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

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

[1] Max Planck Institute for Plant Breeding Research, Department of Chromosome Biology, Carl-von-Linné-Weg 10, 50829 Cologne, Germany.

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

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

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

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

[6] 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.

[7] Faculty of Biology, LMU Munich, Großhaderner Str. 2, 82152, Planegg-Martinsried, Germany.

[*] Present addresses:

Vipul Patel: AlceDiag, 1692 Rue de la Valsière, 34184 Montpellier, France

Abstract

The fate of duplicated genes includes pseudo-, sub-, or neo-functionalization. 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. Using the Arabidopsis multi-parental RIL population, we identified four duplicated-gene based incompatibilities including one previously not reported. Unexpectedly, however, for each of the genetic incompatibilities we found incompatible allele combinations in natural Arabidopsis accessions. Using the incompatible allele combinations as phenotypes for GWAS, we mapped genomic regions containing additional gene copies, which rescue the incompatible allele combinations. Reconstructing the geographic origins of the individual alleles suggested that incompatible alleles co-exist in nature, 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[1]. 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[2–4], 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[5,6]. Gene duplications can provide genetic backup of essential genes and the basis for evolutionary novelties by allowing for new genetic and epigenetic variations[7–9]. 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[10].

Genetic incompatibilities introduced by duplicated genes have been reported within inter/intra-specific hybrids of *Arabidopsis thaliana*[11–13], rice[14–16] or *Mimulus*[17]. 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[11,18], 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[19] and 1001 Epigenomes Project[20] 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[21] 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[19], where these combinations were surprisingly common. Using the incompatible allele

75 combinations as phenotypes we mapped modifiers of all four incompatibilities using GWA.
76 The loss-of-function alleles from duplicated genes were geographically widely distributed, and
77 co-existed with additional gene copies in the same regions showing how additional gene
78 copies can overcome differential copy loss in a population.
79

80 # Results

81 **Identification of incompatible gene pairs within an intercross population**

82 We used the Arabidopsis Multiparent RIL (AMPRIL) population to find incompatible
83 alleles that arose from duplicated genes. The eight AMPRIL founder accessions (An-1, C24,
84 Col-0, Cvi-0, Eri-1, Kyo, L*er* and Sha) were selected across the entire geographic distribution
85 of *A. thaliana* including the Northern hemisphere and the Cape Verde Islands. Recently, we
86 generated chromosome-level genome assemblies of all seven, non-reference founder
87 genomes[22]. The first release of the AMPRIL population (AMPRIL I) contained six
88 subpopulations (referred to as ABBA, ACCA, ADDA, BCCB, BDDB, CDDC) derived from
89 reciprocal diallel crosses between four hybrids (A: Col-0 x Kyo, B: Cvi-0 x Sha, C: Eri-1 x An-
90 1, D: L*er* x C24) and the subsequent selfing to the F4 generation by single-seed descent[21].
91 Here we present the extension of the AMPRIL population with six new subpopulations
92 referred to as EFFE, EGGE, EHHE, FGGF, FHHF and GHHG (E: Col-0 x Cvi-0, F: Sha x Kyo,
93 G: Ler x An-1, H: Eri-1 x C24) based on different diallel intercrossing scheme and selfing of
94 the recombinant genomes until the F6 generation (Fig. 1a). Each subpopulation consists of
95 approximately 90 individuals representing recombinants of four founders. In total, 992 RILs
96 from all twelve subpopulations were sequenced and analyzed using RAD-seq[23]
97 (Supplementary Data 1) and genotyped with ~2 million high-quality SNP markers
98 (Supplementary Data 2). We used a Hidden Markov Model to reconstruct the parental
99 haplotypes (identity-by-descent) including residual heterozygous regions[24] (Supplementary
100 Fig. 1 and Supplementary Note 1). The genotyping resulted in 12,878 different recombination
101 breakpoints (on average one breakpoint per 9.3 kb) across the entire population. This allowed
102 us to divide the genome of each progeny into 12,883 haplotype blocks (Supplementary Data
103 3), where each block relates to the haplotype(s) of only one (homozygous regions) or two
104 (heterozygous regions) of the founder haplotypes.

105 We developed a two-step workflow to combine genetic and genomic evidence to quickly
106 identify incompatible alleles of duplicated genes (Fig. 1b). In the first step, we selected 781
107 distal (inter-chromosome) duplicated gene pairs including 612 gene pairs in which the
108 reference sequence contains two copies and other founder genomes feature at least one
109 copy (Supplementary Data 4). In the remaining 169 gene pairs, the reference sequence only
110 has one copy and at least one other parental genome has an additional copy in a different

111 chromosome. As genetic incompatibility leads to the underrepresentation of incompatible
112 allele combinations[25,26], we searched for significant distortions from the expected frequencies
113 of all parental allele combinations across all 781 duplicated gene pairs in all twelve
114 subpopulations, two merged subpopulations (ABBA and EFFE, CDDC and GHHG as they
115 share the same founders), and the whole population (see Methods). These tests revealed
116 significant distortions in 236 gene pairs (χ^2 test, p-value < 0.05, multiple testing corrected) in
117 at least one of the populations.

118 However, the observed distortions do not necessarily result from genetic incompatibilities
119 in the tested gene. Alternatively, such distortions can also occur if the tested gene duplicate
120 is closely linked to a genetic incompatibility. Hence, in a second step, we examined the alleles
121 of the gene pairs in the founder genomes for loss-of-function (LoF) variations or
122 hypermethylated promoters (Fig. 1b, see Methods). This examination revealed three gene
123 pairs with functional disruption in both of the duplicates in at least one of the founder genomes.
124 These three duplicated genes included two, *HISTIDINOL PHOSPHATE*
125 *AMINOTRANSFERASE* (*HPA*)[11] and *FOLATE TRANSPORTER* (*FOLT*)[12], which were
126 already known for their ability to introduce genetic incompatibilities, as well as one gene pair,
127 which so-far was not reported as the genetic basis for a genetic incompatibility, *TIM22* (*TIM22-*
128 *1*: *AT1G18320*, *TIM22-2*: *AT3G10110*) (Fig. 1c). For all other 233 gene pairs we could not
129 identify non-functional alleles in both copies.

130 We noted that another duplicated gene, *tRNA ADENOSINE DEAMINASE 3* (*TAD3)*,
131 which is also know to introduce a genetic incompatibility[13], was not considered by our initial
132 testing even though the genotypes of the AMPRIL founders should lead to incompatible allele
133 combinations in the RIL populations: all founder genomes except for Kyo have a functional
134 *TAD3-1* (Supplementary Table 1), while the Kyo *TAD3-1* gene is silenced most likely due to
135 its methylated promoter (similar to the Nok-1 and Est-1 accessions in which the incompatibility
136 was described originally[13]). The lack of a functional *TAD3-1* in Kyo is counterbalanced by
137 additional copies (*TAD3-2*) (which however were not part of the main chromosome scaffolds
138 of the Kyo genome assembly). Therefore, the *TAD3* gene duplicate was not considered
139 initially, however, we also used this incompatibility for further analysis.

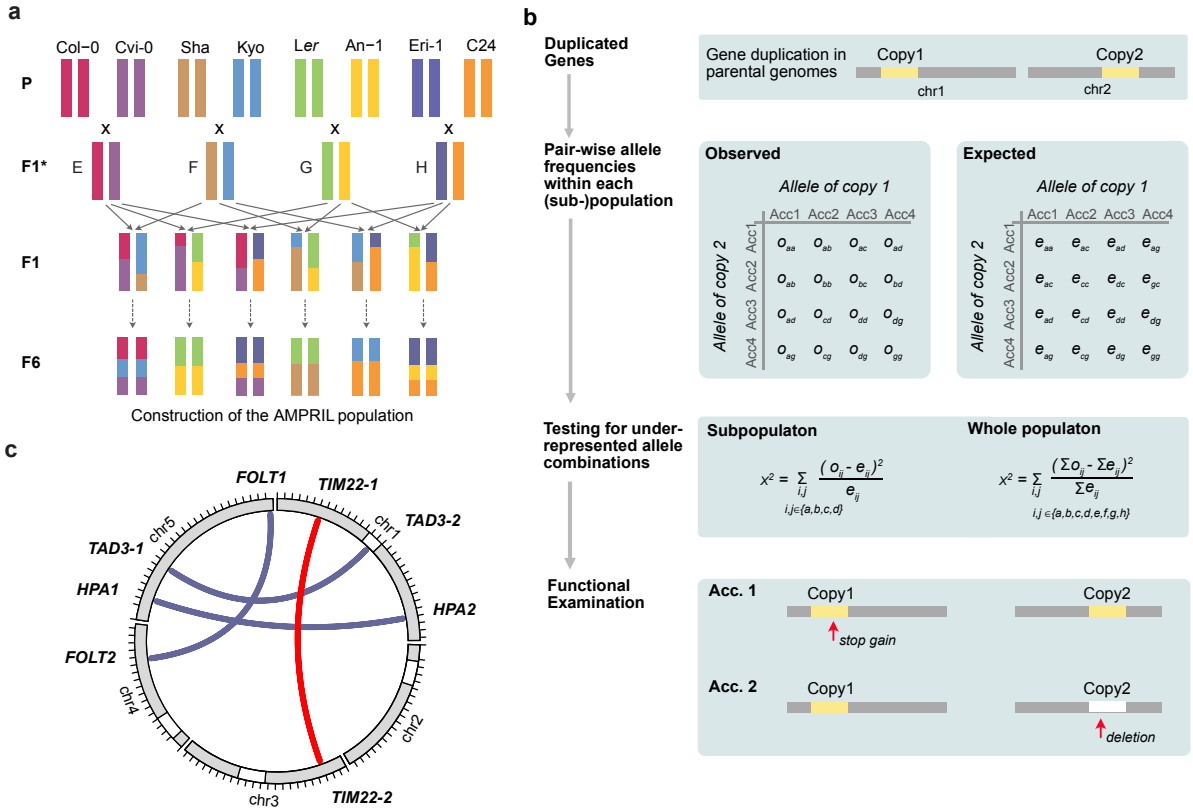


Figure 1. Identification of genetic incompatibilities introduced by duplicated genes in an intercross population (**a**) Construction of the extended AMPRIL population. Eight different *A. thaliana* accessions were used as founder lines. For each of the six subpopulations, two F1* hybrids, which were 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[27]. 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 complete AMPRIL population

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

164 We observed an in-frame premature stop-codon in Col-0 (mis-annotated in the reference 165 annotation) suggesting that *TIM22-1* is not functional in Col-0 (Fig. 2a and Supplementary 166 Table 4). To test if *TIM22-1* is truly non-functional in Col-0, we used the segregation of two T-167 DNA insertion mutants in the two *TIM22* paralogs in Col-0. This showed that *tim22-1* could 168 be homozygous for the T-DNA insertion allele but the T-DNA in *tim22-2* could not be found in 169 homozygous state (Fig. 2c). This suggests that, in Col-0, *TIM22-1* is not functional and *TIM22-*170 *2* is the only functional copy.

171 The *TIM22* paralogs co-located within the regions of a previously reported genetic 172 incompatibility (hereafter named as LD2: LD2.1 for the locus at chromosome 1 and LD2.3 for 173 the locus at chromosome 3) which was mapped in a Cvi-0 x Col-0 RIL population[28]. The 174 genetic underpinnings of this incompatibility however were still unknown. This incompatibility 175 was expressed by a striking underrepresentation of homozygous LD2.1^{Col-0} combined with 176 homozygous LD2.3^{Cvi-0}, which was in agreement with the reduced allele combinations in the 177 AMPRIL population (Supplementary Table 2 and 3). Therefore, we generated heterogeneous 178 inbred family (HIF) lines from Cvi-0 x Col-0 RILs to fine-map LD2.1 and LD2.3 to respectively 179 70kb and 34kb intervals (Supplementary Fig 2). The two candidates *TIM22-1* and *TIM22-2* 180 remained within the intervals.

181 To validate their causative role, we conducted a complementation cross between a 182 heterozygous T-DNA mutant in *TIM22-2* in a Col-0 background (i.e. *TIM22-1* was non-183 functional) and the original HIF line in which *TIM22-1* was homozygous for the Col-0 genotype 184 (i.e. also non-functional) and *TIM22-2* was heterozygous for Col-0/Cvi-0 (Fig. 2d). Within 123 185 hybrids of the offspring, among the four possible allelic combinations at LD2.3, we did not find 186 any hybrids combining a Cvi-0 and a T-DNA alleles at *TIM22-2* (Fig. 2d), providing strong 187 genetic evidence that the Cvi-0 allele cannot complement a knockout (T-DNA) allele at *TIM22-*188 *2* (in a background without other functional *TIM22* allele) and is thus non-functional.

189 Collectively, these segregation and complementation crosses show that the combination 190 of different non-functional alleles of the *TIM22* copies leads to a drastic reduction of allelic 191 combinations in offspring populations, and thus evidence the causative role of *TIM22* in this 192 genetic incompatibility.

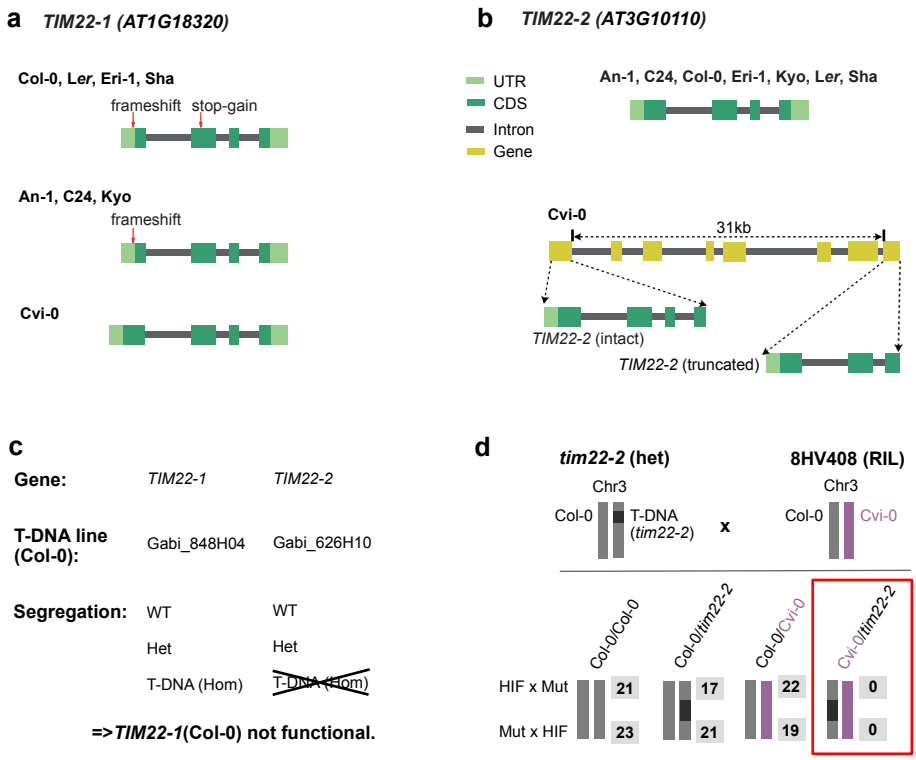
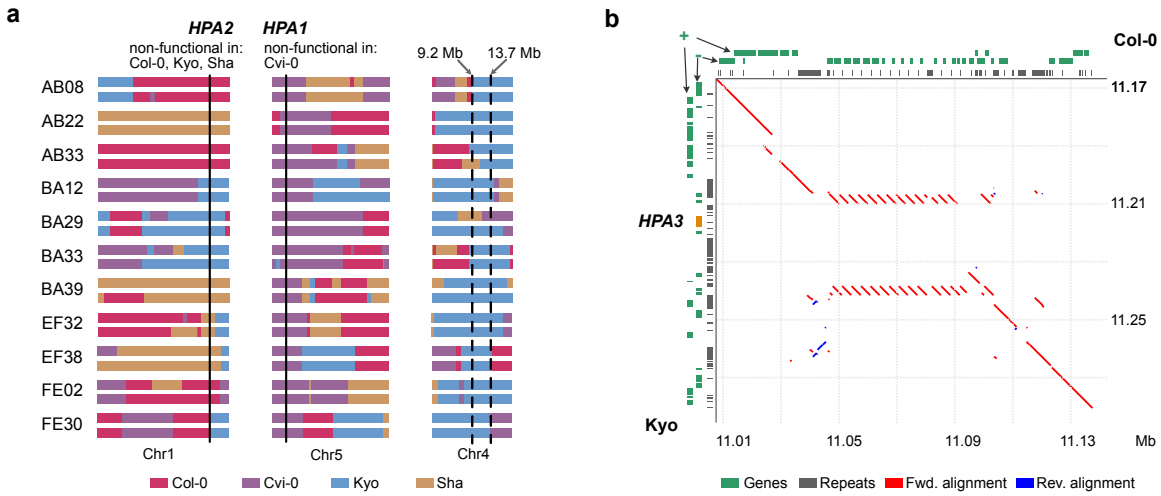


Figure 2. Genomic and genetic evidence of incompatible *TIM22* alleles. **(a)** Gene structure of *TIM22-1* in the genomes of the eight AMPRIL parents. The loss-of-function variants (1bp frameshift indel and a premature stop-codon) relative to the intact *TIM22-1*$^{Cvi-0}$ are shown. The deleterious 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[11] 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[29]. 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 hypermethylated promoter, while Cvi-0

217 carries a functional *HPA2* allele, but does not carry *HPA1* at all (Supplementary Table 5).
218 Unexpectedly, however, we did observe homozygous *HPA1/HPA2* incompatible allele
219 combinations ($HPA2^{-/-}HPA1^{-/-}$) in eleven of the AMPRIL lines within the ABBA and EFFE
220 subpopulations which were derived from Col-0, Cvi-0, Kyo and Sha (Supplementary Table 6-
221 8). Further analysis of these populations revealed an extremely high frequency of the Kyo
222 allele on chromosome 4 (Supplementary Fig. 3), making us recognize that the eleven AMPRIL
223 lines with the incompatible allele combinations all carried at least one Kyo allele at chr4:9.2-
224 13.7 Mb (Fig. 3a). This suggested that Kyo might contain a modifying allele of the
225 incompatibility complementing the lethal allele combination in this region (a similar feature
226 was described as "conditional incompatibility" by Bikard et al. when analyzing crosses
227 between Jea and Col-0[11]). Indeed, when we checked the long-read assembly of the Kyo
228 genome, we identified an additional *HPA* copy (hereafter named as *HPA3*) which co-located
229 with the mapping interval at chr4:11.11 Mb (Fig. 3b), while we could not find this allele in any
230 of the other AMPRIL founder genomes.
231



232
233 **Figure 3. Incompatible allele combinations of *HPA* rescued by an additional gene copy. (a)**
234 Schematic of the genomes of 11 AMPRILs with presumably incompatible allele combinations of *HPA1*
235 and *HPA2*. All these genotypes carry at least one Kyo allele at chr4: 9.2 – 13.7 Mb, suggesting that a
236 Kyo allele in this region can rescue the incompatibility. **(b)** Sequence alignment around the *HPA3* locus
237 on chromosome 4 between Col-0 and Kyo. The position of the third *HPA3* copy in Kyo is marked in
238 orange. Red line: forward alignment; blue line: reverse alignment. Genes arrangement at forward (+)
239 and reverse (-) strands, and repeat annotations are shown at the top (Col-0) or left (Kyo) axes.
240
241 **Mapping modifiers of incompatible allele combinations in natural populations**
242 This analysis showed that the negative effects of incompatible allele combinations can
243 be overcome if they are rescued by modifying alleles. Therefore, the virtual absence of

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

247 To do this, we searched for incompatible allele combinations (separately for all four
248 incompatibilities) in each of the 1,135 accessions of the 1001 Genomes Project[19] and used
249 these allele combinations as phenotype for a GWA (Fig. 4). To define non-functional alleles,
250 we used the resequencing data to search for LoF variations, and the methylome data from
251 the 1001 Epigenomes Project[30] to identify methylated (silenced) promoters (Fig. 4a, see
252 Methods). Additionally, RNA-seq data were explored to distinguish pseudo-heterozygous
253 variants (which exist due to inaccurate short-read alignment at duplicated genes) and to check
254 for gene silencing.

255 For the first incompatibility, in *HPA*, we found pseudo-heterozygous LoF variations
256 including two premature stop-codons and three frameshifts at both reference copies of *HPA*
257 (Fig. 4c and Supplementary Table 9) due to the repetitiveness of the gene sequences. As
258 some accessions did not show expression of *HPA2* most likely due to hypermethylation of
259 their promoters, we tested which of the *HPA* alleles was present in the RNA-seq read data to
260 assign the LoF to either of the *HPA* copies (Fig. 4b). With this, we could assign one stop-
261 codon gain (LoF-300) and two frameshifts (LoF-170, LoF-230) to *HPA2* because these LoF
262 alleles were absent in the RNA-seq data in the accessions which lacked *HPA2* expression.
263 Notably, the frameshift LoF-170 could be rescued by alternative splicing as observed in RNA-
264 seq read mapping (Supplementary Fig. 4). Furthermore, cytosine methylation profiles
265 revealed hypermethylated promoters of *HPA2* in 340 accessions and of *HPA1* in 50
266 accessions (Fig 4d, e), suggesting gene silencing in these accessions, which was in
267 agreement with the absence of pseudo-heterozygous variants in the RNA-seq read data (Fig.
268 4b and Supplementary Fig. 4).

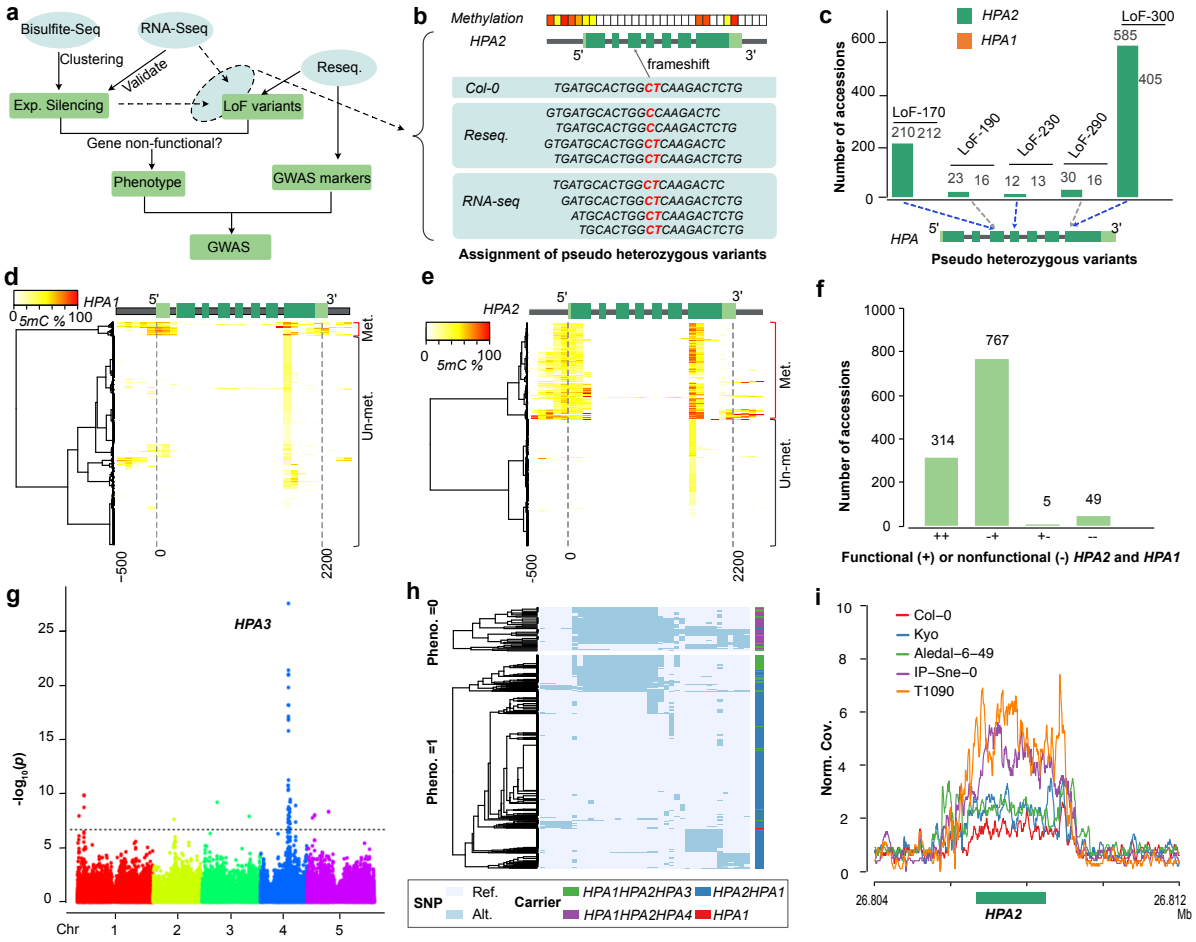


Figure 4: Mapping natural modifiers of genetic incompatibilities using GWAS. (a) Workflow for the identification of modifying alleles of genetic incompatibility using GWAS. **(b)** Schematic example to illustrate how genome-wide methylation and RNA-seq data are used to assign pseudo-heterozygous variants 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[19]) 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[20] (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 *HPA2* 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 significantly associated markers around the *HPA3* locus)

289 shown for accessions with (below) or without (up) functional copies of *HPA2* or *HPA1*. **(i)** The
290 normalized short read mapping coverage (Norm. Cov.: average mapping coverage at *HPA2* divided by
291 average mapping coverage at whole genome) around *HPA2* based on read mappings against the Cvi-
292 0 genome including only one *HPA* gene. (Data from five accessions are shown as examples to illustrate
293 patterns of different *HPA* copies: Col-0: *HPA2HPA1*; Kyo and Aledal-6-49: *HPA2HPA1HPA3*; IP-Sne-
294 0 and T1090: *HPA2HPA1HPA4*).
295
296 By combining the LoF variant genotyping and the methylation analyses, we found 767
297 accessions with a non-functional *HPA2* ($HPA2^{-/-}$ $HPA1^{+/+}$) allele and five accessions with a
298 non-functional *HPA1* ($HPA2^{+/+}HPA1^{-/-}$) allele (Fig. 4f and Supplementary Table 10). Also, 49
299 accessions did not feature any functional alleles ($HPA2^{-/-}HPA1^{-/-}$) and were expected to carry
300 additional modifier(s) to complement the loss of functional copies. To find the locations of
301 these modifiers, we used the absence of functional copies as phenotype (Supplementary
302 Data 5) to run a GWA under the mixed linear model using the SNP markers from the 1001
303 Genomes Project (www.1001genomes.org). This GWAS revealed a significantly associated
304 region at chr4:11.00-11.15 Mb (Fig. 4g and Supplementary Fig. 5, alpha level of 0.05,
305 Bonferroni correction), corresponding to the *HPA3* locus found in Kyo (chr4:11.11 Mb).
306 Though other peaks in unlinked regions were present, these additional loci explained only a
307 small proportion of the heritability.
308 Analyzing the haplotypes at the 39 significantly associated SNP markers at *HPA3* locus
309 revealed a somewhat, but not entirely homogenous haplotype in many of the $HPA2^{-/-}$ $HPA1^{-/-}$
310 carriers (Fig. 4h). To confirm that the modifying haplotype is still identical to the Kyo allele, we
311 aligned the short reads of all accessions of the 1001 Genomes Project against the Kyo
312 reference sequence and found that overall 162 accessions carried the *HPA3* allele
313 (Supplementary Data 5). However, unexpectedly the Kyo *HPA3* was only found in 13 of the
314 49 accessions without functional *HPA2* and *HPA1* alleles, suggesting the presence of two
315 different modifiers at this locus on chromosome 4. To find more support for this, we ran a new
316 GWA without the 162 *HPA3* (Kyo-like allele) carriers, which still led to a significantly
317 associated locus at the region of *HPA3* (Supplementary Fig. 6). Further short-read mappings
318 of all 1,135 genomes against the Cvi-0 reference sequence (where only one copy of *HPA*
319 exists) revealed the presence of (at least) two additional copies (in addition to *HPA1* and
320 *HPA2*) in a total of 42 accessions including all the 36 accessions with the unknown rescuing
321 alleles (Fig. 4i and Supplementary Data 5). Together this suggested that *HPA3* also rescues
322 incompatible *HPA1* and *HPA2* allele combinations in natural populations and that the locus
323 of *HPA3* contains an additional haplotype (hereafter named *HPA4*) which also rescues the
324 incompatibility between *HPA1* and *HPA2*.
325 We continued to apply the same approach to the other three incompatibilities. For *TIM22*,

326 16 accessions of the 1001 Genome Project revealed non-functional allele combinations
327 (Supplementary Table 11 and Supplementary Data 6), again indicating the existence of
328 modifying alleles. However, a GWAS using the presumably-incompatible allele combinations
329 as molecular phenotype did not reveal only one, but numerous significantly associated loci
330 (Supplementary Fig. 7). This might be explained by the low number of incompatible allele
331 carriers, which could affect the power of association mapping leading to false-positive
332 associations. However, even though we could not locate the modifying allele, further analysis
333 of read mapping coverage in *TIM22* revealed that all 16 accessions with non-functional
334 *TIM22-1/TIM22-2* allele combinations carried at least one additional third copy of *TIM22*, while
335 among all 1,119 other genomes of the 1001 Genomes Project only three genomes carried
336 additional copies. This suggests that, like for *HPA*, known non-functional allele combinations
337 of *TIM22* are in fact rescued by additional copies. Moreover, in ten of the 16 accessions,
338 *TIM22-2* showed not only one but multiple additional copies (hereafter named as *TIM22-3*),
339 which was further supported by the genome assembly of Ty-1
340 (https://genomevolution.org/CoGe/GenomeInfo.pl?gid=54584, unpublished), where we could
341 find a cluster of four tandemly arranged *TIM22* gene copies at the *TIM22-2* locus.
342 Because the reference sequence only contained one copy of *TAD3 (TAD3-1)*[13] and *FOLT*
343 *(FOLT1)*[12], we modified our GWA method and only used non-functional alleles at the
344 reference gene as the phenotype to map modifiers for the two remaining incompatibilities (Fig.
345 5 and Supplementary Fig. 8). Due to this modification we would expect to map also the
346 location of the duplicated genes as we had found them in the AMPRIL founders. For *TAD3*, an
347 essential ortholog of the yeast tRNA Adenosine deaminase 3[31], we did not find any
348 accessions with LoF alleles, but we found 150 accessions with a hypermethylated promoter
349 in *TAD3-1* similar to the methylated promoters found in Nok-1 and Est-1, which are known to
350 be non-functional due to this methylated promotor region[13] (Fig. 5a, b). The GWA result
351 revealed one significant peak (Fig. 5c,d), which as expected co-located with the *TAD3-2*
352 locus[13]. All the 150 accessions carried multiple additional copies of *TAD3-2* (Supplementary
353 Data 7), similar to Nok-1 and Est-1. This suggests that the expression silencing of *TAD3-1* is
354 common in natural populations and that the rescue of this loss-of-function allele is generally
355 mediated by additional gene copies at the *TAD3-2* locus as it was shown in the original
356 description of this incompatibility[13].

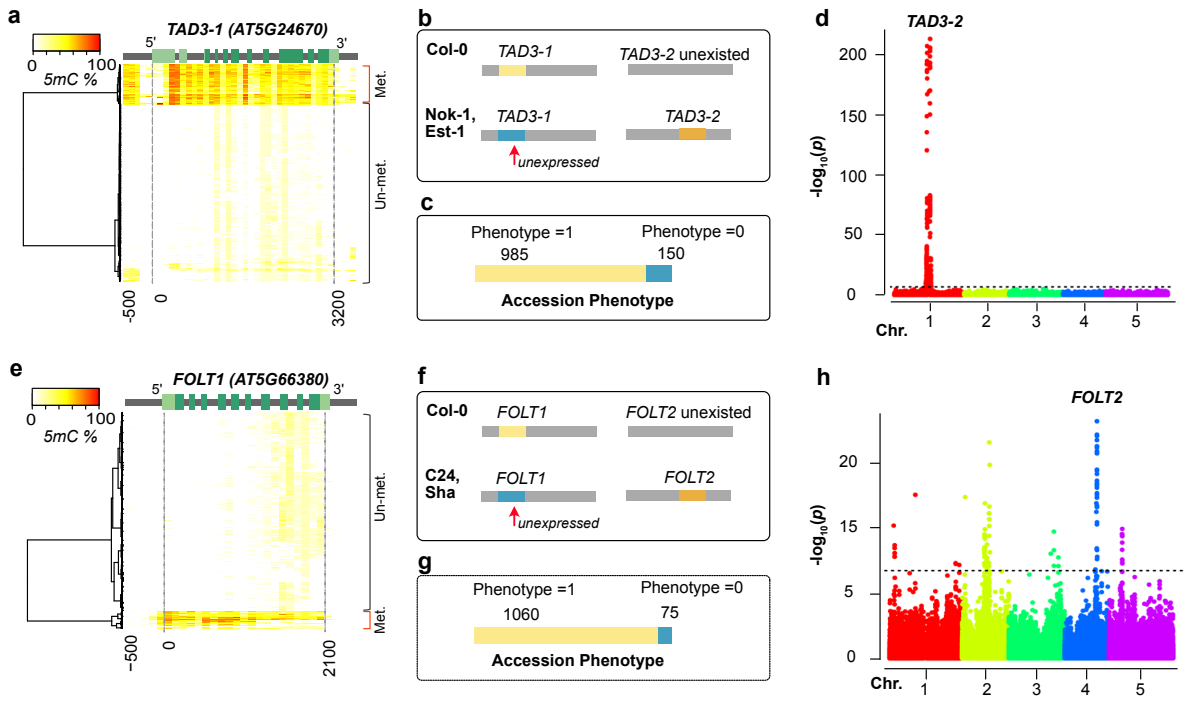
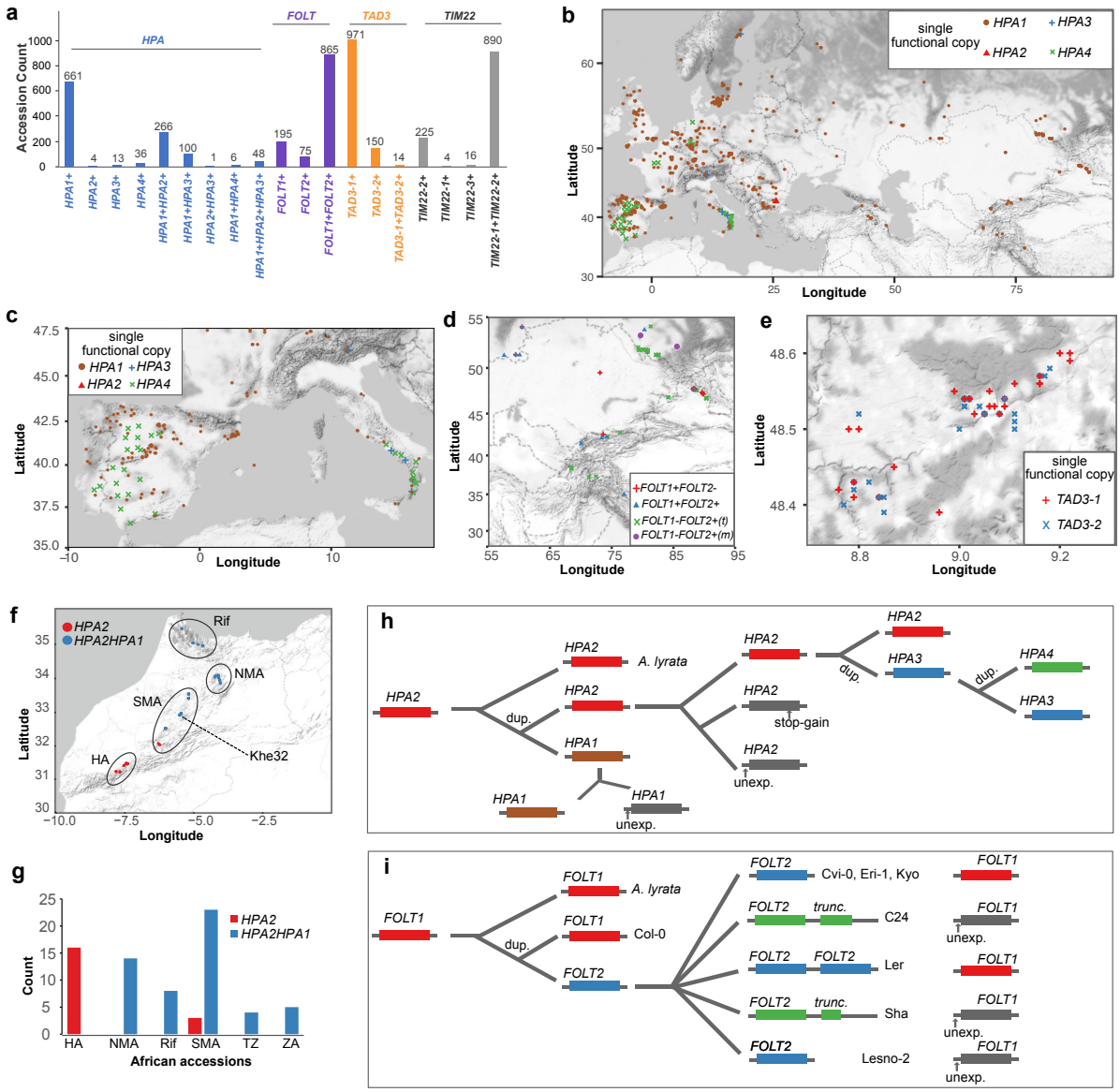


Figure 5: Mapping non-reference gene copies of incompatible alleles using GWAS. (a, e) Hierarchical clustering of cytosine methylation profiles in *TAD3-1* (a) and *FOLT1* (e) based on 888 *A. thaliana* accessions from 1001 Epigenomes Project[20]. The methylation profile was calculated based on 100 bp 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[12,13]. **(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[12,28,32]. 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[12], 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 methylated promoters of *FOLT1* likely leading to expression silencing (Fig. 5e), which was supported by the lack of *FOLT1*-specific pseudo-heterozygous

381 SNPs in the RNA-seq data (Supplementary Fig. 9). Besides evidencing the expression
382 silencing of *FOLT1*, this also suggested the existence of additional *FOLT* gene copies in these
383 75 accessions. When we repeated our GWA approach to find these interacting loci of *FOLT1*,
384 we found multiple significantly associated loci including one region corresponding to *FOLT2*
385 (Fig. 5g, h).
386 Further analyses of the *FOLT* gene copies using short read mapping against the C24 and
387 Sha reference sequences revealed the presence of *FOLT2* in all the 75 accessions with
388 methylated promoter of *FOLT1* and truncated copies of *FOLT2* genes in only 46 out of such
389 75 accessions (Fig. 5e, Supplementary Table 13 and Supplementary Data 8) suggesting that
390 the methylation of the *FOLT1* promoter remains stable even after the truncated copies are
391 segregated out. This is coherent to what was shown in successive generations of Sha x Col-
392 0 RILs where the inducing locus was segregated away six generations ago[12].
393 Taken together we found evidence that all four incompatible allele combinations, initially
394 identified within artificial intercross population, also occur in nature, and that the incompatible
395 allele combinations of three of them were surprisingly common. For all incompatible allele
396 combination carriers, we found evidence for the existence of additional copies, and could
397 even map the locations of some of those using GWA.
398
399 **Geographic distribution of incompatible allele combinations**
400 We next asked how prevalent the potential for incompatible allele combinations was
401 within the natural population of *A. thaliana* by analyzing the presence of different haplotypes
402 in different geographic regions. For this, we first analyzed the allele frequencies of different
403 haplotypes (i.e. the different combinations of functional alleles within individual plants), which
404 varied substantially across the accessions of the 1001 Genomes Project (Fig. 6).
405 For example, among the accessions that carried only one functional copy of *HPA*, we
406 found that 661, 4, 13 and 36 accessions with a single functional copy of *HPA1*, *HPA2*, *HPA3*,
407 or *HPA4*, respectively (Fig. 6a, b, Supplementary Fig. 10). Similarly, most accessions only
408 featured one functional copy of *TAD3* including 971 accessions with only *TAD3-1* and 150
409 with only *TAD3-2* (Fig. 6a, Supplementary Table 14 and Supplementary Fig. 11). In contrast,
410 most of the accessions included multiple functional alleles of *FOLT* as well as of *TIM22* (Fig.
411 6a).
412 Genetic incompatibilities become effective in the offspring of accessions with
413 incompatible alleles. In particular the offspring of accessions with only one functional gene
414 copy will lead to the highest reduction in fitness. We found numerous accessions with the
415 potential to create incompatible allele combinations in their offspring in geographically close
416 regions (Fig. 6c, d, e). For example, among 30 accessions in the South of Italy we found 9
417 accessions with only a functional *HPA1* closely located with 14 accessions with either a

418 functional *HPA3* or a functional *HPA4* only (Fig. 6c). Likewise, accessions with contrasting
419 functional copies of *FOLT* and *TAD3* were collected from the same regions in Central Asia
420 and Germany (Fig. 6d and e) evidencing that incompatible alleles could segregate within the
421 same populations.
422



423

424 **Figure 6. Distribution, origin and evolution of incompatible alleles. (a)** The number of accessions
425 with different functional copies of *HPA*, *FOLT*, *TAD3* and *TIM22* across 1,135 *A. thaliana* accessions
426 from the 1001 Genomes Project. +: functional. **(b)** Geographic distribution of accessions only with one
427 functional copy of *HPA*. **(c, d, e)** Examples of geographically close accessions with only one functional
428 copy of *HPA* in Southern Europe (c), *FOLT* in Central Asia (d) and *TAD3* in Germany (e). *FOLT1-*
429 *FOLT2*+(t): accessions with a functional *FOLT2*, a truncated *FOLT2* and an unexpressed *FOLT1*.
430 *FOLT1-FOLT2*+(m): accessions with a functional *FOLT2* and an unexpressed *FOLT1*, but without
431 truncated *FOLT2* copies. **(f, g)** Gene copies of *HPA* in African *A. thaliana* accessions. TZ: Tanzania,
432 ZA: South Africa. **(h, i)** Schematic of a possible and parsimonious evolutionary history of gene
433 duplication and non-functionalization of *HPA* (h) and *FOLT* (i). unexp.: unexpressed. stop-gain:
434 premature stop codon gained SNP. dup.: gene duplication.
435
436 Interestingly, within some of these hybrid zones of accessions with incompatible alleles,
437 we also found accessions with multiple functional gene copies. For example, in South Italy
438 there were 7 of the 30 accessions that featured multiple functional copies of *HPA,* and
439 similarly in Central Asia, we could find multiple accessions with two functional copies of *FOLT*
440 in addition to the accessions with only one functional copy (Fig. 6d and Supplementary Fig.
441 10-12). In contrast, however, we only observed a few accessions with multiple functional
442 alleles in the hybrid zone of incompatible *TAD3* alleles, most likely because only 14 such
443 accessions were found in the entire set of accessions. Even though we could not find
444 evidence that such hybrid zones are enriched for additional copies or that those additional
445 copies would evolve in these regions, the presence of haplotypes with additional gene copies
446 in these hybrid zones have the potential to mediate the gene flow between the haplotypes
447 with the incompatible allele combinations.
448
449 **Origin and evolution of incompatible alleles**
450 To figure out when these incompatible alleles originated and how they evolved, we
451 investigated their ancestral genotypes within African genotypes which likely represent the
452 ancestral populations of the Eurasian accessions[33].
453 While almost all (99.7%) of the Eurasian accessions had both *HPA2* and *HPA1* (either
454 functional or not), we found that only 27% (20 out of 75) of the African accessions carried only
455 *HPA2* (Fig. 6f, g and Supplementary Data 6). This suggested that *HPA2* was the ancestral
456 copy of *HPA* which was further supported by the synteny alignment with the close relative
457 *Arabidopsis lyrata*[18], which only featured a single *HPA* gene in a region syntenic to *HPA2* and
458 that the duplication events leading to *HPA1* happened early on in Africa (Fig. 6h). One
459 accession, Khe32, from Morocco carried *HPA1* along with the most frequent LoF variant LoF-
460 300 in *HPA2* in the Eurasian accessions, suggesting that the first accession with only a
461 functional *HPA1* possibly arose in North-West Africa and thereby suggests that the genetic

incompatibility with *HPA2* carriers was already possible before the Eurasia colonization of *A. thaliana*.

In contrast, *HPA3* and *HPA4* only occur in Eurasian accessions, most likely indicating that the additional duplication events happened later (Fig. 6h). Comparing the gene sequences of the *HPA* genes revealed that *HPA3* was duplicated from *HPA2* (Supplementary Fig. 13), while *HPA4* might be a tandem duplicate of *HPA3* as it is likely located close to *HPA3*. Interestingly, *HPA3* and *HPA4* carriers segregated for the ancestral LoF-300 variant at *HPA2* (*HPA3*: 68 with and 94 without; *HPA4*: 10 with and 32 without, Supplementary Fig. 14), suggesting free segregation of different *HPA* alleles. In striking contrast to this, the inactive alleles of *HPA1* (i.e. alleles with a non-functional *HPA1* sequence, but excluding the accessions with full deletion alleles as found in Cvi-0) were almost perfectly coupled (48 of 50) with inactive alleles of *HPA2*. Such carriers nearly all had *HPA4* (35) or *HPA3* (13) (Supplementary Fig. 15) and were mainly located in the Iberian Peninsula and Southern Italy suggesting that the additional *HPA* copies were necessary to buffer the incompatibility and to allow the foundation of these populations (Fig. 6c).

Unlike the incompatibility in *HPA*, genetic incompatibilities can also arise in recent population history. All African accessions only had the *TAD3-1* copy, while accessions with multiple copies of *TAD3* can only be observed in Eurasia (Supplementary Fig. 10). Similarly, 38 of the African accessions only have *FOLT1*, while the other 37 accessions have both *FOLT1* and *FOLT2*, but none of the accessions featured a truncated copy of *FOLT2* (*FOLT2tr*), which is the mechanistic origin of the incompatible alleles at *FOLT1* and *FOLT2* (Supplementary Fig. 16 and Supplementary Data 8). In contrast, within the Eurasian accession, we even found three different haplotypes of *FOLT2tr* within a total of 46 accessions based on short read alignments against the eight *A. thaliana* genomes[22] (Supplementary Fig. 17 and Supplementary Data 8). This together suggests that also the genetic incompatibility that is based on *FOLT2tr* has evolved recently after the migration *A. thaliana* to Eurasia (Fig. 6i).

Discussion

Although gene duplication provides genetic backup of essential genes, duplicated genes can also lead to incompatible allele combinations when the duplicated genes undergo reciprocal pseudo-functionalization in separate genomes. Here we studied incompatible allele combinations of four duplicated gene pairs by integrating genetic and genomic information using a multi-parental intercross population leading to the identification of four genetic incompatibilities.

Unexpectedly in this population we identified some lines with what initially looked like an

498 incompatible allele combination of one of the incompatibilities, as they could be rescued by
499 an additional, so-far unknown gene copy. Encouraged by this, we developed a GWA method
500 to map modifying alleles also in natural populations by integrating genomic and epigenomic
501 data to generate a molecular phenotype that describes potentially incompatible allele
502 combinations. With this we identified many natural accessions with putatively incompatible
503 allele combinations and could elucidate the genetics of these incompatibilities as they occur
504 in natural population using the 1,135 accessions released by the 1001 Genomes Project[19].
505 This implies that besides presence/absence analysis as performed by standard pan-genome
506 analysis, gene location is an essential feature of a gene (or gene family).

507 Based on multi-omics data from hundreds of *A. thaliana* accessions, we could
508 comprehensively describe the origin and evolution of several incompatible allele combinations
509 of the four incompatibilities. We found that incompatible alleles are surprisingly frequent in
510 nature and also occur in sympatry, suggesting that the evolution of genetic incompatibilities
511 does not require separated populations as proposed by BDM model[2–4]. Moreover, such
512 incompatible allele combinations can persist over long periods as some of the alleles studied
513 here may have originated before *A. thaliana* colonized Eurasia.

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

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

528

529 # Methods

530 **AMPRIL construction**

531 The eight *A. thaliana* accessions An-1, C24, Col-0, Cvi-0, Eri-1, Kyo, L*er*, and Sha were
532 selected as the AMPRIL founders. We previously constructed a first version of AMPRIL
533 population (AMPRIL I) including six RIL subpopulations[21]. Here AMPRIL I was extended with

534 six additional subpopulations (called AMPRIL II) based on different pairwise intercrosses of
535 the eight founders (Fig. 1a). Each subpopulation contains approximately 90 individuals
536 (Supplementary Data 1). All plants were grown under the normal growth conditions in
537 greenhouse at the MPI-PZ[21]. DNA of 1,100 samples was extracted from the flower buds and
538 prepared for RAD-seq sequencing[23].
539
540 **RAD-seq Library preparation and sequencing**
541 All plants were grown in the greenhouse. The total DNA from each of the 1,100 samples
542 was extracted from the flower buds using DNeasy Plant Kit 96 (Qiagen) and eluted in 200 µl
543 Elution buffer (EB). DNA of each genotype was isolated twice. Both genotype samples were
544 pooled in 1.5 ml tube and the DNA was concentrated by isopropanol precipitation for 2h at -
545 20°C. The samples were centrifuged at 12000 *g* for 5 min at 4°C, the supernatant was
546 removed and the DNA pellet was washed with ice-cold 70% ethanol. Centrifugation was
547 repeated, the supernatant was removed. Air-dried DNA was resuspended in a nuclease free
548 water to 26 ng/µl and stored at -20°C until use. RAD-seq sequencing libraries were prepared
549 as described[34] with modifications. Per genotype, 500 ng DNA were digested with 10 units of
550 CviQI (NEB, cutting site G'TAC) at 25°C for 2 h. The number of expected cutting sites was
551 estimated around 236,000 based on the Col-0 genome sequence. Cut DNA was purified using
552 96 DNA clean and concentrator kit (Zymo) a diluted in 25 µl EB. The 192 different (selected
553 out of total 210 designed, Supplementary Data 1) P1 adapters (200 nM) containing unique 12
554 bp barcodes were ligated by incubation with T4 ligase (NEB) at room temperature for 30 min
555 and the reaction was terminated by 20 min at 65°C. After 30 min at room temperature, 5 µl
556 from each 192 P1-barcoded sample were combined in a 2 ml low bind tube. 3 x 130 µl aliquots
557 were transferred to fresh tubes and DNA was fragmented to average size of 500 bp using
558 Covaris. Sheared DNA was purified using QiaMinElute columns (Qiagen), eluted in 10 µl EB,
559 and the three samples were first pooled and then divided into two 15 µl samples that were
560 run on 1% agarose gel. Regions of 300-500 bp fragments were dissected and DNA was
561 isolated using MinElute gel extraction kit (Quiagen), eluted in 10 µl EB, the samples were
562 pooled, DNA fragment ends were repaired using Quick Blunting™ kit (NEB), purified with
563 QIAquick column (Qiagen) and eluted in 43 µl EB. 3' deoxy-adenine overhangs were added
564 using Klenow Fragment (NEB), the sample was purified with QIAquick column, eluted in 45
565 µl EB, the P2 adapter was ligated, the sample was purified with QIAquick column and eluted
566 in 53 µl EB. To determine the library quality, 10 µl RAD library were PCR amplified (1: 98°C
567 30 sec; 2: 14x 98°C 10 sec, 67°C 30 sec, 68°C 30 sec; 3: 68°C 5min, 4: 4°C hold) using NEB
568 Next High-Fidelity master mix (NEB) in 25 µl reaction volume using RAD-Marker for/ RAD-
569 Marker rev primers (25 nM each) and 5 µl PCR product were loaded into 1% agarose gel next
570 to the 1 µl RAD library template. If the PCR product smear was at least twice as intense as

571 the template smear, the library was considered as of high quality and the amplification was
572 repeated in 50 µl reaction volume. The product was cleaned using AMPure magnetic beads
573 (Beckman Coulter) and dissolved in 20 µl EB. Finally, the sample was run on 1% agarose gel,
574 the region of 300-500 bp fragments was cut out, DNA was isolated using MinElute Gel
575 extraction kit (Qiagen) and eluted in 20 µl EB. The library was sequenced in an Illumina
576 Hiseq2000 sequencing machine. Sequencing reads were demultiplexed according to the
577 barcodes (Supplementary Data 1).
578
579 **AMPRIL genotyping**
580 We used previously released whole-genome short read data[22] to generate markers for
581 genotyping the AMPRILs. We mapped the reads to the reference sequence and called SNPs
582 using SHORE (version 0.9)[35] with default parameter settings. Only homozygous SNP calls
583 were selected after removing SNP calls with low quality (quality < 30), in the repetitive regions
584 or in regions with low mapping quality (quality < 30). The actual marker sets for each of the
585 twelve subpopulations (ABBA, ACCA … GHHG) were selected based on respective parental
586 genomes (Supplementary Data 2), exclusively selecting bi-allelic SNP markers. After
587 excluding samples with replicates or too few sequencing reads, we performed genotyping on
588 992 AMPRILs using a Hidden Markov Model-based approach similar to the recently
589 presented method for the reconstruction of genotypes derived from two parental genomes[24]
590 (for a detailed description see Supplementary Note 1 and Supplementary Data 3).
591
592 **Identifying genetic incompatibilities based on duplicated genes**
593 Duplicated gene pairs were selected based on gene family clustering of protein-coding
594 genes from all eight parental genomes using OrthoFinder (version 2.2.6)[36]. We only selected
595 inter-chromosome duplicated genes to avoid the effects of intra-chromosome linkage
596 (Supplementary Data 4). For each duplicated gene pair, we required two copies in the
597 reference sequence and at least one copy in one of the other genomes (DupGene2), or one
598 copy in reference sequence and at least one copy on a different chromosome in at least one
599 of the other parental genomes (DupGene1). We assessed the genotypes of each of the gene
600 copies in each AMPRIL using the genotypes predicted in the middle of the respective
601 reference gene, or in the case a gene was not present in the reference sequence – using the
602 midpoint between the two closest flanking syntenic regions of the non-reference gene copy
603 (based on synteny calculations from a previous study[22]).
604 We predicted candidate genetic incompatibilities using a two-steps approach. In the first
605 step, we performed chi-square tests (Equation 1) to check whether the frequency of allele
606 pairs in duplicated genes was significantly distorted in any of the subpopulations or in any of
607 two merged subpopulations (ABBA and EFFE or CDDC and GHHG which shared the same

608 four founders, respectively).

609 $$\chi^2 = \sum_{i,j \in \{a,b,c,d\}} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \quad (1)$$

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

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

615 $$x^2 = \sum_{i,j \in \{a,b,c,d,e,f,g,h\}} \frac{(\sum o_{ij} - \sum e_{ij})^2}{\sum e_{ij}} \quad (2)$$

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

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

629

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

631 We mapped all whole-genome resequencing reads of 1,135 accessions from the 1001
632 Genomes Project[19], 75 accessions from Africa[33] and 118 accessions from China[37], to the Col-
633 0 reference genome (TAIR10)[38,39] using BWA (version 0.7.15)[40] with the default parameter
634 settings. SNPs and small indels were called using SAMtools (version 1.9) using default
635 parameters[41]. Homozygous variants with mapping quality of more than 20 and with at least
636 four reads aligned were kept. Pseudo-heterozygous variants in *HPA* and *TIM22* were also
637 recorded. The large insertion and deletions in the 40kb extended genic region of focal genes
638 were predicted using Pindel (version 0.2.5)[42] with parameter settings "-T 1 -x 5 -k -r -j" and
639 Delly (version 0.8.1)[43] with parameter settings "delly call -q 20 -r 20 -n -u 20 -g". The functional
640 effects of these variations were annotated using SnpEff (version 4.3p)[44] using the default
641 parameter settings. The loss-of-function (LoF) effects include loss of start codon, loss of stop
642 codon, gain of premature stop codon, damage of splicing acceptor or donor sites, frameshift

643 and CDS loss.

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

645 Cytosine methylation data (the tab separated file of methylated cytosine positions) of
646 1,211 samples from the 1001 Epigenomes project[20] were downloaded from NCBI (927
647 samples under GEO accession GSE43857 and 284 samples under GEO accession
648 GSE54292). After removing the redundant data sets 888 and 161 data sets from GSE43857
649 and GSE54292, respectively, were retained. For each sample, we calculated the percentage
650 of methylated cytosines in CG, CHG and CHH contexts from 500bp upstream of the
651 transcription start sites to 300bp downstream of the transcription termination sites of each
652 candidate gene in 100bp non-overlapping sliding windows. These methylation profiles were
653 hierarchically clustered using the hclust function implemented in R (version 3.5.1). The
654 pairwise, Euclidean distances between all methylation profiles were calculated and Ward's
655 method was used to cluster the samples into two groups (hypermethylated and unmethylated).
656 This clustering was performed for the samples of GSE43857 and GSE54292 separately as
657 these two data sets were processed in two studies with different piplines[20,45]. The heatmap of
658 methylation patterns was drawn in R using the heatmap.2 function.

659

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

661 For six accessions (Col-0, An-1, C24, Cvi-0, Ler, Kyo), we downloaded the whole-
662 genome DNA methylation data from NCBI from the 1001 Epigenomes Project[20] (GSE43857).
663 For Eri-1 and Sha, DNA methylation data was generated using whole-genome bisulfite
664 sequencing by the Max Planck Genome center. DNA was extracted from plants grown in the
665 greenhouse under standard conditions using the Qiagen DNEasy Plant Mini Kit (Qiagen,
666 Germany) and a sequencing library was prepared using the NEXTflex Bisulfite Library Prep
667 Kit. This library was sequenced on an Illumina HiSeq2000 machine. Sequencing reads were
668 aligned the reference sequence using Bismark (version 0.20.0)[46] with these parameters "-q -
669 -bowtie2 -N 1 -L 24 -p 20". The cytosine methylation profiles in candidate genes were
670 calculated using the same sliding window method as described above. The cytosine
671 methylation profiles together with the profiles from GSE43857 were then clustered again with
672 the same clustering method as described above.

673

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

675 We predicted the presence and absence of functional copies of duplicated genes (*HPA*,
676 *TIM22*) in each of 1,135 accessions from the 1001 Genomes Project according to the
677 annotations of loss-of-function variations and the clustering of cytosine methylation profiles in
678 promoters (Supplementary Data 5-8). For accessions without available methylation
679 sequencing data, we assume the focal genes are expressed. The presence or absence of

680 any functional copies of the reference genes were used as binary phenotype (presence: 1,
681 absence: 0). For the association, we selected 238,166 high-quality SNP markers (minor allele
682 frequency >0.05 and missing rate < 0.1) from 1001 Genomes Project and imputed missing
683 alleles with IMPUTE2[47,48]. An in-house R script implementing the mixed linear model with
684 correction of kinship bias was used to perform the GWA (Bonferroni correction, P < 0.05).
685
686 **Copy number variation analysis**
687 To test for the existence of *HPA3* in a genome, we mapped whole-genome short reads
688 of 1,135 accessions from 1001 Genomes Project[19], of 75 accessions from Africa[33] and of 118
689 accessions from China[37] to the Kyo genome assembly[22] using BWA (version 0.7.15) with
690 default parameters. Accessions with an average mapping coverage >= 5 along the *HPA3*
691 region and duplicated region breakpoints (identified based on the sequence alignment against
692 Col-0 genome using SyRI[49] (version 1.0) with default parameters) were considered as *HPA3*
693 carriers.
694 The copy number of duplicated genes was estimated by the ratio between the average
695 mapping coverage within the focal gene and the average mapping coverage across the whole
696 genome. To estimate the copy number of *HPA*, we mapped the short reads to the Cvi-0
697 genome assembly using BWA (version 0.7.15) with default parameters as the Cvi-0 genome
698 only has one copy of *HPA*. For *TAD3*, *FOLT* and *TIM22*, the copy number was predicted
699 based on short reads mapping against the reference genome where both *TAD3* and *FOLT*
700 only have one copy. For the *TIM22*, copy number was estimated based on the average
701 mapping coverage at both *TIM22-1* and *TIM22-2*.
702
703 **Fine-mapping of incompatible allele at LD2 in Cvi-0 x Col-0 RIL**
704 After the observation that it was not possible to obtain a RIL homozygous for the Col-0
705 allele at LD2.1 while homozygous for the Cvi-0 allele at LD2.3[28], we derived two distinct
706 heterogeneous inbred families (HIF) from two segregating RILs from this population
707 (Supplementary Fig. 2). 8RV467 is segregating for a region largely encompassing LD2.1
708 while fixed Cvi-0 at LD2.3. 8RV408 is segregating for LD2.3 while fixed Col-0 at LD2.1. In
709 each derived HIF family, it is again not possible to fix the incompatible allele combination, so
710 the families were used to exclude intervals that do not contribute to the incompatible
711 interaction. Two rounds of fine-mapping were conducted by genotypically screening
712 increasing populations of descendants (at the seedling stage) for recombinants in the interval.
713 Gradually, the causal interval is reduced and delineated by markers exploiting known SNPs
714 and indels in the Cvi-0 sequence.
715

716 **Segregation of T-DNA mutant line at gene *TIM22-1* and *TIM22-2***

717 T-DNA lines GABI_848H04 segregates for an insertion in *TIM22-1*, while GABI_626H10
718 segregates for an insertion in *TIM22-2*, both in a Col-0 background. Descendance were
719 screened genotypically at the seedling stage to characterize the segregation of the T-DNA
720 insertion allele.

721

722 **Complementation cross to validate the incompatible alleles of *TIM22***

723 A LD2.3 HIF line (8HV408-Het) was crossed with a GABI_626H10 line, both at the
724 heterozygous state, in order to segregate potentially for all four possible hybrid allelic
725 combinations at LD2.3, while maintaining a fixed Col-0 allele at LD2.1. The F1 descendance
726 (123 individuals, from both cross directions) was screened genotypically at the seedling stage
727 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[19] was downloaded from NCBI under the project ID SRP056687. RNA-seq and methylation data from the 1001 Epigenomes Project[20] 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/. Data supporting the findings of this work are available within the paper and its Supplementary Information files. A reporting summary for this article is available as a Supplementary Information file. The source data underlying Figures 1-6 and Supplementary Figures are provided as a Source Data file.

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 Todd Michael (The Salk Institute, USA) for releasing Ty-1 assembly prior to publication, Mehmet Goktay (MPI-PZ, Cologne, Germany) for help with the interpretation of genetic variation in Africa, Pádraic J. Flood 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

763 generated the RAD-seq data. OL, MF, IG and CC performed the mapping and
764 complementation experiments of LD2/*TIM22*. SE and MK generated the AMPRIL population.
765 WBJ and KS wrote the paper.
766

767 ## Competing interests

768 The authors declare no competing interests.

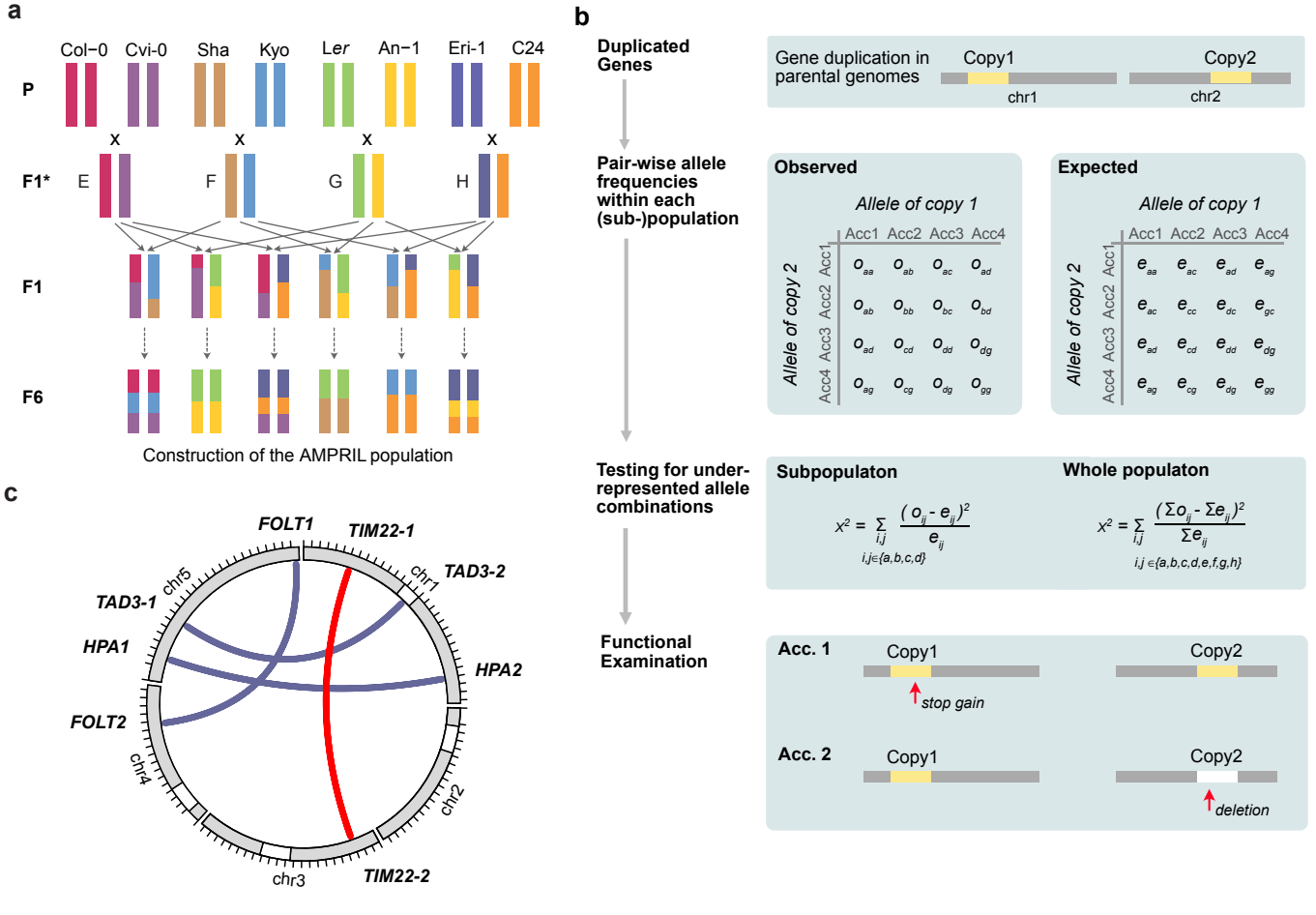
769 ## References

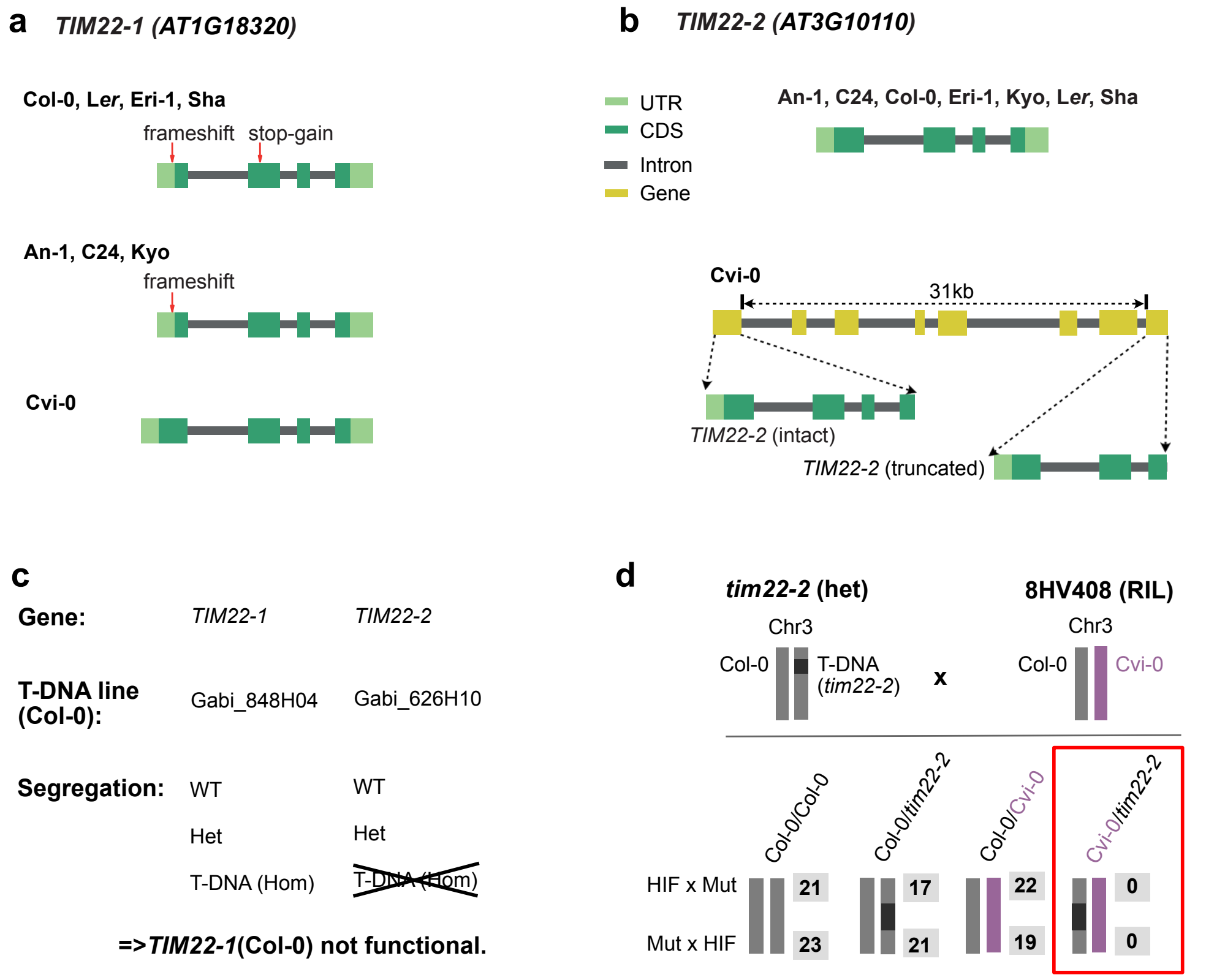
770 1. Maheshwari, S. & Barbash, D. A. The Genetics of Hybrid Incompatibilities. *Annu. Rev.*
771 *Genet.* **45**, 331–355 (2011).
772 2. Bateson, W. Heredity and variation in modern lights. in *Darwin and Modern Science:*
773 *Essays in Commemoration of the Centenary of the Birth of Charles Darwin and of the*
774 *Fiftieth Anniversary of the Publication of the Origin of Species* 85–101 (In: Seward AC
775 (ed). Darwin and Modern Science. Cambridge University, 1909).
776 doi:10.1017/CBO9780511693953.006.
777 3. Dobzhansky, T. *Genetics and the Origin of Species*. (Columbia University Press, 1937).
778 4. Muller, H. J. Isolating mechanisms, evolution and temperature. *Biol. Symp* **6**, 71–125
779 (1942).
780 5. Fishman, L. & Sweigart, A. L. When Two Rights Make a Wrong: The Evolutionary
781 Genetics of Plant Hybrid Incompatibilities. *Annu. Rev. Plant Biol.* **69**, 17.1–17.25 (2018).
782 6. Vaid, N. & Laitinen, R. A. E. Diverse paths to hybrid incompatibility in Arabidopsis. *Plant*
783 *J.* **97**, 199–213 (2019).
784 7. Panchy, N., Lehti-Shiu, M. & Shiu, S.-H. Evolution of Gene Duplication in Plants. *Plant*
785 *Physiol.* **171**, 2294–316 (2016).
786 8. Conant, G. C. & Wolfe, K. H. Turning a hobby into a job: How duplicated genes find
787 new functions. *Nat. Rev. Genet.* **9**, 938–950 (2008).
788 9. Kondrashov, F. A. Gene duplication as a mechanism of genomic adaptation to a
789 changing environment. *Proc. R. Soc. B Biol. Sci.* **279**, 5048–5057 (2012).
790 10. Lynch, M. & Force, A. G. The Origin of Interspecific Genomic Incompatibility via Gene
791 Duplication. *Am. Nat.* **156**, 590–605 (2000).
792 11. Bikard, D. *et al.* Divergent Evolution of Duplicate Genes Leads to Genetic
793 Incompatibilities Within A. thaliana. *Science* **323**, 623–626 (2009).
794 12. Durand, S., Bouché, N., Perez Strand, E., Loudet, O. & Camilleri, C. Rapid
795 establishment of genetic incompatibility through natural epigenetic variation. *Curr. Biol.*
796 **22**, 326–331 (2012).
797 13. Agorio, A. *et al.* An Arabidopsis Natural Epiallele Maintained by a Feed-Forward
798 Silencing Loop between Histone and DNA. *PLoS Genet.* **13**, 1–23 (2017).
799 14. Mizuta, Y., Harushima, Y. & Kurata, N. Rice pollen hybrid incompatibility caused by
800 reciprocal gene loss of duplicated genes. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 20417–
801 20422 (2010).
802 15. Yamagata, Y. *et al.* Mitochondrial gene in the nuclear genome induces reproductive
803 barrier in rice. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 1494–1499 (2010).
804 16. Nguyen, G. N., Yamagata, Y., Shigematsu, Y., Watanabe, M. & Miyazaki, Y.

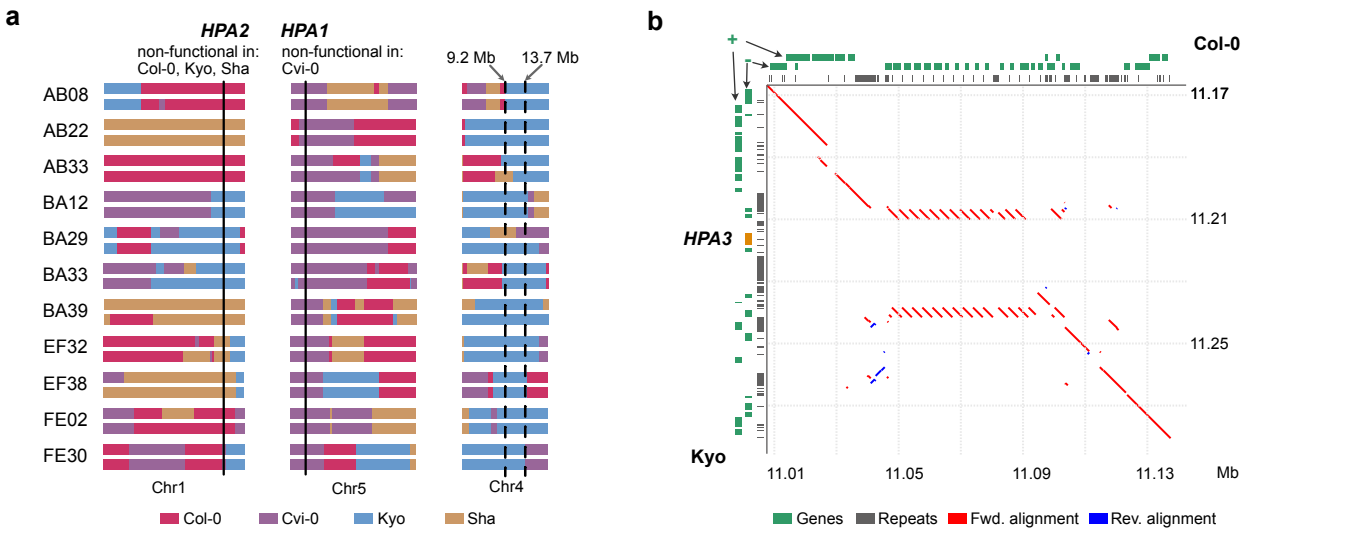
805 Duplication and Loss of Function of Genes Encoding RNA Polymerase III Subunit C4
806 Causes Hybrid Incompatibility in Rice. *G3* **7**, 2565–2575 (2017).
807 17. Zuellig, M. P., Sweigart, A. L., States, U. & E-mail, C. A. Gene duplicates cause hybrid
808 lethality between sympatric species of Mimulus. *PLoS Genet.* **14**, e1007130 (2017).
809 18. Blevins, T., Wang, J., Pflieger, D., Pontvianne, F. & Pikaard, C. S. Hybrid
810 incompatibility caused by an epiallele. *Proc. Natl. Acad. Sci.* **114**, 3702–3707 (2017).
811 19. Alonso-Blanco, C. *et al.* 1,135 Genomes Reveal the Global Pattern of Polymorphism
812 in Arabidopsis thaliana. *Cell* **166**, 481–491 (2016).
813 20. Kawakatsu, T. *et al.* Epigenomic Diversity in a Global Collection of Arabidopsis thaliana
814 Accessions. *Cell* **166**, 492–506 (2016).
815 21. Huang, X. *et al.* Analysis of natural allelic variation in *Arabidopsis* using a multiparent
816 recombinant inbred line population. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 4488–4493
817 (2011).
818 22. Jiao, W. & Schneeberger, K. Chromosome-level assemblies of multiple Arabidopsis
819 genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *Nat.*
820 *Commun.* **11**, 989 (2020).
821 23. Baird, N. A. *et al.* Rapid SNP discovery and genetic mapping using sequenced RAD
822 markers. *PLoS One* **3**, 1–7 (2008).
823 24. Rowan, B. A., Patel, V., Weigel, D. & Schneeberger, K. Rapid and inexpensive whole-
824 genome genotyping-by-sequencing for crossover localization and fine-scale genetic
825 mapping. *G3 Genes, Genomes, Genet.* **5**, 385–398 (2015).
826 25. Ackermann, M. & Beyer, A. Systematic detection of epistatic interactions based on
827 allele pair frequencies. *PLoS Genet.* **8**, e1002463 (2012).
828 26. Corbett-Detig, R. B., Zhou, J., Clark, A. G., Hartl, D. L. & Ayroles, J. F. Genetic
829 incompatibilities are widespread within species. *Nature* **504**, 135–137 (2013).
830 27. Murcha, M. W. *et al.* Characterization of the preprotein and amino acid transporter gene
831 family in Arabidopsis. *Plant Physiol.* **143**, 199–212 (2007).
832 28. Simon, M. *et al.* Quantitative trait loci mapping in five new large recombinant inbred
833 line populations of Arabidopsis thaliana genotyped with consensus single-nucleotide
834 polymorphism markers. *Genetics* **178**, 2253–2264 (2008).
835 29. Muralla, R., Sweeney, C., Stepansky, A., Leustek, T. & Meinke, D. Genetic dissection
836 of histidine biosynthesis in Arabidopsis. *Plant Physiol.* **144**, 890–903 (2007).
837 30. Kawakatsu, T. *et al.* Epigenomic Diversity in a Global Collection of Arabidopsis thaliana
838 Accessions. *Cell* **166**, 492–506 (2016).
839 31. Gerber, A. P. & Keller, W. An adenosine deaminase that generates inosine at the
840 wobble position of tRNAs. *Science* **286**, 1146–1149 (1999).
841 32. Törjék, O. *et al.* Segregation distortion in Arabidopsis C24/Col-0 and Col-0/C24

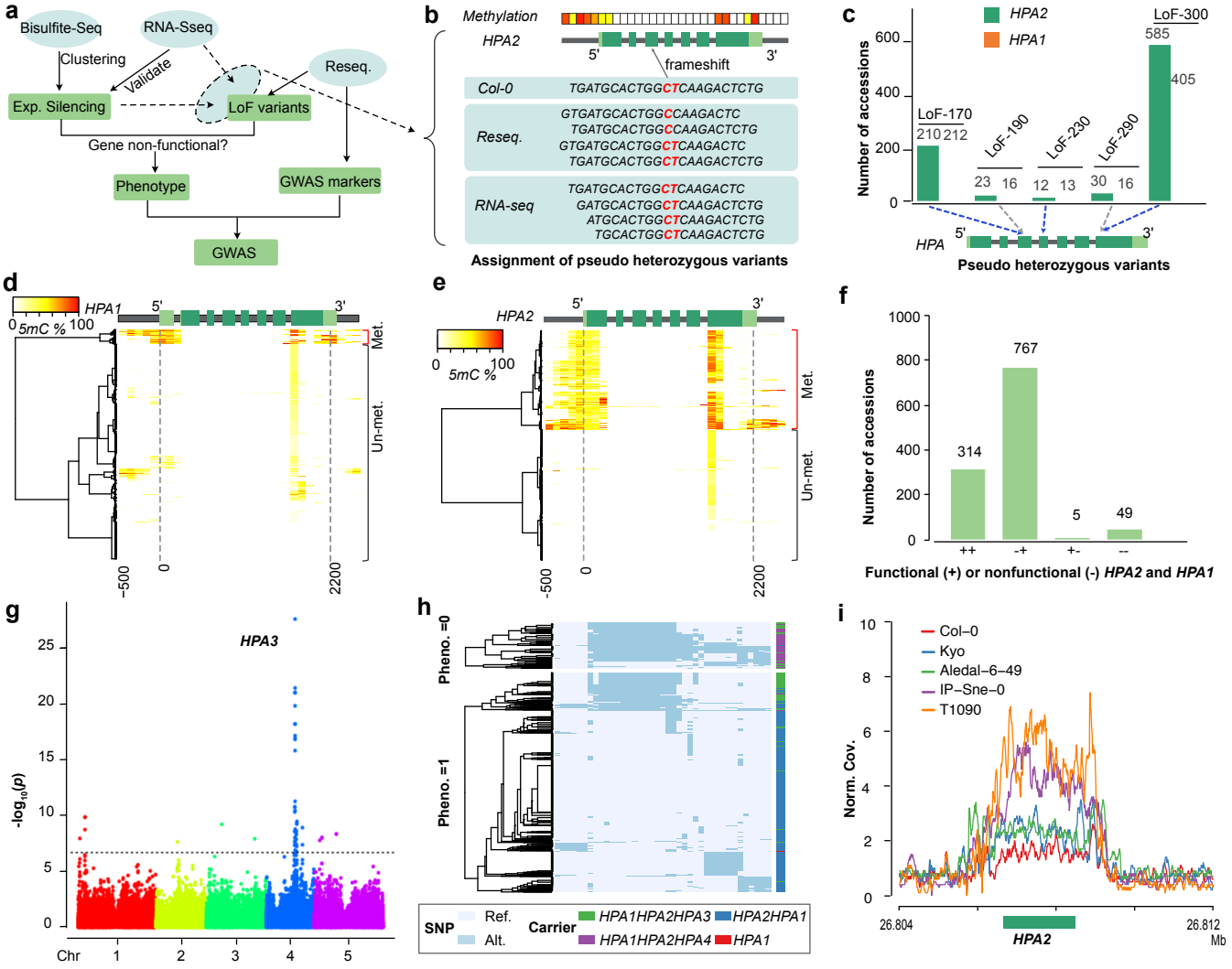
842 recombinant inbred line populations is due to reduced fertility caused by epistatic
843 interaction of two loci. *Theor. Appl. Genet.* **113**, 1551–1561 (2006).
844 33. Durvasula, A. *et al.* African genomes illuminate the early history and transition to selfing
845 in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci.* **114**, 5213–5218 (2017).
846 34. Etter, P. D., Bassham, S., Hohenlohe, P. A., Johnson, E. A. & Cresko, W. A. SNP
847 Discovery and Genotyping for Evolutionary Genetics Using RAD Sequencing BT -
848 Molecular Methods for Evolutionary Genetics. in (eds. Orgogozo, V. & Rockman, M. V)
849 157–178 (Humana Press, 2011). doi:10.1007/978-1-61779-228-1_9.
850 35. Ossowski, S. *et al.* Sequencing of natural strains of Arabidopsis thaliana with short
851 reads. *Genome Res.* (2008) doi:10.1101/gr.080200.108.
852 36. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome
853 comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**,
854 157 (2015).
855 37. Zou, Y.-P. *et al.* Adaptation of Arabidopsis thaliana to the Yangtze River basin.
856 *Genome Biol.* **18**, 239 (2017).
857 38. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering
858 plant Arabidopsis thaliana. *Nature* **408**, 796–815 (2000).
859 39. Lamesch, P. *et al.* The Arabidopsis Information Resource (TAIR): Improved gene
860 annotation and new tools. *Nucleic Acids Res.* **40**, (2012).
861 40. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler
862 transform. *Bioinformatics* **25**, 1754–1760 (2009).
863 41. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**,
864 2078–2079 (2009).
865 42. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: A pattern growth
866 approach to detect break points of large deletions and medium sized insertions from
867 paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009).
868 43. Rausch, T. *et al.* DELLY: Structural variant discovery by integrated paired-end and
869 split-read analysis. *Bioinformatics* **28**, 333–339 (2012).
870 44. Cingolani, P. *et al.* A program for annotating and predicting the effects of single
871 nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster
872 strain w1118; iso-2; iso-3. *Fly (Austin).* **6**, 80–92 (2012).
873 45. Dubin, M. J. *et al.* DNA methylation in Arabidopsis has a genetic basis and shows
874 evidence of local adaptation. *Elife* **4**, 1–23 (2015).
875 46. Krueger, F. & Andrews, S. R. Bismark: A flexible aligner and methylation caller for
876 Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
877 47. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation
878 method for the next generation of genome-wide association studies. *PLoS Genet.* **5**,

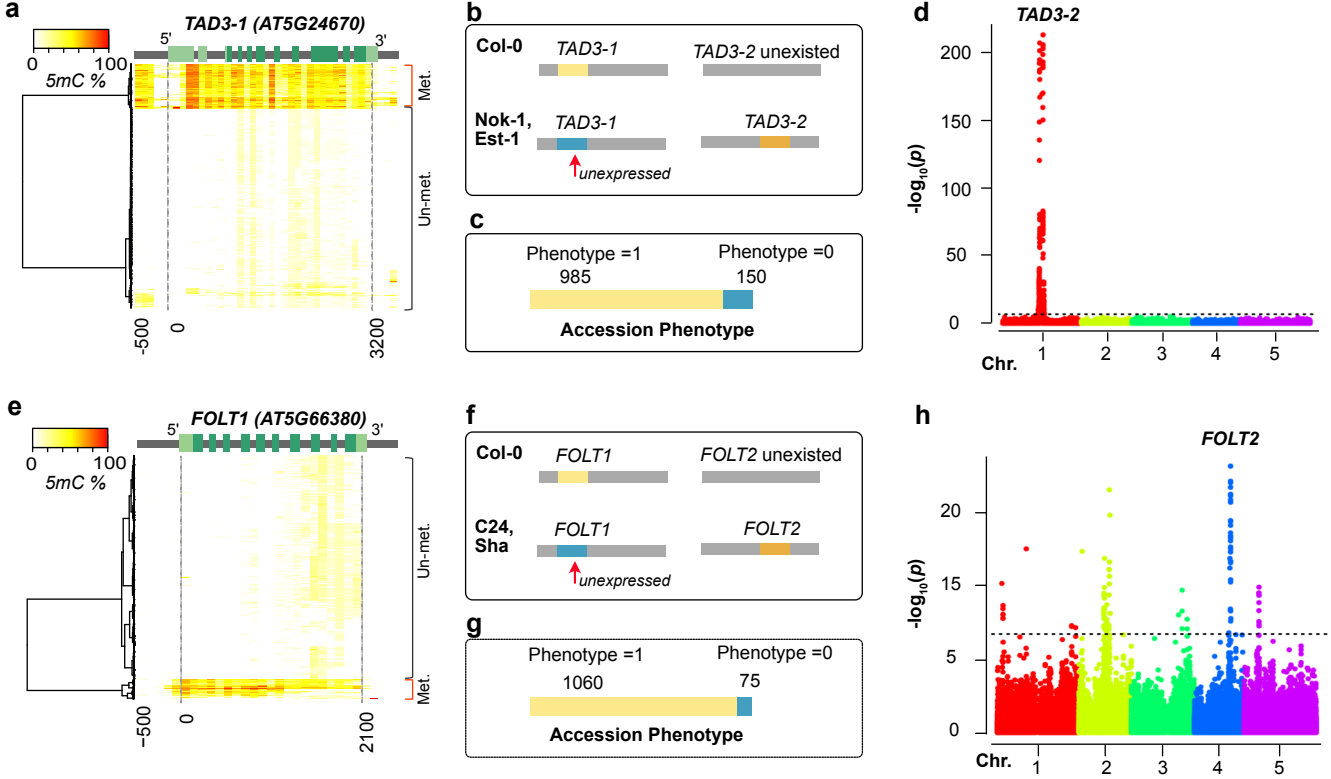
879 (2009).
880 48. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and
881 accurate genotype imputation in genome-wide association studies through pre-phasing.
882 *Nat. Genet.* **44**, 955–959 (2012).
883 49. Goel, M., Sun, H., Jiao, W.-B. & Schneeberger, K. SyRI: finding genomic
884 rearrangements and local sequence differences from whole-genome assemblies.
885 *Genome Biol.* **20**, 277 (2019).
886

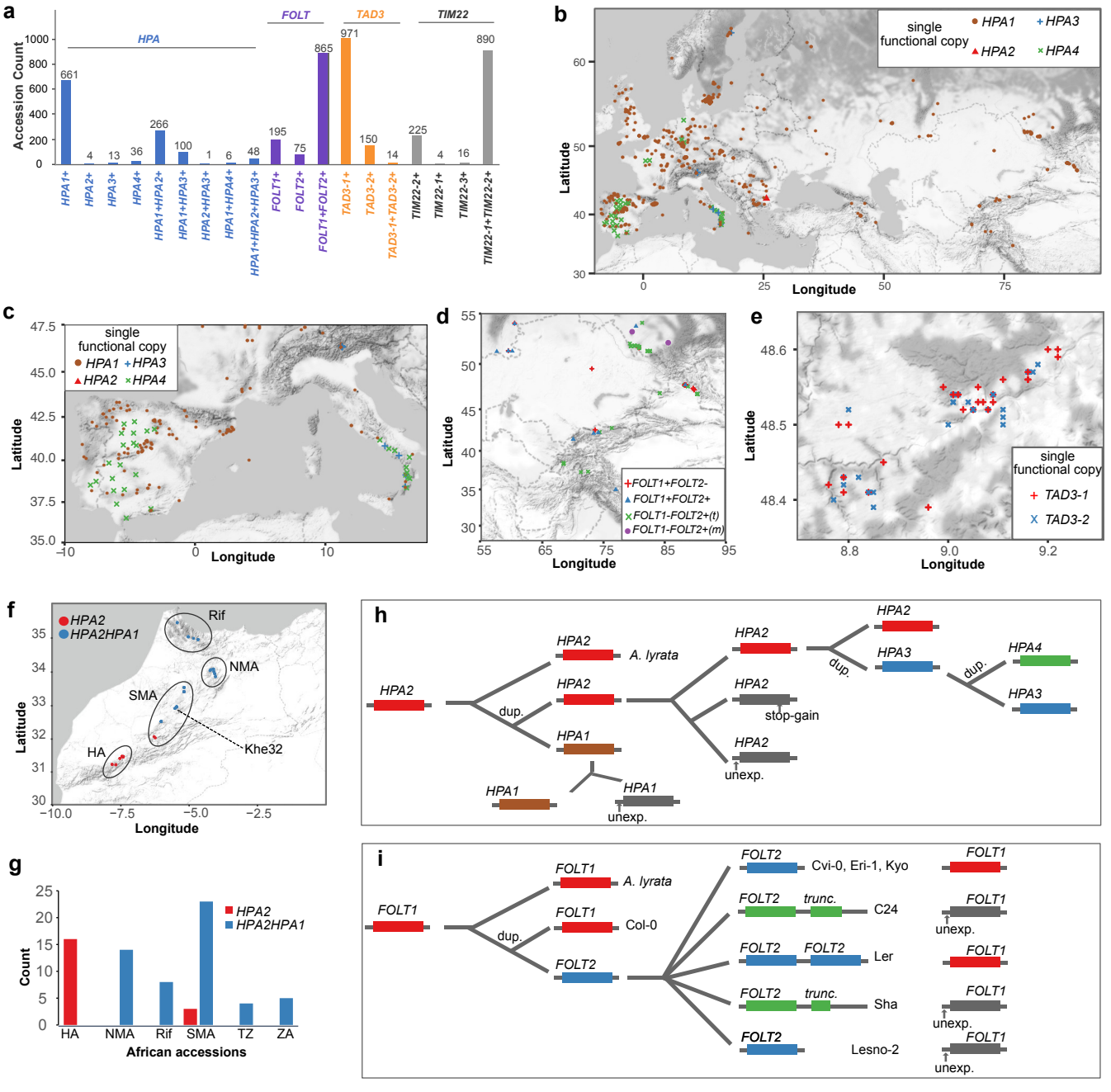

a
Col−0
Cvi-0
Sha
Kyo
Ler
An−1
Eri-1
C24
P
x
F1*
E
F
G
H
F1
F6
Construction of the AMPRIL population
c
FOLT1
TIM22-1
TAD3-2
HPA2
TIM22-2
chr3
FOLT2
HPA1
TAD3-1
chr1
chr2
chr4
chr5
b
Duplicated Genes
Gene duplication in parental genomes
Copy1
Copy2
chr1
chr2
Pair-wise allele frequencies within each (sub-)population
Observed
Expected
Allele of copy 1
Allele of copy 2
Acc1 Acc2 Acc3 Acc4
Testing for under-represented allele combinations
Subpopulaton
Whole populaton
Functional Examination
Acc. 1
Acc. 2
stop gain
deletion


a
TIM22-1 (AT1G18320)
Col-0, Ler, Eri-1, Sha
frameshift
stop-gain
An-1, C24, Kyo
frameshift
Cvi-0
b
TIM22-2 (AT3G10110)
UTR
CDS
Intron
Gene
An-1, C24, Col-0, Eri-1, Kyo, Ler, Sha
Cvi-0
31kb
TIM22-2 (intact)
TIM22-2 (truncated)
c
Gene:
TIM22-1
TIM22-2
T-DNA line (Col-0):
Gabi_848H04
Gabi_626H10
Segregation:
WT
WT
Het
Het
T-DNA (Hom)
T-DNA (Hom)
=>TIM22-1(Col-0) not functional.
d
tim22-2 (het)
8HV408 (RIL)
Chr3
Col-0
T-DNA (tim22-2)
x
Chr3
Col-0
Cvi-0
Col-0/Col-0
Col-0/tim22-2
Col-0/Cvi-0
Cvi-0/tim22-2
HIF x Mut
Mut x HIF
21
23
17
21
22
19
0
0


a
HPA2
non-functional in:
Col-0, Kyo, Sha
HPA1
non-functional in:
Cvi-0
9.2 Mb
13.7 Mb
AB08
AB22
AB33
BA12
BA29
BA33
BA39
EF32
EF38
FE02
FE30
Chr1
Chr5
Chr4
Col-0
Cvi-0
Kyo
Sha
b
Col-0
11.17
11.21
11.25
HPA3
Kyo
11.01
11.05
11.09
11.13
Mb
Genes
Repeats
Fwd. alignment
Rev. alignment


a
Bisulfite-Seq
RNA-Sseq
Clustering
Validate
Reseq.
Exp. Silencing
LoF variants
Gene non-functional?
Phenotype
GWAS markers
GWAS
b
Methylation
HPA2
5'
3'
frameshift
Col-0
TGATGCACTGGCTCAAGACTCTG
Reseq.
GTGATGCACTGGCCAAGACTC
TGATGCACTGGCCAAGACTCTG
GTGATGCACTGGCTCAAGACTC
TGATGCACTGGCTCAAGACTCTG
RNA-seq
TGATGCACTGGCTCAAGACTC
GATGCACTGGCTCAAGACTCTG
ATGCACTGGCTCAAGACTCTG
TGCACTGGCTCAAGACTCTG
Assignment of pseudo heterozygous variants
c
HPA2
HPA1
Number of accessions
LoF-170
210 212
LoF-190
23 16
LoF-230
12 13
LoF-290
30 16
LoF-300
585
405
HPA
5'
3'
Pseudo heterozygous variants
d
HPA1
5mC %
Met.
Un-met.
-500
0
2200
e
HPA2
5mC %
Met.
Un-met.
-500
0
2200
f
Number of accessions
314
767
5
49
++
-+
+-
--
Functional (+) or nonfunctional (-) HPA2 and HPA1
g
HPA3
-log10(p)
Chr
h
Pheno. =0
Pheno. =1
SNP
Ref.
Alt.
Carrier
HPA1HPA2HPA3
HPA2HPA1
HPA1HPA2HPA4
HPA1
i
Col-0
Kyo
Aledal-6-49
IP-Sne-0
T1090
Norm. Cov.
26.804
26.812
Mb
HPA2


a
TAD3-1 (AT5G24670)
5'
3'
0
100
5mC %
Met.
Un-met.
-500
0
3200
b
Col-0
TAD3-1
TAD3-2 unexisted
Nok-1,
Est-1
TAD3-1
TAD3-2
unexpressed
c
Phenotype =1
985
Phenotype =0
150
Accession Phenotype
d
TAD3-2
-log10(p)
200
150
100
50
0
Chr.
1
2
3
4
5
e
FOLT1 (AT5G66380)
5'
3'
0
100
5mC %
Un-met.
Met.
-500
0
2100
f
Col-0
FOLT1
FOLT2 unexisted
C24,
Sha
FOLT1
FOLT2
unexpressed
g
Phenotype =1
1060
Phenotype =0
75
Accession Phenotype
h
FOLT2
-log10(p)
20
15
5
0
Chr.
1
2
3
4
5


a
Accession Count
HPA
FOLT
TAD3
TIM22
661
4
13
36
266
100
1
6
48
195
75
865
971
150
14
225
4
16
890
HPA1+
HPA2+
HPA3+
HPA4+
HPA1+HPA2+
HPA1+HPA3+
HPA2+HPA3+
HPA1+HPA4+
HPA1+HPA2+HPA3+
FOLT1+
FOLT2+
FOLT1+FOLT2+
TAD3-1+
TAD3-2+
TAD3-1+TAD3-2+
TIM22-2+
TIM22-1+
TIM22-3+
TIM22-1+TIM22-2+
b
single functional copy
HPA1
HPA2
HPA3
HPA4
Latitude
Longitude
c
single functional copy
HPA1
HPA2
HPA3
HPA4
Latitude
Longitude
d
FOLT1+FOLT2-
FOLT1+FOLT2+
FOLT1-FOLT2+(t)
FOLT1-FOLT2+(m)
Latitude
Longitude
e
single functional copy
TAD3-1
TAD3-2
Latitude
Longitude
f
HPA2
HPA2HPA1
Rif
NMA
SMA
HA
Khe32
Latitude
Longitude
g
HPA2
HPA2HPA1
Count
HA
NMA
Rif
SMA
TZ
ZA
African accessions
h
HPA2
dup.
HPA2
HPA1
A. lyrata
HPA1
unexp.
HPA2
HPA2
stop-gain
HPA2
unexp.
dup.
HPA2
HPA3
dup.
HPA4
HPA3
i
FOLT1
dup.
FOLT1
FOLT1
FOLT2
A. lyrata
Col-0
FOLT2
Cvi-0, Eri-1, Kyo
FOLT2
trunc.
C24
FOLT2
FOLT2
Ler
FOLT2
trunc.
Sha
FOLT2
Lesno-2
FOLT1
FOLT1
unexp.
FOLT1
FOLT1
unexp.
FOLT1
unexp.