specific
729 de novo repeat library was made with RepeatModeler (v1.0.7) (Smit and Hubley 2008-2015) and
730 combined with the hymenopteran repetitive element database (Nov. 2013) from Repbase (Jurka
731 et al. 2005). The custom library was used by RepeatMasker (v4.0.3) (Smit et al. 2013-2015)
732 (parameters: `-cutoff 250 -s -pa 15 -gc 40 -a -poly`) to identify and mask repetitive elements
733 in the genome, including low-complexity DNA and simple repeats.
734
735 Transposable element (TE) family consensus sequences were identified by rerunning
736 RepeatModeler (Smit and Hubley 2008-2015) on the genome assembly using the "ncbi" search
737 engine. The resulting sequences were provided to RepeatMasker (Smit et al. 2013-2015) as a
738 custom library to locate associated TE copies in the genome (parameters: `-gc 40 -cutoff 250 -`
739 `gff -gccalc -norna -nolow -no_is -poly`). TE families with at least 10 fragments longer than
740 100 bp were extracted for further analysis.
741
742 The sequencing reads were mapped to a concatenation of the masked genome and the consensus
743 TE sequences (BWA MEM (parameters: `-M`) (Li and Durbin 2009)). Families that had at least 1x
744 the median coverage to the reference genome for at least 80% of their sequence (to support at
745 least one full insertion found by RepeatModeler) and at least 2x the maximum coverage of the
746 reference genome (to support multiple insertions of the family) were extracted with
747 genomeCoverageBed (BEDtools (Quinlan and Hall 2010)). We attempted to identify the
748 consensus sequences with BLASTN and BLASTX (Altschul et al. 1990) searches against a
749 database of repeat elements, but the only hits were to the lineage-specific elements identified by
750 RepeatModeler. Sequences were also filtered for BLAST hits to rRNA or mitochondrial
751 sequences.
752
753 We also used dnaPipeTE (Goubert et al. 2015) to identify what proportion of our short reads was
754 composed of repetitive content, we used a random subset of reads corresponding to 1-fold
755 coverage of the genome (331Mb) and took the total for three separate random samplings of reads

(parameters: genome size = 331000000 genome coverage = 1 samples number = 3). We then compared this annotation to the RepeatModeler annotation.

Gene and functional annotation

Automated gene annotation

RNA-Seq data for *N. lecontei* was used to generate training models for gene prediction along with utilization of peptide sequences from other species. PASA (r20130425beta) was used to build a comprehensive transcriptome set from Trinity assembled transcripts along with RNA-Seq read mapping predictions generated from the Tuxedo pipeline. To improve annotation quality, in addition to this *N. lecontei* transcriptome, annotated proteins from *Atta cephalotes* (OGSv1.2), *Acromyrmex echinatior* (OGSv3.8), *Apis melifera* (OGSv3.2), *Athalia rosae* (OGSv1.0), and *Nasonia vitripennis* (OGSv1.0) were provided to Maker (2.09) (Cantarel et al. 2008) as evidence for structural gene prediction. Prior to annotation, the genome was masked using a custom repeat database built using RepeatModeler (v1.0.8) and the annotation was run using the *ab initio* gene predictors Augusts, Genemark-ES and snap in addition to the evidence provided. The functions of the predicted protein-coding genes were putatively established with BLASTP alignments (Altschul et al. 1990) to the Swiss-Prot database (accessed 20 Apr 12) (Apweiler et al. 2004). In cases of multiple matches, the top-ranked alignment was assigned to the gene annotation. Protein motifs and functional domains within the annotations were also identified with an InterProScan (v5.3.46.0) (Jones et al. 2014) search against the InterPro database with gene ontology and IPR lookup (Finn et al. 2016). For the official gene set (OGS), the Maker annotations were filtered by hits to the reference databases and/or a minimum eAED score of 0.1. A second set of gene annotations was generated with the NCBI GNOMON pipeline (annotation release 100 on Nlec1.0 assembly, GCF_001263575.1) (Souvorov et al. 2010).

Because the genome was annotated prior to submission to NCBI, we encountered a problem when the NCBI contamination software flagged vector/adaptor sequences for removal; this would disrupt the coordinates provided by Maker. We used a modified version of GAG (Hall et al. 2014) that could accept the flagged coordinates from NCBI to edit the assembly and update annotation coordinates accordingly.

Chemoreceptor genes

The olfactory (OR) and gustatory (GR) receptor genes were manually curated following Robertson et al. (2003, 2006). Amino acid sequences of manually curated chemoreceptor genes from *Apis mellifera* (Robertson and Wanner 2006; Smith et al. 2011), *Bombus terrestris* (Sadd et al. 2015) and *Cephus cinctus* (Robertson et al. 2018), *Drosophila melanogaster* (Flybase release FB2017_04), and *Nasonia vitripennis* (Robertson et al. 2010) were used as queries in TBLASTN (v2.2.19) (Altschul et al. 1990) searches against the *N. lecontei* draft genome (parameters: `-e 100000 -F F`). Gene models were manually built in TextWrangler (v5.5) (Bare Bones Software), using protein alignment to identify exons and refine the gene structures; alignments were visualized with Clustal X (v2.1) (Larkin et al. 2007). The Neural Network Splice Predictor program from the Berkeley *Drosophila* Genome Project was used to help identify intron splice sites (http://www.fruitfly.org/seq_tools/splice.html). New gene models were added to subsequent TBLASTN searches and this process continued iteratively until new chemoreceptors were no longer found. The gene models were checked against RNAseq reads from tissue-specific

802 transcriptomes (adult antennae, mouthparts, heads, legs, genitalia, and larval heads (Herrig et al.
803 2019)) and against orthologs in the *N. pinetum* draft genome assembly (NCBI accession
804 GCA_004916985.1).
805
806
807 ***Odorant binding proteins***
808 Custom scripts were used to identify Maker gene annotations (see *Automated gene annotation*)
809 that contained the classic/6C, Plus-C, Minus-C, or atypical odorant binding protein (OBP) motif
810 (Xu et al. 2009). These as well as OBPs from *Apis mellifera* and *Nasonia vitripennis* were used
811 as queries for TBLASTN searches against the *N. lecontei* genome; searches did not yield any
812 new OBPs. All genomic regions identified as potential OBPs were manually curated as described
813 for chemoreceptor genes. After manual annotation, duplicate annotations or genes that lacked
814 OBP motifs were removed.
815
816 ***Cytochrome P450 genes***
817 A broad set of 52 insect CYP genes (covering the diversity of insect CYP families) were
818 searched against the *N. lecontei* genome assembly (E-value cutoff 1e3). Scaffolds with hits were
819 then searched against 8782 known insect CYPs. The top 10 hits were returned (later increased to
820 15 to recover more sequences) and filtered for duplicates. An alternative search of the NCBI
821 GNOMON predictions ("Neodiprion lecontei[orgn] AND P450 NOT reductase") was also
822 performed and new sequences were added to the dataset. This approach found all the loci
823 identified by the initial search, indicating that the GNOMON annotation tool was able to
824 comprehensively search for CYP sequences. Finally, the candidate *N. lecontei* CYP sequences
825 were manually curated based on comparison to the best BLAST hits.
826
827 ***Immune-related genes***
828 Because of the relative completeness of its immune annotation, *Drosophila melanogaster*
829 immunity genes were used to guide annotation. Reference immune genes from *D. melanogaster*
830 tagged with the gene ontology term "GO:0002376 – Immune system process" were compiled
831 from Flybase (release 6.13). Orthology with *N. lecontei* proteins was assigned initially with
832 reciprocal BLASTP (Altschul et al. 1990) searches (E-value cutoff 1e-10). Reference *D.*
833 *melanogaster* genes without obvious one-to-one orthologs in *N. lecontei* were examined
834 individually to determine whether closely related paralogs in one or both species interfered with
835 the inference of orthology. If not, they were searched against the *N. lecontei* genome assembly
836 using TBLASTN (Altschul et al. 1990) in an attempt to identify unannotated orthologs.
837
838 Since antimicrobial peptides (AMP) are unlikely to be conserved between *D. melanogaster* and
839 *N. lecontei*, AMPs from three representative hymenopterans *Apis mellifera* (Danihlík et al.
840 2015), *Nasonia vitripennis* (Tian et al. 2010), and *Camponotus floridanus* (Ratzka et al. 2012;
841 Zhang and Zhu 2012; Gupta et al. 2015) were used for BLAST queries. Furthermore, since AMP
842 copy number is fast evolving, we attempted to find all the *N. lecontei* orthologs of each
843 hymenopteran AMP instead of focusing on one-to-one orthology. Once again, BLASTP searches
844 were performed against the annotated proteins and TBLASTN searches were performed against
845 the assembled genome; the TBLASTN search did not reveal additional AMPs. Putative *N.*
846 *lecontei* orthologs were reciprocally blasted against the appropriate hymenopteran proteome to
847 assure that the best hits were indeed AMPs.

848
849 Amino acid and cDNA sequences for chemoreceptor genes, CYP genes, and AMP genes are
850 available in File S1.
851
852 **Glomeruli counts**
853 ***Antennal lobe histology***
854 Whole heads of adult *N. lecontei* of both sexes were fixed in 2% paraformaldehyde, 2%
855 glutaraldehyde in PBS for 5 days. Heads were rinsed for 40 minutes three times and the brains
856 dissected out in cold PBS. Following blocking with goat serum, brains were permeabilized with
857 1% Triton X-100 in PBS (Electron Microscopy Supply, Fort Washington, PA; PBS-TX), rinsed
858 with 0.1% PBS-TX, and incubated on a shaker at 25°C for three nights in primary antibody
859 (1:500 in 2% goat serum in 0.2% PBS-TX). Monoclonal *Drosophila* synapsin I antibody
860 (SYNORF1, AB_2315426) from the Developmental Studies Hybridoma Bank (catalog 3C11)
861 was used to label synapsin. Subsequently, brains were washed in 0.1% PBS-TX and incubated
862 for two nights in Alexa Fluor 568 (ThermoFisher) goat anti-mouse secondary antibody (1:100 in
863 PBS) in the dark at room temperature on a shaker. After secondary incubation, brains were
864 rinsed with distilled water, dehydrated in increasing concentrations of ethanol, and mounted in
865 custom-made aluminum well slides. Brains were cleared by removing ethanol and replacing it
866 with methyl salicylate. Brains were imaged on an inverted Zeiss 880 Laser Scanning Confocal
867 Microscope with a 20X plan-Apochromat 20x 0.8 aperture objective and optically sectioned in
868 the horizontal plane at 3-micron intervals.
869
870 ***Glomeruli segmentation***
871 Whole-brain images of one female and one male were manually segmented using the TrakEM2
872 software package in ImageJ (Cardona et al. 2012; Schindelin et al. 2012). Individual glomeruli
873 were traced in both brain hemispheres. Glomeruli near the center of the antennal lobe can be
874 difficult to distinguish, meaning counts are biased toward fewer glomeruli. Therefore, the largest
875 number of glomeruli confidently detected represents a minimum of the number of expected
876 glomeruli. Male *Neodiprion* have a collection of smaller synaptic clusters in their antennal lobe
877 (Dacks and Nighorn 2011), but the functional significance of this anatomy is not known. There
878 are more than 50 of these smaller synaptic clusters and we suspect they do not represent the
879 traditional one-to-one OR-to-glomerulus organization. Therefore, these structures were not
880 included in counts. Male glomeruli number may be lower if particular OSNs contribute to these
881 clusters instead of forming traditional glomeruli.
882
883
884 **Within-genome signatures of adaptive expansions and contractions**
885 ***Clustering and pseudogene analyses***
886 To evaluate the extent to which members of our five focal gene families were located in tandem
887 arrays, we placed our annotated genes on a linkage-map anchored version of the *N. lecontei*
888 genome assembly described in Linnen et al. 2018. We considered members of a gene family to
889 be clustered if they were located within a genomic region of 20(n - 1) kilobases, where n is the
890 number of genes in the cluster under consideration. This criterion was chosen based on average
891 gene densities in *Nasonia* (Niehuis et al. 2010) and clustering criteria described *Drosophila*
892 (Vieira et al. 2007). For scaffolds that could not be placed on linkage groups, we evaluated
893 clustering only if genes were more than 20 kb from either end of the scaffold.

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

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

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

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

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

For branch tests, the cDNA phylogenies for each *N. lecontei* gene family were used to compare the lineage-specific clade to the rest of the gene family. To determine if the dN/dS of the foreground branch (i.e., the branch with the species-specific expansion) was significantly different from the background (i.e., the rest of the gene family), in codeml we ran a two-ratio

model (Model=2, fix_omega=0) and a one-ratio 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 chemoreceptor, odorant binding protein, cytochrome P450, and antimicrobial peptide gene annotations. If fasta sequence files were available, the number of intact, partial, and pseudogenized genes was determined based on gene names (e.g., labels with "pse" or "partial"); values were compared to those 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.

For comparisons, only the number of putatively functional (intact) genes was used. 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. Gene family size 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 and Erin Scully for advice. We are especially grateful to Hugh Robertson for his guidance on annotating ORs and GRs. This work was supported by the University of Kentucky Center for Computational Sciences and the Lipscomb High Performance Computing Cluster, the United States Department of Agriculture National Institute of Food and Agriculture (2016-67014-2475; CRL), the Kentucky Science and Engineering

Foundation (KSEF-3492-RDE-019; CRL), and the University of Kentucky (Lyman T. Johnson Fellowship; KV).

Data availability

The genome assembly, official gene set (OGS), and transcriptome described in this paper (v1 versions) can be found at https://i5k.nal.usda.gov/neodiprion-lecontei

On GenBank (NCBI), the genome assembly is labeled whole genome shotgun sequencing project accession PRJNA28045 and the genomic sequencing reads are RefSeq accession PRJNA312506. The transcriptome is transcriptome shotgun assembly accession GEDM00000000; this is a combined transcriptome file with sequences from all seven tissue types. The mRNA sequencing reads for each tissue type was submitted separately under BioSample and short read archive accessions SAMN04302192 (adult female head), SAMN04302193 (adult female body), SAMN04302194 (adult male head), SAMN04302195 (adult male body), SAMN04302196 (feeding larval head), SAMN04302197 (feeding larval body), and SAMN04302198 (eonymph body). The predicted gene annotations on NCBI are from Gnomon, the NCBI annotation pipeline, and were not described in this paper. Finally, the clustering analysis was based on a linkage-map anchored version of the genome assembly described in Linnen et al. 2018. This anchored assembly is denoted as v1.1 in NCBI and the *N. lecontei* i5k Workspace@NAL (USDA).

References

Abascal F, Zardoya R, Telford MJ. 2010. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Res 38(suppl 2):W7-W13.

Akhunov ED, Goodyear AW, Geng S, Qi L-L, Echalier B, Gill BS, Gustafson JP, et al. 2003. The organization and rate of evolution of wheat genomes are correlated with recombination rates along chromosome arms. Genome Res 13(5):753-763.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215(3):403-410.

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, et al. 2004. UniProt: the universal protein knowledgebase. Nucleic Acids Res 32(suppl 1):D115-D119.

Arguello JR, Cardoso-Moreira M, Grenier JK, Gottipati S, Clark AG, Benton R. 2016. Extensive local adaptation within the chemosensory system following Drosophila melanogaster's global expansion. Nat Commun 7(1):1-12.

Aronesty E. 2011. ea-utils : "Command-line tools for processing biological sequencing data".

Berenbaum MR. 2002. Postgenomic chemical ecology: from genetic code to ecological interactions. J Chem Ecol 28(5):873-896.

Berenbaum MR, Johnson RM. 2015. Xenobiotic detoxification pathways in honey bees. Curr Opin Insect Sci 10:51-58.

Boetzer M, Pirovano W. 2012. Toward almost closed genomes with GapFiller. Genome Biol 13(6):R56.

Boevé J-L, Blank SM, Meijer G, Nyman T. 2013. Invertebrate and avian predators as drivers of chemical defensive strategies in tenthredinid sawflies. BMC Evol Biol 13(1):198.

1030 Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence
1031 data. Bioinformatics 30(15):2114-2120.
1032 Brand P, Ramírez SR. 2017. The evolutionary dynamics of the odorant receptor gene family in
1033 corbiculate bees. Genome Biol Evol 9(8):2023-2036.
1034 Briscoe AD, Macias-Munoz A, Kozak KM, Walters JR, Yuan F, Jamie GA, Martin SH, et al.
1035 2013. Female behaviour drives expression and evolution of gustatory receptors in butterflies.
1036 PLoS Genet 9(7):e1003620.
1037 Calla B, Noble K, Johnson RM, Walden KKO, Schuler MA, Robertson HM, Berenbaum MR.
1038 2017. Cytochrome P450 diversification and hostplant utilization patterns in specialist and
1039 generalist moths: Birth, death and adaptation. Mol Ecol 26(21):6021-6035.
1040 Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, Holt C, Alvarado AS, Yandell M.
1041 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model organism
1042 genomes. Genome Res 18(1):188-196.
1043 Cao D, Liu Y, Walker WB, Li J, Wang G. 2014. Molecular characterization of the Aphis
1044 gossypii olfactory receptor gene families. PLoS One 9(6):e101187.
1045 Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated
1046 alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25(15):1972-1973.
1047 Cardona A, Saalfeld S, Schindelin J, Arganda-Carreras I, Preibisch S, Longair M, Tomancak P,
1048 Hartenstein V, Douglas RJ. 2012. TrakEM2 software for neural circuit reconstruction. PLoS
1049 One 7(6):e38011.
1050 Castella G, Chapuisat M, Moret Y, Christe P. 2008. The presence of conifer resin decreases the
1051 use of the immune system in wood ants. Ecological Entomology 33(3):408-412.
1052 Comeault AA, Serrato-Capuchina A, Turissini DA, McLaughlin PJ, David JR, Matute DR.
1053 2017. A nonrandom subset of olfactory genes is associated with host preference in the fruit
1054 fly Drosophila orena. Evolution Letters 1(2):73-85.
1055 Consortium iK. 2013. The i5K Initiative: advancing arthropod genomics for knowledge, human
1056 health, agriculture, and the environment. J Hered, 104(5):595-600.
1057 Couto A, Alenius M, Dickson BJ. 2005. Molecular, anatomical, and functional organization of
1058 the Drosophila olfactory system. Curr Biol 15(17):1535-1547.
1059 Cowan MM. 1999. Plant products as antimicrobial agents. Clin Microbiol Rev 12(4):564-582.
1060 Dacks AM, Nighorn AJ. 2011. The organization of the antennal lobe correlates not only with
1061 phylogenetic relationship, but also life history: a basal hymenopteran as exemplar. Chem
1062 Senses 36(2):209-220.
1063 Dahanukar A, Foster K, Carlson JR. 2001. A Gr receptor is required for response to the sugar
1064 trehalose in taste neurons of Drosophila. Nat Neurosci 4(12):1182-1186.
1065 Danihlík J, Aronstein K, Petřivalský M. 2015. Antimicrobial peptides: a key component of
1066 honey bee innate immunity: Physiology, biochemistry, and chemical ecology. Journal of
1067 Apicultural Research 54(2):123-136.
1068 Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models
1069 of protein evolution. Bioinformatics 27(8):1164-1165.
1070 Dayhoff MO. 1976. The origin and evolution of protein superfamilies Federation Proceedings
1071 35(10):2132-2138.
1072 De Bie T, Cristianini N, Demuth JP, Hahn MW. 2006. CAFE: a computational tool for the
1073 study of gene family evolution. Bioinformatics 22(10):1269-1271.
1074 de Brito Sanchez MG. 2011. Taste perception in honey bees. Chem Senses 36(8): 675-692.

1075 Demuth JP, Hahn MW. 2009. The life and death of gene families. Bioessays 31(1):29-39.
1076 Després L, David J-P, Gallet C. 2007. The evolutionary ecology of insect resistance to plant
1077 chemicals. Trends Ecol Evol 22(6):298-307.
1078 Dobler S, Dalla S, Wagschal V, Agrawal AA. 2012. Community-wide convergent evolution in
1079 insect adaptation to toxic cardenolides by substitutions in the Na, K-ATPase. Proc Natl
1080 Acad Sci USA 109(32):13040-13045.
1081 Duffy JE, Macdonald KS. 2010. Kin structure, ecology and the evolution of social organization
1082 in shrimp: a comparative analysis. Proc Royal Soc B 277(1681):575-584.
1083 Engsontia P, Sangket U, Robertson HM, Satasook C. 2015. Diversification of the ant odorant
1084 receptor gene family and positive selection on candidate cuticular hydrocarbon receptors.
1085 BMC Res Notes 8(1):380.
1086 Evans JD, Aronstein K, Chen YP, Hetru C, Imler JL, Jiang H, Kanost M, Thompson GJ, Zou
1087 Z, Hultmark D. 2006. Immune pathways and defence mechanisms in honey bees Apis
1088 mellifera. Insect Mol Biol 15(5):645-656.
1089 Faulkes CG, Bennett NC, Bruford MW, O'Brien HP, Aguilar GH, Jarvis JU. 1997. Ecological
1090 constraints drive social evolution in the African mole–rats. Proc Royal Soc B
1091 264(1388):1619-1627.
1092 Feyereisen R. 2011. Arthropod CYPomes illustrate the tempo and mode in P450 evolution.
1093 Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics 1814(1):19-28.
1094 Feyereisen R. 2012. Insect CYP genes and P450 enzymes. In. Insect Mol Biol and Biochemistry:
1095 Academic Press. p. 236-316.
1096 Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, et al. 2016.
1097 The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res
1098 44(D1):D279-D285.
1099 Fischman BJ, Woodard SH, Robinson GE. 2011. Molecular evolutionary analyses of insect
1100 societies. Proc Natl Acad Sci USA 108(suppl 2):10847-10854.
1101 Fishilevich E, Vosshall LB. 2005. Genetic and functional subdivision of the Drosophila antennal
1102 lobe. Curr Biol 15(17):1548-1553.
1103 Forister ML, Dyer LA, Singer MS, Stireman III JO, Lill JT. 2012. Revisiting the evolution of
1104 ecological specialization, with emphasis on insect–plant interactions. Ecology 93(5):981-991.
1105 Gao Q, Yuan B, Chess A. 2000. Convergent projections of Drosophila olfactory neurons to
1106 specific glomeruli in the antennal lobe. Nat Neurosci 3(8):780-785.
1107 Gardiner A, Barker D, Butlin RK, Jordan WC, Ritchie MG. 2008. Drosophila chemoreceptor
1108 gene evolution: selection, specialization and genome size. Mol Ecol 17(7):1648-1657.
1109 Gaut BS, Wright SI, Rizzon C, Dvorak J, Anderson LK. 2007. Recombination: an
1110 underappreciated factor in the evolution of plant genomes. Nat Rev Genet 8(1):77-84.
1111 Gershenzon J, Dudareva N. 2007. The function of terpene natural products in the natural world.
1112 Nat Chem Biol 3(7):408-414.
1113 Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, et al. 2011.
1114 High-quality draft assemblies of mammalian genomes from massively parallel sequence data.
1115 Proc Natl Acad Sci USA 108(4):1513-1518.
1116 Goldman-Huertas B, Mitchell RF, Lapoint RT, Faucher CP, Hildebrand JG, Whiteman NK.
1117 2015. Evolution of herbivory in Drosophilidae linked to loss of behaviors, antennal responses,
1118 odorant receptors, and ancestral diet. Proc Natl Acad Sci USA 112(10):3026-3031.

1119 Good RT, Gramzow L, Battlay P, Sztal T, Batterham P, Robin C. 2014. The molecular
1120 evolution of cytochrome P450 genes within and between Drosophila species. Genome Biol
1121 Evol 6(5):1118-1134.
1122 Gordon A, Hannon GJ. 2019. "Fastx-toolkit" FASTQ/A short-reads preprocessing tools
1123 (unpublished).
1124 Goubert C, Modolo L, Vieira C, Claire ValienteMoro, Mavingui P, Boulesteix M. 2015. De novo
1125 assembly and annotation of the Asian tiger mosquito (Aedes albopictus) repeatome with
1126 dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever
1127 mosquito (Aedes aegypti). Genome Biol Evol 7(4):1192-1205.
1128 Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, et al. 2011.
1129 Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat
1130 Biotechnol 29(7):644-652.
1131 Grayer RJ, Harborne JB. 1994. A survey of antifungal compounds from higher plants, 1982–
1132 1993. Phytochemistry 37(1):19-42.
1133 Gross JB, Borowsky R, Tabin CJ. 2009. A novel role for Mc1r in the parallel evolution of
1134 depigmentation in independent populations of the cavefish Astyanax mexicanus. PLoS Genet
1135 5(1):e1000326.
1136 Guittard E, Blais C, Maria A, Parvy J-P, Pasricha S, Lumb C, Lafont R, Daborn PJ, Dauphin-
1137 Villemant C. 2011. CYP18A1, a key enzyme of Drosophila steroid hormone inactivation, is
1138 essential for metamorphosis. Dev Bio l349(1):35-45.
1139 Gupta SK, Kupper M, Ratzka C, Feldhaar H, Vilcinskas A, Gross R, Dandekar T, Förster F.
1140 2015. Scrutinizing the immune defence inventory of Camponotus floridanus applying total
1141 transcriptome sequencing. BMC Genom 16(1):1-21.
1142 Ha E-M, Oh C-T, Bae YS, Lee W-J. 2005. A direct role for dual oxidase in Drosophila gut
1143 immunity. Science 310(5749):847-850.
1144 Haas BJ, Papanicolaou A, Yassour M, Grabherr MG, Blood PD, Bowden J, Couger MB, et al.
1145 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform
1146 for reference generation and analysis. Nat Protoc 8(8):1494-1512.
1147 Hahn MW, de Bie T, Stajich JE, Nguyen C, Cristianini N. 2005. Estimating the tempo and
1148 mode of gene family evolution from comparative genomic data. Genome Res 15(8):1153-
1149 1160.
1150 Hahn MW, Han MV, Han S-G. 2007. Gene family evolution across 12 Drosophila genomes.
1151 PLoS Genet 3(11):e197.
1152 Hall B, DeRego T, Geib SM. 2014. GAG: the genome annotation generator (version 1.0).
1153 Han MV, Thomas GW, Lugo-Martinez J, Hahn MW. 2013. Estimating gene gain and loss rates
1154 in the presence of error in genome assembly and annotation using CAFE 3. Mol Biol Evol
1155 30(8):1987-1997.
1156 Hanson MA, Lemaitre B, Unckless RL. 2019. Dynamic evolution of antimicrobial peptides
1157 underscores trade-offs between immunity and ecological fitness. Front Immunol 10:2620.
1158 Harper KE, Bagley RK, Thompson KL, Linnen CR. 2016. Complementary sex determination,
1159 inbreeding depression and inbreeding avoidance in a gregarious sawfly. Heredity 117(5):326-
1160 335.
1161 Helvig C, Koener JF, Unnithan GC, Feyereisen R. 2004. CYP15A1, the cytochrome P450 that
1162 catalyzes epoxidation of methyl farnesoate to juvenile hormone III in cockroach corpora
1163 allata. Proc Natl Acad Sci USA 101(12):4024-4029.

1164 Herrig DK, Vertacnik KL, Linnen CR. 2019. Testing the adaptive decoupling hypothesis in a
1165 hypermetamorphic and sexually dimorphic insect. bioRxiv 2019.12.20.882803.
1166 Himejima M, Hobson KR, Otsuka T, Wood DL, Kubo I. 1992. Antimicrobial terpenes from
1167 oleoresin of ponderosa pine tree Pinus ponderosa: A defense mechanism against microbial
1168 invasion. J Chem Ecol 18(10):1809-1818.
1169 Hughes AL, Nei M. 1992. Maintenance of MHC polymorphism. Nature 335:402–403.
1170 Johnson RM, Harpur BA, Dogantzis KA, Amro Z, Berenbaum MR. 2018. Genomic footprint of
1171 evolution of eusociality in bees: floral food use and CYPome "blooms". Insectes Soc
1172 65(3):445-454.
1173 Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, et al. 2014.
1174 InterProScan 5: genome-scale protein function classification. Bioinformatics 30(9):1236-1240.
1175 Jones WD, Cayirlioglu P, Kadow IG, Vosshall LB. 2007. Two chemosensory receptors together
1176 mediate carbon dioxide detection in Drosophila. Nature 445(7123):86-90.
1177 Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase
1178 Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110(1-4):462-
1179 467.
1180 Karev GP, Wolf YI, Rzhetsky AY, Berezovskaya FS, Koonin EV. 2002. Birth and death of
1181 protein domains: a simple model of evolution explains power law behavior. BMC Evol Biol
1182 2(1):18.
1183 Keeling CI, Yuen MMS, Liao NY, Docking TR, Chan SK, Taylor GA, Palmquist DL, et al.
1184 2013. Draft genome of the mountain pine beetle, Dendroctonus ponderosae Hopkins, a major
1185 forest pest. Genome Biol 14(3):R27.
1186 Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F,
1187 et al. 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23(21):2947-2948.
1188 LeBoeuf AC, Benton R, Keller L. 2013. The molecular basis of social behavior: models, methods
1189 and advances. Curr Opin Neurobiol 23(1):3-10.
1190 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.
1191 2015. Bacterial uracil modulates Drosophila DUOX-dependent gut immunity via Hedgehog-
1192 induced signaling endosomes. Cell Host Microbe 17(2):191-204.
1193 Lee Y, Moon SJ, Montell C. 2009. Multiple gustatory receptors required for the caffeine response
1194 in Drosophila. Proc Natl Acad Sci USA 106(11):4495-4500.
1195 Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or
1196 without a reference genome. BMC Bioinform 12(1):323.
1197 Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform.
1198 Bioinformatics 25(14):1754-1760.
1199 Li R, Yu C, Li Y, Lam T-W, Yiu S-M, Kristiansen K, Wang J. 2009. SOAP2: an improved
1200 ultrafast tool for short read alignment. Bioinformatics 25(15):1966-1967.
1201 Li W, Schuler MA, Berenbaum MR. 2003. Diversification of furanocoumarin-metabolizing
1202 cytochrome P450 monooxygenases in two papilionids: specificity and substrate encounter
1203 rate. Proc Natl Acad Sci USA 100(suppl 2):14593-14598.
1204 Li X, Schuler MA, Berenbaum MR. 2007. Molecular mechanisms of metabolic resistance to
1205 synthetic and natural xenobiotics. Annu Rev Entomol 52:231-253.
1206 Linnen CR, O'Quin CT, Shackleford T, Sears CR, Lindstedt C. 2018. Genetic basis of body
1207 color and spotting pattern in redheaded pine sawfly larvae (Neodiprion lecontei). Genetics
1208 209(1):291-305.

1209 Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of
1210 occurrences of k-mers. Bioinformatics 27(6):764-770.
1211 Martin A, Orgogozo V. 2013. The loci of repeated evolution: a catalog of genetic hotspots of
1212 phenotypic variation. Evolution 67(5):1235-1250.
1213 Matsuo T, Sugaya S, Yasukawa J, Aigaki T, Fuyama Y. 2007. Odorant-binding proteins
1214 OBP57d and OBP57e affect taste perception and host-plant preference in Drosophila
1215 sechellia. PLoS Biology 5(5):e118.
1216 Maxwell DE editor. Proceedings of the 10th International Congress of Entomology. 1958
1217 Montreal, Canada.
1218 McBride CS. 2007. Rapid evolution of smell and taste receptor genes during host specialization
1219 in Drosophila sechellia. Proc Natl Acad Sci USA 104(12):4996-5001.
1220 McBride CS, Arguello RJ. 2007. Five Drosophila genomes reveal nonneutral evolution and the
1221 signature of host specialization in the chemoreceptor superfamily. Genetics 177(3):1395-1416.
1222 McKenna DD, Scully ED, Pauchet Y, Hoover K, Kirsch R, Geib SM, Mitchell RF, et al. 2016.
1223 Genome of the Asian longhorned beetle (Anoplophora glabripennis), a globally significant
1224 invasive species, reveals key functional and evolutionary innovations at the beetle–plant
1225 interface. Genome Biol 17:227
1226 McKenzie SK, Fetter-Pruneda I, Ruta V, Kronauer DJC. 2016. Transcriptomics and
1227 neuroanatomy of the clonal raider ant implicate an expanded clade of odorant receptors in
1228 chemical communication. Proc Natl Acad Sci USA 113(49):14091-14096.
1229 Mittapelly P, Bansal R, Michel A. 2019. Differential expression of cytochrome P450 CYP6 genes
1230 in the brown marmorated stink bug, Halyomorpha halys (Hemiptera: Pentatomidae). J Econ
1231 Entomol 112(3):1403-1410.
1232 Miyamoto T, Slone J, Song X, Amrein H. 2012. A fructose receptor functions as a nutrient
1233 sensor in the Drosophila brain. Cell 151(5):1113-1125.
1234 Moon SJ, Köttgen M, Jiao Y, Xu H, Montell C. 2006. A taste receptor required for the caffeine
1235 response in vivo. Curr Biol 16(18):1812-1817.
1236 Moon SJ, Lee Y, Jiao Y, Montell C. 2009. A Drosophila gustatory receptor essential for aversive
1237 taste and inhibiting male-to-male courtship. Curr Biol 19(19):1623-1627.
1238 Mumm R, Hilker M. 2006. Direct and indirect chemical defence of pine against folivorous
1239 insects. Trends Plant Sci 11(7):351-358.
1240 Nei M. 2007. The new mutation theory of phenotypic evolution. Proc Natl Acad Sci USA
1241 104(30):12235-12242.
1242 Nei M, Rooney AP. 2005. Concerted and birth-and-death evolution of multigene families. Annu
1243 Rev Genet 39:121-152.
1244 Niehuis O, Gibson JD, Rosenberg MS, Pannebakker BA, Koevoets T, Judson AK, Desjardins
1245 CA, et al. 2010. Recombination and its impact on the genome of the haplodiploid parasitoid
1246 wasp Nasonia. PLoS One 5(1):e8597.
1247 Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. 2005. Genomic scans
1248 for selective sweeps using SNP data. Genome Res 15(11):1566-1575.
1249 Nozawa M, Kawahara Y, Nei M. 2007. Genomic drift and copy number variation of sensory
1250 receptor genes in humans. Proc Natl Acad Sci USA 104(51):20421-20426.
1251 Ohno S. 1970. The enormous diversity in genome sizes of fish as a reflection of nature's
1252 extensive experiments with gene duplication. Trans Am Fish Soc 99(1):120-130.

1253 Parra G, Bradnam K, Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in
1254 eukaryotic genomes. Bioinformatics 23(9):1061-1067.
1255 Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, et al. 2007. Diet and
1256 the evolution of human amylase gene copy number variation. Nat Genet 39(10):1256-1260.
1257 Peters RS, Krogmann L, Mayer C, Donath A, Gunkel S, Meusemann K, Kozlov A, et al. 2017.
1258 Evolutionary history of the Hymenoptera. Curr Biol 27(7):1013-1018.
1259 Petrov DA, Hartl DL. 1997. Trash DNA is what gets thrown away: high rate of DNA loss in
1260 Drosophila. Gene 205(1-2):279-289.
1261 Petrov DA, Hartl DL. 1998. High rate of DNA loss in the Drosophila melanogaster and
1262 Drosophila virilis species groups. Mol Biol Evol 15(3):293-302.
1263 Petrov DA, Lozovskaya ER, Hartl DL. 1996. High intrinsic rate of DNA loss in Drosophila.
1264 Nature 384(6607):346-349.
1265 Poelchau M, Childers C, Moore G, Tsavatapalli V, Evans J, Lee C-Y, Lin H, Lin J-W, Hacket
1266 K. 2015. The i5k Workspace@ NAL—enabling genomic data access, visualization and
1267 curation of arthropod genomes. Nucleic Acids Res 43(D1):D714-D719.
1268 Protas ME, Hersey C, Kochanek D, Zhou Y, Wilkens H, Jeffery WR, Zon LI, Borowsky R,
1269 Tabin CJ. 2006. Genetic analysis of cavefish reveals molecular convergence in the evolution
1270 of albinism. Nat Genet 38(1):107-111.
1271 Qiu Y, Tittiger C, Wicker-Thomas C, Le Goff G, Young S, Wajnberg E, Fricaux T, Taquet N,
1272 Blomquist GJ, Feyereisen R. 2012. An insect-specific P450 oxidative decarbonylase for
1273 cuticular hydrocarbon biosynthesis. Proc Natl Acad Sci USA 109(37):14858-14863.
1274 Qu Z, Kenny NJ, Lam HM, Chan TF, Chu KH, Bendena WG, Tobe SS, Hui JHL. 2015. How
1275 did arthropod sesquiterpenoids and ecdysteroids arise? Comparison of hormonal pathway
1276 genes in noninsect arthropod genomes. Genome Biol Evol 7(7):1951-1959.
1277 Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic
1278 features. Bioinformatics 26(6):841-842.
1279 R-Core-Team. 2018. R: A language and environment for statistical computing. Vienna, Austria:
1280 R Foundation for Statistical Computing.
1281 Ratzka C, Förster F, Liang C, Kupper M, Dandekar T, Feldhaar H, Gross R. 2012. Molecular
1282 characterization of antimicrobial peptide genes of the carpenter ant Camponotus floridanus.
1283 PLoS One 7(8):e43036.
1284 Rewitz KF, O'Connor MB, Gilbert LI. 2007. Molecular evolution of the insect Halloween family
1285 of cytochrome P450s: phylogeny, gene organization and functional conservation. Insect
1286 Biochem Mol Biol 37(8):741-753.
1287 Rizzon C, Ponger L, Gaut BS. 2006. Striking similarities in the genomic distribution of
1288 tandemly arrayed genes in Arabidopsis and rice. PLoS Comput Biol 2(9):e115.
1289 Robertson HM, Gadau J, Wanner KW. 2010. The insect chemoreceptor superfamily of the
1290 parasitoid jewel wasp Nasonia vitripennis. Insect Mol Biol 19:121-136.
1291 Robertson HM, Kent LB. 2009. Evolution of the gene lineage encoding the carbon dioxide
1292 receptor in insects. J Insect Sci 9(1):19.
1293 Robertson HM, Wanner KW. 2006. The chemoreceptor superfamily in the honey bee, Apis
1294 mellifera: expansion of the odorant, but not gustatory, receptor family. Genome Res
1295 16(11):1395-1403.
1296 Robertson HM, Waterhouse RM, Walden KKO, Ruzzante L, Reijnders MJMF, Coates BS,
1297 Legeai F, et al. 2018. Genome sequence of the wheat stem sawfly, Cephus cinctus,

representing an early-branching lineage of the Hymenoptera, illuminates evolution of hymenopteran chemoreceptors. Genome Biol Evol 10(11):2997-3011.
Ross L, Gardner A, Hardy N, West SA. 2013. Ecology, not the genetics of sex determination, determines who helps in eusocial populations. Curr Biol 23(23):2383-2387.
Roux J, Privman E, Moretti S, Daub JT, Robinson-Rechavi M, Keller L. 2014. Patterns of positive selection in seven ant genomes. Mol Biol Evol 31(7):1661-1685.
Sadd BM, Barribeau SM, Bloch G, de Graaf DC, Dearden P, Elsik CG, Gadau J, et al. 2015. The genomes of two key bumblebee species with primitive eusocial organization. Genome Biol 16(1):1-32.
Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, Preibisch S, et al. 2012. Fiji: an open-source platform for biological-image analysis. Nature Methods 9(7):676-682.
Sezutsu H, Le Goff G, Feyereisen R. 2013. Origins of P450 diversity. Philos Trans R Soc Lond B Biol Sci 368:20120428.
Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31(19):3210-3212.
Simone M, Evans JD, Spivak M. 2009. Resin collection and social immunity in honey bees. Evolution 63(11):3016-3022.
Singaravelan N, Nee'man G, Inbar M, Izhaki I. 2005. Feeding responses of free-flying honeybees to secondary compounds mimicking floral nectars. J Chem Ecol 31(12):2791-2804.
Slone J, Daniels J, Amrein H. 2007. Sugar receptors in Drosophila. Curr Biol 17(20):1809-1816.
Smadja C, Shi P, Butlin RK, Robertson HM. 2009. Large gene family expansions and adaptive evolution for odorant and gustatory receptors in the pea aphid, Acyrthosiphon pisum. Mol Biol Evol 26(9):2073-2086.
Smit AFA, Hubley R. 2008-2015. RepeatModeler Open-1.0.
Smit AFA, Hubley R, Green P. 2013-2015. RepeatMasker Open-4.0.
Smith CR, Smith CD, Robertson HM, Helmkampf M, Zimin A, Yandell M, Holt C, et al. 2011. Draft genome of the red harvester ant Pogonomyrmex barbatus. Proc Natl Acad Sci USA 108(14):5667-5672.
Smith SG. 1941. A new form of spruce sawfly identified by means of its cytology and parthenogenesis. J Agric Sci 21(5):245-305.
Sohi SS, Ennis TJ. 1981. Chromosomal characterization of cell lines of Neodiprion lecontei (Hymenoptera: Diprionidae). Proc Entomol Soc Ont 112:45–48.
Souvorov A, Kapustin Y, Kiryutin B, Chetvernin V, Tatusova T, Lipman DJ. 2010. Gnomon–NCBI eukaryotic gene prediction tool. National Center for Biotechnology Information.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30(9):1312-1313.
Suzuki HC, Ozaki K, Makino T, Uchiyama H, Yajima S, Kawata M. 2018. Evolution of gustatory receptor gene family provides insights into adaptation to diverse host plants in nymphalid butterflies. Genome Biol Evol 10(6):1351-1362.
Thomas JH. 2006. Analysis of homologous gene clusters in Caenorhabditis elegans reveals striking regional cluster domains. Genetics 172(1):127-143.
Tian C, Gao B, Fang Q, Ye G, Zhu S. 2010. Antimicrobial peptide-like genes in Nasonia vitripennis: a genomic perspective. BMC Genom 11(1):187.

1343 Trapp S, Croteau R. 2001. Defensive resin biosynthesis in conifers. Annu Rev Plant Biol
1344 52(1):689-724.
1345 Vieira FG, Sánchez-Gracia A, Rozas J. 2007. Comparative genomic analysis of the odorant-
1346 binding protein family in 12 Drosophila genomes: purifying selection and birth-and-death
1347 evolution. Genome Biol 8(11):R235.
1348 Vitti JJ, Grossman SR, Sabeti PC. 2013. Detecting natural selection in genomic data. Annu Rev
1349 Genet 4797-120.
1350 Vosshall LB, Wong AM, Axel R. 2000. An olfactory sensory map in the fly brain. Cell
1351 102(2):147-159.
1352 Wanner KW, Robertson HM. 2008. The gustatory receptor family in the silkworm moth
1353 Bombyx mori is characterized by a large expansion of a single lineage of putative bitter
1354 receptors. Insect Mol Biol 17(6):621-629.
1355 Wiens JJ, Lapoint RT, Whiteman NK. 2015. Herbivory increases diversification across insect
1356 clades. Nat Commun 6(1):1-7.
1357 Wilson LF, Wilkinson RC, C AR. 1992. Redheaded pine sawfly: its ecology and management.
1358 In: Service USDoAF, editor. Agriculture Handbook. Washington, DC.
1359 Xu Y-L, He P, Zhang L, Fang S-Q, Dong S-L, Zhang Y-J, Li F. 2009. Large-scale identification
1360 of odorant-binding proteins and chemosensory proteins from expressed sequence tags in
1361 insects. BMC Genom 10(1):632.
1362 Yang J, Chen X, Bai J, Fang D, Qiu Y, Jiang W, Yuan H, et al. 2016. The Sinocyclocheilus
1363 cavefish genome provides insights into cave adaptation. BMC Biol 14(1):1-13.
1364 Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol
1365 24(8):1586-1591.
1366 Zhang J. 2003. Evolution by gene duplication: an update. Trends Ecol Evol 18(6):292-298.
1367 Zhang L, Gaut BS. 2003. Does recombination shape the distribution and evolution of tandemly
1368 arrayed genes (TAGs) in the Arabidopsis thaliana genome? Genome Res 13(12):2533-2540.
1369 Zhang Z, Zhu S. 2012. Comparative genomics analysis of five families of antimicrobial peptide-
1370 like genes in seven ant species. Dev Comp Immunol 38(2):262-274.
1371 Zhen Y, Aardema ML, Medina EM, Schumer M, Andolfatto P. 2012. Parallel molecular
1372 evolution in an herbivore community. Science 337(6102):1634-1637.
1373 Zhou X, Rokas A, Berger SL, Liebig J, Ray A, Zwiebel LJ. 2015. Chemoreceptor evolution in
1374 hymenoptera and its implications for the evolution of eusociality. Genome Biol Evol
1375 7(8):2407-2416.
1376

1377 ## Figure Legends

1378

1379 **Figure 1. Like other diprionids, *N. lecontei* has multiple morphological and behavioral**
1380 **adaptations to *Pinus* foliage. A.** An egg-laying *N. lecontei* female demonstrating several
1381 adaptations for dealing with thick, resinous pine needles, including: a robust saw-like ovipositor
1382 (visible within the needle), a tendency to lay many closely spaced eggs per needles, and a
1383 tendency to cut resin-draining slits on egg-bearing needles (circled). **B.** Prior to hatching, *N.*
1384 *lecontei* eggs absorb water from the host, causing the eggs to swell and the pockets to open.
1385 Throughout development, embryos are in close contact with living host tissue. **C.** Early-instar
1386 larvae have skeletonizing feeding behavior in which only the outer needle tissue is consumed,
1387 leaving the resinous interior intact. This strategy prevents small larvae from being overwhelmed

1388 by sticky resin. **D.** Mid- and late-instar larvae consume the entire pine needle. Larvae sequester
1389 pine resin in specialized pouches for use in self-defense (All photos by R.K. Bagley).
1390
1391 **Figure 2. Optical sections through the antennal lobes of adult female (left) and male (right)**
1392 ***N. lecontei*.** White arrows indicate regions of male-specific synaptic clusters. Scale bars = 500
1393 µm.
1394
1395 **Figure 3. Position of genes belonging to five environmentally responsive gene families along**
1396 **seven *N. lecontei* linkage groups.** Linkage groups (LG) are drawn to scale and ordered as in the
1397 linkage-group anchored assembly described in Linnen et al. 2018 (GenBank accession numbers
1398 are as follows: LG1 = CM009916.1; LG2 = CM009917.1; LG3 = CM009918.1; LG4 =
1399 CM009919.1; LG5 = CM009920.1; LG6 = CM009921.1; LG7 = CM009922.1). Gene family
1400 abbreviations: OR (olfactory receptor), GR (gustatory receptor), OBP (odorant binding protein),
1401 CYP (cytochrome P450), AMP (antimicrobial protein). Each gene family is represented by a
1402 different color. Horizontal lines indicate the approximate locations of genes within LG; diagonal
1403 lines that connect to horizontal lines are used to highlight groups of genes that met our clustering
1404 criteria. Genes that were found on scaffolds that have not been placed on linkage groups are
1405 indicated on the bottom left, with abbreviated scaffold names given in parentheses (e.g., S-210 =
1406 scaffold_210 = LGIB01000210.1 in the assemblies available on NCBI).
1407
1408 **Figure 4. Number of intact genes in hymenopteran genomes for each of five**
1409 **environmentally responsive gene families.** Phylogenetic relationships are as in Moreau et al.
1410 (2006); Hedtke et al. (2013); Roux et al. (2014); Brand et al. (2017); Branstetter et al. (2017);
1411 Peters et al. (2017). Branch lengths are arbitrary. Gene family abbreviations are as in Figure 3.
1412
1413 **Figure 5. Ecological correlates of gene family size in Hymenoptera.** Each point represents the
1414 number of intact genes for a hymenopteran species for which both manually curated gene
1415 annotations and ecological data are available. Asterisks indicate that gene number varies
1416 significantly among the ecological categories under consideration; for significant categories with
1417 >2 groups, letters indicate significance in post-hoc tests (groups that do not share a letter are
1418 significantly different). Gene number and ecological data for all taxa are provided in Table S8.
1419

1420 **Table 1. Summary of within-genome signatures of adaptive expansions and contractions**
1421 **for five environmentally responsive gene families.**

Gene family*	Gene family size					Genomic Clustering		Molecular evolution		
	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

1422

* 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.

1423
1424 **Table 2. Likelihood-ratio tests (LRTs) of positive selection on *Neodiprion*-specific clades**
1425 **(branch models) and on amino acid sites within these clades (site models).**
1426
1427
1428

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

* 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 1 ratio	2.077	1	0.149
		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.

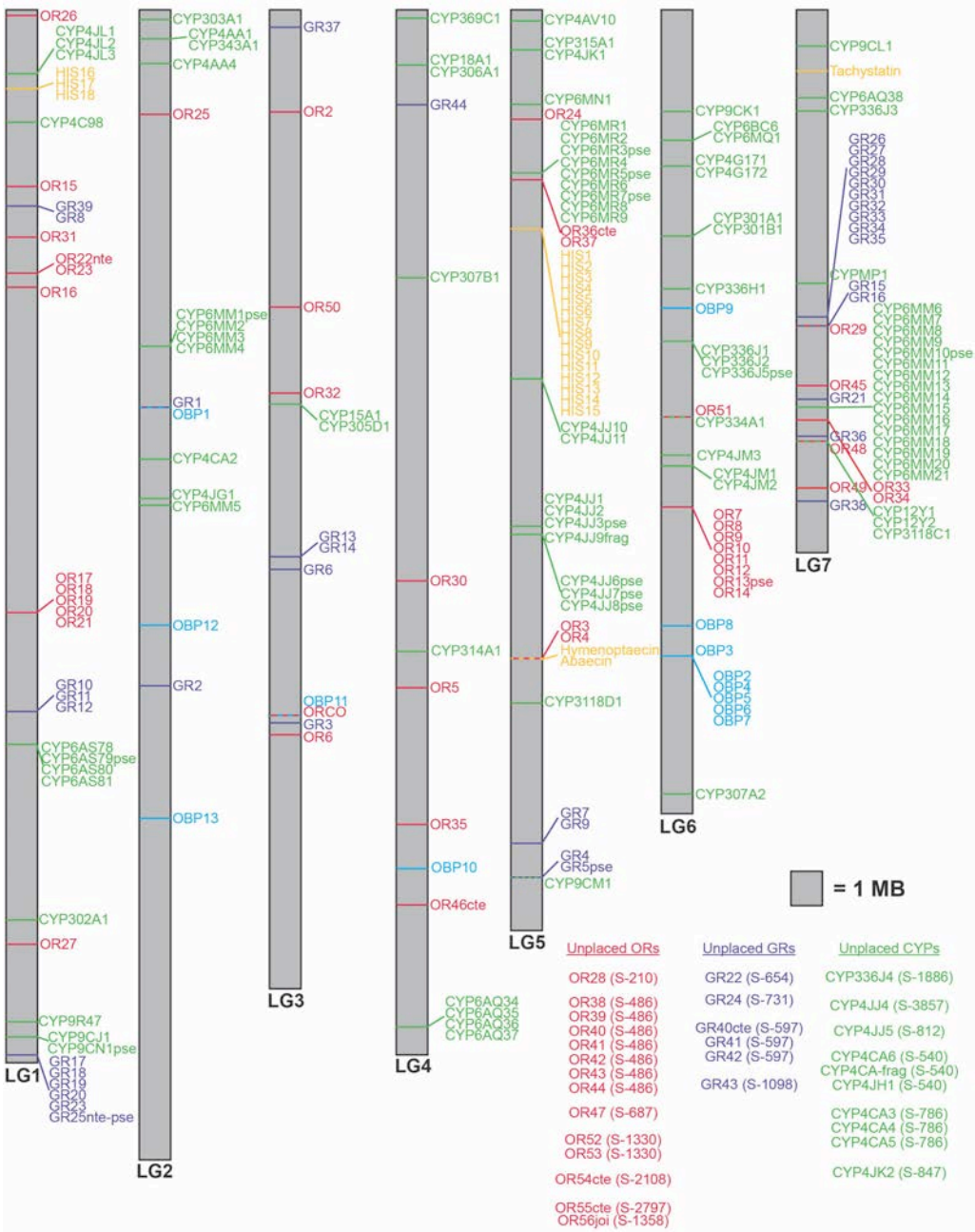

Figure 3. Position of genes belonging to five environmentally responsive gene families along seven *N. lecontei* linkage groups. Linkage groups (LG) are drawn to scale and ordered as in the linkage-group anchored assembly described in Linnen et al. 2018 (GenBank accession numbers are as follows: LG1 = 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	CYP	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	11

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.

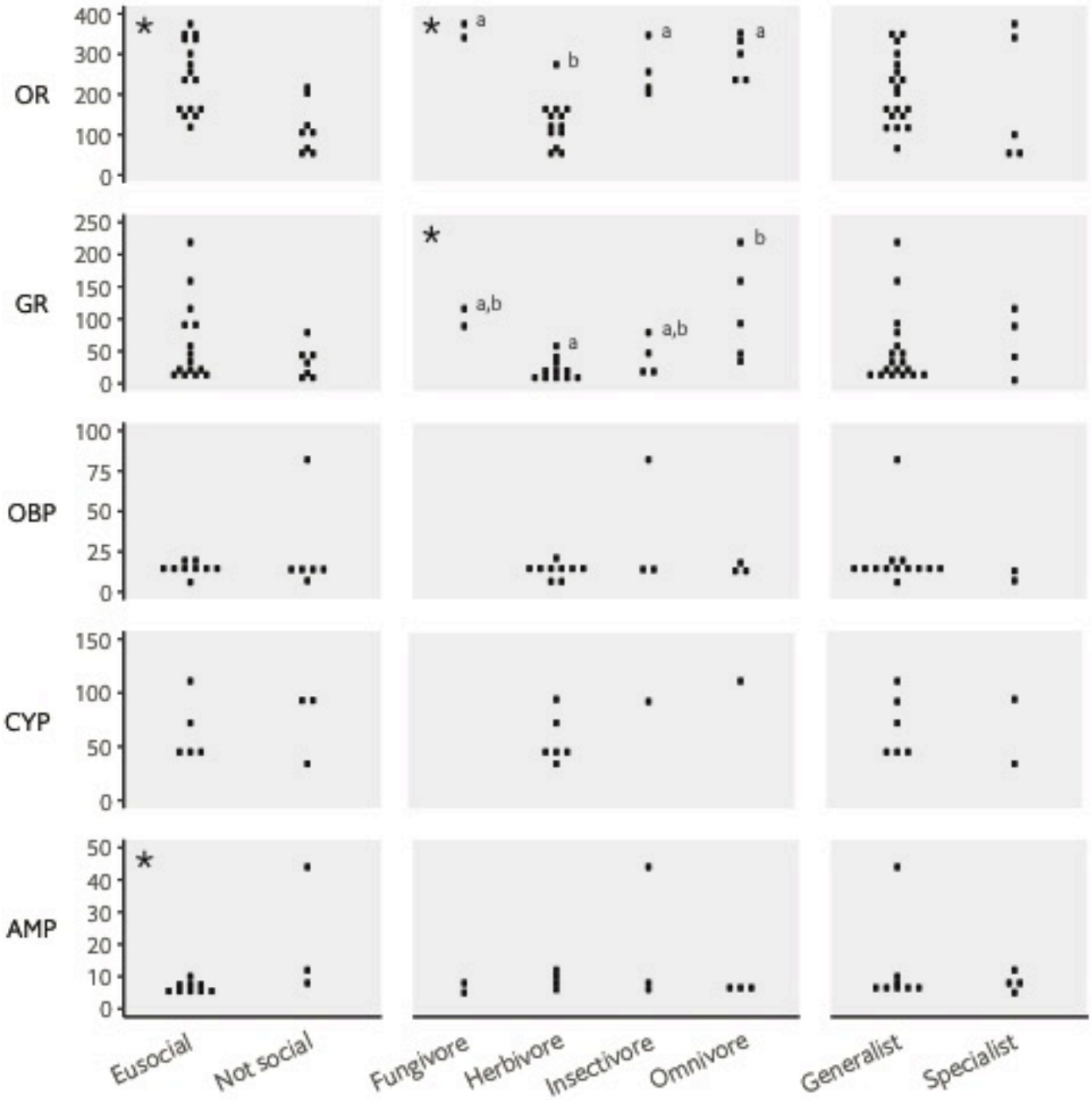

Figure 5. Ecological correlates of gene family size in Hymenoptera. Each point represents the number of intact genes for a hymenopteran species for which both manually curated gene annotations and ecological data are available. Asterisks indicate that gene number varies significantly among the ecological 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.