Article, Discoveries

The evolutionary dynamics of genetic incompatibilities introduced by duplicated genes in *Arabidopsis thaliana*

Wen-Biao Jiao[1], Vipul Patel[2,#], Jonas Klasen[2], Fang Liu[3], Petra Pecinkova[4,5], Marina Ferrand[6], Isabelle Gy[6], Christine Camilleri[6], Sigi Effgen[1], Maarten Koornneef[4,7], Ales Pecinka[4,8], Olivier Loudet[6], Korbinian Schneeberger[1,9]*

[1] Department of Chromosome Biology, Max Planck Institute for Plant Breeding Research, 50829 Cologne, Germany.

[2] Department of Developmental Biology, Max Planck Institute for Plant Breeding Research, 50829 Cologne, Germany.

[3] Department of Breeding Research, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, 06466 Stadt Seeland, Germany.

[4] Department of Plant Breeding and Genetics, Max Planck Institute for Plant Breeding Research, 50829 Cologne, Germany.

[5] Faculty of Science, Palacký University Olomouc, CZ-779 00, Olomouc, Czech Republic.

[6] Institut Jean-Pierre Bourgin, INRAE, AgroParisTech, Université Paris-Saclay, 78000, Versailles, France.

[7] Laboratory of Genetics, Wageningen University & Research., Droevendaalsesteeg 1, 6708 PB, Wageningen, The Netherlands.

[8] Institute of Experimental Botany (IEB), Czech Acad. Sci., Centre of the Region Haná for Biotechnological and Agricultural Research (CRH), CZ-779 00, Olomouc, Czech Republic.

[9] Faculty of Biology, LMU Munich, 82152, Planegg-Martinsried, Germany.

[#] Present addresses: Vipul Patel: AlceDiag, 1692 Rue de la Valsière, 34184 Montpellier, France

* Corresponding author: schneeberger@mpipz.mpg.de

Abstract

Although gene duplications provide genetic backup and allow genomic changes under relaxed selection, they may potentially limit gene flow. When different copies of a duplicated gene are pseudo-functionalized in different genotypes, genetic incompatibilities can arise in their hybrid offspring. While such cases have been reported after manual crosses, it remains unclear whether they occur in nature and how they affect natural populations. Here we identified four duplicated-gene based incompatibilities including one previously not reported within an artificial Arabidopsis intercross population. Unexpectedly, however, for each of the genetic incompatibilities we also identified the incompatible alleles in natural populations based on the genomes of 1,135 Arabidopsis accessions published by the 1001 Genomes Project. Using the presence of incompatible allele combinations as phenotypes for GWAS, we mapped genomic regions which included additional gene copies which likely rescue the genetic incompatibility. Reconstructing the geographic origins and evolutionary trajectories of the individual alleles suggested that incompatible alleles frequently co-exist, even in geographically closed regions, and that their effects can be overcome by additional gene copies collectively shaping the evolutionary dynamics of duplicated genes during population history.

Introduction

Genetic incompatibilities describe the decrease of fitness due to incompatible allele combinations in hybrid individuals (Maheshwari and Barbash 2011). In the hybrid offspring, genetic incompatibilities result in distorted segregation of the incompatible alleles. The evolution of genetic incompatibilities has often been explained by the Bateson-Dobzhansky-Muller (BDM) model (Bateson 1909; Dobzhansky 1937; Muller 1942), where independent mutations in interacting genes get fixed in different populations, which cause deleterious epistasis and reduced fitness in their hybrids. Over the past decades, many studies have elucidated the genetic basis of such genetic incompatibilities including reciprocal pseudo-functionalization (i.e. loss of function) of duplicated genes (Fishman and Sweigart 2018; Vaid and Laitinen 2019). Gene duplications can provide genetic backup of essential genes and the basis for evolutionary novelties by allowing for new genetic and epigenetic variations (Conant and Wolfe 2008; Kondrashov 2012; Panchy et al. 2016). However, in some cases pseudo-functionalization of duplicated essential genes may occur independently in both copies in different individuals. This in turn can lead to the loss of any functional gene copy in hybrid offspring of such individuals, and thereby cause severe genetic incompatibilities (Lynch and Force 2000).

Genetic incompatibilities introduced by duplicated genes have been reported within inter/intra-specific hybrids of *Arabidopsis thaliana* (Bikard et al. 2009; Durand et al. 2012; Agorio et al. 2017), rice (Mizuta et al. 2010; Yamagata et al. 2010; Nguyen et al. 2017) or *Mimulus* (Zuellig et al. 2017). Identification of these incompatible alleles, however, often relied on genetic mapping in experimental populations, which is a time consuming and costly process. As incompatible alleles are frequently introduced by loss-of-function (LoF) (epi)mutations (Bikard et al. 2009; Blevins et al. 2017), initial examination of LoF (epi)mutations within whole-genome sequence data could be a shortcut to quickly target promising candidates.

Although several incompatible alleles from duplicated genes have been identified in *A. thaliana*, it is still unclear how these incompatible alleles originate and evolve in natural populations, and how the populations adapt to the reduction in fitness. Untangling the complex evolutionary process would require accurate (epi)genotypes of incompatible genes across sufficiently large natural populations. The Arabidopsis 1001 Genomes Project (Alonso-Blanco et al. 2016) and 1001 Epigenomes Project (Kawakatsu et al. 2016) have released substantial omics data, which can be used to unravel the evolutionary trajectory of such incompatible alleles.

Here, we created an extended version of the Arabidopsis multiparent RIL population (Huang et al. 2011) to identify genetic incompatibilities between several different genotypes

simultaneously. Based on distorted segregation of duplicated genes, we mapped four genetic incompatibilities. Unexpectedly, however, we identified several, healthy RILs which carried presumably incompatible allele combinations. Further analysis of their genomes revealed additional gene copies rescuing these severe incompatibilities. Encouraged by this, we searched for incompatible allele combinations within 1,135 accessions of the 1001 Genomes Project (Alonso-Blanco et al. 2016), where these combinations were surprisingly common. Using the incompatible allele combinations as phenotypes we mapped modifiers of all four incompatibilities using GWA. The loss-of-function alleles from duplicated genes were geographically widely distributed, and co-existed with additional gene copies in the same regions showing how additional gene copies can overcome differential copy loss in a population.

Results

Identification of incompatible gene pairs within an intercross population

We used the Arabidopsis Multiparent RIL (AMPRIL) population to find incompatible alleles that arose from duplicated genes. The eight AMPRIL founder accessions (An-1, C24, Col-0, Cvi-0, Eri-1, Kyo, L*er* and Sha) were selected across the entire geographic distribution of *A. thaliana* including the Northern hemisphere and the Cape Verde Islands. Recently, we generated chromosome-level genome assemblies of all seven, non-reference founder genomes (Jiao and Schneeberger 2020). The first release of the AMPRIL population (AMPRIL I) contained six subpopulations (referred to as ABBA, ACCA, ADDA, BCCB, BDDB, CDDC) derived from reciprocal diallel crosses between four hybrids (A: Col-0 x Kyo, B: Cvi-0 x Sha, C: Eri-1 x An-1, D: L*er* x C24) and the subsequent selfing to the F4 generation by single-seed descent (Huang et al. 2011). Here we present the extension of the AMPRIL population with six new subpopulations referred to as EFFE, EGGE, EHHE, FGGF, FHHF and GHHG (E: Col-0 x Cvi-0, F: Sha x Kyo, G: Ler x An-1, H: Eri-1 x C24) based on different diallel intercrossing scheme and selfing of the recombinant genomes until the F6 generation (Fig. 1a). Each subpopulation consists of approximately 90 individuals representing recombinants of four founders. In total, 992 RILs from all twelve subpopulations were sequenced and analyzed using RAD-seq (Baird et al. 2008) (Supplementary Data 1) and genotyped with ~2 million high-quality SNP markers. We used a Hidden Markov Model to reconstruct the parental haplotypes (identity-by-descent) including residual heterozygous regions (Rowan et al. 2015) (Supplementary Fig. 1 and Supplementary Note 1). The genotyping resulted in 12,878 different recombination breakpoints (on average one breakpoint per 9.3 kb) across the entire population. This allowed us to divide the genome of each progeny into 12,883 haplotype blocks, where each block relates to the haplotype(s) of

120 only one (homozygous regions) or two (heterozygous regions) of the founder haplotypes.
121 We developed a two-step workflow to combine genetic and genomic evidence to quickly
122 identify incompatible alleles of duplicated genes (Fig. 1b). In the first step, we selected 781
123 distal (inter-chromosome) duplicated gene pairs including 612 gene pairs in which the
124 reference sequence contains two copies and other founder genomes feature at least one
125 copy (Supplementary Data 2). In the remaining 169 gene pairs, the reference sequence only
126 has one copy and at least one other parental genome has an additional copy in a different
127 chromosome. As genetic incompatibility leads to the underrepresentation of incompatible
128 allele combinations (Ackermann and Beyer 2012; Corbett-Detig et al. 2013), we searched for
129 significant distortions from the expected frequencies of all parental allele combinations across
130 all 781 duplicated gene pairs in all twelve subpopulations, two merged subpopulations (ABBA
131 and EFFE, CDDC and GHHG as they share the same founders), and the whole population
132 (see Methods). These tests revealed significant distortions in 236 gene pairs (χ^2 test, p-value
133 < 0.05, multiple testing corrected) in at least one of the populations.
134 However, the observed distortions do not necessarily result from genetic incompatibilities
135 in the tested gene. Alternatively, such distortions can also occur if the tested gene duplicate
136 is closely linked to a genetic incompatibility. Hence, in a second step, we examined the alleles
137 of the gene pairs in the founder genomes for loss-of-function (LoF) variations or
138 hypermethylated promoters (Fig. 1b, see Methods). This examination revealed three gene
139 pairs with functional disruption in both of the duplicates in at least one of the founder genomes.
140 These three duplicated genes included two, *HISTIDINOL PHOSPHATE*
141 *AMINOTRANSFERASE* (*HPA*) (Bikard et al. 2009) and *FOLATE TRANSPORTER* (*FOLT*)
142 (Durand et al. 2012), which were already known for their ability to introduce genetic
143 incompatibilities, as well as one gene pair, which so-far was not reported as the genetic basis
144 for a genetic incompatibility, *TIM22* (*TIM22-1*: *AT1G18320*, *TIM22-2*: *AT3G10110*) (Fig. 1c).
145 For all other 233 gene pairs we could not identify non-functional alleles in both copies.
146 We noted that another duplicated gene, *tRNA ADENOSINE DEAMINASE 3* (*TAD3)*,
147 which is also know to introduce a genetic incompatibility (Agorio et al. 2017), was not
148 considered by our initial testing even though the genotypes of the AMPRIL founders should
149 lead to incompatible allele combinations in the RIL populations: all founder genomes except
150 for Kyo have a functional *TAD3-1* (Supplementary Table 1), while the Kyo *TAD3-1* gene is
151 silenced most likely due to its methylated promoter (similar to the Nok-1 and Est-1 accessions
152 in which the incompatibility was described originally (Agorio et al. 2017)). The lack of a
153 functional *TAD3-1* in Kyo is counterbalanced by additional copies (*TAD3-2*) (which however
154 were not part of the main chromosome scaffolds of the Kyo genome assembly). Therefore,
155 the *TAD3* gene duplicate was not considered initially, however, we also used this
156 incompatibility for further analysis.

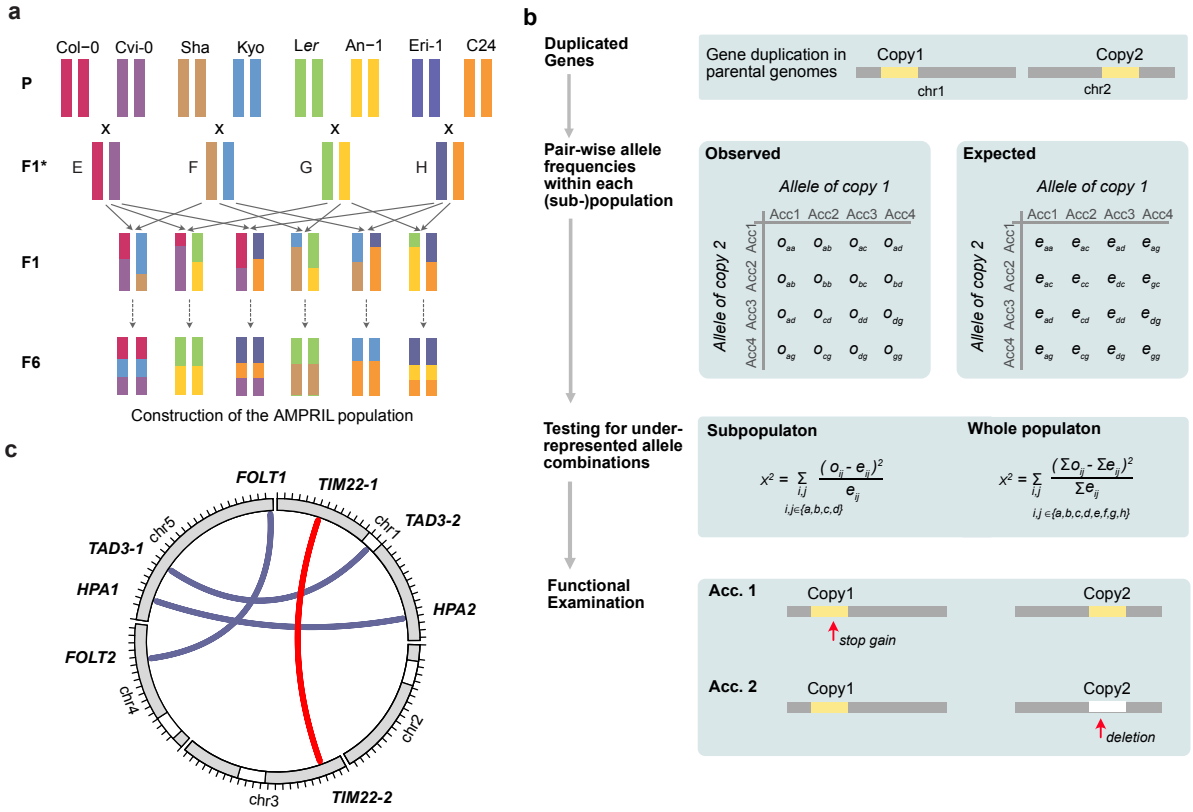


Figure 1. Identification of genetic incompatibilities introduced by duplicated genes in an intercross population (**a**) Construction of the extended AMPRIL population. Eight different *A. thaliana* accessions were used as founder lines. For each of the six subpopulations, two F1* hybrids, which were generated by crossing two founder lines, were again crossed to give rise to the F1 individuals of each population. The F1 individuals were further self-crossed to the F6 generation. (**b**) Workflow for the identification of potentially incompatible alleles in duplicated genes. Unlinked (i.e. on separate chromosomes) duplicated genes were selected and the expected and observed frequencies of all haplotype combinations between the two copies were calculated for each subpopulation and the whole population. Under-represented allele combinations were identified using χ^2 test. Each gene duplication with significantly underrepresented allele combinations was evaluated for non-functionalized or deleted gene copies in the respective parental genomes. (**c**) The location of incompatible alleles in four duplicated gene pairs identified in the AMPRIL population including one so-far unknown incompatibility (red) and one detected a posteriori in an informed way (*TAD3*).

Genetic incompatibility introduced by diverged copies of *TIM22*

Before we started our analysis of the four incompatibilities, we verified that the LoF alleles of *TIM22* are in fact the causal basis of the genetic incompatibility that we observed in the AMPRIL population. *TIM22* encodes for a mitochondrial import inner membrane translocase subunit of the *TIM17*/*TIM22*/*TIM23* family protein (Murcha et al. 2007). All eight founders feature two annotated *TIM22* copies, while Cvi-0 includes an extra truncated copy ~31 kb downstream to *TIM22-2* (Fig. 2a, b). We found significant segregation distortions in the

179 complete AMPRIL population and one subpopulation, EGGE, where the double homozygous
180 allele combination *TIM22-1*$^{Col-0}$*TIM22-2*$^{Cvi-0}$ was significantly underrepresented (Supplementary
181 Table 2 and 3).

182 We observed an in-frame premature stop-codon in Col-0 (mis-annotated in the reference
183 annotation) suggesting that *TIM22-1* is not functional in Col-0 (Fig. 2a and Supplementary
184 Table 4). To test if *TIM22-1* is truly non-functional in Col-0, we used the segregation of two T-
185 DNA insertion mutants in the two *TIM22* paralogs in Col-0. This showed that *tim22-1* could
186 be homozygous for the T-DNA insertion allele but the T-DNA in *tim22-2* could not be found in
187 homozygous state (Fig. 2c). This suggests that, in Col-0, *TIM22-1* is not functional and *TIM22-*
188 *2* is the only functional copy.

189 The *TIM22* paralogs co-located within the regions of a previously reported genetic
190 incompatibility (hereafter named as LD2: LD2.1 for the locus at chromosome 1 and LD2.3 for
191 the locus at chromosome 3) which was mapped in a Cvi-0 x Col-0 RIL population (Simon et
192 al. 2008). The genetic underpinnings of this incompatibility however were still unknown. This
193 incompatibility was expressed by a striking underrepresentation of homozygous LD2.1^{Col-0}
194 combined with homozygous LD2.3^{Cvi-0}, which was in agreement with the reduced allele
195 combinations in the AMPRIL population (Supplementary Table 2 and 3). Therefore, we
196 generated heterogeneous inbred family (HIF) lines from Cvi-0 x Col-0 RILs to fine-map LD2.1
197 and LD2.3 to respectively 70kb and 34kb intervals (Supplementary Fig 2). The two candidates
198 *TIM22-1* and *TIM22-2* remained within the intervals.

199 To validate their causative role, we conducted a complementation cross between a
200 heterozygous T-DNA mutant in *TIM22-2* in a Col-0 background (i.e. *TIM22-1* was non-
201 functional) and the original HIF line in which *TIM22-1* was homozygous for the Col-0 genotype
202 (i.e. also non-functional) and *TIM22-2* was heterozygous for Col-0/Cvi-0 (Fig. 2d). Within 123
203 hybrids of the offspring, among the four possible allelic combinations at LD2.3, we did not find
204 any hybrids combining a Cvi-0 and a T-DNA alleles at *TIM22-2* (Fig. 2d), providing strong
205 genetic evidence that the Cvi-0 allele cannot complement a knockout (T-DNA) allele at *TIM22-*
206 *2* (in a background without other functional *TIM22* allele) and is thus non-functional.

207 Collectively, these segregation and complementation crosses show that the combination
208 of different non-functional alleles of the *TIM22* copies leads to a drastic reduction of allelic
209 combinations in offspring populations, and thus evidence the causative role of *TIM22* in this
210 genetic incompatibility.

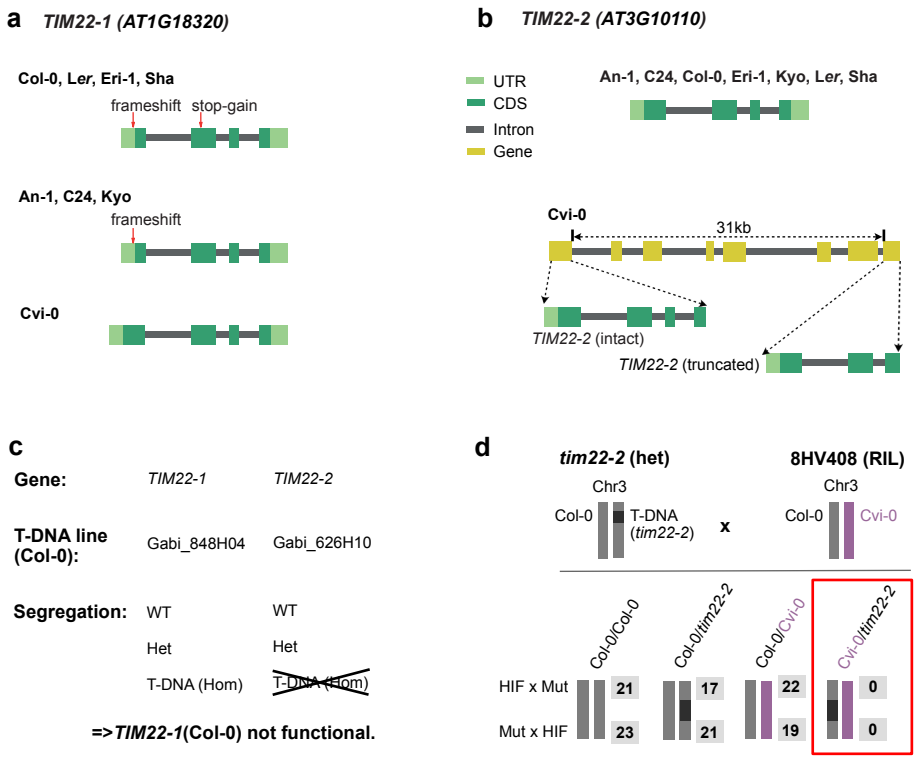
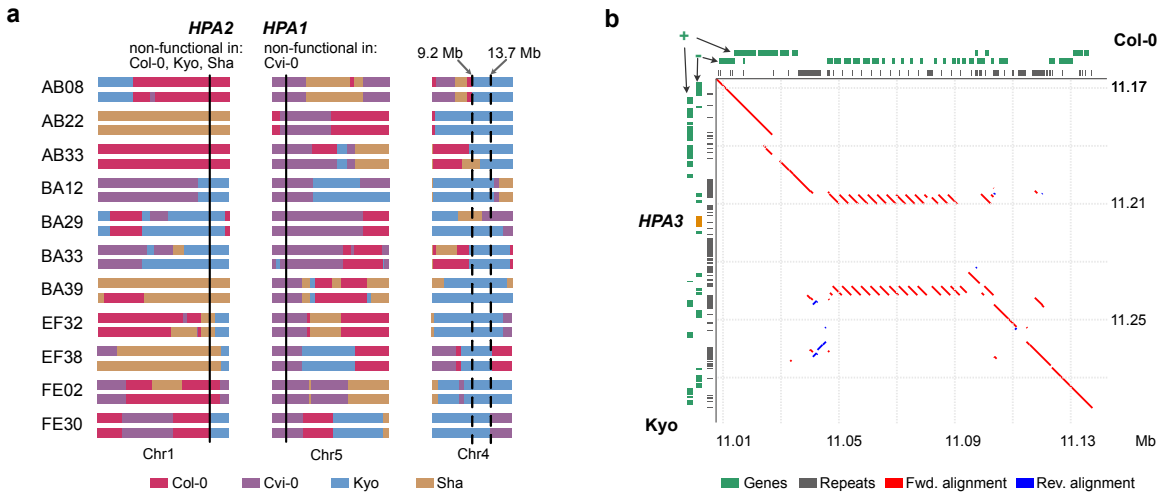


Figure 2. Genomic and genetic evidence of incompatible *TIM22* alleles. **(a)** Gene structure of *TIM22-1* in the genomes of the eight AMPRIL parents. The loss-of-function variants (1bp frameshift indel and a premature stop-codon) relative to the intact *TIM22-1*$^{Cvi-0}$ are shown. The deleterious effect of the frameshift is erased by an alternative translation start site. **(b)** Gene structure of *TIM22-2* in the genomes of the eight AMPRIL parents. The eight accessions share the structure of *TIM22-2* without recognizable loss-of-function variants, however, a truncated copy of *TIM22* could be found in Cvi-0 ~31 kb downstream of *TIM22-2*. **(c)** Segregation of T-DNA alleles within the descendance of 2 segregating Col-0 T-DNA mutant lines (*tim22-1* and *tim22-2*). **(d)** A heterozygous T-DNA line (*tim22-2*) in the Col-0 background (i.e. non-functional for *TIM22-1*) was crossed to 8HV408 (heterozygous Col-0/Cvi-0 at *TIM22-2* and homozygous for the Col-0 allele at *TIM22-1* i.e. nonfunctional for *TIM22-1*). The number of all four possible F1 progenies are shown (in grey) for both cross directions. While the *TIM22-2*$^{Col-0}$ allele can complement the T-DNA, the *TIM22-2*$^{Cvi-0}$ could not, implying that the *TIM22-2*$^{Cvi-0}$ allele is non-functional. HIF: heterogeneous inbred families, Mut: mutant.

Natural modifiers can rescue incompatible allele combinations

Incompatible allele combinations can result in severe phenotypic defects, which lead to the reduction or the full absence of specific allele combinations. For example, the double homozygous non-functional allele combination of *HPA1/HPA2* results in embryo lethality (Bikard et al. 2009) and thereby wipes out all carriers of the incompatible allele combination. *HPA* encodes a histidinol-phosphate amino-transferase for the biosynthesis of histidine, an essential amino acid (Muralla et al. 2007). All eight AMPRIL founders except of Cvi-0 have a functional *HPA1* and a non-functional *HPA2* due to a premature stop codon or a

235 hypermethylated promoter, while Cvi-0 carries a functional *HPA2* allele, but does not carry
236 *HPA1* at all (Supplementary Table 5). Unexpectedly, however, we did observe homozygous
237 *HPA1/HPA2* incompatible allele combinations (*HPA2*$^{-/-}$*HPA1*$^{-/-}$) in eleven of the AMPRIL lines
238 within the ABBA and EFFE subpopulations which were derived from Col-0, Cvi-0, Kyo and
239 Sha (Supplementary Table 6-8). Further analysis of these populations revealed an extremely
240 high frequency of the Kyo allele on chromosome 4 (Supplementary Fig. 3), making us
241 recognize that the eleven AMPRIL lines with the incompatible allele combinations all carried
242 at least one Kyo allele at chr4:9.2-13.7 Mb (Fig. 3a). This suggested that Kyo might contain
243 a modifying allele of the incompatibility complementing the lethal allele combination in this
244 region (a similar feature was described as "conditional incompatibility" by Bikard et al. when
245 analyzing crosses between Jea and Col-0 (Bikard et al. 2009)). Indeed, when we checked the
246 long-read assembly of the Kyo genome, we identified an additional *HPA* copy (hereafter
247 named as *HPA3*) which co-located with the mapping interval at chr4:11.11 Mb (Fig. 3b), while
248 we could not find this allele in any of the other AMPRIL founder genomes.
249



250
251 **Figure 3. Incompatible allele combinations of *HPA* rescued by an additional gene copy. (a)**
252 Schematic of the genomes of 11 AMPRILs with presumably incompatible allele combinations of *HPA1*
253 and *HPA2*. All these genotypes carry at least one Kyo allele at chr4: 9.2 – 13.7 Mb, suggesting that a
254 Kyo allele in this region can rescue the incompatibility. **(b)** Sequence alignment around the *HPA3* locus
255 on chromosome 4 between Col-0 and Kyo. The position of the third *HPA3* copy in Kyo is marked in
256 orange. Red line: forward alignment; blue line: reverse alignment. Genes arrangement at forward (+)
257 and reverse (-) strands, and repeat annotations are shown at the top (Col-0) or left (Kyo) axes.
258
259 **Mapping modifiers of incompatible allele combinations in natural populations**
260 This analysis showed that the negative effects of incompatible allele combinations can
261 be overcome if they are rescued by modifying alleles. Therefore, the virtual absence of

262 functional alleles (among the know loci) of a duplicated gene could act as a molecular
263 phenotype to map the location of additional (rescuing) alleles by genome-wide association
264 (GWA) mapping even in natural populations.

265 To do this, we searched for incompatible allele combinations (separately for all four
266 incompatibilities) in each of the 1,135 accessions of the 1001 Genomes Project (Alonso-
267 Blanco et al. 2016) and used these allele combinations as phenotype for a GWA (Fig. 4). To
268 define non-functional alleles, we used the resequencing data to search for LoF variations,
269 and the methylome data from the 1001 Epigenomes Project (Kawakatsu et al. 2016) to identify
270 methylated (silenced) promoters (Fig. 4a, see Methods). Additionally, RNA-seq data were
271 explored to distinguish pseudo-heterozygous variants (which exist due to inaccurate short-
272 read alignment at duplicated genes) and to check for gene silencing.

273 For the first incompatibility, in *HPA*, we found pseudo-heterozygous LoF variations
274 including two premature stop-codons and three frameshifts at both reference copies of *HPA*
275 (Fig. 4c and Supplementary Table 9) due to the repetitiveness of the gene sequences. As
276 some accessions did not show expression of *HPA2* most likely due to hypermethylation of
277 their promoters, we tested which of the *HPA* alleles was present in the RNA-seq read data to
278 assign the LoF to either of the *HPA* copies (Fig. 4b). With this, we could assign one stop-
279 codon gain (LoF-300) and two frameshifts (LoF-170, LoF-230) to *HPA2* because these LoF
280 alleles were absent in the RNA-seq data in the accessions which lacked *HPA2* expression.
281 Notably, the frameshift LoF-170 could be rescued by alternative splicing as observed in RNA-
282 seq read mapping (Supplementary Fig. 4). Furthermore, cytosine methylation profiles
283 revealed hypermethylated promoters of *HPA2* in 340 accessions and of *HPA1* in 50
284 accessions (Fig 4d, e), suggesting gene silencing in these accessions, which was in
285 agreement with the absence of pseudo-heterozygous variants in the RNA-seq read data (Fig.
286 4b and Supplementary Fig. 4).

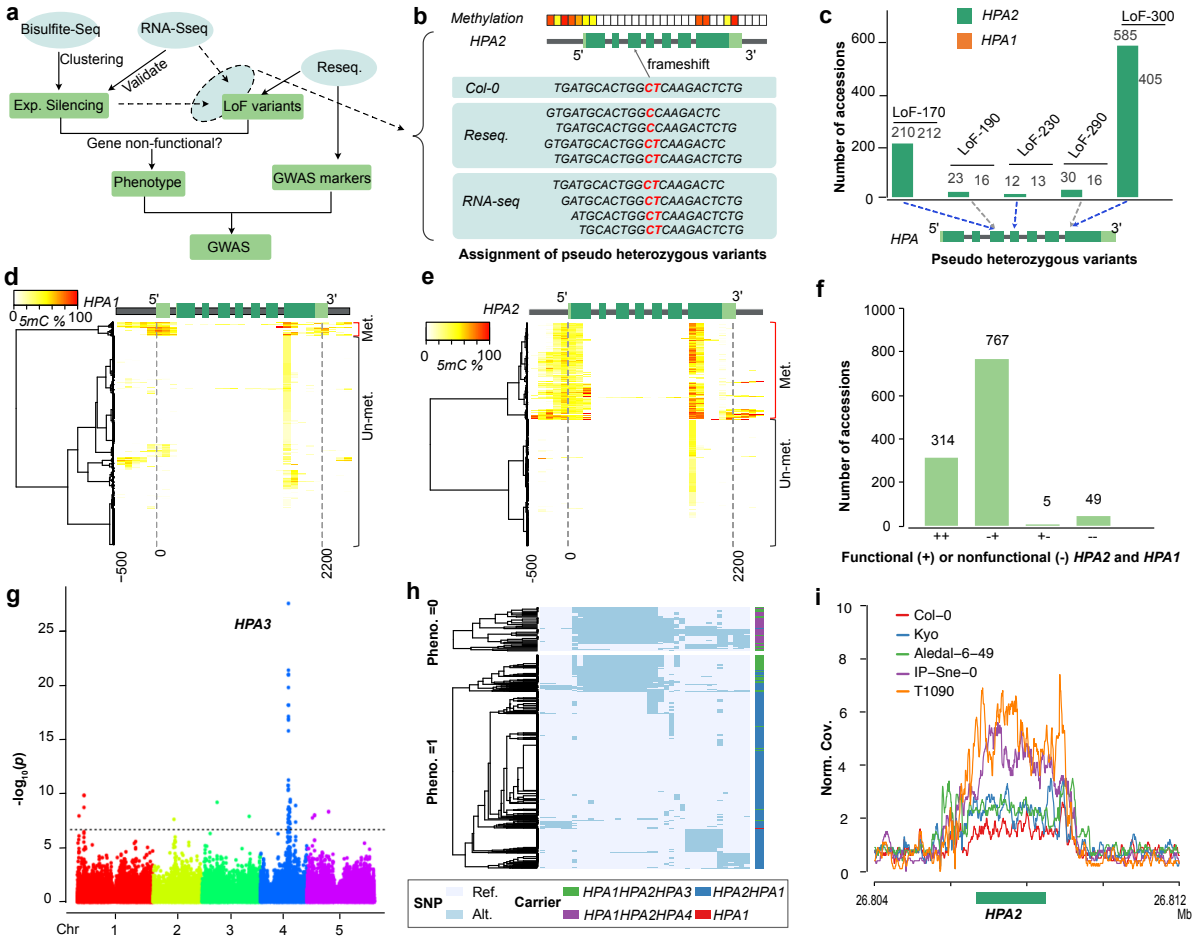


Figure 4: Mapping natural modifiers of genetic incompatibilities using GWAS. (a) Workflow for the identification of modifying alleles of genetic incompatibility using GWAS. **(b)** Schematic example to illustrate how genome-wide methylation and RNA-seq data are used to assign pseudo-heterozygous variants to a specific gene copy despite the repetitive nature of the short-read alignments within duplicated genes. If a specific variation is present in DNA data, but absent in RNA data and one of the gene copies is methylated (i.e. likely expression silenced) the pseudo-heterozygous variation is assigned to this (expression silenced) gene copy. Light green: untranslated region, green: coding region, gray: intron or gene up/down-stream. **(c)** Pseudo-heterozygous LoF variants found in the short read alignments (of 1,135 *A. thaliana* genomes(Alonso-Blanco et al. 2016)) at *HPA2* and *HPA1*. LoF-170, LoF-230, and LoF-300 could be assigned to *HPA2* (using the procedure of (b)), while LoF-190 and LoF290 could not be assigned to either of the gene copies. **(d, e)** Hierarchical clustering of DNA methylation profiles in *HPA1* (d) and *HPA2* (e) based on the methylomes of 888 *A. thaliana* accessions from the 1001 Epigenome Project (Kawakatsu et al. 2016) (NCBI GEO accession: GSE43857). Methylation profiles calculated within 100 bp sliding windows from 500 bp upstream of the transcription start site to 300 bp downstream of the transcription end site. **(f)** The number of accessions with different functional copies of *HPA*2 and *HPA1* across 1,135 *A. thaliana* accessions. **(g)** Manhattan plot of a GWA using the absence of any functional *HPA* gene as phenotype. The most significantly associated locus reveals the region of *HPA3*. The dashed line indicates the significant threshold after multiple testing correction ($-\log_{10}(P)=6.68$). **(h)** Two heatmaps of haplotype clustering (defined by the 39

307 significantly associated markers around the *HPA3* locus) shown for accessions with (below) or without
308 (up) functional copies of *HPA2* or *HPA1*. **(i)** The normalized short read mapping coverage (Norm. Cov.:
309 average mapping coverage at *HPA2* divided by average mapping coverage at whole genome) around
310 *HPA2* based on read mappings against the Cvi-0 genome including only one *HPA* gene. (Data from
311 five accessions are shown as examples to illustrate patterns of different *HPA* copies: Col-0: *HPA2HPA1*;
312 Kyo and Aledal-6-49: *HPA2HPA1HPA3*; IP-Sne-0 and T1090: *HPA2HPA1HPA4*).
313
314 By combining the LoF variant genotyping and the methylation analyses, we found 767
315 accessions with a non-functional *HPA2* (*HPA2*$^{-/-}$ *HPA1*$^{+/+}$) allele and five accessions with a
316 non-functional *HPA1* (*HPA2*$^{+/+}$*HPA1*$^{-/-}$) allele (Fig. 4f and Supplementary Table 10). Also, 49
317 accessions did not feature any functional alleles (*HPA2*$^{-/-}$*HPA1*$^{-/-}$) and were expected to carry
318 additional modifier(s) to complement the loss of functional copies. To find the locations of
319 these modifiers, we used the absence of functional copies as phenotype (Supplementary
320 Data 3) to run a GWA under the mixed linear model using the SNP markers from the 1001
321 Genomes Project (www.1001genomes.org). This GWAS revealed a significantly associated
322 region at chr4:11.00-11.15 Mb (Fig. 4g and Supplementary Fig. 5, alpha level of 0.05,
323 Bonferroni correction), corresponding to the *HPA3* locus found in Kyo (chr4:11.11 Mb).
324 Though other peaks in unlinked regions were present, these additional loci explained only a
325 small proportion of the heritability.
326 Analyzing the haplotypes at the 39 significantly associated SNP markers at *HPA3* locus
327 revealed a somewhat, but not entirely homogenous haplotype in many of the *HPA2*$^{-/-}$ *HPA1*$^{-/-}$
328 carriers (Fig. 4h). To confirm that the modifying haplotype is still identical to the Kyo allele, we
329 aligned the short reads of all accessions of the 1001 Genomes Project against the Kyo
330 reference sequence and found that overall 162 accessions carried the *HPA3* allele
331 (Supplementary Data 3). However, unexpectedly the Kyo *HPA3* was only found in 13 of the
332 49 accessions without functional *HPA2* and *HPA1* alleles, suggesting the presence of two
333 different modifiers at this locus on chromosome 4. To find more support for this, we ran a new
334 GWA without the 162 *HPA3* (Kyo-like allele) carriers, which still led to a significantly
335 associated locus at the region of *HPA3* (Supplementary Fig. 6). Further short-read mappings
336 of all 1,135 genomes against the Cvi-0 reference sequence (where only one copy of *HPA*
337 exists) revealed the presence of (at least) two additional copies (in addition to *HPA1* and
338 *HPA2*) in a total of 42 accessions including all the 36 accessions with the unknown rescuing
339 alleles (Fig. 4i and Supplementary Data 3). Together this suggested that *HPA3* also rescues
340 incompatible *HPA1* and *HPA2* allele combinations in natural populations and that the locus
341 of *HPA3* contains an additional haplotype (hereafter named *HPA4*) which also rescues the
342 incompatibility between *HPA1* and *HPA2*.
343 We continued to apply the same approach to the other three incompatibilities. For *TIM22*,

344 16 accessions of the 1001 Genome Project revealed non-functional allele combinations
345 (Supplementary Table 11 and Supplementary Data 4), again indicating the existence of
346 modifying alleles. However, a GWAS using the presumably-incompatible allele combinations
347 as molecular phenotype did not reveal only one, but numerous significantly associated loci
348 (Supplementary Fig. 7). This might be explained by the low number of incompatible allele
349 carriers, which could affect the power of association mapping leading to false-positive
350 associations. However, even though we could not locate the modifying allele, further analysis
351 of read mapping coverage in *TIM22* revealed that all 16 accessions with non-functional
352 *TIM22-1/TIM22-2* allele combinations carried at least one additional third copy of *TIM22*, while
353 among all 1,119 other genomes of the 1001 Genomes Project only three genomes carried
354 additional copies. This suggests that, like for *HPA*, known non-functional allele combinations
355 of *TIM22* are in fact rescued by additional copies. Moreover, in ten of the 16 accessions,
356 *TIM22-2* showed not only one but multiple additional copies (hereafter named as *TIM22-3*),
357 which was further supported by the genome assembly of Ty-1
358 (https://genomevolution.org/CoGe/GenomeInfo.pl?gid=54584, unpublished), where we could
359 find a cluster of four tandemly arranged *TIM22* gene copies at the *TIM22-2* locus.
360 Because the reference sequence only contained one copy of *TAD3 (TAD3-1)* (Agorio et
361 al. 2017) and *FOLT (FOLT1)* (Durand et al. 2012), we modified our GWA method and only
362 used non-functional alleles at the reference gene as the phenotype to map modifiers for the
363 two remaining incompatibilities (Fig. 5 and Supplementary Fig. 8). Due to this modification we
364 would expect to map also the location of the duplicated genes as we had found them in the
365 AMPRIL founders. For *TAD3*, an essential ortholog of the yeast tRNA Adenosine deaminase
366 3 (Gerber and Keller 1999), we did not find any accessions with LoF alleles, but we found 150
367 accessions with a hypermethylated promoter in *TAD3-1* similar to the methylated promoters
368 found in Nok-1 and Est-1, which are known to be non-functional due to this methylated
369 promotor region (Agorio et al. 2017) (Fig. 5a, b). The GWA result revealed one significant
370 peak (Fig. 5c,d), which as expected co-located with the *TAD3-2* locus (Agorio et al. 2017). All
371 the 150 accessions carried multiple additional copies of *TAD3-2* (Supplementary Data 5),
372 similar to Nok-1 and Est-1. This suggests that the expression silencing of *TAD3-1* is common
373 in natural populations and that the rescue of this loss-of-function allele is generally mediated
374 by additional gene copies at the *TAD3-2* locus as it was shown in the original description of
375 this incompatibility (Agorio et al. 2017).

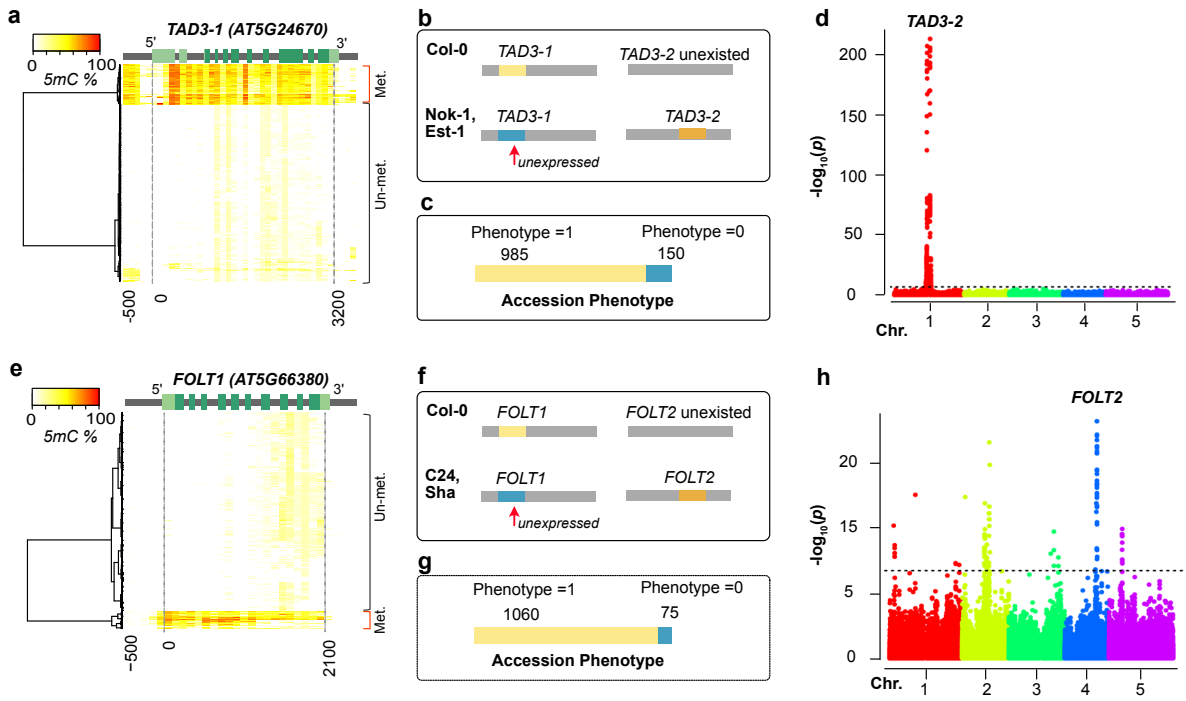
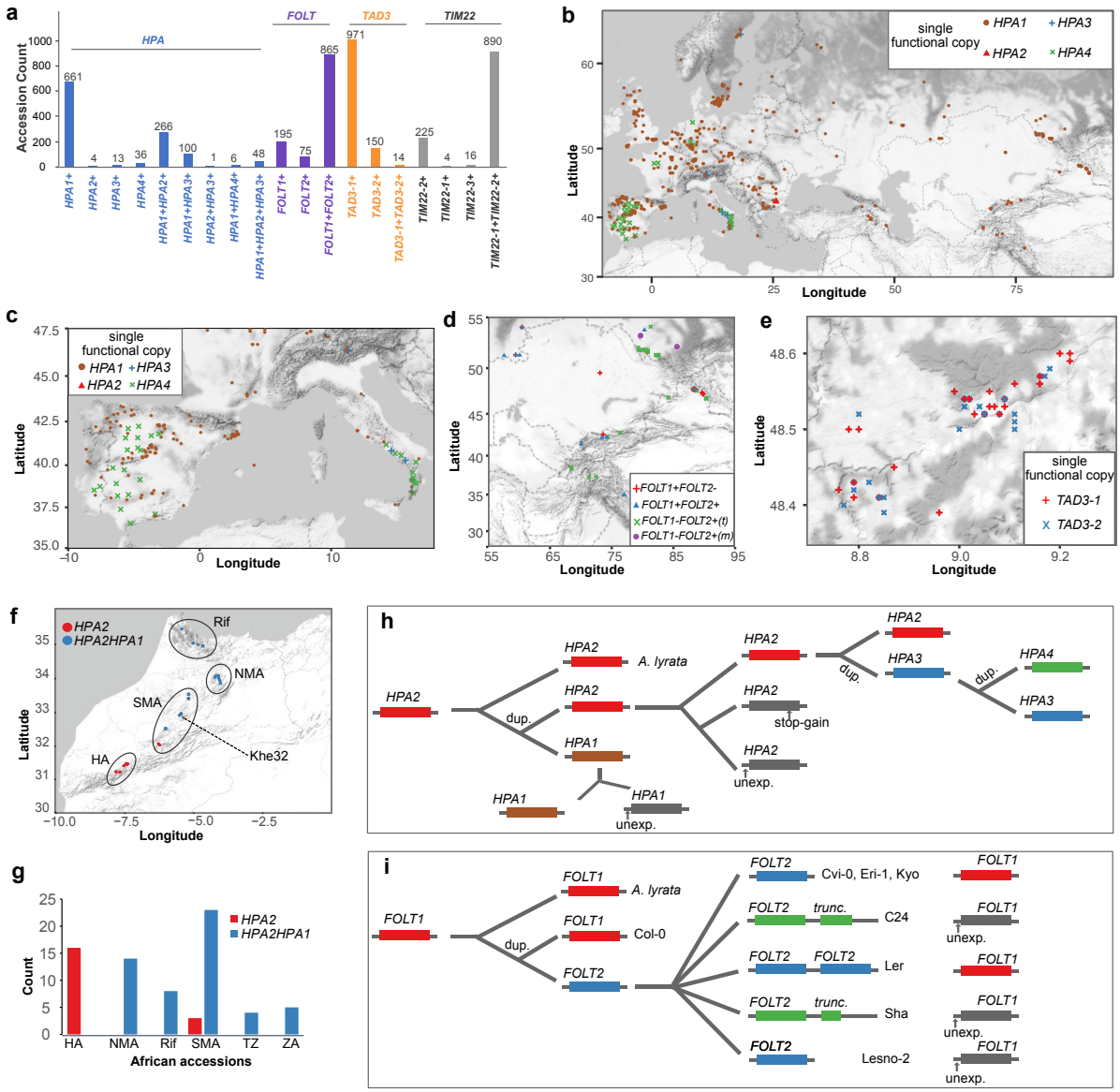


Figure 5: Mapping non-reference gene copies of incompatible alleles using GWAS. (a, e) Hierarchical clustering of cytosine methylation profiles in *TAD3-1* (a) and *FOLT1* (e) based on 888 *A. thaliana* accessions from 1001 Epigenomes Project (Kawakatsu et al. 2016). The methylation profile was calculated based on 100 bp sliding windows from the 500 bp upstream of transcription start site to the 300 bp downstream of the transcription end site**.** Light green: UTR, green: coding region, gray: intron or up/down-stream. **(b, f)** The genetic incompatibilities introduced by *TAD3* (b) in hybrids between Col-0 and Nok-1/Est-1, and by *FOLT* (f) in hybrids between Col-0 and C24/Sha as shown previously (Durand et al. 2012; Agorio et al. 2017). **(c, g)** The number of accessions with functional (Phenotype=1) or non-functional (Phenotype=0) *TAD3-1* (c) and *FOLT1* (g) gene copies. **(d, h)** Manhattan plots of the GWA using the absence of functional *TAD3-1* (d) or *FOLT1* (h) gene copies as phenotype. The dashed line indicates the significance threshold after multiple testing correction ($-\log_{10}P$=6.68).

The last of the four genetic incompatibilities, introduced by *FOLT* encoding for a folate transporter, was previously discovered in hybrids from crosses between Col-0 x C24/Sha (Törjék et al. 2006; Simon et al. 2008; Durand et al. 2012). Col-0 has only one copy of *FOLT* (*FOLT1*) at chromosome 5, whereas the C24 and Sha have an additional copy, *FOLT2,* at chromosome 4 including some extra truncated copies near *FOLT2* (Supplementary Table 12). In the earlier study, the truncated copies were shown to express siRNAs and activate the RNA-directed DNA methylation pathway to silence *FOLT1* in C24 and Sha (Durand et al. 2012), which resulted in a lethal allele combination in F2 hybrids between Col-0 ($FOLT1^{+/+}$) and C24/Sha ($FOLT1^{-/-}$ $FOLT2^{+/+}$) (Fig. 5f). Analyzing the accessions of the 1001 Genome Project for functional and non-functional alleles of *FOLT*, we found 75 accessions with

400 methylated promoters of *FOLT1* likely leading to expression silencing (Fig. 5e), which was
401 supported by the lack of *FOLT1*-specific pseudo-heterozygous SNPs in the RNA-seq data
402 (Supplementary Fig. 9). Besides evidencing the expression silencing of *FOLT1*, this also
403 suggested the existence of additional *FOLT* gene copies in these 75 accessions. When we
404 repeated our GWA approach to find these interacting loci of *FOLT1*, we found multiple
405 significantly associated loci including one region corresponding to *FOLT2* (Fig. 5g, h).
406 Further analyses of the *FOLT* gene copies using short read mapping against the C24 and
407 Sha reference sequences revealed the presence of *FOLT2* in all the 75 accessions with
408 methylated promoter of *FOLT1* and truncated copies of *FOLT2* genes in only 46 out of such
409 75 accessions (Fig. 5e, Supplementary Table 13 and Supplementary Data 6) suggesting that
410 the methylation of the *FOLT1* promoter remains stable even after the truncated copies are
411 segregated out. This is coherent to what was shown in successive generations of Sha x Col-
412 0 RILs where the inducing locus was segregated away six generations ago (Durand et al.
413 2012).
414 Taken together we found evidence that all four incompatible allele combinations, initially
415 identified within artificial intercross population, also occur in nature, and that the incompatible
416 allele combinations of three of them were surprisingly common. For all incompatible allele
417 combination carriers, we found evidence for the existence of additional copies, and could
418 even map the locations of some of those using GWA.
419
420 **Geographic distribution of incompatible allele combinations**
421 We next asked how prevalent the potential for incompatible allele combinations was
422 within the natural population of *A. thaliana* by analyzing the presence of different haplotypes
423 in different geographic regions. For this, we first analyzed the allele frequencies of different
424 haplotypes (i.e. the different combinations of functional alleles within individual plants), which
425 varied substantially across the accessions of the 1001 Genomes Project (Fig. 6).
426 For example, among the accessions that carried only one functional copy of *HPA*, we
427 found that 661, 4, 13 and 36 accessions with a single functional copy of *HPA1*, *HPA2*, *HPA3*,
428 or *HPA4*, respectively (Fig. 6a, b, Supplementary Fig. 10). Similarly, most accessions only
429 featured one functional copy of *TAD3* including 971 accessions with only *TAD3-1* and 150
430 with only *TAD3-2* (Fig. 6a, Supplementary Table 14 and Supplementary Fig. 11). In contrast,
431 most of the accessions included multiple functional alleles of *FOLT* as well as of *TIM22* (Fig.
432 6a).
433 Genetic incompatibilities become effective in the offspring of accessions with
434 incompatible alleles. In particular the offspring of accessions with only one functional gene
435 copy will lead to the highest reduction in fitness. We found numerous accessions with the
436 potential to create incompatible allele combinations in their offspring in geographically close

regions (Fig. 6c, d, e). For example, among 30 accessions in the South of Italy we found 9 accessions with only a functional *HPA1* closely located with 14 accessions with either a functional *HPA3* or a functional *HPA4* only (Fig. 6c). Likewise, accessions with contrasting functional copies of *FOLT* and *TAD3* were collected from the same regions in Central Asia and Germany (Fig. 6d and e) evidencing that incompatible alleles could segregate within the same populations.



445 **Figure 6. Distribution, origin and evolution of incompatible alleles. (a)** The number of accessions
446 with different functional copies of *HPA*, *FOLT*, *TAD3* and *TIM22* across 1,135 *A. thaliana* accessions
447 from the 1001 Genomes Project. +: functional. **(b)** Geographic distribution of accessions only with one
448 functional copy of *HPA*. **(c, d, e)** Examples of geographically close accessions with only one functional
449 copy of *HPA* in Southern Europe (c), *FOLT* in Central Asia (d) and *TAD3* in Germany (e). *FOLT1-*
450 *FOLT2*+(t): accessions with a functional *FOLT2*, a truncated *FOLT2* and an unexpressed *FOLT1*.
451 *FOLT1-FOLT2*+(m): accessions with a functional *FOLT2* and an unexpressed *FOLT1*, but without
452 truncated *FOLT2* copies. **(f, g)** Gene copies of *HPA* in African *A. thaliana* accessions. TZ: Tanzania,
453 ZA: South Africa. **(h, i)** Schematic of a possible and parsimonious evolutionary history of gene
454 duplication and non-functionalization of *HPA* (h) and *FOLT* (i). unexp.: unexpressed. stop-gain:
455 premature stop codon gained SNP. dup.: gene duplication.
456
457 Interestingly, within some of these hybrid zones of accessions with incompatible alleles,
458 we also found accessions with multiple functional gene copies. For example, in South Italy
459 there were 7 of the 30 accessions that featured multiple functional copies of *HPA,* and
460 similarly in Central Asia, we could find multiple accessions with two functional copies of *FOLT*
461 in addition to the accessions with only one functional copy (Fig. 6d and Supplementary Fig.
462 10-12). In contrast, however, we only observed a few accessions with multiple functional
463 alleles in the hybrid zone of incompatible *TAD3* alleles, most likely because only 14 such
464 accessions were found in the entire set of accessions. Even though we could not find
465 evidence that such hybrid zones are enriched for additional copies or that those additional
466 copies would evolve in these regions, the presence of haplotypes with additional gene copies
467 in these hybrid zones have the potential to mediate the gene flow between the haplotypes
468 with the incompatible allele combinations.
469
470 **Origin and evolution of incompatible alleles**
471 To figure out when these incompatible alleles originated and how they evolved, we
472 investigated their ancestral genotypes within African genotypes which likely represent the
473 ancestral populations of the Eurasian accessions (Durvasula et al. 2017).
474 While almost all (99.7%) of the Eurasian accessions had both *HPA2* and *HPA1* (either
475 functional or not), we found that only 27% (20 out of 75) of the African accessions carried only
476 *HPA2* (Fig. 6f, g and Supplementary Data 3). This suggested that *HPA2* was the ancestral
477 copy of *HPA* which was further supported by the synteny alignment with the close relative
478 *Arabidopsis lyrata* (Blevins et al. 2017), which only featured a single *HPA* gene in a region
479 syntenic to *HPA2* and that the duplication events leading to *HPA1* happened early on in Africa
480 (Fig. 6h). One accession, Khe32, from Morocco carried *HPA1* along with the most frequent
481 LoF variant LoF-300 in *HPA2* in the Eurasian accessions, suggesting that the first accession
482 with only a functional *HPA1* possibly arose in North-West Africa and thereby suggests that

483 the genetic incompatibility with *HPA2* carriers was already possible before the Eurasia
484 colonization of *A. thaliana*.
485 In contrast, *HPA3* and *HPA4* only occur in Eurasian accessions, most likely indicating
486 that the additional duplication events happened later (Fig. 6h). Comparing the gene
487 sequences of the *HPA* genes revealed that *HPA3* was duplicated from *HPA2* (Supplementary
488 Fig. 13), while *HPA4* might be a tandem duplicate of *HPA3* as it is likely located close to *HPA3*.
489 Interestingly, *HPA3* and *HPA4* carriers segregated for the ancestral LoF-300 variant at *HPA2*
490 (*HPA3*: 68 with and 94 without; *HPA4*: 10 with and 32 without, Supplementary Fig. 14),
491 suggesting free segregation of different *HPA* alleles. In striking contrast to this, the inactive
492 alleles of *HPA1* (i.e. alleles with a non-functional *HPA1* sequence, but excluding the
493 accessions with full deletion alleles as found in Cvi-0) were almost perfectly coupled (48 of
494 50) with inactive alleles of *HPA2*. Such carriers nearly all had *HPA4* (35) or *HPA3* (13)
495 (Supplementary Fig. 15) and were mainly located in the Iberian Peninsula and Southern Italy
496 suggesting that the additional *HPA* copies were necessary to buffer the incompatibility and to
497 allow the foundation of these populations (Fig. 6c).
498 Unlike the incompatibility in *HPA*, genetic incompatibilities can also arise in recent
499 population history. All African accessions only had the *TAD3-1* copy, while accessions with
500 multiple copies of *TAD3* can only be observed in Eurasia (Supplementary Fig. 10). Similarly,
501 38 of the African accessions only have *FOLT1*, while the other 37 accessions have both
502 *FOLT1* and *FOLT2*, but none of the accessions featured a truncated copy of *FOLT2* (*FOLT2tr*),
503 which is the mechanistic origin of the incompatible alleles at *FOLT1* and *FOLT2*
504 (Supplementary Fig. 16 and Supplementary Data 6). In contrast, within the Eurasian
505 accession, we even found three different haplotypes of *FOLT2tr* within a total of 46 accessions
506 based on short read alignments against the eight *A. thaliana* genomes (Jiao and
507 Schneeberger 2020) (Supplementary Fig. 17 and Supplementary Data 6). This together
508 suggests that also the genetic incompatibility that is based on *FOLT2tr* has evolved recently
509 after the migration *A. thaliana* to Eurasia (Fig. 6i).
510

511 ## Discussion

512 Although gene duplication provides genetic backup of essential genes, duplicated genes
513 can also lead to incompatible allele combinations when the duplicated genes undergo
514 reciprocal pseudo-functionalization in separate genomes. Here we studied incompatible allele
515 combinations of four duplicated gene pairs by integrating genetic and genomic information
516 using a multi-parental intercross population leading to the identification of four genetic
517 incompatibilities.
518 Unexpectedly in this population we identified some lines with what initially looked like an

519 incompatible allele combination of one of the incompatibilities, as they could be rescued by
520 an additional, so-far unknown gene copy. Encouraged by this, we developed a GWA method
521 to map modifying alleles also in natural populations by integrating genomic and epigenomic
522 data to generate a molecular phenotype that describes potentially incompatible allele
523 combinations. With this we identified many natural accessions with putatively incompatible
524 allele combinations and could elucidate the genetics of these incompatibilities as they occur
525 in natural population using the 1,135 accessions released by the 1001 Genomes Project
526 (Alonso-Blanco et al. 2016). This implies that besides presence/absence analysis as
527 performed by standard pan-genome analysis, gene location is an essential feature of a gene
528 (or gene family).

529 Based on multi-omics data from hundreds of *A. thaliana* accessions, we could
530 comprehensively describe the origin and evolution of several incompatible allele combinations
531 of the four incompatibilities. We found that incompatible alleles are surprisingly frequent in
532 nature and also occur in sympatry, suggesting that the evolution of genetic incompatibilities
533 does not require separated populations as proposed by BDM model (Bateson 1909;
534 Dobzhansky 1937; Muller 1942). Moreover, such incompatible allele combinations can persist
535 over long periods as some of the alleles studied here may have originated before *A. thaliana*
536 colonized Eurasia.

537 Here, new gene copies arose either from distal duplications (in *HPA* and *FOLT*) or
538 tandem duplications (in all four cases) and counteract incompatibilities. Even though
539 additional copies reduce the frequency and impact of genetic incompatibilities, they could also
540 increase the potential for more incompatible allele combinations. Subsequent sub-
541 functionalization could be a way out of the trap of genetic incompatibilities, as there would be
542 a selective pressure to keep both gene copies. While we have assumed functional
543 redundancy of all gene copies in this work, an extensive number of accessions shared the
544 functional copies of bot *FOLT* genes (865 of 1,135), which could indicate that these copies
545 are not fully redundant and thereby limit the establishment of incompatible allele combinations.

546 Taken together, our work demonstrates that the potential for genetic incompatibilities due
547 to duplicated essential genes is surprisingly high in nature. However, the effects of such
548 incompatibilities are counteracted by additional gene copies, which undergo dynamic
549 changes shaped by the recurrent events of gene duplication and non-functionalization during
550 population history.

551

552 # Materials and Methods

553 **AMPRIL construction**

554 The eight *A. thaliana* accessions An-1, C24, Col-0, Cvi-0, Eri-1, Kyo, L*er*, and Sha were

555 selected as the AMPRIL founders. We previously constructed a first version of AMPRIL
556 population (AMPRIL I) including six RIL subpopulations (Huang et al. 2011). Here AMPRIL I
557 was extended with six additional subpopulations (called AMPRIL II) based on different
558 pairwise intercrosses of the eight founders (Fig. 1a). Each subpopulation contains
559 approximately 90 individuals (Supplementary Data 1). All plants were grown under the normal
560 growth conditions in greenhouse at the MPI-PZ (Huang et al. 2011). DNA of 1,100 samples
561 was extracted from the flower buds and prepared for RAD-seq sequencing (Baird et al. 2008).
562
563 **RAD-seq Library preparation and sequencing**
564 All plants were grown in the greenhouse. The total DNA from each of the 1,100 samples
565 was extracted from the flower buds using DNeasy Plant Kit 96 (Qiagen) and eluted in 200 µl
566 Elution buffer (EB). DNA of each genotype was isolated twice. Both genotype samples were
567 pooled in 1.5 ml tube and the DNA was concentrated by isopropanol precipitation for 2h at -
568 20°C. The samples were centrifuged at 12000 *g* for 5 min at 4°C, the supernatant was
569 removed and the DNA pellet was washed with ice-cold 70% ethanol. Centrifugation was
570 repeated, the supernatant was removed. Air-dried DNA was resuspended in a nuclease free
571 water to 26 ng/µl and stored at -20°C until use. RAD-seq sequencing libraries were prepared
572 as described (Etter et al. 2011) with modifications. Per genotype, 500 ng DNA were digested
573 with 10 units of CviQI (NEB, cutting site G'TAC) at 25°C for 2 h. The number of expected
574 cutting sites was estimated around 236,000 based on the Col-0 genome sequence. Cut DNA
575 was purified using 96 DNA clean and concentrator kit (Zymo) a diluted in 25 µl EB. The 192
576 different (selected out of total 210 designed, Supplementary Data 1) P1 adapters (200 nM)
577 containing unique 12 bp barcodes were ligated by incubation with T4 ligase (NEB) at room
578 temperature for 30 min and the reaction was terminated by 20 min at 65°C. After 30 min at
579 room temperature, 5 µl from each 192 P1-barcoded sample were combined in a 2 ml low bind
580 tube. 3 x 130 µl aliquots were transferred to fresh tubes and DNA was fragmented to average
581 size of 500 bp using Covaris. Sheared DNA was purified using QiaMinElute columns (Qiagen),
582 eluted in 10 µl EB, and the three samples were first pooled and then divided into two 15 µl
583 samples that were run on 1% agarose gel. Regions of 300-500 bp fragments were dissected
584 and DNA was isolated using MinElute gel extraction kit (Quiagen), eluted in 10 µl EB, the
585 samples were pooled, DNA fragment ends were repaired using Quick Blunting™ kit (NEB),
586 purified with QIAquick column (Qiagen) and eluted in 43 µl EB. 3' deoxy-adenine overhangs
587 were added using Klenow Fragment (NEB), the sample was purified with QIAquick column,
588 eluted in 45 µl EB, the P2 adapter was ligated, the sample was purified with QIAquick column
589 and eluted in 53 µl EB. To determine the library quality, 10 µl RAD library were PCR amplified
590 (1: 98°C 30 sec; 2: 14x 98°C 10 sec, 67°C 30 sec, 68°C 30 sec; 3: 68°C 5min, 4: 4°C hold)
591 using NEB Next High-Fidelity master mix (NEB) in 25 µl reaction volume using RAD-Marker

592 for/ RAD-Marker rev primers (25 nM each) and 5 µl PCR product were loaded into 1% agarose
593 gel next to the 1 µl RAD library template. If the PCR product smear was at least twice as
594 intense as the template smear, the library was considered as of high quality and the
595 amplification was repeated in 50 µl reaction volume. The product was cleaned using AMPure
596 magnetic beads (Beckman Coulter) and dissolved in 20 µl EB. Finally, the sample was run on
597 1% agarose gel, the region of 300-500 bp fragments was cut out, DNA was isolated using
598 MinElute Gel extraction kit (Qiagen) and eluted in 20 µl EB. The library was sequenced in an
599 Illumina Hiseq2000 sequencing machine. Sequencing reads were demultiplexed according
600 to the barcodes (Supplementary Data 1).
601
602 **AMPRIL genotyping**
603 We used previously released whole-genome short read data(Jiao and Schneeberger
604 2020) to generate markers for genotyping the AMPRILs. We mapped the reads to the
605 reference sequence and called SNPs using SHORE (version 0.9) (Ossowski et al. 2008) with
606 default parameter settings. Only homozygous SNP calls were selected after removing SNP
607 calls with low quality (quality < 30), in the repetitive regions or in regions with low mapping
608 quality (quality < 30). The actual marker sets for each of the twelve subpopulations (ABBA,
609 ACCA … GHHG) were selected based on respective parental genomes, exclusively selecting
610 bi-allelic SNP markers. After excluding samples with replicates or too few sequencing reads,
611 we performed genotyping on 992 AMPRILs using a Hidden Markov Model-based approach
612 similar to the recently presented method for the reconstruction of genotypes derived from two
613 parental genomes (Rowan et al. 2015) (for a detailed description see Supplementary Note 1).
614
615 **Identifying genetic incompatibilities based on duplicated genes**
616 Duplicated gene pairs were selected based on gene family clustering of protein-coding
617 genes from all eight parental genomes using OrthoFinder (version 2.2.6) (Emms and Kelly
618 2015). We only selected inter-chromosome duplicated genes to avoid the effects of intra-
619 chromosome linkage (Supplementary Data 2). For each duplicated gene pair, we required
620 two copies in the reference sequence and at least one copy in one of the other genomes
621 (DupGene2), or one copy in reference sequence and at least one copy on a different
622 chromosome in at least one of the other parental genomes (DupGene1). We assessed the
623 genotypes of each of the gene copies in each AMPRIL using the genotypes predicted in the
624 middle of the respective reference gene, or in the case a gene was not present in the
625 reference sequence – using the midpoint between the two closest flanking syntenic regions
626 of the non-reference gene copy (based on synteny calculations from a previous study (Jiao
627 and Schneeberger 2020)).
628 We predicted candidate genetic incompatibilities using a two-steps approach. In the first

629 step, we performed chi-square tests (Equation 1) to check whether the frequency of allele
630 pairs in duplicated genes was significantly distorted in any of the subpopulations or in any of
631 two merged subpopulations (ABBA and EFFE or CDDC and GHHG which shared the same
632 four founders, respectively).

633 $$\chi^2 = \sum_{i,j \in \{a,b,c,d\}} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \quad (1)$$

634 Here, the o_{ij} and e_{ij} represent the observed and expected allele pair frequency of
635 duplicated genes, respectively, and *a*, *b*, *c*, *d* represent the parental genotypes in each
636 subpopulation.

637 Additionally, we applied a modified chi-square test (Equation 2) for all the duplicated
638 genes in the whole AMPRIL population by considering the effects of population structure.

639 $$x^2 = \sum_{i,j \in \{a,b,c,d,e,f,g,h\}} \frac{(\sum o_{ij} - \sum e_{ij})^2}{\sum e_{ij}} \quad (2)$$

640 Here, the o_{ij} and e_{ij} represent the observed and expected allele pair frequency of inter-
641 chromosome duplicated genes, respectively, and *a*, *b*, *c*, *d, e, f, g,* and *h* represent all parental
642 genotypes. The observed and expected allele pair frequency in the whole population was the
643 sum of observed ($\sum o_{ij}$) and expected ($\sum e_{ij}$) allele pair frequencies in each of the
644 subpopulations.

645 The gene pairs with at least one significant segregation distortion in their observed allele
646 combinations (FDR < 0.05) were kept for the next step. In this second step, we checked
647 whether the respective gene copies contained loss-of-function variation or methylated
648 promoters (as described below). To address alternative splicing which is known to rescue
649 loss of-function variation, we also checked the gene annotation within the parental genome
650 assemblies to confirm the loss-of-function. Only duplicated genes with confirmed loss-of-
651 function alleles in both copies of the duplication (in at least one of the parental genomes) were
652 kept as candidates for genetic incompatibilities based on duplicated genes.

653

654 **Identification of LoF variants in candidate genes**

655 We mapped all whole-genome resequencing reads of 1,135 accessions from the 1001
656 Genomes Project (Alonso-Blanco et al. 2016), 75 accessions from Africa (Durvasula et al.
657 2017) and 118 accessions from China (Zou et al. 2017), to the Col-0 reference genome
658 (TAIR10) (The Arabidopsis Genome Initiative 2000; Lamesch et al. 2012) using BWA (version
659 0.7.15) (Li and Durbin 2009) with the default parameter settings. SNPs and small indels were
660 called using SAMtools (version 1.9) using default parameters (Li et al. 2009). Homozygous
661 variants with mapping quality of more than 20 and with at least four reads aligned were kept.
662 Pseudo-heterozygous variants in *HPA* and *TIM22* were also recorded. The large insertion
663 and deletions in the 40kb extended genic region of focal genes were predicted using Pindel

664 (version 0.2.5) (Ye et al. 2009) with parameter settings "-T 1 -x 5 -k -r -j" and Delly (version
665 0.8.1) (Rausch et al. 2012) with parameter settings "delly call -q 20 -r 20 -n -u 20 -g". The
666 functional effects of these variations were annotated using SnpEff (version 4.3p) (Cingolani
667 et al. 2012) using the default parameter settings. The loss-of-function (LoF) effects include
668 loss of start codon, loss of stop codon, gain of premature stop codon, damage of splicing
669 acceptor or donor sites, frameshift and CDS loss.

670 **Clustering of cytosine methylation profile in gene promoter**

671 Cytosine methylation data (the tab separated file of methylated cytosine positions) of
672 1,211 samples from the 1001 Epigenomes project (Kawakatsu et al. 2016) were downloaded
673 from NCBI (927 samples under GEO accession GSE43857 and 284 samples under GEO
674 accession GSE54292). After removing the redundant data sets 888 and 161 data sets from
675 GSE43857 and GSE54292, respectively, were retained. For each sample, we calculated the
676 percentage of methylated cytosines in CG, CHG and CHH contexts from 500bp upstream of
677 the transcription start sites to 300bp downstream of the transcription termination sites of each
678 candidate gene in 100bp non-overlapping sliding windows. These methylation profiles were
679 hierarchically clustered using the hclust function implemented in R (version 3.5.1). The
680 pairwise, Euclidean distances between all methylation profiles were calculated and Ward's
681 method was used to cluster the samples into two groups (hypermethylated and unmethylated).
682 This clustering was performed for the samples of GSE43857 and GSE54292 separately as
683 these two data sets were processed in two studies with different pipelines (Dubin et al. 2015;
684 Kawakatsu et al. 2016). The heatmap of methylation patterns was drawn in R using the
685 heatmap.2 function.

686

687 **Analysis of DNA methylation within the genomes of the AMPRIL founders**

688 For six accessions (Col-0, An-1, C24, Cvi-0, Ler, Kyo), we downloaded the whole-
689 genome DNA methylation data from NCBI from the 1001 Epigenomes Project (Kawakatsu et
690 al. 2016) (GSE43857). For Eri-1 and Sha, DNA methylation data was generated using whole-
691 genome bisulfite sequencing by the Max Planck Genome center. DNA was extracted from
692 plants grown in the greenhouse under standard conditions using the Qiagen DNEasy Plant
693 Mini Kit (Qiagen, Germany) and a sequencing library was prepared using the NEXTflex
694 Bisulfite Library Prep Kit. This library was sequenced on an Illumina HiSeq2000 machine.
695 Sequencing reads were aligned the reference sequence using Bismark (version 0.20.0)
696 (Krueger and Andrews 2011) with these parameters "-q --bowtie2 -N 1 -L 24 -p 20". The
697 cytosine methylation profiles in candidate genes were calculated using the same sliding
698 window method as described above. The cytosine methylation profiles together with the
699 profiles from GSE43857 were then clustered again with the same clustering method as
700 described above.

701

702 **Mapping modifiers of incompatible alleles using GWA**

703 We predicted the presence and absence of functional copies of duplicated genes (*HPA*,
704 *TIM22*) in each of 1,135 accessions from the 1001 Genomes Project according to the
705 annotations of loss-of-function variations and the clustering of cytosine methylation profiles in
706 promoters (Supplementary Data 3-6). For accessions without available methylation
707 sequencing data, we assume the focal genes are expressed. The presence or absence of
708 any functional copies of the reference genes were used as binary phenotype (presence: 1,
709 absence: 0). For the association, we selected 238,166 high-quality SNP markers (minor allele
710 frequency >0.05 and missing rate < 0.1) from 1001 Genomes Project and imputed missing
711 alleles with IMPUTE2 (Howie et al. 2009; Howie et al. 2012). An in-house R script
712 implementing the mixed linear model with correction of kinship bias was used to perform the
713 GWA (Bonferroni correction, P < 0.05).

714

715 **Copy number variation analysis**

716 To test for the existence of *HPA3* in a genome, we mapped whole-genome short reads
717 of 1,135 accessions from 1001 Genomes Project (Alonso-Blanco et al. 2016), of 75
718 accessions from Africa (Durvasula et al. 2017) and of 118 accessions from China (Zou et al.
719 2017) to the Kyo genome assembly (Jiao and Schneeberger 2020) using BWA (version
720 0.7.15) with default parameters. Accessions with an average mapping coverage >= 5 along
721 the *HPA3* region and duplicated region breakpoints (identified based on the sequence
722 alignment against Col-0 genome using SyRI (Goel et al. 2019) (version 1.0) with default
723 parameters) were considered as *HPA3* carriers.

724 The copy number of duplicated genes was estimated by the ratio between the average
725 mapping coverage within the focal gene and the average mapping coverage across the whole
726 genome. To estimate the copy number of *HPA*, we mapped the short reads to the Cvi-0
727 genome assembly using BWA (version 0.7.15) with default parameters as the Cvi-0 genome
728 only has one copy of *HPA*. For *TAD3*, *FOLT* and *TIM22*, the copy number was predicted
729 based on short reads mapping against the reference genome where both *TAD3* and *FOLT*
730 only have one copy. For the *TIM22*, copy number was estimated based on the average
731 mapping coverage at both *TIM22-1* and *TIM22-2*.

732

733 **Fine-mapping of incompatible allele at LD2 in Cvi-0 x Col-0 RIL**

734 After the observation that it was not possible to obtain a RIL homozygous for the Col-0
735 allele at LD2.1 while homozygous for the Cvi-0 allele at LD2.3 (Simon et al. 2008), we derived
736 two distinct heterogeneous inbred families (HIF) from two segregating RILs from this
737 population (Supplementary Fig. 2). 8RV467 is segregating for a region largely encompassing

738 LD2.1 while fixed Cvi-0 at LD2.3. 8RV408 is segregating for LD2.3 while fixed Col-0 at LD2.1.
739 In each derived HIF family, it is again not possible to fix the incompatible allele combination,
740 so the families were used to exclude intervals that do not contribute to the incompatible
741 interaction. Two rounds of fine-mapping were conducted by genotypically screening
742 increasing populations of descendants (at the seedling stage) for recombinants in the interval.
743 Gradually, the causal interval is reduced and delineated by markers exploiting known SNPs
744 and indels in the Cvi-0 sequence.
745
746 **Segregation of T-DNA mutant line at gene *TIM22-1* and *TIM22-2***
747 T-DNA lines GABI_848H04 segregates for an insertion in *TIM22-1*, while GABI_626H10
748 segregates for an insertion in *TIM22-2*, both in a Col-0 background. Descendance were
749 screened genotypically at the seedling stage to characterize the segregation of the T-DNA
750 insertion allele.
751
752 **Complementation cross to validate the incompatible alleles of *TIM22***
753 A LD2.3 HIF line (8HV408-Het) was crossed with a GABI_626H10 line, both at the
754 heterozygous state, in order to segregate potentially for all four possible hybrid allelic
755 combinations at LD2.3, while maintaining a fixed Col-0 allele at LD2.1. The F1 descendance
756 (123 individuals, from both cross directions) was screened genotypically at the seedling stage
757 for the presence/absence of both the T-DNA insertion and the Cvi-0 allele.

Data availability

Raw RAD-seq data of the AMPRIL population was deposited to European Nucleotide Archive (ENA) under the project accession ID PRJEB39883. Genome resequencing data of all eight founders and BS-seq data of Eri-1 and Sha can be accessed in ENA under the project accession ID PRJEB31147 and PRJEB38624. Genome resequencing data generated in 1001 Genomes Project (Alonso-Blanco et al. 2016) was downloaded from NCBI under the project ID SRP056687. RNA-seq and methylation data from the 1001 Epigenomes Project (Kawakatsu et al. 2016) were downloaded from NCBI under the project accession ID GSE80744, GSE54680, GSE43857 and GSE54292. The SNP markers from the 1001 Genomes Project were downloaded from https://1001genomes.org/data/GMI-MPI/releases/v3.1/.

Code availability

Custom code used in this study can be freely accessed at *https://github.com/schneebergerlab/AMPRIL-GI*

Seed availability

We are currently preparing seeds of new AMPRIL populations for submission to NASC. (http://arabidopsis.info/). Note, heterozygous regions as reported in the genotype data might be fixed for one of the alleles in the requested seeds.

Acknowledgements

The authors would like to thank Detlef Weigel (Max Planck Institute for Developmental Biology, MPI-EB), Felix Bemm (MPI-EB), Todd Michael (Salk Institute), Joe Ecker (Salk Institute), Florian Jupe (Salk Institute) for releasing Ty-1 assembly prior to publication, Mehmet Goktay (Max Planck Institute for Plant Breeding Research, MPI-PZ) for help with the interpretation of genetic variation in Africa, Pádraic J. Flood (MPI-PZ) for helpful comments on the manuscript, the Max Planck Genome center for their sequencing efforts and the 1001 Genomes (1001genomes.org) and Epigenomes consortia for releasing their data. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy – EXC 2048/1– 390686111 (K.S.), and the European Research Council (ERC) Grant “INTERACT” (802629) (K.S.).

Authors contributions

WBJ and KS designed this project. WBJ, VP, JK, and FL analyzed the data. PP and AP

generated the RAD-seq data. OL, MF, IG and CC performed the mapping and complementation experiments of LD2/*TIM22*. SE and MK generated the AMPRIL population. WBJ and KS wrote the paper.

Competing interests

The authors declare no competing interests.

References

Ackermann M, Beyer A. 2012. Systematic detection of epistatic interactions based on allele pair frequencies. *PLoS Genet.* 8:e1002463.

Agorio A, Durand S, Fiume E, Brousse C, Gy I, Simon M, Anava S, Rechavi O, Loudet O, Camilleri C, et al. 2017. An Arabidopsis Natural Epiallele Maintained by a Feed-Forward Silencing Loop between Histone and DNA. *PLoS Genet.* 13:1–23.

Alonso-Blanco C, Andrade J, Becker C, Bemm F, Bergelson J, Borgwardt KMM, Cao J, Chae E, Dezwaan TMM, Ding W, et al. 2016. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell* 166:481–491.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* 3:1–7.

Bateson W. 1909. Heredity and variation in modern lights. In: Darwin and Modern Science: Essays in Commemoration of the Centenary of the Birth of Charles Darwin and of the Fiftieth Anniversary of the Publication of the Origin of Species. In: Seward AC (ed). Darwin and Modern Science. Cambridge University. p. 85–101.

Bikard D, Patel D, Le Mette C, Giorgi V, Camilleri C, Bennett MJ, Loudet O. 2009. Divergent Evolution of Duplicate Genes Leads to Genetic Incompatibilities Within A. thaliana. *Science* 323:623–626.

Blevins T, Wang J, Pflieger D, Pontvianne F, Pikaard CS. 2017. Hybrid incompatibility caused by an epiallele. *Proc. Natl. Acad. Sci.* 114:3702–3707.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin).* 6:80–92.

Conant GC, Wolfe KH. 2008. Turning a hobby into a job: How duplicated genes find new functions. *Nat. Rev. Genet.* 9:938–950.

Corbett-Detig RB, Zhou J, Clark AG, Hartl DL, Ayroles JF. 2013. Genetic incompatibilities are widespread within species. *Nature* 504:135–137.

Dobzhansky T. 1937. Genetics and the Origin of Species. New York.: Columbia University Press

Dubin MJ, Zhang P, Meng D, Remigereau MS, Osborne EJ, Casale FP, Drewe P, Kahles A, Jean G, Vilhjálmsson B, et al. 2015. DNA methylation in Arabidopsis has a genetic basis and shows evidence of local adaptation. *Elife* 4:1–23.

Durand S, Bouché N, Perez Strand E, Loudet O, Camilleri C. 2012. Rapid establishment of genetic incompatibility through natural epigenetic variation. *Curr. Biol.* 22:326–331.

835 Durvasula A, Fulgione A, Gutaker RM, Alacakaptan SI, Flood PJ, Neto C, Tsuchimatsu T,
836 Burbano HA, Picó FX, Alonso-Blanco C, et al. 2017. African genomes illuminate the early
837 history and transition to selfing in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci.* 114:5213–
838 5218.
839 Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome
840 comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.*
841 16:157.
842 Etter PD, Bassham S, Hohenlohe PA, Johnson EA, Cresko WA. 2011. SNP Discovery and
843 Genotyping for Evolutionary Genetics Using RAD Sequencing BT - Molecular Methods
844 for Evolutionary Genetics. In: Orgogozo V, Rockman M V, editors. Totowa, NJ: Humana
845 Press. p. 157–178.
846 Fishman L, Sweigart AL. 2018. When Two Rights Make a Wrong: The Evolutionary Genetics
847 of Plant Hybrid Incompatibilities. *Annu. Rev. Plant Biol.* 69:17.1–17.25.
848 Gerber AP, Keller W. 1999. An adenosine deaminase that generates inosine at the wobble
849 position of tRNAs. *Science* 286:1146–1149.
850 Goel M, Sun H, Jiao W-B, Schneeberger K. 2019. SyRI: finding genomic rearrangements and
851 local sequence differences from whole-genome assemblies. *Genome Biol.* 20:277.
852 Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. 2012. Fast and accurate
853 genotype imputation in genome-wide association studies through pre-phasing. *Nat.*
854 *Genet.* 44:955–959.
855 Howie BN, Donnelly P, Marchini J. 2009. A flexible and accurate genotype imputation method
856 for the next generation of genome-wide association studies. *PLoS Genet.* 5.
857 Huang X, Paulo M-J, Boer M, Effgen S, Keizer P, Koornneef M, van Eeuwijk F a. 2011.
858 Analysis of natural allelic variation in *Arabidopsis* using a multiparent recombinant inbred
859 line population. *Proc. Natl. Acad. Sci. U. S. A.* 108:4488–4493.
860 Jiao W, Schneeberger K. 2020. Chromosome-level assemblies of multiple Arabidopsis
861 genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *Nat.*
862 *Commun.* 11:989.
863 Kawakatsu T, Huang S shan??Carol, Jupe F, Sasaki E, Schmitz RJ, Urich MA, Castanon R,
864 Nery JR, Barragan C, He Y, et al. 2016. Epigenomic Diversity in a Global Collection of
865 Arabidopsis thaliana Accessions. *Cell* 166:492–506.
866 Kawakatsu T, Huang S shan C, Jupe F, Sasaki E, Schmitz RJJ, Urich MA a., Castanon R,
867 Nery JRR, Barragan C, He Y, et al. 2016. Epigenomic Diversity in a Global Collection of
868 Arabidopsis thaliana Accessions. *Cell* 166:492–506.
869 Kondrashov FA. 2012. Gene duplication as a mechanism of genomic adaptation to a
870 changing environment. *Proc. R. Soc. B Biol. Sci.* 279:5048–5057.
871 Krueger F, Andrews SR. 2011. Bismark: A flexible aligner and methylation caller for Bisulfite-

872 Seq applications. *Bioinformatics* 27:1571–1572.
873 Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K,
874 Alexander DL, Garcia-Hernandez M, et al. 2012. The Arabidopsis Information Resource
875 (TAIR): Improved gene annotation and new tools. *Nucleic Acids Res.* 40.
876 Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform.
877 *Bioinformatics* 25:1754–1760.
878 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R.
879 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–
880 2079.
881 Lynch M, Force AG. 2000. The Origin of Interspecific Genomic Incompatibility via Gene
882 Duplication. *Am. Nat.* 156:590–605.
883 Maheshwari S, Barbash DA. 2011. The Genetics of Hybrid Incompatibilities. *Annu. Rev.*
884 *Genet.* 45:331–355.
885 Mizuta Y, Harushima Y, Kurata N. 2010. Rice pollen hybrid incompatibility caused by
886 reciprocal gene loss of duplicated genes. *Proc. Natl. Acad. Sci. U. S. A.* 107:20417–
887 20422.
888 Muller HJ. 1942. Isolating mechanisms, evolution and temperature. *Biol. Symp* 6:71–125.
889 Muralla R, Sweeney C, Stepansky A, Leustek T, Meinke D. 2007. Genetic dissection of
890 histidine biosynthesis in Arabidopsis. *Plant Physiol.* 144:890–903.
891 Murcha MW, Elhafez D, Lister R, Tonti-Filippini J, Baumgartner M, Philippar K, Carrie C,
892 Mokranjac D, Soll J, Whelan J. 2007. Characterization of the preprotein and amino acid
893 transporter gene family in Arabidopsis. *Plant Physiol.* 143:199–212.
894 Nguyen GN, Yamagata Y, Shigematsu Y, Watanabe M, Miyazaki Y. 2017. Duplication and
895 Loss of Function of Genes Encoding RNA Polymerase III Subunit C4 Causes Hybrid
896 Incompatibility in Rice. *G3* 7:2565–2575.
897 Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, Weigel D. 2008. Sequencing
898 of natural strains of Arabidopsis thaliana with short reads. *Genome Res.*
899 Panchy N, Lehti-Shiu M, Shiu S-H. 2016. Evolution of Gene Duplication in Plants. *Plant*
900 *Physiol.* 171:2294–2316.
901 Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. 2012. DELLY: Structural
902 variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*
903 28:333–339.
904 Rowan BA, Patel V, Weigel D, Schneeberger K. 2015. Rapid and inexpensive whole-genome
905 genotyping-by-sequencing for crossover localization and fine-scale genetic mapping. *G3*
906 *Genes, Genomes, Genet.* 5:385–398.
907 Simon M, Loudet O, Durand S, Bérard A, Brunel D, Sennesal FX, Durand-Tardif M, Pelletier
908 G, Camilleri C. 2008. Quantitative trait loci mapping in five new large recombinant inbred

909 line populations of Arabidopsis thaliana genotyped with consensus single-nucleotide
910 polymorphism markers. *Genetics* 178:2253–2264.
911 The Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering
912 plant Arabidopsis thaliana. *Nature* 408:796–815.
913 Törjék O, Witucka-Wall H, Meyer RC, Von Korff M, Kusterer B, Rautengarten C, Altmann T.
914 2006. Segregation distortion in Arabidopsis C24/Col-0 and Col-0/C24 recombinant
915 inbred line populations is due to reduced fertility caused by epistatic interaction of two
916 loci. *Theor. Appl. Genet.* 113:1551–1561.
917 Vaid N, Laitinen RAE. 2019. Diverse paths to hybrid incompatibility in Arabidopsis. *Plant J.*
918 97:199–213.
919 Yamagata Y, Yamamoto E, Aya K, Win KT, Doi K, Sobrizal, Ito T, Kanamori H, Wu J,
920 Matsumoto T, et al. 2010. Mitochondrial gene in the nuclear genome induces
921 reproductive barrier in rice. *Proc. Natl. Acad. Sci. U. S. A.* 107:1494–1499.
922 Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. 2009. Pindel: A pattern growth approach to
923 detect break points of large deletions and medium sized insertions from paired-end short
924 reads. *Bioinformatics* 25:2865–2871.
925 Zou Y-P, Hou X-H, Wu Q, Chen J-F, Li Z-W, Han T-S, Niu X-M, Yang L, Xu Y-C, Zhang J, et
926 al. 2017. Adaptation of Arabidopsis thaliana to the Yangtze River basin. *Genome Biol.*
927 18:239.
928 Zuellig MP, Sweigart AL, States U, E-mail CA. 2017. Gene duplicates cause hybrid lethality
929 between sympatric species of Mimulus. *PLoS Genet.* 14:e1007130.
930