Plant Cell Advance Publication. Published on March 18, 2019, doi:10.1105/tpc.18.00791

RESEARCH ARTICLE

Adaptation and Phenotypic Diversification in Arabidopsis through Loss-of-function Mutations in Protein-coding Genes

Yong-Chao Xu[1,2], Xiao-Min Niu[1,2], Xin-Xin Li[1,2], Wenrong He[3], Jia-Fu Chen[1,2], Yu-Pan Zou[1,2], Qiong Wu[1], Yong E. Zhang[2,4], Wolfgang Busch[3], Ya-Long Guo[1,2*]

[1] State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China

[2] University of Chinese Academy of Sciences, Beijing 100049, China

[3] Salk Institute for Biological Studies, Plant Molecular and Cellular Biology Laboratory, La Jolla, CA 92037, USA

[4] State Key Laboratory of Integrated Management of Pest Insects and Rodents & Key Laboratory of the Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing, China

* Corresponding Author: yalong.guo@ibcas.ac.cn

Short title: Adaptation via loss-of-function mutations

One-sentence summary: A large-scale study exploring how gene loss contributes to evolutionary change reveals that loss-of-function mutations contribute to the adaptation and phenotypic diversification of plants.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) is: yalong.guo@ibcas.ac.cn.

ABSTRACT

According to the “less-is-more” hypothesis, gene loss is an engine for evolutionary change. Loss-of-function (LoF) mutations resulting in the natural knockout of protein-

coding genes not only provide information about gene function, but they also play important roles in adaptation and phenotypic diversification. Although the “less-is-more” hypothesis was proposed two decades ago, it remains to be explored on a large scale. In this study, we identified 60,819 LoF variants in 1,071 *Arabidopsis thaliana* genomes and found that 34% of *A. thaliana* protein-coding genes annotated in the Col-0 genome do not have any LoF variants. We found that nucleotide diversity, transposable element density, and gene family size are strongly correlated with the presence of LoF variants. Intriguingly, 0.9% of LoF variants with minor allele frequency larger than 0.5% are associated with climate change. In addition, in the Yangtze River basin population, 1% of genes with LoF mutations were under positive selection, providing important insights into the contribution of LoF mutations to adaptation. In particular, our results demonstrate that LoF mutations shape diverse phenotypic traits. Overall, our results highlight the importance of the LoF variants for the adaptation and phenotypic diversification of plants.

INTRODUCTION

Variation in gene copy number plays important roles in adaptation and diversification. The two major processes that result in the gain of genes or gene copies are *de novo* gene formation and gene duplication (Chen et al., 2010; Carvunis et al., 2012; Guo, 2013; Palmieri et al., 2014; Zhao et al., 2014; McLysaght and Guerzoni, 2015; Li et al., 2016). However, the “less-is-more” hypothesis proposes that gene loss also contributes to evolutionary change in the context of gene copy number variation (Olson, 1999). The dosage balance hypothesis suggests that altering the stoichiometry of members of macromolecular complexes can affect the function of the whole complex and could ultimately affect phenotype and evolutionary fitness (Edger and Pires, 2009; Birchler and Veitia, 2012; Hao et al., 2018). Therefore, the dosage balance hypothesis provides a more compelling explanation of the advantages and disadvantages of removing gene copies from a genome and explains the “less-is-more” hypothesis in a mechanistic way. There are two major mechanisms underlying gene loss: the physical removal of a gene by recombination, and gene inactivation by loss-of-function (LoF) mutations (Albalat and Canestro, 2016). The latter mechanism can occur through the gain of a premature stop codon, splice site disruption, or disruption of a transcript reading

frame. LoF mutations are frequent and have recently gained attention owing to the improved detection power of advanced sequencing techniques (MacArthur et al., 2012; Narasimhan et al., 2016).

Consistent with the “less-is-more” hypothesis, many cases of gene loss have been reported in diverse organisms, including unicellular organisms such as bacteria (Will et al., 2010; Hottes et al., 2013), and multicellular organisms, such as plants (Zufall and Rausher, 2004; Hoballah et al., 2007; Song et al., 2007; Tang et al., 2007; Gujas et al., 2012; Amrad et al., 2016; Sas et al., 2016; Wu et al., 2017) and animals (Greenberg et al., 2003; Hodgson et al., 2014; Goldman-Huertas et al., 2015), including humans (MacArthur et al., 2012; Narasimhan et al., 2016). LoF mutations have been shown to affect important biological processes, such as development and stress resistance in plants (Gujas et al., 2012; Wu et al., 2017), and intellectual disabilities in humans (Green et al., 2017). However, the extent to which LoF mutations are correlated with adaptation and phenotypic diversification is largely unknown (Albalat and Canestro, 2016). Even though loss of function mutations with produce large and detrimental effects on fitness will be depleted in inbreeding species compared to outcrossing plant and animal species, inbreeders are an excellent system to study the effect of natural homozygous knockouts, as inbreeding leads to a high frequency of homozygous LoF mutations (Saleheen et al., 2017). The best-characterized selfing species is *Arabidopsis thaliana*, which is naturally occurring in almost all parts of the world.

In this study, to investigate the evolutionary patterns of LoF mutations in natural populations, we explored 1,071 genomes of *A. thaliana* from the 1001 Genomes Project (The 1001 Genomes Consortium, 2016), the African genomes project (Durvasula et al., 2017), and our own sequencing project (Zou et al., 2017). Across these 1,071 Arabidopsis genomes, we identified 60,819 LoF variants, including 17,453 stop codon-introducing (stop-gain) variants, 37,935 disruptions of a transcript reading frame (frameshift), and 5,431 splice site-disrupting single nucleotide variants (splice site). We found that the presence of LoF mutations is correlated with the level of nucleotide diversity, the density of transposable elements (TEs), and gene family size. Intriguingly, 1% of genes with LoF mutations are under positive selection in the Yangtze River basin

population, suggesting that these genes with LoF mutations are crucial for adaptation. Overall, this study highlights the importance of LoF mutations for adaptation and phenotypic diversification.

RESULTS

The Presence of LoF Variants is Correlated with Nucleotide Diversity, TE density, and Gene Family Size

We scanned the genomes of 1,071 accessions for LoF variants (Figure 1A and 1B), including 893 genomes downloaded from the 1001 Genomes Project (Supplemental Data Set 1) (The 1001 Genomes Consortium, 2016), 61 from the African genomes project (Supplemental Data Set 2) (Durvasula et al., 2017), and 117 from our own genome project (Supplemental Data Set 3) (Zou et al., 2017). Given that LoF mutations annotated by mapping sequences to a reference genome will have high false positive rates, a series of stringent filtering steps were implemented, which were similar to those employed in previous studies of human genomes (MacArthur et al., 2012; Narasimhan et al., 2016). First, single nucleotide variants (SNVs) and indels (insertions/deletions) were removed if they matched multiple times to the reference genome or were located in short tandem repeats or if the SNVs were in close proximity (3 bp or less) to an indel. Second, stop-gain variants were excluded if they were within a codon that was linked to other SNVs, and frameshift variants were removed if they were occurring together with other frameshifts that resulted in the restoration of the reading frame. Third, SNVs and indels were removed if the inferred LoF mutations occurred within the last 5% of the transcript or were observed in the ancestral states inferred from *Arabidopsis lyrata* and *Capsella rubella* (see Methods). Fourth, LoF variants that affect all known transcripts of the protein-coding gene in the reference genome (Col-0) were retained for subsequent analysis. After filtering according to these rules, we detected 17,453 (214.7 per accession) premature stop codons that were caused by SNVs (stop-gain), 37,935 (512.2 per accession) insertion/deletion-induced frameshift variants (frameshift) leading to the disruption of a transcript reading frame, and 5,431 (103.0) splice site-disrupting SNVs (splice site) (Table 1). Taking all of these stop-gain, frameshift,

and splice site variants together, we identified 60,819 (829.9 per accession) LoF variants across the 1,071 accessions (Table 1).

To assess the power of our detection method and to estimate the false discovery rate (FDR) for LoF variants, we used RNA sequencing data for eight *de-novo* assembled genomes (Gan et al., 2011), the Landsberg erecta genome assembled with PacBio Sequel data (Zapata et al., 2016), and the KBS-Mac-74 genome assembled with nanopore data (Michael et al., 2018). The results suggest that our method has a very low FDR (2.6% in each accession on average) (Supplemental Data Set 4). Given that the accumulated FDR across 1,071 accessions might be higher than that in a single genome, we calculated the accumulated false positive rates based on validation data in these 10 accessions. While with the addition of genomes to the analysis, the accumulated false positive LoF as well as the total number of LoF increased, the accumulated FDR was largely stable (Supplemental Figure 1). Therefore, we estimated that in each *A. thaliana* accession, an average of 808.2 genes have LoF mutations (3.0% of all protein-coding genes in the reference Col-0) with a FDR of 2.6%. In particular, 80 false positive LoF mutations (22 stop-gains, 56 frameshifts, and 2 splice sites) identified in the validation process were excluded from the subsequent analysis.

The density of genes with stop-gain, frameshift, splice site, and LoF variants (summed up for all the stop-gain, frameshift, and splice site variants), as calculated by dividing the number of genes with LoF variants (genes with LoF variants detected in any of the 1,071 accessions) by the gene number in each 200 kb window in the Col-0 reference genome, varied across the genome (Figure 1C, Supplemental Figure 2). By contrast, the density of stop-gain and frameshift variants within gene bodies was roughly evenly distributed, with a slightly higher frequency at the beginning and at the end of genes (Supplemental Figure 3). This pattern is similar to the distribution in human and great ape genomes (MacArthur et al., 2012; de Valles-Ibanez et al., 2016).

To infer mechanisms that favor LoF, we analyzed the correlations between LoF and various gene features. The density of genes with LoF mutations (stop-gain, frameshift, splice site and LoF variants) across the Col-0 genome showed a strong correlation with nucleotide diversity (π) across the 1,071 genomes and

the TE density in the Col-0 genome (Figure 1D). Large gene families tend to have high gene redundancy (Wagner, 2005), and gene loss in large gene families may be an advantage according to the dosage balance hypothesis (Edger and Pires, 2009; Birchler and Veitia, 2012; Hao et al., 2018). We therefore evaluated whether these genes were prone to acquiring LoF mutations. We classified the 27,206 protein-coding genes of the reference Col-0 genome into 7,430 gene families based on amino acid sequence similarity. We then estimated the correlation between gene family size and the proportion of genes with common LoF variants (variants displaying a minor allele frequency [MAF] larger than 5% across the 1,071 accessions). For all gene families, there was a positive correlation between gene family size and the LoF ratio (the percentage of genes with LoF variants in each gene family) (Spearman $r = 0.21$, $p = 2.2 \times 10^{-16}$). This suggests that genes belonging to larger gene families are more likely to have LoF variants. Similarly, the median sizes of gene families for genes with and without LoF variants were 11 and 6, respectively (Figure 2A, Wilcoxon sum test, $p < 2.2 \times 10^{-16}$). This indicates that genes belonging to larger gene families are more likely to acquire LoF mutations. Furthermore, analogous to previous studies (Prachumwat and Li, 2008; Guo, 2013), we divided gene families into three classes: one gene or singleton (class 1: 3,825 families), two genes (class 2: 1,353 families), and three genes or more (class 3: 2,252 families). The genes in class 3 were more likely to contain LoF variants than the others (Supplemental Figure 4, class 3 vs. class 1 or 2, Wilcoxon sum test, all $p < 0.01$). Therefore, the presence of LoF variants is correlated with nucleotide diversity, TE density, and gene family size.

A Large Fraction of *A. thaliana* Genes could Affect Fitness in Natural Environments

A genome-wide study of LoF variants can provide insight into gene lethality by analyzing the frequency of all null alleles for a given gene in natural environments (population gene lethality) (Albalat and Canestro, 2016). We found that 9,249 protein-coding genes (34.0% of all protein-coding genes in the Col-0 reference genome) do not have any LoF variant within our panel of 1,071 accessions. To ask whether known essential genes are depleted of LoF variants, we used a database of 358 lethal genes identified in laboratory studies (Meinke

et al., 2008). 88.3% of the lethal genes do not have any natural LoF mutation, and 7.3% of the lethal genes display a LoF allele frequency less than 0.5% across the 1,071 accessions (Figure 2B). This suggests that lethal genes in *A. thaliana* are significantly depleted of natural LoF mutations (Pearson's chi-squared test, $p < 2.2 \times 10^{-16}$). We then performed gene ontology (GO) enrichment analysis of the four different gene classes based on the LoF allele frequency in the 1,071 accessions (genes without LoF, with LoF, with a LoF allele frequency of <0.05 and ≥0.05). Although both the gene sets (with and without LoF variants) were enriched for the GO term 'unknown function', the only gene set highly enriched for many other GO terms was the set without LoF variants (Figure 2C). These results suggest a model in which natural knockouts of 9,249 *A. thaliana* genes without LoF variants would reduce fitness in natural environments, and individuals with LoF mutations in these non-essential genes can likely be removed from populations by purifying selection.

LoF Variants are Correlated with Climate Variables

LoF variants could be functionally important and associated with climate variables. To investigate this notion, we performed a genome-wide association study (GWAS) of 14,270 LoF variants with MAF larger than 0.5% and 20 environmental variables, including latitude and 19 climate variables (Supplemental Data Set 5). To validate that the associations were not random, we compared the total number of significantly associated LoF variants with the number obtained from 1,000 GWAS permutations based on 14,270 randomly resampled SNPs (single-nucleotide polymorphisms; MAF > 0.5%) in the coding regions of protein-coding genes across the whole genome for each of the 20 environmental variables. If LoF variants were randomly associated with environmental variables, then we would expect the number of associated variants obtained in the permutation analysis and the observed value for each environmental variable to be similar. For 10 of the 20 environmental variables, the observed number was significantly higher than the number identified by permutation analysis (Supplemental Figure 5, one-sample Wilcoxon signed rank test, $p < 0.05$), suggesting that LoF variants are more prone to associate with climate variables. 125 of the 14,270 LoF variants (0.9%) were significantly associated with at least one of the 10 environmental variables, involving 124

184 genes (Supplemental Figures 6-8 and Supplemental Data Set 6). Of these
185 genes, some are known to be functionally important, such as *CRK36*
186 (Supplemental Data Set 6). *CRK36* encodes an abiotic stress-inducible receptor-
187 like protein kinases that negatively regulates abscisic acid signaling (Tanaka et
188 al., 2012).

189 We used RNA-seq data for 541 accessions to compare the expression levels
190 of alleles with and without LoF mutations for each of these environmentally
191 associated genes (Kawakatsu et al., 2016) (Figure 2D). Among the 124
192 associated genes, 49 genes have both LoF and non-LoF alleles in the 541
193 accessions for which RNA-seq data were available. There were no LoF variants
194 of the other 75 genes in this set of accessions. We therefore focused on the 49
195 genes with both LoF and non-LoF alleles (different LoF variants of the same
196 gene were combined). The transcript levels of 16.4% of these genes were
197 significantly altered (8.2% upregulated and 8.2% downregulated) in accessions
198 harboring the LoF variants (Wilcoxon sum test, FDR corrected $p < 0.05$) (Figure
199 2D). Furthermore, genes with LoF variants associated with environmental
200 variables exhibited greater rates of upregulation (8.2% vs. 3.0%) or
201 downregulation (8.2% vs. 6.5%) than genes with LoF variants at the genome
202 level (Figure 2E). To determine whether the expression changes were caused by
203 the different genetic backgrounds of various natural accessions, we calculated
204 the up- and downregulated gene numbers based on 30,000 randomly selected
205 alleles with non-LoF variants in coding regions across the 541 accessions. A
206 higher proportion of genes with LoF variants associated with environmental
207 variables are up- (8.2% vs. 1.7%) or downregulated (8.2% vs. 2.4%) compared
208 to those detected when only the genetic background was taken into account
209 (Supplemental Figure 9). The higher proportion of upregulated genes with LoF
210 variants associated with environmental variables could result from diverse
211 mechanisms, such as a dominant gain-of-function effect, in which LoF mutations
212 are effectively gain-of-function mutations, thereby mimicking increased gene
213 function (Schild et al., 1995; Xie et al., 1998).

214 **LoF Variants of *KUK*, *PRR5*, and *LAZY1* Shape Phenotypic Variation**

215 Common LoF alleles are nearly always less deleterious than rare LoF alleles,
216 but some could have an impact on phenotypes (MacArthur et al., 2012).

Considering that stop-gain mutations can exert severe effects by directly truncating genes, we performed an additional GWAS analysis for stop-gain mutations and the 10 environmental variables that are significantly associated with LoF variants (Supplemental Figure 10). We found that many genes with stop-gain mutations are associated with environmental variables (Supplemental Data Set 7). One gene with a significant GWAS signal for a precipitation-related trait (precipitation in the wettest month) is the F-box protein gene *KUK* (Figure 2F, Supplemental Figures 10 and 11), which is involved in regulating root development in Arabidopsis (Meijon et al., 2014). *KUK* alleles with and without LoF variants are widely distributed across Eurasia, and most accessions from the Yangtze River basin carry stop-gain variants (Supplemental Figure 12A). Based on the phenotypic data for 129 accessions available from a previous study (Figure 3A and Supplemental Data Set 8) (Meijon et al., 2014), the meristematic zone and mature cortical cells of accessions with LoF variants (stop-gain, frameshift, or total LoF; no putative splice variation was observed) were significantly longer than those of accessions without LoF variants (Figure 3B and Supplemental File 1; population structure-corrected ANOVA, $p < 0.05$).

To independently test the hypothesis that truncated versions of *KUK* are associated with increased meristem and mature cell sizes, we phenotyped accessions containing specific LoF alleles that were not evaluated in the previous study (with one exception: Bur-0) (Meijon et al., 2014). We chose three accessions that did not contain a stop codon (control), three accessions with the most common stop codon located in the first half of *KUK* (111 G/A: TGG-TGA, SG111), the only two accessions in the 1001 Genomes panel that had a distal stop codon (652 C/T: CGA-TGA, SG652), including one of the accessions with the largest meristem and mature cell traits (Bur-0) according to a previous study (Meijon et al., 2014), and two accessions with a distal frameshift (738 T/-, FS738) (Supplemental Figure 13 and Supplemental Data Set 9). Compared with the control, all three classes showed significant increases (pairwise Wilcoxon rank sum test, $p < 0.05$) in meristem length, with the most common stop codon associated with the smallest increase (Supplemental Figure 13 and Supplemental Data Set 9). Overall, these data suggest that truncated alleles of *KUK* influence cellular traits, most likely via a dominant mechanism. Further

transgenic studies should be performed to confirm the effects of these genetic variants in the same genetic background.

In addition to the LoF alleles of genes that were associated with environmental variables based on the GWAS, we studied *PSEUDO-RESPONSE REGULATOR5* (*PRR5*), which is involved in regulating various circadian-associated biological events, such as flowering time (Nakamichi et al., 2016). *PRR5* alleles with and without LoF variants are widely distributed in our set of 1,071 accessions (Supplemental Figure 12B). Of the 812 accessions with flowering time data (The 1001 Genomes Consortium, 2016) (Figure 3C and Supplemental Data Set 10), the flowering time of accessions with *PRR5* LoF variants (frameshift, or total LoF; no putative splice variation was observed) was significantly delayed compared to accessions without LoF variants at either 10°C or 16°C (Figure 3D and Supplemental File 1; population structure-corrected ANOVA, $p < 0.05$). We note that while flowering time for the stop-gain allele was delayed relative to the non-LoF allele, the difference was not significant, probably owing to the small number of accessions.

Given the above two cases of genes having common LoF variants, we further studied *LAZY1* as an example for a rare LoF variant. Only one accession from the Yangtze River basin population (29-8) showed a frameshift mutation in *LAZY1*, which that was caused by a 20 bp deletion in exon 3 and resulted in a premature stop codon (Figure 4D and Supplemental Figure 14). *LAZY1* mutant plants display a much larger branch angle (81°) than wild type (42°) (Yoshihara et al., 2013). As expected, we found that accession 29-8 has a large branch angle (74°) (Figure 4A). To validate that this frameshift mutation was the causal mutation, we crossed accession 29-8 with accessions displaying smaller branch angles: Col-0 (30°) and 3-2 (30°). In the Col-0 × 29-8 F_2 population, the segregation ratio for branch angles was roughly 3:1, suggesting that a major, recessively acting gene was responsible for the large branch angle of 29-8 (Figure 4B). Using a mapping-by-sequencing approach based on 57 F_2 plants from the Col-0 × 29-8 F_2 population with large branch angles (>85°), we identified a causal region on chromosome 5 (Figure 4C). To refine the target interval, we used F_2 plants in both Col-0 × 29-8 and 3-2 × 29-8 populations, and finally narrowed down the causal locus to a 240 kb interval containing *LAZY1*

(Figure 4D, Supplemental Data Set 11 and 12). The truncated transcript of *LAZY1* in the 29-8 accession caused the loss of the only functional domain of the nuclear localization signal (NLS), which is located between conserved regions 3 and 4. This NLS is a key component during the nuclear import process (Supplemental Figure 14) (Yoshihara et al., 2013). Overall, these results provide evidence that the frameshift variant of *LAZY1* is likely responsible for the larger branch angle in accession 29-8.

A subset of LoF variants exhibits a signature of natural selection

Compared with other SNP variants across the whole genome in the 1,071 accessions, LoF variants are biased towards low frequencies (Figure 5A). This suggests that many natural LoF variants might be deleterious in Arabidopsis and are under purifying selection. Similarly, compared with 1,000 permutation results of non-LoF variants associated with environmental variables (GWAS non-LoF variants), LoF variants associated with environmental variables (GWAS LoF variants) were biased towards low allele frequencies (Figure 5A). This suggests that a large fraction of LoF variants that are associated with environmental variables are likely under purifying selection as well.

Although LoF mutations are usually deleterious, adaptive LoF mutations can occur and spread rapidly in small populations (Olson, 1999). To determine whether genes with LoF mutations were influenced by natural selection, we focused on the Yangtze River basin population (86 accessions, Supplemental Data Set 3), a population that recently adapted to this region (Zou et al., 2017). In total, 2,709 genes had LoF variants (3,349 LoF variants) in this population; among them, 54 genes (64 LoF variants) were under positive selection based on a selection sweep analysis in our recent study (Zou et al., 2017) (Figure 5B, see Supplemental Data Set 13 and 14 for details). Of the 54 genes in these selected regions, 27 genes (1.0% of the genes with LoF variants) had a LoF allele frequency exceeding 95% and were more likely under positive selection, including *KUK*, in which one LoF allele (111 G/A: TGG-TGA type, Supplemental Figure 11) was fixed in the Yangtze River basin population (Supplemental Data Set 14).

Finally, we performed ecological niche modeling of non-LoF (571) and LoF (500) alleles of *KUK*. In niche identity tests, the simulated values were significantly higher than the observed value (Io = 0.877, p < 0.001), indicating that there is significant ecological differentiation between the two groups (Figure 5C, Supplemental Figure 15). This result suggests that the LoF allele of *KUK* is associated with adaptation to the Yangtze River basin. Overall, our results indicate that LoF mutations can be associated with adaptation and phenotypic diversification; more importantly, at least 1.0% of the genes with LoF variants are under positive selection in the Yangtze River basin.

DISCUSSION

The “less-is-more” hypothesis proposes an evolutionary process involving adaptive gene loss (Olson, 1999). LoF mutations have gained increasing attention (MacArthur et al., 2012; Yang et al., 2015; Narasimhan et al., 2016). For example, recent studies have demonstrated the functional importance of natural LoF mutations (Gujas et al., 2012; Lek et al., 2016; Saleheen et al., 2017; Wu et al., 2017). However, our understanding of the evolutionary pattern of LoF mutations at the genome level and the adaptive effects of LoF variants at the population level is highly limited. In this study, we investigated the evolutionary pattern of LoF mutations in *A. thaliana* and found that a high level of nucleotide diversity, high TE density, and high gene redundancy (large gene family size) are associated with LoF mutations. The latter two factors were also observed in a previous study of LoF mutations in human populations (MacArthur et al., 2012). Given that these three factors themselves are largely correlated with rates of evolution, rapid sequence evolution could be associated with a high frequency of LoF mutations. More importantly, we hypothesize that 34% of *A. thaliana* genes can affect fitness in natural environments, and furthermore, 1% of genes with LoF variants are under positive selection in the Yangtze River basin population.

The mechanisms by which LoF mutations can produce functional effects are complicated. Some of the profound effects of LoF can be explained by the dosage balance hypothesis (Birchler and Veitia, 2007; Hou et al., 2018; Kremling

et al., 2018). A simple mechanism for a beneficial effect of a null mutation is the removal of a protein that is detrimental in the current environment (Hottes et al., 2013). For example, *brx* (*brevis radix*) LoF alleles in Arabidopsis help roots adapt to acidic soil (Gujas et al., 2012). However, it is also possible that LoF variants can act as dominant mutations. For instance, stop-gain variants of the SOX9 transcription factor MiniSOX9 act in a dominant-negative manner, thereby counteracting the activity of the wild-type variant (Abdel-Samad et al., 2011). Other stop-gain and missense variants can act as dominant gain-of-function mutations, thereby mimicking increased gene function (Schild et al., 1995; Xie et al., 1998). In the case of *KUK*, a gene that acts as a positive regulator of meristem and cell size (Meijon et al., 2014), a hypermorphic gain-of-function is a more likely explanation than a LoF. Although most premature stop codons result in a LoF or dominant effects, more complex possibilities exist. For instance, in the fruit fly *Drosophila sechellia*, an olfactory receptor pseudogene encodes a functional receptor as a result of translation read-through of the premature termination codon (Prieto-Godino et al., 2016).

Of the many LoF variants (or pseudogenes) that we identified in over 1,000 natural accessions, some are presumably functional, with a minor or altered function, rather than non-functional. More efforts are needed to understand their functional effects. For example, transgenic analyses are needed to clarify the functional differences between the predicted pseudogenes and their ancestral genes. While more function research is needed, LoF mutational variation also provides a way of investigating the phenotypic consequences of the loss of specific genes within diverse genomes while at the same time establishing their role in the evolutionary process. Overall, our study highlights the importance of natural knockouts for adaptation and phenotypic diversification and emphasizes the need for more in-depth studies of LoF variants.

METHODS

Data analyzed in this study

The genomes of 1,071 accessions were used, including 893 representative genomes of *A. thaliana* from the 1001 Genomes Project (The 1001 Genomes

Consortium, 2016) (Supplemental Data Set 1), 61 from the African sequenced genomes project (Durvasula et al., 2017) (Supplemental Data Set 2), and 117 genomes from our own sequencing project (Zou et al., 2017) (Supplemental Data Set 3). The 541 transcriptomes used in this study were obtained from a previous study, as indicated in Supplemental Data Set 1 (Kawakatsu et al., 2016). All 1,071 accessions with clean data were mapped against the Col-0 reference genome, and SNPs and indels were called using Genome Analysis Toolkit (GATK version 2.1.8) with default parameters (DePristo et al., 2011).

Identification of candidate LoF variants

High-quality homozygous SNPs and indels (quality (Q) ≥ 30, quality-by-depth ratio (QD) ≥ 10 (QD ≥ 5 for indels), ReadPosRankSum ≥ -8.0, depth of coverage (DP) ≥ 3, probability of strand bias (FS) ≤ 10.0 (FS ≤ 200.0 for indels)) were used to identify LoF mutations using SnpEff software (SnpEff 3.3f) (McLaren et al., 2010). SNPs annotated as STOP GAINED, SPLICE SITE DONOR, and SPLICE SITE ACCEPTOR (SPLICE SITE DONOR and SPLICE SITE ACCEPTOR were combined as splice site), and indels annotated as FRAMESHIFT, were regarded as LoF variants and included in subsequent analyses.

Filtering of candidate LoF variants

To remove false positive LoF variants caused by sequencing and mapping errors, annotation errors, and the effect of nearby variants, a series of stringent filters were used. First, a sequence context filter was applied. Variants that could be mapped to multiple regions of the reference genome (Col-0) were removed using Genome Multitool (Derrien et al., 2012; Marco-Sola et al., 2012). SNVs (stop-gain and splice site) and indels (frameshift) overlapping with tandem repeats and SNVs in close proximity (3 bp or less) to a known indel were excluded. Second, multi-nucleotide polymorphism filters were applied. To filter incorrect premature stop codon annotation, stop-gain variants that were found within the same codon linked to other SNVs were removed. Frameshift variants were filtered out if two frameshift variants resulted in the restoration of the reading frame. Three or more frameshift variants occurring in each gene of the same accession were excluded; most resulted in the restoration of the reading frame. Third, annotation filters were applied. SNVs and indels were filtered out if

410 the inferred LoF allele was also the ancestral state, as this implies that such
411 variants were “gain-of-function” or the sequencing errors of the gene model at
412 this location in the reference.

413 Two related species, *Arabidopsis lyrata* (MN47) and *Capsella rubella* (MTE),
414 were used to infer the ancestral state of SNVs and indels. For SNVs and indels,
415 ancestral states were determined using alignments of Col-0 with the *A. lyrata*
416 and *C. rubella* reference genomes, as described in our previous study (Li et al.,
417 2016). SNVs and indels were filtered out if either of the two outgroups showed
418 the same stop-gain or frameshift mutations observed in the 1,071 accessions.

419 The disrupted fraction of the coding sequence of the longest transcript caused
420 by stop-gain and frameshift variants was calculated, and stop-gain or frameshift
421 mutations occurring within the last 5% of the transcript were removed. Because
422 we did not determine the position of splice site disruption on the final transcript
423 (MacArthur et al., 2012), we were not able to perform this analysis for splice site
424 variation. Finally, if there were two or more LoF variants in the same allele, only
425 the LoF variant closest to the start codon (ATG) was considered.

426 **Validation of LoF mutations**

427 We estimated the FDR using the assembled transcripts of eight accessions from
428 a previous study (Gan et al., 2011), as well as gene sequences that were
429 annotated using two published long reads assembled genomes: Landsberg
430 erecta assembled from PacBio Sequel data (Zapata et al., 2016) and KBS-Mac-
431 74 assembled from nanopore data (Michael et al., 2018). The genes of KBS-
432 Mac-74 and Landsberg erecta were annotated using exonerate v2.2.0 (Slater
433 and Birney, 2005) with CDS (coding sequences) of Col-0. All assembled
434 transcript sequences for genes with LoF mutations in the eight accessions, and
435 annotated gene sequences of KBS-Mac-74, Landsberg erecta, and Col-0 gene
436 sequences were aligned using MUSCLE v3.8.31 (Edgar, 2004). All LoF
437 mutations were manually checked based on the aligned sequences. Stop-gain
438 and frameshift variants of the 10 accessions were used to assess the curve of
439 accumulated false positive rate (all false positive LoF variants found in each
440 accession combined). Splice site variants were not used in the analysis of
441 accumulated false positive rate, because assembled transcripts of eight

accessions were assembled based on short reads, which are not robust enough to validate splice sites. Nevertheless, the FDRs for splice sites in the two long reads assembled genomes are similar or even lower than either stop-gain or frameshift. The accumulated false positive LoF number, accumulated total LoF number (the number of LoF variants with transcriptional or long reads data in each accession), and accumulated false positive LoF ratio were calculated with all probable accumulated groups.

LoF polymorphism variation

All parameters were calculated using a 200 kb window size and 10 kb step size along the Col-0 reference genome. The density of genes with stop-gain variants was calculated by dividing the number of genes with stop-gains by the number of genes in each 200 kb window in the Col-0 reference genome. The same rationale was used to calculate the density of frameshift, splice site, and total LoF variants (summed over all the three LoF variants). Nucleotide diversity (π) was calculated based on the SNP matrix of 1,071 accessions in each 200 kb window. The GC content was calculated based on the Col-0 genome in each 200 kb window. Exon number, CDS length, and gene expression level were calculated for each gene based on the Col-0 genome in each 200 kb window. Gene expression levels in Col-0 were obtained from a previous study (Kawakatsu et al., 2016). TE density was calculated as the number of TEs in each 200 kb window in the Col-0 reference genome with TE annotations downloaded from TAIR10.

Gene family, sequence, and expression analysis

Gene family analysis was performed according to our previous study (Guo, 2013). GO term enrichment analysis was performed using agriGO (Tian et al., 2017). All genes (5,050) with candidate LoF variants were removed from the non-LoF category in the GO analysis. PCR was performed to obtain the *LAZY1* upstream region (~1.5 kb), gene body, and downstream regions (~0.5 kb) from Col-0, 3-2, and 29-8 (marked with an asterisk in Supplemental Data Set 3) using Q5 polymerase (New England Biolabs, Ipswich, MA, USA). Sequences were aligned using Lasergene Seqman (DNASTAR, Madison, WI, USA). Markers used in the analysis of plants with large branch angles are listed in Supplemental

Data Set 11 and 12. Primers used for the sequencing and amplification are listed in Supplemental Data Set 15.

A total of 541 transcriptomes were available from the 1001 Genomes Project (Supplemental Data Set 1) (Kawakatsu et al., 2016), and 6,675 genes with LoF variants whose mean expression levels (FPKM) were larger than 3 across 541 accessions were kept for subsequent analysis (Meng et al., 2016). For each of the 6,675 genes with LoF variants, 541 alleles were divided into two groups: alleles with and without LoF variants. Then, the Wilcoxon sum test was used to evaluate the expression data. All *p* values were adjusted for multiple testing by computing their FDR (Benjamini and Hochberg, 1995). Up- or downregulation was calculated using the mean expression data in the two groups (alleles with and without LoF variants).

Genome-wide association study

GWAS was performed by using the compressed mixed linear model (Zhang et al., 2010) from the GAPIT R package (Lipka et al., 2012) to associate data for 20 environmental variables (latitude and 19 climate variables from the WORLDCLIM database) (Supplemental Data Set 5) with the 14,270 LoF variants (MAF > 0.5%) of the 1,071 accessions. Principal components ("Q" matrix) and a kinship matrix ("K" matrix) were included to account for population structure. Bayesian information criterion-based model selection as implemented in the GAPIT R package was used to find the optimal number of principal components for each trait. The whole-genome significance cutoff for associations was set to 0.01/total LoF variants ($-\log_{10}(p) = 6.12$). For the GWAS permutation analysis, the SNP matrix that contained 14,270 SNPs (MAF > 0.5%) was randomly resampled 1,000 times for those SNPs in the coding region of protein-coding genes across the whole genome. Whether the actual value is larger than the permutation results was tested using the one-sample Wilcoxon signed rank test. GWAS analysis was also performed using 3,822 stop-gain variants (MAF >0.5%) with these 20 environmental variables within the 1,071 accessions. The significance cutoff in this analysis was set to 0.05/total LoF variants ($-\log_{10}(p) = 4.88$).

Ecological niche modeling analysis

MaxEnt 3.3.3 was used to perform ecological niche modeling via maximum entropy with default settings (Phillips et al., 2006) to examine ecological divergence between the two *KUK* groups (excluding redundant and unreliable records) combined with data for 19 ecological variables downloaded from the WORLDCLIM database (Phillips and Dudik, 2008). For the analysis, default settings of MaxEnt 3.3.3 were used. Ten of the environmental variables with pairwise Pearson correlation coefficients of $-0.7 < r < 0.7$ were selected for final analysis (marked red in Supplemental Data Set 5) with outputs set at 25% for testing and 75% for training the model. The model accuracy was assessed based on the area under the curve (AUC), with scores between 0.7 and 0.9 indicating good discrimination, and AUC > 0.9 indicating reliable discrimination (Swets, 1988). The range of suitable distributions was drawn using DIVA-GIS v7.5 (Hijmans et al., 2005).

ENMTools 1.3 was used to perform niche identity test by calculating Warren's *I*, which ranged from 0 (no niche overlap) to 1 (identical niches), with 100 pseudo-replicates, between the LoF and non-LoF *KUK* groups using the 10 environmental variables that were used in MaxEnt (Warren et al., 2010). The observed *I* (*Io*) was calculated using the "ecological niche overlap" function implemented in ENMTools 1.3.

Measurements of root traits

Measurements were performed using 3.5-day-old seedlings grown on 1× MS medium with 1% sucrose in 16 h light/8 h darkness at 22°C (120 μmol m^{-2} s^{-1} light intensity). The photon flux density was 120 μmol m-2 s-1 (five cold white fluorescent Philips T5 28W/840 light bulbs and one warm white and yellow Philips T5 28W/830 light bulb). For root length measurement, roots were imaged using CCD scanners and measured using Fiji software (Schindelin et al., 2012). More than 15 seedlings were measured per line. To measure cellular traits, each seedling was treated with 10 μg/mL propidium iodide (PI) solution for 2 min and washed with water twice. Each root tip was subsequently imaged under a Zeiss 710 confocal microscope with a 20× objective. Both meristematic zone and mature cell sizes were measured using Fiji software (Schindelin et al., 2012). The length of the meristematic zone of each seedling was determined by measuring the distance from the quiescent center to the root cortex cell below

539 the first cortex cell that was twice as long as the cell below. The values for the
540 left and right side of root longitudinal sections were averaged to obtain the length
541 of the meristematic zone. The length of mature cells of each seedling was
542 determined by averaging cell lengths of three mature cortex cells in the zone in
543 which the xylem strands became visible. Five seedlings were measured per
544 accession.

545 **Mapping-by-sequencing of branch angle**

546 Plants of three natural accessions (Col-0, 3-2, and 29-8), and F_2 populations,
547 were grown under LD conditions (16 h light/8 h dark, 120 μmol m^{-2} s^{-1} light
548 intensity) at 20°C. When the lateral shoot was longer than 5 cm, a protractor was
549 used to measure the angle between the lateral branch and the main stem. The
550 branch angles were measured in 2016 and 2017.

551 DNA was extracted from pooled leaves of 57 plants with a large branch angle
552 (>85°) selected from a group of 684 F_2 plants of Col-0 × 29-8 with a large branch
553 angle (>85°). The pooled DNA sample was sequenced using Illumina HiSeq
554 Xten (150 bp pair-end reads, insert size of 450 bp). In total, 41,757,716 total
555 reads (approximately 34.5-fold coverage) were mapped to the Col-0 genome
556 (TAIR10), and SNPs were called using SHORE. The SNPs from pooled
557 populations were used as markers to identify regions with an excess of
558 homozygous alleles across the genome using SHOREmap (Schneeberger et al.,
559 2009).

560 **Statistical analysis**

561 Statistical analysis was performed using R (http://www.r-project.org/). Population
562 structure-corrected ANOVAs were performed according to a previous study (Li et
563 al., 2014). Significance was calculated by ANOVA with phenotypic differences
564 between haplotypes on the residuals after subtracting the best linear unbiased
565 predictors (BLUPs). EMMAX was used to estimate the BLUPs with a kinship
566 matrix calculated using an SNP matrix available based on mixed linear models
567 from the 1001 Genomes Project (Kang et al., 2010).

568 **Accession numbers**

569 Sequence data from this article can be found in the GenBank/EMBL libraries
570 under the following accession numbers: The *LAZY1* (AT5G14090) sequences of

Col-0, 3-2, and 29-8, under accession numbers MF621897–MF621899. Other genes studied in this article can be found in the Arabidopsis TAIR database (https://www.arabidopsis.org) under the following accession numbers: *CRK36* (AT4G04490), *KUK* (AT1G60370), *PRR5* (AT5G24470). The genome data for the pooled Col-0 × 29-8 F_2 populations have been deposited in the NCBI Sequence Read Archive (SRA) under accession number: SRR5947598.

Supplemental Data

Supplemental Figure 1. Accumulated false positive LoF rate.

Supplemental Figure 2. Variation in TE density (TE), nucleotide diversity (π), and the density of genes with stop-gain (stop-gain), frameshift (Frameshift), splice site (Splice), and LoF variants across the chromosomes.

Supplemental Figure 3. Distribution of stop-gain and frameshift variants at gene positions in all 1,071 accessions, distributed in 5% bins.

Supplemental Figure 4. Gene families with multiple genes have more LoF variants than gene families with one or two genes based on amino acid sequence similarity in the Col-0 genome.

Supplemental Figure 5. The number of observed LoF variants based on GWAS compared with the distribution from permutations based on non-LoF.

Supplemental Figure 6. Histogram distributions for each bioclimatic trait used for GWAS.

Supplemental Figure 7. The narrow sense of estimated heritability in GWAS analysis.

Supplemental Figure 8. Genome-wide association study of LoF variants and 10 climate parameters.

Supplemental Figure 9. Expression level variation of 30,000 random selected non-LoF SNP variants in 541 accessions based on RNA-seq data.

Supplemental Figure 10. Genome-wide association study of stop-gain variants and 10 climate parameters.

Supplemental Figure 11. Examples of the origin of LoF variants in the *KUK* gene based on 129 accessions with phenotypic data.

Supplemental Figure 12. Geographic distribution of LoF variants for two genes in 1,071 accessions.

Supplemental Figure 13. Cellular root trait variation for *KUK* alleles with and without LoF.

Supplemental Figure 14. Alignment of Col-0, 3-2 and 29-8 LAZY1 amino acid sequences.

Supplemental Figure 15. The results of niche identity tests (*I*) for niches between non-LoF and LoF groups.

Supplemental Data Set 1. Summary of the 893 samples used in this study from the 1001 Genomes Project.

Supplemental Data Set 2. Summary of the 61 samples from African sequenced genomes.

Supplemental Data Set 3. Summary of the 117 samples from our genome project used in this study.

Supplemental Data Set 4. LoF mutation validation based on 10 assembled genomes.

Supplemental Data Set 5. Environmental variables used in GWAS and ecological niche modeling.

Supplemental Data Set 6. LoF variants with significant signals in GWAS.

Supplemental Data Set 7. Stop-gain variants with significant signals in GWAS.

Supplemental Data Set 8. The 129 samples used in the root length analysis from a previous study.

Supplemental Data Set 9. Cellular root trait measurements for 10 individuals per accession.

Supplemental Data Set 10. The 812 samples used in the flowering time analysis.

Supplemental Data Set 11. Markers used in the analysis of plants with large branch angles of the F_2 generation of Col-0 × 29-8.

Supplemental Data Set 12. Markers used in the analysis of plants with large branch angles of the F_2 generation of 3-2 × 29-8.

Supplemental Data Set 13. 530 Genes located in the 48 regions under positive selection in the Yangtze River basin population.

Supplemental Data Set 14. Genes with LoF variants located within the regions under positive selection in the Yangtze River population.

Supplemental Data Set 15. Primers used for the sequencing and amplification of *LAZY1*.

Supplemental File 1. ANOVA/test tables.

ACKNOWLEDGEMENTS

We would like to thank Magnus Nordborg and Huijing Ma for helpful suggestions about the study, and members of the Guo lab for suggestions and comments about this work. Especially we thank the anonymous reviewers for their help improving the manuscript. This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB27010305), the Innovative Academy of Seed Design, Chinese Academy of Sciences, the National Natural Science Foundation of China (91731306 to Y.-L.G.), and start-up funds from the Salk Institute for Biological Research (W.B.).

AUTHOR CONTRIBUTIONS

Y.-L.G. conceived the study. Y.-C.X., J.-F.C., Y.-P.Z., Q.W., Y.E.Z., and Y.-L.G. analyzed the data. X.-M.N. and X.-X.L. conducted *LAZY1* experiments; W.H. and W.B. measured cellular traits for natural accessions. Y.-C.X. and Y.-L.G. wrote the paper with contributions from all authors.

REFERENCES

Abdel-Samad, R., Zalzali, H., Rammah, C., Giraud, J., Naudin, C., Dupasquier, S., Poulat, F., Boizet-Bonhoure, B., Lumbroso, S., Mouzat, K., Bonnans, C., Pignodel, C., Raynaud, P., Fort, P., Quittau-Prevostel, C., and Blache, P. (2011). MiniSOX9, a dominant-negative variant in colon cancer cells. Oncogene **30,** 2493-2503.

Albalat, R., and Canestro, C. (2016). Evolution by gene loss. Nat Rev Genet **17,** 379-391.

Amrad, A., Moser, M., Mandel, T., de Vries, M., Schuurink, R.C., Freitas, L., and Kuhlemeier, C. (2016). Gain and loss of floral scent production through changes in structural genes during pollinator-mediated speciation. Curr Biol **26,** 3303-3312.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate - a practical and powerful approach to multiple testing. J R Stat Soc B **57,** 289-300.

Birchler, J.A., and Veitia, R.A. (2007). The gene balance hypothesis: from classical genetics to modern genomics. Plant Cell **19,** 395-402.

Birchler, J.A., and Veitia, R.A. (2012). Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. Proceedings of the National Academy of Sciences of the United States of America **109,** 14746-14753.

Carvunis, A.R., Rolland, T., Wapinski, I., Calderwood, M.A., Yildirim, M.A., Simonis, N., Charloteaux, B., Hidalgo, C.A., Barbette, J., Santhanam, B., Brar, G.A., Weissman, J.S., Regev, A., Thierry-Mieg, N., Cusick, M.E., and Vidal, M. (2012). Proto-genes and de novo gene birth. Nature **487,** 370-374.

Chen, S., Zhang, Y.E., and Long, M. (2010). New genes in Drosophila quickly become essential. Science **330,** 1682-1685.

de Valles-Ibanez, G., Hernandez-Rodriguez, J., Prado-Martinez, J., Luisi, P., Marques-Bonet, T., and Casals, F. (2016). Genetic load of loss-of-function polymorphic variants in great apes. Genome Biol Evol **8,** 871-877.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., McKenna, A., Fennell, T.J., Kernytsky, A.M., Sivachenko, A.Y., Cibulskis, K., Gabriel, S.B., Altshuler, D., and Daly, M.J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet **43,** 491-498.

Derrien, T., Estelle, J., Marco Sola, S., Knowles, D.G., Raineri, E., Guigo, R., and Ribeca, P. (2012). Fast computation and applications of genome mappability. PLoS One **7,** e30377.

Durvasula, A., Fulgione, A., Gutaker, R.M., Alacakaptan, S.I., Flood, P.J., Neto, C., Tsuchimatsu, T., Burbano, H.A., Pico, F.X., Alonso-Blanco, C., and Hancock, A.M. (2017). African genomes illuminate the early history and transition to selfing in *Arabidopsis thaliana*. Proceedings of the National Academy of Sciences of the United States of America **114,** 5213-5218.

Edgar, R.C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics **5,** 113.

Edger, P.P., and Pires, J.C. (2009). Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. Chromosome Res **17,** 699-717.

Gan, X., Stegle, O., Behr, J., Steffen, J.G., Drewe, P., Hildebrand, K.L., Lyngsoe, R., Schultheiss, S.J., Osborne, E.J., Sreedharan, V.T., Kahles, A., Bohnert, R., Jean, G., Derwent, P., Kersey, P., Belfield, E.J., Harberd, N.P., Kemen, E., Toomajian, C., Kover, P.X., Clark, R.M., Ratsch, G., and Mott, R. (2011). Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. Nature **477,** 419-423.

712 **Goldman-Huertas, B., Mitchell, R.F., Lapoint, R.T., Faucher, C.P.,**
713 **Hildebrand, J.G., and Whiteman, N.K.** (2015). Evolution of herbivory in
714 Drosophilidae linked to loss of behaviors, antennal responses, odorant
715 receptors, and ancestral diet. Proceedings of the National Academy of
716 Sciences of the United States of America **112,** 3026-3031.

717 **Green, C., Willoughby, J., Study, D.D.D., and Balasubramanian, M.** (2017).
718 De novo *SETD5* loss-of-function variant as a cause for intellectual disability
719 in a 10-year old boy with an aberrant blind ending bronchus. Am J Med
720 Genet A **173,** 3165-3171.

721 **Greenberg, A.J., Moran, J.R., Coyne, J.A., and Wu, C.I.** (2003). Ecological
722 adaptation during incipient speciation revealed by precise gene replacement.
723 Science **302,** 1754-1757.

724 **Gujas, B., Alonso-Blanco, C., and Hardtke, C.S.** (2012). Natural Arabidopsis
725 *brx* loss-of-function alleles confer root adaptation to acidic soil. Curr Biol **22,**
726 1962-1968.

727 **Guo, Y.L.** (2013). Gene family evolution in green plants with emphasis on the
728 origination and evolution of *Arabidopsis thaliana* genes. Plant J **73,** 941-951.

729 **Hao, Y., Washburn, J.D., Rosenthal, J., Nielsen, B., Lyons, E., Edger, P.P.,**
730 **Pires, J.C., and Conant, G.C.** (2018). Patterns of population variation in two
731 paleopolyploid eudicot lineages suggest that dosage-based selection on
732 homeologs Is long-lived. Genome Biol Evol **10,** 999-1011.

733 **Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., and Jarvis, A.** (2005).
734 Very high resolution interpolated climate surfaces for global land areas. Int J
735 Climatol **25,** 1965-1978.

736 **Hoballah, M.E., Gubitz, T., Stuurman, J., Broger, L., Barone, M., Mandel, T.,**
737 **Dell'Olivo, A., Arnold, M., and Kuhlemeier, C.** (2007). Single gene-
738 mediated shift in pollinator attraction in *Petunia*. Plant Cell **19,** 779-790.

739 **Hodgson, J.A., Pickrell, J.K., Pearson, L.N., Quillen, E.E., Prista, A., Rocha,**
740 **J., Soodyall, H., Shriver, M.D., and Perry, G.H.** (2014). Natural selection for
741 the Duffy-null allele in the recently admixed people of Madagascar. Proc Biol
742 Sci **281,** 20140930.

743 **Hottes, A.K., Freddolino, P.L., Khare, A., Donnell, Z.N., Liu, J.C., and**
744 **Tavazoie, S.** (2013). Bacterial adaptation through loss of function. Plos
745 Genet **9,** e1003617.

746 **Hou, J., Shi, X., Chen, C., Islam, M.S., Johnson, A.F., Kanno, T., Huettel, B.,**
747 **Yen, M.R., Hsu, F.M., Ji, T., Chen, P.Y., Matzke, M., Matzke, A.J.M.,**
748 **Cheng, J., and Birchler, J.A.** (2018). Global impacts of chromosomal
749 imbalance on gene expression in Arabidopsis and other taxa. Proceedings of
750 the National Academy of Sciences of the United States of America **115,**
751 E11321-E11330.

Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.Y., Freimer, N.B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. Nat Genet **42,** 348-354.

Kawakatsu, T., Huang, S.C., Jupe, F., Sasaki, E., Schmitz, R.J., Urich, M.A., Castanon, R., Nery, J.R., Barragan, C., He, Y., Chen, H., Dubin, M., Lee, C.R., Wang, C., Bemm, F., Becker, C., O'Neil, R., O'Malley, R.C., Quarless, D.X., Genomes, C., Schork, N.J., Weigel, D., Nordborg, M., and Ecker, J.R. (2016). Epigenomic diversity in a global collection of *Arabidopsis thaliana* accessions. Cell **166,** 492-505.

Kremling, K.A.G., Chen, S.Y., Su, M.H., Lepak, N.K., Romay, M.C., Swarts, K.L., Lu, F., Lorant, A., Bradbury, P.J., and Buckler, E.S. (2018). Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. Nature **555,** 520-523.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., Tukiainen, T., Birnbaum, D.P., Kosmicki, J.A., Duncan, L.E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., Cooper, D.N., Deflaux, N., DePristo, M., Do, R., Flannick, J., Fromer, M., Gauthier, L., Goldstein, J., Gupta, N., Howrigan, D., Kiezun, A., Kurki, M.I., Moonshine, A.L., Natarajan, P., Orozco, L., Peloso, G.M., Poplin, R., Rivas, M.A., Ruano-Rubio, V., Rose, S.A., Ruderfer, D.M., Shakir, K., Stenson, P.D., Stevens, C., Thomas, B.P., Tiao, G., Tusie-Luna, M.T., Weisburd, B., Won, H.H., Yu, D., Altshuler, D.M., Ardissino, D., Boehnke, M., Danesh, J., Donnelly, S., Elosua, R., Florez, J.C., Gabriel, S.B., Getz, G., Glatt, S.J., Hultman, C.M., Kathiresan, S., Laakso, M., McCarroll, S., McCarthy, M.I., McGovern, D., McPherson, R., Neale, B.M., Palotie, A., Purcell, S.M., Saleheen, D., Scharf, J.M., Sklar, P., Sullivan, P.F., Tuomilehto, J., Tsuang, M.T., Watkins, H.C., Wilson, J.G., Daly, M.J., MacArthur, D.G., and Exome Aggregation, C. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature **536,** 285-291.

Li, P., Filiault, D., Box, M.S., Kerdaffrec, E., van Oosterhout, C., Wilczek, A.M., Schmitt, J., McMullan, M., Bergelson, J., Nordborg, M., and Dean, C. (2014). Multiple *FLC* haplotypes defined by independent cis-regulatory variation underpin life history diversity in *Arabidopsis thaliana*. Genes Dev **28,** 1635-1640.

Li, Z.W., Chen, X., Wu, Q., Hagmann, J., Han, T.S., Zou, Y.P., Ge, S., and Guo, Y.L. (2016). On the origin of de novo genes in *Arabidopsis thaliana* populations. Genome Biol Evol **8,** 2190-2202.

Lipka, A.E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P.J., Gore, M.A., Buckler, E.S., and Zhang, Z. (2012). GAPIT: genome association and prediction integrated tool. Bioinformatics **28,** 2397-2399.

MacArthur, D.G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J.K., Montgomery, S.B.,

Albers, C.A., Zhang, Z.D., Conrad, D.F., Lunter, G., Zheng, H., Ayub, Q., DePristo, M.A., Banks, E., Hu, M., Handsaker, R.E., Rosenfeld, J.A., Fromer, M., Jin, M., Mu, X.J., Khurana, E., Ye, K., Kay, M., Saunders, G.I., Suner, M.M., Hunt, T., Barnes, I.H., Amid, C., Carvalho-Silva, D.R., Bignell, A.H., Snow, C., Yngvadottir, B., Bumpstead, S., Cooper, D.N., Xue, Y., Romero, I.G., Genomes Project, C., Wang, J., Li, Y., Gibbs, R.A., McCarroll, S.A., Dermitzakis, E.T., Pritchard, J.K., Barrett, J.C., Harrow, J., Hurles, M.E., Gerstein, M.B., and Tyler-Smith, C. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. Science **335,** 823-828.

Marco-Sola, S., Sammeth, M., Guigo, R., and Ribeca, P. (2012). The GEM mapper: fast, accurate and versatile alignment by filtration. Nat Methods **9,** 1185-1188.

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., and Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics **26,** 2069-2070.

McLysaght, A., and Guerzoni, D. (2015). New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. Philos Trans R Soc Lond B Biol Sci **370,** 20140332.

Meijon, M., Satbhai, S.B., Tsuchimatsu, T., and Busch, W. (2014). Genome-wide association study using cellular traits identifies a new regulator of root development in Arabidopsis. Nat Genet **46,** 77-81.

Meinke, D., Muralla, R., Sweeney, C., and Dickerman, A. (2008). Identifying essential genes in *Arabidopsis thaliana*. Trends Plant Sci **13,** 483-491.

Meng, D., Dubin, M., Zhang, P., Osborne, E.J., Stegle, O., Clark, R.M., and Nordborg, M. (2016). Limited contribution of DNA methylation variation to expression regulation in *Arabidopsis thaliana*. Plos Genet **12,** e1006141.

Michael, T.P., Jupe, F., Bemm, F., Motley, S.T., Sandoval, J.P., Lanz, C., Loudet, O., Weigel, D., and Ecker, J.R. (2018). High contiguity *Arabidopsis thaliana* genome assembly with a single nanopore flow cell. Nat Commun **9,** 541.

Nakamichi, N., Takao, S., Kudo, T., Kiba, T., Wang, Y., Kinoshita, T., and Sakakibara, H. (2016). Improvement of Arabidopsis biomass and cold, drought and salinity stress tolerance by modified circadian clock-associated PSEUDO-RESPONSE REGULATORs. Plant Cell Physiol **57,** 1085-1097.

Narasimhan, V.M., Hunt, K.A., Mason, D., Baker, C.L., Karczewski, K.J., Barnes, M.R., Barnett, A.H., Bates, C., Bellary, S., Bockett, N.A., Giorda, K., Griffiths, C.J., Hemingway, H., Jia, Z., Kelly, M.A., Khawaja, H.A., Lek, M., McCarthy, S., McEachan, R., O'Donnell-Luria, A., Paigen, K., Parisinos, C.A., Sheridan, E., Southgate, L., Tee, L., Thomas, M., Xue, Y., Schnall-Levin, M., Petkov, P.M., Tyler-Smith, C., Maher, E.R., Trembath, R.C., MacArthur, D.G., Wright, J., Durbin, R., and van Heel, D.A. (2016).

Health and population effects of rare gene knockouts in adult humans with related parents. Science **352,** 474-477.

Olson, M.V. (1999). When less is more: gene loss as an engine of evolutionary change. Am J Hum Genet **64,** 18-23.

Palmieri, N., Kosiol, C., and Schlötterer, C. (2014). The life cycle of Drosophila orphan genes. eLife **3,** e01311.

Phillips, S.J., and Dudik, M. (2008). Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography **31,** 161-175.

Phillips, S.J., Anderson, R.P., and Schapire, R.E. (2006). Maximum entropy modeling of species geographic distributions. Ecol Model **190,** 231-259.

Prachumwat, A., and Li, W.H. (2008). Gene number expansion and contraction in vertebrate genomes with respect to invertebrate genomes. Genome Res **18,** 221-232.

Prieto-Godino, L.L., Rytz, R., Bargeton, B., Abuin, L., Arguello, J.R., Peraro, M.D., and Benton, R. (2016). Olfactory receptor pseudo-pseudogenes. Nature **539,** 93-97.

Saleheen, D., Natarajan, P., Armean, I.M., Zhao, W., Rasheed, A., Khetarpal, S.A., Won, H.H., Karczewski, K.J., O'Donnell-Luria, A.H., Samocha, K.E., Weisburd, B., Gupta, N., Zaidi, M., Samuel, M., Imran, A., Abbas, S., Majeed, F., Ishaq, M., Akhtar, S., Trindade, K., Mucksavage, M., Qamar, N., Zaman, K.S., Yaqoob, Z., Saghir, T., Rizvi, S.N., Memon, A., Hayyat Mallick, N., Ishaq, M., Rasheed, S.Z., Memon, F.U., Mahmood, K., Ahmed, N., Do, R., Krauss, R.M., MacArthur, D.G., Gabriel, S., Lander, E.S., Daly, M.J., Frossard, P., Danesh, J., Rader, D.J., and Kathiresan, S. (2017). Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. Nature **544,** 235-239.

Sas, C., Muller, F., Kappel, C., Kent, T.V., Wright, S.I., Hilker, M., and Lenhard, M. (2016). Repeated inactivation of the first committed enzyme underlies the loss of benzaldehyde emission after the selfing transition in *Capsella*. Curr Biol **26,** 3313-3319.

Schild, L., Canessa, C.M., Shimkets, R.A., Gautschi, I., Lifton, R.P., and Rossier, B.C. (1995). A mutation in the epithelial sodium channel causing Liddle disease increases channel activity in the *Xenopus laevis* oocyte expression system. Proceedings of the National Academy of Sciences of the United States of America **92,** 5699-5703.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J.Y., White, D.J., Hartenstein, V., Eliceiri, K., Tomancak, P., and Cardona, A. (2012). Fiji: an open-source platform for biological-image analysis. Nat Methods **9,** 676-682.

Schneeberger, K., Ossowski, S., Lanz, C., Juul, T., Petersen, A.H., Nielsen, K.L., Jorgensen, J.E., Weigel, D., and Andersen, S.U. (2009). SHOREmap: simultaneous mapping and mutation identification by deep sequencing. Nat Methods **6,** 550-551.

Slater, G.S., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics **6,** 31.

Song, X.J., Huang, W., Shi, M., Zhu, M.Z., and Lin, H.X. (2007). A QTL for rice grain width and weight encodes a previously unknown RING-type E3 ubiquitin ligase. Nat Genet **39,** 623-630.

Swets, J.A. (1988). Measuring the accuracy of diagnostic systems. Science **240,** 1285-1293.

Tanaka, H., Osakabe, Y., Katsura, S., Mizuno, S., Maruyama, K., Kusakabe, K., Mizoi, J., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2012). Abiotic stress-inducible receptor-like kinases negatively control ABA signaling in Arabidopsis. Plant J **70,** 599-613.

Tang, C., Toomajian, C., Sherman-Broyles, S., Plagnol, V., Guo, Y.L., Hu, T.T., Clark, R.M., Nasrallah, J.B., Weigel, D., and Nordborg, M. (2007). The evolution of selfing in *Arabidopsis thaliana*. Science **317,** 1070-1072.

The 1001 Genomes Consortium. (2016). 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. Cell **166,** 481-491.

Tian, T., Liu, Y., Yan, H., You, Q., Yi, X., Du, Z., Xu, W., and Su, Z. (2017). agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. Nucleic Acids Res **45,** W122-W129.

Wagner, A. (2005). Distributed robustness versus redundancy as causes of mutational robustness. Bioessays **27,** 176-188.

Warren, D.L., Glor, R.E., and Turelli, M. (2010). ENMTools: A toolbox for comparative studies of environmental niche models. Ecography **33,** 607-611.

Will, J.L., Kim, H.S., Clarke, J., Painter, J.C., Fay, J.C., and Gasch, A.P. (2010). Incipient balancing selection through adaptive loss of aquaporins in natural *Saccharomyces cerevisiae* populations. Plos Genet **6,** e1000893.

Wu, W., Liu, X., Wang, M., Meyer, R.S., Luo, X., Ndjiondjop, M.N., Tan, L., Zhang, J., Wu, J., Cai, H., Sun, C., Wang, X., Wing, R.A., and Zhu, Z. (2017). A single-nucleotide polymorphism causes smaller grain size and loss of seed shattering during African rice domestication. Nat Plants **3,** 17064.

Xie, J., Murone, M., Luoh, S.M., Ryan, A., Gu, Q., Zhang, C., Bonifas, J.M., Lam, C.W., Hynes, M., Goddard, A., Rosenthal, A., Epstein, E.H., Jr., and de Sauvage, F.J. (1998). Activating *Smoothened* mutations in sporadic basal-cell carcinoma. Nature **391,** 90-92.

Yang, H., He, B.Z., Ma, H., Tsaur, S.C., Ma, C., Wu, Y., Ting, C.T., and Zhang, Y.E. (2015). Expression profile and gene age jointly shaped the genome-wide distribution of premature termination codons in a *Drosophila melanogaster* population. Mol Biol Evol **32,** 216-228.

Yoshihara, T., Spalding, E.P., and Iino, M. (2013). *AtLAZY1* is a signaling component required for gravitropism of the *Arabidopsis thaliana* inflorescence. Plant J **74,** 267-279.

Zapata, L., Ding, J., Willing, E.M., Hartwig, B., Bezdan, D., Jiao, W.B., Patel, V., Velikkakam James, G., Koornneef, M., Ossowski, S., and Schneeberger, K. (2016). Chromosome-level assembly of Arabidopsis thaliana Ler reveals the extent of translocation and inversion polymorphisms. Proceedings of the National Academy of Sciences of the United States of America **113,** E4052-4060.

Zhang, Z., Ersoz, E., Lai, C.Q., Todhunter, R.J., Tiwari, H.K., Gore, M.A., Bradbury, P.J., Yu, J., Arnett, D.K., Ordovas, J.M., and Buckler, E.S. (2010). Mixed linear model approach adapted for genome-wide association studies. Nat Genet **42,** 355-360.

Zhao, L., Saelao, P., Jones, C.D., and Begun, D.J. (2014). Origin and spread of de novo genes in *Drosophila melanogaster* populations. Science **343,** 769-772.

Zou, Y.P., Hou, X.H., Wu, Q., Chen, J.F., Li, Z.W., Han, T.S., Niu, X.M., Yang, L., Xu, Y.C., Zhang, J., Zhang, F.M., Tan, D., Tian, Z., Gu, H., and Guo, Y.L. (2017). Adaptation of *Arabidopsis thaliana* to the Yangtze River basin. Genome Biol **18,** 239.

Zufall, R.A., and Rausher, M.D. (2004). Genetic changes associated with floral adaptation restrict future evolutionary potential. Nature **428,** 847-850.

Table 1. Total number of different LoF variants in the 1,071 genomes. Numbers in parentheses indicate the gene number; some genes have multiple LoF variants.

	Before filtering		After filtering	
Variant type	**Total**	**Per sample**	**Total**	**Per sample**
Stop-gain	28,944 (10,350)	432.1 (403.5)	17,453 (7,264)	214.7 (214.7)
Frameshift	72,335 (14,994)	1,448.4 (1,178.9)	37,935 (10,800)	512.2 (512.2)
Splice site	10,414 (5,769)	277.5 (264.9)	5,431 (3,449)	103.0 (103.0)
Total	111,693 (17,957)	2,157.9 (1,732.8)	60,819 (12,918)	829.9 (829.9)

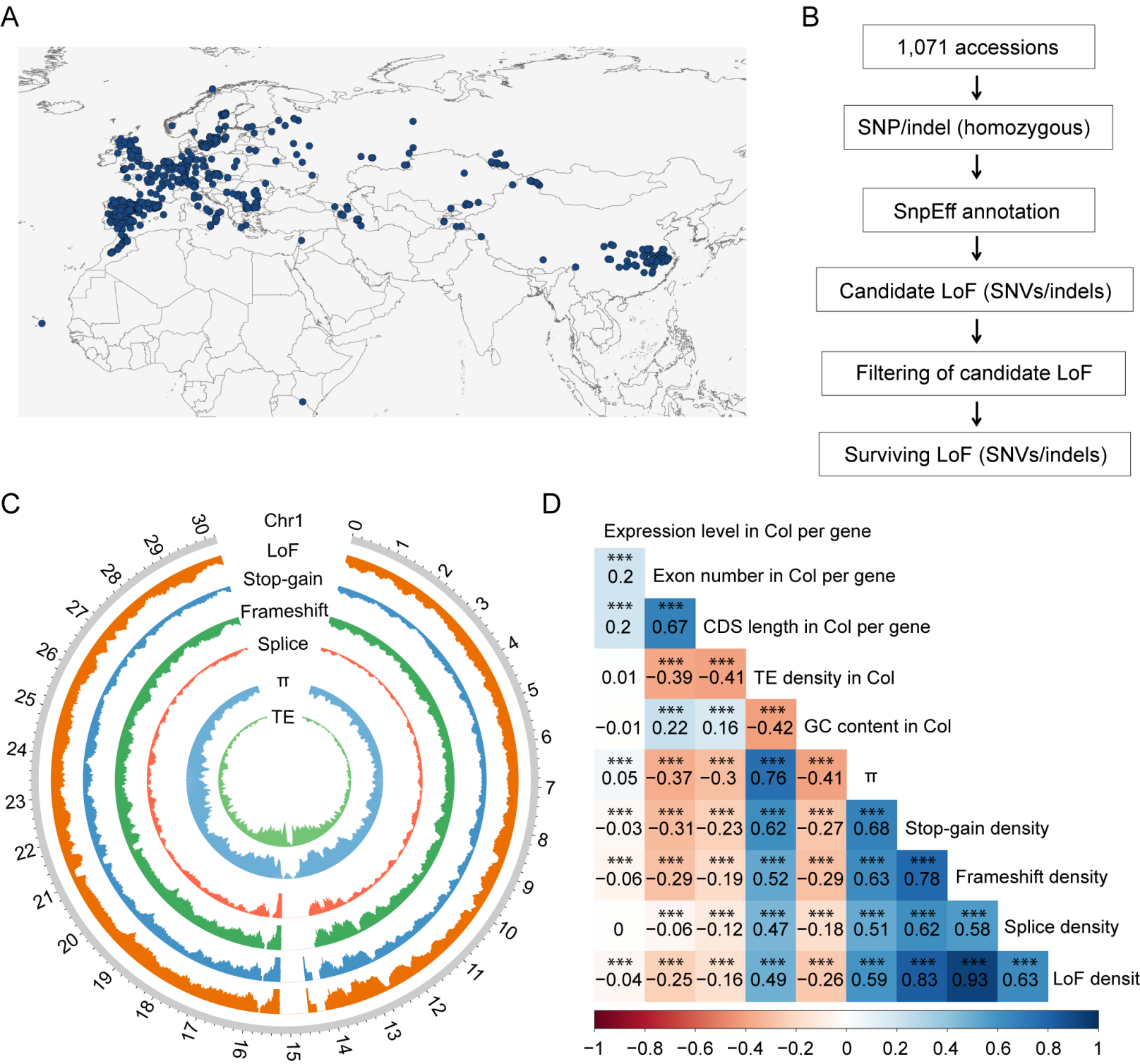

Figure 1. Identification of LoF mutations in 1,071 Arabidopsis thaliana genomes. (A) Geographic distribution of the 1,071 accessions. (B) Process to identify LoF variants. (C) Variation in TE density (TE), nucleotide diversity (π), and the density of genes with stop-gain (Stop-gain), frameshift (Frameshift), splice site (Splice), and LoF variants across chromosome 1. The other four chromosomes are shown in Supplemental Figure 1. (D) Spearman correlations adjusted for multiple testing between different gene features across the whole genome, including expression level, exon number, and CDS length per gene in Col-0; GC content in Col-0; TE density in Col-0; nucleotide diversity (π); and the density of genes with stop-gain (Stop-gain density), frameshift (Frameshift density), splice site (Splice density), and LoF (LoF density). Values marked with asterisks indicate strong correlation coefficients (*** $p < 0.001$).

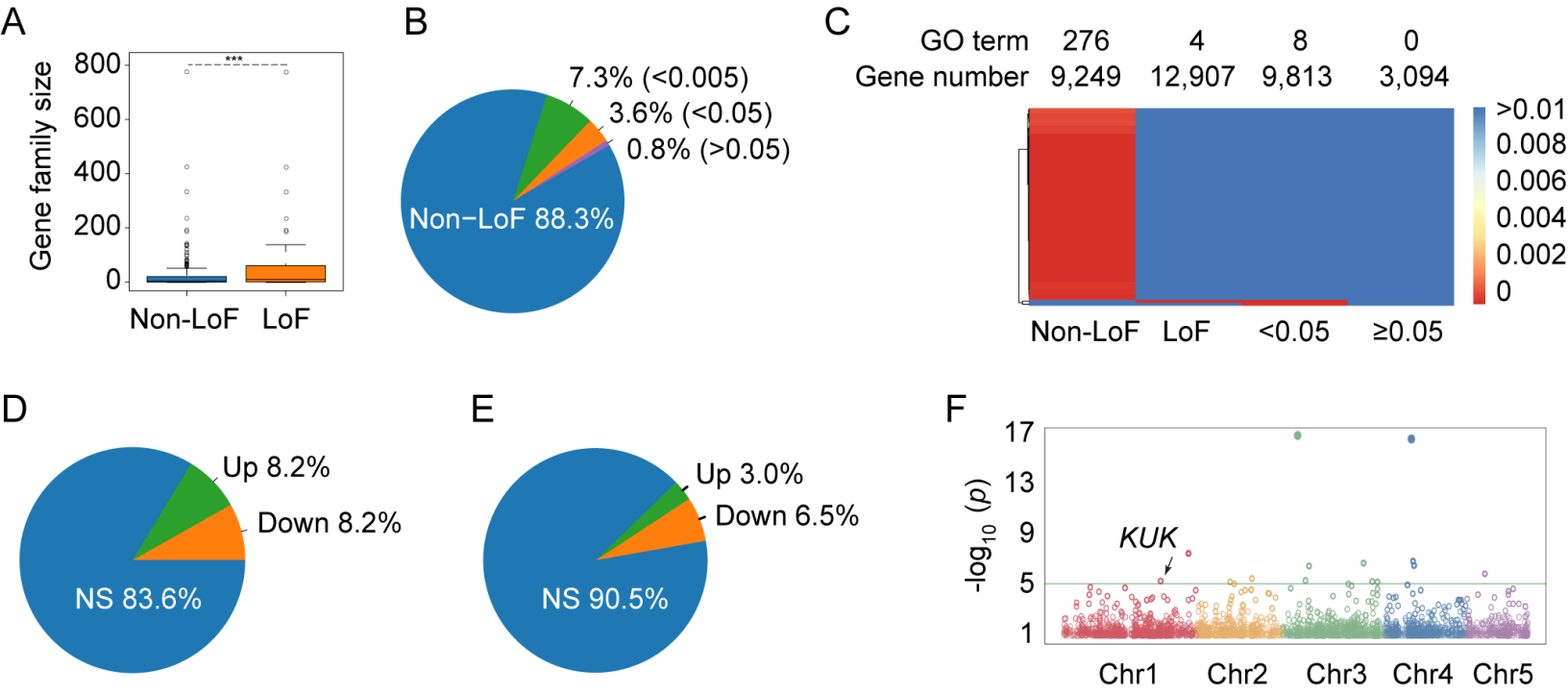

Figure 2. Functional analyses of genes with and without LoF variants. (A) Genes from larger gene families tend to have LoF variants based on the Col-0 genome. In the boxplot, the box equals the difference between the 75th and 25th percentiles. The thick line indicates the median. The two whiskers indicate 1.5 interquartile range of the lower quartile or the upper quartile. Outliers are plotted as individual points. *** $p < 0.001$. (B) LoF mutations are under-represented in lethal genes in A. thaliana. "<0.005" indicates a minor allele frequency (MAF) less than 0.5%, "<0.05" indicates MAF < 5%, and ">0.05" indicates MAF > 5%. (C) Gene Ontology enrichment analysis of genes with LoF, with different LoF allele frequencies, and those without LoF variants across the 1,071 genomes. Color indicates p values of GO enrichment analysis. (D) Expression level variation between alleles with and without LoF variants of genes with significant GWAS signals based on RNA-seq data for 541 accessions. (E) Expression level variation between alleles with and without LoF variants (6,675 genes) based on RNA-seq data for 541 accessions. NS indicates insignificant difference. (F) Genome-wide association study of stop-gain variants with regard to precipitation in the wettest month.

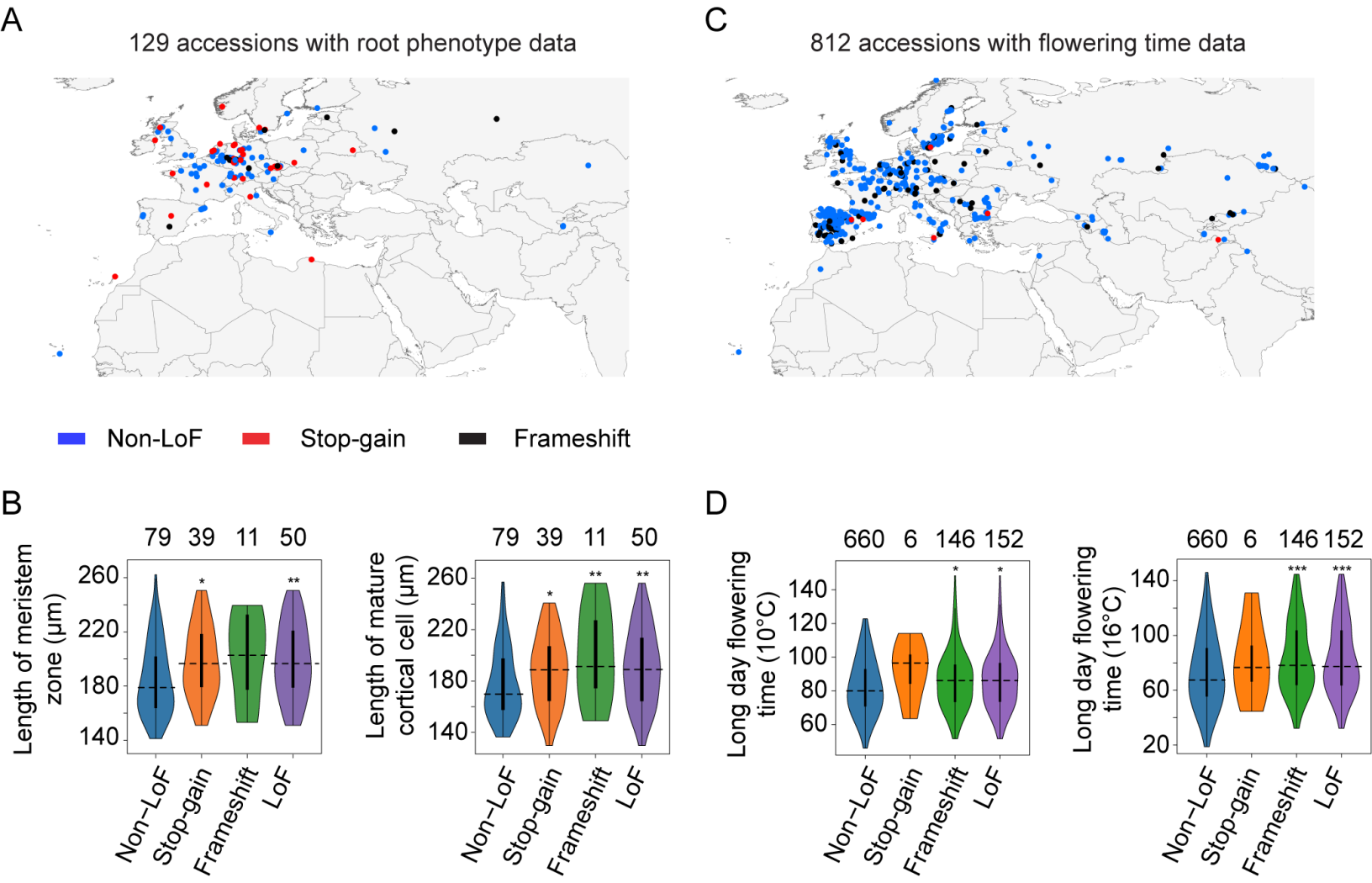

Figure 3. Functional differentiation of alleles with or without LoF. (A) Geographic distribution of different KUK alleles in 129 accessions with phenotypic data. (B) Phenotypic variation for KUK alleles with and without LoF. (C) Geographic distribution of PRR5 alleles in 812 accessions with flowering time data. (D) Flowering time variation for PRR5 alleles with and without LoF. The numbers above each violin plot indicate the number of accessions containing the particular variant type. The dashed line indicates the median value of the phenotypic trait. Population structure-corrected ANOVA, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

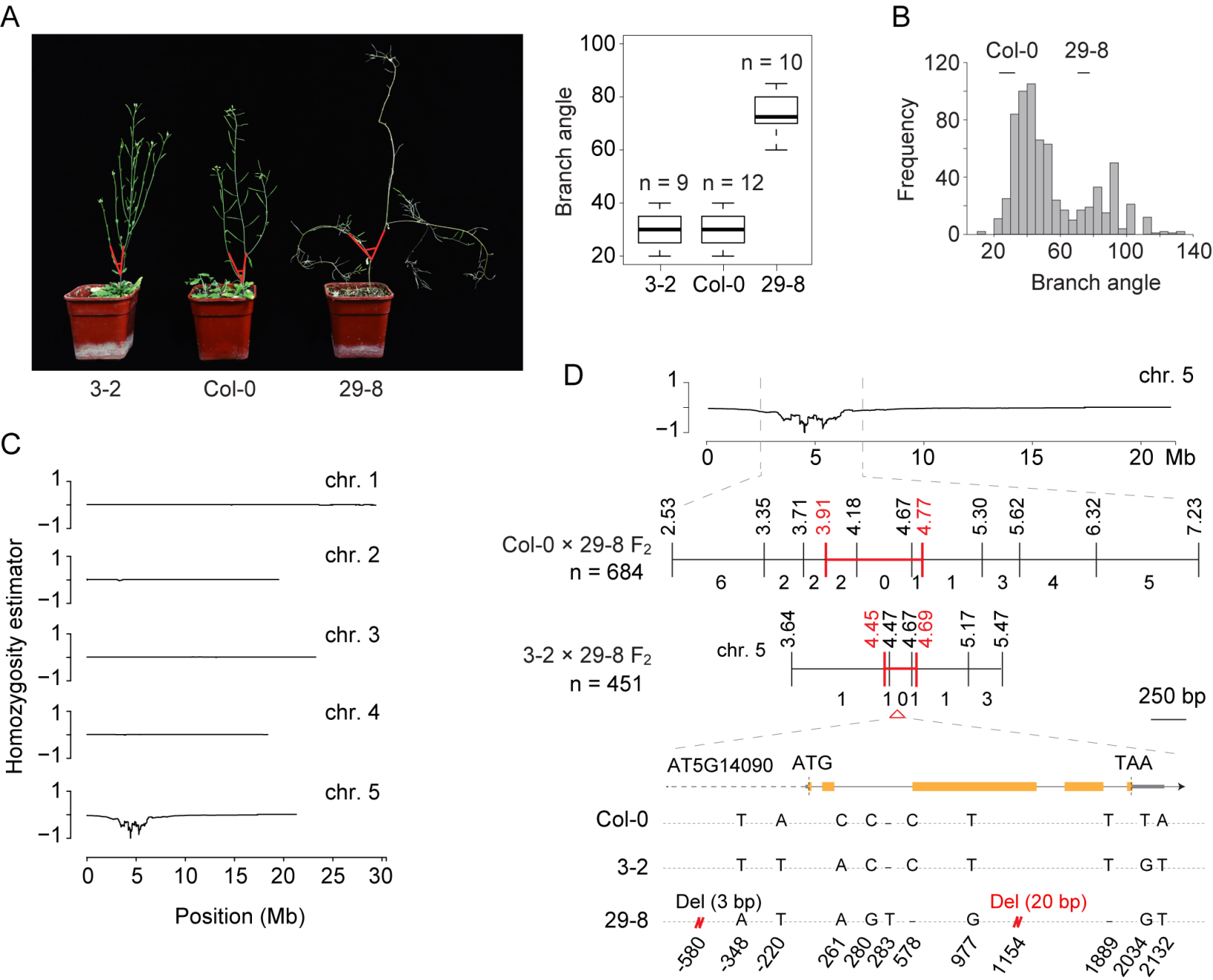

Figure 4. Frameshift in LAZY1 increases branch angle. (A) Representative accessions with different branch angles. The red triangles show the tangent lines drawn for the measurement. The boxplot (right) indicates distribution of branch angles in different accessions. The borders of the box indicate the 75th and 25th percentiles. The thick line marks the median. The two whiskers extend to the most extreme data points. n indicates the number of individuals measured for each accession. (B) Distribution of branch angle frequency of the 684 F2 individuals from the Col-0 × 29-8 cross. Average and range of branch angles of two natural accessions are indicated. (C) SHOREmap analysis of branch angle. The homozygosity estimator is 0 at even allele frequencies for both natural accessions, 1 when homozygous for the small angle accession Col-0, and -1 when homozygous for the large angle accession 29-8. (D) Fine mapping by genetic-linkage analysis and sequence variation in the candidate gene LAZY1. The causal locus was narrowed down to the region between markers 4.45 M and 4.69 M. The number of recombinants between the markers and the causal locus is indicated on the bottom of the linkage map. Schematic illustration of sequence variation in LAZY1 is shown for Col-0, 3-2, and 29-8. LAZY1 is shown at the top, and the position of ATG is defined as +1. Thin lines indicate introns; thick yellow lines indicate protein-coding sequences; the frameshift induced by the 20 bp deletion is indicated in red.

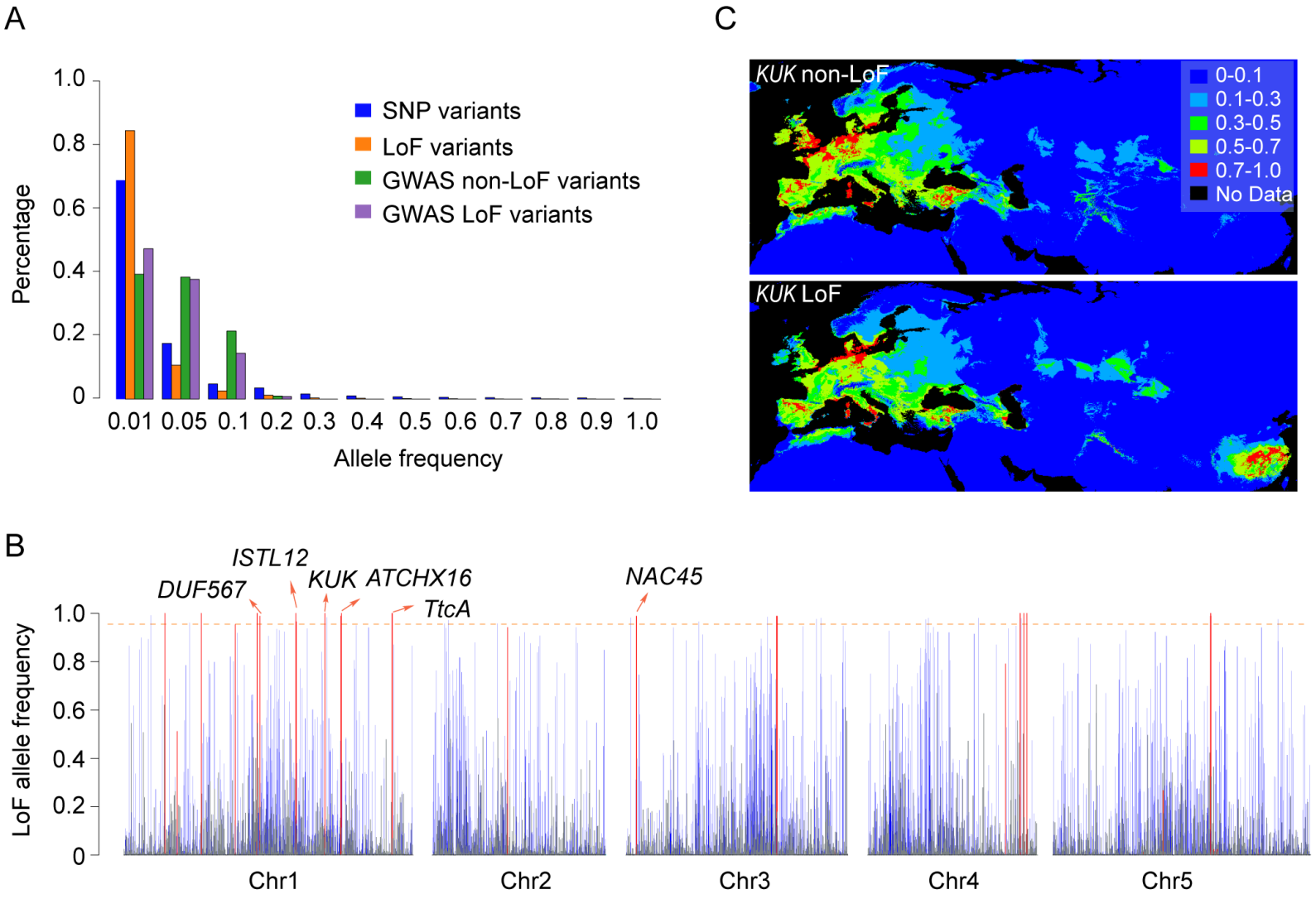

Figure 5. LoF variants are under natural selection. (A) Distribution of allele frequencies of all SNP variants and LoF variants, LoF variants (GWAS LoF variants) and permutated non-LoF variants (GWAS non-LoF variants) associated with environmental variables, at the genome level across 1,071 accessions. (B) Genes under positive selection in the Yangtze River population based on a selective sweep analysis. Gray lines indicate allele frequencies of 60,819 LoF variants across 1,071 accessions; blue lines indicate genes with 3,349 LoF variants in the Yangtze River basin population; red lines indicate genes with LoF variants in regions with positive selection in the Yangtze River basin population. The horizontal dashed line indicates an allele frequency of 95%. (C) Ecological niche modeling of *KUK*. The niche overlap between the LoF and non-LoF haplotype groups assessed by Warren's I similarity statistics (*Io* = 0.877, $p < 0.001$); areas of suitable habitats are marked in different colors corresponding to a scale from 0 to 1, based on 100 pseudo-replicates.