Transposon accumulation at xenobiotic gene family loci: a comparative genomic analysis in aphids

Tobias Baril[1,2]*, Adam Pym[1,3], Chris Bass[1,4], Alex Hayward[1,5*]

[1] Centre for Ecology and Conservation, University of Exeter, Penryn Campus, Cornwall, TR10 9FE, UK

[2] ORCID: 0000-0002-5936-7531

[3] ORCID: 0000-0001-8754-4016

[4] ORCID: 0000-0002-2590-1492

[5] ORCID: 0000-0001-7413-718X

* Corresponding Author information:

Tobias Baril: tb529@exeter.ac.uk

Alex Hayward: alex.hayward@exeter.ac.uk

Abstract

The evolution of resistance is a major challenge for the sustainable control of pests and pathogens. Thus, a deeper understanding of the evolutionary and genomic mechanisms underpinning resistance evolution is required to safeguard health and food production. Several studies have implicated transposable elements (TEs) in xenobiotic resistance evolution in insects. However, analyses are generally restricted to one insect species and/or one or a few xenobiotic gene families (XGFs). We examine evidence for TE accumulation at XGFs by performing a comparative genomic analysis across 20 aphid genomes, considering major subsets of XGFs involved in metabolic resistance to insecticides: Cytochrome P450s, glutathione S-transferases, esterases, UDP-glucuronosyltransferases, and ABC transporters. We find that TEs are significantly enriched at XGFs compared to other genes. XGFs show similar levels of TE enrichment to housekeeping genes. But unlike housekeeping genes, XGFs are not constitutively expressed in germline cells, supporting the selective enrichment of TEs at XGFs, rather than enrichment due to chromatin availability. Hotspots of extreme TE enrichment occur around certain XGFs. In aphids of agricultural importance, we find particular enrichment of TEs around cytochrome P450 genes with known functions in the detoxification of synthetic insecticides. Our results provide evidence supporting a general role for TEs as a source of genomic variation at host XGFs, and highlight the existence of considerable variability in TE content across XGFs and host species. These findings demonstrate the need for detailed functional verification analyses to clarify the significance of individual TE insertions and elucidate underlying mechanisms at TE-XGF hotspots.

Introduction

Xenobiotics are substances that are foreign to an organism, including naturally occurring plant allelochemicals and man-made insecticides (Li et al. 2007). Xenobiotic toxicity can present a strong selective pressure, leading to the rapid evolution of resistance via efficient

46 xenobiotic removal, metabolism, and tolerance (Chung et al. 2007; Li et al. 2007).
47 Consequently, numerous insects have evolved resistance to both plant allelochemicals,
48 produced for plant defence, and synthetic insecticides, developed to protect economically
49 important crops from insect pests, or humans and livestock from insect vectored diseases
50 (Sparks et al. 2021; Sternberg and Thomas 2018). The intensive use of numerous synthetic
51 insecticides has led many insects to evolve resistance to multiple insecticide classes (Li et
52 al. 2007; Bradshaw et al. 2016), severely affecting our ability to control target insects (Singh
53 et al. 2021). The emergence of multiple insecticide resistance is a major global challenge,
54 threatening global food security in the case of agricultural pests, and human and animal
55 health in the case of disease vectors (Worner and Gevrey 2006; Deutsch et al. 2018).

56 Two major mechanisms are frequently implicated in conferring xenobiotic resistance
57 across a wide range of arthropods (Li et al. 2007; Kliot and Ghanim 2012): (i) 'target-site
58 resistance', involving structural changes (mutations) in the gene encoding the insecticide
59 target protein, that make it less sensitive to the toxic effect of the insecticide; (ii) 'metabolic
60 resistance', involving the increased production or activity of metabolic enzymes that break
61 down or sequester the insecticide. Regarding the latter, several major gene families are
62 associated with metabolic resistance, which act during three distinct phases of xenobiotic
63 metabolism (Kennedy and Tierney 2013): (i) Cytochrome P450 monooxygenases (CYP
64 genes); (ii) glutathione S-transferases (GSTs), esterases, and UDP-
65 glucuronosyltransferases (UGTs); (iii) ATP-binding cassette transporters (ABC transporters).
66 We briefly summarise these three phases below. During phase I, CYP genes catalyse
67 functionalisation reactions, where hydrophobic xenobiotics are typically converted to more
68 water-soluble metabolites through the addition of functional groups (Kennedy and Tierney
69 2013). CYP genes are diverse enzymes involved in several purposes from biosynthesis to
70 metabolism, and are sometimes referred to as 'nature's blowtorch', due to their high-valence
71 iron chemistry oxidation mechanism (Guengerich 2009). CYP genes can mediate resistance
72 to all classes of insecticides due to their broad substrate specificity and versatility, whilst they

73 are also involved in several other processes, such as juvenile hormone and ecdysteroid
74 metabolism, and fatty acid biosynthesis (Feyereisen 2005; Kennedy and Tierney 2013;
75 Feyereisen 1999). During phase II, GSTs, UGTs, and esterases catalyse conjugation
76 reactions between phase I substrates and endogenous molecules to form water-soluble
77 metabolites (Kennedy and Tierney 2013). During phase III, ABC transporters, which are cell
78 membrane transport proteins that efflux toxins and modified toxins from the cell, export
79 xenobiotics and products of phase I and II (Gott et al. 2017). In addition, ABC transporters
80 can also block the cellular import of xenobiotics to protect the host, which is termed phase 0.
81 (Gott et al. 2017). Insecticide resistance can arise via mutations acting on the genes
82 involved in the processes described above through various genetic mechanisms, including
83 upregulation, changes to coding sequence, and gene amplification (Li et al. 2007). It is also
84 important to note that many XGF members are involved in other metabolic processes
85 beyond xenobiotic resistance. For example, some GSTs have roles in processes including
86 eye pigmentation, intracellular transport, and cell signalling pathways (Ketterman et al. 2011;
87 Ranson and Hemingway 2005), whilst ABC transporters have numerous roles in processes
88 including heme biosynthesis, iron homeostasis, and protection against oxidative stress
89 (Dermauw and Van Leeuwen 2014). UGTs also have roles in the modulation of endobiotics
90 and olfactory processes (Ahn et al. 2012), and esterases have roles in neurodevelopment
91 and pheromone signalling (Gilbert and Gill 2014).

92 There is growing evidence that transposable elements (TEs) can play an important
93 role in the evolution of xenobiotic resistance in insects (Gilbert et al. 2021; Rostant et al.
94 2012), with examples from several major lineages: Diptera: *Drosophila* (Feyereisen 2005;
95 Bogwitz et al. 2005; Schmidt et al. 2010; Aminetzach et al. 2005; Remnant et al. 2013;
96 Mateo et al. 2014; Salces-Ortiz et al. 2020), *Culex* (Itokawa et al. 2011; Darboux et al. 2007),
97 *Musca domestica* (Li et al. 2007; Kasai and Scott 2001), *Anopheles funestus* (Weedall et al.
98 2020) Lepidoptera: *Heliothis* (Gahan et al. 2001; Yang et al. 2007; Zhao et al. 2010; Chen
99 and Li 2007), *Pectinophora gossypiella* (Fabrick et al. 2011; Wang et al. 2019) Hemiptera:

100 *Myzus persicae* (Singh et al. 2020; Panini et al. 2021). TEs are DNA sequences capable of
101 moving from one genomic location to another within the genome. TEs occur in nearly all
102 eukaryotic genomes, and are implicated in the evolution of host genomic novelty through
103 diverse processes including the modification of regulatory networks, chromosomal
104 rearrangements, exon shuffling, and donation of coding sequence (Bourque et al. 2018;
105 Wells and Feschotte 2020; Cosby et al. 2019; Sundaram et al. 2014). In the case of
106 xenobiotic resistance, TEs are reported to have contributed to evolvability via myriad
107 mechanisms, including: gene amplification and duplication (Singh et al. 2020; Remnant et al.
108 2013), knockout of a susceptible allele in heterozygous individuals (Panini et al. 2021),
109 increases in detoxification gene expression (Bogwitz et al. 2005; Itokawa et al. 2011), allelic
110 succession leading to increases in resistance gene copy number (Schmidt et al. 2010), and
111 alternative splicing and production of truncated proteins that prevent interactions with
112 xenobiotics (Darboux et al. 2007; Gahan et al. 2001; Yang et al. 2007; Zhao et al. 2010).

113 Studies examining the contribution of TEs to resistance evolution suggest that there
114 can be an additive effect of successive TE insertions at focal host loci over time. For
115 example, in the well-known case of the cytochrome P450 (CYP) gene *CYP6G1* and DDT
116 resistance in *Drosophila*, successive TE insertions are linked with an increasing ability to
117 detoxify the insecticide (Catania et al. 2004; Daborn et al. 2002; Chung et al. 2007; Schmidt
118 et al. 2010). A similar process involving the accumulation of TEs, in this case at *CYP6CY3,*
119 is implicated in the evolution of resistance to nicotine, following a host plant shift to tobacco
120 in the aphid *Myzus persicae nicotianae* (Puinean et al. 2010; de Little et al. 2017). More
121 widely, there is evidence that the selective accumulation of TEs at rapidly evolving host loci
122 under strong selection may represent a general evolutionary process. For example, studies
123 have reported the enrichment of TEs at other gene classes, such as immune genes and
124 those involved in responses to external stimuli (van de Lagemaat et al. 2003; Song et al.
125 1998). Existing studies strongly suggest that TEs are involved in individual cases of
126 resistance to certain insecticides, but there is currently very limited understanding of how

127 associations with TEs vary among different classes of XGFs, or among whole clades of
128 insects. However, patterns in TE accumulation remain poorly described, and it is unclear to
129 what extent TEs may be enriched at xenobiotic gene families (XGFs) compared to other host
130 genes, or how much variability exists across XGFs.

131 Here we consider patterns of TE accumulation at XGFs in the genomes of 20 species
132 sampled from across aphid phylogeny, for which high-quality genome assemblies are
133 available. The aphid family is a relevant clade within which to explore the genomic bases of
134 insecticide resistance, as it includes numerous globally important crop pests responsible for
135 causing billions of US dollars of crop losses annually (Van Emden and Harrington 2017), and
136 members that have been intensively targeted with insecticides and evolved resistance. For
137 example, the severely damaging global crop pest *Myzus persicae* (the green peach aphid) is
138 resistant to at least 82 insecticides, via at least eight different resistance mechanisms (Bass
139 et al. 2014; Mota-Sanchez and Wise 2023; Singh et al. 2021). We hypothesise similar
140 broadscale patterns for TE accumulation at XGFs among aphid genomes, given similar
141 challenges from xenobiotics that target conserved biological pathways. Meanwhile, given the
142 strong selection pressure that can arise from xenobiotic exposure, and repeated xenobiotic
143 challenges over time, we anticipate an ability to detect patterns at the species-level.

144 Since the TE landscapes of aphids are poorly described, we begin by characterising
145 patterns of TE content across our focal genomes. After this, we examine key fundamental
146 questions relating to the pattern of associations between TEs and XGFs in aphid genomes.
147 (i) Firstly, we test whether XGFs are enriched for TEs compared to other host genes, to
148 evaluate the overall signature of TE accumulation at XGFs. (ii) Secondly, we examine the
149 extent to which TE content varies among individual XGFs and among major XGF
150 classifications, to determine if contributions from TEs are more pronounced at certain types
151 of XGF. (iii) Thirdly, we consider if specific TE classifications are enriched at XGFs (e.g DNA
152 TEs, rolling circles, *Penelope*-like elements, LINEs, SINEs, and LTRs), to explore if
153 particular TEs are predisposed to contribute to the evolution of host resistance. (iv) Fourthly,

154 we assess whether patterns of TE accumulation at XGFs vary among aphid species to
155 examine if patterns vary across aphid phylogeny. (v) Fifthly, we investigate TE enrichment at
156 XGFs that have demonstrated associations with the detoxification of synthetic insecticides,
157 to test if these show particularly strong signatures of TE recruitment. (vi) Lastly, we explore
158 whether our results suggest a selective role for patterns in TE accumulation at XGFs versus
159 the alternative explanation of increased availability for TE insertion due to chromatin
160 availability.

161 Results

162 *Transposable Element Landscapes in Aphids*

163 TEs make considerable contributions to genome size in eukaryotes (Kidwell 2002; Gregory
164 2005). In aphids, we find that TE content varies from 8.21% (*Melanaphis sacchari*) to
165 35.52% (*Metopolophium dirhodum*) of total genome size (Figure 1B, Supplemental Table
166 S1). This is in accordance with levels of TE content identified among a large-scale survey of
167 196 insect species, where TE content varied from 0.08% to 53.93% of total genome size
168 (Peccoud et al. 2017) . Similarly, consistent with findings in other eukaryotic lineages, we
169 report a significant strong positive association between genome size and TE content in
170 aphids (Linear regression, $F_{1,19}$=46.35, $p<0.01$, R^2=0.69, Supplemental Figure S1). Aphids
171 show variation in sexual and asexual lifecycles (Van Emden and Harrington 2017). It is
172 hypothesised that asexual lineages would suffer from Mullers ratchet as deleterious
173 mutations accumulate (Muller 1964). Despite this, we find no significant difference in TE
174 content between holocyclic and anholocyclic lineages (Wilcoxon Rank Sum, $W_{14,6}$ = 35,
175 $p>0.05$) (Figure 1).

176 TEs have a non-random distribution within host genomes (Bourque et al. 2018). As
177 expected, given the predominantly deleterious effects of TE insertion into host genes
178 (Bourque et al. 2018; Sultana et al. 2017), exonic regions exhibit the lowest levels of TE

179 sequence content in aphid genomes (mean TE content in exons = 1.90% of total genome
180 size). Across 19 of the 20 aphid genomes considered here, the greatest TE coverage
181 (expressed as proportion of total genome size) is found in 5' and 3' gene flanking regions,
182 where it varies between 3.49% (*M. sacchari*) and 17.94% (*M. dirhodum*) (Figure 2).
183 However, this pattern is not universal across all aphid species. In *Sipha flava,* the majority of
184 TE coverage is present in intronic regions (10.60%), while in *M. cerasi*, the majority occurs in
185 intergenic space (11.28%) (Figure 2).

186 The amount of TE sequence and the frequency of TE insertions in genic versus
187 intergenic regions in aphid genomes is not significantly correlated with genome
188 compactness, described as the ratio of genic (intron and exon) to intergenic (gene flanks and
189 intergenic space) base pairs (Spearman's Rank, S=1680, p>0.05, rho=-0.09, Supplemental
190 Figure S2, Supplemental Table S2). *S. flava* has the most compact genome, with a
191 genic:intergenic base pair ratio of 1.50:1 (250.0 Mb:166.8 Mb), and so presumably a higher
192 likelihood of TE insertion into genic regions compared to intergenic space, leading to the
193 observation that the majority of TE insertions are found in intronic regions. In contrast, *M.*
194 *cerasi* has the least compact genome, with a much lower genic:intergenic base pair ratio of
195 0.13:1 (76.1 Mb:559.2 Mb). In less compact genomes such as *M. cerasi*, TEs can
196 accumulate in expanded intergenic regions that act as 'genomic safe havens', where
197 insertion is less likely to exert deleterious effects (Arkhipova 2018). For context, the human
198 genome (GCF_009914755.1), has a TE content of 44%-69% (Nurk et al. 2022; de Koning et
199 al. 2011) and a genic:intergenic base pair ratio of 0.55:1 (1717.7 Mb:3117.3 Mb), making it
200 several times more compact than *M. cerasi*, but much less compact than *S. flava*.

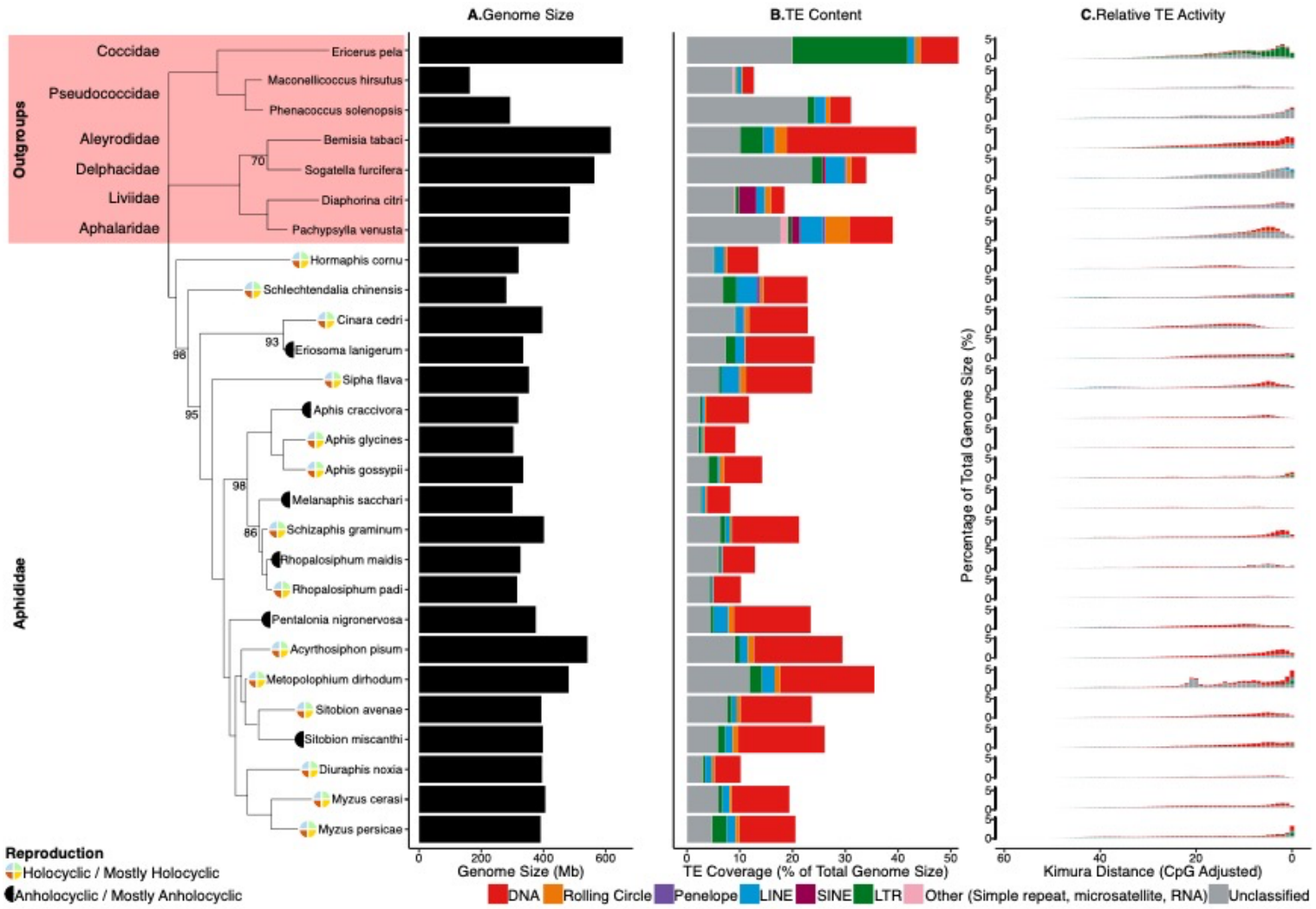


Figure 1. Summary of TE content and activity in aphid genomes, plotted in phylogenetic order. Nodes in the aphid phylogeny with bootstrap support <100 are labelled. Reproductive mode (Van Emden and Harrington 2017) is labelled next to species names. Genome size is represented by the black bars. TE content is expressed as percentage of total genome size for each species, with major TE classifications represented by the colours indicated in the key. Kimura 2-parameter distance (CpG adjusted) from each TE family consensus is used as a proxy for relative TE activity, where a lower Kimura distance indicates more

recent TE activity. Due to challenges with accurately estimating TE age, divergence from consensus better reflects relative TE activity, with lower Kimura distances signalling more recent TE activity. Activity is organised such that recent TE activity is towards the RHS of the X axis.

208 Across all aphid species and in most individual genomes, the dominant TE
209 classification is DNA TEs, which comprise between 36.6% (*Schlechtendalia chinensis*) and
210 68.9% (*Aphis craccivora*) of total TE content (μ = 45.27%, SD = 18.00%, Figure 1B, Table 1,
211 Supplemental Table S1). SINEs are present only in the genomes of 7 of the 20 species
212 considered (Supplemental Table S1) and, where present, they comprise just 0.009% (*A.*
213 *pisum*) to 0.1% (*A. craccivora*) of total TE content.

TE Classification	Mean Share of TE Content (% of total TE content)	Standard Deviation of Share of TE Content (% of total TE content)
DNA	45.27	18.00
Unclassified	35.95	14.66
LINE	7.52	3.76
LTR	6.19	8.00
Rolling Circle	3.46	2.16
SINE	2.14	5.24
Penelope	0.40	0.75

214 Table 1. Mean and standard deviation of the percentage of total TE content attributed to
215 each of the main TE classifications for the 20 aphid genomes included in this study.

216

217 There is little evidence of ancient TE activity in aphid genomes, as indicated by the
218 relatively low number of TEs showing high divergence from their respective TE consensus
219 models in repeat landscape plots (Figure 1C). On the other hand, there is considerable
220 evidence of relatively recent TE activity in most aphid species, with particular evidence of
221 very recent TE activity in *Myzus persicae* and *M. dirhodum*, indicated by large numbers of
222 TEs with low genetic distance from their respective consensus sequences (Figure 1C). The
223 relative absence of ancient TE activity is presumably a consequence of a relatively rapid
224 genomic turnover rate in aphids, and contrasts strongly with TE landscapes in mammals,
225 where a low rate of genomic turnover leads to considerable accumulations of ageing TEs

226 (Blass et al. 2012). Instead, the dynamics we observe for aphids are consistent with patterns
227 reported for other insects, such as lepidopterans (Lavoie et al. 2013; Baril and Hayward
228 2022).

229 The extent to which TE content differs among the genomes of closely related species
230 can vary greatly. We find no significant effect of phylogeny on TE abundance (No. of TEs
231 present in a genome) or TE diversity (No. distinct TE families present in a genome (Wicker
232 et al. 2007)) in aphids (Phylogenetic GLMM, TE abundance: posterior mean = 0.592, 95%
233 HPD = 0.005,0.999; TE diversity: posterior mean = 0.022, 95% HPD = 0.000,0.111). Thus,
234 more closely related aphid species do not possess similar numbers of TEs or a similar level
235 of TE diversity. We also find that most TE families present in aphid genomes are species-
236 specific, with almost half of the TE families present in a particular aphid genome being
237 unique to that species (μ = 46.97%) (Figure 2). On average, 10.02% of the TE families
238 identified in an aphid genome are shared among all 20 species (Figure 2). The remaining
239 43.01% of TE families are shared in similar proportions among 2 to 19 aphid species (i.e. μ =
240 2.53% per category, Figure 2). The strong signature of species-specificity observed for aphid
241 TE families suggests that the independent gain of new TE families is a major factor driving
242 aphid TE landscapes. Collectively, these findings suggest a dynamic repeat landscape in
243 aphid genomes, characterised by relatively frequent gain and turnover of TEs.

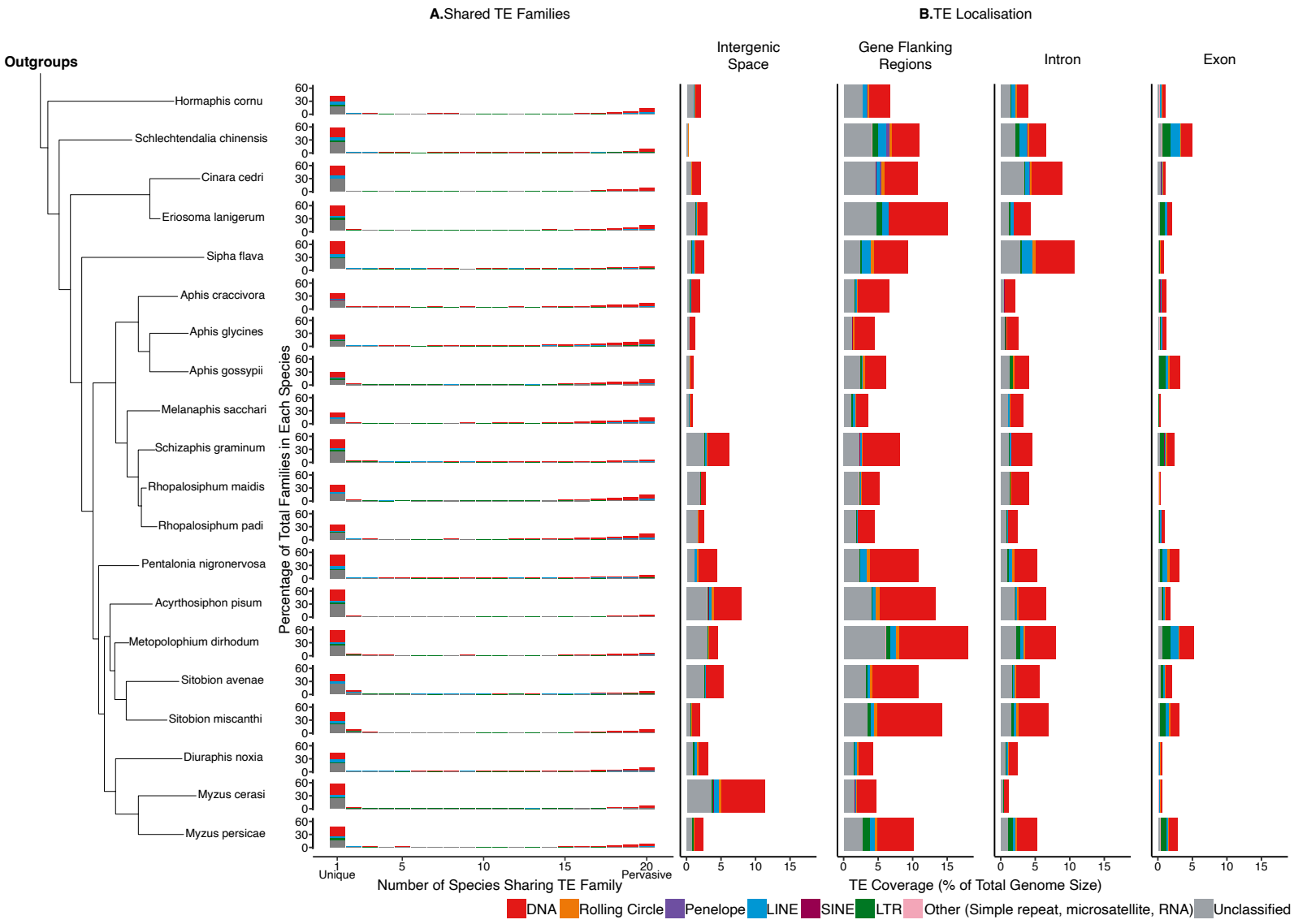


Figure 2. Overview of the extent to which TE families are shared among aphid genomes, and their insertion locations. **A**. Quantification of shared and unique TE families in aphids. Main TE classifications are represented by the colours indicated in the key. **B**. TE genome compartment occupancy, expressed by TEs as a percentage of total genome size. Gene flanking regions are defined as 20kb directly upstream or downstream of the gene body

248 *TEs are enriched at XGFs compared to other host genes*

249 Previous studies have implicated TEs in the evolution of xenobiotic resistance across a
250 range of insects and resistance loci (Li et al. 2007). However, the extent to which TE
251 association at XGFs represents a general mechanism for the evolution of new resistance
252 phenotypes remains unclear.

253 Across all aphid genomes considered here, we find significant enrichment of TEs,
254 both in TE coverage and TE count, at XGFs in comparison to all other genes excluding
255 XGFs (TE coverage: Wilcoxon Rank Sum, $W_{6488,401787}$ = 732,129,770, $p<0.01$; TE count:
256 Wilcoxon Rank Sum, $W_{6488,401787}$ = 770,552,556, $p<0.01$) (Figure 3, Supplemental Figure
257 S3). Specifically, XGFs in aphids have a mean of 2.4x TE coverage and 2.1x TE count
258 compared to all other genes.

259

260 *TE content varies among individual XGF loci*

261 We observe an enrichment of TEs around all major XGF categories considered compared to
262 other host genes: Cytochrome P450s (CYPs), 2.4x TE coverage, 2.1x TE count; glutathione
263 S-transferases (GSTs), 2.7x TE coverage, 2.3x TE count; esterases, 2.0x TE coverage, 1.8x
264 TE count; UDP-glucuronosyltransferases (UGTs), 2.0x TE coverage, 1.7x TE count; ABC
265 transporters (ABCs), 2.8x TE coverage, 2.4x TE count (Figure 3, Supplemental Figure S3,
266 Supplemental Table S3). Whilst all major XGF categories are enriched for TEs, there is no
267 significant difference in relative enrichment among XGF types, suggesting relatively equal
268 levels of enrichment across XGF types (Kruskal-Wallis, TE sequence: $\chi^2_4 = 7.72$, $p = 0.10$;
269 TE count: $\chi^2_4 = 4.25$, $p = 0.37$). However, considerable variability in TE content is present
270 within XGF categories, attributable to extremely large accumulations of TEs at certain
271 individual XGFs (Table 2, Figure 4). Some of these XGFs are unusually large in terms of
272 sequence length compared to the mean size for the XGF type in question, as indicated by

273 gene size Z-scores > 3 (Table 2), which is potentially a consequence of the increased
274 presence of TEs in their intronic regions (Figure 4). To investigate this, we consider gene
275 sizes with genic TEs removed. We find that 5/18 of the most TE-enriched XGFs have
276 significantly inflated gene sizes when TE insertions are removed, indicated by a Z-score of >
277 3. This suggests that these loci were significantly larger than average before accounting for
278 TE contributions. However, the remaining 13 highly TE-enriched XGFs have gene sizes
279 within the expected range in the absence of TEs, suggesting that TEs are responsible for the
280 significant increases in gene size at the majority of TE-enriched XGF loci.

281 Individual XGFs with the greatest enrichment of TE coverage and TE count are listed
282 in table 2 (i.e. >35kb TE sequence and >80 TE insertions, compared to a non-XGF mean of
283 1.5kb TE sequence and 1.48 TE insertions). Whilst all XGF types are represented in the TE
284 hotspots listed in Table 2, there are markedly more hotspots at UGTs and CYP genes, and
285 fewer at ABCs, GSTs, and Esterases.

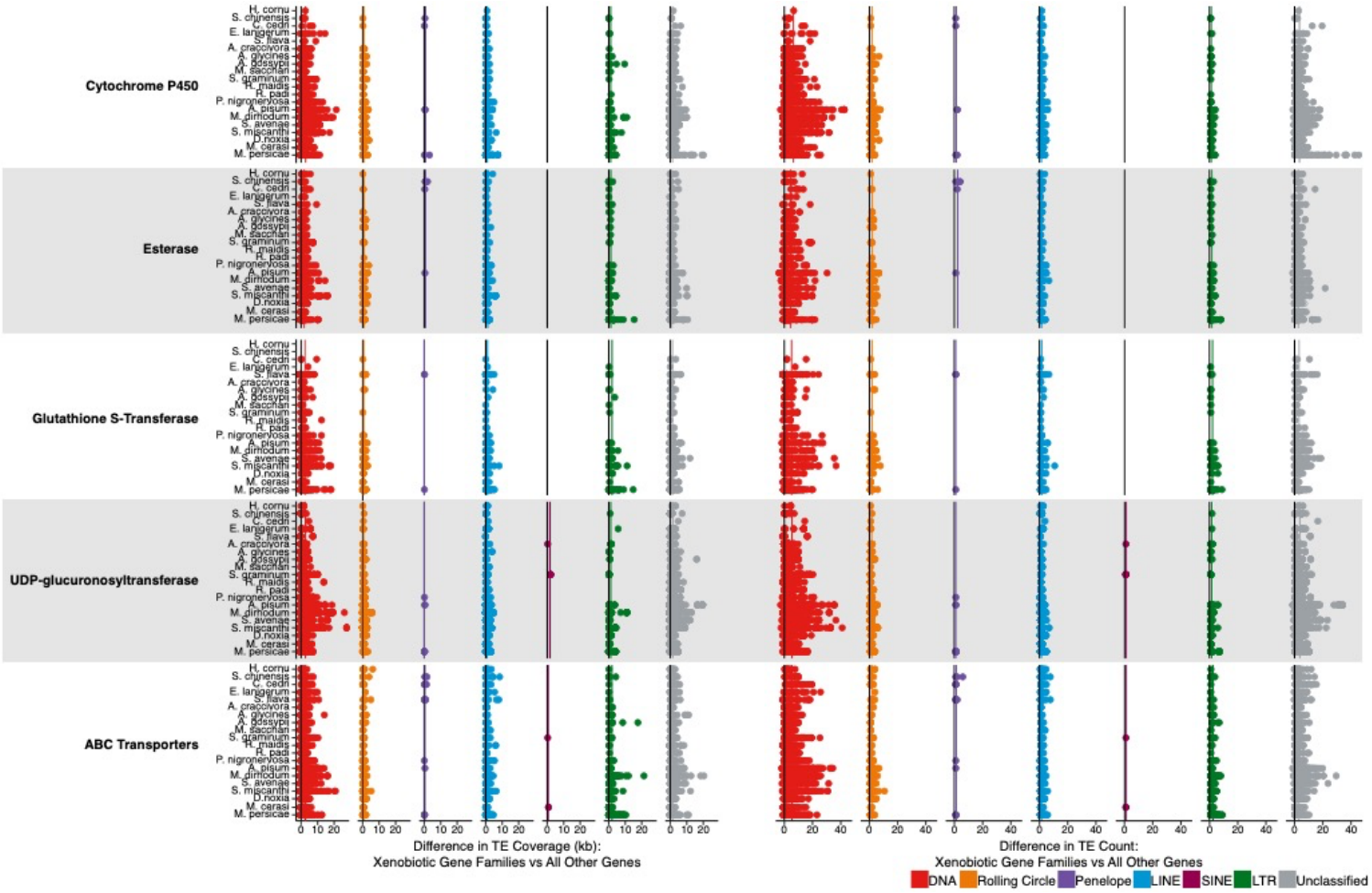


Figure 3. Differences in TE coverage and TE count at XGFs compared to all other host genes. Coverage and count of genic (exon and intron) XGF sequence plus flanking regions (20kb directly upstream or downstream of the gene body) is shown. Each point represents a single XGF locus. Black lines indicate the expected TE coverage difference whereby XGFs and other genes are equally enriched for TEs (i.e. 0 enrichment). Coloured lines indicate species mean coverage and count for each XGF locus and TE type. Major TE classifications are indicated by the colours in the key.

Aphid Species	Gene ID	XGF Type	Mean Gene Size (by gene type)	Gene Size in bp (Z-score)	XGF associated TE Coverage (bp)	Fold-increase above aphid-wide enrichment (XGF-type specific)	Fold-increase above aphid-wide enrichment (non-XGF genes)	XGF associated TE Count	Fold-increase above aphid-wide enrichment (XGF-type specific)	Fold-increase above aphid-wide enrichment (non-XGF genes)	Gene Size after removing genic TEs (Z-score)
Greatest Enrichment of TE Coverage											
Acyrthosiphon pisum	**2399_UGT344E8_v2**	**UDP-glucuronosyltransferase**	**8,302**	**55,119 (4.63)**	**55,257**	**9.05**	**36.38**	**118**	**9.08**	**79.73**	**24,991 (2.45)**
Acyrthosiphon pisum	**2354_UGT344D11_v2**	**UDP-glucuronosyltransferase**	**8,302**	**57,143 (4.83)**	**45,208**	**7.40**	**29.76**	**96**	**7.38**	**64.86**	**26,620 (2.68)**
Myzus persicae	**125_CYP380C52**	**Cytochrome P450**	**7,221**	**100,885 (9.70)**	**44,205**	**6.28**	**29.10**	**102**	**6.00**	**68.92**	**56,825 (5.97)**
Myzus persicae	**109_CYP380C31**	**Cytochrome P450**	**7,221**	**99,455 (9.55)**	**42,790**	**6.08**	**28.17**	**94**	**5.53**	**63.51**	**57,937 (6.10)**
Myzus persicae	352_gst_g525.t1	Glutathione S-Transferase	3,450	16,231 (2.65)	40,940	5.02	26.95	45	2.50	30.41	10,854 (1.54)
Myzus persicae	**105_CYP380C19**	**Cytochrome P450**	**7,221**	**61,287 (5.60)**	**39,133**	**5.56**	**25.76**	**100**	**5.88**	**67.57**	**34,033 (3.17)**
Metopolophium dirhodum	2360_UGT344A13	UDP-glucuronosyltransferase	8,302	18,807 (1.04)	38,003	6.22	25.02	45	3.46	30.41	8,327 (0.11)

Metopolophium dirhodum	1793_UGT345A3	UDP-glucuronosyltransferase	8,302	6,504 (-0.18)	37,746	6.18	24.85	29	2.23	19.59	4,684 (-0.40)
Sitobion miscanthi	690_UGT341A5_v2	UDP-glucuronosyltransferase	8,302	4,437 (-0.38)	36,666	6.00	24.14	42	3.23	28.38	3,950 (-0.50)
Sitobion miscanthi	687_UGT341A5	UDP-glucuronosyltransferase	8,302	4,456 (-0.38)	36,179	5.92	23.82	39	3.00	26.35	3,969 (-0.50)
Metopolophium dirhodum	13_abc_g2222.t1	ABC Transporter (ABC Subfamily G)	11,211	14,244 (0.32)	35,841	4.30	23.60	35	1.84	23.65	12,513 (0.06)
Sitobion miscanthi	1630_UGT344E8_v2	UDP-glucuronosyltransferase	8,302	3,249 (-0.50)	35,637	5.84	23.46	57	4.38	38.51	2,150 (-0.75)
Sitobion miscanthi	72_est_g6279.t1	Esterase	10,888	35,579 (1.82)	35,569	5.90	23.42	63	4.50	42.57	23,910 (1.03)
Greatest Enrichment of TE Count											
Acyrthosiphon pisum	**2399_UGT344E8_v2**	**UDP-glucuronosyltransferase**	**8,302**	**55,119 (4.63)**	**55,257**	**9.05**	**36.38**	**118**	**9.08**	**79.73**	**24,991 (2.45)**
Myzus persicae	**125_CYP380C52**	**Cytochrome P450**	**7,221**	**100,885 (9.70)**	**44,205**	**6.28**	**29.10**	**102**	**6.00**	**68.92**	**56,825 (5.97)**
Myzus persicae	**105_CYP380C19**	**Cytochrome P450**	**7,221**	**61,287 (5.60)**	**39,133**	**5.56**	**25.76**	**100**	**5.88**	**67.57**	**34,033 (3.17)**
Acyrthosiphon pisum	**2354_UGT344D11_v2**	**UDP-glucuronosyltransferase**	**8,302**	**57,143 (4.83)**	**45,208**	**7.40**	**29.76**	**96**	**7.38**	**64.86**	**26,620 (2.68)**

Myzus persicae	**109_CYP380C31**	**Cytochrome P450**	**7,221**	**99,455 (9.55)**	**42,790**	**6.08**	**28.17**	**94**	**5.53**	**63.51**	**57,937 (6.10)**
Acyrthosiphon pisum	2423_abc_g11808.t1	ABC Transporter (ABC Subfamily A)	11,211	25,566 (1.53)	25,867	3.11	17.03	84	4.42	56.76	19,631 (0.83)
Sitobion avenae	1033_UGT344A27	UDP-glucuronosyltransferase	8,302	62,310 (5.34)	33,066	5.41	21.77	82	6.31	55.41	45,645 (5.35)
Acyrthosiphon pisum	1088_CYP380C8	Cytochrome P450	7,221	10,900 (0.38)	28,620	4.07	18.84	82	4.82	55.41	5,975 (-0.27)
Sitobion miscanthi	911_UGT344A27	UDP-glucuronosyltransferase	8,302	64,596 (5.57)	27,402	4.49	18.04	81	6.23	54.73	38,974 (4.41)
Acyrthosiphon pisum	1_abc_g11808.t1	ABC Transporter (ABC Subfamily A)	11,211	39,350 (3.00)	31,743	3.81	20.90	81	4.26	54.73	25,849 (1.50)

Table 2. XGF loci with the greatest enrichment for TE coverage and TE count in aphids. Z-scores indicate the number of standard deviations away from mean gene size for individual genes. Z-scores <-3 or >3 indicate statistically unusually small or large genes compared to mean XGF size, respectively. Entries in bold are XGFs that are highly enriched for both TE coverage and TE count. ABC transporter subfamilies have been labelled where appropriate (Dermauw and Van Leeuwen 2014).

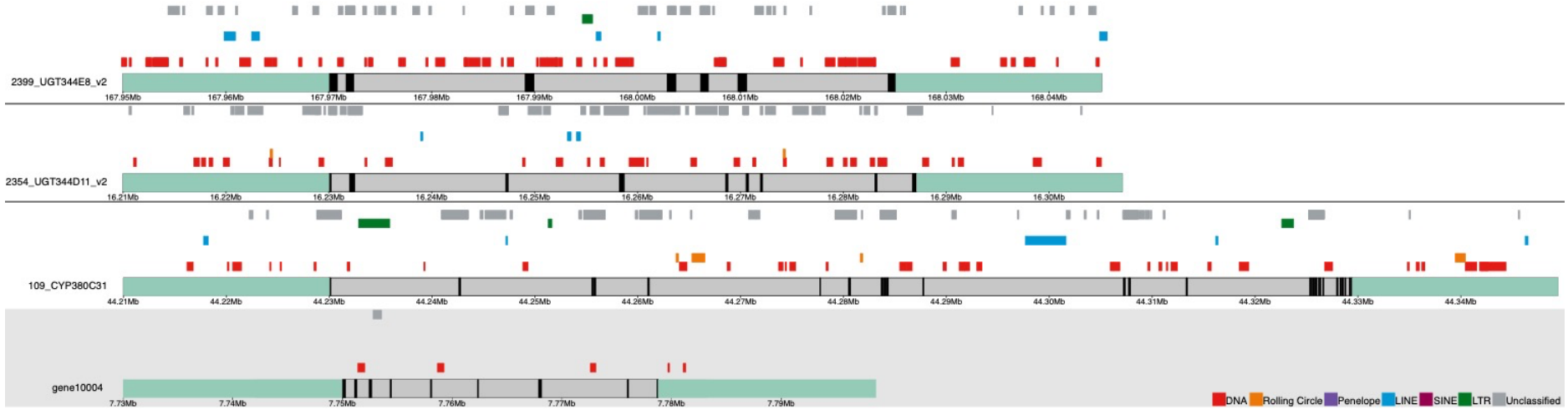


Figure 4. Karyoplots illustrating the most TE-dense XGF loci, corresponding to the XGFs highlighted in bold in Table 2. For comparison, the gene region represented in the shaded box (gene10004) shows a non-XGF from *M. persicae* with representative TE coverage close to the mean for non-XGFs (1,519bp). XGFs 125_CYP380C52 and 105_CYP380C19 are not shown, as these are nested at the CYP cluster containing 109_CYP380C31. Sea green shaded regions indicate gene flanking regions. Grey bars represent gene bodies, and black bars show exonic regions. TEs are annotated above their respective gene tracks, with colours indicating the main TE classifications, as depicted in the key at the bottom of the figure.

310 *Specific TE types are enriched at* XGFs

311 While most TE types contribute to enrichment at XGFs, overall enrichment is primarily
312 driven by increases in DNA TE content, due to the high frequency of DNA TEs in aphid
313 genomes, where they comprise almost half of total TE content (Table 3).

314 The TE type with the greatest degree of enrichment in coverage around XGFs in
315 comparison to other genes is SINEs, although there is a slight non-significant reduction in
316 SINE count at XGFs (**TE coverage**, 3.2x enrichment: XGFs μ = 1,144bp, all other genes μ =
317 358bp; Wilcoxon Rank Sum, $W_{9,482} = 439$, $p < 0.01$; **TE count,** 0.82x enrichment: XGFs μ =
318 1.00, all other genes μ = 1.22; Wilcoxon Rank Sum, $W_{9,482} = 2592$, $p = 0.14$). This is notable,
319 since SINEs are generally very scarce in aphid genomes, and suggests a particular retention
320 of SINE sequence at XGFs (Figure 1, Figure 3, Table 1, Supplemental Table S1). However,
321 given their scarcity, the relative contribution that SINEs make to TE enrichment at XGFs
322 compared to other TE types is extremely low. Conversely, DNA TEs are the most abundant
323 TE type in aphid genomes, and they are significantly enriched at XGFs both in TE coverage
324 and TE count (**TE coverage**, 2.0x enrichment: XGFs μ = 4,026bp, all other genes μ =
325 1,974bp; Wilcoxon Rank Sum, $W_{6156,336473} = 616,739,338$, $p < 0.01$; **TE count**, 6.4x
326 enrichment: XGFs μ = 10.05, all other genes μ = 1.57; Wilcoxon Rank Sum, $W_{6156,336473}$ =
327 1,267,581,920, $p < 0.01$) (Figure 3). Rolling circle TEs show significant, but lower levels of
328 enrichment (**TE coverage**, 1.1x enrichment: XGFs μ = 794bp, all other genes μ = 728bp;
329 Wilcoxon Rank Sum, $W_{1801,55720} = 46,195,509$, $p<0.01$; **TE count**, 1.8x enrichment: XGFs μ =
330 2.14, all other genes μ = 1.22; Wilcoxon Rank Sum, $W_{1801,55720} = 56,336,944$, $p < 0.01$). In
331 contrast, LINEs do not show significant levels of enrichment at XGFs in terms of TE
332 coverage, but there is a significant enrichment of LINE TE count (**TE coverage**, 1.03x
333 enrichment: XGFs μ = 992bp, all other genes μ = 959bp; Wilcoxon Rank Sum, $W_{2575,101686}$ =
334 128,128,748, $p = 0.06$; **TE count**, 1.3x enrichment: XGFs μ = 1.79, all other genes μ = 1.41;
335 Wilcoxon Rank Sum, $W_{2575,101686} = 167,286,786$, $p < 0.01$). Similarly, we find only a slight,

336 non-significant, enrichment of LTR TE coverage at XGFs, but we do identify a significant
337 enrichment of LTR TE count (**TE coverage**, 1.4x enrichment: XGFs μ = 1,817bp, all other
338 genes μ = 1,238bp; Wilcoxon Rank Sum, $W_{1004,51486}$ = 25,656,118, p = 0.69; **TE count**, 1.7x
339 enrichment: XGFs μ = 2.54, all other genes μ = 1.51; Wilcoxon Rank Sum, $W_{1004,51486}$ =
340 32,328,754, $p < 0.01$). Only PLEs show a depletion, albeit insignificant, in TE coverage
341 around XGFs, but they also show significant enrichment in count around XGFs (**TE**
342 **coverage**, 0.7x enrichment: XGF μ = 456bp, all other genes μ = 608bp; $W_{47,6823}$ = 185,654, p
343 = 0.06; **TE count**, 1.3x enrichment: XGF μ = 1.43, all other genes μ = 1.11; Wilcoxon Rank
344 Sum, $W_{47,6823}$ = 174,958, $p < 0.05$). However, like SINEs, PLEs account for a very small
345 proportion of total TE content (Supplemental Table S1).

TE Type	Non-XGF	All XGFs	CYPs	GSTs	Esterases	UGTs	ABCs
DNA	1,974 (2)	4,026(10)	4,445 (11)	4,800 (11)	3,513 (9)	3,743 (8)	4,609 (11)
Rolling Circle	728 (1)	794 (2)	836 (2)	849 (2)	1,126 (3)	706 (2)	756 (2)
Penelope	608 (1)	456 (1)	840 (2)	175 (1)	848 (2)	185 (1)	477 (1)
LINE	959 (1)	992 (2)	965 (2)	1,225 (2)	1,492 (2)	859 (2)	1,126 (2)
SINE	358 (1)	1,144 (1)	-	-	-	1,686 (1)	602 (1)
LTR	1,238 (2)	1,817 (3)	1,115 (1)	2,528 (3)	1,397 (2)	1,124(2)	3,826 (5)
Unclassified	1,307 (1)	2,255 (5)	2,114 (5)	2,290 (6)	1,754 (4)	2,508 (6)	2,384 (6)

346 Table 3. Mean base pair coverage of each TE classification at XGFs among aphids.
347 Numbers in brackets indicate mean TE copy number at each XGF for each TE classification.

348

349 *XGF enrichment varies among aphid genomes*

350 We find no significant difference in TE enrichment among different XGF types (i.e CYP
351 genes, esterases, UGTs, GSTs, ABCs) (Kruskal-Wallis, TE coverage: $\chi^2_4 = 7.72$, $p = 0.10$;
352 TE count: $\chi^2_4 = 4.25$, $p = 0.37$, Figure 3). However, we do find significant variation in the
353 magnitude of TE enrichment at XGFs among aphid species (linear regression, TE coverage:
354 $F_{19,78} = 37.49$, $p < 0.01$, adjusted $R^2 = 0.88$, TE count: $F_{19,78} = 28.28$, $p < 0.01$, adjusted $R^2 =$
355 0.84). Specifically, in terms of TE coverage, *C. cedri* shows the highest levels of enrichment
356 across all XGFs, with a mean enrichment of 15.9x compared to non-XGFs (Figure 5,
357 Supplemental Table S3). This is considerably higher than the lowest level of mean TE
358 coverage enrichment, which was found in *M. sacchari* with 1.9x TE coverage enrichment at
359 XGFs compared to non-XGFs (Figure 5, Supplemental Table S3). In terms of TE count, the
360 highest level of enrichment at XGFs compared to non-XGFs was also found in *C. cedri*, with
361 a mean TE count enrichment of 28.5x the number of TE insertions, whilst the lowest level
362 was found in *A. craccivora*, with a mean enrichment of 4.6x the number of TE insertions at
363 XGFs compared to non-XGFs (Figure 5, Supplemental Table S3)

364 Overall, whilst we observe enrichment in both TE coverage and TE count at XGFs
365 compared to non-XGFs in aphid genomes, fold-change increases are larger for TE count.
366 High enrichment for TE count combined with lower increases in TE coverage suggest that
367 enrichment is driven by the presence of large numbers of TE fragments at XGFs, as
368 opposed to full-length TEs (Figure 6). Indeed, mean TE length at XGFs across aphids is just
369 536bp, which is much shorter than most full length TEs, excluding SINEs (Wells and
370 Feschotte 2020; Wicker et al. 2007). Further, 90.40% of TEs across XGFs are below
371 1,000bp in length (Figure 6). Therefore, it is likely that a general pattern in TE accumulation
372 at XGFs is TE truncation with retention of specific TE domains, as previously observed for
373 certain individual cases of TE co-option during the evolution of host insecticide resistance
374 (Singh et al. 2020; Remnant et al. 2013; Panini et al. 2021; Bogwitz et al. 2005; Itokawa et

al. 2011; Schmidt et al. 2010; Darboux et al. 2007; Gahan et al. 2001; Yang et al. 2007; Zhao et al. 2010).

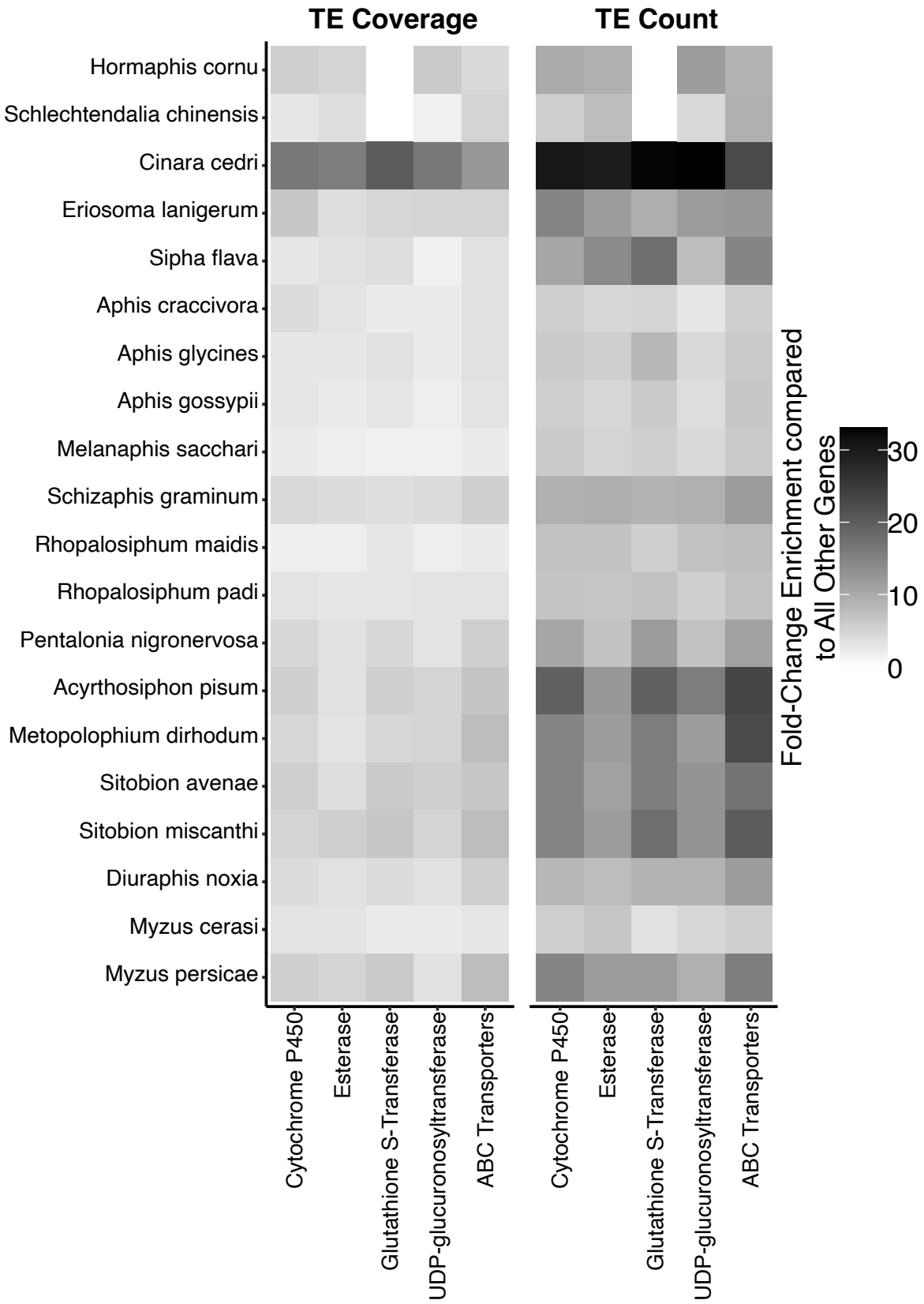
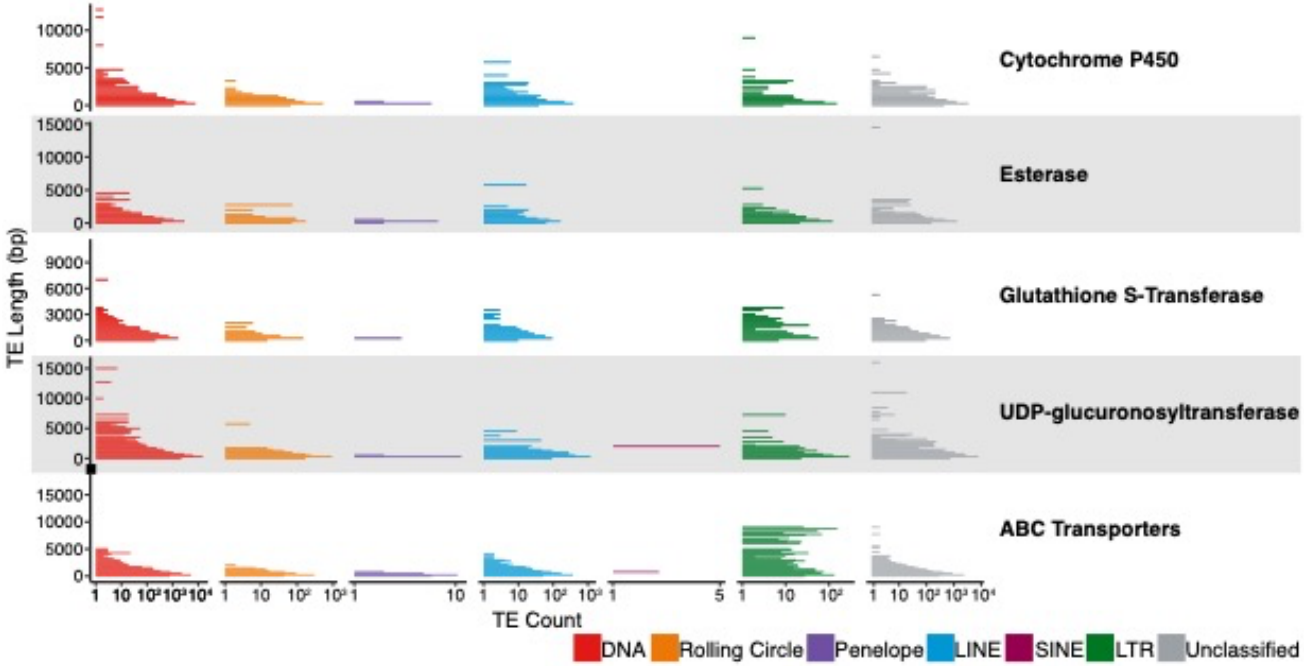


Figure 5. TE enrichment at XGFs compared to genome-wide levels. Fold-change enrichment for TE coverage and TE count at XGFs compared to within-genome levels at non-XGFs across 20 aphid species.

383



384

385 Figure 6. Fragmented TEs at XGFs across all aphid species, presented as TE count by TE length with a bin width of 250bp. TE count is
386 presented on a $\log_{10}$ scale. The main TE classifications are indicated by the colours indicated in the key. Mean TE length across all TEs at
387 XGFs is 536bp.

388 *TEs are enriched at XGFs associated with the detoxification of synthetic*
389 *insecticides in certain aphid species*

390 Cytochrome P450 genes (CYP genes) are a highly important gene family for the
391 detoxification of xenobiotics, including insecticides (Katsavou et al. 2022; Feyereisen 2005).
392 Indeed, increases in the expression of certain CYP genes can provide dramatic increases in
393 insecticide resistance (Nauen et al. 2022). Additionally, several cases of TE interactions with
394 CYP genes have been characterised (Chung et al. 2007; Chen and Li 2007; Daborn et al.
395 2002; Itokawa et al. 2011; Schmidt et al. 2010; Karunker et al. 2008; Nauen et al. 2022).
396 Therefore, we undertook a targeted analysis to examine evidence for the accumulation of
397 TEs at CYP genes with putative functions in xenobiotic resistance. For this, aphid CYP
398 genes were labelled based on the availability of evidence to support their xenobiotic
399 resistance functions from (Katsavou et al. 2022), as some CYP gene family members play
400 other roles, such as hormone breakdown. This resulted in the assignment of 2-29 CYP
401 genes per aphid genome to the xenobiotic detoxification category (Supplemental Table S5).
402 In *S. flava*, *C. cedri*, and *M. cerasi*, we find significant enrichment of TEs surrounding CYP
403 genes in CYP gene (sub)families frequently implicated in the breakdown of xenobiotics
404 compared to those with other functions (Figure 7B). However, we find the opposite pattern in
405 *S. graminum*, *M. dirhodum*, and *M. persicae*, where overall there is a significant depletion of
406 TEs surrounding CYP genes associated with xenobiotic resistance (Figure 7).

407 Next, we considered patterns of TE accumulation at CYP genes specifically involved
408 in resistance to synthetic insecticides, rather than resistance to xenobiotics more generally.
409 For this, CYP genes associated with xenobiotic resistance functions were labelled to reflect
410 the compounds to which they confer resistance. CYP genes that confer resistance to
411 insecticides were then considered separately from CYP genes associated with the
412 detoxification of other xenobiotics, such as those offering resistance to naturally occurring
413 plant allelochemicals. We find a significant enrichment of TEs around CYP genes associated

414 with resistance to synthetic insecticides in *P. nigronervosa*, *A. pisum*, *M. cerasi*, and *M.*
415 *persicae* (Figure 7). With the exception of *P. nigronervosa*, these are all species of "*most*
416 *agricultural importance*" (Van Emden and Harrington 2017), and are therefore likely to have
417 experienced particularly high exposure to synthetic insecticides, and extremely strong
418 selection to evolve resistance. The significant enrichment of TEs at XGFs associated with
419 resistance to synthetic insecticides in these species is thus consistent with their selective
420 recruitment for resistance evolution. However, no significant enrichment of TEs around CYP
421 genes associated with resistance to synthetic insecticides was found in 8 other species of
422 "*most agricultural importance*" included in our analysis: *A. craccivora, A. gossypii, D. noxia*,
423 *M. dirhodum, R. padi, S. graminum,* and *S. avenae* (Van Emden and Harrington 2017).

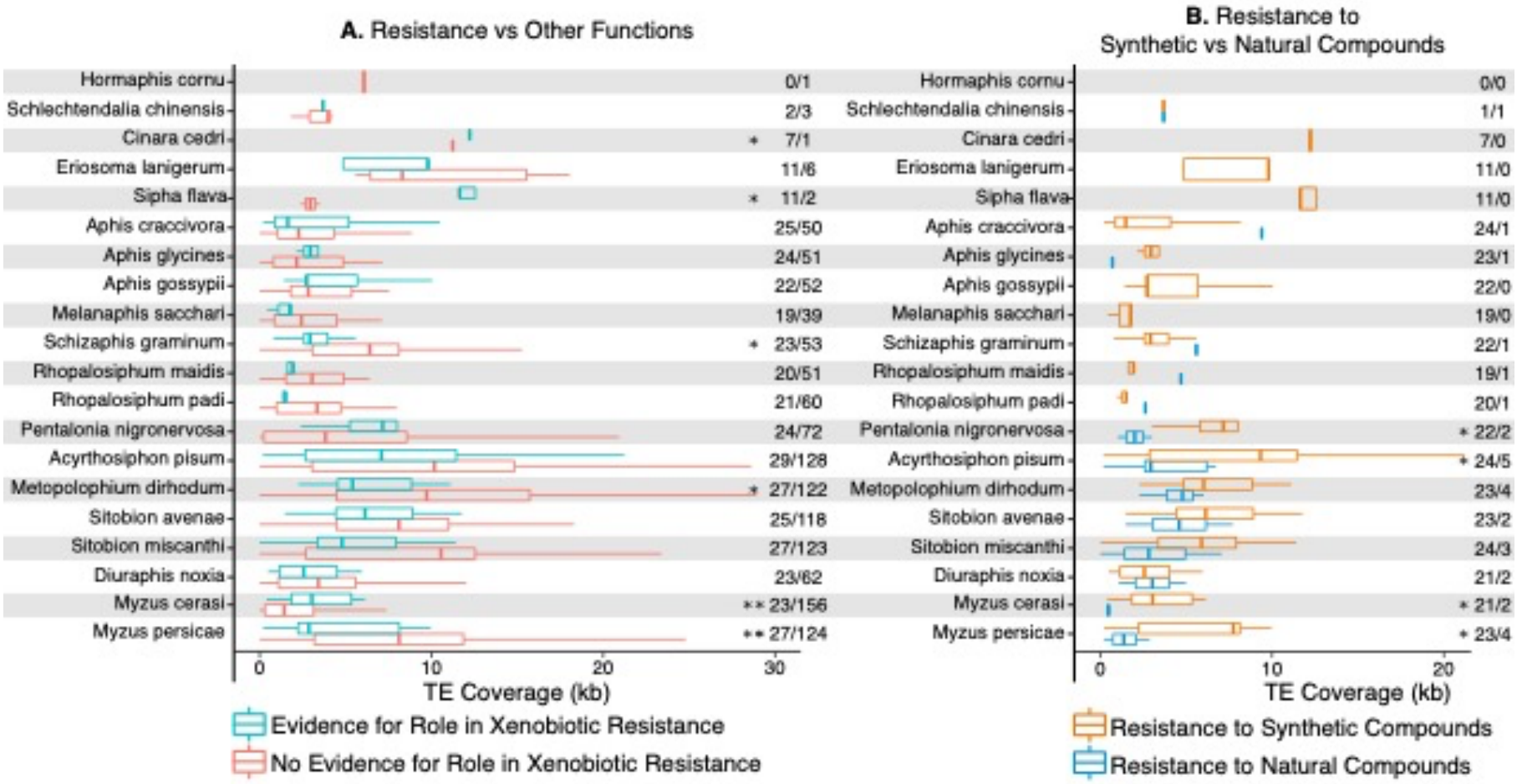


Figure 7. TE association with cytochrome P450 genes in aphids. **A**. TE coverage around CYP genes linked to xenobiotic resistance versus those with other functions. Numbers on the right indicate the number of genes in each group being compared. **B**.TE coverage around CYP genes linked to resistance to synthetic insecticides versus those conferring resistance to natural xenobiotics. Numbers on the right indicate the number of genes in each group under comparison.

429 *Evidence for the selective enrichment of TEs at XGFs*

430 We hypothesised that TE enrichment at resistance loci could arise through two
431 distinct processes: (i) Chromatin availability, whereby XGFs are surrounded by open
432 chromatin due to frequent transcriptional activity or constitutive expression, and so represent
433 locations that are more accessible for transposition; (ii) Selective retention of beneficial TE
434 insertions. Given that XGFs are essential for the detoxification of harmful plant metabolites
435 and insecticides, and so are of key importance for survival and reproduction, TE insertions
436 that modify XGFs in beneficial ways may be selectively retained and so spread through host
437 populations. Consequently, the combined effects of XGFs as selective hotspots (especially
438 considering the extremely strong selection mediated by intensive insecticide regimes), and
439 the demonstrated evolutionary potential of TEs, could result in the selective retention of TE
440 insertions surrounding XGFs over time. Such effects may be magnified by the influence of
441 host stress under insecticide exposure, which has been implicated in increases in TE activity
442 (Chénais et al. 2012; Stapley et al. 2015; Horváth et al. 2017). Specifically, increased TE
443 activity may result in a higher likelihood of TE insertions occurring at any genomic location,
444 including XGFs. However, TE insertions at XGFs may be more likely to be selectively
445 retained, if they offer host benefits in the form of resistance mutations, due to their capacity
446 to contribute to host evolvability through diverse genetic mechanisms (Bourque et al. 2018;
447 Rebollo et al. 2012).

448 Only germline TE activity, as opposed to somatic TE activity, is relevant from the
449 perspective of TE accumulation at XGFs, since only TE insertions that occur in the germline
450 have the potential to be transmitted to the next generation (novel TE insertions in the
451 genomes of somatic cells are an evolutionary dead-end). TE density at a gene, defined as
452 the proportion of its sequence occupied by TEs, is positively correlated with germline gene
453 expression, suggesting that TEs preferentially insert into regions that are actively transcribed
454 in the germline (McVicker and Green 2010). Therefore, a key question to distinguish

455 between alternative hypotheses that explain TE enrichment at XGFs is, to what extent are
456 XGFs expressed in the germline? We addressed this question using RNA-seq data from a
457 recent study on *A. pisum* (Jaquiéry et al. 2022) to examine gene expression in testes and
458 ovaries for XGFs and a large panel of housekeeping genes. Housekeeping genes are
459 constitutively expressed in all cell types and so represent a 'maximally accessible' case with
460 which to examine TE enrichment (Butte et al. 2001) (Supplemental Table S4). We find that
461 germline housekeeping gene expression is significantly higher compared to XGFs
462 considering mean RPKM (reads per kilobase of transcript per million base pairs sequenced)
463 in *A. pisum* (Wilcoxon Rank Sum, $W_{612,112} = 52252$, $p < 0.01$, Figure 8). The observed
464 pattern of expression of housekeeping genes and XGFs is also present in *Drosophila*
465 *melanogaster*, suggesting conservation across insect diversity (Supplemental Figure S3).
466 Consistent with chromatin availability as an explanation for TE enrichment, we find that TEs
467 are significantly enriched at housekeeping genes compared to all other host genes,
468 excluding XGFs (TE coverage: Wilcoxon Rank Sum, $W_{546,475492} = 218,813,232$, $p < 0.01$; TE
469 count: Wilcoxon Rank Sum, $W_{546,475492} = 170,683,160$, $p < 0.01$), to a similar level as that
470 observed for XGFs versus all other host genes (Figure 8, Supplemental Figures S4 and S5).
471 However, given that germline expression of XGFs is consistently low across all XGF types
472 (Figure 8), we conclude that germline expression cannot explain the observed enrichment of
473 TEs at XGFs. Instead, our findings support the alternative explanation for the enrichment of
474 TEs at XGFs, which is selective retention of TE sequences due to genetic contributions
475 towards the evolution of xenobiotic resistance mutations. There is no mechanistic basis to
476 suggest that TEs specifically target the regions surrounding XGFs as insertion sites,
477 compared to other genomic locations. Instead, under this model TE insertions at XGFs are
478 more likely to be retained and spread through the host population (compared to TE
479 insertions at other loci), due to the combination of their capacity to contribute to host
480 evolvability, and the extremely strong selection pressure imposed by insecticide treatment.

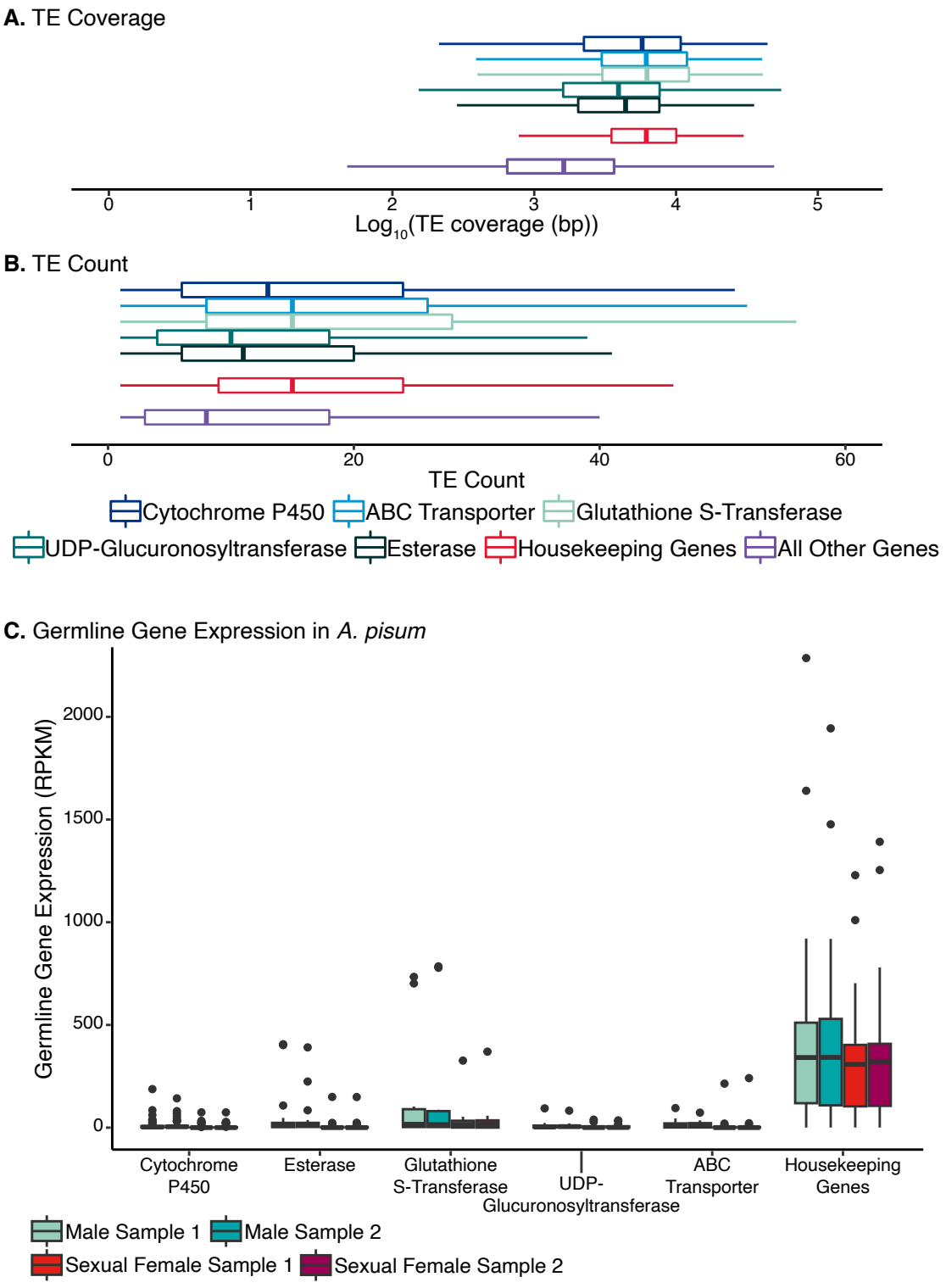


Figure 8. TE occupancy at XGFs, housekeeping genes, and all other genes, and expression of XGFs and housekeeping genes in *A. pisum* (Jaquiéry et al. 2022). **A**. TE coverage around XGFs, housekeeping genes, and all other genes. Box limits indicate upper and lower quartiles, whilst central lines indicate mean TE coverage. **B**. TE count around XGFs, housekeeping genes, and all other genes. Box limits indicate upper and lower quartiles, whilst central lines indicate mean TE count. **C**. Expression of XGFs and housekeeping genes in germline tissues (testis and ovary) in *A. pisum*, expressed as reads per kilobase of transcript per million base pairs sequenced (RPKM), to determine expected expression of XGFs in germline cell types. Gene types are indicated by colours in the key.

491 ## Discussion

492 Resistance evolution remains a major societal challenge that greatly affects efforts to control
493 harmful pests and pathogens (Van Emden and Harrington 2017). To develop new
494 approaches to overcome the challenge of resistance evolution, a deeper understanding of
495 the genomic mechanisms leading to the repeated emergence of resistant phenotypes is
496 required. Whilst isolated examples of TE involvement in the evolution of xenobiotic
497 resistance exist, the extent to which TEs act as a source of genomic novelty for resistance
498 evolution, and host evolution more generally, remains an open question.

499 We examine whether genes implicated in xenobiotic resistance are enriched for TE
500 insertions, potentially due to the generation of novel advantageous mutations for resistance
501 evolution. We find an enrichment of TEs, both in terms of TE sequence coverage and
502 number of TE insertions, at XGFs across all aphids and all XGF types. Further, we provide
503 evidence that this enrichment is the consequence of selection at XGFs, as opposed to
504 elevated insertion rates due to chromatin availability at XGFs in the germline.

505 Our results also uncover considerable variability in the enrichment of TEs around
506 XGFs both within and among aphid species and XGF types. This demonstrates that TEs are
507 not globally enriched at every XGF, but instead are enriched only at certain XGFs and
508 depleted at others. Consequently, the patterns we observe are driven by hotspots of TE
509 enrichment around certain XGF types, and in some cases individual XGFs, which show
510 extremely high levels of enrichment compared to other XGFs. For example, considering all
511 aphid species examined, we identify significant enrichment of TEs around GSTs. However,
512 this result is driven by a subset of aphids, with 11 of the 20 species showing no evidence of
513 significant enrichment around this XGF class (Supplemental Figure S4). This is consistent
514 with varying levels of selection acting on particular detoxification genes and particular aphid
515 species. Meanwhile, we find markedly more TE hotspots at UGTs and CYP genes, whereby
516 certain XGFs are massively enriched for TE fragments. These observations demonstrate the
517 nuances involved in patterns of TE accumulation at XGFs and underline a need for

518 functional validation analyses to test the impact of individual TE insertions on XGF
519 expression, and whether highly TE-enriched loci are indeed involved in xenobiotic
520 resistance.

521 Important crop pests are likely to experience much higher levels of insecticide stress
522 than non-target species, demonstrated by the increased number and diversity of insecticide
523 resistance mutations in target species, compared to non-target species (Mota-Sanchez and
524 Wise 2023). Consequently, variability in TE enrichment is expected among aphid species
525 and XGFs, depending on variation in the degree of insecticide exposure. This prediction is
526 confirmed when considering CYP genes, for which data clarifying the relationship to specific
527 insecticides is available (rather than just to xenobiotics in general). Specifically, we find that
528 TE enrichment is significantly higher around CYP genes that are implicated in the
529 detoxification of synthetic insecticides, compared to those involved in the detoxification of
530 natural xenobiotics, but only in three out of ten aphids of "*major agricultural importance*" (Van
531 Emden and Harrington 2017) included in our study (*A. pisum, M. cerasi*, and *M. persicae*).
532 Consequently, either CYP genes are less important for survival in the other species (perhaps
533 due to differences in pest control strategies, such as a greater reliance on forms of non-
534 insecticide based control, such as biocontrol strategies), or there are genomic mechanisms
535 that limit the co-option of TEs for resistance evolution in these species.

536 Demonstrated contributions of TEs to xenobiotic resistance often involve small
537 fragments of the original TE insertions, and contain motifs that act as host regulatory
538 elements (Singh et al. 2020; Remnant et al. 2013; Panini et al. 2021; Bogwitz et al. 2005;
539 Itokawa et al. 2011; Schmidt et al. 2010; Darboux et al. 2007; Gahan et al. 2001; Yang et al.
540 2007). Meanwhile, the emergence of a resistant phenotype frequently appears to result from
541 a single, or up to a few, TE insertions at each specific XGF locus, rather than large
542 accumulations of TE DNA (Schmidt et al. 2010; Daborn et al. 2002; Chung et al. 2007;
543 Panini et al. 2021). This presumably occurs due to genomic and selective processes that act
544 to retain the core TE regions that convey a selective benefit, while purging long repetitive

545 and potentially deleterious regions (e.g. that promote non-homologous recombination with
546 similar repetitive elements). Meanwhile, the impacts of these TE insertions can occur via
547 complex multistep processes including the retention of several truncated TE fragments (such
548 as for CYP6G1, Schmidt *et al*. (2010)). Thus, a key consideration is that TE-associated
549 resistance mechanisms go beyond simple enrichment at specific loci.

550 We find that DNA TEs are highly enriched at XGFs, despite the more widely
551 acknowledged role of TEs with more complex cis-regulatory elements, such as LTR
552 retrotransposons, in altering host gene expression (Chung et al. 2007). This may suggest
553 that other processes beyond TE-mediated increases in expression are involved in the
554 patterns of enrichment we uncover. Although, it is entirely possible that the DNA TEs
555 involved also contain sequences relevant for the recruitment of transcription factors that
556 increase expression. Meanwhile, at present we cannot currently explain the repeated cases
557 whereby certain XGFs are massively enriched in terms of TE coverage and TE count
558 compared to other XGFs. Thus, unravelling the influence of TEs at these loci represents a
559 key target for future research.

560 Gaining a deeper understanding of the mechanisms that drive patterns of association
561 between TEs and XGFs will require contributions from complementary approaches to the
562 broadscale comparative genomic analyses performed here. Of particular relevance are
563 functional validation studies to examine the influence of individual TE insertions on nearby
564 XGF loci, which our results provide considerable scope to address. Further analyses are
565 also required to elucidate the underlying mechanisms by which TE sequences may
566 contribute to resistance phenotypes, such as screening to investigate if implicated TE
567 sequences contain transcription factor binding sites, as well as information on chromatin
568 conformation and the availability of DNA within TE sequences at XGFs. Meanwhile, detailed
569 intra-species analyses can provide valuable evolutionary insights into individual TE
570 insertions of interest (e.g. (Aminetzach et al. 2005)), and information on TE occupancy within
571 populations (e.g. (Barrón et al. 2014)), particularly from strains displaying different resistance

572 phenotypes using high-quality long-read datasets and individual TE annotations.
573 Collectively, such approaches will ultimately allow evaluation of the extent to which TEs are
574 utilised in resistance mutations, versus the contributions arising from other classes of genetic
575 variation such as SNPs and CNVs.

576

577 **Materials and Methods**

578 *Estimating Host Phylogeny*

579 The genome assemblies of 20 aphid species, and 7 diverse hemipteran outgroup species
580 (*Ericerus pela*, *Maconellicoccus hirsutus*, *Phenacoccus solenopsis, Bemisa tabaci, Sogatella*
581 *furcifera, Diaphorina citri, Pachypsylla venusta*), were obtained along with gene annotations
582 where available (Supplemental Table S6). To estimate host phylogeny, a supermatrix
583 approach was used. Specifically, for each genome assembly, BUSCO (version 5.2.2) (Manni
584 et al. 2021a, 2021b) was used with the hemiptera_odb10 geneset to identify conserved gene
585 orthologs in each genome assembly. The identifiers for all complete genes were extracted
586 and those present in less than three genomes were removed. The amino acid sequences for
587 each gene from all species were extracted and saved in individual fasta files. For each gene,
588 sequences were aligned using MAFFT (version 7.453) (Katoh and Standley 2013) (--auto).
589 The subsequent alignments were concatenated using Phykit create_concat (Steenwyk et al.
590 2021) to generate a supermatrix. To generate the species phylogeny, we performed 1,000
591 ultrafast bootstrap repetitions (-bb 1000) in IQ-tree (version 1.6.12)(Nguyen et al. 2015),
592 using the best fit amino acid model identified by Modelfinder (Kalyaanamoorthy et al. 2017)
593 (Supplemental File S1).

594

595 *Transposable Element Annotation*

596 To curate and annotate TEs, each genome assembly was annotated with the Earl Grey TE
597 annotation pipeline (version 1.3) (https://github.com/TobyBaril/EarlGrey) (Baril et al. 2022).
598 Briefly, known TEs from Arthropoda (-r arthropoda) were first annotated using both the Dfam
599 3.4 (Hubley et al. 2016) and RepBase RepeatMasker Edition (version 20181026) (Jurka et
600 al. 2005; Kapitonov and Jurka 2008) TE databases. Following this, Earl Grey identified *de*
601 *novo* TE families and refined these using an iterative 'BLAST, Extract, Extend' process (Platt
602 et al. 2016). Following final annotation with the combined library of known and *de novo* TE
603 families, annotations are processed to remove overlapping annotations and to defragment
604 annotations likely to originate from the same TE insertion. TE annotation GFF files are
605 provided in Supplemental File S2.

606 To characterise TEs as either shared or species-specific, *de novo* TE libraries from
607 each genome assembly were clustered to the TE family definition of (Wicker et al. 2007),
608 implemented as described by (Goubert et al. 2022) using cd-hit-est (-d 0 -aS 0.8 -c 0.8 -G 0 -
609 g 1 -b 500 -r 1) (Li and Godzik 2006; Fu et al. 2012). Sequences within each cluster were
610 designated a number between 0 and 20 to determine the number of other species that the
611 TE family was shared with, with 0 identifying TEs unique to a species, and 19 identifying TEs
612 shared across all 20 aphid species considered in this study. For known TE families from
613 Dfam and RepBase, each TE was labelled using the same system based on the number of
614 species the TE family was identified in.

615 Genetic distance from TE consensus sequence is used as a proxy for estimated
616 timing of TE activity (i.e TE age). Using genetic distance from TE consensus as a proxy for
617 insertion time should be taken with caution, as this can be influenced by the methodology
618 used to generate TE consensus sequences, and whether known TE sequences from other
619 organisms are used as a reference or purely *de novo* TE consensus sequences from the

620 species of interest. However, this metric can still be used to provide a rough relative estimate
621 of TE activity.

622

623 *Phylogenetic Heritability of TE Abundance and Diversity*

624 To determine the phylogenetic heritability of TE abundance, expressed as total TE count,
625 and TE diversity, expressed as total number of distinct TE families, we measured the
626 amount of variation explained by phylogenetic relationships using Bayesian phylogenetic
627 mixed models with Markov chain Monte Carlo estimations, with an intercept fitted as a fixed
628 effect and the phylogeny fitted as a random effect, in the MCMCglmm package (Hadfield
629 2010) in R (version 4.2.1) (Team 2013) using the Rstudio IDE (Team 2015; Racine 2013). A
630 poisson error distribution and log link function were used, and inverse gamma priors were
631 specified for all R and G-side random effects (V = 1, v = 0.002). Models were run for
632 11,000,000 iterations with a burn-in of 1,000,000 and a thinning interval of 1,000. This
633 approach generated 10,000 posterior samples from which posterior mode and 95% CIs were
634 calculated. The proportion of between-species variation explained by phylogeny was
635 calculated from the model using the equation $V_p / (V_p + V_s)$ where V_p and V_s represent the
636 phylogenetic and species-specific components of between-species variance (Freckleton et
637 al. 2002), which is equivalent to phylogenetic heritability.

638

639 *Gene Predictions and Transposable Element Association with Genomic*
640 *Compartments*

641 For genome assemblies lacking a gene annotation GFF file at the time of analysis (*A.*
642 *gossypii*, *H. cornu, M. dirhodum, P. nigronervosa, R. padi, S. graminum, S. chinensis, S.*

643 *avenae, S. miscanthi*), genes were annotated using Augustus (version 3.3.3) (Stanke et al.
644 2008; Keller et al. 2011) using the pea aphid reference gene set (--species=pea_aphid --
645 strand=both --genemodel=partial --introns=on --exonnames=on --gff3=on).

646 To prepare genome annotation files for intersection with TE loci, each gene
647 annotation GFF was processed to generate a GFF file containing coordinates of exons,
648 introns, 5' and 3' flanking regions, and intergenic space. Gene flanking regions are defined
649 here as 20 kb directly upstream or downstream of the gene body. These flanking regions are
650 determined to identify TEs at a distance that could be described as the proximate promoter
651 region, rather than just accounting for the core promoter region, to include both promoter
652 and more distal enhancer regions for genes (Lenhard et al. 2012). Intergenic space
653 coordinates were generated using bedtools complement (Quinlan and Hall 2010), to
654 generate inverse coordinates to those of genic compartments and flanking regions. To
655 quantify TE occupation in different genomic compartments (introns, exons, 5' and 3' flanking
656 regions, and intergenic space), bedtools intersect was used to calculate overlap (-wao)
657 between all compartments and TEs.

658

659 *Identification of Xenobiotic Gene Family Loci and Housekeeping Genes*

660 The gene identities for all ABC transporters, cytochrome P450s, UDP-
661 glucuronosyltransferases, esterases, and glutathione S-transferases were manually curated
662 from the functional annotation of the *M. persicae* G006v2 genome (Singh et al. 2021). Using
663 these identifiers, exon coordinates were obtained from the *M. persicae* gene annotation
664 GFF3 file, and corresponding nucleotide sequences were extracted from the genome
665 assembly. The resulting exon sequences from *M. persicae* were used as queries to identify
666 XGFs in all other aphid assemblies using blastn (Camacho et al. 2009) with an e-value
667 threshold of $1x10^{-10}$. Blast hits for each species were manually inspected to extract those

that best represented each XGF, with the inclusion of gene duplications where appropriate. Following this, a bed file of XGF coordinates for each species was generated including the annotation of flanking regions as described above (Supplemental File S3).

Housekeeping genes (HKGs) are required for basic cellular functions and are typically expressed in all cell types (Butte et al. 2001). Amino acid sequences of 28 HKGs characterised for aphids (Bansal et al. 2012; Koramutla et al. 2016; Li et al. 2020; Kang et al. 2017; Yang et al. 2015, 2014) were obtained from GenBank (Benson et al. 2013) for *M. persicae* (Supplemental File S4). Subsequently, a blastx search was performed, with default parameters, against each aphid genome to identify HKG loci in the other 19 species. For each species, HKGs were manually curated and a bed file of HKG coordinates, including the annotation of flanking regions, was generated (Supplemental File S5). To quantify TE occupation around XGFs and HKGs, bedtools intersect was used to calculate overlap (-wao) between annotated TEs and XGFs.

Competing Interests

The authors declare no competing interests.

ACKNOWLEDGEMENTS AND FUNDING

TB was supported by a studentship from the Biotechnology and Biological Sciences Research Council-funded South West Biosciences Doctoral Training Partnership (grant number BB/M009122/1).

AP was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement number 646625).

CB was supported by a Biotechnology and Biological Sciences Research Council (BBSRC) grant (grant number BB/S006060/1) and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement number 646625)

AH was supported by a Biotechnology and Biological Sciences Research Council (BBSRC) David Phillips Fellowship (grant number BB/N020146/1).

Myzus cerasi DNA sequencing data were downloaded from AphidBase. Funding for the sequencing was provided by ERC Starting Grant APHIDHOST-310190 awarded to Jorunn Bos at the James Hutton Institute, United Kingdom.

AUTHOR CONTRIBUTIONS

TB performed TE annotation and downstream analyses, produced the figures, and drafted the manuscript. AP manually curated xenobiotic resistance genes from *M. persicae.* CB and AH contributed to the analytical design and data analysis and participated in writing the manuscript. AH conceived and coordinated the study. All authors read and approved the final manuscript.

REFERENCES

Ahn S-J, Vogel H, Heckel DG. 2012. Comparative analysis of the UDP-glycosyltransferase multigene family in insects. *Insect Biochem Mol Biol* **42**: 133–147.

Aminetzach YT, Macpherson JM, Petrov DA. 2005. Pesticide resistance via transposition-mediated adaptive gene truncation in Drosophila. *Science* **309**: 764–767.

Arkhipova IR. 2018. Neutral theory, transposable elements, and eukaryotic genome evolution. *Mol Biol Evol* **35**: 1332–1337.

Bansal R, Mamidala P, Mian MAR, Mittapalli O, Michel AP. 2012. Validation of reference genes for gene expression studies in Aphis glycines (Hemiptera: Aphididae). *J Econ Entomol* **105**: 1432–1438.

Baril T, Hayward A. 2022. Migrators within migrators: exploring transposable element dynamics in the monarch butterfly, Danaus plexippus. *Mob DNA* **13**: 5.

Baril T, Imrie RM, Hayward A. 2022. Earl Grey: a fully automated user-friendly transposable element annotation and analysis pipeline. https://www.researchsquare.com/article/rs-1812599/latest.pdf.

Barrón MG, Fiston-Lavier A-S, Petrov DA, González J. 2014. Population genomics of transposable elements in Drosophila. *Annu Rev Genet* **48**: 561–581.

Bass C, Puinean AM, Nauen R, Gutbrod O, Denholm I, Williamson MS, Slater R, Field LM, Zimmer CT, Foster SP. 2014. The evolution of insecticide resistance in the peach potato aphid, Myzus persicae. *Insect Biochem Mol Biol* **51**: 41–51.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2013. GenBank. *Nucleic Acids Res* **41**: D36-42.

Blass E, Bell M, Boissinot S. 2012. Accumulation and rapid decay of non-LTR retrotransposons in the genome of the three-spine stickleback. *Genome Biol Evol* **4**: 687–702.

Bogwitz MR, Chung H, Magoc L, Rigby S, Wong W, O'Keefe M, McKenzie JA, Batterham P, Daborn PJ. 2005. Cyp12a4 confers lufenuron resistance in a natural population of Drosophila melanogaster. *Proc Natl Acad Sci U S A* **102**: 12807–12812.

Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, Imbeault M, Izsvák Z, Levin HL, Macfarlan TS, et al. 2018. Ten things you should know about transposable elements. *Genome Biol* **19**: 199.

Bradshaw CJA, Leroy B, Bellard C, Roiz D, Albert C, Fournier A, Barbet-Massin M, Salles J-M, Simard F, Courchamp F. 2016. Massive yet grossly underestimated global costs of invasive insects. *Nat Commun* **7**: 12986.

Butte AJ, Dzau VJ, Glueck SB. 2001. Further defining housekeeping, or "maintenance," genes Focus on "A compendium of gene expression in normal human tissues." *Physiol Genomics* **7**: 95–96.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: Architecture and applications. *BMC Bioinformatics* **10**: 1–9.

Catania F, Kauer MO, Daborn PJ, Yen JL, Ffrench-Constant RH, Schlötterer C. 2004.

759 World-wide survey of an Accord insertion and its association with DDT resistance in
760 Drosophila melanogaster. *Mol Ecol* **13**: 2491–2504.

761 Chen S, Li X. 2007. Transposable elements are enriched within or in close proximity to
762 xenobiotic-metabolizing cytochrome P450 genes. *BMC Evol Biol* **7**: 46.

763 Chénais B, Caruso A, Hiard S, Casse N. 2012. The impact of transposable elements on
764 eukaryotic genomes: From genome size increase to genetic adaptation to stressful
765 environments. *Gene* **509**: 7–15.

766 Chung H, Bogwitz MR, McCart C, Andrianopoulos A, Ffrench-Constant RH, Batterham P,
767 Daborn PJ. 2007. Cis-regulatory elements in the accord retrotransposon result in
768 tissue-specific expression of the Drosophila melanogaster insecticide resistance
769 gene Cyp6g1. *Genetics* **175**: 1071–1077.

770 Cosby RL, Chang N-C, Feschotte C. 2019. Host–transposon interactions: conflict,
771 cooperation, and cooption. *Genes Dev* **33**: 1098–1116.

772 Daborn PJ, Yen JL, Bogwitz MR, Le Goff G, Feil E, Jeffers S, Tijet N, Perry T, Heckel D,
773 Batterham P, et al. 2002. A single p450 allele associated with insecticide resistance
774 in Drosophila. *Science* **297**: 2253–2256.

775 Darboux I, Charles J-F, Pauchet Y, Warot S, Pauron D. 2007. Transposon-mediated
776 resistance to Bacillus sphaericus in a field-evolved population of Culex pipiens
777 (Diptera: Culicidae). *Cell Microbiol* **9**: 2022–2029.

778 de Koning APJ, Gu W, Castoe TA, Batzer MA, Pollock DD. 2011. Repetitive elements may
779 comprise over Two-Thirds of the human genome. *PLoS Genet* **7**.
780 http://dx.doi.org/10.1371/journal.pgen.1002384.

781 de Little SC, Edwards O, van Rooyen AR, Weeks A, Umina PA. 2017. Discovery of
782 metabolic resistance to neonicotinoids in green peach aphids (Myzus persicae) in
783 Australia. *Pest Manag Sci* **73**: 1611–1617.

784 Dermauw W, Van Leeuwen T. 2014. The ABC gene family in arthropods: Comparative
785 genomics and role in insecticide transport and resistance. *Insect Biochem Mol Biol*
786 **45**: 89–110.

787 Deutsch CA, Tewksbury JJ, Tigchelaar M, Battisti DS, Merrill SC, Huey RB, Naylor RL.
788 2018. Increase in crop losses to insect pests in a warming climate. *Science* **361**:
789 916–919.

790 Fabrick JA, Mathew LG, Tabashnik BE, Li X. 2011. Insertion of an intact CR1
791 retrotransposon in a cadherin gene linked with Bt resistance in the pink bollworm,
792 Pectinophora gossypiella. *Insect Molecular Biology* **20**: 651–665.
793 http://dx.doi.org/10.1111/j.1365-2583.2011.01095.x.

794 Feyereisen R. 2005. Insect cytochrome P450. *Comprehensive molecular insect science*.

795 Feyereisen R. 1999. Insect P450 enzymes. *Annu Rev Entomol* **44**: 507–533.

796 Freckleton RP, Harvey PH, Pagel M. 2002. Phylogenetic analysis and comparative data: a
797 test and review of evidence. *Am Nat* **160**: 712–726.

798 Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-
799 generation sequencing data. *Bioinformatics* **28**: 3150–3152.

800 Gahan LJ, Gould F, Heckel DG. 2001. Identification of a gene associated with Bt resistance
801 in Heliothis virescens. *Science* **293**: 857–860.

802 Gilbert C, Peccoud J, Cordaux R. 2021. Transposable Elements and the Evolution of
803 Insects. *Annu Rev Entomol* **66**: 355–372.

804 Gilbert LI, Gill SS, eds. 2014. *Insect pharmacology*. Academic Press.

805 Gott RC, Kunkel GR, Zobel ES, Lovett BR, Hawthorne DJ. 2017. Implicating ABC
806 Transporters in Insecticide Resistance: Research Strategies and a Decision
807 Framework. *J Econ Entomol* **110**: 667–677.

808 Goubert C, Craig RJ, Bilat AF, Peona V, Vogan AA, Protasio AV. 2022. Correction: A
809 beginner's guide to manual curation of transposable elements. *Mob DNA* **13**: 15.

810 Gregory TR. 2005. Synergy between sequence and size in large-scale genomics. *Nat Rev*
811 *Genet* **6**: 699–708.

812 Guengerich FP. 2009. Cataloging the Repertoire of Nature's Blowtorch, P450. *Chem Biol* **16**:
813 1215–1216.

814 Hadfield JD. 2010. MCMC Methods for Multi-Response Generalized Linear Mixed Models:
815 The MCMCglmm R Package. *J Stat Softw* **33**: 1–22.

816 Horváth V, Merenciano M, González J. 2017. Revisiting the Relationship between
817 Transposable Elements and the Eukaryotic Stress Response. *Trends Genet* **33**:
818 832–841.

819 Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, Smit AFA, Wheeler TJ. 2016.
820 The Dfam database of repetitive DNA families. *Nucleic Acids Res* **44**: D81–D89.

821 Itokawa K, Komagata O, Kasai S, Masada M, Tomita T. 2011. Cis-acting mutation and
822 duplication: History of molecular evolution in a P450 haplotype responsible for
823 insecticide resistance in Culex quinquefasciatus. *Insect Biochem Mol Biol* **41**: 503–
824 512.

825 Jaquiéry J, Simon J-C, Robin S, Richard G, Peccoud J, Boulain H, Legeai F, Tanguy S,
826 Prunier-Leterme N, Letrionnaire G. 2022. Masculinization of the X-chromosome in
827 aphid soma and gonads. *Peer Community J* **2**.
828 http://dx.doi.org/10.24072/pcjournal.166.

829 Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase
830 Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110**:
831 462–467.

832 Kalyaanamoorthy S, Minh BQ, Wong TKF, Haeseler A von, Jermiin LS. 2017. ModelFinder:
833 fast model selection for accurate phylogenetic estimates. *Nature Methods 2017 14:6*
834 **14**: 587–589.

835 Kang Z-W, Liu F-H, Tian H-G, Zhang M, Guo S-S, Liu T-X. 2017. Evaluation of the reference
836 genes for expression analysis using quantitative real-time polymerase chain reaction
837 in the green peach aphid, Myzus persicae. *Insect Sci* **24**: 222–234.

838 Kapitonov VV, Jurka J. 2008. A universal classification of eukaryotic transposable elements
839 implemented in Repbase. *Nat Rev Genet* **9**: 411–412.

840 Karunker I, Benting J, Lueke B, Ponge T, Nauen R, Roditakis E, Vontas J, Gorman K,

841 Denholm I, Morin S. 2008. Over-expression of cytochrome P450 CYP6CM1 is
842 associated with high resistance to imidacloprid in the B and Q biotypes of Bemisia
843 tabaci (Hemiptera: Aleyrodidae). *Insect Biochem Mol Biol* **38**: 634–644.

844 Kasai S, Scott JG. 2001. Cytochrome P450s CYP6D3 and CYP6D1 are part of a P450 gene
845 cluster on autosome 1 in the house fly. *Insect Mol Biol* **10**: 191–196.

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

848 Katsavou E, Riga M, Ioannidis P, King R, Zimmer CT, Vontas J. 2022. Functionally
849 characterized arthropod pest and pollinator cytochrome P450s associated with
850 xenobiotic metabolism. *Pesticide Biochemistry and Physiology* **181**: 105005.
851 http://dx.doi.org/10.1016/j.pestbp.2021.105005.

852 Keller O, Kollmar M, Stanke M, Waack S. 2011. A novel hybrid gene prediction method
853 employing protein multiple sequence alignments. *Bioinformatics* **27**: 757–763.

854 Kennedy CJ, Tierney KB. 2013. Xenobiotic Protection/Resistance Mechanisms in
855 Organisms. In *Environmental Toxicology: Selected Entries from the Encyclopedia of*
856 *Sustainability Science and Technology* (ed. E.A. Laws), pp. 689–721, Springer New
857 York, New York, NY.

858 Ketterman AJ, Saisawang C, Wongsantichon J. 2011. Insect glutathione transferases. *Drug*
859 *Metab Rev* **43**: 253–265.

860 Kidwell MG. 2002. Transposable elements and the evolution of genome size in eukaryotes.
861 *Genetica* **115**: 49–63.

862 Kliot A, Ghanim M. 2012. Fitness costs associated with insecticide resistance. *Pest Manag*
863 *Sci* **68**: 1431–1437.

864 Koramutla MK, Aminedi R, Bhattacharya R. 2016. Comprehensive evaluation of candidate
865 reference genes for qRT-PCR studies of gene expression in mustard aphid, Lipaphis
866 erysimi (Kalt). *Sci Rep* **6**: 25883.

867 Krause SA, Overend G, Dow JAT, Leader DP. 2022. FlyAtlas 2 in 2022: enhancements to
868 the Drosophila melanogaster expression atlas. *Nucleic Acids Res* **50**: D1010–D1015.

869 Lavoie CA, Ii RNP, Novick PA, Counterman BA, Ray DA. 2013. Transposable element
870 evolution in Heliconius suggests genome diversity within Lepidoptera. *Mob DNA* **21**:
871 1–10.

872 Lenhard B, Sandelin A, Carninci P. 2012. Metazoan promoters: emerging characteristics and
873 insights into transcriptional regulation. *Nat Rev Genet* **13**: 233–245.

874 Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of
875 protein or nucleotide sequences. *Bioinformatics* **22**: 1658–1659.

876 Li X, Gong P, Wang B, Wang C, Li M, Zhang Y, Li X, Gao H, Ju J, Zhu X. 2020. Selection
877 and validation of experimental condition-specific reference genes for qRT-PCR in
878 Metopolophium dirhodum (Walker) (Hemiptera: Aphididae). *Scientific Reports* **10**.
879 http://dx.doi.org/10.1038/s41598-020-78974-z.

880 Li X, Schuler MA, Berenbaum MR. 2007. Molecular mechanisms of metabolic resistance to
881 synthetic and natural xenobiotics. *Annu Rev Entomol* **52**: 231–253.

882 Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021a. BUSCO Update: Novel
883 and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage
884 for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol Biol Evol* **38**: 4647–
885 4654.

886 Manni M, Berkeley MR, Seppey M, Zdobnov EM. 2021b. BUSCO: Assessing Genomic Data
887 Quality and Beyond. *Curr Protoc* **1**: e323.

888 Mateo L, Ullastres A, González J. 2014. A transposable element insertion confers xenobiotic
889 resistance in Drosophila. *PLoS Genet* **10**: e1004560.

890 McVicker G, Green P. 2010. Genomic signatures of germline gene expression. *Genome Res*
891 **20**: 1503–1511.

892 Mota-Sanchez D, Wise JC. 2023. The Arthropod pesticide resistance database.
893 http://www.pesticideresistance.org.

894 Muller HJ. 1964. The relation of recombination to mutational advance. *Mutat Res* **106**: 2–9.

895 Nauen R, Bass C, Feyereisen R, Vontas J. 2022. The Role of Cytochrome P450s in Insect
896 Toxicology and Resistance. *Annu Rev Entomol* **67**: 105–124.

897 Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A Fast and Effective
898 Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol Evol*
899 **32**: 268–274.

900 Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose
901 N, Uralsky L, Gershman A, et al. 2022. The complete sequence of a human genome.
902 *Science* **376**: 44–53.

903 Panini M, Chiesa O, Troczka BJ, Mallott M, Manicardi GC, Cassanelli S, Cominelli F,
904 Hayward A, Mazzoni E, Bass C. 2021. Transposon-mediated insertional mutagenesis
905 unmasks recessive insecticide resistance in the aphid Myzus persicae. *Proc Natl*
906 *Acad Sci U S A* **118**. http://dx.doi.org/10.1073/pnas.2100559118.

907 Peccoud J, Loiseau V, Cordaux R, Gilbert C. 2017. Massive horizontal transfer of
908 transposable elements in insects. *Proceedings of the National Academy of Sciences*
909 **114**: 4721–4726.

910 Platt RN, Blanco-Berdugo L, Ray DA. 2016. Accurate transposable element annotation is
911 vital when analyzing new genome assemblies. *Genome Biol Evol* **8**: 403–410.

912 Puinean AM, Foster SP, Oliphant L, Denholm I, Field LM, Millar NS, Williamson MS, Bass C.
913 2010. Amplification of a cytochrome P450 gene is associated with resistance to
914 neonicotinoid insecticides in the aphid Myzus persicae. *PLoS Genet* **6**: 1–11.

915 Quinlan AR, Hall IM. 2010. BEDTools: A flexible suite of utilities for comparing genomic
916 features. *Bioinformatics* **26**: 841–842.

917 Racine JS. 2013. RSTUDIO: A PLATFORM-INDEPENDENT IDE FOR R AND SWEAVE. *J*
918 *Appl Econometrics* **27**: 167–172.

919 Ranson H, Hemingway J. 2005. Mosquito glutathione transferases. *Methods Enzymol* **401**:
920 226–241.

921 Rebollo R, Romanish MT, Mager DL. 2012. Transposable Elements: An Abundant and
922 Natural Source of Regulatory Sequences for Host Genes. *Annu Rev Genet* **46**: 21–

923 42.

924 Remnant EJ, Good RT, Schmidt JM, Lumb C, Robin C, Daborn PJ, Batterham P. 2013.
925 Gene duplication in the major insecticide target site, Rdl, in Drosophila melanogaster.
926 *Proc Natl Acad Sci U S A* **110**: 14705–14710.

927 Rostant WG, Wedell N, Hosken DJ. 2012. Transposable Elements and Insecticide
928 Resistance. *Advances in Genetics* 169–201. http://dx.doi.org/10.1016/b978-0-12-
929 394394-1.00002-x.

930 Salces-Ortiz J, Vargas-Chavez C, Guio L, Rech GE, González J. 2020. Transposable
931 elements contribute to the genomic response to insecticides in Drosophila
932 melanogaster. *Philos Trans R Soc Lond B Biol Sci* **375**: 20190341.

933 Schmidt JM, Good RT, Appleton B, Sherrard J, Raymant GC, Bogwitz MR, Martin J, Daborn
934 PJ, Goddard ME, Batterham P, et al. 2010. Copy number variation and transposable
935 elements feature in recent, ongoing adaptation at the Cyp6g1 locus. *PLoS Genet* **6**:
936 e1000998.

937 Singh KS, Cordeiro EMG, Troczka BJ, Pym A, Mackisack J, Mathers TC, Duarte A, Legeai
938 F, Robin S, Bielza P, et al. 2021. Global patterns in genomic diversity underpinning
939 the evolution of insecticide resistance in the aphid crop pest Myzus persicae.
940 *Commun Biol* **4**: 847.

941 Singh KS, Troczka BJ, Duarte A, Balabanidou V, Trissi N, Carabajal Paladino LZ, Nguyen P,
942 Zimmer CT, Papapostolou KM, Randall E, et al. 2020. The genetic architecture of a
943 host shift: An adaptive walk protected an aphid and its endosymbiont from plant
944 chemical defenses. *Sci Adv* **6**: eaba1070.

945 Song WY, Pi LY, Bureau TE, Ronald PC. 1998. Identification and characterization of 14
946 transposon-like elements in the noncoding regions of members of the Xa21 family of
947 disease resistance genes in rice. *Mol Gen Genet* **258**: 449–456.

948 Sparks TC, Storer N, Porter A, Slater R, Nauen R. 2021. Insecticide resistance management
949 and industry: the origins and evolution of the Insecticide Resistance Action
950 Committee (IRAC) and the mode of action classification scheme. *Pest Manag Sci* **77**:
951 2609–2619.

952 Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically
953 mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**: 637–
954 644.

955 Stapley J, Santure AW, Dennis SR. 2015. Transposable elements as agents of rapid
956 adaptation may explain the genetic paradox of invasive species. *Mol Ecol* **24**: 2241–
957 2252.

958 Steenwyk JL, Buida TJ, Labella AL, Li Y, Shen X-X, Rokas A. 2021. PhyKIT: a broadly
959 applicable UNIX shell toolkit for processing and analyzing phylogenomic data.
960 *Bioinformatics*. http://dx.doi.org/10.1093/bioinformatics/btab096.

961 Sternberg ED, Thomas MB. 2018. Insights from agriculture for the management of
962 insecticide resistance in disease vectors. *Evol Appl* **11**: 404–414.

963 Sultana T, Zamborlini A, Cristofari G, Lesage P. 2017. Integration site selection by
964 retroviruses and transposable elements in eukaryotes. *Nat Rev Genet* **18**: 292–308.

965 Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, Snyder MP, Wang T. 2014. Widespread
966 contribution of transposable elements to the innovation of gene regulatory networks.
967 *Genome Res* **24**: 1963–1976.

968 Team R. 2015. RStudio: integrated development for R. *RStudio, Inc, Boston, MA URL*
969 *http://www rstudio com*.

970 Team RC. 2013. R: A language and environment for statistical computing.

971 van de Lagemaat LN, Landry J-R, Mager DL, Medstrand P. 2003. Transposable elements in
972 mammals promote regulatory variation and diversification of genes with specialized
973 functions. *Trends Genet* **19**: 530–536.

974 Van Emden HF, Harrington R. 2017. *Aphids as crop pests*. Cabi.

975 Wang L, Wang J, Ma Y, Wan P, Liu K, Cong S, Xiao Y, Xu D, Wu K, Fabrick JA, et al. 2019.
976 Transposon insertion causes cadherin mis-splicing and confers resistance to Bt
977 cotton in pink bollworm from China. *Sci Rep* **9**: 7479.

978 Weedall GD, Riveron JM, Hearn J, Irving H, Kamdem C, Fouet C, White BJ, Wondji CS.
979 2020. An Africa-wide genomic evolution of insecticide resistance in the malaria vector
980 Anopheles funestus involves selective sweeps, copy number variations, gene
981 conversion and transposons. *PLoS Genet* **16**: e1008822.

982 Wells JN, Feschotte C. 2020. A Field Guide to Eukaryotic Transposable Elements. *Annu*
983 *Rev Genet* **54**: 539–561.

984 Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P,
985 Morgante M, Panaud O, et al. 2007. A unified classification system for eukaryotic
986 transposable elements. *Nat Rev Genet* **8**: 973–982.

987 Worner SP, Gevrey M. 2006. Modelling global insect pest species assemblages to
988 determine risk of invasion. *J Appl Ecol* **43**: 858–867.

989 Yang C, Pan H, Liu Y, Zhou X. 2014. Selection of Reference Genes for Expression Analysis
990 Using Quantitative Real-Time PCR in the Pea Aphid, Acyrthosiphon pisum (Harris)
991 (Hemiptera, Aphidiae). *PLoS ONE* **9**: e110454.
992 http://dx.doi.org/10.1371/journal.pone.0110454.

993 Yang C, Pan H, Liu Y, Zhou X. 2015. Temperature and Development Impacts on
994 Housekeeping Gene Expression in Cowpea Aphid, Aphis craccivora (Hemiptera:
995 Aphidiae). *PLoS One* **10**: e0130593.

996 Yang Y, Chen H, Wu Y, Yang Y, Wu S. 2007. Mutated cadherin alleles from a field
997 population of Helicoverpa armigera confer resistance to Bacillus thuringiensis toxin
998 Cry1Ac. *Appl Environ Microbiol* **73**: 6939–6944.

999 Zhao J, Jin L, Yang Y, Wu Y. 2010. Diverse cadherin mutations conferring resistance to
1000 Bacillus thuringiensis toxin Cry1Ac in Helicoverpa armigera. *Insect Biochem Mol Biol*
1001 **40**: 113–118.

1002

1003

1004 ## Supplemental File Captions

1005 Supplemental Figure S1. The relationship between TE content and genome size in
1006 aphids. Blue line indicates the relationship calculated using linear regression. Grey
1007 area indicates 95% confidence intervals.
1008
1009 Supplemental Figure S2. Correlation between genic:intergenic base pairs ratio and
1010 TE content in aphids.
1011
1012 Supplemental Figure S3. Expression of XGFs and housekeeping genes in germline
1013 tissues (testis and ovary) in *D. melanogaster* using data from FlyAtlas2 (Krause et al.
1014 2022). Expression is measured as fragments per kilobase of transcript per million
1015 base pairs sequenced (FPKM), to determine expected expression of XGFs in
1016 germline cell types. Gene types are indicated by colours in the key.
1017
1018 Supplemental Figure S4. TE abundance around xenobiotic gene family loci
1019 compared to all other genes and housekeeping genes, split by xenobiotic resistance
1020 gene type.
1021
1022 Supplemental Figure S5. TE coverage around xenobiotic gene family loci compared
1023 to housekeeping genes.
1024
1025 Supplemental File S1. Aphid species phylogeny rooted with 7 outgroup species in
1026 Nexus format.
1027
1028 Supplemental File S2. Tar archive containing TE annotations in GFF format for all

1029 aphid genomes in this study.

1030

1031 Supplemental File S3. Tar archive containing XGF coordinates in GFF format for all

1032 aphid genomes in this study.

1033

1034 Supplemental File S4. *Myzus persicae* housekeeping amino acid sequences for

1035 each aphid genome in FASTA format.

1036

1037 Supplemental File S5. Tar archive containing housekeeping gene coordinates in

1038 GFF format for all aphid genomes in this study.

1039

1040 Supplemental File S6. Cytochrome P450 amino acid sequences for each aphid

1041 genome in FASTA format.

1042

1043 Supplemental Table S1. High-level TE quantification for the 21 aphid and 7 outgroup

1044 genome assemblies used in this study.

1045

1046 Supplemental Table S2. Genic:Intergenic bases ratios for measures of genome

1047 compactness

1048

1049 Supplemental Table S3. TE enrichment at XGFs among aphids.

1050

1051 Supplemental Table S4. Germline expression for XGF and housekeeping genes in

1052 *A. pisum* from (Jaquiéry et al. 2022)

1053

1054 Supplemental Table S5. Resistance status for each identified Cytochrome P450

1055 gene, using evidence from (Katsavou et al. 2022).

1056

1057 Supplemental Table S6. Genome assembly accessions and BUSCO annotation

1058 information for all aphid and outgroup genome assemblies included in this study.

1059