Patterns of transposable element accumulation at resistance 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. A deeper understanding of the evolutionary and genomic mechanisms underpinning resistance evolution is required. Several studies have implicated TEs in xenobiotic resistance evolution in insects. However, analyses are generally restricted to one insect species, and/or small numbers of xenobiotic resistance loci (XRL). We examine evidence for the accumulation of transposable elements (TEs) at XRL by performing a comparative genomic analysis across 21 aphid genomes. We consider major subsets of XRL involved in metabolic resistance to insecticides: Cytochrome P450s, glutathione S-transferases, esterases, UDP-glucuronosyltransferases, and ABC transporters. TEs are significantly enriched at XRL compared to other genes. XRL show similar levels of TE enrichment to housekeeping genes. But unlike housekeeping genes, XRL are not constitutively expressed in germline cells, supporting the selective enrichment of TEs at XRL, rather than enrichment due to chromatin availability. Hotspots of TE enrichment occur around certain XRL, whilst others are depleted. In aphids of agricultural importance, we find particular enrichment of TEs around cytochrome P450 genes with known functions in the detoxification of synthetic insecticides, compared to genes involved in other processes. Our results provide evidence supporting a general role for TEs as a source of genomic variation at host XRL, and highlight the existence of considerable variability in TE content across XRL and host species. These findings demonstrate the need for detailed functional verification analyses to clarify the significance of individual insertions and elucidate underlying mechanisms at TE-XRL 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 to many insects evolving 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 There is growing evidence that transposable elements (TEs) can play an important
57 role in the evolution of xenobiotic resistance in insects (Gilbert et al. 2021; Rostant et al.
58 2012), with examples from several major lineages: Diptera: *Drosophila* (Feyereisen 2005;
59 Bogwitz et al. 2005; Schmidt et al. 2010; Aminetzach et al. 2005; Remnant et al. 2013;
60 Mateo et al. 2014; Salces-Ortiz et al. 2020), *Culex (Itokawa et al. 2011; Darboux et al. 2007),*
61 *Musca domestica* (Li et al. 2007; Kasai and Scott 2001), *Anopheles funestus* (Weedall et al.
62 2020) Lepidoptera: *Heliothis (Gahan et al. 2001; Yang et al. 2007; Zhao et al. 2010; Chen*
63 *and Li 2007)*, *Pectinophora gossypiella* (Fabrick et al. 2011; Wang et al. 2019) Hemiptera:
64 *Myzus persicae* (Singh et al. 2020; Panini et al. 2021). TEs are DNA sequences capable of
65 moving from one genomic location to another within the genome. TEs occur in nearly all
66 eukaryotic genomes, and are implicated in the evolution of host genomic novelty through
67 diverse processes including the modification of regulatory networks, chromosomal
68 rearrangements, exon shuffling, and donation of coding sequence (Bourque et al. 2018;
69 Wells and Feschotte 2020; Cosby et al. 2019; Sundaram et al. 2014). In the case of
70 xenobiotic resistance, TEs are reported to have contributed to evolvability via myriad
71 mechanisms, including: gene amplification and duplication (Singh et al. 2020; Remnant et al.
72 2013), knockout of a susceptible allele in heterozygous individuals (Panini et al. 2021),

73 increases in detoxification gene expression (Bogwitz et al. 2005; Itokawa et al. 2011), allelic
74 succession leading to increases in resistance gene copy number (Schmidt et al. 2010), and
75 alternative splicing and production of truncated proteins that prevent interactions with
76 xenobiotics (Darboux et al. 2007; Gahan et al. 2001; Yang et al. 2007; Zhao et al. 2010).

77 Studies examining the contribution of TEs to resistance evolution suggest that there
78 can be an additive effect of successive TE insertions at focal host loci over time. For
79 example, in the well-known case of the cytochrome P450 (CYP) gene *CYP6G1* and DDT
80 resistance in *Drosophila*, successive TE insertions are linked with an increasing ability to
81 detoxify the insecticide *(Catania et al. 2004; Daborn et al. 2002; Chung et al. 2007; Schmidt*
82 *et al. 2010)*. A similar process involving the accumulation of TEs, in this case at *CYP6CY3,*
83 is implicated in the evolution of resistance to nicotine, following a host plant shift to tobacco
84 in the aphid *Myzus persicae nicotianae (Puinean et al. 2010; de Little et al. 2017)*. More
85 widely, there is evidence that the selective accumulation of TEs at rapidly evolving host loci
86 under strong selection may represent a general evolutionary process. For example, studies
87 have reported the enrichment of TEs at other gene classes, such as immune genes and
88 those involved in responses to external stimuli (van de Lagemaat et al. 2003; Song et al.
89 1998). However, patterns in TE accumulation remain poorly described, and it is unclear to
90 what extent TEs may be enriched at xenobiotic resistance loci (XRL) compared to other host
91 genes, or how much variability exists across XRL. Meanwhile, existing studies strongly
92 suggest that TEs are involved in individual cases of resistance to certain insecticides, but
93 there is currently very limited understanding of how associations with TEs vary among
94 different classes of xenobiotic resistance loci (XRL), or among whole clades of insects.

95 Here we consider patterns of TE accumulation at XRL in the genomes of 21 species
96 sampled from across aphid phylogeny, for which high-quality genome assemblies are
97 available. The aphid family is a relevant clade within which to explore the genomic bases of
98 insecticide resistance, as it includes numerous globally important crop pests responsible for
99 causing billions of US dollars of crop losses annually (Van Emden and Harrington 2017), and

100 members that have been intensively targeted with insecticides and evolved resistance. For
101 example, the severely damaging global crop pest *Myzus persicae* (the green peach aphid) is
102 resistant to at least 82 insecticides, via at least eight different resistance mechanisms (Bass
103 et al. 2014; Whalon et al.; Singh et al. 2021).

104 Since the TE landscapes of aphids are poorly described, we begin by characterising
105 patterns of TE content across our focal genomes. After this, we examine key fundamental
106 questions relating to the pattern of associations between TEs and XRL in aphid genomes. (i)
107 Firstly, we test whether XRL are enriched for TEs compared to other host genes, to evaluate
108 the overall signature of TE accumulation at XRL. (ii) Secondly, we examine the extent to
109 which TE content varies among individual XRL and among major XRL classifications, to
110 determine if contributions from TEs are more pronounced at certain types of XRL. (iii)
111 Thirdly, we consider if specific TE classifications are enriched at XRL (e.g DNA TEs, rolling
112 circles, *Penelope*-like elements, LINEs, SINEs, and LTRs), to explore if particular TEs are
113 predisposed to contribute to the evolution of host resistance. (iv) Fourthly, we assess
114 whether patterns of TE accumulation at XRL vary among aphid species to examine if
115 patterns vary across aphid phylogeny. (v) Fifthly, we investigate TE enrichment at XRL that
116 have demonstrated associations with the detoxification of synthetic insecticides, to test if
117 these show particularly strong signatures of TE recruitment. (vi) Lastly, we explore whether
118 our results suggest a selective role for patterns in TE accumulation at XRL versus the
119 alternative explanation of increased availability for TE insertion due to chromatin availability.

120 We find that XRL are enriched for TEs compared to the mean level for other gene
121 types in aphid genomes. Evidence from *Drosophila* suggests that XRL are not constitutively
122 expressed in germline cell types, and so are not more available to transposition than other
123 host genes, providing evidence supporting a role for the selective enrichment of TEs at XRL.
124 DNA TEs and SINEs show the highest enrichment at XRL, and we find that TE enrichment
125 at XRL varies among XRL types and across aphid species, both in terms of TE coverage
126 and count. We also find particularly high enrichment of TEs at cytochrome P450 genes with

127 known functions in the detoxification of synthetic insecticides in three aphids of major
128 agricultural importance (*A. pisum*, *M. cerasi*, and *M. persicae*). Overall, these findings are
129 consistent with a role for TE accumulation at strongly selected XRL, and highlight the
130 extremely high levels of TE enrichment (both in terms of TE coverage and copy number) that
131 can accumulate at certain XRL, in turn supporting an important and general role for TEs in
132 the evolution of insecticide resistance.

133 Results

134 *Transposable Element Landscapes in Aphids*

135 TEs make considerable contributions to genome size in eukaryotes (Kidwell 2002; Gregory
136 2005). In aphids, we find that TE content varies from 8.21% (*Melanaphis sacchari*) to
137 35.52% (*Metopolophium dirhodum*) of total genome size (Figure 1B, Supplementary Table
138 1). This is in accordance with levels of TE content identified among a large-scale survey of
139 196 insect species, where TE content varied from 0.08% to 53.93% of total genome size
140 (Peccoud et al. 2017). Similarly, consistent with findings in other eukaryotic lineages, we
141 report a significant strong positive association between genome size and TE content in
142 aphids (Linear regression, $F_{1,26}$=100.7, $p<0.01$, R^2=0.79, Supplementary Figure 1).

143 TEs have a non-random distribution within host genomes (Bourque et al. 2018). As
144 expected, given the predominantly deleterious effects of TE insertion into host genes
145 (Bourque et al. 2018; Sultana et al. 2017), exonic regions exhibit the lowest levels of TE
146 sequence content in aphid genomes (mean TE content in exons = 2.08% of total genome
147 size). Across 19 of the 21 aphid genomes considered here, the greatest TE coverage
148 (expressed as proportion of total genome size) is found in 5' and 3' gene flanking regions,
149 where it varies between 3.49% (*M. sacchari*) and 17.94% (*M. dirhodum*) (Figure 2).
150 However, this pattern is not universal across all aphid species. In *Sipha flava,* the majority of

151 TE coverage occurs in intronic regions (10.60%), while in *M. cerasi*, the majority occurs in
152 intergenic space (11.28%) (Figure 2).

153 The amount of TE sequence and the frequency of TE insertions in genic versus
154 intergenic regions in aphid genomes is not significantly correlated with genome
155 compactness, described as the ratio of genic (intron and exon) to intergenic (gene flanks and
156 intergenic space) base pairs (Spearman's Rank, S=1680, p>0.05, rho=-0.09, Supplementary
157 Figure 2, Supplementary Table 2). *S. flava* has the most compact genome, with a
158 genic:intergenic base pair ratio of 1.50:1 (250.0 Mb:166.8 Mb), and presumably a higher
159 likelihood of TE insertion into genic regions compared to intergenic space, leading to the
160 observation that the majority of TE insertions are found in intronic regions. In contrast, *M.*
161 *cerasi* has the least compact genome, with a much lower genic:intergenic base pair ratio of
162 0.13:1 (76.1 Mb:559.2 Mb). In less compact genomes such as *M. cerasi*, TEs can
163 accumulate in expanded intergenic regions that act as 'genomic safe havens', where
164 insertion is less likely to exert deleterious effects (Arkhipova 2018). For context, the human
165 genome (GCF_009914755.1), has a TE content of 44%-69% (Nurk et al. 2022; de Koning et
166 al. 2011) and a genic:intergenic base pair ratio of 0.55:1 (1717.7 Mb:3117.3 Mb), making it
167 several times more compact than *M. cerasi*, but much less compact than *S. flava*.

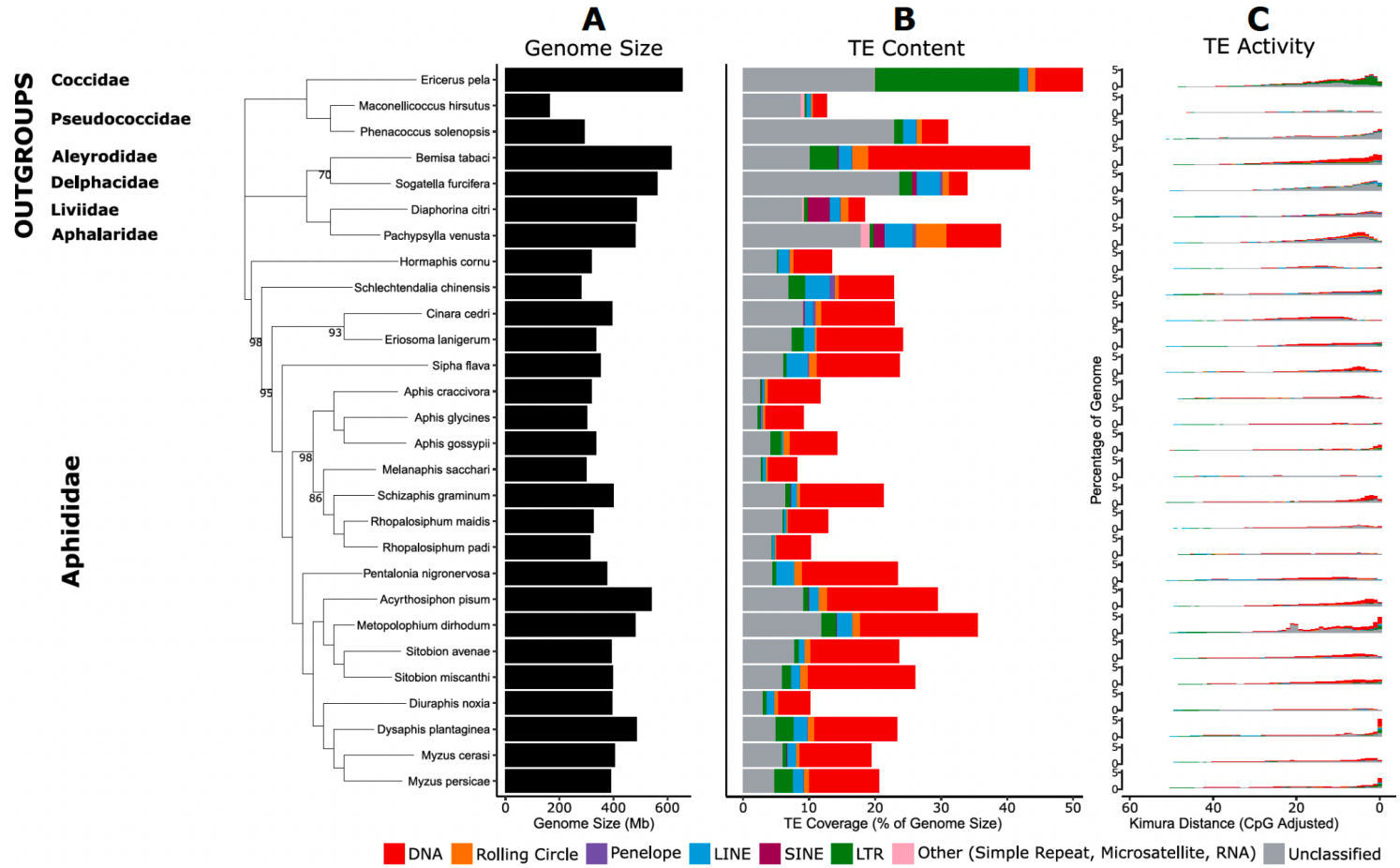

Figure 1. Summary of TE content and activity in aphid genomes, plotted in phylogenetic order. Nodes in the aphid phylogeny with bootstrap support <100 are labelled. Genome size is represented by the black bars. TE content is expressed as percentage of total genome size for each species, with major TE 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 TE activity, where a lower Kimura distance indicates more recent TE activity. Activity is organised such that recent TE activity is towards the RHS of the X axis.

174 Across all aphid species and in most individual genomes, the dominant TE
175 classification is DNA TEs, which comprise between 36.6% (*Schlechtendalia chinensis*) and
176 68.9% (*Aphis craccivora*) of total TE content (μ = 45.58%, SD = 17.74%, Figure 1B, Table 1,
177 Supplementary Table 1). SINEs are present only in the genomes of 7 of the 21 species
178 considered (Supplementary Table 1) and, where present, they comprise just 0.009% (*A.*
179 *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	45.58%	17.74%
Unclassified	35.40%	14.67%
LINE	7.58%	3.70%
LTR	6.40%	7.93%
Rolling Circle	3.48%	2.12%
SINE	2.14%	5.24%
Penelope	0.39%	0.73%

180 Table 1. Means and standard deviations of the percentage of total TE content attributed to
181 each of the main TE classifications in the 21 aphid genomes included in this study.

182

183 There is little evidence of ancient TE activity in aphid genomes, as indicated by the
184 relatively low number of TEs showing high divergence from their respective TE consensus
185 models in repeat landscape plots (Figure 1C). On the other hand, there is considerable
186 evidence of relatively recent TE activity in most aphid species, with particular evidence of
187 very recent TE activity in *Myzus persicae*, *Dysaphis plantaginea*, and *M. dirhodum*, indicated
188 by large numbers of TEs with low genetic distance from their respective consensus
189 sequences (Figure 1C). The relative absence of ancient TE activity is presumably a
190 consequence of a relatively rapid genomic turnover rate in aphids, and contrasts strongly
191 with TE landscapes in mammals, where a low rate of genomic turnover leads to

192 considerable accumulations of ageing TEs (Blass et al. 2012). Instead, the dynamics we
193 observe for aphids are consistent with patterns reported for other insects, such as
194 lepidopterans (Lavoie et al. 2013; Baril and Hayward 2022).

195 The extent to which TE content differs among the genomes of closely related species
196 can vary greatly. We find no significant effect of phylogeny on TE abundance (No. of TEs
197 present in a genome) or diversity (No. distinct TE families present in a genome (Wicker et al.
198 2007)) in aphids (Phylogenetic GLMM, TE abundance: posterior mean = 0.599, 95% HPD =
199 0.005,0.999; TE diversity: posterior mean = 0.023, 95% HPD = 0.000,0.111). Thus, more
200 closely related aphid species do not possess similar numbers of TEs or a similar level of TE
201 diversity. We also find that the majority of TE families present in aphid genomes are species-
202 specific, with almost half of the TE families present in a particular aphid genome being
203 unique to that species (μ = 47.32%) (Figure 2). Considering TE families that occur in two or
204 more aphid species, the set containing the highest number of shared TEs is 21 (i.e. all
205 species considered) with a mean percentage of 9.92% of TE families (Figure 2). The
206 remaining 42.76% of TE families are shared in similar proportions among 2 to 20 aphid
207 species (i.e. μ = 2.38% per category, Figure 2). The strong signature of species-specificity
208 observed for aphid TE families suggests that the independent gain of new TE families is a
209 major factor driving aphid TE landscapes. Collectively, these findings suggest a dynamic
210 repeat landscape in aphid genomes, characterised by relatively frequent gain and turnover
211 of TEs.

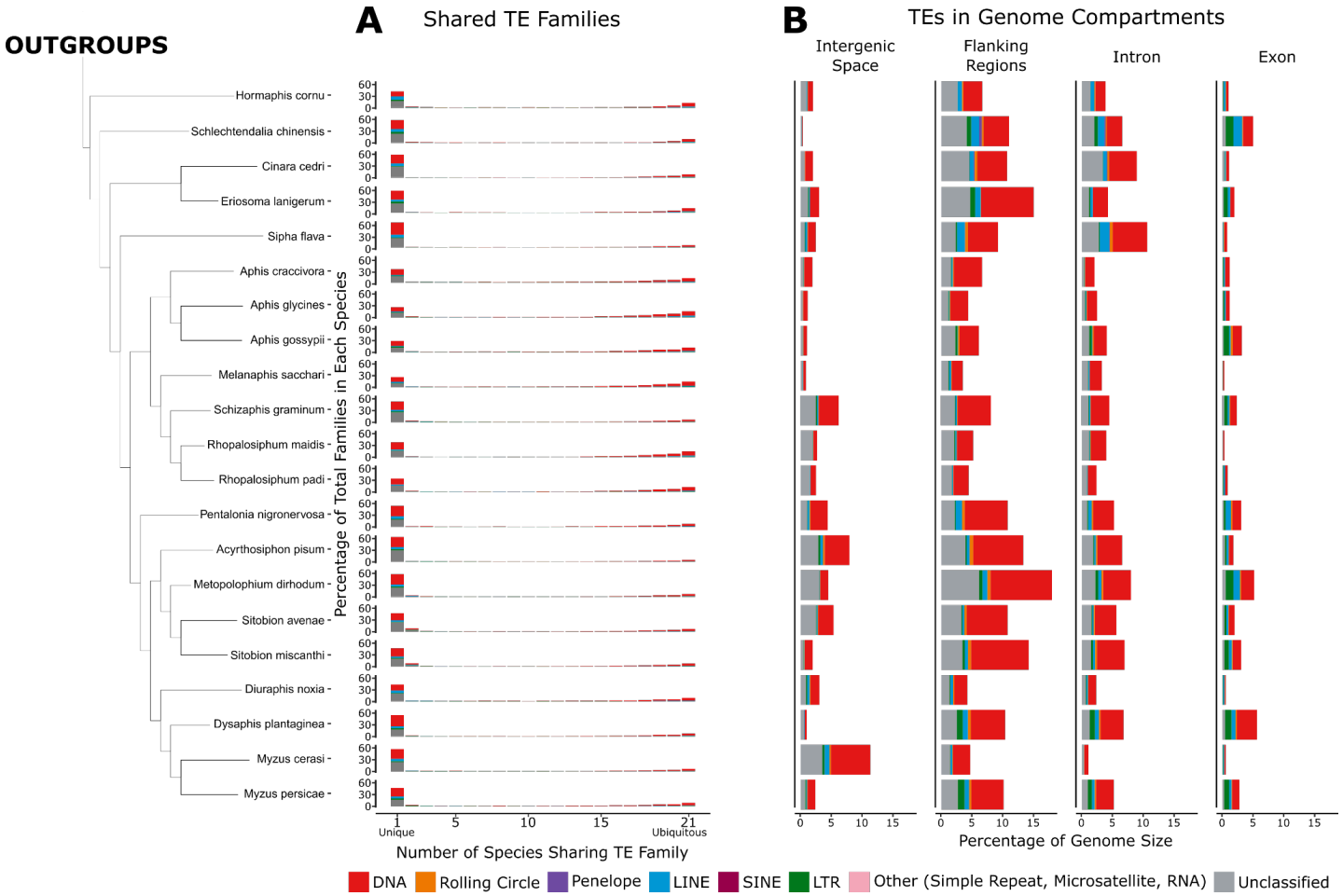

Figure 2. Overview of the extent to which TE families are shared among aphid genomes, and their insertion locations. **A**. Quantification of shared and unique TE families in aphids. Main TE classifications are represented by the colours indicated in the key. **B**. TE genome compartment occupancy, expressed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

217 *TEs are enriched at XRL compared to other host genes*

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

222 Across all aphid genomes, we find significant enrichment of TEs, both in TE
223 coverage and TE count, at XRL in comparison to all other genes excluding XRL (TE
224 coverage: Wilcoxon Rank Sum, $W_{7095,439790}$ = 874,611,740, $p<0.01$; TE count: Wilcoxon Rank
225 Sum, $W_{7095,439790}$ = 931,410,344, $p<0.01$) (Figure 3, Supplementary Figure 3). Specifically,
226 aphids have a mean of 2.4x TE coverage and 2.1x TE count at XRL compared to all other
227 genes.

228

229 *TE content varies among individual XRL*

230 We observe an enrichment of TEs around all major XRL categories considered compared to
231 other host genes: Cytochrome P450s (CYPs), 2.4x TE coverage, 2.1x TE count; glutathione
232 S-transferases (GSTs), 2.7x TE coverage, 2.3x TE count; esterases, 2.0x TE coverage, 1.8x
233 TE count; UDP-glucuronosyltransferases (UGTs), 2.0x TE coverage, 1.7x TE count; ABC
234 transporters (ABCs), 2.8x TE coverage, 2.4x TE count (Figure 3, Supplementary Figure 3,
235 Supplementary Table 3). Whilst all major XRL categories are enriched for TEs, there is no
236 significant difference in relative enrichment among XRL types, suggesting relatively equal
237 levels of enrichment across XRL types (Kruskal-Wallis, TE sequence: $\chi^2_4 = 9.05$, $p = 0.06$;
238 TE count: $\chi^2_4 = 4.98$, $p = 0.29$). However, considerable variability in TE content is present
239 within XRL categories, attributable to extremely large accumulations of TEs at certain
240 individual XRL (Table 2, Figure 4). Some of these XRL are unusually large in terms of
241 sequence length compared to the mean size for the XRL type in question, as indicated by

242 gene size Z-scores > 3 (Table 2), which is potentially a consequence of the increased
243 presence of TEs in their intronic regions (Figure 4). To investigate this, we consider gene
244 sizes with genic TEs removed, finding that only 5/18 of the most TE-enriched XRL have
245 significantly inflated gene sizes, indicated by a Z-score of > 3. This suggests that these loci
246 were significantly larger than average before accounting for TE contributions. However, the
247 remaining 13 highly enriched XRL have gene sizes within the expected range in the absence
248 of TEs, suggesting that TEs are indeed responsible for the significant increases in gene size
249 at the majority of these loci.

250 Individual XRL with the greatest enrichment of TE coverage and TE count (i.e. >35kb
251 TE sequence and >80 TE insertions, compared to a non-XRL mean of 1.5kb TE sequence
252 and 1.48 TE insertions) are listed in Table 2. Whilst all XRL types are represented in the TE
253 hotspots listed in Table 2, there are markedly more hotspots at UGTs and CYP genes, and
254 fewer at ABCs, GSTs, and Esterases.

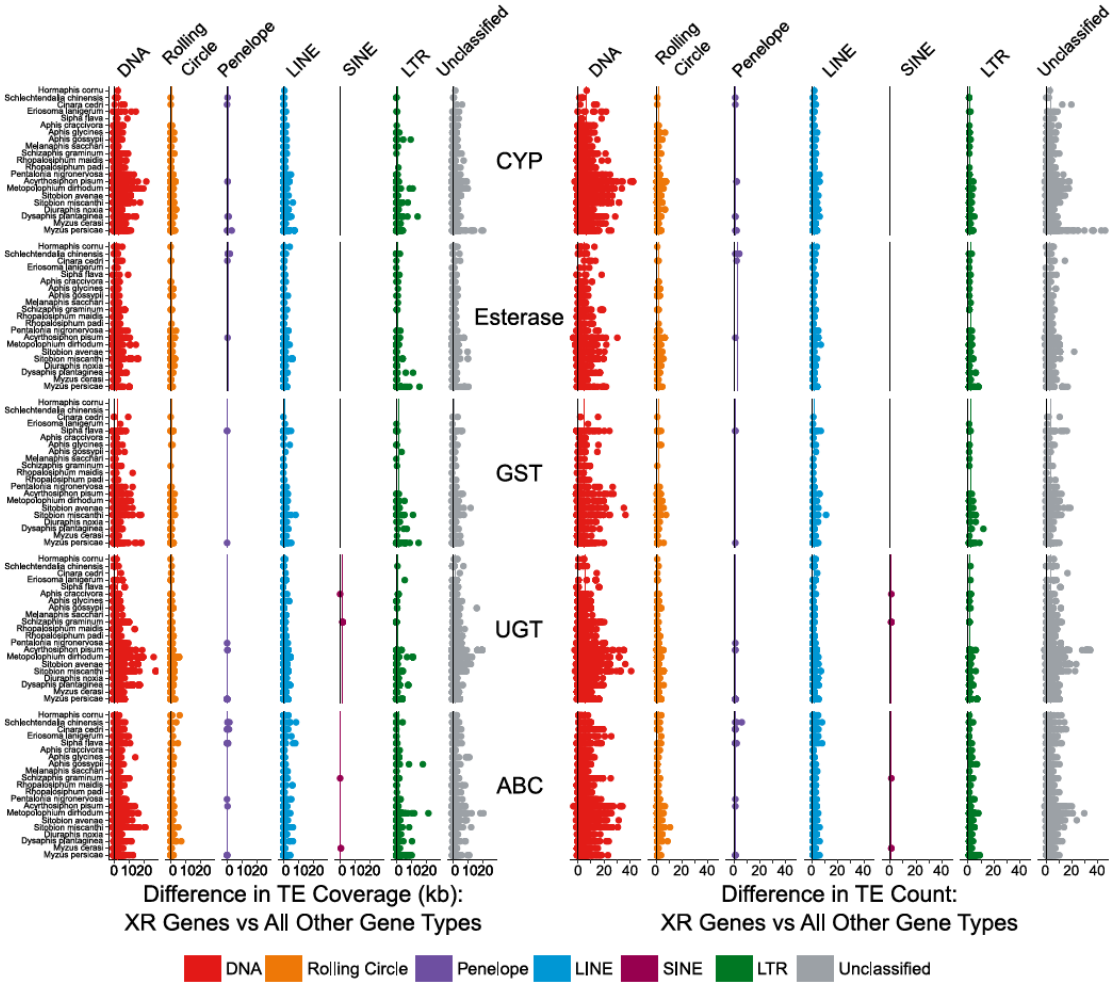

Figure 3. Differences in TE coverage and TE count at XRL compared to all other host genes. Coverage and count of genic (exon and intron) XRL sequence plus flanking regions (20kb directly upstream or downstream of the gene body) is shown. Each point represents a single XRL. Black lines indicate the expected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

Aphid Species	Gene ID	XRL Type	Mean Gene Size (by gene type)	Gene Size in bp (Z-score)	XRL associated TE Coverage (bp)	Fold-increase above aphid-wide enrichment (XRL-type specific)	Fold-increase above aphid-wide enrichment (non-XRL genes)	XRL associated TE Count	Fold-increase above aphid-wide enrichment (XRL-type specific)	Fold-increase above aphid-wide enrichment (non-XRL 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	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	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	11,211	39,350 (3.00)	31,743	3.81	20.90	81	4.26	54.73	25,849 (1.50)

Table 2. XRL 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 XRL size, respectively. Entries in bold are XRLs that are highly enriched for both TE coverage and TE count.

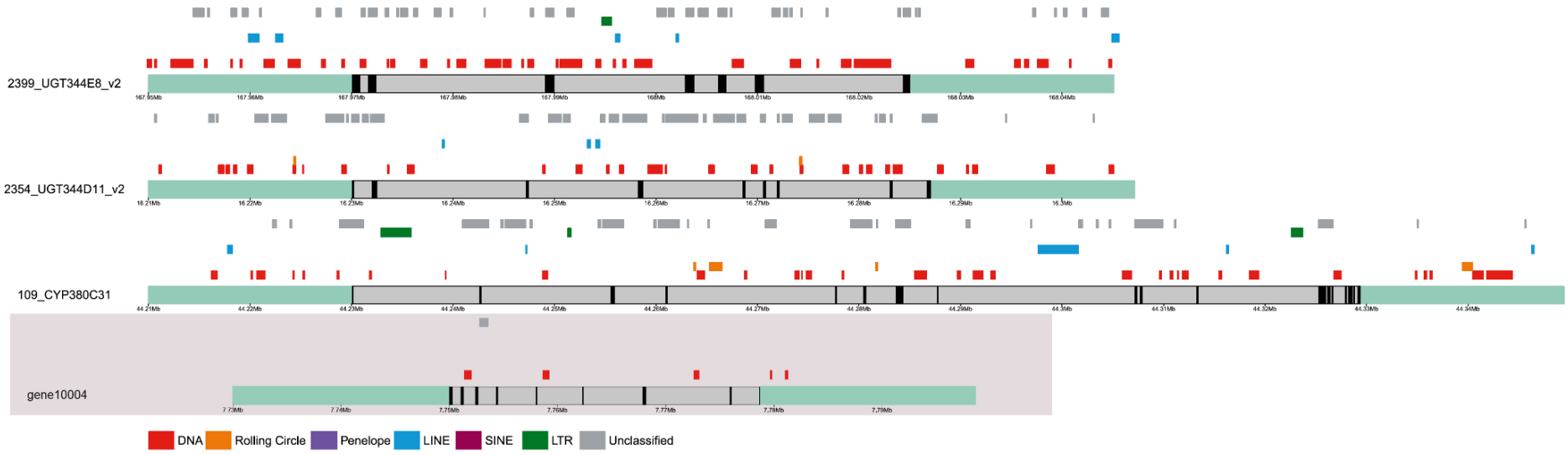

Figure 4. Karyoplots illustrating the most TE-dense XRL, corresponding to the XRL highlighted in bold in Table 2. For comparison, the gene region represented in the shaded box (gene10004) shows a non-XRL from *M. persicae* with representative TE coverage close to the mean for non-XRL (1,519bp). XRLs 125_CYP380C52 and 105_CYP380C19 are not shown, as these are gene isoforms in the same region as 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.

278 *Specific TE types are enriched at XRL*

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

282 The TE type with the greatest degree of enrichment in coverage around XRL in
283 comparison to other genes is SINEs, although there is a slight non-significant reduction in
284 SINE count at XRL (**TE coverage**, 3.2x enrichment: XRL μ = 1,144bp, all other genes μ =
285 358bp; Wilcoxon Rank Sum, $W_{9,482} = 439$, $p < 0.01$; **TE count,** 0.82x enrichment: XRL μ =
286 1.00, all other genes μ = 1.22; Wilcoxon Rank Sum, $W_{9,482} = 2592$, $p = 0.14$). This is notable,
287 since SINEs are generally very scarce in aphid genomes, and suggests a particular retention
288 of SINE sequence at XRL (Figure 1, Figure 3, Table 1, Supplementary Table 1). However,
289 given their scarcity, the relative contribution that SINEs make to TE enrichment at XRL
290 compared to other TE types is extremely low. Conversely, DNA TEs are the most abundant
291 TE type in aphid genomes, and they are significantly enriched at XRL both in TE coverage
292 and TE count (**TE coverage**, 2.1x enrichment: XRL μ = 4,225bp, all other genes μ =
293 1,970bp; Wilcoxon Rank Sum, $W_{6703,368474} = 732,223,460$, $p < 0.01$; **TE count**, 6.2x
294 enrichment: XRL μ = 10.04, all other genes μ = 1.61; Wilcoxon Rank Sum, $W_{6703,368474}$ =
295 1,538,514,776, $p < 0.01$) (Figure 3). Rolling circle TEs show significant, but lower levels of
296 enrichment (**TE coverage**, 1.2x enrichment: XRL μ = 851bp, all other genes μ = 735bp;
297 Wilcoxon Rank Sum, $W_{2013,61439} = 57,121,576$, $p<0.01$; **TE count**, 1.7x enrichment: XRL μ =
298 2.15, all other genes μ = 1.24; Wilcoxon Rank Sum, $W_{2013,61439} = 70,189,210$, $p < 0.01$).
299 LINEs also show significant but lower levels of enrichment at XRL (**TE coverage**, 1.2x
300 enrichment: XRL μ = 1,154bp, all other genes μ = 978bp; Wilcoxon Rank Sum, $W_{2866,110685}$ =
301 152,229,422, $p < 0.01$; **TE count**, 1.3x enrichment: XRL μ = 1.79, all other genes μ = 1.43;
302 Wilcoxon Rank Sum, $W_{2866,110685} = 203,186.510$, $p < 0.01$). In contrast, we find only a slight,
303 non-significant, enrichment of LTR TE coverage at XRL, but we do identify a significant

304 enrichment of LTR TE count (**TE coverage**, 1.4x enrichment: XRL μ = 1,853bp, all other
305 genes μ = 1,319bp; Wilcoxon Rank Sum, $W_{1127,58711}$ = 32,412,424, p = 0.24; **TE count**, 1.7x
306 enrichment: XRL μ = 2.54, all other genes μ = 1.53; Wilcoxon Rank Sum, $W_{1127,58711}$ =
307 41,974,654, $p < 0.01$). Only PLEs show a depletion, albeit insignificant, in TE coverage
308 around XRL, but they also show significant enrichment in count around XRL (**TE coverage**,
309 0.7x enrichment: XRL μ = 472bp, all other genes μ = 621bp; $W_{49,7075}$ = 197,672, p = 0.09; **TE**
310 **count**, 1.3x enrichment: XRL μ = 1.43, all other genes μ = 1.12; Wilcoxon Rank Sum,
311 $W_{49,7075}$ = 190,634, $p < 0.05$). However, like SINEs, PLEs account for a very small proportion
312 of total TE content (Supplementary Table 1).

TE Type	Non-XRL	All XRL	CYPs	GSTs	Esterases	UGTs	ABCs
DNA	1,969bp (2)	4,225bp (10)	4,481bp (11)	4,757bp (11)	3,489bp (8)	3,748bp (8)	4,651bp (11)
Rolling Circle	735bp (1)	851bp (2)	823bp (2)	814bp (2)	1,115bp (3)	706bp (2)	798bp (2)
Penelope	621bp (1)	505bp (1)	841bp (1)	175bp (1)	848bp (2)	185bp (1)	477bp (1)
LINE	978 (1)	1,154bp (2)	1,089bp (2)	1,217bp (2)	1,408bp (2)	944bp (2)	1,114bp (2)
SINE	358 (1)	1,144bp (1)	-	-	-	1,686bp (1)	602bp (1)
LTR	1,319bp (2)	2,055bp (3)	1,214bp (1)	2,635bp (3)	1,443bp (2)	1,264bp (2)	3,718bp (5)
Unclassified	1,294 (1)	2,163bp (5)	2,079bp (5)	2,248bp (6)	1,718bp (4)	2,392bp (6)	2,380bp (6)

313 Table 3. Mean base pair coverage of each TE classification at XRL among aphids. Numbers
314 in brackets indicate mean TE copy number at each XRL for each TE classification.

315

316 *XRL enrichment varies among aphid genomes*

317 We find no significant difference in TE enrichment among different XRL types (i.e CYP
318 genes, esterases, UGTs, GSTs, ABCs) (Kruskal-Wallis, TE coverage: $\chi^2_4 = 9.05$, $p = 0.06$;
319 TE count: $\chi^2_4 = 4.98$, $p = 0.29$, Figure 3). However, we do find significant variation in the
320 magnitude of TE enrichment at XRL among aphid species (linear regression, TE coverage:
321 $F_{20,82} = 35.28$, $p < 0.01$, adjusted $R^2 = 0.87$, TE count: $F_{20,82} = 27.19$, $p < 0.01$, adjusted $R^2 =$
322 0.84). Specifically, in terms of TE coverage, *C. cedri* shows the highest levels of enrichment
323 across all XRL, with a mean enrichment of 15.9x compared to non-XRL (Figure 5,
324 Supplementary Table 3). This is considerably higher than the lowest level of mean TE
325 coverage enrichment, which was found in *M. sacchari* with 1.9x TE coverage enrichment at
326 XRL compared to non-XRL (Figure 5, Supplementary Table 3). In terms of TE count, the
327 highest level of enrichment at XRL compared to non-XRL was also found in *C. cedri*, with a
328 mean TE count enrichment of 28.5x the number of TE insertions, whilst the lowest level was
329 found in *A. craccivora*, with a mean enrichment of 4.6x the number of TE insertions at XRL
330 compared to non-XRL (Figure 5, Supplementary Table 3)

331 Overall, whilst we observe enrichment in both TE coverage and TE count at XRL
332 compared to non-XRL in aphid genomes, fold-change increases are larger for TE count.
333 High enrichment for TE count combined with lower increases in TE coverage suggest that
334 enrichment is driven by the presence of large numbers of TE fragments at XRL, as opposed
335 to full-length TEs (Figure 6). Indeed, mean TE length at XRL across aphids is just 444bp,
336 which is much shorter than most full length TEs, excluding SINEs (Wells and Feschotte
337 2020; Wicker et al. 2007). Further, 90.23% of TEs across XRL are below 1,000bp in length
338 (Figure 6). Therefore, it is likely that a general pattern in TE accumulation at XRL is the
339 truncation of TEs with retention of specific TE domains, as previously observed for certain
340 individual cases of TE co-option during the evolution of host insecticide resistance (Singh et
341 al. 2020; Remnant et al. 2013; Panini et al. 2021; Bogwitz et al. 2005; Itokawa et al. 2011;

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

343 2010).

344

345

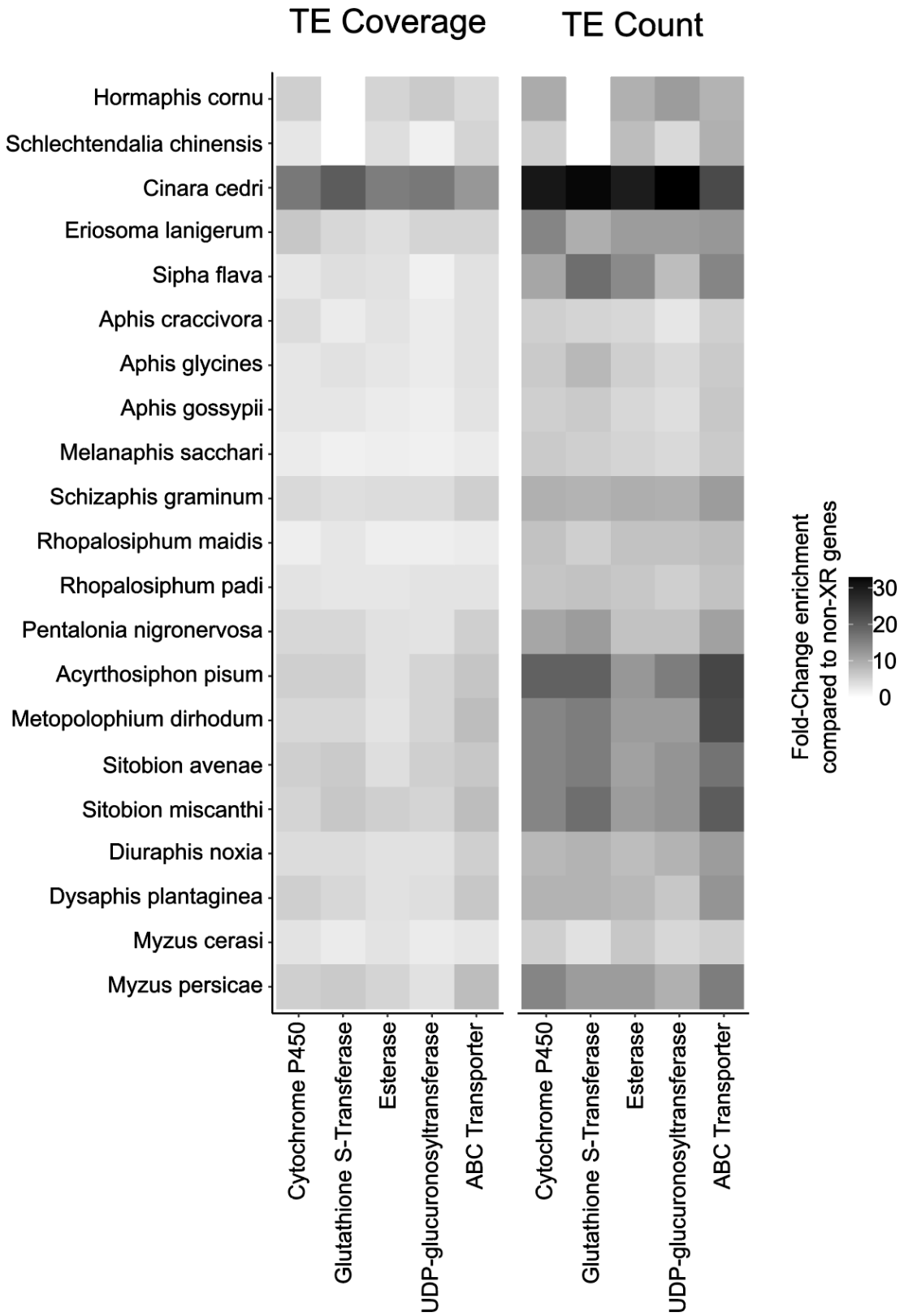

Figure 5. TE enrichment at XRL compared to genome-wide levels. Fold-change enrichment for TE coverage and TE count at XRL compared to within-genome levels at non-XR genes across 21 aphid species.

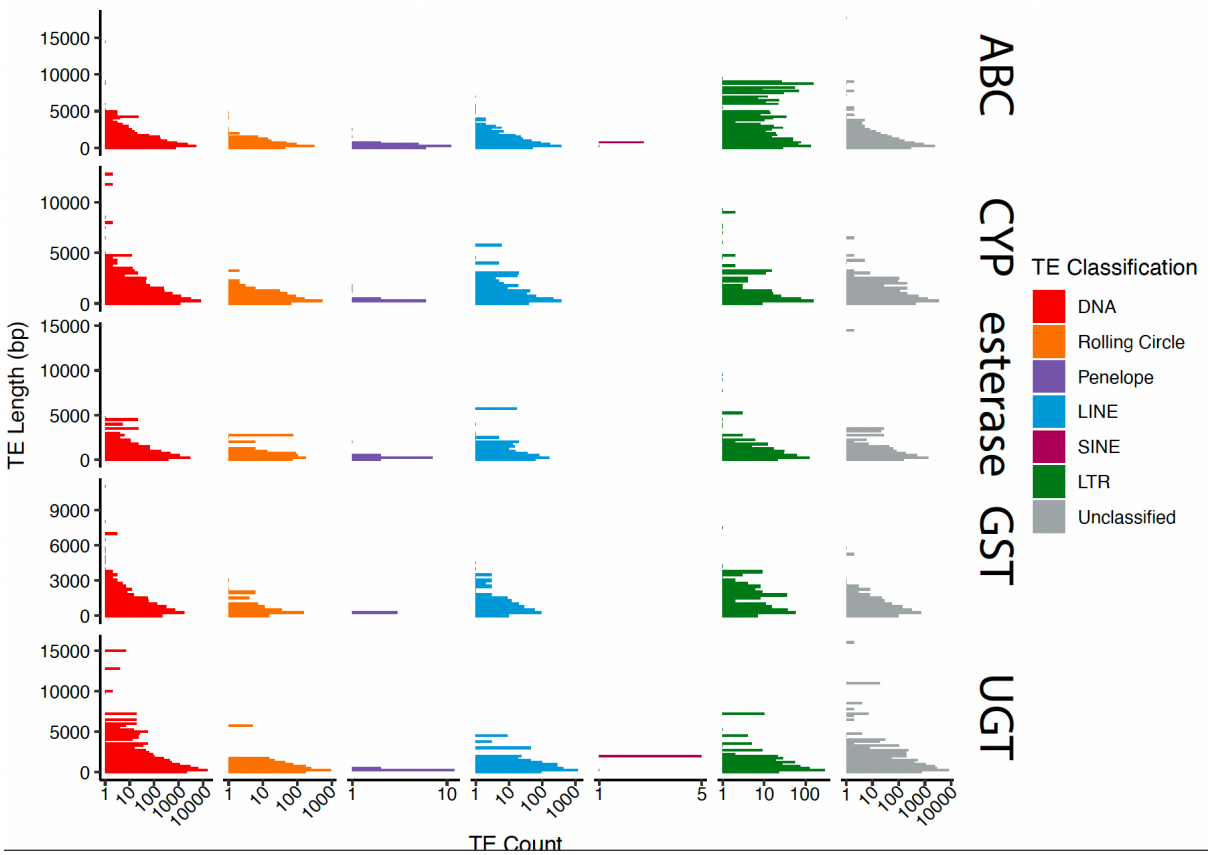

Figure 6. Fragmented TEs at XRL across all aphid species, presented as TE count by TE length with a bin width of 250bp. TE count is presented on a $\log_{10}$ scale. The main TE classifications are indicated by the colours indicated in the key. Mean TE length across all TEs at XRL is 444bp.

TEs are enriched at XRL associated with the detoxification of synthetic insecticides in certain aphid species

Cytochrome P450 genes (CYP genes) are a highly important gene family for the detoxification of xenobiotics, including insecticides (Katsavou et al. 2022; Feyereisen 2005). Indeed, increases in the expression of certain CYP genes can provide dramatic increases in insecticide resistance (Nauen et al. 2022). Additionally, several cases of TE interactions with CYP genes have been characterised (Chung et al. 2007; Chen and Li 2007; Daborn et al. 2002; Itokawa et al. 2011; Schmidt et al. 2010; Karunker et al. 2008; Nauen et al. 2022). Therefore, we undertook a targeted analysis to examine evidence for the accumulation of TEs at CYP genes with putative functions in xenobiotic resistance. For this, aphid CYP

366 genes were labelled based on the availability of evidence to support their xenobiotic
367 resistance functions (Katsavou et al. 2022), as some CYP gene family members play other
368 roles, such as hormone breakdown, resulting in the assignment of 2-29 CYP genes per
369 aphid genome to the xenobiotic detoxification category (Supplementary Table 5). In *S. flava*,
370 *C. cedri*, and *M. cerasi*, we find significant enrichment of TEs surrounding CYP genes in
371 CYP gene (sub)families frequently implicated in the breakdown of xenobiotics compared to
372 those with other functions (Figure 7B). However, we found the opposite pattern in *S.*
373 *graminum*, *M. dirhodum*, *D. plantiginea*, and *M. persicae*, where overall there is significant
374 depletion of TEs surrounding CYP genes associated with xenobiotic resistance (Figure 7B).

375 Next, we considered patterns of TE accumulation at CYP genes specifically involved
376 in resistance to synthetic insecticides, rather than resistance to xenobiotics more generally.
377 For this, CYP genes associated with xenobiotic resistance functions were labelled to reflect
378 the compounds to which they confer resistance. CYP genes that confer resistance to
379 insecticides were then considered separately from CYP genes associated with the
380 detoxification of other xenobiotics, such as those offering resistance to naturally occurring
381 plant allelochemicals. We find a significant enrichment of TEs around CYP genes associated
382 with resistance to synthetic insecticides in *P. nigronervosa*, *A. pisum*, *M. cerasi*, and *M.*
383 *persicae* (Figure 7C). Of these four aphid species, the latter three are all species of "*most*
384 *agricultural importance*" (Van Emden and Harrington 2017), and are therefore likely to have
385 experienced particularly high exposure to synthetic insecticides, and extremely strong
386 selection to evolve resistance. The significant enrichment of TEs at XRL associated with
387 resistance to synthetic insecticides in these species is thus consistent with their selective
388 recruitment for resistance evolution. However, no significant enrichment of TEs around CYP
389 genes associated with resistance to synthetic insecticides was found in the other 8 species
390 of "*most agricultural importance*" included in our analysis: *A. craccivora, A. gossypii, D.*
391 *noxia*, *M. dirhodum, R. padi, S. graminum,* and *S. avenae (Van Emden and Harrington*
392 *2017)*.

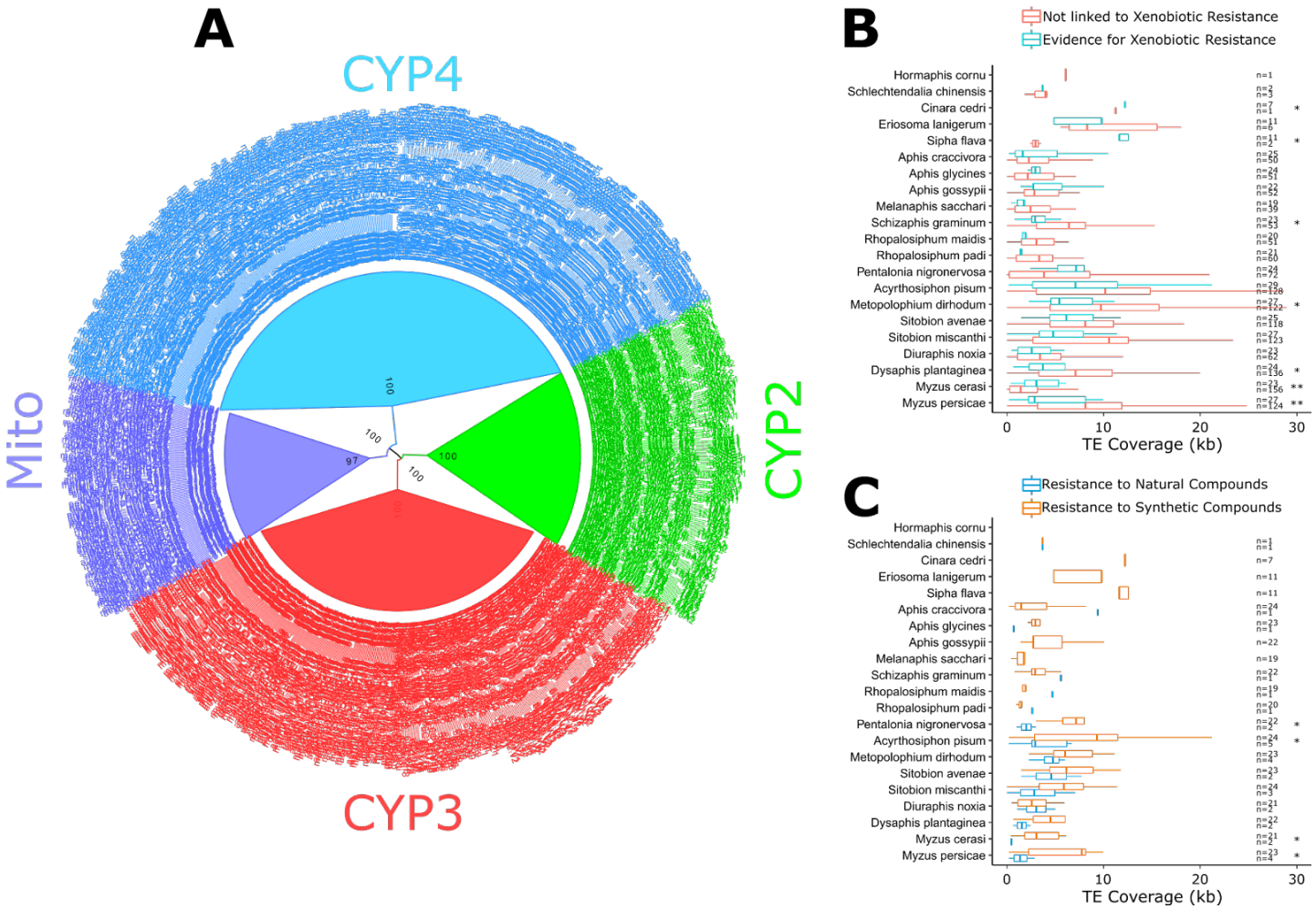

Figure 7. TE association with cytochrome P450 genes in aphids. **A**. Midpoint-rooted phylogeny of CYP genes identified in aphid genomes in this study. Individual genes are coloured by their CYP clan. **B**. TE coverage around CYP genes linked to xenobiotic resistance versus those with other functions. Numbers on the right indicate the number of genes in each group being compared. **C**.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.

399 *Evidence for the selective enrichment of TEs at XRL*

400 We hypothesised that TE enrichment at resistance loci could arise through two
401 distinct processes: (i) Chromatin availability, whereby XRL may be surrounded by open
402 chromatin due to frequent transcriptional activity or constitutive expression, and so represent
403 locations that are more accessible for transposition; (ii) Selective retention of beneficial TE
404 insertions. Given that XRL are essential for the detoxification of harmful plant metabolites
405 and insecticides, and so are of key importance for survival and reproduction, TE insertions
406 that modify XRL in beneficial ways may be selectively retained and so spread through host
407 populations. Consequently, the combined effects of XRL as selective hotspots (especially
408 considering the extremely strong selection mediated by intensive insecticide regimes), and
409 the demonstrated evolutionary potential of TEs, could result in the selective retention of TE
410 insertions surrounding XRL over time. Such effects may be magnified by the influence of
411 host stress under insecticide exposure, which has been implicated in increases in TE activity
412 (Chénais et al. 2012; Stapley et al. 2015; Horváth et al. 2017). Specifically, increased TE
413 activity may result in a higher likelihood of TE insertions occurring at any genomic location,
414 including XRL. However, TE insertions at XRL may be more likely to be selectively retained,
415 if they offer host benefits in the form of resistance mutations, due to their capacity to
416 contribute to host evolvability through diverse genetic mechanisms (Bourque et al. 2018;
417 Rebollo et al. 2012).

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

425 between alternative hypotheses that explain TE enrichment at XRL is, to what extent are
426 XRL expressed in the germline? We addressed this question using data obtained from the
427 *Drosophila* gene expression atlas, FlyAtlas2 (Krause et al. 2022), to examine gene
428 expression in testes and ovaries for XRL and a large panel of housekeeping genes.
429 Housekeeping genes are constitutively expressed in all cell types and so represent a
430 'maximally accessible' case with which to examine TE enrichment (Butte et al. 2001)
431 (Supplementary Table 4). We find that germline housekeeping gene expression is
432 significantly higher compared to XRL considering mean FPKM (fragments per kilobase of
433 transcript per million base pairs sequenced) data derived from FlyAtlas2 (Wilcoxon Rank
434 Sum, $W_{456,51} = 1105$, $p < 0.01$, Figure 8). Consistent with chromatin availability as an
435 explanation for TE enrichment, we find that TEs are significantly enriched at housekeeping
436 genes compared to all other host genes, excluding XRL (TE coverage: Wilcoxon Rank Sum,
437 $W_{573,518117} = 250,570,019$, $p < 0.01$; TE count: Wilcoxon Rank Sum, $W_{573,518117} =$
438 $194,478,652$, $p < 0.01$), to a similar level as that observed for XRL versus all other host
439 genes (Figure 8, Supplementary Figures 3 and 4). However, given that germline expression
440 of XRL is consistently low across all XRL types (Figure 8), we conclude that germline
441 expression cannot explain the observed enrichment of TEs at XRLs. Instead, our findings
442 support the alternative explanation for the enrichment of TEs at XRLs, which is selective
443 retention of TE sequences due to genetic contributions towards the evolution of xenobiotic
444 resistance mutations. There is no mechanistic basis to suggest that TEs specifically target
445 the regions surrounding XRLs as insertion sites, compared to other genomic locations.
446 Instead, under this model TE insertions at XRLs are more likely to be retained and spread
447 through the host population (compared to TE insertions at other loci), due to the combination
448 of their capacity to contribute to host evolvability, and the extremely strong selection
449 pressure imposed by insecticide treatment.

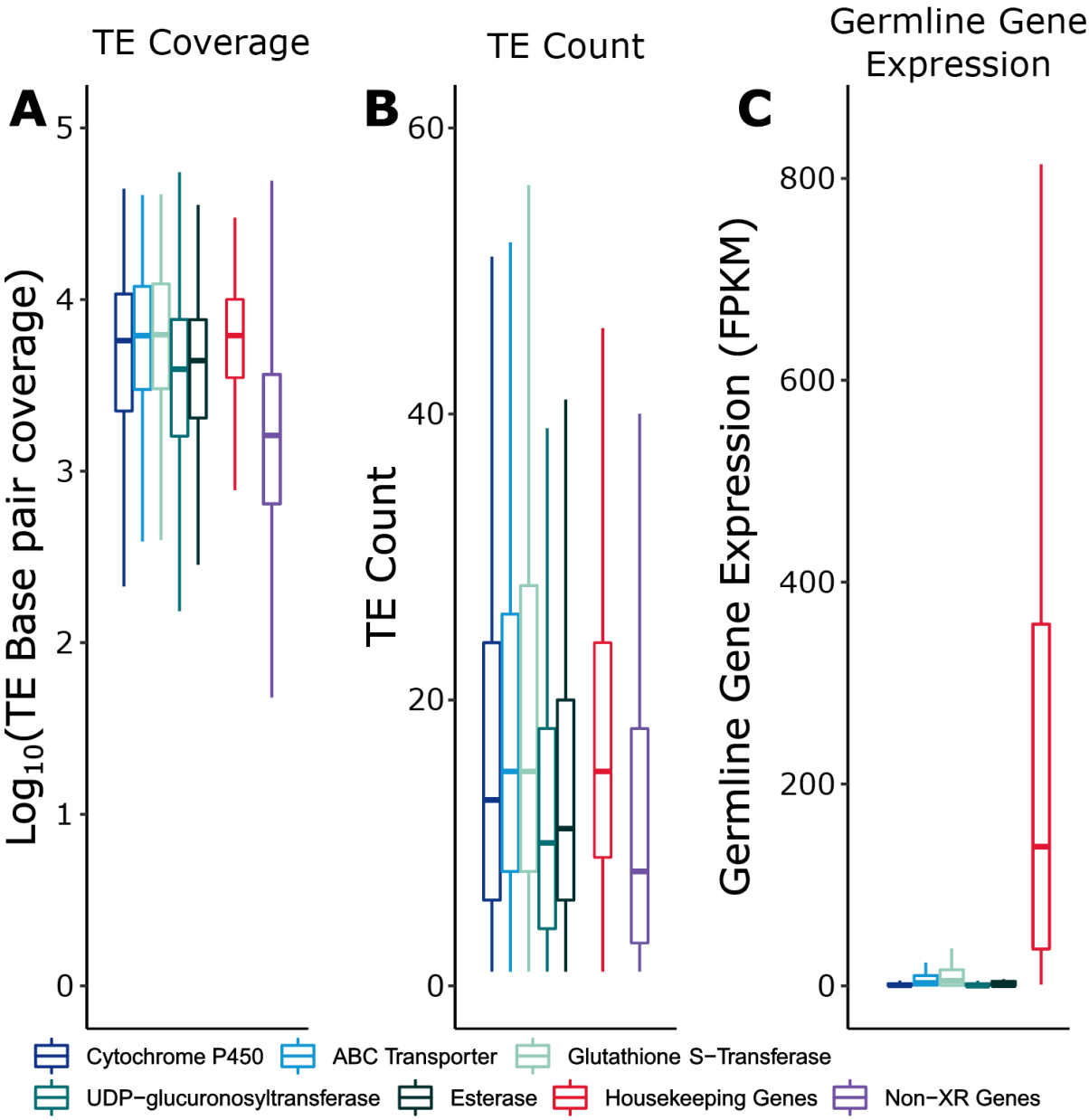

Figure 8. TE occupancy at XRL, housekeeping genes, and all other genes, and expression of XRL and housekeeping genes in *D. melanogaster*. **A**. TE coverage around XRL, housekeeping genes, and all other genes. Box limits indicate upper and lower quartiles, whilst central lines indicate mean TE coverage. **B**. TE count around XRL, housekeeping genes, and all other genes. Box limits indicate upper and lower quartiles, whilst central lines indicate mean TE count. **C**. Expression of XRL and housekeeping genes in germline tissues (testis and ovary) in *D. melanogaster* from FlyAtlas2, expressed as fragments per kilobase of transcript per million base pairs sequenced (FPKM), to determine expected expression of XRL in germline cell types. Gene types are indicated by colours in the key.

463 ## Discussion

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

471 We examine whether genes implicated in xenobiotic resistance are enriched for TE
472 insertions, potentially due to the generation of novel advantageous mutations for resistance
473 evolution. We find an enrichment of TEs, both in terms of TE sequence coverage and
474 number of TE insertions, at XRL across all aphids and all XRL types. Further, we provide
475 evidence that this enrichment is the consequence of selection at XRL, as opposed to
476 elevated insertion rates due to chromatin availability at XRL in the germline.

477 Our results also uncover considerable variability in the enrichment of TEs around
478 XRL both within and among aphid species and XRL types. This demonstrates that TEs are
479 not globally enriched at every XRL, but instead are enriched only at certain XRL and
480 depleted at others. Consequently, the patterns we observe are driven by hotspots of TE
481 enrichment around certain XRL types, and in some cases individual XRL, which show
482 extremely high levels of enrichment compared to other XRL. For example, considering all
483 aphid species examined, we identify significant enrichment of TEs around GSTs. However,
484 this result is driven by a subset of aphids, with 11 of the 21 species showing no evidence of
485 significant enrichment around this XRL class (Supplementary Figure 3). This is consistent
486 with varying levels of selection acting on particular detoxification genes and particular aphid
487 species. Meanwhile, we find markedly more TE hotspots at UGTs and CYP genes, whereby
488 certain XRL are massively enriched for TE fragments. These observations demonstrate the

489 nuances involved in patterns of TE accumulation at XRL, and underline a need for functional
490 validation analyses to test the impact of individual TE insertions on XRL expression.

491 Important crop pests are likely to experience much higher levels of insecticide stress
492 than non-target species, demonstrated by the increased number and diversity of insecticide
493 resistance mutations in target species, compared to non-target species (Whalon et al.).
494 Consequently, variability in TE enrichment is expected among aphid species and XRL,
495 depending on variation in the degree of insecticide exposure. This prediction is confirmed
496 when considering CYP genes, for which data clarifying the relationship to specific
497 insecticides is available (rather than just to xenobiotics in general). Specifically, we find that
498 TE enrichment is significantly higher around CYP genes that are implicated in the
499 detoxification of synthetic insecticides, compared to those involved in the detoxification of
500 natural xenobiotics, but only in three out of ten aphids of "*major agricultural importance*" (Van
501 Emden and Harrington 2017) included in our study (*A. pisum, M. cerasi*, and *M. persicae*).
502 Consequently, either CYP genes are less important for survival in the other species (perhaps
503 due to differences in pest control strategies, such as a greater reliance on forms of non-
504 insecticide based control, for instance the use of biocontrol strategies), or there are genomic
505 mechanisms that limit the co-option of TEs for resistance evolution in these species.

506 Demonstrated contributions of TEs to xenobiotic resistance often involve small
507 fragments of the original TE insertions, and contain motifs that act as host regulatory
508 elements (Singh et al. 2020; Remnant et al. 2013; Panini et al. 2021; Bogwitz et al. 2005;
509 Itokawa et al. 2011; Schmidt et al. 2010; Darboux et al. 2007; Gahan et al. 2001; Yang et al.
510 2007). Meanwhile, the emergence of a resistant phenotype frequently appears to result from
511 a single, or up to a few, TE insertions at each specific XRL, rather than large accumulations
512 of TE DNA (Schmidt et al. 2010; Daborn et al. 2002; Chung et al. 2007; Panini et al. 2021).
513 This presumably occurs due to genomic and selective processes that act to retain the core
514 TE regions that convey a selective benefit, while purging long repetitive and potentially
515 deleterious regions (e.g. that promote non-homologous recombination with similar repetitive

516 elements). Meanwhile, the impacts of these TE insertions can occur via complex multistep
517 processes including the retention of several truncated TE fragments (such as for CYP6G1,
518 Schmidt *et al*. (2010)). Thus, a key consideration is that TE-associated resistance
519 mechanisms go beyond simple enrichment at specific loci.

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

530 Additional data are required to elucidate the potential underlying mechanisms
531 through which the TE sequences identified here may play a role in resistance phenotypes,
532 such as: information on chromatin conformation and the availability of DNA within TE
533 sequences at XRL, screening to investigate if TE sequences contain key transcription factor
534 binding sites, and the results of functional validation experiments. This will ultimately allow
535 the evaluation of the extent to which TEs are utilised in resistance mutations, versus the
536 contributions arising from other classes of genetic variation such as SNPs and CNVs.

537

538 **Materials and Methods**

539 *Estimating Host Phylogeny*

540 The genome assemblies of 21 aphid species, and 7 diverse hemipteran outgroup species
541 (*Ericerus pela*, *Maconellicoccus hirsutus*, *Phenacoccus solenopsis, Bemisa tabaci, Sogatella*
542 *furcifera, Diaphorina citri, Pachypsylla venusta*), were obtained along with gene annotations
543 where available (Supplementary Table 6). To estimate host phylogeny, a supermatrix
544 approach was used. Specifically, for each genome assembly, BUSCO (version 5.2.2) (Manni
545 et al. 2021a, 2021b) was used with the hemiptera_odb10 geneset to identify conserved gene
546 orthologs in each genome assembly. The identifiers for all complete genes were extracted
547 and those present in less than three genomes were removed. The amino acid sequences for
548 each gene from all species were extracted and saved in individual fasta files. For each gene,
549 sequences were aligned using MAFFT (version 7.453) (Katoh and Standley 2013) (--auto).
550 The subsequent alignments were concatenated using Phykit create_concat (Steenwyk et al.
551 2021) to generate a supermatrix. To generate the species phylogeny, we performed 1,000
552 ultrafast bootstrap repetitions (-bb 1000) in IQ-tree (version 1.6.12)(Nguyen et al. 2015),
553 using the best fit amino acid model identified by Modelfinder (Kalyaanamoorthy et al. 2017)
554 (Supplementary File 1).

555

556 *Transposable Element Annotation*

557 To curate and annotate TEs, each genome assembly was annotated with the Earl Grey TE
558 annotation pipeline (version 1.3) (https://github.com/TobyBaril/EarlGrey) (Baril et al. 2022).
559 Briefly, known TEs from Arthropoda (-r arthropoda) were first annotated using both the Dfam
560 3.4 (Hubley et al. 2016) and RepBase RepeatMasker Edition (version 20181026) (Jurka et
561 al. 2005; Kapitonov and Jurka 2008) TE databases. Following this, Earl Grey identified *de*

562 *novo* TE families and refined these using an iterative 'BLAST, Extract, Extend' process (Platt
563 et al. 2016). Following final annotation with the combined library of known and *de novo* TE
564 families, annotations are processed to remove overlapping annotations and to defragment
565 annotations likely to originate from the same TE insertion. TE annotation GFF files are
566 provided in Supplementary File 2.

567 To characterise TEs as either shared or species-specific, *de novo* TE libraries from
568 each genome assembly were clustered to the TE family definition of (Wicker et al. 2007),
569 implemented as described by (Goubert et al. 2022) using cd-hit-est (-d 0 -aS 0.8 -c 0.8 -G 0 -
570 g 1 -b 500 -r 1) (Li and Godzik 2006; Fu et al. 2012). Sequences within each cluster were
571 designated a number between 0 and 20 to determine the number of other species that the
572 TE family was shared with, with 0 identifying TEs unique to a species, and 20 identifying TEs
573 shared across all 21 aphid species considered in this study. For known TE families from
574 Dfam and RepBase, each TE was labelled using the same system based on the number of
575 species the TE family was identified in.

576 Genetic distance from TE consensus sequence is used as a proxy for estimated
577 timing of TE activity (i.e TE age). Using genetic distance from TE consensus as a proxy for
578 insertion time should be taken with caution, as this can be influenced by the methodology
579 used to generate TE consensus sequences, and whether known TE sequences from other
580 organisms are used as a reference or purely *de novo* TE consensus sequences from the
581 species of interest. However, this metric can still be used to provide a rough relative estimate
582 of TE activity.

583

584 ### *Heritability of TE Abundance and Diversity*

585 To determine the heritability of TE abundance, expressed as total TE count, and TE
586 diversity, expressed as total number of distinct TE families, we measured the amount of

587 variation explained by phylogenetic relationships using Bayesian phylogenetic mixed models
588 with Markov chain Monte Carlo estimations, with an intercept fitted as a fixed effect and the
589 phylogeny fitted as a random effect, in the MCMCglmm package (Hadfield 2010) in R
590 (version 4.2.1) (Team 2013) using the Rstudio IDE (Team 2015; Racine 2013). A poisson
591 error distribution and log link function were used, and inverse gamma priors were specified
592 for all R and G-side random effects (V = 1, v = 0.002). Models were run for 11,000,000
593 iterations with a burn-in of 1,000,000 and a thinning interval of 1,000. This approach
594 generated 10,000 posterior samples from which posterior mode and 95% CIs were
595 calculated. The proportion of between-species variation explained by phylogeny was
596 calculated from the model using the equation $V_p / (V_p + V_s)$ where V_p and V_s represent the
597 phylogenetic and species-specific components of between-species variance (Freckleton et
598 al. 2002), which is equivalent to phylogenetic heritability.

599

600 *Gene Predictions and Transposable Element Association with Genomic*
601 *Compartments*

602 For genome assemblies lacking a gene annotation GFF file (*A. gossypii*, *H. cornu, M.*
603 *dirhodum, P. nigronervosa, R. padi, S. graminum, S. chinensis, S. avenae, S. miscanthi*),
604 genes were annotated using Augustus (version 3.3.3) (Stanke et al. 2008; Keller et al. 2011)
605 using the pea aphid reference gene set (--species=pea_aphid --strand=both --
606 genemodel=partial --introns=on --exonnames=on --gff3=on).

607 To prepare genome annotation files for intersection with TE loci, each gene
608 annotation GFF was processed to generate a GFF file containing coordinates of exons,
609 introns, 5' and 3' flanking regions, and intergenic space. Gene flanking regions are defined
610 here as 20 kb directly upstream or downstream of the gene body. These flanking regions are
611 determined to identify TEs at a distance that could be described as the proximate promoter

612 region, rather than just accounting for the core promoter region, to include both promoter
613 and more distal enhancer regions for genes (Lenhard et al. 2012). Intergenic space
614 coordinates were generated using bedtools complement (Quinlan and Hall 2010), to
615 generate inverse coordinates to those of genic compartments and flanking regions. To
616 quantify TE occupation in different genomic compartments (introns, exons, 5' and 3' flanking
617 regions, and intergenic space), bedtools intersect was used to calculate overlap (-wao)
618 between all compartments and TEs.

619

620 *Identification of Xenobiotic Resistance Genes and Housekeeping Genes*

621 Two major mechanisms are frequently implicated in conferring xenobiotic resistance in a
622 wide range of arthropods (Li et al. 2007; Kliot and Ghanim 2012): (i) 'target-site resistance',
623 involving structural changes (mutations) in the gene encoding the insecticide target protein,
624 that make it less sensitive to the toxic effect of the insecticide; (ii) 'metabolic resistance',
625 involving the increased production or activity of metabolic enzymes that break down or
626 sequester the insecticide. Regarding the latter, several major gene types are associated with
627 metabolic resistance, which act during three distinct phases of xenobiotic metabolism
628 (Kennedy and Tierney 2013): (i) Cytochrome P450 monooxygenases (CYP genes); (ii)
629 glutathione S-transferases (GSTs), esterases, and UDP-glucuronosyltransferases (UGTs);
630 (iii) ATP-binding cassette transporters (ABC transporters). We briefly summarise these three
631 phases below. During phase I, CYP genes catalyse functionalisation reactions, where
632 hydrophobic xenobiotics are typically converted to more water-soluble metabolites through
633 the addition of functional groups (Kennedy and Tierney 2013). CYP genes are diverse
634 enzymes involved in several purposes from biosynthesis to metabolism, and are sometimes
635 referred to as 'nature's blowtorch', due to their high-valence iron chemistry oxidation
636 mechanism (Guengerich 2009). CYP genes can mediate resistance to all classes of
637 insecticides due to their broad substrate specificity and versatility, whilst they are also

638 involved in several other processes, such as juvenile hormone and ecdysteroid metabolism,
639 and fatty acid biosynthesis (Feyereisen 2005; Kennedy and Tierney 2013; Feyereisen 1999).
640 During phase II, GSTs, UGTs, and esterases catalyse conjugation reactions between phase
641 I substrates and endogenous molecules to form water-soluble metabolites (Kennedy and
642 Tierney 2013). During phase III, ABC transporters, which are cell membrane transport
643 proteins that efflux toxins and modified toxins from the cell, export xenobiotics and products
644 of phase I and II (Gott et al. 2017). In addition, ABC transporters can also block the cellular
645 import of xenobiotics to protect the host, which is termed phase 0. (Gott et al. 2017).
646 Insecticide resistance can arise via mutations acting on the genes involved in the processes
647 described above through various genetic mechanisms, including upregulation, changes to
648 coding sequence, and gene amplification (Li et al. 2007).

649 The gene identities for all ABC transporters, cytochrome P450s, UDP-
650 glucuronosyltransferases, esterases, and glutathione S-transferases were manually curated
651 from the functional annotation of the *M. persicae* G006v2 genome (Singh et al. 2021). Using
652 these identifiers, exon coordinates were obtained from the *M. persicae* gene annotation
653 GFF3 file, and corresponding nucleotide sequences were extracted from the genome
654 assembly. The resulting exon sequences from *M. persicae* were used as queries to identify
655 XRL in all other aphid assemblies using blastn (Camacho et al. 2009) with an e-value
656 threshold of $1x10^{-10}$. Blast hits for each species were manually inspected to extract those
657 that best represented each XRL, with the inclusion of gene duplications where appropriate.
658 Following this, a bed file of XRL coordinates for each species was generated including the
659 annotation of flanking regions as described above (Supplementary File 3).

660 Housekeeping genes (HKGs) are required for basic cellular functions and are
661 typically expressed in all cell types (Butte et al. 2001). Amino acid sequences of 28 HKGs
662 characterised for aphids (Bansal et al. 2012; Koramutla et al. 2016; Li et al. 2020; Kang et al.
663 2017; Yang et al. 2015, 2014) were obtained from GenBank (Benson et al. 2013) for *M.*
664 *persicae* (Supplementary File 4). Subsequently, a blastx search was performed, with default

665 parameters, against each aphid genome to identify HKG loci in the other 20 species. For
666 each species, HKGs were manually curated and a bed file of HKG coordinates, including the
667 annotation of flanking regions, was generated (Supplementary File 5). To quantify TE
668 occupation around XRL and HKGs, bedtools intersect was used to calculate overlap (-wao)
669 between annotated TEs and XRL.

670

671 *Estimation of Aphid Cytochrome P450 Phylogeny*

672 For each aphid species, the sequence of each CYP exon was extracted using the
673 appropriate XRL bed file and the bedtools command getfasta. Each exon was translated in
674 all six reading frames using EMBOSS transeq (Rice et al. 2000). To determine the correct
675 frame for each exon, each exon was used as a query in a blastp search, with default
676 parameters, against the amino acid sequences of *M. persicae* CYP genes from (Dermauw et
677 al. 2020). Blast hits were manually inspected to determine the correct frame for each exon
678 and the amino acid sequences for each gene were appended to generate a single fasta
679 entry for each gene from each species (Supplementary File 6). Sequences were aligned
680 using MAFFT (--auto) and the CYP phylogeny was generated using 1,000 ultrafast bootstrap
681 repetitions (-bb 1000) in IQ-tree, using the best fit amino acid model identified by Modelfinder
682 (Kalyaanamoorthy et al. 2017) (Supplementary File 1). The resultant tree was midpoint
683 rooted and each sequence was annotated based on the CYP clan it clustered with using
684 figtree (http://tree.bio.ed.ac.uk/software/figtree/). Individual CYP genes were also labelled
685 based on their implication in resistance functions, including compounds they confer
686 resistance to, as described by (Katsavou et al. 2022) (Supplementary Table 4).

687

688 Data Access

689 All data generated and used in this study is provided as supplementary files.

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.

Dysaphis plantaginea DNA sequencing data were kindly provided by Amandine Cornille from Université Paris-Saclay.

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

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.

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

746 EW. 2013. GenBank. *Nucleic Acids Res* **41**: D36–42.

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

750 Bogwitz MR, Chung H, Magoc L, Rigby S, Wong W, O'Keefe M, McKenzie JA,
751 Batterham P, Daborn PJ. 2005. Cyp12a4 confers lufenuron resistance in a
752 natural population of Drosophila melanogaster. *Proc Natl Acad Sci U S A* **102**:
753 12807–12812.

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

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

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

763 Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL.
764 2009. BLAST+: Architecture and applications. *BMC Bioinformatics* **10**: 1–9.

765 Catania F, Kauer MO, Daborn PJ, Yen JL, Ffrench-Constant RH, Schlötterer C.
766 2004. World-wide survey of an Accord insertion and its association with DDT
767 resistance in Drosophila melanogaster. *Mol Ecol* **13**: 2491–2504.

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

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

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

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

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

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

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

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

791 Dermauw W, Van Leeuwen T, Feyereisen R. 2020. Diversity and evolution of the
792 P450 family in arthropods. *Insect Biochem Mol Biol* **127**: 103490.

793 Deutsch CA, Tewksbury JJ, Tigchelaar M, Battisti DS, Merrill SC, Huey RB, Naylor
794 RL. 2018. Increase in crop losses to insect pests in a warming climate. *Science*
795 **361**: 916–919.

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

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

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

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

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

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

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

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

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

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

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

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

822 Horváth V, Merenciano M, González J. 2017. Revisiting the Relationship between
823 Transposable Elements and the Eukaryotic Stress Response. *Trends Genet* **33**:
824 832–841.

825 Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, Smit AFA, Wheeler TJ.
826 2016. The Dfam database of repetitive DNA families. *Nucleic Acids Res* **44**:
827 D81–D89.

828 Itokawa K, Komagata O, Kasai S, Masada M, Tomita T. 2011. Cis-acting mutation
829 and duplication: History of molecular evolution in a P450 haplotype responsible
830 for insecticide resistance in Culex quinquefasciatus. *Insect Biochem Mol Biol* **41**:
831 503–512.

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

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

838 Kang Z-W, Liu F-H, Tian H-G, Zhang M, Guo S-S, Liu T-X. 2017. Evaluation of the
839 reference genes for expression analysis using quantitative real-time polymerase
840 chain reaction in the green peach aphid, Myzus persicae. *Insect Sci* **24**: 222–
841 234.

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

844 Karunker I, Benting J, Lueke B, Ponge T, Nauen R, Roditakis E, Vontas J, Gorman
845 K, Denholm I, Morin S. 2008. Over-expression of cytochrome P450 CYP6CM1
846 is associated with high resistance to imidacloprid in the B and Q biotypes of
847 Bemisia tabaci (Hemiptera: Aleyrodidae). *Insect Biochem Mol Biol* **38**: 634–644.

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

850 Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version
851 7: Improvements in performance and usability. *Mol Biol Evol* **30**: 772–780.

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

856 Keller O, Kollmar M, Stanke M, Waack S. 2011. A novel hybrid gene prediction
857 method employing protein multiple sequence alignments. *Bioinformatics* **27**:
858 757–763.

859 Kennedy CJ, Tierney KB. 2013. Xenobiotic Protection/Resistance Mechanisms in
860 Organisms. In *Environmental Toxicology: Selected Entries from the*
861 *Encyclopedia of Sustainability Science and Technology* (ed. E.A. Laws), pp.

862 689–721, Springer New York, New York, NY.

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

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

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

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

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

876 Lenhard B, Sandelin A, Carninci P. 2012. Metazoan promoters: emerging
877 characteristics and insights into transcriptional regulation. *Nat Rev Genet* **13**:
878 233–245.

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

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

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

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

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

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

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

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

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

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

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

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

913 Platt RN, Blanco-Berdugo L, Ray DA. 2016. Accurate transposable element
914 annotation is vital when analyzing new genome assemblies. *Genome Biol Evol*
915 **8**: 403–410.

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

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

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

924 Rebollo R, Romanish MT, Mager DL. 2012. Transposable Elements: An Abundant
925 and Natural Source of Regulatory Sequences for Host Genes. *Annu Rev Genet*
926 **46**: 21–42.

927 Remnant EJ, Good RT, Schmidt JM, Lumb C, Robin C, Daborn PJ, Batterham P.
928 2013. Gene duplication in the major insecticide target site, Rdl, in Drosophila
929 melanogaster. *Proc Natl Acad Sci U S A* **110**: 14705–14710.

930 Rice P, Longden L, Bleasby A. 2000. EMBOSS: The European Molecular Biology
931 Open Software Suite. *Trends Genet* **16**: 276–277.

932 Rostant WG, Wedell N, Hosken DJ. 2012. Transposable Elements and Insecticide
933 Resistance. *Advances in Genetics* 169–201. http://dx.doi.org/10.1016/b978-0-
934 12-394394-1.00002-x.

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

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

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

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

950 Song WY, Pi LY, Bureau TE, Ronald PC. 1998. Identification and characterization of
951 14 transposon-like elements in the noncoding regions of members of the Xa21
952 family of disease resistance genes in rice. *Mol Gen Genet* **258**: 449–456.

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

957 Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically
958 mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**:
959 637–644.

960 Stapley J, Santure AW, Dennis SR. 2015. Transposable elements as agents of rapid
961 adaptation may explain the genetic paradox of invasive species. *Mol Ecol* **24**:
962 2241–2252.

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

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

968 Sultana T, Zamborlini A, Cristofari G, Lesage P. 2017. Integration site selection by
969 retroviruses and transposable elements in eukaryotes. *Nat Rev Genet* **18**: 292–
970 308.

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

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

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

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

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

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

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

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

990 Whalon, Mota-Sanchez, Hollingworth. Arthropod pesticide resistance database.
991 *Michigan State University*.

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

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

997 Yang C, Pan H, Liu Y, Zhou X. 2014. Selection of Reference Genes for Expression
998 Analysis Using Quantitative Real-Time PCR in the Pea Aphid, Acyrthosiphon
999 pisum (Harris) (Hemiptera, Aphidiae). *PLoS ONE* **9**: e110454.
1000 http://dx.doi.org/10.1371/journal.pone.0110454.

1001 Yang C, Pan H, Liu Y, Zhou X. 2015. Temperature and Development Impacts on
1002 Housekeeping Gene Expression in Cowpea Aphid, Aphis craccivora (Hemiptera:
1003 Aphidiae). *PLoS One* **10**: e0130593.

1004 Yang Y, Chen H, Wu Y, Yang Y, Wu S. 2007. Mutated cadherin alleles from a field
1005 population of Helicoverpa armigera confer resistance to Bacillus thuringiensis
1006 toxin Cry1Ac. *Appl Environ Microbiol* **73**: 6939–6944.

1007 Zhao J, Jin L, Yang Y, Wu Y. 2010. Diverse cadherin mutations conferring resistance
1008 to Bacillus thuringiensis toxin Cry1Ac in Helicoverpa armigera. *Insect Biochem*
1009 *Mol Biol* **40**: 113–118.

1010

1011

1012 ## Supplementary File Captions

1013 Supplementary Figure 1. The relationship between TE content and genome size in
1014 aphids. Blue line indicates the relationship calculated using linear regression. Grey
1015 area indicates 95% confidence intervals.
1016
1017 Supplementary Figure 2. Correlation between genic:intergenic base pairs ratio and
1018 TE content in aphids.
1019
1020 Supplementary Figure 3. TE abundance around xenobiotic resistance genes
1021 compared to all other genes and housekeeping genes, split by xenobiotic resistance
1022 gene type.
1023
1024 Supplementary Figure 4. TE coverage around xenobiotic resistance genes
1025 compared to housekeeping genes.
1026
1027 Supplementary File 1. Midpoint-rooted phylogenetic tree of CYP genes identified in
1028 aphids in NEXUS format.
1029
1030 Supplementary File 2. Tar archive containing TE annotations in GFF format for all
1031 aphid genomes in this study.
1032
1033 Supplementary File 3. Tar archive containing XRL coordinates in GFF format for all
1034 aphid genomes in this study.
1035
1036 Supplementary File 4. *Myzus persicae* housekeeping amino acid sequences for each

1037 aphid genome in FASTA format.

1038

1039 Supplementary File 5. Tar archive containing housekeeping gene coordinates in

1040 GFF format for all aphid genomes in this study.

1041

1042 Supplementary File 6. Cytochrome P450 amino acid sequences for each aphid

1043 genome in FASTA format.

1044

1045 Supplementary File 7. Supplementary Tables 1-6, including contents page.

1046