Expansion of a core regulon in specialized metabolism by mobile genetic elements promotes chemical diversity in *Arabidopsis thaliana*

Brenden Barco*, Yoseph Kim**, and Nicole K. Clay*

* Department of Molecular, Cellular & Developmental Biology, Yale University, Kline Biology Tower 734, 219 Prospect St., New Haven, CT 06511

** Hopkins School, 986 Forest Rd, New Haven, CT 06515

Author Contributions: B.B. and N.K.C performed pathogen assays and ChIP-PCR experiments. B.B. and Y.K. profiled accessions and species. B.B. performed all other experiments. B.B. and N.K.C. interpreted the results and wrote the paper.

Abstract

Plant specialized metabolites are ecologically specialized, mostly lineage-specific molecules whose chemical diversity has been exploited by humans for medical, agriculture, and industrial applications. The mechanisms that gave rise to these phenotypic novelties are unclear, particularly those involving the co-option of recently duplicated genes into functional modules. Here, we show that a LINE retrotransposon (*EPCOT3*) is responsible for the recruitment of newly duplicated gene *CYP82C2* into the WRKY33 regulon and the indole-3-carbonylnitrile (ICN) biosynthetic pathway. WRKY33 is an ancient regulator of plant specialized metabolism, functionally conserved since the gymnosperm-angiosperm split over 300 million years ago. Preferred WRKY33 binding sites are carried by *EPCOT3*, which inserted upstream of *CYP82C2* and underwent chromatin remodeling to become an enhancer that coordinately regulates *CYP82C2* gene expression in response to pathogen effectors. The regulatory neofunctionalization of *CYP82C2* gave rise to pathogen-inducible expression of species-specific metabolite 4-hydroxy-ICN, which is required for antibacterial defense in *Arabidopsis thaliana*. Our results suggest that the transposable element *EPCOT3* contributed clade/species-specific innovations to a core regulon that functions as an extended regulon in specialized metabolism and plant innate immunity.

Keywords: specialized/secondary metabolism | regulatory neofunctionalization | LINE retrotransposon | WRKY33

Summary

Plant secondary or specialized metabolites are essential for plant survival in complex environments and collectively number in the hundreds of thousands. The genetics and epigenetics of chemical diversity in plant specialized metabolism remain unclear. Here, we describe an expansion of the core interactions between an ancient transcription factor and its target biosynthetic genes by mobile genetic elements that disseminate transcription factor binding sites in the genome and undergo chromatin remodeling to become transcriptional enhancers. The extended interactions led to the biosynthesis of a species-specific antimicrobial metabolite important for plant survival. Our findings contribute to a growing understanding of chemical innovation, a critically important but poorly understood process in evolutionary biology.

49 Plant specialized metabolites exist as adaptations towards co-evolving biotic and
50 fluctuating abiotic environments. Consequently, plant specialized metabolism is under
51 constant selective pressure towards chemical innovation (Chae *et al*., 2014; Weng *et*
52 *al*., 2012). The evolutionary process of chemical innovation resulted in the collective
53 synthesis of hundreds of thousands of chemically diverse and ecologically specialized
54 metabolites, many of which exhibit narrower taxonomic distributions compared to
55 primary metabolites (Dixon, 2001; Wink, 2003). Plant specialized metabolic diversity is
56 thought to evolve through the co-option of pre-existing genes. This is mainly
57 accomplished via gene duplication of primary or specialized metabolic enzyme-
58 encoding genes and the neofunctionalization of one or both paralogs to produce
59 enzymes with new expression patterns and/or protein functions, including an increased
60 ability to carry out alternate but latent reactions on novel substrates (Ohno, 1970; Force
61 *et al*., 1999; Weng *et al*., 2012).
62
63 To respond appropriately to changing environments, plant specialized metabolism must
64 be highly dynamic and tightly controlled. This is accomplished in large part by
65 interactions between transcription factors (TFs) and the constituent genes in specialized
66 metabolic pathways, which are often organized under common TFs into regulons and
67 thus more co-expressed than those in non-specialized metabolism (Omranian *et al*.,
68 2015). This organization is thought to fine-tune the timing, amplitude, and tissue-specific
69 expression of pathway genes and subsequent metabolite accumulation (Grotewold,
70 2005; Hartmann, 2007; Martin *et al*., 2010; Tohge & Fernie, 2012). However, very little

71 is known about how newly duplicated genes enter into these core regulons to promote
72 specialized metabolic diversity.
73
74 Changes in *cis*-regulatory sequences such as enhancers and promoters are a major
75 driver of phenotypic diversity (Levine and Davidson, 2005; Prud'homme *et al.*, 2007;
76 Wray, 2007; Wittkopp & Kalay, 2012; Rogers *et al.*, 2013) and likely accelerate the
77 capture of newly duplicated biosynthetic genes into regulons. Enhancers, which consist
78 of TF binding sites (TFBSs) that alter the transcription of target genes independent of
79 orientation and distance (Spitz & Furlong, 2012), are most responsible for *cis*-regulatory
80 divergence (Wittkopp & Kalay, 2012). Enhancers are derived either through mutation or
81 transposable element (TE) insertion. TEs can provide TFBSs and other regulatory
82 innovations upon which selection acts to drive their co-option or exaptation as *cis*-
83 regulatory elements. The identification of ancient TE co-option or exaptation events by
84 sequence conservation has led to the hypothesis that TE exaptation events were largely
85 responsible for a rapid transcriptional rewiring of gene regulatory networks in higher
86 plants, mammals, and other vertebrates (de Souza *et al.*, 2013; Henaff *et al.*, 2014).
87
88 Bacteria elicit two primary immune defense modes in plants, Pattern- and Effector-
89 triggered immunity (PTI and ETI) (Jones & Dangl, 2006). Pathogenic bacteria
90 additionally compromise PTI via specific virulence effector proteins (effector-triggered
91 susceptibility, ETS; Jones & Dangl, 2006). PTI involves the extracellular perception of
92 conserved molecules known as microbe-associated molecular patterns (MAMPs),

93 whereas ETI involves the cytosolic perception of effectors. Although ETI results in the
94 formation of more rapid and robust pathogen-specific response (Jones & Dangl, 2006),
95 both result in the ability of naïve host cells to generate, through non-self perception and
96 subsequent transcriptional reprogramming, specialized metabolites necessary for
97 pathogen defense (Hammerschmidt, 1999; Mansfield, 2000; Clay *et al*., 2009).
98
99 Pathogen-inducible specialized metabolites can be synthesized *de novo* in the plant in
100 an active state (phytoalexins) or constitutively as precursors that become activated
101 (phytoanticipins) (VanEtten *et al*., 1994). Glucosinolates are b-thioglucoside
102 phytoanticipins produced by plants in the mustard family (Brassicaceae) (Rodman *et al*.,
103 1998; Mithen *et al*., 2010). In *A. thaliana,* pathogen inoculation or MAMP treatment
104 directs the biosynthesis of tryptophan (Trp)-derived indole glucosinolates to 4-
105 hydroxyindol- and 4-methoxyindol-3-ylmethylglucosionolate (4OH-I3M, 4M-I3M), and
106 triggers the biosynthesis of phytoalexins 3-thiazol-2'yl-indole (camalexin) and 4-
107 hydroxyindole-3-carbonylnitrile (4OH-ICN) (Tsuji *et al*., 1992; Bednarek *et al*., 2009;
108 Clay *et al*., 2009; Rajniak *et al*., 2015). These molecules are critical for defense
109 responses against against *Pseudomonas syringae* pv, *tomato* DC3000 (*Pst*) (Clay *et al*.,
110 2009 Rajniak *et al*., 2015). Unlike 4M-I3M, whose immune function appears conserved
111 across Brassicaceae, camalexin biosynthesis is restricted to the Camelineae tribe of
112 Brassicaceae, thus representing a clade-specific diversification of pathogen-inducible
113 Trp-derived specialized metabolism (Bednarek *et al*., 2011). The phylogenetic
114 conservation of 4OH-ICN biosynthesis has not yet been investigated.

115

116 The indole glucosinolate, camalexin and 4OH-ICN biosynthetic pathways share the
117 conversion of Trp to indole-3-actetaldoxime (IAOx) via the genetically redundant P450
118 monoxygenases CYP79B2 and CYP79B3 (**Fig. 1a**) (Zhao *et al*., 2002; Glawischnig *et*
119 *al*., 2004; Rajniak *et al*., 2015). The camalexin and 4OH-ICN pathways share the
120 conversion of IAOx to indole-3-cyanohydrin (ICY) by partially redundant P450s
121 CYP71A12 and CYP71A13 (**Fig. 1a**) (Nafisi *et al*., 2007; Klein *et al*., 2013; Rajniak *et*
122 *al*., 2015). CYP71A13 and CYP71B15/PAD3 catalyze further reactions, leading to
123 camalexin production, whereas the flavin-dependent oxidase FOX1/AtBBE3/Re-Tox1
124 and P450 CYP82C2 convert ICY to 4OH-ICN (**Fig. 1a**) (Nafisi *et al*., 2007; Böttcher *et*
125 *al*., 2009; Rajniak *et al*., 2015).
126

127 WRKY transcription factors (TFs) are key regulators of PTI and ETI (Eulgem &
128 Somssich, 2007), and of plant specialized metabolism (Schluttenhofer & Yuan, 2015).
129 WRKY33 is the most ancient regulator of plant specialized metabolism; its ortholog in
130 the green alga *Chlamydomonas reinhardtii* may be ancestral to all higher plant WRKYs
131 (Rinerson *et al*., 2015; Schluttenhofer & Yuan, 2015). Predicted WRKY33 orthologs in
132 higher plants regulate the biosynthesis of lineage-specific alkaloids, terpenes, and
133 phenylpropanoids (Schluttenhofer & Yuan, 2015). In *A. thaliana*, WRKY33 directly
134 activates nearly all 4OH-ICN and camalexin pathway genes in response to bacterial
135 flagellin and fungal pathogen *Botrytis cinerea* (Liu *et al*., 2015, Birkenbihl *et al*., 2017)
136 and may directly activate *CYP71A13* in response to *Pst* and *Pst* carrying the pathogen

137 effector *avrRpm1* (*Psta*) (Qiu *et al*., 2008). WRKY TFs bind to the W-box core motif
138 [TTGAC(T/C)] and require additional motifs to encode binding specificity (Rushton *et al*.,
139 2010; Liu *et al*., 2015). WRKY33 has been shown recently to preferentially bind W-
140 boxes that are within 500nt of the motif [(T/G)TTGAAT] (hereafter referred to as the
141 WRKY33-specific motif) in response to *B. cinerea* (Liu *et al*., 2015).
142
143 Here, we show that 4OH-ICN is an *A. thaliana*-specific metabolite that is preferentially
144 metabolized under ETI conditions, displays WRKY33-dependent intraspecific variation,
145 and is directly correlated with the insertion of a WRKY33-binding LINE retrotransposon
146 sequence (*EPCOT3*) upstream of *CYP82C2*. Phylogenetic and epigenetic analyses of
147 related TEs reveal that the preferred WRKY33 binding sequence on *EPCOT3* likely
148 formed pre-insertion, and that chromatin remodeling occurred post-insertion, leading to
149 *EPCOT3*'s co-option as a *CYP82C2* enhancer. *EPCOT3* likely contributed to the
150 expansion of an ancient core regulon in specialized metabolism to accommodate a
151 clade/species-specific innovation within the conserved framework of pathogen-inducible
152 Trp-derived metabolism.
153
154 **Results**
155 **4OH-ICN is specific to ETI**
156 To identify the major Trp-derived specialized metabolites induced under ETI in *A.*
157 *thaliana*, we compared host transcriptional and metabolic responses to the PTI-eliciting
158 MAMP flg22, the PTI/ETS-eliciting pathogen *Pst*, and the ETI-eliciting pathogen *Psta*

159 under similar conditions as those of previous studies (Denoux *et al*., 2008; Clay *et al*.,
160 2009). Both flg22 and *Psta* induced genes involved in 4OH-ICN, camalexin and 4M-I3M
161 biosynthesis, with 4OH-ICN and camalexin biosynthetic genes having a higher level of
162 induction than those of 4M-I3M in *Psta*-inoculated plants (**SI Appendix, Fig. S1a**;
163 Denoux *et al*., 2008). This result is consistent with the largely quantitative differences
164 observed in transcriptional responses between PTI and ETI (Tao *et al*., 2003; Navarro
165 *et al*., 2004). By contrast, the metabolite responses between PTI and ETI differed
166 qualitatively. 4OH-I3M and 4M-I3M were present in uninfected plants and accumulated
167 to modest levels at the expense of parent metabolite I3M in flg22- and *Psta*-inoculated
168 plants (**SI Appendix, Fig. S1b**) (Clay *et al*., 2009). By comparison, ICN, 4OH-ICN, and
169 camalexin were absent in uninfected plants and at trace levels in flg22-inoculated
170 plants. ICN and camalexin accumulated to high levels in *Pst*- and *Psta*-inoculated
171 plants, whereas 4OH-ICN solely accumulated to high levels in *Psta*-challenged plants
172 (**Fig. 1b; SI Appendix, Fig. S1c**). These results suggest that 4OH-I3M, 4M-I3M,
173 camalexin, and ICN are synthesized in response to multiple PTI elicitors, whereas 4OH-
174 ICN biosynthesis is specific to ETI.
175

176 ***WRKY33* required and sufficient to activate 4OH-ICN**

177 4OH-ICN biosynthetic genes are highly co-expressed with each other (Rajniak *et al*.,
178 2015) and with camalexin biosynthetic genes (**SI Appendix, Fig. S1d**), which are in the
179 WRKY33 regulon (Qiu *et al*., 2008; Birkenbihl *et al*., 2012). To determine whether 4OH-
180 ICN biosynthetic genes are also in the WRKY33 regulon, we compared camalexin, ICN

181 and 4OH-ICN levels between wild-type and a *wrky33* loss-of-function mutant that
182 encodes two differently truncated proteins (**Fig. 2a**; Zheng *et al*., 2006). Consistent with
183 a previous report (Qiu *et al*., 2008), *wrky33* was impaired in camalexin biosynthesis in
184 response to *Psta* and *Pst* carrying the ETI-eliciting pathogen effector *avrRps4* (*Pst*
185 *avrRps4*) (**Fig. 2b; SI Appendix, Fig. S2a**). The *wrky33* mutant was similarly impaired
186 in 4OH-ICN biosynthesis (**Fig. 2b; SI Appendix, Fig. S2a**). These results indicate that
187 WRKY33 is required for camalexin and 4OH-ICN biosynthesis in response to multiple
188 ETI elicitors.
189
190 To investigate whether *WRKY33* is sufficient to activate the 4OH-ICN pathway, we used
191 a two-component glucocorticoid-inducible system to generate *wrky33* plants that in the
192 presence of the glucocorticoid hormone dexamethasone (dex) express a wild-type copy
193 of WRKY33 with a C-terminal fusion to 1x flag epitope (*wrky33/DEX:WRKY33-flag*; **SI**
194 **Appendix, Figs. S2b-c**). Induced expression of *WRKY33-flag* restored camalexin and
195 4OH-ICN biosynthesis in *Psta*-challenged *wrky33* plants to greater than wild-type levels
196 (**SI Appendix, Fig. S2d**). These results indicate that *WRKY33* is required and sufficient
197 to activate camalexin and 4OH-ICN biosynthesis.
198
199 **Intraspecific variation in *WRKY33* affects 4OH-ICN and immunity**
200 Intraspecific variation in TFs can contribute to gain or loss of phenotypes, such as
201 branching in maize (Studer *et al*., 2011) or pelvic loss in three-spined stickleback fish
202 (Chan *et al*., 2010). In addition, the wide variation in camalexin biosynthesis reported

203 among natural accessions of *A. thaliana* (Kagan & Hammerschmidt, 2002) suggests
204 that a similar variation in 4OH-ICN biosynthesis may exist. To identify additional
205 transcriptional activators of 4OH-ICN biosynthesis that otherwise might be refractory to
206 traditional genetic approaches, we compared intraspecific variation in *Psta*-induced
207 camalexin, ICN and 4OH-ICN among 35 re-sequenced accessions and *wrky33* (Col-0
208 accession). We found camalexin and 4OH-ICN levels to be positively correlated among
209 accessions (R^2 = 0.37**; SI Appendix, Fig. S2e**), lending further support to their co-
210 regulation by WRKY33. Accession Dijon-G (Di-G) was identified to produce less
211 camalexin and 4OH-ICN and more ICN than its near-isogenic relatives, the Landsberg
212 accessions L*er*-0 and L*er*-1 (**Fig. 2b; SI Appendix, Figs. S2e-f**). In addition, differences
213 observed in the metabolite response between Landsberg accessions and Di-G most
214 closely resembled those between Col-0 and *wrky33* mutant (**Fig. 2b, SI Appendix, Fig.**
215 **S2e**). These results led us to hypothesize that genetic variation in a regulatory gene, as
216 opposed to an immune signaling gene, is responsible for the metabolite phenotypes
217 observed in Di-G. To test this hypothesis, genetic variation between Di-G and three
218 sequenced Landsberg accessions (La-0, L*er*-0, and L*er*-1) were used to identify 354
219 genes that were differentially mutated to high effect in Di-G (**SI Appendix, Fig. S2g**).
220 Twenty-eight of these mutated Di-G genes were annotated by Gene Ontology to have
221 roles in defense, including *WRKY33* (**SI Appendix, Table S1**). We confirmed by Sanger
222 sequencing that Di-G *WRKY33* harbors a nonsense mutation early in the N-terminal
223 DNA-binding motif (**Fig. 2a**), likely abolishing protein function. Our findings indicate that
224 camalexin and 4OH-ICN are sensitive to intraspecific variation in *WRKY33*.

225

226 *Pst* infection reduces plant fitness (Kover & Schaal, 2002), and camalexin and 4OH-ICN
227 promote plant fitness by contributing non-redundantly to disease resistance to *Pst*
228 (Rajniak *et al*., 2015). To confirm that disease resistance to *Pst* is also sensitive to
229 intraspecific variation in *WRKY33*, we measured bacterial growth in adult leaves of
230 *wkry33* and Di-G and their respective (near-)isogenic accessions Col-0 and L*er*-1.
231 *wrky33* and Di-G were more susceptible to *Pst* than their (near)isogenic relatives and
232 comparable to the 4OH-ICN biosynthetic mutant *cyp82C2* (**Fig. 2c**; Rajniak *et al*.,
233 2015). In addition, induced expression of *WRKY33-flag* restored wild-type levels of
234 resistance in *wrky33* (**Fig. 2c**). Together, our results support a specific role of *WRKY33*
235 in antibacterial defense as an activator of Trp-derived specialized metabolism.
236

237 **WRKY33 activates 4OH-ICN biosynthesis**

238 To confirm that the 4OH-ICN biosynthetic pathway is in the WRKY33 regulon, we first
239 compared *WRKY33*, *CYP71A13, CYP71B15, FOX1* and *CYP82C2* transcript levels
240 among WT, *wrky33*, and *wrky33/DEX:WRKY33-flag*. Consistent with previous reports
241 (Qiu *et al*., 2008), *CYP71A13* and *CYP71B15* expression was down-regulated in *wrky33*
242 plants in response to *Psta* and upregulated in *wrky33/DEX:WRKY33-flag* (**SI Appendix,**
243 **Fig. S3a**). Similarly, *FOX1* and *CYP82C2* expression were unchanged or down-
244 regulated in *wrky33* plants in response to *Psta*, and significantly upregulated in
245 *wrky33/DEX:WRKY33-flag* (**Fig. 3a**). Together with our metabolite analysis, these

246 findings indicate that WRKY33 mediates camalexin and 4OH-ICN biosynthesis in
247 response to pathogen effectors.
248
249 We then tested for WRKY33 binding to W-box-containing regions upstream of
250 camalexin and 4OH-ICN biosynthetic genes in dex-treated and *Psta*-infected
251 *wrky33/DEX:WRKY33-flag* seedlings by chromatin immunoprecipitation (ChIP)-PCR.
252 WRKY33 has been shown to bind to a W-box region upstream of *CYP71A12* (Birkenbihl
253 *et al*., 2017), a region that also contains three WRKY33-specific motifs and is consistent
254 with WRKY33's reported binding site preference (Liu *et al*., 2015). We additionally
255 observed that *Psta*-induced WRKY33 bound strongly (greater than 5-fold enrichment) to
256 a single W-box region upstream of *FOX1* and *CYP82C2* (W2 and W4, respectively;
257 **Figs. 3b-c; SI Appendix, Fig. S3b**). Both regions also contain one to three WRKY33-
258 specific motifs. Together with our expression analysis, our findings indicate that
259 WRKY33 uses preferred WRKY33 binding sites to directly activate 4OH-ICN
260 biosynthetic genes in response to pathogen effectors.
261
262 Interestingly, *Psta*-induced WRKY33 did not bind to the W5 region upstream of
263 *CYP82C2* (**Fig. 3c**), a W-box region that does not contain any WRKY33-specific motifs
264 and is just upstream of neighboring gene of unknown function *At4g31960* (**Fig. 3b**).
265 WRKY33 reportedly binds to W5 in response to flg22 and *B. cinerea* (Liu *et al*., 2015;
266 Birkenbihl *et al*., 2017). By contrast, *Psta*-induced WRKY33 bound strongly to W1
267 region upstream of *CYP71B15* (**SI Appendix, Fig. S3c-d)**, a W-box region that also

268 does not contain any WRKY33-specific motifs. WRKY33 reportedly binds to a region
269 encompassing W1 in response to flg22 and *Psta* (Qiu *et al*., 2008; Birkenbihl *et al*.,
270 2012). These findings suggest that WRKY33 may use W-box extended motifs or novel
271 specificity motifs to target camalexin biosynthetic genes in response to pathogen
272 effectors, or 4OH-ICN biosynthetic genes in response to MAMPs or fungal pathogens.
273

274 ***CYP82C2* underwent regulatory neofunctionalization**

275 CYP82C2 catalyzes the last step in 4OH-ICN biosynthesis, hydroxylating ICN to form
276 4OH-ICN (Rajniak *et al*., 2015), and likely was the last 4OH-ICN pathway gene to be
277 recruited to the WRKY33 regulon in *A. thaliana*. To explore the phylogenetic distribution
278 pattern of 4OH-ICN biosynthesis, we profiled ICN and 4OH-ICN metabolites in close
279 and distant relatives of *A. thaliana* in response to *Psta*. While ICN biosynthesis was
280 observed across multiple close relatives, 4OH-ICN was only detected in *A. thaliana*
281 (**Fig. 4a; SI Appendix, Fig. S4a**). This result suggests that 4OH-ICN manifests a
282 species-specific diversification of pathogen-inducible Trp-derived metabolism in the
283 mustard family.
284

285 In *A. thaliana*, *CYP82C2* resides in a near-tandem cluster with paralogs *CYP82C3* and
286 *CYP82C4* (**Fig. 4b**). We performed phylogenetic and syntenic analyses to identify
287 putative *CYP82C2* orthologs in ICN-synthesizing species. All identified homologs are
288 syntenic to *CYP82C2* or *CYP82C4*, and encode proteins with >88% identity to one
289 another (**Fig. 4b; SI Appendix, Figs. S4b-c**). *CYP82C3* is present only in *A. thaliana*,

290 and although more similar to *CYP82C2* than *CYP82C4* in sequence, it is not functionally
291 redundant with *CYP82C2* (**Fig. 4b; SI Appendix, Fig. S4b**; Rajniak *et al*., 2015).
292 *CYP82C4* is required for the biosynthesis of sideretin, a widely-conserved
293 phenylalanine-derived metabolite required for iron acquisition (Rajniak *et al*., 2018).
294 *CYP82C4* has syntenic orthologs in the mustard family, correlating with the distribution
295 of sideretin biosynthesis (**Fig. 4b; SI Appendix, Fig. S4b;** Rajniak *et al*., 2018). By
296 contrast, *CYP82C2* has syntenic orthologs only within the *Arabidopsis* genus (**Fig. 4b;**
297 **SI Appendix, Fig. S4b**). These results suggest that *CYP82C2* duplicated from
298 *CYP82C4* prior to the formation of the *Arabidopsis* genus and then acquired a new
299 expression pattern and/or catalytic function prior to *A. thaliana* speciation approx. 2
300 million years later (Hu *et al*., 2011; Hohmann *et al*., 2015).
301
302 CYP82C2 and CYP82C4 were previously characterized to 5-hydroxylate with equal
303 efficiency the specialized metabolite 8-methoxypsoralen, a molecule structurally
304 reminiscent of ICN and sideretin (Kruse *et al*., 2008). The apparent similarities in
305 substrate specificity and catalytic function suggest that *CYP82C2* may have diverged
306 from *CYP82C4* in expression but not protein function. To test this, we first compared the
307 expression of *CYP82C2* and *CYP82C4* in *A. lyrata* and *A. thaliana* in response to *Psta*.
308 4OH-ICN biosynthetic genes *CYP79B2, CYP71A12* and *FOX1* were upregulated in both
309 species, consistent with the common presence of ICN (**Figs. 4a,c**). By contrast,
310 *CYP82C2* levels were respectively upregulated and unchanged in *A. thaliana* and *A.*
311 *lyrata*, correlating with the distribution of 4OH-ICN in these species (**Figs. 4a,c**).

312 *CYP82C4* expression was unchanged in both species (**Fig. 4c**). These results indicate
313 that 4OH-ICN biosynthesis is linked with pathogen-induced expression of *CYP82C2.*
314
315 We then compared the aligned upstream sequences of *CYP82C2* and *CYP82C4* in *A.*
316 *lyrata* and *A. thaliana* and observed good sequence conservation among orthologs but
317 poor conservation among paralogs (**SI Appendix, Fig. S4d**), indicating that sequences
318 upstream of *CYP82C4* and *CYP82C2* were independently derived. We performed
319 expression analysis in *A. thaliana* to confirm that *CYP82C2* and *CYP82C4* have
320 different expression patterns. *CYP82C2* expression is upregulated in response to *Psta*
321 and unchanged under iron deficiency (**Figs. 4c-d; SI Appendix, Fig. S1a;** Rajniak *et*
322 *al*., 2015). Conversely, *CYP82C4* is upregulated under iron deficiency and unchanged
323 in response to *Psta* (**Figs. 4c-d**; Murgia *et al*., 2011; Rajniak *et al*., 2018). Finally,
324 *CYP82C4* was unchanged in *Psta*-challenged *wrky33* and *wrky33/DEX:WRKY33-flag*
325 (**SI Appendix, Fig. S4e**). Our findings suggest that *CYP82C2* diverged from *CYP82C4*
326 by acquiring WRKY33 regulation for its pathogen-induced expression.
327
328 We next assessed dN/dS ratios along branches of the CYP82C phylogenetic tree (**SI**
329 **Appendix, Fig. S4b**) and found good support for purifying selection acting on CYP82C
330 enzymes (ω=0.21), and no support for positive selection acting on CYP82C2/3 enzymes
331 (**SI Appendix, Table S2**). Lastly, we identified non-conserved amino acid residues
332 among CYP82C homologs and mapped this information onto a homology model of
333 CYP82C2. The protein inner core, which encompasses the active site and substrate

334 channel, is highly conserved among CYP82C homologs (**SI Appendix, Fig. S4g**), and
335 is consistent with CYP82C2 and CYP82C4's reportedly redundant catalytic functions
336 (Kruse *et al*., 2008). Altogether, our findings suggest that *CYP82C2* underwent
337 regulatory neofunctionalization (Moore & Purugganan, 2005), diverging from *CYP82C4*
338 in expression but not protein function.
339

340 **TE *EPCOT3* is an *CYP82C2* enhancer**

341 WRKY33 regulation of *CYP82C2* is mediated by a WRKY33 TFBS in the W4 region
342 (**Figs. 3, 5a; SI Appendix, Fig. S3c)**. Preferential WRKY33 binding at this region
343 should also be influenced by chromatin features associated with *cis*-regulatory elements
344 like enhancers and basal promoters (Slattery *et al*., 2014). To investigate how
345 *CYP82C2* acquired WRKY33 binding for its pathogen-induced expression, we
346 compared the aligned upstream sequences of *CYP82C* homologs in ICN-synthesizing
347 species. We observed three large upstream sequences specific to *A. thaliana*
348 *CYP82C2*, hereafter named ***E**ighty-two-C2 **P**romoter **C**ontained **O**nly in A. **T**haliana1-3*
349 (*EPCOT1-3*; **Fig. 5a**). *EPCOT3* in particular is a 240nt region that completely
350 encompasses W4 (**Fig. 5a**), indicating that the WRKY33's regulation of *CYP82C2* in
351 response to *Psta* may be species-specific. Further bioinformatics analysis revealed that
352 *EPCOT3* has the epigenetic signature of an active enhancer (Roudier *et al*., 2011; Liu *et*
353 *al*., 2018). Relative to neighboring sequences, *EPCOT3* is enriched with activating
354 histone mark H3K4me2 and lacks the repressive histone mark H3K27me3 (**Fig. 5b**)
355 (Heintzman *et al*., 2007; Hoffman *et al*., 2010; Roudier *et al*., 2011; Bonn *et al*., 2012;

356 Wang *et al*., 2014). Our findings suggest that *EPCOT3* functions as an enhancer that
357 mediates WRKY33 binding and activation of *CYP82C2* in response to pathogen
358 effectors.
359
360 *EPCOT3* contains a 3' poly-A tail and is flanked by variable-length target site
361 duplications (**Fig. 5c; SI Appendix, Fig. S5a**), which are hallmarks of eukaryotic LINE
362 retrotransposons (Malik *et al*., 1999). LINE retrotransposition (reverse transcription and
363 integration) results in frequent 5'-truncation of retrocopies (Luan *et al*., 1993). We
364 identified eleven variably truncated retrocopies similar to *EPCOT3* throughout the
365 genome, including *Ta22*, one of the first LINEs characterized in *A. thaliana* (**Fig. 5c**; **SI**
366 **Appendix, Figs. S5a-b, Table S3**; Wright *et al*., 1996). *EPCOT3*-related LINEs were
367 sorted into two groups roughly correspondent to their phylogenetic placement:
368 *EPCOT3-LIKE* (*EPL*) for those with high identity (>65%) to *EPCOT3* and *Ta22* or *Ta22-*
369 *LIKE* (*Ta22L*) for the remainder (**SI Appendix, Fig. S5a; Table S3**). Only *Ta22* and
370 *Ta22L1* are full-length LINEs (**Fig. 5c**), presumably encoding the proteins necessary for
371 their own transposition and for the transposition of nonautonomous family members like
372 *EPCOT3*. We also identified two syntenic species-specific *Ta22L*s, but no *EPLs*, in *A.*
373 *lyrata* (**SI Appendix, Table S3**). Given the 80% overall sequence identity between *A.*
374 *thaliana* and *A. lyrata* (Hu *et al*., 2011), this data indicates that *EPCOT3* and *EPLs*
375 arose from retrotransposition following the speciation of *A. thaliana.*
376

377 Of all the retrocopies, *EPL1* is most similar to *EPCOT3* (85.4% identity), sharing the W-
378 box and WRKY33-specific motif, whereas *EPL2* is less similar (67%) and lacks the
379 WRKY33-specific motif (**Fig. 5c; SI Appendix, Table S3, Fig. S5a**). *EPL1* and *EPL2*
380 are much less truncated than *EPCOT3* (**Fig. 5c**), and lack epigenetic signatures typical
381 of *cis*-regulatory sequences (**SI Appendix, Fig. S5c**) (Roudier *et al*., 2011; Liu *et al*.,
382 2018). To investigate whether the sequence information and chromatin features
383 associated with *EPL*s are sufficient for WRKY33 binding, we tested for WRKY33 binding
384 to *EPL* sequences homologous to the W4 region of *EPCOT3* in dex-treated, *Psta*-
385 infected *wrky33/DEX:WRKY33-flag* plants by ChIP-(q)PCR. Compared to *EPCOT3*
386 (**Fig. 3c**), WRKY33 respectively bound weakly or not at all to *EPL1* and *EPL2* (**Fig. 5d;**
387 **SI Appendix, Fig. S5d**). Our findings suggest the following history: (1) *EPL1* likely
388 retroduplicated from *EPL2* or its progenitor, which already contained a W-box; (2) *EPL1*
389 then acquired a WRKY33-specific motif by mutation; (3) *EPCOT3* likely retroduplicated
390 from *EPL1* and then acquired epigenetic signatures of an enhancer, thereby allowing
391 selection to act on standing variation rather than *de novo* mutation for *CYP82C2*
392 recruitment into the 4OH-ICN biosynthetic pathway.
393

394 **Discussion**

395 TEs were originally conceived to act as “controlling elements” of several loci in the
396 genome (McClintock, 1956), and are now well understood to provide TFBSs and other
397 innovations upon which selection acts to drive their exaptation as *cis*-regulatory
398 elements. The identification of ancient TE-mediated exaptation events by sequence

399 indicates TEs were largely responsible for the rapid transcriptional rewiring of gene
400 regulatory networks in many eukaryotes (de Souza *et al*., 2013). Relatively recent
401 exaptations of TEs into enhancers of host genes have been described for the beta-like
402 globin and IFNL1 genes in primates, Cyp6g1 in fruit flies, and tb1 in maize (Pi *et al*.,
403 2004; Thomson *et al*., 2009; Schmidt *et al*., 2010; Studer *et al*., 2011; de Souza *et al*.,
404 2013). However, it is unclear whether TE exaptations contribute to physiological
405 changes that result in interspecific variation/innovation and fitness benefit (de Souza *et*
406 *al*., 2013). In this study, we show that *EPCOT3* is a TE-derived enhancer that mediates
407 WRKY33 binding and activation of *CYP82C2,* leading to 4OH-ICN biosynthesis and
408 increased disease resistance to a bacterial pathogen. This is to date the first report of a
409 recent TE exaptation event that resulted in a clade-specific expansion of a core regulon
410 in plant specialized metabolism.
411
412 Although the *EPL1/EPCOT3* progenitor retrotransposed a preferred WRKY33 TFBS in
413 the form of *EPCOT3* upstream of *CYP82C2*, a further series of epigenetic modifications
414 were needed to facilitate optimal access of *EPCOT3* by WRKY33. *EPL1* exists in a
415 silenced heterochromatin state (**SI Appendix, Fig. S5c**), typical for TEs (Slotkin &
416 Martienssen, 2007), and is bound weakly by WRKY33 (**Fig. 5d**), whereas *EPCOT3* is in
417 an open chromatin state (**Fig. 5b**; Roudier *et al*., 2011; Liu *et al*., 2018) and bound
418 strongly by WRKY33 (**Fig. 3c**). The more severe 5'-truncation of *EPCOT3* could
419 account for its release from TE silencing mechanisms, and the initially weak WRKY33
420 binding could provide a ‘seed’ for chromatin remodelers to drive the exaptation of newly

421 retrotransposed *EPCOT3* into a *bona fide* enhancer. Further epigenomic sampling
422 within *Arabidopsis* is needed to better clarify epigenetic transformations underlying the
423 *EPCOT3* exaptation event.
424
425 Accession Di-G is closely related to Landsberg accessions (SI Appendix, Figs. S2e-g;
426 Hardtke *et al*., 1996), but synthesizes less camalexin and 4OH-ICN (**Fig. 2b**; Kagan &
427 Hammerschmidt, 2002), is more susceptible to a range of bacterial and fungal
428 pathogens (**Fig. 2c**) (Hugouvieux *et al*., 1998; Kagan & Hammerschmidt, 2002;
429 Mukherjee *et al*., 2009), and is more sensitive to ethylene phytohormone (Chatfield *et*
430 *al*., 2008). WRKY33 has been implicated in camalexin biosynthesis (Qiu *et al*., 2008),
431 antifungal defense (Zheng *et al*., 2006), and ethylene biosynthesis (Li *et al*., 2012). We
432 identified WRKY33 as causal for some if not all of these phenotypes in Di-G. This is the
433 first report of WRKY33's involvement in antibacterial defense and is consistent with the
434 contribution of camalexin and 4OH-ICN towards antibacterial defense (Rajniak *et al*.,
435 2015).
436
437 WRKY33 is an ancient transcription factor responsible for many fitness-promoting traits
438 in plants, thus it is unexpected that an *A. thaliana* accession would have a naturally
439 occurring *wrky33* mutation (C536T transversion). Di-G is the sole member of 1,135
440 sequenced accessions to have a high-effect single nucleotide polymorphism (SNP) in
441 *WRKY33* (1001 Genomes Consortium, 2016). Di-G and L*er*-0 have long been models
442 for studies in mutagenesis (Rédei, 1962, Müller, 1966), and thus a possibility exists that

443 Di-G may have originated from an ethyl methanesulfonate (EMS) mutagenesis screen
444 of L*er*-0. Historical EMS mutagenesis experiments generated upwards of tens of
445 thousands of mutations per cell (Müller 1966; Rédei & Koncz, 1993; Camara *et al*.,
446 2000), well within the range of ~25,000 SNPs that are not concordant between Di-G and
447 L*er*-0 (**SI Appendix, Fig. S2f**). These findings open the possibility that Di-G is not a
448 natural accession but an artificially-derived L*er*-0 *wrky33* mutant.
449

450 **METHODS**

451 Details of plant materials and growth conditions, plant binary vector construction and
452 transformation, bacterial infection assays, RNA extraction and qPCR analysis,
453 extraction and LC-DAD-MS analysis of indole phytoalexins, extraction and LC-DAD-
454 FLD-MS analysis of glucosinolates, chromatin immunoprecipitation and (q)PCR,
455 phylogenetic analysis and bioinformatics analysis can be found in **SI Appendix**.
456

457 **ACKNOWLEDGEMENTS**

458 We thank E.S. Sattely for ICN/ICN-ME, 4OH-ICA/4OH-ICA-ME and camalexin
459 standards. This work was supported by T32-GM007499 (to B.B.) and
460 Elsevier/Phytochemistry Young Investigator Award (to N.K.C.).
461

462 **REFERENCES**

463 1001 Genomes Consortium (2016) 1,135 Genomes Reveal the Global Pattern of
464 Polymorphism in Arabidopsis thaliana. Cell, 166(2), 481-491.

465

466 Birkenbihl RP, Diezel C, Somssich, IE (2012) Arabidopsis WRKY33 is a key
467 transcriptional regulator of hormonal and metabolic responses toward Botrytis cinerea
468 infection. Plant Physiology, 159(1), 266-285.

469

470 Birkenbihl RP, Kracher B, Somssich, IE (2017) Induced Genome-Wide Binding of Three
471 Arabidopsis WRKY Transcription Factors during Early MAMP-Triggered Immunity. The
472 Plant Cell, 29(1), 20-38.

473

474 Bonn S, *et al*. (2012) Tissue-specific analysis of chromatin state identifies temporal
475 signatures of enhancer activity during embryonic development. Nat Genet 44(2): 148-
476 156.

477

478 Böttcher C, *et al*. (2009) The multifunctional enzyme CYP71B15 (PHYTOALEXIN
479 DEFICIENT3) converts cysteine-indole-3-acetonitrile to camalexin in the indole-3-
480 acetonitrile metabolic network of Arabidopsis thaliana. Plant Cell 21(6): 1830-1845.

481

482 Camara, MD, Ancell CA, Pigliucci M (2000) Induced mutations: a novel tool to study
483 phenotypic integration and evolutionary constraints in Arabidopsis thaliana. Evolutionary
484 Ecology Research, 2(8), 1009-1029.

485

486 Chae L, Kim T, Nilo-Poyanco R, Rhee SY (2014) Genomic signatures of specialized
487 metabolism in plants. Science 344: 510-513.
488
489 Chatfield SP, Raizada MM (2008) Ethylene and shoot regeneration: hookless1
490 modulates de novo shoot organogenesis in Arabidopsis thaliana. Plant Cell Reports,
491 27(4), 655-666.
492
493 Chan YF, *et al*. (2010) Adaptive evolution of pelvic reduction in sticklebacks by recurrent
494 deletion of a Pitx1 enhancer. Science 237: 302-305.
495
496 Clay NK, Adio AM, Denoux C, Jander G, Ausubel FM (2009) Glucosinolate metabolites
497 required for an Arabidopsis innate immune response. Science 323: 95-101.
498
499 de Souza, FS. J Franchini, LF Rubinstein M (2013) Exaptation of transposable elements
500 into novel cis-regulatory elements: is the evidence always strong? Mol Biol Evol 30(6),
501 1239-1251.
502
503 Dixon RA (2001) Natural products and plant disease resistance. Nature 411: 843-847.
504
505 Denoux C, *et al*. (2008) Activation of defense response pathways by OGs and Flg22
506 elicitors in Arabidopsis seedlings. Mol Plant 1: 423-445.
507

508 Eulgem T, Somssich IE (2007) Networks of WRKY transcription factors in defense
509 signaling. Curr Opin Plant Biol 10: 366-371.
510
511 Force AM, *et al*. (1999) Preservation of duplicate genes by complementary,
512 degenerative mutations. Genetics 151: 1531-1545.
513
514 Glawischnig E, Hansen BG, Olsen CE, Halkier BA (2004) Camalexin is
515 synthesized from indole-3-acetaldoxime, a key branching point between primary and
516 secondary metabolism in Arabidopsis. Proc Natl Acad Sci USA 101: 8245-8250.
517
518 Grotewold E (2005) Plant metabolic diversity: a regulatory perspective. Trends Plant Sci
519 10: 57-62.
520
521 Greene EA, *et al*. (2003) Spectrum of chemically induced mutations from a large-scale
522 reverse-genetic screen in Arabidopsis. Genetics, 164(2), 731-740.
523
524 Hammerschmidt R (1999) PHYTOALEXINS: What have we learned after 60 years?
525 Annu Rev Phytopathol 37: 285-306.
526
527 Hardtke CS, Müller J, Berleth T (1996) Genetic similarity among Arabidopsis thaliana
528 ecotypes estimated by DNA sequence comparison. Plant Molecular Biology, 32(5), 915-
529 922.

530

531 Hartmann T (2007) From waste products to ecochemicals: fifty years research of plant
532 secondary metabolism. Phytochemistry 68: 2831-2846.
533

534 Henaff E, *et al*. (2014) Extensive amplification of the E2F transcription factor binding
535 sites by transposons during evolution of Brassica species. Plant J 77: 852-862.
536

537 Heintzman ND, *et al*. (2007) Distinct and predictive chromatin signatures of
538 transcriptional promoters and enhancers in the human genome. Nat Genet, 39(3), 311.
539

540 Hoffman BG, *et al*. (2010) Locus co-occupancy, nucleosome positioning, and H3K4me1
541 regulate the functionality of FOXA2-, HNF4A-, and PDX1-bound loci in islets and liver.
542 Genome Res 20(8): 1037-1061.
543

544 Hohmann N, Wolf EM, Lysak MA, Koch MA (2015) A time-calibrated road map of
545 brassicaceae species radiation and evolutionary history. Plant Cell 27(10): 2770-2784.
546

547 Hu TT, *et al*. (2011) The Arabidopsis lyrata genome sequence and the basis of rapid
548 genome size change. Nat Genet 43(5): 476-481.
549

550 Hugouvieux V, Barber, CE, Daniels, MM (1998) Entry of Xanthomonas campestris pv.
551 campestris into hydathodes of Arabidopsis thaliana leaves: a system for studying early
552 infection events in bacterial pathogenesis. MPMI, 11(6), 537-543.
553
554 Jones JDG, Dangl JL (2006) The plant immune system. Nature 444: 323-329.
555
556 Kagan IA, Hammerschmidt R (2002) Arabidopsis ecotype variability in camalexin
557 production and reaction to infection by Alternaria brassicicola. J Chem Ecol 28(11):
558 2121-2140.
559
560 Klein AP, Anarat-Cappillino G, Sattely ES (2013) Minimum set of cytochromes P450 for
561 reconstituting the biosynthesis of camalexin, a major Arabidopsis antibiotic.
562 Angewandte Chemie 52: 13625-13628.
563
564 Kover PX, Schaal BA (2002) Genetic variation for disease resistance and tolerance
565 among Arabidopsis thaliana accessions. Proc Natl Acad Sci USA 99(17): 11270-11274.
566
567 Kruse T, *et al*. (2008) In planta biocatalysis screen of P450s identifies 8-
568 methoxypsoralen as a substrate for the CYP82C subfamily yielding original chemical
569 structures. Chem Biol 15: 149-156.
570

571 Levine M, Davidson EH (2005) Gene regulatory networks for development. Proc Natl
572 Acad Sci USA 102: 4936-4942.
573
574 Li G, *et al*. (2012). Dual-level regulation of ACC synthase activity by MPK3/MPK6
575 cascade and its downstream WRKY transcription factor during ethylene induction in
576 Arabidopsis. PLoS genetics, 8(6), e1002767.
577
578 Liu S, Kracher B, Ziegler J, Birkenbihl RP, Somssich IE (2015) Negative regulation of
579 ABA signaling by WRKY33 is critical for Arabidopsis immunity towards Botrytis cinerea
580 2100. Elife 4: e07295.
581
582 Liu Y, *et al*. (2018) PCSD: a plant chromatin state database. Nucleic Acids Research,
583 46(D1), D1157-D1167.
584
585 Luan DD, Korman MH, Jakubczak JL, Eickbush TH (1993) Reverse transcription of
586 R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-
587 LTR retrotransposition. Cell 72: 595-605.
588
589 Malik HS, Burke WD, Eickbush TH (1999) The age and evolution of non-LTR
590 retrotransposable elements. Mol Biol Evol 16: 793-805.
591

592 Mansfield JW (2000) Antimicrobial compounds and resistance: the role of phytoalexins
593 and phytoanticipins. In Mechanisms of Resistance to Plant Diseases, A. Slusarenko,
594 RSS Fraser and LC. van Loon eds (Kluwer Academic Publishers Dordrecht, The
595 Netherlands), pp. 325-370.
596
597 Martin C, Ellis N, Rook F (2010) Do transcription factors play special roles in adaptive
598 variation? Plant Physiol 154: 506-511.
599
600 McClintock B (1956) Controlling elements and the gene. Cold Spring Harb Symp Quant
601 Biol 21: 197-216.
602
603 Mithen R, Bennett R, Marquez J (2010) Glucosinolate biochemical diversity and
604 innovation in the Brassicales. Phytochemistry 71: 2074-2086.
605
606 Moore RC, Purugganan MD (2005) The evolutionary dynamics of plant duplicate genes.
607 Curr Opin Plant Biol, 8(2), 122-128.
608
609 Mukherjee AK, Lev S, Gepstein S, Horwitz, BB (2009) A compatible interaction of
610 Alternaria brassicicola with Arabidopsis thaliana ecotype DiG: evidence for a specific
611 transcriptional signature. BMC Plant Biology, 9(1), 31.
612

613 Müller, AA (1966) Die Induktion von rezessiven Letalmutationen durch
614 Äthylmethansulfonat beiArabidopsis. Theoretical and Applied Genetics, 36(5), 201-220.
615
616 Murgia I, Tarantino D, Soave C, Morandini P (2011) Arabidopsis CYP82C4 expression
617 is dependent on Fe availability and circadian rhythm, and correlates with genes involved
618 in the early Fe deficiency response. J Plant Physiol 168: 894-902.
619
620 Nafisi M, *et al*. (2007) Arabidopsis cytochrome P450 monooxygenase 71A13 catalyzes
621 the conversion of indole-3-acetaldoxime in camalexin synthesis. Plant Cell 19: 2039-
622 2052.
623
624 Navarro L, *et al*. (2004) The transcriptional innate immune response to flg22. Interplay
625 and overlap with Avr gene-dependent defense responses and bacterial pathogenesis.
626 Plant Physiol 135: 1113-1128.
627
628 Ohno S (1970) Evolution by Gene Duplication. Springer-Verlag: Heidelberg, Germany.
629
630 Omranian N, *et al*. (2015) Differential metabolic and coexpression networks of plant
631 metabolism. Trends Plant Sci 20: 266-268
632
633 Pi W, *et al*. (2004) The LTR enhancer of ERV-9 human endogenous retrovirus is active
634 in oocytes and progenitor cells in transgenic zebrafish and humans. Proc Natl Acad Sci
635 USA 101: 805-810.

636

637 Prud'homme B, Gompel N, Carroll SB (2007) Emerging principles of regulatory
638 evolution. Proc Natl Acad Sci USA 104 Suppl 1: 8605-8612.
639

640 Qiu JL, *et al*. (2008) Arabidopsis MAP kinase 4 regulates gene expression through
641 transcription factor release in the nucleus. EMBO J 27: 2214-2221.
642

643 Rajniak J, Barco B, Clay NK, Sattely ES (2015) A new cyanogenic metabolite in
644 Arabidopsis required for inducible pathogen defence. Nature 525: 376-379.
645

646 Rajniak J, *et al*. (2018) Biosynthesis of redox-active metabolites in response to iron
647 deficiency in plants. Nat Chem Biol 14: 442-450.
648

649 Rédei GP (1962) Single locus heterosis. Zeitschrift für Vererbungslehre, 93(1), 164-170.
650

651 Rédei GP, Koncz C (1993) Classical mutagenesis. In *Methods in Arabidopsis research*
652 (pp. 16-82).
653

654 Rinerson CI, Rabara R, Tripathi P, Shen QJ, Rushton PJ (2015) The evolution of WRKY
655 transcription factors. BMC Plant Biology 15: 66.
656

657 Rodman J, Soltis R, Soltis D, Sytsma K, Karol K. (1998) Parallel evolution of

658 glucosinolate biosynthesis inferred from congruent nuclear and plastid gene
659 phylogenies. Am J Bot 85: 997.
660
661 Rogers WA, *et al*. (2013) Recurrent modification of a conserved cis-regulatory element
662 underlies fruit fly pigmentation diversity. PLoS Genetics, 9(8), e1003740.
663
664 Roudier F, *et al*. (2011) Integrative epigenomic mapping defines four main chromatin
665 states in Arabidopsis. EMBO J 30(10): 1928-1938.
666
667 Rushton PJ, Somssich IE, Ringler P, Shen QJ (2010) WRKY transcription factors.
668 Trends Plant Sci 15: 247-258.
669
670 Schluttenhofer C, Yuan L (2015) Regulation of specialized metabolism by WRKY
671 transcription factors. Plant Physiol 167: 295-306.
672
673 Schmidt JM, *et al*. (2010) Copy number variation and transposable elements feature in
674 recent, ongoing adaptation at the Cyp6g1 locus. PLoS Genetics 175: 1071-1077.
675
676 Slattery M, *et al*. (2014) Absence of a simple code: how transcription factors read the
677 genome. Trends Biochem Sci 39: 381-399.
678

679 Slotkin, R. K, Martienssen R (2007) Transposable elements and the epigenetic
680 regulation of the genome. Nat Rev Genet, 8(4), 272.
681
682 Spitz, F. Furlong EE (2012) Transcription factors: from enhancer binding to
683 developmental control. Nat Rev Genet 13: 613-626.
684
685 Studer A, Zhao Q, Ross-Ibarra J, Doebley J (2011) Identification of a functional
686 transposon insertion in the maize domestication gene tb1. Nat Genet 43: 1160-1163.
687
688 Tam OH, *et al*. (2008) Pseudogene-derived small interfering RNAs regulate gene
689 expression in mouse oocytes. Nature 453: 534-538.
690
691 Tao Y, *et al*. (2003) Quantitative nature of Arabidopsis responses during compatible and
692 incompatible interactions with the bacterial pathogen Pseudomonas syringae. Plant Cell
693 15: 317-330.
694
695 Thomson SJ, *et al*. (2009) The role of transposable elements in the regulation of IFN-λ1
696 gene expression. Proc Natl Acad Sci USA 106(28): 11564-11569.
697
698 Tsuji J, Jackson EP, Gage DA, Hammerschmidt R, Somerville SC (1992)
699 Phytoalexin accumulation in Arabidopsis thaliana during the hypersensitive reaction to
700 Pseudomonas syringae pv syringae. Plant Physiol 98: 1304-1309.

701

702 Watanabe T, *et al*. (2008) Endogenous siRNAs from naturally formed dsRNAs regulate
703 transcripts in mouse oocytes. Nature 453: 539-543.
704

705 Tohge T, Fernie AR (2012) Co-expression and co-responses: within and beyond
706 transcription. Front Plant Sci 3: 248.
707

708 VanEtten HD, Mansfield JW, Bailey JA, Farmer EE (1994) Two classes of plant
709 antibiotics: phytoalexins versus “phytoanticipins.” Plant Cell 6: 1191-1192.
710

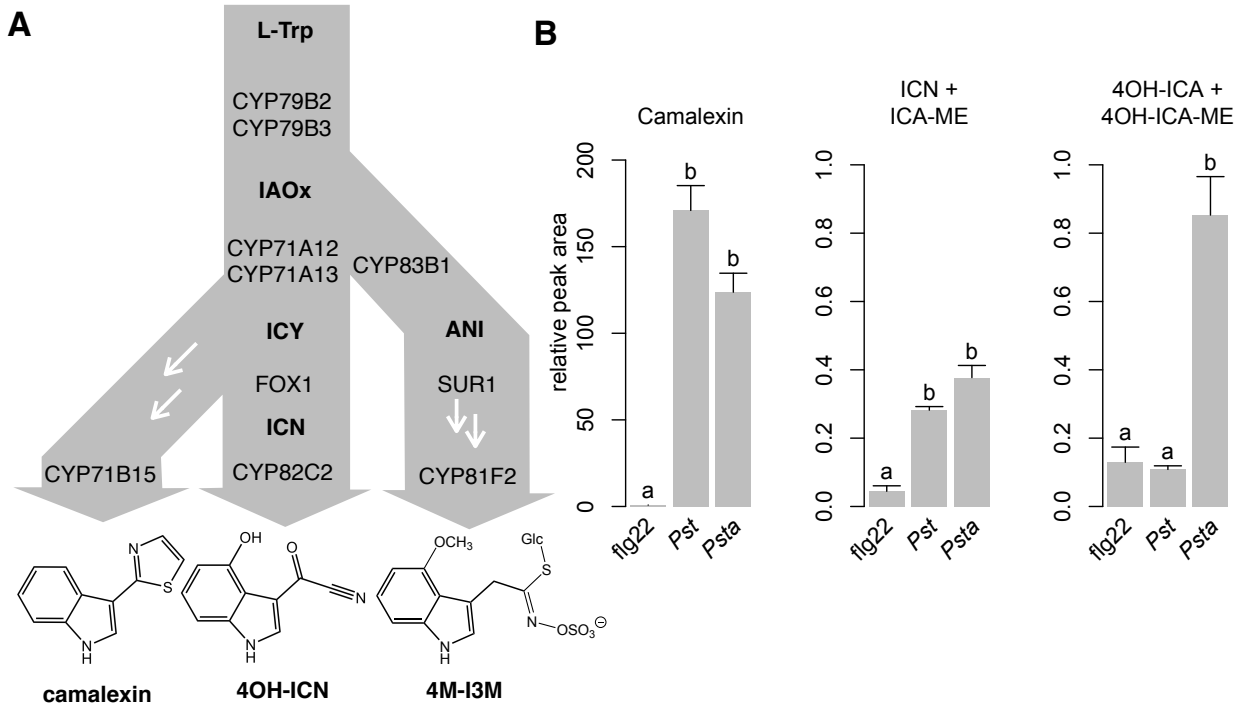
711 Wang Y, Li X, Hu H (2014) H3K4me2 reliably defines transcription factor binding
712 regions in different cells. Genomics, 103(2), 222-228.
713

714 Watanabe T, *et al*. (2008) Endogenous siRNAs from naturally formed dsRNAs regulate
715 transcripts in mouse oocytes. Nature 453: 539-543.
716

717 Weng JK, Philippe RN, Noel JP (2012) The rise of chemodiversity in plants. Science
718 336: 1667-1670.
719

720 Wicker T, *et al*. (2007) A unified classification system for eukaryotic transposable
721 elements. Nat Rev Genet 8: 973-982.
722

723 Wink M (2003) Evolution of secondary metabolites from an ecological and molecular
724 phylogenetic perspective. Phytochemistry 64: 3-19.
725
726 Wittkopp PJ, Kalay G (2012) Cis-regulatory elements: molecular mechanisms and
727 evolutionary processes underlying divergence. Nat Rev Genet 13: 59-69
728
729 Wray GA (2007) The evolutionary significance of cis-regulatory mutations. Nat Rev
730 Genet 8: 206-216.
731
732 Zhao Y, *et al*. (2002) Trp-dependent auxin biosynthesis in Arabidopsis:
733 involvement of cytochrome P450s CYP79B2 and CYP79B3. Genes Dev 16: 3100-3112.
734
735 Zheng Z, Qamar SA, Chen Z, Mengiste T (2006) Arabidopsis WRKY33 transcription
736 factor is required for resistance to necrotrophic fungal pathogens. Plant J 48: 592-605.
737
738



739

740 **Fig. 1. 4OH-ICN biosynthesis is specific to ETI**

741 **A.** Schematic of tryptophan (L-Trp)-derived specialized metabolism in *A. thaliana*. White

742 arrows denote the presence of additional enzymes. ICY, indole cyanohydrin; ANI, *aci*-

743 nitro indole.

744 **B.** LC-DAD-MS analysis of camalexin, ICN, and 4OH-ICN in seedlings elicited with

745 flg22, *Pst*, or *Psta* for 24 hr. Data represent mean ± SE of four biological replicates.

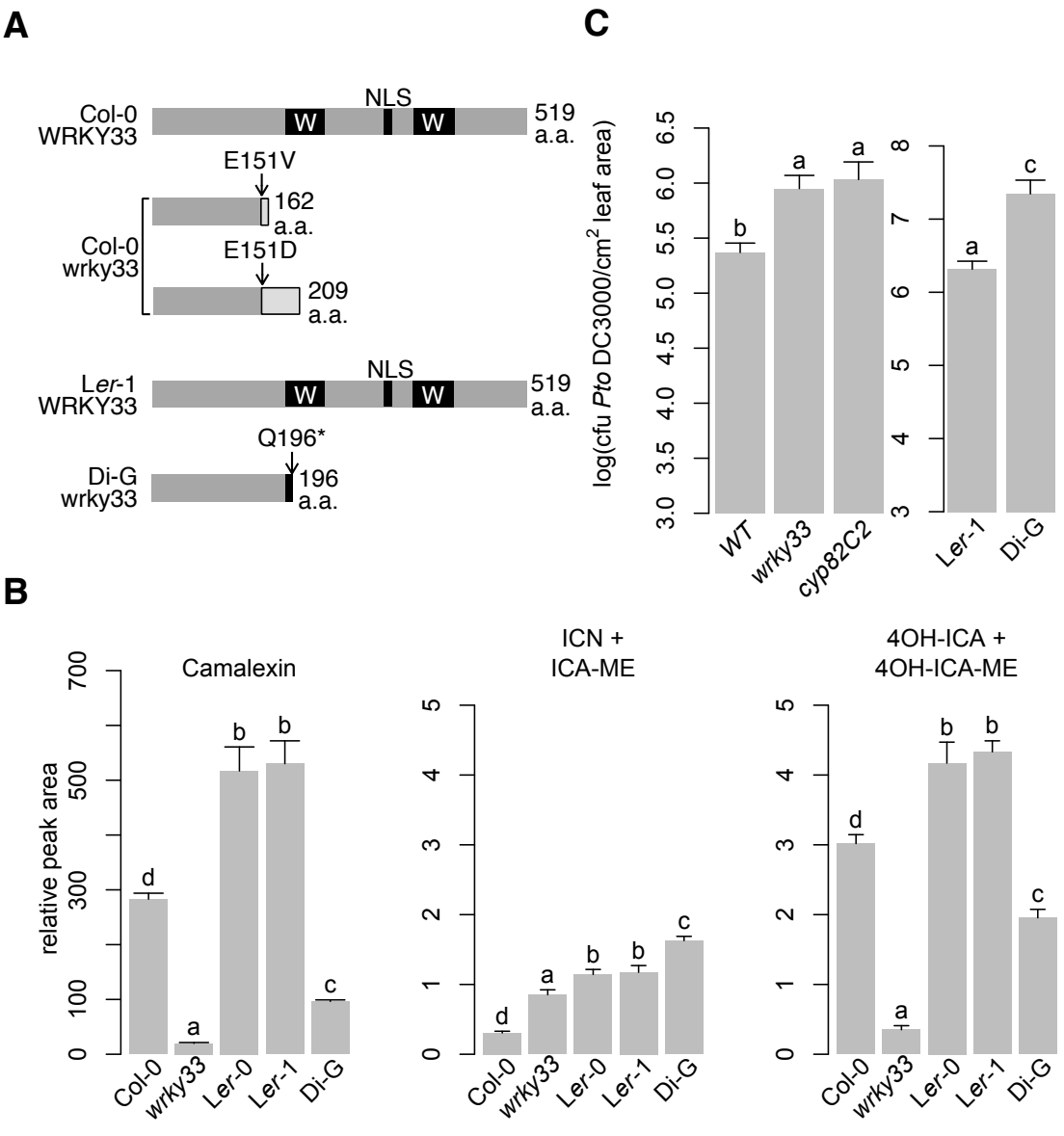
746 Different letters denote statistically significant differences ($P < 0.05$, two-tailed *t*-test).

747 ICA-ME and 4OH-ICA-ME are methanolic degradation products of ICN and 4OH-ICN,

748 respectively. 4OH-ICA is an aqueous degradation product of 4OH-ICN.

749

750



751

752 **Fig. 2. Intraspecific variation in *WRKY33* affects 4OH-ICN and immunity**

753 **A.** Schematic of WRKY33 proteins in Col-0, Col-0 *wrky33*, L*er*-1 and Di-G. Black boxes

754 denote WRKY domain (W) or nuclear localization signal (NLS).

755 **B.** LC-DAD-MS analysis of camalexin, ICN, and 4OH-ICN in seedlings inoculated with

756 *Psta* for 24 hr. Data represent mean ± SE of four replicates.

757 **C.** Bacterial growth analysis of *Pst* in surface-inoculated leaves pre-treated with 20 μM
758 dex for 6-8 hr. Data represent mean ± SE of 8-12 biological replicates. CFU, colony-
759 forming units. Different letters in **(B-C**) denote statistically significant differences ($P <$
760 0.05, two-tailed *t*-test). Experiments in (**B-C**) were performed at least twice, producing
761 similar results.
762
763

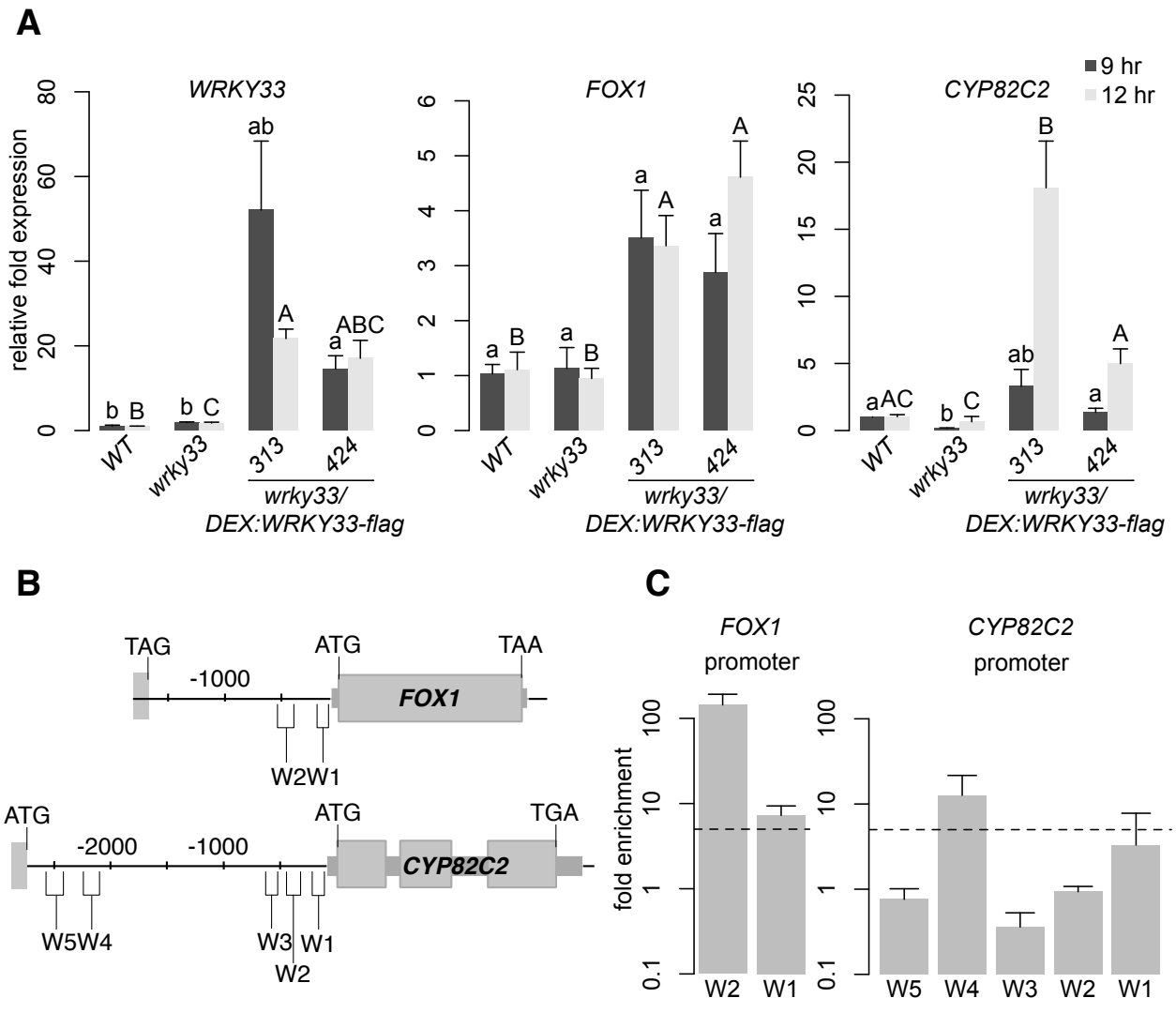


Fig. 3. WRKY33 directly activates 4OH-ICN biosynthetic genes.

A. qPCR analysis of 4OH-ICN regulatory and biosynthetic genes in seedlings inoculated with 20 μM dex and *Psta* for 9 and 12 hr.

B. Schematic of *FOX1* and *CYP82C2* loci, indicating nt positions of W-box-containing regions (W).

C. ChIP-PCR analysis of W-box-containing regions upstream of *FOX1* and *CYP82C2* in *wrky33/DEX:WRKY33-flag* plants co-treated with 20 μM dex (D) or mock solution (M)

772 and *Psta* for 9 hr. Dashed line represents the 5-fold cutoff between weak and strong TF-

773 DNA interactions. Data in (**B-C**) represent mean ± SE of four replicates.

774

775

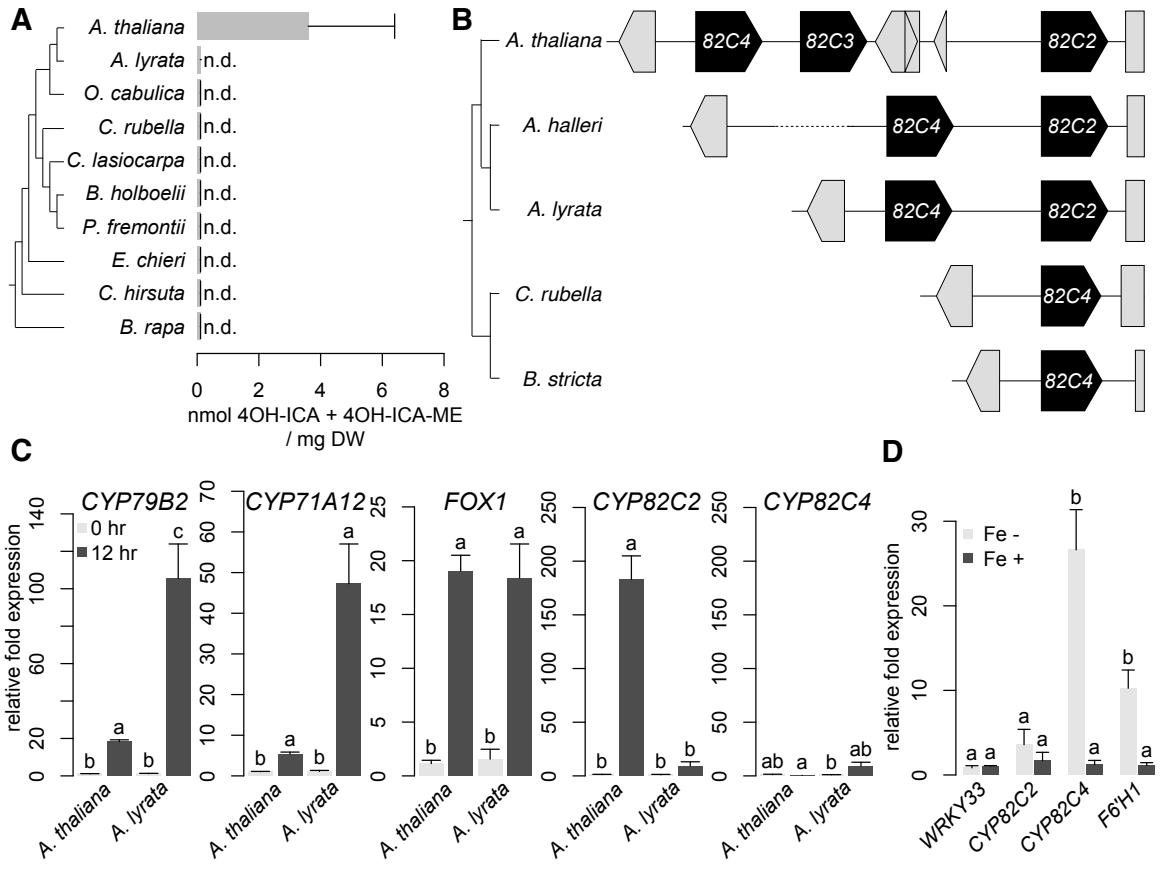


Fig. 4. Regulatory neofunctionalization of *CYP82C2*

A. (Right) HPLC-DAD analysis of 4OH-ICN in seedlings inoculated with *Psta* for 30 hr. (Left) Phylogenetic species tree. Data represent mean ± SE of three independent experiments (n = 4 biological replicates), each with *A. thaliana* as a positive control. 4OH-ICA and 4OH-ICA-ME are aqueous and methanolic degradations products of 4OH-ICN, respectively. DW, dry weight; n.d., not detected. Experiments were performed twice, producing similar results.

B. (Right) Synteny map of the *CYP82C* genes. Grey arrows represent non-*CYP82C* genes. Grey dotted lines represent large (>500nt) sequence gaps. (Left) phylogenetic species tree.

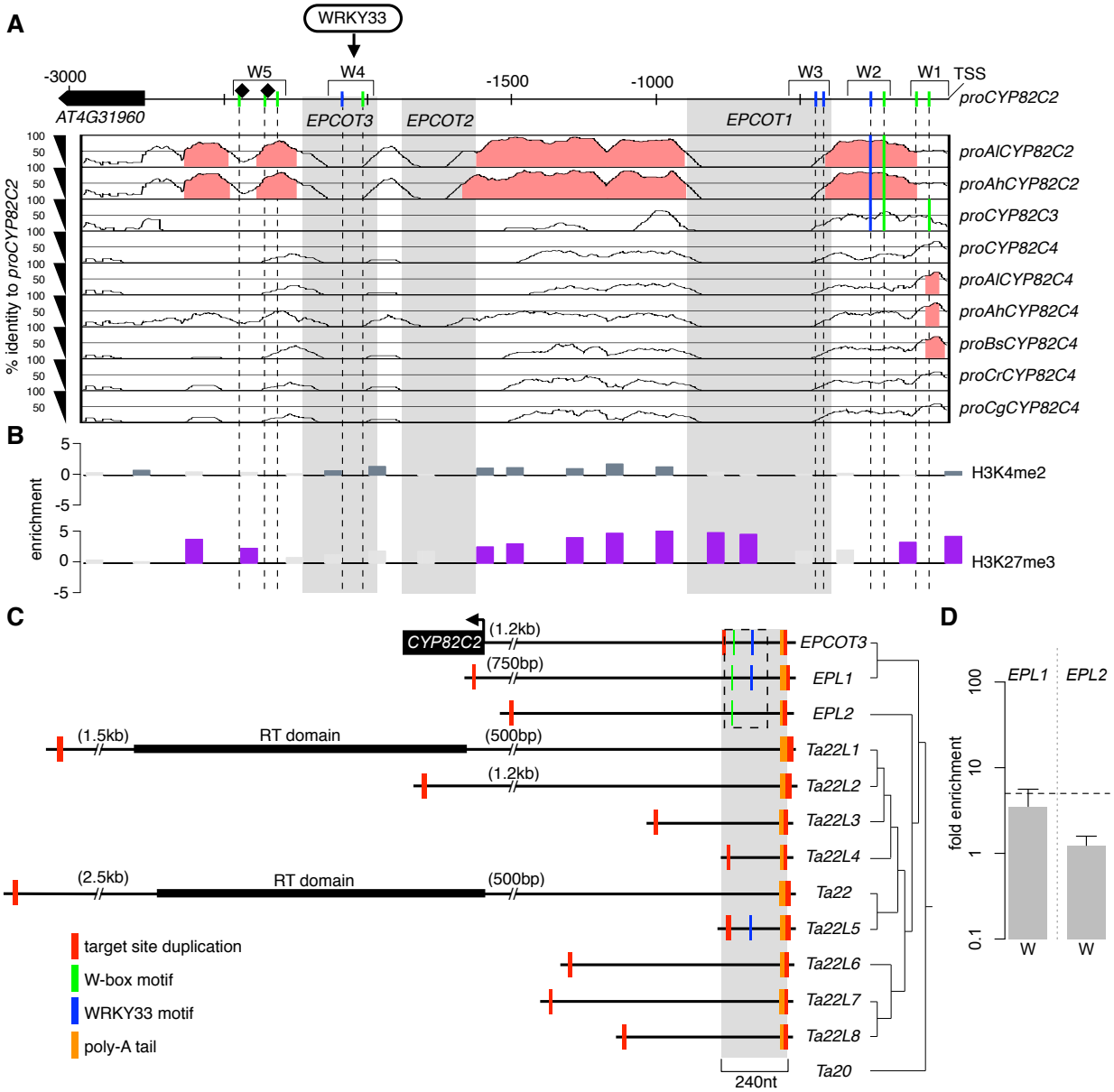
787 **C-D.** qPCR analysis of 4OH-ICN and sideretin biosynthetic genes in seedlings

788 inoculated with *Psta* (**C**) or grown in iron-deficient medium (**D**). Data represents the

789 mean ± SE of four biological replicates.

790

791



792

793 **Fig. 5. TE *EPCOT3* is a *CYP82C2* enhancer**

794 **A**. mVISTA plot of *CYP82C2* upstream sequence, indicating nt positions of unique

795 (*EPCOT1–3;* gray boxes) and conserved regions (≥70% sequence identity; pink) among

796 homologous sequences. Also indicated are positions of W-boxes (green) and WRKY33-

797 specific motifs (blue) that are present (solid lines) or absent (dashed lines) in each

798 homologous sequence, previously known WRKY33 TFBSs (diamonds) and ChIP-tested

799 regions (W1-5). TSS, transciptional start site; *Al, Arabidopsis lyrata; Ah, Arabidopsis*
800 *halleri*; *Cr, Capsella rubella*; *Bs, Boechera stricta; Cg, Capsella grandiflora*.
801 **B.** Epigenetic map of *CYP82C2* upstream sequence, indicating nt positions of
802 significant amounts of H3K4me2 (blue-gray bars), and H3K27me3 (purple bars).
803 **C.** (Left) Schematic of *EPCOT3* and related LINE retrotransposons in *A. thaliana* drawn
804 to scale, indicating nt positions of *CYP82C2* and reverse transcriptase (RT) domain. A
805 text file of the alignment and a more detailed tree are available as **Datasets S2-3**.
806 (Right) Phylogenetic maximum likelihood tree. Dashed box represent region containing
807 W-boxes (green lines) and/or WRKY33-binding motifs (blue lines) within *EPCOT3,*
808 *EPL1* and *EPL2*.
809 **D.** ChIP-PCR analysis of W-box-containing regions (W) within *EPL1* and *EPL2* in
810 *wrky33/DEX:WRKY33-flag* plants co-treated with 20 μM dex (D) or mock solution (M)
811 and *Psta* for 9 hr. Data represent mean ± SE of four replicates. Dashed line represents
812 the 5-fold cutoff between weak and strong TF-DNA interactions.
813
814

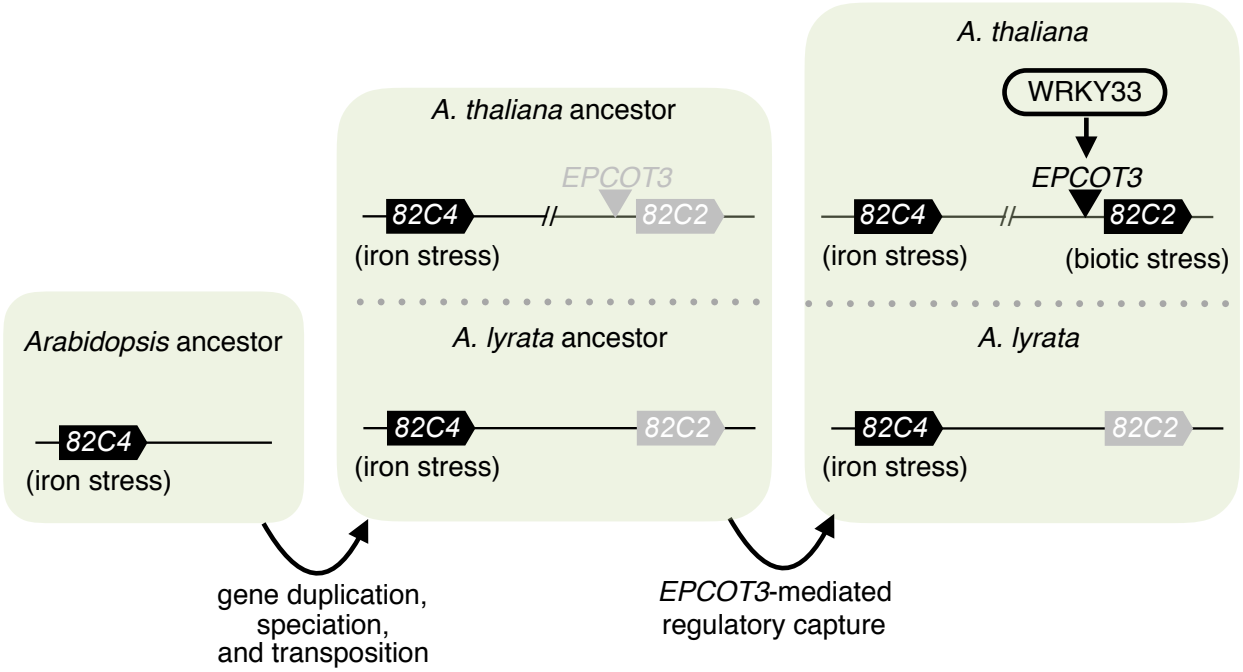


Fig. 6. Model of regulatory neofunctionalization of *CYP82C2*.