bioRxiv preprint doi: https://doi.org/10.1101/399808; this version posted August 26, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

1 # Genomic signatures accompanying the
2 # dietary shift to phytophagy in
3 # polyphagous beetles

4

5 Mathieu Seppey, Panagiotis Ioannidis, Brent C. Emerson, Camille Pitteloud, Marc Robinson-Rechavi,
6 Julien Roux, Hermes E. Escalona, Duane D. McKenna, Bernhard Misof, Seunggwan Shin, Xin Zhou,
7 Robert M. Waterhouse*, Nadir Alvarez*
8
9 MS: Department of Ecology & Evolution, University of Lausanne, 1015 Lausanne, Switzerland. Present address: Department of
10 Genetic Medicine and Development, University of Geneva and Swiss Institute of Bioinformatics, rue Michel-Servet 1, 1211
11 Geneva, Switzerland. ORCID: 0000-0003-3248-011X.
12
13 PI: Department of Genetic Medicine and Development, University of Geneva and Swiss Institute of Bioinformatics, rue Michel-
14 Servet 1, 1211 Geneva, Switzerland. Present address: Institute of Molecular Biology and Biotechnology, Foundation for Research
15 and Technology - Hellas, N. Plastira 100, Vassilika Vouton, 70013 Heraklion, Greece. ORCID: 0000-0003-0939-6745.
16
17 BCE: Island Ecology and Evolution Research Group, Instituto de Productos Naturales y Agrobiología (IPNA-CSIC), Avda.
18 Astrofísico Francisco Sánchez 3, 38206 San Cristóbal de la Laguna, Santa Cruz de Tenerife, Spain. School of Biological
19 Sciences, University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, UK.
20
21 CP: Department of Ecology & Evolution, University of Lausanne, 1015 Lausanne, Switzerland. Present address: Dep. of
22 Environmental Systems Science, ETHZ, Universitätstrasse 16, 8092 Zurich
23
24 MRR: Department of Ecology & Evolution, University of Lausanne, and Swiss Institute of Bioinformatics, 1015 Lausanne,
25 Switzerland. ORCID: 0000-0002-3437-3329.
26
27 JR: Department of Ecology & Evolution, University of Lausanne, and Swiss Institute of Bioinformatics, 1015 Lausanne,
28 Switzerland. Present address: Department of Biomedicine, University of Basel, Basel, Switzerland. ORCID: 0000-0002-4192-
29 5099.
30
31 HEE: Center for Molecular Biodiversity Research (ZMB), Zoological Research Museum Alexander Koenig,
32 Adenaueralle 160, D-53113 Bonn, Germany
33
34 DDM: Department of Biological Sciences. University of Memphis, 3700 Walker Ave., Memphis TN, U.S.A. ORCID: 0000-0002-
35 7823-8727.
36
37 BM: Center for Molecular Biodiversity Research (ZMB), Zoological Research Museum Alexander Koenig,
38 Adenaueralle 160, D-53113 Bonn, Germany
39
40 SS: Department of Biological Sciences. University of Memphis, 3700 Walker Ave., Memphis TN, U.S.A. ORCID: 0000-0002-
41 4258-8661
42
43 XZ: Department of Entomology, College of Plant Protection, China Agricultural University, Beijing, China 100093. ORCID: 0000-
44 0002-1407-7952.
45
46 RMW: Department of Ecology & Evolution, University of Lausanne, and Swiss Institute of Bioinformatics, 1015 Lausanne,
47 Switzerland. ORCID: 0000-0003-4199-9052
48
49 NA: Department of Ecology & Evolution, University of Lausanne, 1015 Lausanne, Switzerland, and Museum of Natural History
50 of the City of Geneva, 1 Rte de Malagnou, 1208 Genève, Switzerland. ORCID: 0000-0002-0729-166X.
51
52 * These authors contributed to the work equally
53
54 Corresponding authors: robert.waterhouse@unil.ch, nadir.alvarez@ville-ge.ch

bioRxiv preprint doi: https://doi.org/10.1101/399808; this version posted August 26, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

55 **Abstract**

56

57 The diversity and evolutionary success of beetles (Coleoptera) are proposed to have arisen from
58 millions of years of specialized trophic interactions with land plants. In particular, ingestion of toxic plant
59 allelochemicals may impose selective pressures that drive genomic diversification and speciation in
60 phytophagous beetles. However, evidence of changes in beetle gene repertoires driven by these
61 interactions remains largely anecdotal and without explicit hypothesis testing. To address this, we
62 explored the genomic consequences of beetle-plant trophic interactions by performing comparative
63 gene family analyses across 18 species representing the two most speciose beetle suborders. By
64 contrasting gene content of species from the phytophagous-rich suborder Polyphaga with those of the
65 mainly predatory Adephaga, we identified families of detoxification enzymes that underwent adaptive
66 expansions in Polyphaga. These genomic signatures that accompany the dietary shift to phytophagy in
67 polyphagous beetles suggest a key role for interactions with plant chemical defenses in driving beetle
68 diversification.

bioRxiv preprint doi: https://doi.org/10.1101/399808; this version posted August 26, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

69 **Introduction**

70

71 Species richness among eukaryotes varies substantially, with some clades having only a few
72 representatives and others comprising hundreds of thousands of extant species. In particular, the class
73 Insecta outnumbers all other classes with more than half of all described extant species (Farrell, 1998;
74 Grimaldi and Engel, 2005). Beetles (Coleoptera) encompass approximately 380,000 described species,
75 representing ca. 40% of described insect diversity (Slipinski et al., 2011). Several hypotheses have
76 been proposed to explain this richness, notably their complex interactions with flowering plants (Farrell,
77 1998; Leschen and Buckley, 2007; McKenna et al., 2009, 2015; Zhang et al., 2018) and a high lineage
78 survival rate (Hunt et al., 2007). Nevertheless, detailed supporting evidence from molecular genetic
79 studies remains sparse, making it difficult to assess the relative importance of these and other
80 potentially important contributing factors (Barraclough et al., 1998; Suchan and Alvarez, 2015).

81

82 The remarkable evolutionary success of beetles may have been driven by the interplay between their
83 trophic niche and their genomic content and architecture. This is based on the premise that
84 environmental and ecological conditions are likely to be predominant factors influencing the fate of
85 genetic variation in populations under natural selection (Barrick and Lenski, 2013), eventually driving
86 divergence into distinct species (Seehausen et al., 2014). Among all components of the biotic
87 environment, the trophic niche (principal source of nourishment) of an organism plays a crucial role in
88 shaping the evolution of phenotypic innovations and their underlying genomic changes, e.g. feeding
89 modes in cichlid fishes (Parsons et al., 2016), mouth development in *Pristionchus* nematodes (Ragsdale
90 et al., 2013), and bitter taste receptors in vertebrates (Li and Zhang, 2014). Among several hypotheses
91 explaining the tremendous diversity among beetles, a shift from an ancestral diet as saprophages
92 (detritus-feeding) or mycophages (fungi-feeding) (Betz et al., 2003) to phytophagy (feeding on living
93 plant material in a broad sense) is often evoked (Farrell, 1998; Leschen and Buckley, 2007; McKenna
94 et al., 2009). While the suborder Adephaga (~45,000 species) comprises mostly predatory species,
95 including ground beetles and diving beetles, the largest beetle suborder, Polyphaga (~315,000), is
96 predominantly comprised of phytophagous clades, among which the most species-rich families are
97 weevils (Curculionidae, ~51,000), longhorn beetles (Cerambycidae, ~30,000), and leaf beetles
98 (Chrysomelidae, ~32,000) (Slipinski et al., 2011). Phytophagy appeared approximately 425 million

bioRxiv preprint doi: https://doi.org/10.1101/399808; this version posted August 26, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

99 years ago, quickly after terrestrial life was established (Labandeira, 2002). It progressively diversified
100 to target most plant tissues (Labandeira, 2013), shortly before the radiation of flowering plants 120-100
101 million years ago (Grimaldi, 1999). In response, plants have evolved diverse strategies to protect
102 themselves, which in turn impose selective pressures on the animals that feed on them.
103
104 While many biological processes are likely to play a role in this evolutionary battle, a key weapon in the
105 arsenal of phytophagous insects' adaptations is their ability to neutralize or minimize the effects of plant
106 secondary compounds. Protein families known to be crucial for eliminating harmful plant toxins are
107 cytochrome P450 monooxygenases (P450s), carboxylesterases (CEs), UDP-glycosyltransferases
108 (UGTs), and glutathione S-transferases (GSTs) (Voelckel and Jander, 2014). While P450s and CEs
109 modify residues to make compounds more hydrophilic, UGTs and GSTs conjugate xenobiotic
110 compounds to hydrophilic molecules. Detoxification is completed by membrane transporters, such as
111 ATP-binding cassette (ABCs) transporters, which move xenobiotic compounds to where they can either
112 be excreted, or less frequently sequestered in order to be reused as a defense mechanism (Voelckel
113 and Jander, 2014). Additionally, to prevent phytophagous insects from digesting their tissues, plants
114 produce enzyme inhibitors that block catalytic sites or compete with the substrates of enzymes involved
115 in digestion. The major families affected are endopeptidases, such as cysteine, and serine proteases,
116 as well as more specific enzymes such as glycoside hydrolases (GHs), certain types of which are able
117 to break down polysaccharide molecules, including cellulose, hemicellulose and pectin in plant cell walls
118 (McKenna et al., 2016; Pauchet et al., 2010). Other adaptations to phytophagy include repertoires of
119 chemoreceptors that are crucial for finding appropriate food sources (Goldman-Huertas et al., 2015),
120 and the specialization of mouthparts in response to plant mechanical barriers, which are highly
121 diversified in insects (Labandeira, 1997).
122
123 As lineages diverge, their genomes accumulate changes, some of which are expected to be directly
124 linked to functional adaptations. Identifying such genomic features and linking them to phenotypic
125 differences, whilst robustly distinguishing between the effects of stochastic changes and natural
126 selection (Hurst, 2009), is critical to deciphering the genomic drivers of species radiations (Shaw and
127 Lesnick, 2009). Changes include point substitutions, which may affect existing functional elements, but
128 also larger-scale changes such as duplications, from individual genes to entire genomes, which by
129 adding new members to the repertoires of key gene families may constitute an ideal mechanism to

bioRxiv preprint doi: https://doi.org/10.1101/399808; this version posted August 26, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

130 facilitate the emergence of novel functions leading to successful phytophagy (Kondrashov, 2012).
131 Whereas newly generated gene copies are usually redundant or deleterious and pseudogenized,
132 rendering the gene copy non-functional (Innan and Kondrashov, 2010), they are sometimes maintained.
133 Particularly interesting cases of gene family expansions are the ones restricted to specific lineages,
134 resulting in lineage specific expansion (LSE). Evolutionary mechanisms causing LSE are numerous
135 and not all adaptive (see Innan and Kondrashov, 2010 for a comprehensive review). However,
136 duplicated gene copies may provide an immediate selective advantage and be maintained by selection.
137 This can be due to an increased dosage of the gene, or to changes following the duplication being
138 selected in one gene copy but not the other, which might allow evolution towards a different function in
139 so-called neo-functionalization processes. Enzymes are considered particularly relevant candidates for
140 such evolutionary processes as they could expand their range of substrates (Francino, 2005).
141
142 Here we apply a comparative genomics approach to examine the evolution of genes putatively involved
143 in plant-insect interactions by sampling from the two largest beetle suborders, which, generally-
144 speaking, present contrasting trophic niches. We contrast exemplars from the characteristically
145 predaceous Adephaga with exemplars from Polyphaga and we hypothesize that plant-insect
146 interactions during the dietary shift to phytophagy should be accompanied by genomic evolutionary
147 signatures visible at the subordinal scale. Using genomic and transcriptomic data from 18 beetle
148 species, we estimate ancestral gene family content, taking into account gene gains and losses across
149 the species phylogeny, to identify significant LSEs of gene families related to phytophagy and
150 signatures of adaptive expansions in these families. Ignoring sensory receptors, as their evolution might
151 be driven by agents other than those related strictly to trophic niche (Brito et al., 2016), and
152 morphological genes, as their inferred association with diet is less robust, we focus on genes coding
153 for enzymes, for which adaptive LSE specific to Polyphaga would suggest a role for detoxification and
154 digestive pathways in driving adaptation and speciation.

bioRxiv preprint doi: https://doi.org/10.1101/399808; this version posted August 26, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

155 ## Results

156 *A representative sampling of the two major coleopteran suborders*

157 Reliable estimation of gene gain and loss events requires a robust evolutionary framework, i.e. a
158 phylogeny that includes the species studied, as well as the characterization of gene families across
159 complete gene sets from these same species. To study adaptation to phytophagy, we sampled from
160 both Adephaga (mostly predaceous) and Polyphaga (with diverse trophic habits, including a very large
161 number of phytophagous species). A balanced sampling of each suborder was achieved comprising
162 twelve transcriptomes and six genomes, with Benchmarking Universal Single-Copy Ortholog (BUSCO)
163 completeness estimates (Simão et al., 2015; Waterhouse et al., 2018) ranging from 71.9% to 97%
164 (Figure 1, Table 1). The species phylogeny was estimated using 405 BUSCO genes found to be
165 complete in all species and the strepsipteran outgroup, *Stylops melittae* (Figure 1). Protein-coding
166 sequence predictions ranged from 9,844 to 24,671 genes per beetle species. These sequences
167 matched 14,908 Arthropoda orthologous groups (OGs) containing at least one species of Coleoptera
168 in the OrthoDB v8 catalogue (Kriventseva et al., 2015). This represented a minimum of 6,742 and a
169 maximum of 11,149 OGs for *Carabus frigidus* and *Leptinotarsa decemlineata,* respectively. OGs
170 containing genes from only one of the two sampled suborders were excluded, resulting in a total of
171 9,720 OGs for the analysis that have evolutionary histories traceable to the last common ancestor of
172 beetles. Functional annotations of the sequences within these OGs were used to identify and assign
173 several of them to enzyme families relevant to the tested hypothesis. These candidate OGs comprised
174 four UGTs, 22 P450s, 19 CEs, six GSTs, four SERs, seven CYSs, 28 ABCs, and one GH, for a total of
175 91 candidate OGs from eight families of genes (i.e. functional categories) (Table 2).

176 **Table 1.** Beetle genomes and transcriptomes included in the study. Taxonomic classifications are listed with data sources, as well as completeness (Benchmarking
177 Universal Single-Copy Ortholog, BUSCO, score, C=Complete, S=Complete Single-Copy, D=Complete Duplicated, F=Fragmented, M=Missing), number of predicted
178 proteins, and number of orthologous groups (OGs) with genes from each species. The outgroup species used in the phylogeny, *Stylops melittae*, belongs to the order
179 Strepsiptera, which is the sister group of Coleoptera (Niehuis et al., 2012).

Species	Short form	Suborder	Family	Type of assembly	Accession	Bioproject	Source	BUSCO score (1658 genes)	Predicted proteins	OrthoDB groups (OG)
Cicindela hybrida	CHYBR	Adephaga	Carabidae	Transcriptome	GDMH01000000	PRJNA286505	1KITE, this study	C:90.4[S:84.9%,D:5.5%],F:3.4%,M:6.2%	13,916	8,111
Calosoma frigidum	CFRIG	Adephaga	Carabidae	Transcriptome	GDLF01000000	PRJNA286499	1KITE, this study	C:78.3%[S:73.5%,D:4.8%],F:9.8%,M:11.9%	9,844	6,742
Elaphrus aureus	EAURE	Adephaga	Carabidae	Transcriptome	GDPI01000000	PRJNA286520	1KITE, this study	C:92.4%[S:85.9%,D:6.5%],F:2.2%,M:5.4%	12,808	8,020
Noterus clavicornis	NCLAV	Adephaga	Noteridae	Transcriptome	GDNA01000000	PRJNA286561	1KITE, this study	C:90.9%[S:84.1%,D:6.8%],F:4.4%,M:4.7%	12,981	7,918
Haliplus fluviatilis	HFLUV	Adephaga	Haliplidae	Transcriptome	GDMW01000000	PRJNA286525	1KITE, this study	C:91.8%[S:76.3%,D:15.5%],F:4.3%,M:3.9%	19,408	8,528
Cybister lateralimarginalis	CLATE	Adephaga	Dytiscidae	Transcriptome	GDLH01000000	PRJNA286512	1KITE, this study	C:88.2%[S:83.1%,D:5.1%],F:3.7%,M:8.1%	13,916	7,256
Sinaspidytes wrasei	SWRAS	Adephaga	Aspidytidae	Transcriptome	GDNH01000000	PRJNA286492	1KITE, this study	C:87.0%[S:76.7%,D:10.3%],F:4.0%,M:9.0%	13,392	7,721
***Dineutes* sp.**	DINEU	Adephaga	Gyrinidae	Transcriptome	GDNB01000000	PRJNA286516	1KITE, this study	C:71.9%[S:51.6%,D:20.3%],F:13.6%,M:14.5%	14,644	7,089
Gyrinus marinus	GMARI	Adephaga	Gyrinidae	Transcriptome	GAUY01000000	PRJNA219564	1KITE, Misof et al., 2014	C:81.5%[S:79.0%,D:2.5%],F:8.1%,M:10.4%	13,867	7,663
Aleochara curtula	ACURT	Polyphaga	Staphylinidae	Transcriptome	GATW01000000	PRJNA219522	1KITE, Misof et al., 2014	C:91.4%[S:87.5%,D:3.9%],F:4.0%,M:4.6%	20,280	8,513
Anoplophora glabripennis	AGLAB	Polyphaga	Cerambycidae	Genome	GCF_000390285	PRJNA167479	I5k, McKenna et al., 2016	C:96.9%[S:95.8%,D:1.1%],F:2.7%,M:0.4%	22,035	10,959
Agrilus planipennis	APLAN	Polyphaga	Buprestidae	Genome	GCF_000699045	PRJNA230921	I5k, unpublished	C:92.5%[S:91.2%,D:1.3%],F:4.5%,M:3.0%	15,497	9,089
Dendroctonus ponderosae	DPOND	Polyphaga	Curculionidae	Genome	GCF_000355655	PRJNA162621	Keeling et al., 2013	C:91.2%[S:86.0%,D:5.2%],F:4.1%,M:4.7%	13,457	8,518
Leptinotarsa decemlineata	LDECE	Polyphaga	Chrysomelidae	Genome	GCF_000500325	PRJNA171749	I5k, Schoville et al., 2018	C:88.9%[S:87.5%,D:1.4%],F:9.9%,M:1.2%	24,671	11,149
Laparocerus tessellatus	LTESS	Polyphaga	Curculionidae	Transcriptome	10.5281/zenodo.1336288	N/A	This study	C:93.8%[S:91.9%,D:1.9%],F:1.4%,M:4.8%	18,448	8,616
Meloe violaceus	MVIOL	Polyphaga	Meloidae	Transcriptome	GATA01000000	PRNJA219578	1KITE, Misof et al., 2014	C:90.3%[S:85.6%,D:4.7%],F:5.9%,M:3.8%	14,295	8,480
Onthophagus taurus	OTAUR	Polyphaga	Scarabaeidae	Genome	GCF_000648695	PRJNA167478	I5k, unpublished	C:96.2%[S:93.9%,D:2.3%],F:2.5%,M:1.3%	17,483	9,315
Tribolium castaneum	TCAST	Polyphaga	Tenebrionidae	Genome	GCF_000002335	PRJNA12540	Tribolium Genome Sequencing Consortium et al., 2008	C:97.0%[S:96.5%,D:0.5%],F:1.6%,M:1.4%	16,645	9,429
Stylops melittae	SMELI	Outgroup	Stylopidae	Transcriptome	GAZM02000000	PRNJA219603	1KITE, Misof et al., 2014	C:76.5%[S:55.0%,D:21.5%],F:7.1%,M:16.4%	13,026	6,104

180

bioRxiv preprint doi: https://doi.org/10.1101/399808; this version posted August 26, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

181 **Table 2.** Candidate gene categories with the keywords and identifiers used to select them from the full
182 sets of sequences annotated with InterProScan. To be included as candidate orthologous groups (OGs)
183 in the category, OGs were required to have at least one sequence matching both a UniRef and an
184 InterProScan entry, and an additional gene ontology term in the case of serine proteases.

185

Category	InterProScan (Pfam or InterPro identifiers) or gene ontology	UnifRef KeyWord	Number of OGs
UDP-glycosyltransferases (UGTs)	PF00201	name:"cluster UDP glucuronosyltransferase" OR name:"cluster UDP glycosyltransferase"	4
Cytochrome P450 oxidases (P450s)	PF00067	name:"cluster Cytochrome P450"	22
Carboxylesterases (CEs)	PF02230, PF00135	name:"cluster carboxylesterase" OR name:"carboxylic ester hydrolase"	19
Glutathione S-transferases (GSTs)	PF00043, PF02798	name:"cluster Glutathione S-transferase"	6
Serine proteases (SERs)	PF00450,PF12146, PF05577,GO:0008236	name:"cluster Serine protease" OR name:"cluster Serine peptidase"	4
Cysteine proteases (CYSs)	PF00112	name:"cluster cysteine protease" OR name:"cluster cystein protease" OR name:"cluster Papain"	7
ABC transporters (ABCs)	PPF00005, PF00664	name:"cluster ABC"	28
Glycoside hydrolases (GHs)	IPR000334, IPR000743, IPR001360, IPR001547	name:"cluster Glycoside hydrolase"	1
Total			91

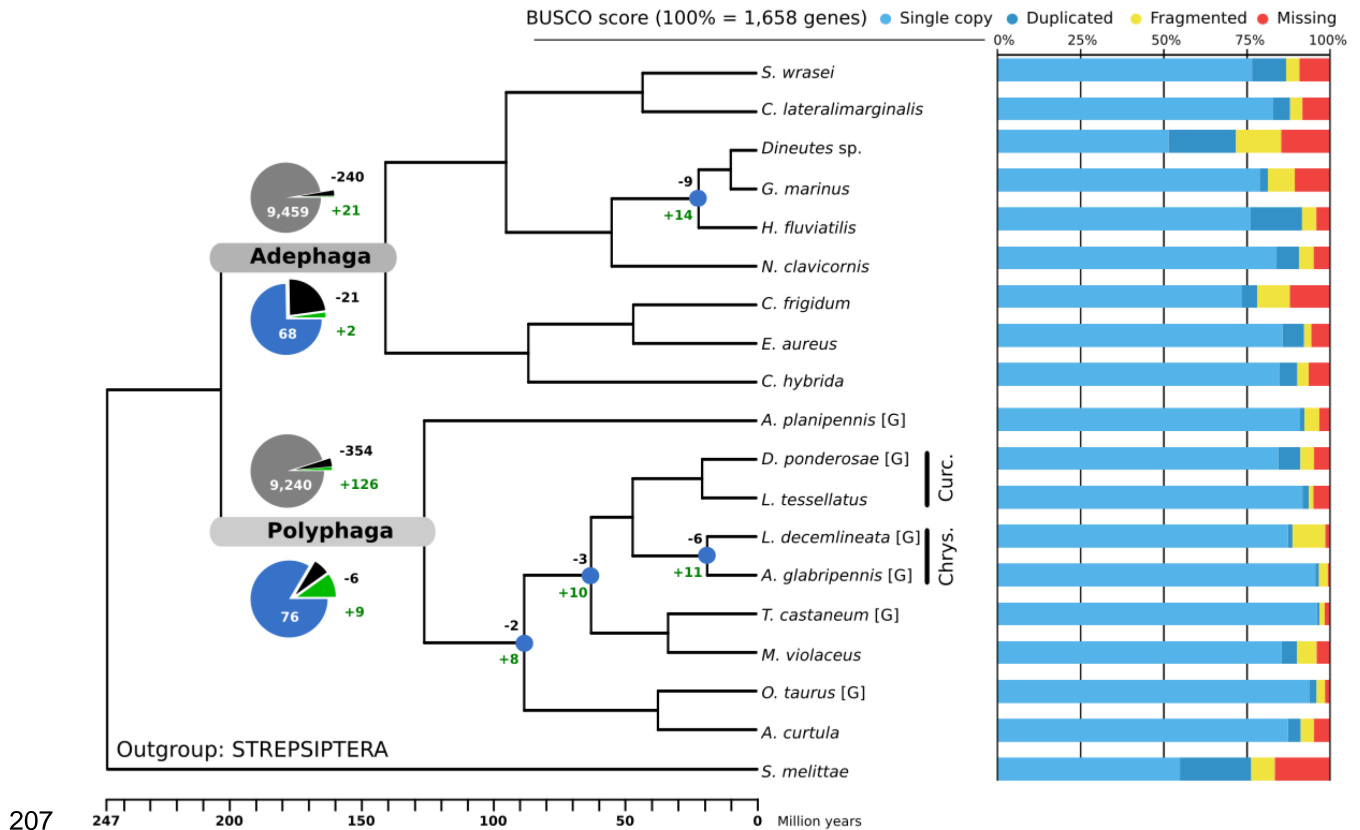
186

bioRxiv preprint doi: https://doi.org/10.1101/399808; this version posted August 26, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

187 *Polyphaga exhibit more frequent gains across a larger set of OGs*

188 Analysis of per-species gene counts of the complete set of 9,720 OGs was performed with the
189 Computational Analysis of gene Family Evolution (CAFE v3) (Han et al., 2013) tool. The mode
190 considering distinct gene gain (λ=0.0019 gain/gene/million years) and gene loss (μ=0.0018
191 loss/gene/million years) was preferred over a single value for λ and μ, having a significantly greater
192 maximum likelihood score (see Methods). The λ (gain) and μ (loss) values predicted when CAFE was
193 run on each suborder separately were λ=0.0020 and μ=0.0027 for Adephaga, versus λ=0.0023 and
194 μ=0.0021 for Polyphaga, showing a tendency for Adephaga to lose genes and for Polyphaga to gain
195 genes. Among the 9,720 OGs were 21 with reported expansions originating at the Adephaga root and
196 126 at the Polyphaga root (see Figure 1 to locate the nodes). Conversely, 240 OGs showed gene losses
197 for Adephaga and 354 for Polyphaga. Two expansions and 21 losses affected the candidate OGs for
198 Adephaga, and nine expansions and six losses for Polyphaga. Other polyphagan nodes leading to
199 polyphagous-rich clades (i.e., Chrysomeloidea and Curculionidae) also exhibited more candidate OGs
200 expanding than contracting (Figure 1). All counts of gene gains and losses per node are presented in
201 Supplementary Figures 1 and 2. Additionally, CAFE assigned individual OG p-values of < 0.01 to a
202 subset of 910 (9.3%) OGs, which, according to De Bie et al., 2006, indicates gene families likely to have
203 experienced accelerated rates of gain and loss. These are interesting to investigate further as they may
204 represent large OGs of potentially unequal size between the suborders. Among these were 26 of the
205 91 candidate OGs (28.6%), a significantly larger proportion (2-sample test for equality of proportions,
206 chi-square test, p-value < 0.0001) compared with just 9.3% of non-candidate OGs.

bioRxiv preprint doi: https://doi.org/10.1101/399808; this version posted August 26, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



207

208

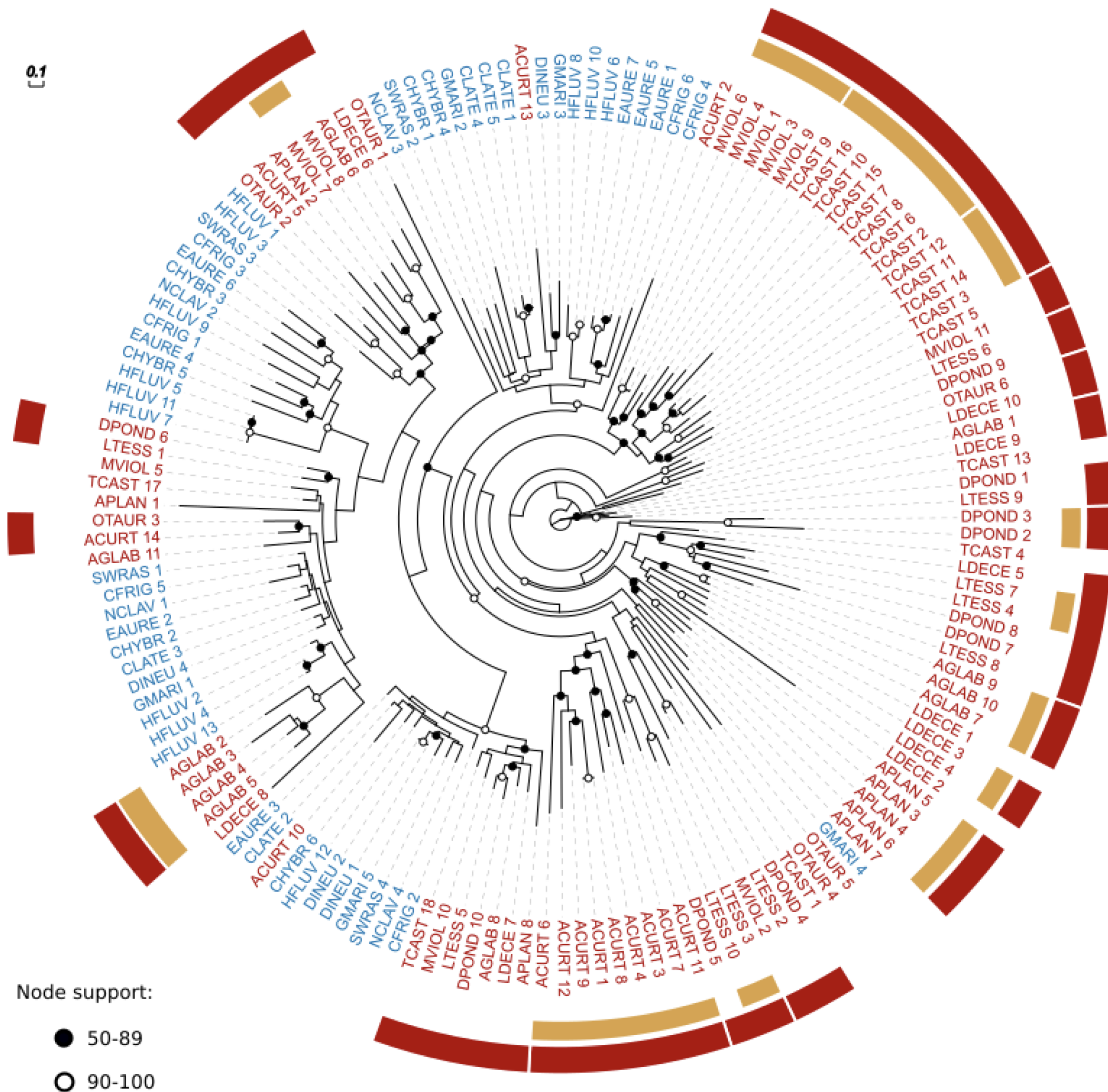
209 **Figure 1.** The ultrametric species phylogeny with gene family expansions and contractions quantified
210 for nodes of interest and bar charts showing completeness of the genomic and transcriptomic datasets
211 studied. The species tree was built from 405 single-copy orthologs and constrained to have
212 Geadephaga (*C. frigidum*, *E. aureus*, *C. hybridia*) and Hydradephaga (the six other Adephaga) as
213 monophyletic sister clades (e.g., following McKenna et al., 2015). Branch lengths are scaled in millions
214 of years. Maximum likelihood bootstrap support was 99 or 100% for all branches. [G] symbol indicates
215 data from species with sequenced genomes with the remaining species being from transcriptomes. The
216 numbers of orthologous groups (OGs) with expansions (+) and contractions (-) are displayed at the root
217 node of each suborder. Pie charts show proportions of OGs with gene losses (black) and gene gains
218 (green) with respect to OGs with no significant losses or gains for all considered OGs (grey) and only
219 the candidate OGs (blue). While gains constitute only a small subset of all OGs in both suborders, the
220 proportion of gains is much larger among candidate OGs in Polyphaga. The nodes indicated by blue
221 circles in the Polyphaga sub-tree lead to species-rich clades containing species that are largely
222 phytophagous (e.g., Chrysomelidae and Curculionidae, respectively Chrys. and Curc.) and experienced
223 larger proportions of gains among the candidate OGs. The Benchmarking Universal Single-Copy
224 Ortholog (BUSCO) scores indicate the relative levels of completeness and putative gene duplications
225 for the genome-based and transcriptome-based datasets in terms of 1,658 BUSCOs from the
226 insecta_odb9 assessment dataset.

bioRxiv preprint doi: https://doi.org/10.1101/399808; this version posted August 26, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

227 *Signatures of adaptive expansion are more prevalent in Polyphaga*

228 All 910 OGs with significant variations in their gene content were tested for signatures of adaptive
229 expansion in each suborder, by comparing Brownian motion (BM, neutral) to Ornstein-Uhlenbeck (OU,
230 selective pressure) evolutionary models (Beaulieu et al., 2012). As mentioned previously, these
231 included 26 OGs that belong to one of the functional categories listed in Table 2 ("candidate" OGs). The
232 models consider per-species gene count as a trait that can evolve towards a value, which may or may
233 not differ between the two suborders and may or may not be guided by selective pressure; we call this
234 the "optimum" value in models integrating selection. In total, 21 OGs displayed a higher optimum for
235 Adephaga (0.2% of the initial 9,720 OGs) and 88 for Polyphaga (0.9%). Eight of these 88 polyphagan
236 OGs (Table 3 and gene trees in Figure 2 and Supplementary Figures 3-9) are candidate OGs belonging
237 to one of the candidate gene families of Table 2, while none of the 21 adephagan OGs belong to any
238 of the candidate gene families. The proportion of OGs with expansions and higher optima in the
239 background (all "candidate" and remaining "control" OGs) was significantly larger for Polyphaga
240 compared to Adephaga (2-sample test for equality of proportions, chi-squared, 88/9720 vs. 21/9720, p-
241 value < 1e-09), indicating that Polyphaga have experienced globally more LSE under selection on
242 protein-coding genes. Furthermore, a test for enrichment (see Methods) of OGs with LSE under
243 selection from the candidate families (Table 2) compared to the background was significant for
244 Polyphaga (8/91 vs. 88/9720, p-value < 1e-09). The same test applied individually on each candidate
245 gene family within the candidate dataset demonstrated that categories enriched for LSE under selection
246 in Polyphaga were GSTs (3/6 positive tests, fdr-corrected p-value < 1e-09) and CEs (3/19, fdr-corrected
247 p-value < 0.005), as shown in detail in Table 4.

bioRxiv preprint doi: https://doi.org/10.1101/399808; this version posted August 26, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



248

249

250 **Figure 2.** Molecular phylogeny from the largest glutathione S-transferase (GST) orthologous group
251 among those exhibiting lineage-specific expansions driven by selection. Red labels indicate genes
252 belonging to species of Polyphaga, accounting for 98 out of 152 genes (their Ornstein-Uhlenbeck per-
253 species optimum is 11.69 vs. 6.85 for Adephaga (blue labels), see Table 3). The presence of several
254 clades of polyphagan and adephagan genes delineates duplication events following the divergence of
255 the two suborders. Encircling the gene labels are red bars that highlight polyphagan clades with
256 bootstrap support of >50% and yellow bars that highlight intra-specific duplications with bootstrap
257 support of >50%. Corresponding full names of species are given in Table 1. Branch lengths represent
258 substitutions per site and bootstrap support below 50% is not displayed.

bioRxiv preprint doi: https://doi.org/10.1101/399808; this version posted August 26, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

259

260 **Table 3.** Candidate orthologous groups (OGs) with CAFE overall p-values p<0.01 for which a model
261 favoring selection for larger sizes in Polyphaga showed a greater likelihood. OG identifiers for functional
262 category cytochrome P450s (P450), carboxylesterases (CE), glutathione S-transferases (GST), and
263 cysteine proteases (CYS) are from OrthoDB v8 (ODB8 ID). Small-sample-size corrected Akaike
264 Information Criterion (AICc) values are reported for all tested models. BM1 (Brownian motion with a
265 single rate for the whole tree), BMS (Brownian motion with different rates for each regime), OU1
266 (selection towards the same optimum for both regimes), all representing the null hypothesis. OUM
267 (selection towards two optima, same variance) and OUMV (selection towards two optima, two
268 variances), representing the alternative hypotheses. The mean values in each suborder (Adephaga
269 versus Polyphaga) are presented in the last two columns. Values in bold font indicate the preferred
270 (maximum likelihood) model. A delta AICc > 2 is required for H1 to be retained.

271

Category	ODB8 ID	BM1 AICc H0.1	BMS AICc H0.2	OU1 AICc H0.3	OUM AICc H1.1	OUMV AICc H1.2	Mean Adephaga	Mean Polyphaga
P450	EOG805VG7	148.37	153.21	143.35	148.10	**137.95**	34.13	34.47
CE	EOG87DCWX	143.23	143.77	143.63	**138.91**	141.15	6.55	18.78
CE	EOG8KD911	87.08	91.23	82.90	89.10	**79.74**	0.89	2.86
CE	EOG876NDC	80.64	85.67	80.08	87.42	**77.23**	1.72	3.48
GST	EOG87WR3Z	86.24	87.76	76.05	74.40	**72.12**	1.71	3.16
GST	EOG81RS7Z	108.77	114.44	109.19	113.76	**103.79**	6.85	11.69
GST	EOG85F05D	117.62	115.88	111.53	107.62	**106.32**	5.69	9.16
CYS	EOG8JDKNM	91.85	91.62	89.74	**85.66**	88.25	1.80	3.78

272
273

bioRxiv preprint doi: https://doi.org/10.1101/399808; this version posted August 26, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 4. Gene family category and candidate orthologous group (OG) enrichments among positive results. The top panel presents the statistical significance of each test for enrichment of candidate gene families among the positive results when compared to the background, for Polyphaga. The lower panel indicates the number of positive results in both suborders, for candidate OGs and background. Significant values at the 0.05 threshold are shown in bold.

Category	Positive/Total OGs	Category enrichment FDR
P450	1/22	0.36268
CE	3/19	**0.00252**
GST	3/6	**0.00016**
CYS	1/7	0.16627
UGT	0/4	1.00000
SER	0/4	1.00000
ABC	0/28	1.00000
GH	0/1	1.00000
Category	**Positive/Total OGs**	**Candidate vs. background enrichment p-value**
Background (Polyphaga)	88/9,720	**0**
Candidates (Polyphaga)	8/91	
Background (Adephaga)	21/9,720	1
Candidates (Adephaga)	0/91	

Discussion

Comparative genomic analyses often highlight expanded gene families and link these expansions to biological functions peculiar to, or of special interest in, their focal organism(s). However, these analyses usually do not explicitly test for any hypothesized evolutionary model that might support such links. Here we test a specific hypothesis of adaptation to a phytophagous diet, by comparing candidate gene family repertoires from nine adephagan (a mostly predaceous suborder) and nine polyphagan (a highly

bioRxiv preprint doi: https://doi.org/10.1101/399808; this version posted August 26, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

292 phytophagous suborder) beetle species. These candidate families are putatively involved in
293 detoxification of plant allelochemicals and digestion of plant tissues. Specifically, we identify evidence
294 for potentially adaptive gene family expansions in the species rich Polyphaga. This result is robust to
295 potentially confounding factors that could arise from combining genomic and transcriptomic datasets,
296 conservative definitions of candidate gene families, or the greater species richness of the Polyphaga
297 (see discussion points below and Supplementary Information). Through explicitly testing for adaptive
298 LSEs, these results offer conclusive supporting evidence for the key evolutionary role of the
299 phytophagous trophic niche in driving gene family expansions in Coleoptera (specifically Polyphaga), a
300 feature that likely facilitated adaptation of polyphagan beetles to specialized plant feeding.
301
302 *Dataset heterogeneity*
303 For the comparison of gene repertoires between the two groups to be unbiased, the gene content of all
304 analyzed species should be of similar accuracy and completeness. The number of predicted proteins
305 for the genomic resources for each beetle species (Table 1, mean 15,977 and standard deviation 3,748)
306 was within the range expected of insects (see Waterhouse, 2015). The average total gene count for
307 Adephaga species (all transcriptomes) was about 4,200 fewer than for Polyphaga, which include two
308 genomes with more than 22,000 genes. This difference in average gene counts is reduced to just 1,384
309 when considering only genes assigned to the 9,720 OGs selected for the analysis. Our conservative
310 orthology filtering therefore ensured that the comparisons focused on gene families with reliably
311 traceable evolutionary histories that span both groups of beetles. Secondly, assessments of
312 completeness showed that the majority of the datasets contained more than 90% of complete BUSCOs
313 (Figure 1, Table 1). While the dynamically evolving families that are the focus of this study are clearly
314 not universal single-copy orthologs, the high levels of BUSCO completeness support the assumption
315 that the datasets represent good coverage of the species' gene content. Re-analyses of our data that
316 exclude the two adephagan beetle species with fewer than 80% complete BUSCOs reduced the power
317 of the model tests but nevertheless still identified the three GST OGs that favor a model with a higher
318 optima for Polyphaga (see Supplementary Results). Three of the adephagan transcriptomes showed
319 more than 10% of duplicated BUSCOs, which could have arisen from suboptimal filtering of the
320 transcriptomes, i.e. failure to remove alternative transcripts of the same gene. While such potentially
321 inflated gene counts for these adephagans might prevent the identification of some true expansions in

bioRxiv preprint doi: https://doi.org/10.1101/399808; this version posted August 26, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

322 Polyphaga, they do not invalidate those that were identified. Finally, half of the OGs representing
323 positive results showed a higher mean value for polyphagan transcriptomes than genomes, including
324 the three GST OGs (Supplementary Table 1), and explicitly testing for effects due to using both genome
325 and transcriptome data for the species of Polyphaga, by performing a modified OUwie analysis with
326 data type as the regime under selection, identified only one CE (EOG8KD911) for which the favored
327 model linked gene family expansion to species with genomes (see Supplementary Results).

328

329 *Candidate OG identification*

330 The annotation strategy was designed to link OGs to candidate gene families based on manually
331 selected keywords used to filter sequence search results, as well as Pfam and InterPro identifiers (Table
332 2), with the aim of excluding false positives (see Methods). This conservative strategy may not have
333 fully captured all possible candidate OGs, which would therefore have remained in the background set
334 of OGs that were used as controls. For example, we identified nine GST OGs (six were retained as
335 candidates after filtering) while ten subclasses have been identified in arthropods (Roncalli et al., 2015).
336 While the strict (conservative) strategy we employed to identify candidate OGs may have resulted in an
337 underestimate of the extent of the observed effects, this does not invalidate those that were identified.
338 In addition, filtering the OGs to retain only those with genes from both Adephaga and Polyphaga
339 excluded from the analyses any genes that were specific to either suborder. These might include genes
340 with key roles in phytophagy, e.g. enzymes acquired by horizontal gene transfer identified from the *A.*
341 *planipennis*, *A. glabripennis*, and *D. ponderosae* genomes (McKenna et al., 2016). While
342 acknowledging their importance, here we explicitly tested for adaptive LSE in one lineage versus the
343 other so gene evolutionary histories were required to span the two suborders and thus be traceable to
344 their last common ancestor.

345

346 *The more speciose Polyphaga exhibit more dynamic gene repertoire evolution*

347 The μ and λ values reported by CAFE on all 9,720 OGs are consistent with assessments of other insect
348 clades (Hahn et al., 2007; Neafsey et al., 2015). Although the overall gain rate is slightly higher than
349 the loss rate, the number of OGs losing genes reported by CAFE at each individual node is generally
350 larger than the number of OGs with gains. This is reconciled by considering that across Coleoptera
351 many OGs lost a few genes while few families gained many genes. As most OGs display a low number

bioRxiv preprint doi: https://doi.org/10.1101/399808; this version posted August 26, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

352 of genes per species, i.e. they are evolving under 'single-copy control' (Waterhouse et al., 2011), losing
353 more than one ortholog per species is understandably rare, while there is no theoretical limit for an OG
354 to gain new members. Comparing the two clades, Polyphaga has a higher rate of gene gain and six
355 times more OGs with gains, and while the Adephaga rate of gene loss is higher, Polyphaga have 1.5
356 times more OGs that have experienced gene losses. Hence, the gene repertoires of Polyphaga exhibit
357 a more dynamic evolutionary history with more gains (rate) in more OGs (counts) and fewer losses
358 (rate) spread out over more OGs (counts). It is possible that this greater dynamism may be generally
359 linked to the greater species richness of Polyphaga, with no specific role for phytophagy underpinning
360 this trend. However, among the candidate OGs for detoxification and digestion there are also more
361 gains in Polyphaga, and, in contrast to the background, fewer losses. Thus, both gain and maintenance
362 are higher for candidates in Polyphaga, which is consistent with a key role for phytophagy in driving
363 dynamic gene repertoire evolution, and particularly LSEs.
364
365 *Evidence for adaptive expansions of gene families involved in detoxification in polyphagan beetles*

366 In addition to observing more expansions among candidate OGs in the suborder Polyphaga, the positive
367 results from the OUwie analysis support the hypothesis that selective pressures drive detoxification
368 enzymes towards larger gene family sizes. This is especially pronounced for GSTs, for which half of
369 the OGs tested positive and for which a significant enrichment compared to the background was found.
370 The importance of GSTs in dietary shifts to phytophagy has been noted in mustard-feeding flies, where
371 duplicated GSTs involved in the mercapturic acid pathway showed signatures of positive selection
372 (Gloss et al., 2014). Our results therefore suggest that comparable phenomena have been acting at the
373 level of polyphagan beetles. The CEs also show a statistically significant enrichment for positive results
374 compared to the background, further supporting the diet detoxification hypothesis. The other positive
375 results include a P450 OG and a CYS OG, neither of which led to a category enrichment compared to
376 the background. The P450 OG is by far the largest among the positive results (Supplementary Figure
377 3), highlighting the importance of P450s in beetle (and generally insect) physiology with diverse roles
378 beyond detoxification, e.g. hormone biosynthesis (Kong et al., 2014). However, while considered as
379 significantly expanded and under selection by our model, the actual mean values in the suborders are
380 not dramatically different. Importantly, the enrichment of positive results among candidates still holds if
381 this OG is excluded (see Supplementary Results). The involvement of P450s in many other processes

bioRxiv preprint doi: https://doi.org/10.1101/399808; this version posted August 26, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

382 may explain why a broader difference between the suborders was not identified. Apart from one positive
383 result among the cysteine proteases (no significant category enrichment), our study did not highlight
384 additional expansions in other digestive enzymes or in transporters within a suborder. The lack of
385 evidence for expansion in Polyphaga with respect to ABC transporters, which is the candidate functional
386 category encompassing the highest number of OGs, may indicate that the ancestral diversity of
387 transporters was sufficient for maintaining the excretion of toxins, despite variations in the substrates
388 imposing a selective pressure on early stages of the detoxification pathway. Alternatively, if such
389 pressure were acting on later stages of the pathway, i.e. transporters, its strength could have been too
390 low for the detection power of our methods and data, unlike for GSTs or CEs.
391
392 *Conclusions*
393 By comparing the degree of expansion among gene families involved in detoxification of plant
394 secondary compounds in two suborders of beetles characterized by generally contrasting trophic niches
395 (i.e., Polyphaga contain a high proportion of phytophagous species while Adephaga encompass mostly
396 predacious species), we provide molecular genetic evidence supporting the popular hypothesis that
397 Coleoptera species richness may be in part explained by their interaction with land plants. Candidate
398 OGs of GSTs, CEs, P450s, and CYSs tested positive for adaptive LSEs in the phytophagous
399 polyphagan beetle lineage, and categories of GSTs and CEs in particular, were enriched for OGs with
400 such adaptive LSEs. Moreover, across all OGs tested, Polyphaga exhibited significantly more adaptive
401 LSEs than Adephaga. This indicates that genes other than the candidate detoxification and digestion
402 enzymes, which could include genes with functions less obviously related or unrelated to phytophagy,
403 are also likely to have played a role in the adaptive success of Polyphaga. While this suggests that
404 additional functional categories remain to be explored, contrasting gene family evolution across the two
405 major suborders of beetles demonstrates a role for interactions with plant secondary compounds, and
406 supports a role for phytophagy in general, as important drivers of the remarkable radiation of
407 polyphagan beetles.
408
409
410

bioRxiv preprint doi: https://doi.org/10.1101/399808; this version posted August 26, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

411 **Methods**

412 A chart summarizing the main steps of the analysis is available as Supplementary Figure 10.

413

414 *Data sources*

415 This study included six genomes and 13 transcriptomes representing a balanced sampling of
416 polyphagan and adephagan beetles, along with one representative of the sister group to Coleoptera,
417 Strepsiptera, to root the species phylogeny. Annotated gene sets from four genomes were sourced from
418 the i5k pilot project datasets (Robinson et al., 2011) (*Anoplophora glabripennis* v0.5.3 (McKenna et al.,
419 2016), *Leptinotarsa decemlineata* v0.5.3 *(Schoville et al., 2018), Onthophagus taurus* v0.5.3*, Agrilus*
420 *planipennis* v0.5.3) and two were independently published *Dendroctonus ponderosae* Ensembl
421 Metazoa v1.0 (Keeling et al., 2013) and *Tribolium castaneum* Ensembl Metazoa v3.22 (Tribolium
422 Genome Sequencing Consortium et al., 2008). One Polyphaga transcriptome, *Laparocerus tessellatus*
423 (Supplementary Methods), was sequenced for this project and the others were provided by the 1KITE
424 project (Supplementary Methods, http://www.1kite.org, Misof et al., 2014, Peters et al., 2017). A detailed
425 list is presented in Table 1.

426

427 *Coding sequence predictions, transcriptome and genome quality assessments*

428 Coding sequences and peptide sequences were predicted from all transcriptomes using TransDecoder
429 (v2.0.1 https://transdecoder.github.io [last accessed May 4th, 2017]) along with a custom python script
430 to retain the best-scoring entry among overlapping predictions. The coding sequences and peptide
431 sequences from the genomes were retrieved from their official annotated gene sets. All genome and
432 transcriptome gene sets were assessed using BUSCO (v2.0, python 3.4.1, dataset insecta_odb9/2016-
433 10-21, mode proteins) (Waterhouse et al., 2018). This tool identifies near-universal single-copy
434 orthologs by using hidden Markov model profiles from amino acid alignments. CD-HIT-EST v4.6.1 (Li
435 and Godzik, 2006) was run on the protein sequences with a 97.5 percent identity threshold to ensure
436 that all species datasets were filtered to select a single isoform per gene.

437

438 *Orthology delineation*

439 The OrthoDB (Kriventseva et al., 2015) hierarchical orthology delineation procedure was employed to
440 predict orthologous protein groups (OGs). Briefly, protein sequence alignments are assessed to identify

bioRxiv preprint doi: https://doi.org/10.1101/399808; this version posted August 26, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

441 all best reciprocal hits (BRHs) between genes from each pair of species, which are then clustered into
442 OGs following a graph-based approach that starts with BRH triangulation. The annotated proteins from
443 the genomes of *A. planipennis*, *O. taurus* and all transcriptomes were mapped to OrthoDB v8 at the
444 Arthropoda level (with 87 species including four of the beetles with sequenced genomes). Mapping
445 uses the same BRH-based clustering procedure but only allows genes from mapped species to join
446 existing OGs. These OGs were then filtered to identify the 9,720 OGs with representatives from both
447 Polyphaga and Adephaga to focus the study on OGs with evolutionary histories traceable to the last
448 common ancestor of all the beetles, i.e. 5,188 OGs with genes from only one of the two suborders were
449 removed.
450
451 *Species phylogeny*
452 To build an ultrametric phylogeny required for the CAFE analyses, the maximum likelihood molecular
453 species phylogeny was first estimated based on the concatenated superalignment of orthologous amino
454 acid sequences from each of the datasets. Protein sequences of single-copy BUSCO genes and the
455 best-scoring duplicated genes present in all species were individually aligned for each set of BUSCO-
456 identified orthologs using MAFFT with the --auto parameter (Katoh and Standley, 2013) and each result
457 was manually reviewed to exclude poor-quality alignments. Four hundred and five alignments were
458 retained out of 436 and concatenated into a superalignment, partitioned according to the best model
459 for each set of orthologs using aminosan 1.0.2015.01.23 (Tanabe, 2011). RAxML v8.1.2 (-f a -m
460 PROTGAMMA -N 1000) (Stamatakis, 2014), and used to compute the maximum likelihood tree. The
461 monophyly of Geadephaga and Hydradephaga was constrained to match the generally accepted
462 resolution of Adephaga (as in McKenna et al., 2015). The chronos function of the R package ape (v3.4
463 on R 3.2.1, relaxed model) (Paradis et al., 2004) was used to obtain an ultrametric tree and the tip to
464 root length was adjusted to match the approximately 250 million year evolutionary history of crown
465 group Coleoptera (McKenna et al., 2015).
466
467 *Functional annotation and definition of candidate genes*
468 InterProScan was run on all species protein sets (-appl Pfam --goterms, 5.16.55) (Jones et al., 2014)
469 to identify protein families. Additionally, blastp 2.3.0 (Altschul et al., 1997; Camacho et al., 2009) was
470 run against uniref50 (version Jun 22, 2016; Suzek et al., 2015) with an e-value cut-off of 1e-20. An OG

bioRxiv preprint doi: https://doi.org/10.1101/399808; this version posted August 26, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

471 was included in the set of candidate OGs when it had a match to both the uniref50 clusters and Pfam
472 families (Finn et al., 2016) or gene ontologies (The Gene Ontology Consortium, 2017) as detailed in
473 Table 2.
474
475 *CAFE analysis*

476 The number of genes in OGs for each species were counted. All candidates and remaining (control)
477 OGs were pooled together and processed with CAFE 3.1 (Han et al., 2013), to infer gene family
478 evolution in terms of gene gains and losses. First, the python script provided by CAFE was used to
479 estimate the error in our dataset. The CAFE software was then run using the mode in which the gain
480 and loss rates are estimated together (λ) and a second mode in which they are estimated separately
481 (gains=λ, losses=μ). The more complex model was retained as it reached a significantly better score (-
482 199,989 for a single estimated parameter and -199,981 for two distinct estimated parameters, 2x delta
483 log-likelihood = 16, chi-squared distribution, df=1). For the entire analysis, the CAFE overall p-value
484 threshold was kept at its default value (0.01). To run CAFE on each suborder separately, the newick
485 file was pruned to retain only required species using newick utils 1.1.0 (Junier and Zdobnov, 2010).
486
487 *Evolutionary models*

488 To evaluate adaptive OG expansion, the likelihood of the count data was tested by optimizing
489 parameters considering two methods provided by the OUwie R package v1.51 (Beaulieu and O'Meara,
490 2016; Beaulieu et al., 2012). First, a Brownian motion (BM) approach was used, which assumes no
491 selection and thus differences result from a stochastic process whose rate is estimated. Second,
492 Ornstein-Uhlenbeck (OU) models were used. They take into account an optimal family size which is
493 obtained by selective pressure. Two groups were defined in the phylogeny, namely Polyphaga and
494 Adephaga, to which the two different regimes to consider were assigned, plus a third regime to the root.
495 This represents a simplified scenario allowing for the comparison of gene contents between one group
496 and the other rather than attempting to estimate 'levels' of phytophagy or zoophagy across the
497 phylogeny. The models BM1 (Brownian motion with a single rate for the whole tree), BMS (Brownian
498 motion with different rates for each group), OU1 (selection towards the same optimum for both groups)
499 were optimized as null hypotheses (H0) and compared to OUM (selection towards two optima, same
500 variance) and OUMV (selection towards two optima, two variances) models as alternative hypotheses

bioRxiv preprint doi: https://doi.org/10.1101/399808; this version posted August 26, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

501 (H1). The Akaike information criterion corrected for sample-size (AICc) (Hurvich and Tsai, 1989) was
502 used to compare models and an AICc>2 between the best H0 and the best H1 model was considered
503 as significant to prefer the H1 model.
504
505 *Statistical enrichment*
506 All results for candidates and controls were pooled together to obtain a background distribution of
507 positive and negative results. Positive results are those OGs that passed the OUwie analysis, and
508 negative results are all of the 9,720 OGs that did not obtain a significant overall CAFE p-value or did
509 not pass the OUwie analysis. Then, 100,000 random draws (using the R function sample, without
510 replacement) having the sample size of the candidate category to test for enrichment were taken from
511 the background and the significant outcomes for Polyphaga and Adephaga were counted. A p-value
512 was calculated for each group as follows: the number of random draws reaching the amount of
513 significant outcomes found for the candidate category, or more, divided by 100,000. Additionally, the
514 multiple tests conducted on each individual candidate category were corrected for false discovery rate
515 (FDR) using the R p.adjust function (method BH, Benjamini Hochberg).
516
517 *Gene trees*
518 The alignments for the gene trees were produced using MAFFT with the --auto parameter. The gene
519 trees were computed with RAxML v8.1.2 (-f a -m PROTGAMMALGF -N 100) and plotted with EvolView
520 (He et al., 2016; Zhang et al., 2012).

bioRxiv preprint doi: https://doi.org/10.1101/399808; this version posted August 26, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

521 ## Acknowledgements

522 The authors thank Rolf Beutel, Adam Slipinski, Kai Schuette, Ralph Peters, and Eric Anton for providing
523 specimens for the 1KITE samples, as well as Michael Balke, Jia Fenglong, Xu Shengquan, and
524 Alexandros Vasilikopoulos for the *Sinaspidytes wrasei* data. The authors also thank Shanlin Liu,
525 Alexander Donath, Lars Podsiadlowski, Karen Meusemann, and the 1KITE Coleoptera group for
526 providing transcriptome assemblies of 1KITE species. The authors gratefully acknowledge pre-
527 publication access to gene predictions for *Onthophagus taurus* and *Agrilus planipennis* from the Baylor
528 College of Medicine Human Genome Sequencing Center's i5k pilot project, represented by Stephen
529 Richards. The authors additionally thank Catherine Berney who contributed to the production of the
530 *Laparocerus tessellatus* transcriptome, and Nicolas Salamin for advice on the applied methods and
531 feedback on the manuscript.
532
533 This work was supported by the United States National Science Foundation (DEB 1355169) and the
534 United States Department of Agriculture (cooperative agreement 8130-0547-CA) to DDM, the Spanish
535 grant CGL2013-42589-P, awarded by the MINECO and co-financed by FEDER to BCE, the Science
536 Foundation DFG grant BA2152/11-1, 2, the BGI-Shenzhen, the China National Genebank, and the
537 following Swiss National Science Foundation grants: 31003A_143936 (PI), 31003A_173048 (MRR),
538 PP00P3_170664 (RMW), PP00P3_144870 (NA) and PP00P3_172899 (NA). Funding for open access
539 charge: Museum of Natural History of Geneva.
540
541 Some of the computations were performed at the Vital-IT (http://www.vital-it.ch) center for high-
542 performance computing of the Swiss Institute of Bioinformatics.
543

544 ## Author Contributions

545 MS and NA conceived the study. PI delineated orthology. CP prepared the libraries for *L. tessellatus*
546 transcriptome sequencing. HEE, DDM, BM, SS, and XZ provided access to 1KITE transcriptome data.
547 MS conducted the analyses. BCE, MRR, JR, RMW and NA supervised the analyses. MS, RMW, and
548 NA wrote the manuscript, with input from all authors.

bioRxiv preprint doi: https://doi.org/10.1101/399808; this version posted August 26, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

549 **References**

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Barraclough, T.G., Barclay, M.V., and Vogler, A.P. (1998). Species richness: does flower power explain beetle-mania? Curr. Biol. CB *8*, R843-845.

Barrick, J.E., and Lenski, R.E. (2013). Genome dynamics during experimental evolution. Nat. Rev. Genet. *14*, 827–839.

Beaulieu, J.M., and O'Meara, B. (2016). OUwie: Analysis of Evolutionary Rates in an OU Framework.

Beaulieu, J.M., Jhwueng, D.-C., Boettiger, C., and O'Meara, B.C. (2012). Modeling Stabilizing Selection: Expanding the Ornstein–Uhlenbeck Model of Adaptive Evolution. Evolution *66*, 2369–2383.

Betz, O., Thayer, M.K., and Newton, A.F. (2003). Comparative morphology and evolutionary pathways of the mouthparts in spore-feeding Staphylinoidea (Coleoptera). Acta Zool. *84*, 179–238.

Brito, N.F., Moreira, M.F., and Melo, A.C.A. (2016). A look inside odorant-binding proteins in insect chemoreception. J. Insect Physiol. *95*, 51–65.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421.

De Bie, T., Cristianini, N., Demuth, J.P., and Hahn, M.W. (2006). CAFE: a computational tool for the study of gene family evolution. Bioinformatics *22*, 1269–1271.

Farrell, B.D. (1998). "Inordinate Fondness" explained: why are there So many beetles? Science *281*, 555–559.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., et al. (2016). The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. *44*, D279–D285.

Francino, M.P. (2005). An adaptive radiation model for the origin of new gene functions. Nat. Genet. *37*, 573–577.

Gloss, A.D., Vassão, D.G., Hailey, A.L., Nelson Dittrich, A.C., Schramm, K., Reichelt, M., Rast, T.J., Weichsel, A., Cravens, M.G., Gershenzon, J., et al. (2014). Evolution in an ancient detoxification pathway is coupled with a transition to herbivory in the drosophilidae. Mol. Biol. Evol. *31*, 2441–2456.

Goldman-Huertas, B., Mitchell, R.F., Lapoint, R.T., Faucher, C.P., Hildebrand, J.G., and Whiteman, N.K. (2015). Evolution of herbivory in Drosophilidae linked to loss of behaviors, antennal responses, odorant receptors, and ancestral diet. Proc. Natl. Acad. Sci. U. S. A. *112*, 3026–3031.

Grimaldi, D. (1999). The Co-Radiations of Pollinating Insects and Angiosperms in the Cretaceous. Ann. Mo. Bot. Gard. *86*, 373.

Grimaldi, D.A., and Engel, M.S. (2005). Evolution of the insects (Cambridge [U.K.] ; New York: Cambridge University Press).

Hahn, M.W., Han, M.V., and Han, S.-G. (2007). Gene Family Evolution across 12 Drosophila Genomes. PLoS Genet. *3*, e197.

Han, M.V., Thomas, G.W.C., Lugo-Martinez, J., and Hahn, M.W. (2013). Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. Mol. Biol. Evol. *30*,

bioRxiv preprint doi: https://doi.org/10.1101/399808; this version posted August 26, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

1987–1997.

He, Z., Zhang, H., Gao, S., Lercher, M.J., Chen, W.-H., and Hu, S. (2016). Evolview v2: an online visualization and management tool for customized and annotated phylogenetic trees. Nucleic Acids Res. *44*, W236–W241.

Hunt, T., Bergsten, J., Levkanicova, Z., Papadopoulou, A., John, O.S., Wild, R., Hammond, P.M., Ahrens, D., Balke, M., Caterino, M.S., et al. (2007). A comprehensive phylogeny of beetles reveals the evolutionary origins of a superradiation. Science *318*, 1913–1916.

Hurst, L.D. (2009). Fundamental concepts in genetics: genetics and the understanding of selection. Nat. Rev. Genet. *10*, 83–93.

Hurvich, C.M., and Tsai, C.-L. (1989). Regression and time series model selection in small samples. Biometrika *76*, 297–307.

Innan, H., and Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. Nat. Rev. Genet. *11*, 97–108.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: genome-scale protein function classification. Bioinforma. Oxf. Engl. *30*, 1236–1240.

Junier, T., and Zdobnov, E.M. (2010). The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. Bioinforma. Oxf. Engl. *26*, 1669–1670.

Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780.

Keeling, C.I., Yuen, M.M.S., Liao, N.Y., Docking, T.R., Chan, S.K., Taylor, G.A., Palmquist, D.L., Jackman, S.D., Nguyen, A., Li, M., et al. (2013). Draft genome of the mountain pine beetle, Dendroctonus ponderosae Hopkins, a major forest pest. Genome Biol. *14*, R27.

Kondrashov, F.A. (2012). Gene duplication as a mechanism of genomic adaptation to a changing environment. Proc. R. Soc. B Biol. Sci. *279*, 5048–5057.

Kong, Y., Liu, X.-P., Wan, P.-J., Shi, X.-Q., Guo, W.-C., and Li, G.-Q. (2014). The P450 enzyme Shade mediates the hydroxylation of ecdysone to 20-hydroxyecdysone in the Colorado potato beetle, Leptinotarsa decemlineata. Insect Mol. Biol. *23*, 632–643.

Kriventseva, E.V., Tegenfeldt, F., Petty, T.J., Waterhouse, R.M., Simão, F.A., Pozdnyakov, I.A., Ioannidis, P., and Zdobnov, E.M. (2015). OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. Nucleic Acids Res. *43*, D250-256.

Labandeira, C.C. (1997). INSECT MOUTHPARTS: Ascertaining the Paleobiology of Insect Feeding Strategies. Annu. Rev. Ecol. Syst. *28*, 153–193.

Labandeira, C.C. (2002). The history of associations between plants and animals. In: Herrera, C. M., Pellmyr, O.M. (Eds.), Plant-Animal Interactions: An Evolutionary Approach, London.

Labandeira, C.C. (2013). Deep-time patterns of tissue consumption by terrestrial arthropod herbivores. Naturwissenschaften *100*, 355–364.

Leschen, R.A.B., and Buckley, T.R. (2007). Multistate characters and diet shifts: evolution of Erotylidae (Coleoptera). Syst. Biol. *56*, 97–112.

Li, D., and Zhang, J. (2014). Diet shapes the evolution of the vertebrate bitter taste receptor gene repertoire. Mol. Biol. Evol. *31*, 303–309.

bioRxiv preprint doi: https://doi.org/10.1101/399808; this version posted August 26, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinforma. Oxf. Engl. *22*, 1658–1659.

McKenna, D.D., Sequeira, A.S., Marvaldi, A.E., and Farrell, B.D. (2009). Temporal lags and overlap in the diversification of weevils and flowering plants. Proc. Natl. Acad. Sci. U. S. A. *106*, 7083–7088.

McKenna, D.D., Wild, A.L., Kanda, K., Bellamy, C.L., Beutel, R.G., Caterino, M.S., Farnum, C.W., Hawks, D.C., Ivie, M.A., Jameson, M.L., et al. (2015). The beetle tree of life reveals that Coleoptera survived end-Permian mass extinction to diversify during the Cretaceous terrestrial revolution: Phylogeny and evolution of Coleoptera (beetles). Syst. Entomol. *40*, 835–880.

McKenna, D.D., Scully, E.D., Pauchet, Y., Hoover, K., Kirsch, R., Geib, S.M., Mitchell, R.F., Waterhouse, R.M., Ahn, S.-J., Arsala, D., et al. (2016). Genome of the Asian longhorned beetle (Anoplophora glabripennis), a globally significant invasive species, reveals key functional and evolutionary innovations at the beetle-plant interface. Genome Biol. *17*, 227.

Misof, B., Liu, S., Meusemann, K., Peters, R.S., Donath, A., Mayer, C., Frandsen, P.B., Ware, J., Flouri, T., Beutel, R.G., et al. (2014). Phylogenomics resolves the timing and pattern of insect evolution. Science *346*, 763–767.

Neafsey, D.E., Waterhouse, R.M., Abai, M.R., Aganezov, S.S., Alekseyev, M.A., Allen, J.E., Amon, J., Arcà, B., Arensburger, P., Artemov, G., et al. (2015). Highly evolvable malaria vectors: The genomes of 16 Anopheles mosquitoes. Science *347*, 1258522.

Niehuis, O., Hartig, G., Grath, S., Pohl, H., Lehmann, J., Tafer, H., Donath, A., Krauss, V., Eisenhardt, C., Hertel, J., et al. (2012). Genomic and morphological evidence converge to resolve the enigma of Strepsiptera. Curr Biol. *22*, 1309–1313.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. Bioinforma. Oxf. Engl. *20*, 289–290.

Parsons, K.J., Concannon, M., Navon, D., Wang, J., Ea, I., Groveas, K., Campbell, C., and Albertson, R.C. (2016). Foraging environment determines the genetic architecture and evolutionary potential of trophic morphology in cichlid fishes. Mol. Ecol. *25*, 6012–6023.

Pauchet, Y., Wilkinson, P., Chauhan, R., and ffrench-Constant, R.H. (2010). Diversity of Beetle Genes Encoding Novel Plant Cell Wall Degrading Enzymes. PLOS ONE *5*, e15635.

Peters, R.S., Krogmann, L., Mayer, C., Donath, A., Gunkel, S., Meusemann, K., Kozlov, A., Podsiadlowski, L., Petersen, M., Lanfear, R., et al. (2017). Evolutionary History of the Hymenoptera. Curr Biol. *27*, 1013–1018.

Ragsdale, E.J., Müller, M.R., Rödelsperger, C., and Sommer, R.J. (2013). A developmental switch coupled to the evolution of plasticity acts through a sulfatase. Cell *155*, 922–933.

Robinson, G.E., Hackett, K.J., Purcell-Miramontes, M., Brown, S.J., Evans, J.D., Goldsmith, M.R., Lawson, D., Okamuro, J., Robertson, H.M., and Schneider, D.J. (2011). Creating a Buzz About Insect Genomes. Science *331*, 1386–1386.

Roncalli, V., Cieslak, M.C., Passamaneck, Y., Christie, A.E., and Lenz, P.H. (2015). Glutathione S-Transferase (GST) Gene Diversity in the Crustacean Calanus finmarchicus – Contributors to Cellular Detoxification. PLOS ONE *10*, e0123322.

Schoville, S.D., Chen, Y.H., Andersson, M.N., Benoit, J.B., Bhandari, A., Bowsher, J.H., Brevik, K., Cappelle, K., Chen, M.-J.M., Childers, A.K., et al. (2018). A model species for agricultural pest genomics: the genome of the Colorado potato beetle, Leptinotarsa decemlineata (Coleoptera: Chrysomelidae). Sci. Rep. *8*, 1931.

bioRxiv preprint doi: https://doi.org/10.1101/399808; this version posted August 26, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Seehausen, O., Butlin, R.K., Keller, I., Wagner, C.E., Boughman, J.W., Hohenlohe, P.A., Peichel, C.L., Saetre, G.-P., Bank, C., Brännström, A., et al. (2014). Genomics and the origin of species. Nat. Rev. Genet. *15*, 176–192.

Shaw, K.L., and Lesnick, S.C. (2009). Genomic linkage of male song and female acoustic preference QTL underlying a rapid species radiation. Proc. Natl. Acad. Sci. U. S. A. *106*, 9737–9742.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics *31*, 3210–3212.

Slipinski, S.A., Leschen, R.A.B., and Lawrence, J.F. (2011). Order Coleoptera Linnaeus, 1758. In: Zhang, Z.-Q. (Ed.) Animal biodiversity: An outline of higher-level classification and survey of taxonomic richness. 6.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinforma. Oxf. Engl. *30*, 1312–1313.

Suchan, T., and Alvarez, N. (2015). Fifty years after Ehrlich and Raven, is there support for plant–insect coevolution as a major driver of species diversification? Entomol. Exp. Appl. *157*, 98–112.

Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B., Wu, C.H., and UniProt Consortium (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinforma. Oxf. Engl. *31*, 926–932.

Tanabe, A.S. (2011). Kakusan4 and Aminosan: two programs for comparing nonpartitioned, proportional and separate models for combined molecular phylogenetic analyses of multilocus sequence data. Mol. Ecol. Resour. *11*, 914–921.

The Gene Ontology Consortium (2017). Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Res. *45*, D331–D338.

Tribolium Genome Sequencing Consortium, Richards, S., Gibbs, R.A., Weinstock, G.M., Brown, S.J., Denell, R., Beeman, R.W., Gibbs, R., Beeman, R.W., Brown, S.J., et al. (2008). The genome of the model beetle and pest Tribolium castaneum. Nature *452*, 949–955.

Voelckel, C., and Jander, G. (2014). Annual Plant Reviews, Insect-Plant Interactions.

Waterhouse, R.M. (2015). A maturing understanding of the composition of the insect gene repertoire. Curr. Opin. Insect Sci. *7*, 15–23.

Waterhouse, R.M., Zdobnov, E.M., and Kriventseva, E.V. (2011). Correlating traits of gene retention, sequence divergence, duplicability and essentiality in vertebrates, arthropods, and fungi. Genome Biol. Evol. *3*, 75–86.

Waterhouse, R.M., Seppey, M., Simão, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E.V., and Zdobnov, E.M. (2018). BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. Mol. Biol. Evol. *35*, 543–548.

Zhang, H., Gao, S., Lercher, M.J., Hu, S., and Chen, W.-H. (2012). EvolView, an online tool for visualizing, annotating and managing phylogenetic trees. Nucleic Acids Res. *40*, W569–W572.

Zhang, S.-Q., Che, L.-H., Li, Y., Liang, D., Pang, H., Ślipiński, A., and Zhang, P. (2018). Evolutionary history of Coleoptera revealed by extensive sampling of genes and species. Nat. Commun. *9*, 205.

550