The Evolutionary Dynamics of Genetic Incompatibilities Introduced by Duplicated Genes in Arabinopsis 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 and Korbinian Schneeberger*1,9

1Department of Chromosome Biology, Max Planck Institute for Plant Breeding Research, Cologne, Germany
2Department of Plant Developmental Biology, Max Planck Institute for Plant Breeding Research, Cologne, Germany
3Department of Breeding Research, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, Stadt Seeland, Germany
4Department of Plant Breeding and Genetics, Max Planck Institute for Plant Breeding Research, Cologne, Germany
5Department of Cell Biology and Genetics, Faculty of Science, Palacký University Olomouc, Olomouc, Czech Republic.
6Institut Jean-Pierre Bourgin, INRAE, AgroParisTech, Université Paris-Saclay, Versailles, France
7Laboratory of Genetics, Wageningen University & Research, Wageningen, The Netherlands
8Institute of Experimental Botany (IEB), Czech Academy of Sciences, Centre of the Region Haná for Biotechnological and Agricultural Research (CRH), Olomouc, Czech Republic
9Faculty of Biology, LMU Munich, Planegg-Martinsried, Germany
†Present address: AlceDiag, Montpellier, France
*Corresponding author: E-mail: schneeberger@mpipz.mpg.de.
Associate editor: Stephen Wright

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 pseudofunctionalized in different genotypes, genetic incompatibilities can arise in their hybrid offspring. Although 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 that 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 coexist, 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.

Key words: genetic incompatibility, duplicated gene, HPA, TIM22, genome-wide association study, loss of function.

Introduction

Genetic incompatibilities describe the decrease of fitness due to dysfunctional allele combinations in hybrid individuals (Maheshwari and Barbash 2011). 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 pseudofunctionalization (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, pseudofunctionalization 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 the 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/intraspecific hybrids of Arabinopsis thaliana (Bikard et al. 2009; Durand et al. 2012; Agorio et al. 2017), rice (Mizuta et al. 2010;
Identification of incompatible alleles, however, often relied on genetic mapping in experimental populations, which is a time-consuming and costly process. As incompatible alleles at duplicated genes are frequently introduced by loss-of-function (LoF) mutations (such as stop-codon gain, frameshift, gene deletion) and epimutations (Bikard et al. 2009; Blevins et al. 2017), initial examination of LoF (epi)mutations within whole-(epi)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 as phenotypes, we mapped modifiers of all four incompatibilities using GWA. The LoF alleles from duplicated genes were geographically widely distributed, and coexisted within 1,135 accessions of the 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.

We developed a two-step workflow to combine genetic and genomic evidence to quickly identify incompatible alleles of duplicated genes (fig. 1b). In the first step, we selected 781 distal (interchromosome) duplicated gene pairs including 612 gene pairs in which the reference sequence contains two copies and other founder genomes feature at least one copy (supplementary data 2, Supplementary Material online). In the remaining 169 gene pairs, the reference sequence only has one copy and at least one other parental genome has an additional copy in a different chromosome. As genetic incompatibility leads to the underrepresentation of incompatible allele combinations (Ackermann and Beyer 2012; Corbett-Detig et al. 2013), we searched for significant distortions from the expected frequencies of all parental allele combinations across all 781 duplicated gene pairs in all 12 subpopulations, two merged subpopulations (ABBA and EFFE, CDDC and GHHG as they share the same founders), and the whole population (see Materials and Methods). These tests revealed significant distortions in 236 gene pairs ($\chi^2$ test, $P$ value < 0.05, multiple testing corrected) in at least one of the populations.

However, the observed distortions do not necessarily result from genetic incompatibilities in the tested gene. Alternatively, such distortions can also occur if the tested gene duplicate is closely linked to a genetic incompatibility. Hence, in a second step, we examined the alleles of the gene pairs in the founder genomes for LoF variations or hypermethylated promoters (fig. 1b, see Materials and Methods). This examination revealed three gene pairs with functional disruption in both of the duplicates in at least one of the founder genomes. These three duplicated genes included two, HISTIDINOL PHOSPHATE AMINOTRANSFERASE (HPA) (Bikard et al. 2009) and FOLATE TRANSPORTER (FOLT) (Durand et al. 2012), which were already known for their ability to introduce genetic incompatibilities, as well as one gene pair, which so far was not reported as the genetic basis
for a genetic incompatibility, TIM22 (TIM22-1: AT1G18320, TIM22-2: AT3G10110) (fig. 1c). For all other 233 gene pairs, we could not identify nonfunctional alleles in both copies.

We noted that another duplicated gene, tRNA ADENOSINE DEAMINASE 3 (TAD3), which is also known to introduce a genetic incompatibility (Agorio et al. 2017), was not considered by our initial testing even though the genotypes of the AMPRIL founders lead to incompatible allele combinations in the RIL populations: all founder genomes except for Kyo have a functional TAD3-1 (supplementary table 1, Supplementary Material online), whereas the Kyo TAD3-1 gene is silenced most likely due to its methylated promoter (similar to the Nok-1 and Est-1 accessions in which the incompatibility was described originally; Agorio et al. 2017). The lack of a functional TAD3-1 in Kyo is counterbalanced by additional copies (TAD3-2) (which however were not part of the main chromosome scaffolds of the Kyo genome assembly). Therefore, the TAD3 gene duplicate was not considered initially, however, we also used this incompatibility for further analysis.

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, whereas Cvi-0 includes an extra truncated copy ~31 kb downstream to TIM22-2 (fig. 2a and b). We found significant segregation distortions in the complete AMPRIL population and one subpopulation, EGGE, where the double homozygous allele combination TIM22-1<sup>Col-0</sup>/TIM22-2<sup>Cvi-0</sup> was significantly underrepresented (supplementary tables 2 and 3, Supplementary Material online).

We observed an in-frame premature stop-codon in Col-0 (mis-annotated in the reference annotation) suggesting that TIM22-1 is not functional in Col-0 (fig. 2a and supplementary table 4, Supplementary Material online). To test if TIM22-1 is...
The loss-of-function variants (1-bp frameshift indel and a premature stop-codon) relative to the intact TIM22 allele is nonfunctional. HIF, heterogeneous inbred families; Mut, mutant.

Fig. 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 (1-bp frameshift indel and a premature stop-codon) relative to the intact TIM22-1 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 about 31 kb downstream of TIM22-2. (c) Segregation of T-DNA alleles within the descendence of two 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., nonfunctional for TIM22-2) 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 possible F1 progenies are shown (in gray) for both cross directions. Although the TIM22-2 Col-0 allele can complement the T-DNA, the TIM22-2 Cvi-0 allele could not, implying that the TIM22-2 Cvi-0 allele is nonfunctional. HIF, heterogeneous inbred families; Mut, mutant.

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

Collectively, these segregation and complementation crosses show that the combination of different nonfunctional alleles of the TIM22 copies leads to a drastic reduction of allelic combinations in offspring populations, and thus evidence the causative role of TIM22 in this genetic incompatibility.

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 nonfunctional 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 nonfunctional HPA2 due to a premature stop codon or a hypermethylated promoter, whereas Cvi-0 carries a functional HPA2 allele, but does not carry HPA1 at all (supplementary table 5, Supplementary Material online). Unexpectedly, however, we did observe homozygous HPA1/HPA2 incompatible allele combinations (HPA2+/−/HPA1+/−) in 11 of the AMPRIL lines within the ABBA and EFFE subpopulations which were derived from Col-0, Cvi-0, Kyo, and Sha (supplementary tables 6–8, Supplementary Material online). Further analysis of these populations revealed an extremely high frequency of the Kyo allele on chromosome 4 (supplementary fig. 3, Supplementary Material online), making us recognize that the 11 AMPRIL lines with the incompatible allele combinations all carried at least one Kyo allele at chr4:9.2–13.7 Mb, suggesting that a Kyo allele in this region can rescue the incompatibility. This analysis showed that the negative effects of incompatible allele combinations can be overcome if they are rescued by modifying alleles. Therefore, the virtual absence of functional alleles (among the known loci) of a duplicated gene could act as a molecular phenotype to map the location of additional (rescuing) alleles by genome-wide association (GWA) mapping even in natural populations.

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

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

By combining the LoF variant genotyping and the methylation analyses, we found 767 accessions with a nonfunctional HPA2 (HPA2−/− HPA1+/+) allele and five accessions
with a nonfunctional HPA1 (HPA2+/-HPA1+-/-) allele (fig. 4f and supplementary table 10, Supplementary Material online). In addition, 49 accessions did not feature any functional alleles (HPA2-/-HPA1-/-) and were expected to carry additional modifier(s) to complement the loss of functional copies. To find the locations of these modifiers, we used the absence of functional copies as phenotype (supplementary data 3, Supplementary Material online) to run a GWA under the mixed linear model using the SNP markers from the 1001 Genomes Project (www.1001genomes.org). This GWA revealed a significantly associated region at chr4:11.00-11.15 Mb (fig. 4g and supplementary fig. 5, Supplementary Material online, alpha level of 0.05, Bonferroni correction), corresponding to the HPA3 locus found in Kyo (chr4:11.11Mb). Though other peaks in unlinked regions were present, these additional loci explained only a small proportion of the heritability.

Analyzing the haplotypes at the 39 significantly associated SNP markers at HPA3 locus revealed a somewhat, but not entirely homogenous haplotype in many of the HPA2-/-HPA1-/- carriers (fig. 4h). To confirm that the modifying haplotype is still identical to the Kyo allele, we aligned the short reads of all accessions of the 1001 Genomes Project against the Kyo reference sequence and found that overall 162 accessions carried the HPA3 allele (supplementary data 3, Supplementary Material online). However, unexpectedly, the Kyo HPA3 was only found in 13 of the 49 accessions without functional HPA2 and HPA1 alleles, suggesting the presence of two different modifiers at this locus on chromosome 4. To find more support for this, we ran a new GWA without the 162 HPA3 (Kyo-like allele) carriers, which still led to a significantly associated locus at the region of HPA3 (supplementary fig. 6, Supplementary Material online). Further short-read mappings of all 1,135 genomes against the Cvi-0 reference sequence revealed multiple additional copies of TIM22-1/TIM22-2 allele combinations carried at least one additional third copy of TIM22, whereas among all 1,119 other genomes of the 1001 Genomes Project only three genomes carried additional copies. This suggests that, like for HPA, known nonfunctional allele combinations of TIM22 are in fact rescued by additional copies. Moreover, in ten of the 16 accessions, TIM22-2 showed not only one but multiple additional copies (hereafter named as TIM22-3), which was further supported by the genome assembly of Ty-1 (https://genomeweb.org/CoGe/Genomelnfo.pl?gid=54584, unpublished), where we could find a cluster of four tandemly arranged TIM22 gene copies at the TIM22-2 locus.

Because the reference sequence only contained one copy of TAD3 (TAD3-1) (Agorio et al. 2017) and FOLT (FOLT1) (Durand et al. 2012), we modified our GWA method and only used nonfunctional alleles at the reference gene as the phenotype to map modifiers for the two remaining incompatibilities (fig. 5 and supplementary fig. 8, Supplementary Material online). Due to this modification we would expect to map also the location of the duplicated genes as we had found them in the AMPRIL founders. For TAD3, an essential ortholog of the yeast tRNA Adenosine deaminase 3 (Gerber and Keller 1999), we did not find any accessions with LoF alleles, but we found 150 accessions with a hypermethylated promoter in TAD3-1 similar to the methylated promoters found in Nok-1 and Est-1, which are known to be nonfunctional due to this methylated promoter region (Agorio et al. 2017) (fig. 5a and b). The GWA result revealed one significant peak (fig. 5c and d), which as expected colocated with the TAD3-2 locus (Agorio et al. 2017). All the 150 accessions carried multiple additional copies of TAD3-2 (supplementary data 5, Supplementary Material online), similar to Nok-1 and Est-1. This suggests that the expression silencing of TAD3-1 is common in natural populations and that the rescue of this loss-of-function allele is generally mediated by additional gene copies at the TAD3-2 locus as it was shown in the original description of this incompatibility (Agorio et al. 2017).

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, Supplementary Material online). 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 nonfunctional 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 pseudoheterozygous SNPs in the RNA-seq data.
methylated promoter of FOLT1 and FOLT2 genes in only 46 out of such 75 accessions (fig. 5). Revealed the presence of mapping against the C24 and Sha reference sequences FOLT1 Material online) suggesting that the methylation of the gion corresponding to GWA approach to find these interacting loci of FOLT1 gene copies in these 75 accessions. When we repeated our FOLT1 Besides providing evidence for the expression silencing of successions of Sha alleles are segregated out. This is consistent to what was shown in 1232 (supplementary fig. 9, Supplementary Material online). Similarly, most accessions which varied substantially across the accessions of the 1001 Genomes Project (fig. 6).

For example, among the accessions that carried only one functional copy of HPA, we found 661, 4, 13, and 36 accessions with a single functional copy of HPA1, HPA2, HPA3, or HPA4, respectively (fig. 6a and b and supplementary fig. 10, Supplementary Material online). Similarly, most accessions only featured one functional copy of TAD3 including 971 accessions with only TAD3-1 and 150 with only TAD3-2 (fig. 6a and supplementary table 14 and supplementary fig. 11, Supplementary Material online). In contrast, most of the

(supplementary fig. 9, Supplementary Material online). Besides providing evidence for the expression silencing of FOLT1, this also suggested the existence of additional FOLT gene copies in these 75 accessions. When we repeated our GWA approach to find these interacting loci of FOLT1, we found multiple significantly associated loci including one region corresponding to FOLT2 (fig. 5g and h).

Further analyses of the FOLT gene copies using short read mapping against the C24 and Sha reference sequences revealed the presence of FOLT2 in all the 75 accessions with methylated promoter of FOLT1 and truncated copies of FOLT2 genes in only 46 out of such 75 accessions (fig. 5e and supplementary table 13 and data 6, Supplementary Material online) suggesting that the methylation of the FOLT1 promoter remains stable even after the truncated copies are segregated out. This is consistent to what was shown in successive generations of Sha × Col-0 RILs where the inducing locus was segregated away six generations ago (Durand et al. 2012).

Taken together, we found evidence that all four incompatible allele combinations, initially identified within an artificial intercross population, also occur in nature, and that the incompatible allele combinations of three of them were surprisingly common. For all incompatible allele combination carriers, we found evidence for the existence of additional copies, and could even map the locations of some of those using GWA.

Geographic Distribution of Incompatible Allele Combinations
We next asked how prevalent the potential for incompatible allele combinations was within natural populations of A. thaliana by analyzing the presence of different haplotypes in different geographic regions. For this, we first analyzed the allele frequencies of different haplotypes (i.e., the different combinations of functional alleles within individual plants), which varied substantially across the accessions of the 1001 Genomes Project (fig. 6).

Fig. 5. Mapping nonreference gene copies of incompatible alleles using GWAS. (a and e) Hierarchical clustering of cytosine methylation profiles in TAD3-1 (a) and FOLT1 (e) based on 888 Arabidopsis 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 and 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/Shan as shown previously (Durand et al. 2012; Agorio et al. 2017). (c and g) The number of accessions with functional (Phenotype = 1) or nonfunctional (Phenotype = 0) TAD3-1 (c) and FOLT1 (g) gene copies. (d and 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 (−log10P = 6.68).
accessions included multiple functional alleles of FOLT as well as of TIM22 (fig. 6a).

Genetic incompatibilities become effective in the offspring of accessions with incompatible alleles. In particular, the offspring of accessions with only one functional gene copy will lead to the highest reduction in fitness. We found numerous accessions with the potential to create incompatible allele combinations in their offspring in geographically close regions (fig. 6c–e). For example, among 30 accessions in the South of Italy we found nine 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.

Interestingly, within some of these contact zones of accessions with incompatible alleles, we also found accessions with
multiple functional gene copies. For example, in South Italy there were seven of the 30 accessions that featured multiple functional copies of HPA, and similarly in Central Asia, we could find multiple accessions with two functional copies of FOLT in addition to the accessions with only one functional copy (fig. 6d and supplementary figs. 10–12, Supplementary Material online). In contrast, however, we only observed a few accessions with multiple functional alleles in the contact zone of incompatible TAD3 alleles, most likely because only 14 such accessions were found in the entire set of accessions. Even though we could not find evidence that such contact zones are enriched for additional copies or that those additional copies would evolve in these regions, the presence of haplotypes with additional gene copies in these contact zones have the potential to mediate the gene flow between the haplotypes with the incompatible allele combinations.

Origin and Evolution of Incompatible Alleles
To figure out when these incompatible alleles originated and how they evolved, we investigated their ancestral genotypes within African genotypes which likely represent the ancestral populations of the Eurasian accessions (Durvasula et al. 2017). Although almost all (99.7%) of the Eurasian accessions had both HPA2 and HPA1 (either functional or not), we found that only 27% (20 out of 75) of the African accessions carried only HPA2 (fig. 6f and g; supplementary data 3, Supplementary Material online). This suggested that HPA2 was the ancestral copy of HPA which was further supported by the synteny alignment with the close relative Arabidopsis lyrata (Blevins et al. 2017), which only featured a single HPA gene in a region syntenic to HPA2 and that the duplication events leading to HPA1 happened early on in Africa (fig. 6h). One accession, Khe32, from Morocco carried HPA1 along with the most frequent LoF variant LoF-300 in HPA2 in the Eurasian accessions, suggesting that the first accession with only a 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, Supplementary Material online), whereas 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, Supplementary Material online), suggesting free segregation of different HPA alleles. In striking contrast to this, the inactive alleles of HPA1 (i.e., alleles with a nonfunctional 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, Supplementary Material online) 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, whereas accessions with multiple copies of TAD3 can only be observed in Eurasia (supplementary fig. 10, Supplementary Material online). Similarly, 38 of the African accessions only have FOLT1, whereas 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 data 6, Supplementary Material online). 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 (Jiao and Schneeberger 2020) (supplementary fig. 17 and data 6, Supplementary Material online). 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 pseudofunctionalization in separate genomes (Lynch and Force 2000). Here, we first identified incompatible allele combinations of four duplicated gene pairs by integrating genetic and genomic information of a multiparental intercross population. Unexpectedly, when we searched the genomes of 1,135 accessions released by the 1001 Genomes Project (Alonso-Blanco et al. 2016), we identified many natural accessions with putatively incompatible allele combinations of these four genes and could elucidate the genetics of their incompatibilities and modifiers as they occur in natural populations using GWA. As duplicated genes are common in plant genomes (Lynch and Conery 2000; Maere et al. 2005; Panchy et al. 2016), duplicated gene based incompatibilities might be more common than typically anticipated, suggesting that our GWA approach could be further applied to search for more candidates in other genes.

Using multiomics data from hundreds of A. thaliana accessions, we found that incompatible alleles are surprisingly frequent in nature and also occur in sympatry, implying that the evolution of genetic incompatibilities does not require separated populations as proposed by BDM model (Bateson 1909; Dobzhansky 1937; Muller 1942). Moreover, incompatible allele combinations can persist over long periods as some of the alleles studied here may have originated before A. thaliana colonized Eurasia. This might be due to the inefficient elimination of nonfunctional alleles in a predominantly selfing species (Bustamante et al. 2002). Arabidopsis thaliana wild populations have an outcrossing rate of around 1%, which however, can reach up to 15% in some outcrossing hotspots (Bakker et al. 2006; Bomblyies et al. 2010), suggesting that the incompatible alleles might be recombined in the same
individuals at some regions. Some recombined alleles (such as double homozygous recessive alleles) will be removed when no mutation of extrafunctional copy appears, which in turn can reduce the frequency of incompatible allele combinations.

As incompatibilities introduced by duplicated genes are likely caused by neutral mutations and genetic drift (Lynch and Force 2000), such neutral incompatibility alleles of duplicated genes are supposed to be purged even under low level of gene flow (Bank et al. 2012). Alternatively, when duplicated alleles are linked with other regions under constrains of selection, they could be also maintained during evolution. Taken together, the frequency and distribution of incompatible alleles at duplicated genes are dynamically affected by multiple factors including mating system, gene flow, and perhaps selection.

Although it remains unclear if incompatibilities in sympatry are common in other species as well, they have also been identified in Minulus species (Zuellig et al. 2018; Zuellig and Sweigart 2018). Using genotyping data from globally distributed individuals of other species such as rice where major resource have been generated (Huang et al. 2012; Wang et al. 2018) have the potential to reveal the frequency and distribution of incompatible duplicated alleles in these species in future (Mizuta et al. 2010; Yamagata et al. 2010; Nguyen et al. 2017).

Here, new gene copies arose either from distal duplications (in HPA and FOLT) or tandem duplications (in all four cases) and counteract incompatibilities. Even though additional copies reduce the frequency and impact of genetic incompatibilities, they could also increase the potential for more incompatible allele combinations. Subsequent subfunctionalization could be a way out of the trap of genetic incompatibilities, as there would be a selective pressure to keep both gene copies. Although we have assumed functional redundancy of all gene copies in this work, an extensive number of accessions shared the functional copies of both FOLT genes (865 of 1,135), which could indicate that these copies are not fully redundant and thereby limit the establishment of incompatible allele combinations. Thus, the incompatible alleles in sympatry may slow down neofunctionalization or pseudogenization of apparently redundant copies in multiple-copy carriers, and may partially explain why more duplicated genes are preserved than one would expect from theory (Lynch and Conery 2003; Maere et al. 2005).

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

**Materials and Methods**

**AMPRIL Construction**

The eight A. thaliana accessions An-1, C24, Col-0, Cvi-0, Eri-1, Kyo, Ler, and Sha were selected as the AMPRIL founders. We previously constructed a first version of AMPRIL population (AMPRIL I) including six RIL subpopulations (Huang et al. 2011). Here, AMPRIL I was extended with six additional subpopulations (called AMPRIL II) based on different pairwise intercrosses of the eight founders (fig. 1a). Each subpopulation contains approximately 90 individuals (supplementary data 1, Supplementary Material online). All plants were grown under the normal growth conditions in greenhouse at the MPI-PZ (Huang et al. 2011). DNA of 1,100 samples was extracted from the flower buds and prepared for RAD-seq sequencing (Baird et al. 2008).

**RAD-Seq Library Preparation and Sequencing**

All plants were grown in the greenhouse. The total DNA from each of the 1,100 samples was extracted from the flower buds using DNeasy Plant Kit 96 (Qiagen) and eluted in 200 μl Elution buffer (EB). DNA of each genotype was isolated twice. Both genotype samples were pooled in 1.5 ml tube and the DNA was concentrated by isopropanol precipitation for 2 h at −20 °C. The samples were centrifuged at 12,000 × g for 5 min at 4 °C, the supernatant was removed, and the DNA pellet was washed with ice-cold 70% ethanol. Centrifugation was repeated, the supernatant was removed. Air-dried DNA was resuspended in a nuclease free water to 26 ng/μl and stored at −20°C until use. RAD-seq sequencing libraries were prepared as described (Etter et al. 2011) with modifications. Per genotype, 500 ng DNA were digested with 10 units of CviQI (NEB, cutting site G’TAC) at 25 °C for 2 h. The number of expected cutting sites was estimated around 236,000 based on the Col-0 genome sequence. Cut DNA was purified using 96 DNA clean and concentrator kit (Zymo) diluted in 25 μl EB. The 192 different (selected out of total 210 designed, supplementary data 1, Supplementary Material online) P1 adapters (200 nM) containing unique 12-bp barcodes were ligated by incubation with T4 ligase (NEB) at room temperature for 30 min and the reaction was terminated by 20 min at 65 °C. After 30 min at room temperature, 5 μl from each 192 P1-barcoded sample were combined in a 2 ml low bind tube. About 3 × 130 μl aliquots were transferred to fresh tubes and DNA was fragmented to average size of 500 bp using Covaris. Sheared DNA was purified using QiaMinElute columns (Qiagen), eluted in 10 μl EB, and the three samples were first pooled and then divided into two 15 μl samples that were run on 1% agarose gel. Regions of 300- to 500-bp fragments were dissected and DNA was isolated using MinElute gel extraction kit (Qiagen), eluted in 10 μl EB, the samples were pooled, DNA fragment ends were repaired using Quick Blunting kit (NEB), purified with QiAquick column (Qiagen), and eluted in 43 μl EB. 3’-deoxy-adenine overhangs were added using Klenow Fragment (NEB), the sample was purified with QiAquick column, eluted in 45 μl EB, the P2 adapter was ligated, the sample was purified with QiAquick column and eluted in 53 μl EB. To determine the library quality, 10 μl RAD library were PCR amplified (1: 98 °C 30 s; 2: 14 × 98 °C 10 s, 67 °C 30 s, 68 °C 30 s; 3: 68 °C 5 min, 4: 4 °C hold) using NEB Next High-Fidelity master mix (NEB) in 25 μl reaction volume using RAD-Marker for/RAD-Marker rev primers (25 nM each) and 5 μl PCR product were loaded into 1%
Sequencing reads were demultiplexed according to the barcode sequenced in an Illumina HiSeq 2000 sequencing machine. Amplification was repeated in 50 µl (Beckman Coulter) and dissolved in 20 µl extraction kit (Qiagen), and eluted in 20 µl. DNA was isolated using MinElute Gel extraction kit (Qiagen), and sequenced next to the 1236 markers from a previous study; Jiao and Schneeberger 2020). For example, the parental haplotype of one progeny in the subpopulation EGGE can be found in supplementary table 3, Supplementary Material online. There are 88 individuals in this subpopulation. If the genotype at these two loci in one individual genotype was aabb, the count of the allele combination “ab” was increased by two. However, if the genotype was heterozygous, for example, “abbc,” the counts of the allele combinations “ab” and “bc” were both increased by 1. The P value of the $\chi^2$ test is 2.5e-09, suggesting a significant distortion of allele combinations.

Additionally, we applied a modified $\chi^2$ test (eq. 2) for all the duplicated genes in the whole AMPRIL population by considering the effects of population structure.

$$\chi^2 = \sum_{ij \in \{a, b, c, d, e, f, g, h\}} \frac{(\sum o_{ij} - \sum e_{ij})^2}{\sum e_{ij}}.$$  

Here, the $o_{ij}$ and $e_{ij}$ represent the observed and expected allele pair frequency of interchromosome duplicated genes, respectively, and $a, b, c, d, e, f, g,$ and $h$ represent all parental genotypes. The observed and expected allele pair frequencies in the whole population was the sum of observed (eq. 1) and expected (eq. 2) allele pair frequencies in each of the subpopulations.

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

Identification of LoF Variants in Candidate Genes
We mapped all whole-genome resequencing reads of 1,135 accessions from the 1001 Genomes Project (Alonso-Blanco et al. 2016), 75 accessions from Africa (Durvasula et al. 2017), and 118 accessions from China (Zou et al. 2017), to the Col-0 reference genome (TAIR10) (The Arabidopsis Genome Initiative 2000; Lamesch et al. 2012) using BWA (version 0.7.15) (Li and Durbin 2009) with the default parameter settings. SNPs and small indels were called using SAMtools (version 1.9) using default parameters (Li et al. 2009). Homozygous variants with mapping quality of more than 20 and with at least four reads aligned were kept. Pseudoheterozygous variants in HPA and TIM22 were also recorded. The large insertion and deletions in the 40-kb extended genic region of focal genes were predicted using Pindel (version 0.2.5) (Ye et al. 2009) with parameter settings “-T 1 -x 5 -k -r -j” and Delly (version 0.8.1) (Rausch et al. 2012) with parameter settings “delly call -q 20 -r 20 -n -u 20 –g.” The functional effects of these variations were annotated using SnpEff (version 4.3p) (Cingolani et al. 2012) using the default parameter settings. The LoF effects include loss of start codon, loss of stop codon, gain of premature stop codon, damage of splicing acceptor or donor sites, frameshift, and CDS loss.

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

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

Mapping Modifiers of Incompatible Alleles Using GWA
We predicted the presence and absence of functional copies of duplicated genes (HPA, TIM22) in each of 1,135 accessions from the 1001 Genomes Project according to the annotations of LoF variations and the clustering of cytosine methylation profiles in promoters (supplementary data 3–6, Supplementary Material online). For accessions without available methylation sequencing data, we assume the focal genes are expressed. The presence or absence of any functional copies of the reference genes were used as binary phenotype (presence: 1, absence: 0). For the association, we selected 238,166 high-quality SNP markers (minor allele frequency >0.05 and missing rate <0.1) from 1001 Genomes Project and imputed missing alleles with IMPUTE2 (Howie et al. 2009; Howie et al. 2012). An in-house R script (https://github.com/schneeebergerlab/AMPRIL-GI/blob/master/GWAS/gwas.LM.multiproxr) implementing the mixed linear model with correction of kinship bias was used to perform the GWA (Bonferroni correction, P < 0.05).

Copy Number Variation Analysis
To test for the existence of HPA3 in a genome, we mapped whole-genome short reads of 1,135 accessions from 1001 Genomes Project (Alonso-Blanco et al. 2016), of 75 accessions from Africa (Durvasula et al. 2017) and of 118 accessions from China (Zou et al. 2017) to the Kyogeno assembly (Jiao and Schneeberger 2020) using BWA (version 0.7.15) with default parameters. Accessions with an average mapping coverage <5 along the HPA3 region and duplicated region breakpoints (identified based on the sequence alignment against Col-0 genome using SyRI; Goel et al. 2019; version 1.0; with default parameters) were considered as HPA3 carriers. The copy number of duplicated genes was estimated by the ratio between the average mapping coverage within the focal gene and the average mapping coverage across the whole genome. To estimate the copy number of HPA, we mapped the short reads to the Cvi-0 genome assembly using BWA (version 0.7.15) with default parameters as the Cvi-0 genome only has one copy of HPA. For TAD3, FOLT, and TIM22, the copy number was predicted based on short reads mapping against the reference genome where both TAD3 and FOLT only have one copy. For the TIM22, copy number
was estimated based on the average mapping coverage at both TIM22-1 and TIM22-2.

Fine-Mapping of Incompatible Allele at LD2 in Cvi-0 × Col-0 RIL
After the observation that it was not possible to obtain a RIL homozygous for the Col-0 allele at LD2.1 while homozygous for the Cvi-0 allele at LD2.3 (Simon et al. 2008), we derived two distinct heterogeneous inbred families (HIF) from two segregating RILs from this population (supplementary fig. 2, two distinct heterogeneous inbred families (HIF) from two sequence.

Gradually, the causal interval is reduced and delineated by ants (at the seedling stage) for recombinants in the interval. To exclude intervals that do not contribute to the incompatible allele combination, so the families were used in each derived HIF family, it is again not possible to fix the incompatible allele combination, so the families were used to exclude intervals that do not contribute to the incompatible interaction. Two rounds of fine-mapping were conducted by genotypically screening increasing populations of descendants (at the seedling stage) for recombinants in the interval. Gradually, the causal interval is reduced and delineated by markers exploiting known SNPs and indels in the Cvi-0 sequence.

Segregation of T-DNA Mutant Line at Gene TIM22-1 and TIM22-2
T-DNA lines GABI_848H04 segregates for an insertion in TIM22-1, whereas GABI_626H10 segregates for an insertion in TIM22-2, both in a Col-0 background. Descendant was screened genotypically at the seedling stage to characterize the segregation of the T-DNA insertion allele.

Complementation Cross to Validate the Incompatible Alleles of TIM22
A LD2.3 HIF line (8HV408-Het) was crossed with a GABI_626H10 line, both at the heterozygous state, in order to segregate potentially for all four possible hybrid allelic combinations at LD2.3, while maintaining a fixed Col-0 allele at LD2.1. The F1 descendant (123 individuals, from both cross directions) was screened genotypically at the seedling stage to characterize the presence/absence of both the T-DNA insertion and the Cvi-0 allele.

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.

Supplementary Material
Supplementary data are available at Molecular Biology and Evolution online.

Acknowledgments
The authors would like to thank Detlef Weigel (Max Planck Institute for Developmental Biology, MPI-EB), Felix Bemm (MPI-EB), Todd Michael (Salk Institute), Joseph Ecker (Salk Institute), Florian Jupe (Salk Institute) for releasing the 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 Germanys Excellence Strategy—EXC 2048/1—390686111 (K.S.), and the European Research Council (ERC) Grant “INTERACT” (802629) (K.S.).

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germanys Excellence Strategy—EXC 2048/1—390686111 (K.S.), and the European Research Council (ERC) Grant “INTERACT” (802629) (K.S.).

Author Contributions
W.-B.J. and K.S. designed this project. W.-B.J., V.P., J.K., and F.L. analyzed the data. P.P. and A.P. generated the RAD-seq data. O.L., M.F., I.G., and C.C. performed the mapping and complementation experiments of LD2/TIM22. S.E. and M.K. generated the AMPRIL population. W.-B.J. and K.S. wrote the article.

Data Availability
Raw RAD-seq data of the AMPRIL population were 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) were 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/.

References


