A New Catalog of Structural Variants in 1,301 *A. thaliana* Lines from Africa, Eurasia, and North America Reveals a Signature of Balancing Selection at Defense Response Genes

Mehmet Göktay, Andrea Fulgione, and Angela. M. Hancock ⓘ*

Max Planck Institute for Plant Breeding Research, Cologne, Germany

*Corresponding author: E-mail: hancock@mpipz.mpg.de.
Associate editor Amanda Larracuente

Abstract

Genomic variation in the model plant *Arabidopsis thaliana* has been extensively used to understand evolutionary processes in natural populations, mainly focusing on single-nucleotide polymorphisms. Conversely, structural variation has been largely ignored in spite of its potential to dramatically affect phenotype. Here, we identify 155,440 indels and structural variants ranging in size from 1 bp to 10 kb, including presence/absence variants (PAVs), inversions, and tandem duplications in 1,301 *A. thaliana* natural accessions from Morocco, Madeira, Europe, Asia, and North America. We show evidence for strong purifying selection on PAVs in genes, in particular for housekeeping genes and homeobox genes, and we find that PAVs are concentrated in defense-related genes (R-genes, secondary metabolites) and F-box genes. This implies the presence of a "core" genome underlying basic cellular processes and a "flexible" genome that includes genes that may be important in spatially or temporally varying selection. Further, we find an excess of intermediate frequency PAVs in defense response genes in nearly all populations studied, consistent with a history of balancing selection on this class of genes. Finally, we find that PAVs in genes involved in the cold requirement for flowering (vernalization) and drought response are strongly associated with temperature at the sites of origin.

Key words: presence–absence variation, core genome, dispensable genome, balancing selection, R-genes, F-box.

Introduction

The model plant *Arabidopsis thaliana* has been important for deciphering core physiological, developmental, and adaptive processes (Somerville and Koornneef 2002; Provart et al. 2016). *Arabidopsis thaliana* grows naturally in diverse environments throughout Eurasia and Africa, where populations are exposed to differing selective pressures. Natural variation of *A. thaliana* has mainly been studied using single-nucleotide polymorphisms (SNPs) (Alonso-Blanco et al. 2016; Durvasula et al. 2017; Zou et al. 2017; Fulgione et al. 2018). However, SNPs represent only a subset of the variation in the genome. Structural variants (SVs) are generally defined as genomic variants larger than 50 bp including presence/absence variants (PAVs), tandem duplications, inversions, translocations, and complex SVs (Eisfeldt et al. 2019; Kosugi et al. 2019). Compared with SNPs, SVs are likely to cause more dramatic effects on gene functions and phenotypes (Kosugi et al. 2019). Structural variation can also have indirect effects through repression of meiotic crossovers (Sturtevant 1926; Morgan et al. 2017; Rowan et al. 2019).

SVs are common in humans (estimated at >20K variants per individual) and only a subset has been associated with phenotypes, including disease (Weischenfeldt et al. 2013; Abel et al. 2020; Ho et al. 2020). For example, in humans, an approximately 3-Mb deletion on chromosome 15 (chr15q11-13—paternal) was shown to result in loss of function of multiple genes and cause Prader–Willi syndrome (Weischenfeldt et al. 2013), and Charcot–Marie–Tooth disease is known to be caused by a duplication event on Chromosome 17 (chr17p12), which damages peripheral nerves (Weischenfeldt et al. 2013). In Drosophila, several individual SVs are associated with fitness and display clinal patterns (González et al. 2010; Kapun et al. 2016; Durmaz et al. 2018).

Plant genomes seem to be exceptional at tolerating structural variation. Extensive variation in gene content has been observed across individuals in rice (Wang et al. 2018; Fuentes et al. 2019; Choi et al. 2020), maize (Springer et al. 2009; Sun et al. 2018), *A. thaliana* (Cao et al. 2011; Zmienko et al. 2020), and grapes (Zhou et al. 2019) and several studies have implicated individual SVs in trait variation. Structural variation resulting from transposable element insertions has been shown to play roles in domestication in maize (Studer et al. 2011) and in gene expression divergence between *Arabidopsis* species (Hollister et al. 2011) and copy number variation is linked with several postdomestication traits (Lye and Purugganan 2019). Other studies identified large-scale chromosomal inversions associated with salt tolerance in Mimulus (Lowry and Willis 2010), flowering time in *A. thaliana* (Fransz et al. 2016), berry color in grapes (Zhou et al. 2019), awn length in basmati rice (Choi et al. 2020), and the loss of the
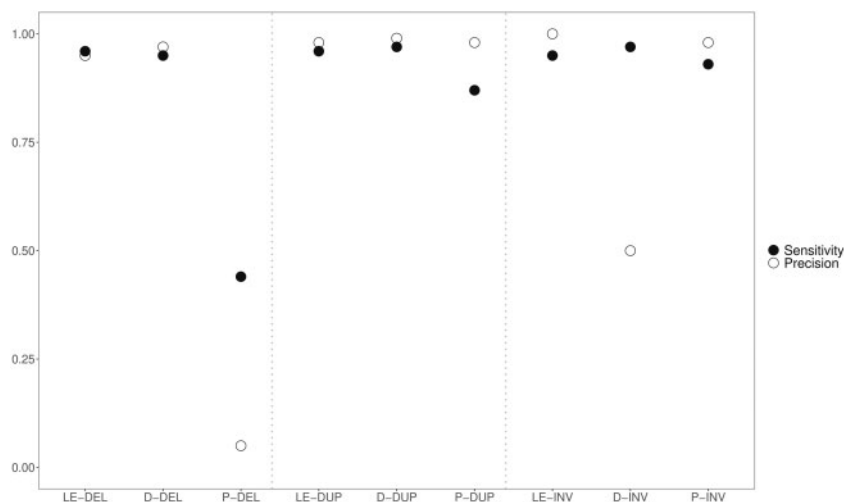
Open Access

FIG. 1. Comparison of SV callers based on simulations and PacBio. Abbreviations along the *x* axis represent the combination of method used and variant type and are defined as follows: LE, LumpyExpress; D, Delly; P, Pindel; DEL, deletion; DUP, duplication; INV, inversion.

jointed fruit pedicel (Soyk et al. 2019) and fruit size (Alonge et al. 2020) in tomato.

In this study, we identify SVs in 1,301 *A. thaliana* accessions from Africa, Europe, Asia, and North America and examine the evolutionary history of these SV polymorphisms. First, we assess the precision and sensitivity of several available methods by comparison to simulations and long-read data to produce an analysis pipeline. Given the limitations of short-read data, we focus our analyses on indels and SVs (including PAVs, tandem duplications, and inversions) smaller than 10 kb. We identify structural variation in *A. thaliana* natural accessions examine their global patterns of polymorphism. We identify sets of "core" highly conserved genes and variable "dispensable" genes, a subset of which show evidence for balancing selection. Finally, we examine evidence of local adaptation on PAVs based on correlations with environmental variables. This analysis reveals strong correlations for PAVs in three genes involved in the vernalization (cold) response for flowering as well as genes involved in drought and heat tolerance. We make our variant calls available together with information about the level of support for each call, which can be used for flexible integration of SVs into existing analysis pipelines.

Results

Performance Comparison of SV Callers

Identifying and calling SVs in short-read data is not a trivial task. To identify the best software for calling SVs, we compared the performance of three popular tools: PINDEL, DELLY, and LUMPYEXPRESS (Ye et al. 2009; Rausch et al. 2012; Layer et al. 2014). We conducted comparisons to SVs produced in simulations as well as to calls made from long-read Pacific Biosciences (hereafter PacBio) data that we generated.

For the simulation-based approach, LUMPYEXPRESS outperformed DELLY and PINDEL (fig. 1). Both precision and sensitivity for LUMPYEXPRESS and DELLY were very high (≥95%) for deletions and tandem duplications, whereas both measures were much lower for PINDEL (44% and 5%). For inversions, LUMPYEXPRESS had much higher precision (100%) than DELLY (50%), which called many false positives and PINDEL performed slightly worse than LUMPYEXPRESS. In silico simulations provide information about the performance of SV detection tools under ideal conditions, but they assume simple scenarios. However, in real population-level data, SVs may be more complex, with multiple SVs and mutations occurring at the same locus. To explore this, we also compared the performance of the three tools relative to calls from long-read data (PacBio) in an empirical case. For this, we used the Cvi-0 accession, which is one of the very most diverged accessions compared with the Col-0 (TAIR10) reference and therefore represents a particularly challenging case, where complex structural variation is likely. We note that this case is challenging both for long-read and short-read callers.

We considered the results from long-read calls (PacBio) to represent a "high confidence" set and then assessed how well calls from short-read data using LUMPYEXPRESS, DELLY, and PINDEL agreed with this high confidence set. Although PacBio data are much more reliable for large SVs, error rates in PacBio data are high for individual SNPs and short INDELs. Therefore, in an effort to maximize the true positive rate for calls with the PacBio data, we only considered SVs between 50 bp and 10 kb in length in this comparative analysis. Supplementary figure S1, Supplementary Material online, shows the performance of LUMPYEXPRESS, DELLY, and PINDEL on short-read data from Cvi-0 relative to the high PacBio-based confidence set. For the tandem duplications and inversions, agreement was never above 45%. However, deletions and associated PAVs were called with relatively high agreement. LUMPYEXPRESS called 62% of the high confidence set of PAVs with 80% agreement to SNIFFLES and DELLY called 64% of PAVs with 74% agreement to SNIFFLES.

Given the discrepancy between the very high sensitivity and precision, we observed in simulations compared with the lower levels with PacBio data, we were interested in better

Fig. 2. Geographical distribution of *Arabidopsis thaliana* samples included in this study.

understanding what drives the concordant and discordant calls in short- and long-read data. We focused on the LUMPYEXPRESS software and examined alignments in randomly chosen regions where calls based on PacBio data (using SNIFFLES) and Illumina data (using LUMPYEXPRESS) agreed and where they differed. Supplementary figures S2–S4, Supplementary Material online, show five randomly selected regions of agreement between calls with Illumina and PacBio data, five randomly selected PAVs identified from Illumina data but not from PacBio data and five randomly selected PAVs identified in PacBio data but not in Illumina data. Somewhat surprisingly, we found that the discrepancies tended to result from false negatives in one or the other technology and/or offsets in the true breakpoints. Based on this sample, the errors appeared to be equally prevalent in long- and short-read data. Discrepancies tended occur in regions where alignments are imperfect so that called PAVs are likely true positives that went undetected by either the short-read or long-read calling approaches. The discrepant regions tended to be complex (possibly involving multiple layered SVs) and/or repetitive regions, which are known to be problematic for calling both SNPs and SVs in long- and short-read data. Overall, the lower matching to PacBio long-read data appears to result from the general problem that calling SVs in complex genomic regions is equally error-prone using short- (Illumina) and long-read (PacBio) data. Based on the results of comparisons to both simulations and real data, we decided to use LUMPYEXPRESS for identifying and calling SVs in the global sample set.

Alignment, SV Calling, and Genotyping in 1,301 Diverse *A. thaliana* Accessions

Next, we examined the patterns of variation in structural polymorphisms across diverse wild *A. thaliana* accessions. Figure 2 shows the geographic distribution of the samples that we included in this study. We identified 155,440 SVs ranging in size from 1 bp to 10 kb among 1,301 accessions using a pipeline consisting of LUMPYEXPRESS, SVTYPER, and SVTOOLS (supplementary fig. S5, Supplementary Material online). This set includes 124,905 PAVs, 25,061 tandem duplications, and 5,474 inversions. Although we make the entire set of SVs publicly available (PRJEB38975), our subsequent analyses focus on PAV polymorphisms, which were called with the greatest precision in our testing data set. For researchers interested in using this data set, we make information available about the strength of evidence for SVs based on the number of reads supporting each SV as well as the type of evidence (split read and/or discordant reads), which could be used for further (i.e., more stringent) filtering.

PAVs Recapitulate Population Clustering Obtained from SNPs

We examined the global pattern of polymorphism in our total set of PAVs in order to assess whether these variants recapitulated the signals found in SNP data. We consider this to be a further test to validate the PAV genotyping because if the quality of PAV genotypes is high, we expect to see global patterns that are similar to those found using SNP data. We clustered a representative subset of diverse accessions using PAVs and compared this with results obtained previously from SNP data (Durvasula et al. 2017). We found that the overall structure of the NJ-tree based on PAVs (fig. 3) recapitulates that based on SNPs (figure 2A in Durvasula et al. 2017). First, both SNPs and PAVs clearly separate the Eurasian nonrelict clade (comprising the majority of Eurasian accessions) from the relict clades (highly diverged groups mainly found in the Iberian Peninsula and Africa). Second, similar to the pattern observed for SNPs, the Eurasian clade has a nearly star-shaped pattern with little reticulation whereas the relict clades are more deeply reticulated with longer internal
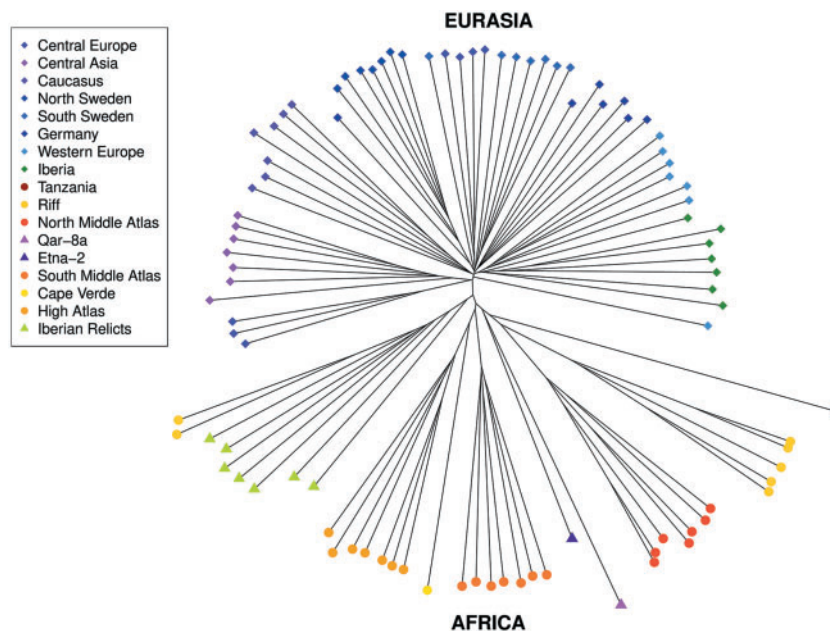
Fig. 3. Neighbor-joining tree including seven representative samples for each population. To produce the neighbor-joining tree, the full set of accessions was down-sampled to achieve similar numbers for different geographic regions. The tree is based on 95 accessions which contain a total of 9,062 PAVs. The tree shows a clear separation between Eurasia and Africa, which further validates PAV calls. PAVs give reliable information to recapitulate previous findings from SNP data.

branches. Finally, clustering using SVs recovered the fine-scale differentiation of population clusters within Africa and most of that within Eurasia.

Identification of Genomic Hotspots of PAVs and Conserved Regions

Although reduction in effective population size due to selfing is expected to reduce the efficiency of selection at weakly deleterious alleles (Schemske and Lande 1985; Charlesworth et al. 1990; Charlesworth and Wright 2001), variants that are lethal or have severe effects in the homozygous state are more often exposed to selection in selfing compared with outcrossing species (Stebbins 1950; Shields 1982; Schemske and Lande 1985; Glémin 2007). To determine whether there is evidence for purifying selection acting to limit PAVs, we examined their frequencies and genome-wide distributions. At a broad scale, PAVs were nearly uniformly distributed across chromosomes, except in pericentromeric regions where they were more common (supplementary fig. S6, Supplementary Material online). When we focused specifically on genic and nongenic regions within chromosome arms, we found that PAVs were enriched by 3.5-fold in nongenic relative to genic regions ($\chi^2 = 1.8E7$, df = 1, $P = 1E$-16), consistent with stronger purifying selection in genic regions (Charlesworth et al. 1993) (table 1). Purifying selection on weakly deleterious alleles is expected to increase the relative proportion of low-frequency variants in functional relative to neutrally evolving genomic regions (Nielsen 2005). We compared the unfolded site frequency spectra (SFS) (supplementary fig. S7, Supplementary Material online) for genic and intergenic loci for all populations where ancestral state could be assigned based on consensus calls in a set of divergent genomes. We did not find

enrichment of low-frequency variants in genic regions relative to nongenic regions; rather, we found a slight deficit in all cases. Taken together our results suggest that new genic PAVs often have very strong deleterious effects and are quickly removed by purifying selection, so that they are not found in segregating variation. Conversely, the set of genic PAVs that are segregating at appreciable frequencies do not show evidence of purifying selection relative to intragenic PAVs.

Next, we asked which genes or pathways showed deficits of PAVs. To do that, we extracted all genes that never contained any PAVs in any samples and performed gene ontology (GO) enrichment analysis. We tested significance using both Fisher's exact tests (FET) and a more conservative permutation-based approach that corrects for clustering of signals in the genome (Gowinda) (Huang et al. 2009; Kofler et al. 2012). Categories that were significantly enriched with both FET and Gowinda include translation (Benjamini-corrected P values for FET = 5.4E-33 and for Gowinda = 9.13E-05) and translational elongation (Benjamini P values for FET = 1.1E-4 and for Gowinda = 2.93E-02). These categories include tRNAs and rRNAs, which are classical examples of housekeeping genes (Rifkind et al. 1976). In addition, the GO term regulation of DNA-templated transcription, which contains homeobox genes and transcription factors was also strongly enriched (Benjamini P value for FET = 6.5E-03 and Gowinda = 9.12E-05). Homeobox genes play an important role in body plan specification of higher organisms during early stages of embryogenesis (Duverger and Morasso 2008). Additional GO term categories that were found to be enriched either in FET or Gowinda analyses are listed in (supplementary tables S2 and S3, Supplementary Material online).

Table 1. Distribution of PAVs across the Genome for 1,301 *Arabidopsis thaliana* Accessions.

Genomic Partition	Mb	Number of PAVs	Mean Length	Median Length	Mb PAVs	Proportion Containing PAVs
Whole genome	119,146,348	124,905	681.8	91	39,696,386	0.333
Intergenic	59,041,854	87,627	547.9	87	30,766,220	0.521
Genic	60,104,494	37,278	996.1	100	8,920,166	0.148

Table 2. Multiple-Test Significant GO Terms for Genes Carrying PAV Polymorphisms Among the Total Set of 1,301 Samples.

GO Term	Enrichment Score	Benjamini (Corrected P Value)	Test Statistics
Signal transduction	1.2	2.2E-6	One-tailed Fisher's exact test
Defense response	1.2	1.4E-3	One-tailed Fisher's exact test
SCF-dependent proteasomal ubiquitin-dependent protein catabolic process	1.4	1.4E-3	One-tailed Fisher's exact test
Cell–cell signaling	2.4	1.84E-3	Permutation-based test
Lipid transport	1.6	1.84E-3	Permutation-based test
Lipid localization	1.6	1.84E-3	Permutation-based test
Lignan metabolic process	2.6	4.46E-3	Permutation-based test
Lignan biosynthetic process	2.6	4.46E-3	Permutation-based test

NOTE.—Two methods (FET and Gowinda) were employed to be able to identify enriched categories with different test statistics.

Table 3. Defense Response Enrichment with Significance Assessed by One-Tailed Fisher's Exact Tests and Gowinda for Each Population.

Population	One-Tailed Fisher's Exact Test		Gowinda	
	Enrichment Score	Benjamini Corrected P Value	Enrichment Score	Benjamini Corrected P Value
Asia	1.4	1.20E-07	1.2	1.83E-03
Central Europe	1.3	2.60E-05	1.1	1.17E-01
Germany	1.4	1.60E-08	1.1	3.38E-01
High Atlas (Morocco)	1.6	9.40E-11	1.4	2.45E-03
Iberian Relicts	1.5	4.20E-10	1.3	2.23E-03
Iberian nonrelicts	1.4	8.30E-10	1.1	7.94E-02
Italy, Balkans, and Caucasus	1.4	2.70E-09	1.2	1.25E-03
North Middle Atlas (Morocco)	1.7	2.00E-13	1.5	5.70-03
North Sweden	1.6	2.30E-13	1.4	5.43E-03
Riff (Morocco)	1.7	8.80E-11	1.4	2.83E-03
South Middle Atlas (Morocco)	1.6	1.60E-12	1.4	2.77E-03
South Sweden	1.4	1.70E-09	1.3	1.55E-03
Western Europe	1.4	1.50E-11	1.2	8.62E-03
Madeira	1.7	2.00E-09	1.4	5.72E-03
Yangtze River Basin (PopY)	1.6	5.50E-12	1.5	6.03E-03
North-Western China (PopN)	1.6	2.40E-08	1.5	4.06E-03

Although genes show a deficit of PAVs relative to nongenic regions overall, some types of genes may be more likely to contain PAVs than others. To identify these, we extracted locations of all genes that overlap at least one PAV among the 1,301 samples and performed GO enrichment analysis using both FET and the Gowinda permutation-based approach (Huang et al. 2009; Kofler et al. 2012). The results are reported in (table 2).

Signal transduction was the most significantly enriched class with FET, and the enrichment was driven largely by defense-related genes mainly consisting of TIR-NBS, TIR-NBS-LRR, LLR, and TIR classes and secondary metabolites. Other enriched classes included the more specific defense response category and a category related to the production of Lignans, a class of secondary metabolites (Bagniewska-Zadworna et al. 2014). In addition, SCF-dependent proteasomal ubiquitin-dependent protein catabolic processes were enriched, which contains many F-box genes. F-box genes have previously been noted to evolve rapidly (Xu et al.

2009) and are known to be involved in several crucial processes related to environmental stress response including embryogenesis, hormonal responses, seedling development, floral organogenesis, senescence, and pathogen resistance (Xu et al. 2009). Based on Gowinda analysis, the most enriched GO terms that we identified are known to be related to stress including cell–cell signaling (Xing and Laroche 2011), lipid transport, and localization (Yeats and Rose 2008). The defense response GO term was also enriched in marginal tests using Gowinda ($P = 2E-2$), but the enrichment was not significant with Benjamini correction. This discrepancy is likely due to the extreme clustering of defense genes across the genome.

Next, we examined the patterns within populations. We extracted all PAVs for each population and performed GO term enrichment. Table 3 shows defense response enrichment from FET and Gowinda. The population-based analysis reveals significant Benjamini enrichment of defense response genes with PAVs across almost all populations for both
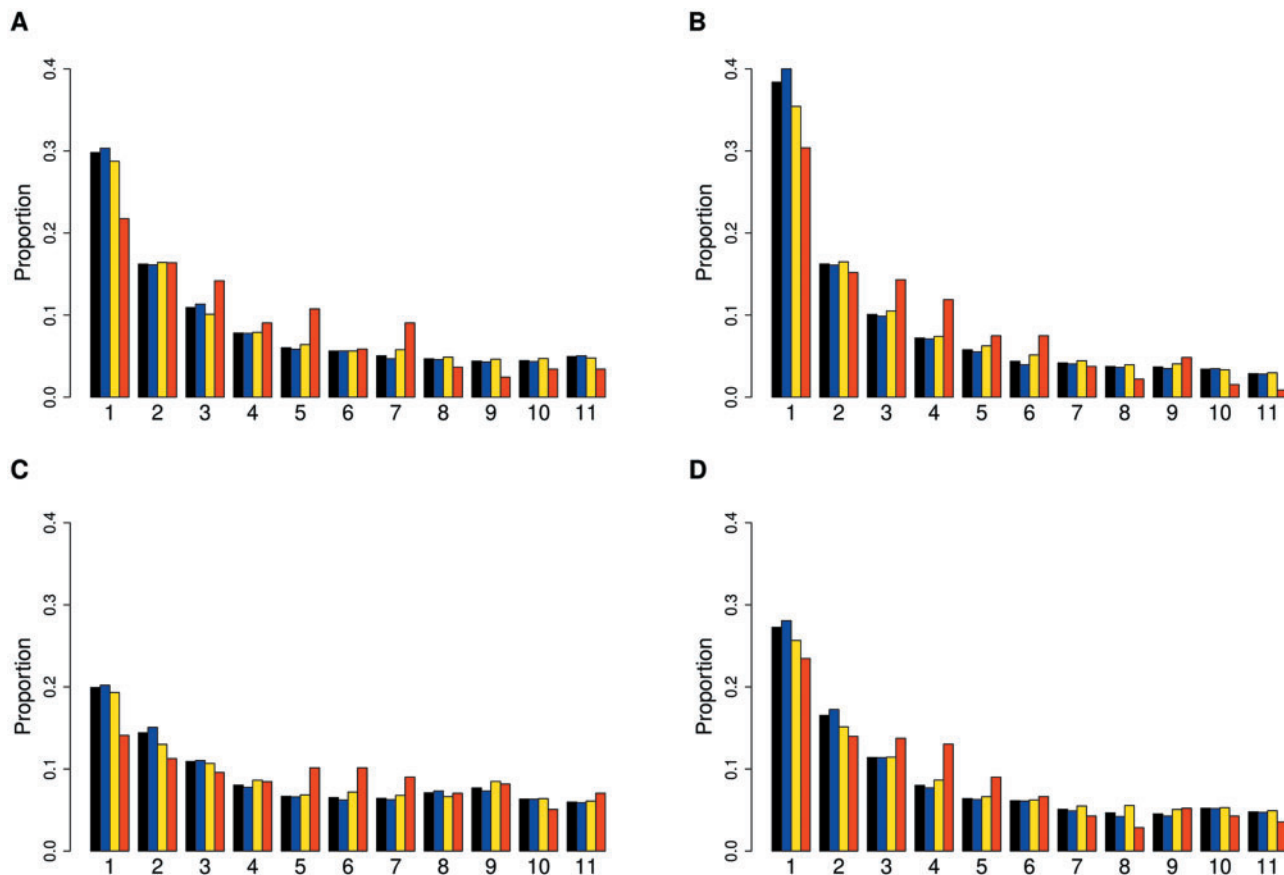
Fig. 4. SFS of PAVs for whole genome (black), intergenic (blue), genic(yellow), and defense genes (red) in four representative populations: (A) South Middle Atlas, (B) Iberian Relicts, (C) Asia (PopN), and (D) North Sweden.

enrichment tests. The signal was somewhat weaker in Central Europe, Germany, and Iberian nonrelicts, where defense response genes were not enriched at the Benjamini significance level using the Gowinda method. Other categories that were often enriched across individual populations are signal transduction and secondary metabolite biosynthetic processes (supplementary tables S4–S19, Supplementary Material online).

Since we observed very strong signals in defense response genes, we were interested in better understanding the evolutionary forces acting on these. We examined the SFS of defense-related genes relative to the complete set of genic and nongenic categories for each population. In most populations, defense PAV frequencies were shifted towards intermediate levels relative to other genes (fig. 4 and supplementary fig. S8, Supplementary Material online). To quantitatively compare the frequency distribution of PAVs overlapping defense genes with the total sets of genes and intergenic regions, we used the Tajima's *D* statistic. This statistic summarizes the information in the SFS such that an excess of intermediate frequency variants results in a more positive Tajima's *D*, and an excess of low-frequency variants results in a more negative Tajima's *D* (Tajima 1989). For each population, Tajima's *D* was more positive for defense genes compared with the total set of genes. To assess significance for this result, we randomly subsampled genic PAVs to match

the number of defense response PAVs 100K times and calculated an empirical *P* value based on the distribution of Tajima's *D* in the sampled data sets. Population bottlenecks and expansions affect the frequency spectrum and therefore the genome-wide value of Tajima's *D*. We found that Tajima's *D* was highly significantly elevated in defense PAVs in all populations except Yangtze River Basin (PopY) (not significant; $P = 0.551$) and Northern Sweden, where the significance was marginal ($P = 0.0254$) (fig. 5 and supplementary table S20, Supplementary Material online). Both populations are known to have experienced strong bottlenecks in the past (Huber et al. 2014; Zou et al. 2017) and accordingly have a higher Tajima's *D* in intergenic regions. We further assessed evidence of long-term balancing selection at defense loci based on a signal of long-range LD among SNPs in these regions. For this, we used the BetaScan method (Siewert and Voight 2017, 2020). We calculated beta scores for each population and tested for enrichment of the defense genes in the 5% tail of the distribution of beta scores. For each population, we found a significant enrichment of these defense genes in the tail of this distribution (supplementary table S21, Supplementary Material online). Taken together, the shift towards intermediate frequency variation resulting in a shift to more positive Tajima's *D* and the observed enrichment of long-range LD in defense genes is consistent with balancing selection on this class of loci.
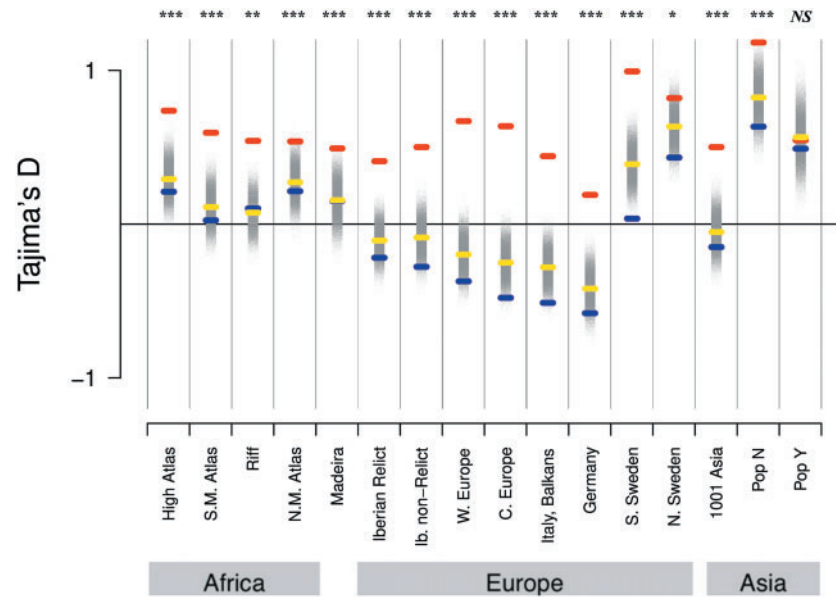
Fig. 5. Tajima's *D* of intergenic (blue), genic (yellow), and defense (red) loci. To assess significance for Tajima's *D* in defense response genes, 100K sets of genic PAVs were randomly subsampled to match the number of defense response PAVs and these distributions are shown in gray. Asterisks denote level of significance; $*P < 0.5 \times 10^{-2}$, $**P < 5 \times 10^{-3}$; $***P < 5 \times 10^{-4}$; and NS, not significant.

Correlation with Environmental Variables (GWAS)

We next asked whether there was evidence for involvement of specific PAVs in local adaptation based on associations with environmental variables. For this, we accessed data for four representative environmental variables (Bio5: Maximum temperature of the warmest month, Bio6: Minimum temperature of the coldest month, Bio13: Precipitation of the wettest month and Bio14: Precipitation of the driest month) from the WorldClim database (Fick and Hijmans 2017). To identify SVs that were correlated with the environment while controlling for potential confounding effects of population structure, we performed GWAS with a linear mixed model that controls for relatedness using a kinship matrix as a random variable (Zhou and Stephens 2012) and environmental variables as phenotypes (fig. 6A–D). We also conducted the same analyses for SNPs so that we could compare the signals between the two classes of variants.

Several genic PAVs (PAVs within 10 kb of genes) strongly associated with minimum temperature in the coldest month (Bio6) were related to the timing of flowering and photosynthesis (fig. 6B). There were strong correlations for multiple PAVs in a *MADS AFFECTING FLOWERING* (*MAF*) gene cluster at the bottom of chromosome 5, including a Bonferroni significant association with a PAV that impacts *MAF3* as well as several less significantly correlated (P value range: 2×10^{-7} to 1×10^{-2}) PAVs that affect other genes in this cluster. *MAF3* and other *MAF* cluster genes work together with *FLM* to repress expression of *FT* and inhibit flowering and play a role in flowering in response to cold (vernalization) (Ratcliffe et al. 2003; Gu et al. 2013). Photosynthetic efficiency is reduced in cold, which can result in damage from build-up of reactive oxygen species (ROS) (Tripathy and Oelmüller 2012; Prinzenberg et al. 2020). Several PAVs in genes involved

in photosynthesis and ROS production were among the most strongly correlated with minimum temperature, including *AT1G22700* (*tetratricopeptide repeat [TPR]-like superfamily protein*) (P value: 4.8×10^{-7}), *AT1G22710* (*SUC2*) (P value: 4.8×10^{-7}), *AT1G43560* (*THIOREDOXIN Y2*) (P value: 1.3×10^{-6}), *AT5G09600* (*Succinate dehydrogenase 3-1*) (P value: 3.5×10^{-5}), and *AT5G22140* (*FAD/NAD(P)-binding oxidoreductase family protein*) (P value: 3.6×10^{-5}).

For maximum temperature in the warmest month (Bio5) PAVs at three genes involved in photomorphogenesis (hypocotyl elongation), a phenotype with a plastic temperature-mediated effect, were among the strongest correlations (fig. 6A). These included AT5G11260 (*ELONGATED HYPOCOTYL 5*) (P value: 1.5×10^{-5}), AT5G42350 (CFK1) (P value: 8.3×10^{-5}), and AT5G58140 (*PHOTOTROPIN 2*) (P value: 2.4×10^{-5}). Among the strongest correlations with precipitation in the wettest month (Bio13) were PAVs within 10 kb of several genes and gene clusters involved in cell wall morphogenesis: *RGXT1* (AT4G01770) and *RGXT2* (AT4G01750) (P value: 4.9×10^{-6}), *Walls Are Thin 1* (AT1G75500) (P value: 1×10^{-6}), and *EXPA18* (AT1G62980) and *KNAT7* (AT1G62990) (P value: 9.5×10^{-5}) (fig. 6C). A PAV in a cluster of several immune response genes (AT1G72890–AT1G72950) was among the most correlated with precipitation in the driest month (Bio14) (P value: 1.9×10^{-5}) (fig. 6D).

Although having information about PAVs can be specifically useful to identify candidates in GWAS for functional follow-up analysis and therefore creates added value compared with SNPs alone, some information may be partially redundant with SNP results in the sense that LD between PAVs and SNPs will result in statistical associations with both types of variants. To examine this, we compared the signals we found with PAVs to those with SNPs in the same regions.
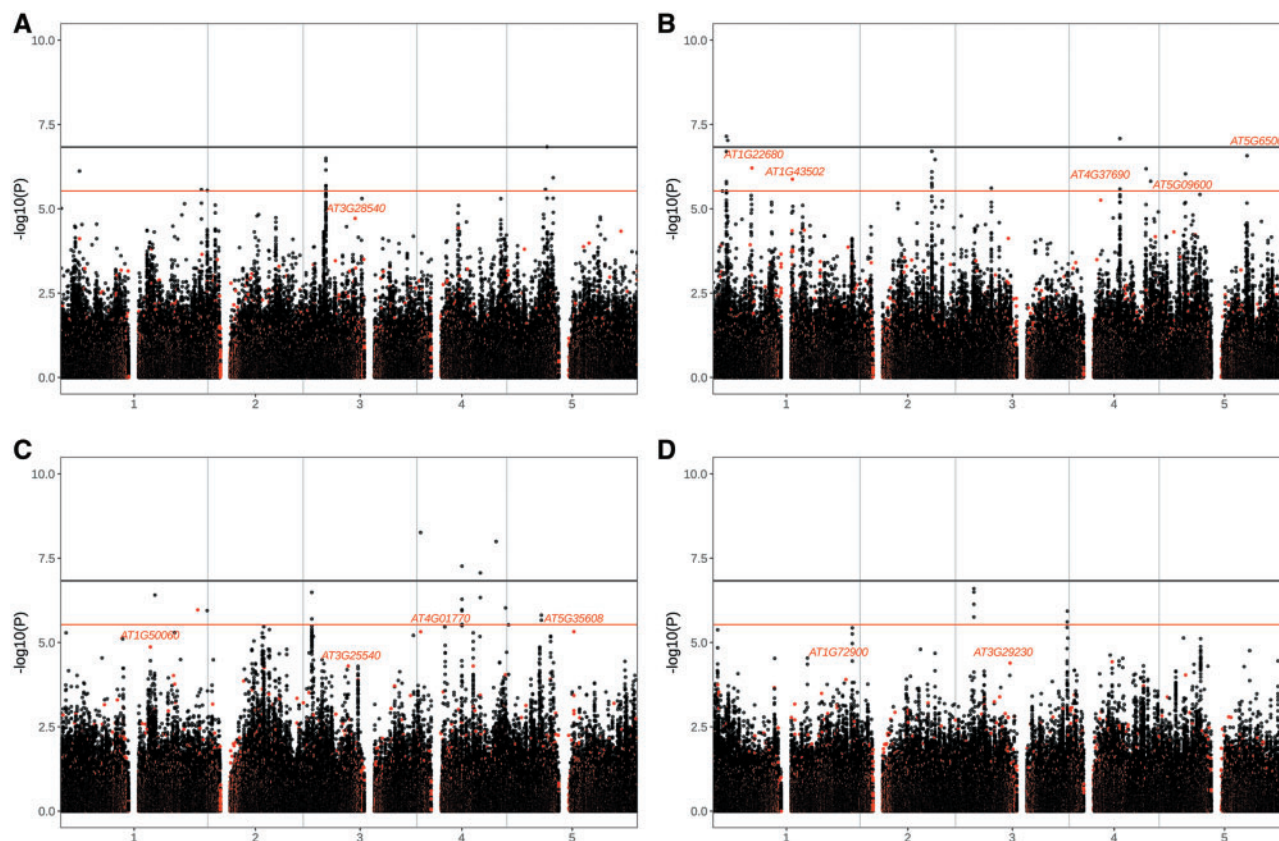
Fig. 6. Plots of $-\log_{10}(P$ values) from genome-wide association mapping between PAVs and SNPs with environmental variables, including (*A*) Bio5: Maximum temperature of the warmest month, (*B*) Bio 6: Minimum temperature of the coldest month, (*C*) Bio13: Precipitation of the wettest month, and (*D*) Bio14: Precipitation of the driest month. Red and black dots indicate PAVs and SNPs, respectively. Horizontal lines show genome-wide significance thresholds with Bonferroni correction for SNPs (black) and PAVs (red).

Overall, there are many fewer PAVs than SNPs in the genome (124905 vs. 8033502). The average distance between two any SNPs is 14.4 bp and the average distance between PAVs is 771 bp, and the average distance from any PAV to the nearest SNP is 332 bp so that LD between any two SNPs tends to be higher than LD between any two PAVs. To determine the extent to which we identified novel variation by including PAVs in our analysis, we examined how well a SNP represents each strongly correlated PAV. To this end, for each PAV with a GWAS P value <0.001, we identified the SNP with the highest LD (r^2) within a 10-kb window. The mean and SD of r^2 between a PAV with this significance level and the most correlated SNP for each of the four traits were as follows: Bio5: mean $= 0.70$, SD $= 0.25$; Bio6: mean $= 0.75$, SD $= 0.24$; Bio13: mean $= 0.70$, SD $= 0.23$; and Bio14: mean $= 0.76$, SD $= 0.25$. The full distributions of these most correlated SNPs are shown in supplementary figure S11, Supplementary Material online. The set of PAVs we identified may be useful specifically to reveal missing heritability in GWAS and for identifying causative variants (when these turn out to be SVs) even in cases where they are tagged by SNPs.

Discussion

In this study, we identified SVs among 1,301 diverse *A. thaliana* accessions using publicly available NGS data and we make the entire set available, which includes 124,905

PAVs, 25,061 tandem duplications, and 5,474 inversions. For researchers interested in using these data, we note that the data set includes information about the strength of evidence for SVs based on the number of reads supporting each SV as well as the type of evidence (split read and/or discordant reads), which could be used for further filtering. Here, we focused our population genetic analyses on PAVs because these were called with the highest accuracy based on our analyses.

Given that the genic PAVs we identify are expected to often disrupt gene function, our results were consistent with the idea that a genome is made up of an essential "core" set of genes as well as a "flexible" set of genes, which may be dispensable depending on the specific local selection pressures faced by an organism. In our analyses, we found evidence for purifying selection based on the low occurrence of PAVs in genic compared with nongenic regions, and the enrichment of core housekeeping genes in genomic regions that were most deficient in PAVs (table 1). Compared with tissue-specific genes, housekeeping genes generally evolve more slowly, have lower Ka/Ks ratio (the rate ratio of nonsynonymous to synonymous substitutions), and tend to evolve under strong purifying selection (Zhang and Li 2004). Among the regions deficient in PAVs, classical housekeeping genes (Zhang and Li 2004) including tRNAs, ribosomal proteins, elongation factors, NAD+ transporters, transcription

factors, homeobox proteins were highly enriched (supplementary tables S2 and S3, Supplementary Material online).

In contrast to this, when we used a traditional method to detect evidence of purifying selection on weakly deleterious alleles based on an excess of low-frequency variants in the allele frequency spectrum, we found no such evidence (supplementary fig. S6, Supplementary Material online). This approach is based on the expectation that selection limits the spread of deleterious variants in the population. In our analysis, we found the opposite trend, in which genic PAVs were underrepresented in the lowest frequency bin relative to nongenic variants. This may be because PAVs in genes tend to be strongly deleterious rather than mildly deleterious. Although the efficiency of selection is expected to be reduced on weakly deleterious variants in selfing species (Schemske and Lande 1985; Charlesworth et al. 1990; Charlesworth and Wright 2001) and this has been shown empirically in A. thaliana and related plant species (Slotte et al. 2013; Slotte 2014; Laenen et al. 2018), purifying selection on loci that are lethal in the homozygous state is expected to be strong in selfing species because these variants are quickly exposed to selection (Stebbins 1950; Shields 1982; Schemske and Lande 1985; Glémin 2007; Charlesworth and Willis 2009; Arunkumar et al. 2015). Further, the overall pattern that we observe in which genic regions are more highly conserved than nongenic regions is consistent with the expected effects of background selection (Charlesworth et al. 1993) and that linkage disequilibrium is increased within genes in A. thaliana (Berger et al. 2015). Taken together our results suggest that newly arising genic PAV mutations tend to be strongly deleterious and therefore often evolve under purifying selection in A. thaliana, whereas those that remain in the population do not evolve under strong purifying selection relative to intergenic PAVs.

However, some categories of genes were much more likely to contain PAVs relative to the genomic background. We found that defense-related genes (R-genes, secondary metabolites) and F-box genes have an excess of PAVs compared with other genes. R-genes and F-box genes are multigene families known to be rapidly evolving (Xu et al. 2009; Yang et al. 2013). Further, plants produce a massive number of metabolites and only a few of these are primary (those common to all organisms); others are known as secondary metabolites (Pichersky and Gang 2000; Labarrere et al. 2019). Many secondary metabolites are thought to be involved in defense against herbivores and pathogens (Isah 2019) and previous work has shown that a subset of these regions that are involved in the production of glucosinolates appears to be rapidly evolving (Kliebenstein et al. 2001; Kliebenstein 2004) that multiple losses have occurred over evolutionary time across Eurasia. (Katz et al. 2020). Our results agree with these findings, and provide evidence that these genes are not only rapidly evolving but also belong to the dispensable genome and carry high levels of structural variation.

We further found that PAVs in defense response genes tend to be present at more intermediate frequencies within populations compared with the genomic background (fig. 4), suggesting they are maintained in populations by some form

of balancing selection. This maintenance of polymorphism can involve spatially- or temporally varying selection and/or fitness trade-offs. The classical gene-for-gene model posits that a specific gene (R-gene) from the host is involved in recognition of a specific pathogen avirulence (avr) gene (Flor 1971). In many cases, the gene-for-gene model in plants can explain maintenance of polymorphism in the evolution of disease resistance genes (Bergelson et al. 2001; Tian et al. 2002; Gao et al. 2009; Karasov et al. 2014). However, it has been shown that R-gene polymorphism in A. thaliana is sometimes more complex (Karasov et al. 2014; Laflamme et al. 2020). The panNLRome in A. thaliana recently showed that although there is high variation in NLR genes this diversity is not unlimited (Van de Weyer et al. 2019). Trade-offs between growth and herbivore or pathogen resistance (Coley et al. 1985; Walling 2009; Huot et al. 2014) also likely contribute to the maintenance of polymorphisms in a population. For example, a hyperactive ACD6 allele is known to strongly increase resistance of A. thaliana to a broad range of pathogens but alters its growth dramatically (Todesco et al. 2010). An additional mechanism that acts to maintain variation in R-genes involves interactions between incompatibility alleles. Several gene combinations, especially for disease resistance genes, are reported as lethal for the plants (Bomblies et al. 2010; Smith et al. 2011; Chae et al. 2014; Tran et al. 2017). This phenomenon involves autoimmunity and hybrid necrosis, which is the opposite of heterosis or hybrid vigor (Chae et al. 2014; Tran et al. 2017). Further, accumulating genomic data from related species suggests that balancing selection may be common in defense response loci in other species as well (Koenig et al. 2019).

The effects of SVs on fitness in the wild may change across environments or over time, as has been shown in previous focused studies (Gao et al. 2009; Huard-chauveau et al. 2013). Recently, it was further shown that loss of function variants may contribute to local adaptation and phenotypic diversity in A. thaliana (Monroe et al. 2018; Xu et al. 2019). Consistent with this, we found strong associations between several genic PAVs and environmental variables, including several involved in response to vernalization with minimum temperature in the coldest month ELF9 (AT5G16260), AGL31(AT5G65050), and MAF3 (AT5G65050) (fig. 6). This is reminiscent of the patterns observed in natural populations at the well-known FRIGIDA locus, where loss of function alleles obviate the need for cold exposure before flowering (Le Corre et al. 2002; Stinchcombe et al. 2004; Shindo et al. 2005; Zhang and Jiménez-Gómez 2020).

The set of SVs identified and genotyped here will be useful alone or in combination with available SNP data to investigate A. thaliana evolution and trait architecture using GWAS or recombinant populations.

Materials and Methods

Samples

We retrieved Illumina short-read data for 1,327 samples from four studies (Alonso-Blanco et al. 2016; Durvasula et al. 2017; Zou et al. 2017; Fulgione et al. 2018). We excluded 26 samples

from the analysis due to low data quality, so that the final data set included 1,301 samples. Besides publicly available data, we also sequenced one sample (PRJNA638240) from Cape Verde Island (Cvi-0) with Pacific Biosciences long-read sequencing technology (PacBio) (supplementary table S1, Supplementary Material online). For this, we sterilized and sowed Cvi-0 seeds on MS (Murashige & Skoog) media supplemented with sucrose. Then we stratified seeds for 6 days. Later we moved seeds to a growth chamber for 2 weeks. Finally, we transferred plants to the dark, where they remained for 3 days before DNA extraction using a NucleoSpin plant II protocol. After quality checks, size selection was performed with a Blue Pippin (Sage Science) (>10 kb) and DNA sequencing was performed with PacBio RS II. DNA extraction, size selection, and PacBio sequencing was performed at Max Planck Genome Center in Cologne, Germany.

Performance Comparison of SV Callers

We tested three popular tools designed for SV identification from short-read data to compare their performance. These included LUMPYEXPRESS (v0.2.13) (Layer et al. 2014), DELLY (v0.8.1) (Rausch et al. 2012), and PINDEL (v0.2.5b8) (Ye et al. 2009). LUMPYEXPRESS is an automated breakpoint detection tool for standard analysis which internally uses LUMPY (v0.2.13) (Layer et al. 2014). Although LUMPYEXPRESS (v0.2.13) uses three different sources of information including Read Pair (RP), Split Read (SR), and Read Depth (RD) information, DELLY (v0.8.1) uses only RP and SR information and PINDEL (v0.2.5b8) relies on only SR information to identify structural variations.

To identify the best software for calling SVs in short-read data, we used two approaches: a simulation-based approach and a comparison to calls from long-read data. For the simulation approach, we introduced 1,000 structural variations (maximum length 10 kb and SNP mutations frequency 0.1) with SURVIVOR (v1.0.6) (Jeffares et al. 2017) on the *A. thaliana* TAIR10 genome for each kind including deletions, tandem duplications, and inversions. We simulated NGS reads by WGSIM (v1.9) (Li et al. 2009) using the following parameters (-h –N 10000000 -1 150 -2 150) from SVs introduced reference. Later, we mapped the simulated reads to original TAIR10 by BWA-MEM (v0.7.17) (Li and Durbin 2009) with default parameters. Finally, we called SVs using LUMPYEXPRESS (v0.2.13), DELLY (v0.8.1), and PINDEL (v0.2.5b8) with default parameters.

For the PacBio approach, we first mapped PacBio reads to TAIR10 using NGMLR (v0.2.7) (Sedlazeck et al. 2018) with default settings. After mapping, we converted SAM (the Sequence Alignment/Map format) files to BAM (Binary Alignment/Map format) files by SAMTOOLS (v1.8) (Li et al. 2009) with the parameter settings "view -Sb." Then, we employed SNIFFLES (v1.0.8) (Sedlazeck et al. 2018) the settings "–genotype, -l 50" to call SVs. Finally, we compared the performance of the three short-read tools based on SNIFFLES (v1.0.8) calls. For both approaches, we considered a minimum of 50% reciprocal overlap to represent true positives then we calculated their sensitivity (True Positives/[True Positives + False Negatives]) and precision (True Positives/[True Positives + False Positives]).

Calling of SVs and SNPs in Natural Populations

We identified SVs among 1,301 accessions using the pipeline (supplementary fig. S1, Supplementary Material online, https://github.com/HancockLab/SVS_A.thaliana) that consists of LUMPYEXPRESS (v0.2.13), SVTYPER (v0.7.0) (Chiang et al. 2015), and SVTOOLS (v0.4.0) (https://doi.org/10.5281/zenodo.1442926, last accessed February 10, 2019). Our pipeline is forked and modified from (https://github.com/arq5x/lumpy-sv, last accessed December 3, 2018). After calling SVs (deletions, tandem duplications, and inversions), we performed two genotyping steps. First, we conducted individual genotyping with SVTYPER (v0.7.0), and then we conducted joint genotyping with SVTOOLS (v0.4.0). This final genotyping allowed us to differentiate missing genotypes from matches to the reference. The discovery set can have an influence on the power to identify variants across populations. Therefore, we included all individuals in the discovery panel to reduce the false-negative rate overall; however, this likely results in a bias towards higher SV discovery rates in deeply sampled populations and regions.

Our pipeline failed for 26 samples, which were thus excluded from the analysis. There was no clear reason why these samples failed, and as they are not clustered into any one population removing them is not expected to bias the data set. These samples include (collection location, sequencing facility) are BRR4 (MPI Tübingen), BRR12 (MPI Tübingen), BRR57 (MPI Tübingen), BRR107 (MPI Tübingen), Bur-0 (IRL, Mott), Can-0 (ESP, Mott), Dem-4 (Salk), KBS-Mac-74 (MPI Tübingen), LI-SET-036 (MPI Tübingen), MSGA-61 (MPI Tübingen), Oy-0 (Mott), Sf-2 (ESP, Mott), Paw-13 (MPI Tübingen), Paw-20 (MPI Tübingen), Rsch-4 (RUS, Mott), Yng-53 (MPI Tübingen), Tsu-0 (Mott), Uod-7 (AUT, Salk), Yng-4 (MPI Tübingen), Zu-0 (SUI, Mott), 11PNA1.14 (MPI Tübingen), 328PNA062 (MPI Tübingen), 87 (CHN, Yangtze Genomes), 36-31 (CHN, Yangtze Genomes), 36-17 (CHN, Yangtze Genomes), and 27-9(CHN, Yangtze Genomes).

The VCF file used for subsequent population genetic analysis was generated by setting an upper size limit of 10 kb, including only polymorphic PAVs. Therefore, the sizes of variants included in this analysis range from 1 bp to 10 kb. The nature of short-read data prevented us from identifying large SVs. We provide two VCF files: one with filtering applied with only PAV variants included, and one with all raw calls including PAVs, inversions, and tandem duplications.

In addition to PAVs, we also called biallelic SNPs for all 1,301 samples following the same pipeline we used previously (Durvasula et al. 2017; Fulgione et al. 2018).

Examining Population Structure Using PAVs

We produced a whole-genome neighbor-joining (NJ) tree in R with the ape (v5.0) package (Paradis and Schliep 2019) using the same set of samples used in (Durvasula et al. 2017) except for herbarium samples, which were sequenced with single-end sequencing data and could therefore not be included in our SV-calling pipeline.

Identifying Genomic Hotspots and Conserved Regions

We defined genic and nongenic regions based on TAIR10 annotation. Any region that overlaps with a gene is treated as a genic region and the rest is treated as a nongenic region. We compared genic and nongenic regions to see if there is any difference based on PAVs. We found more PAVs in intergenic regions compared with genic regions. To test the significance of finding a higher proportion of PAVs in intergenic regions than genic regions, we used the χ^2 test. We set the expected genic region to 60,104,494 bases (sum of all genes length without overlap) and expected intergenic region to 59,041,854 bases (sum of all nongenic regions). If the distribution of PAVs throughout the genome was random, we would not expect to see any difference between genic and intergenic regions. Our observations for genic regions and intergenic regions overlapping with PAVs were 8,920,166 bases and 30,766,220 bases, respectively.

Later we focused on the genic regions to be able to see the gene sets with excesses or deficits of PAVs. Genes that overlap with PAVs were extracted and performed enrichment analysis. Same enrichment analysis was also done for the genes that never overlap with a PAV. We tested for GO term (Carbon et al. 2019) enrichment analysis using one-tailed FET with Benjamini correction as well as Gowinda (v1.0), a more conservative test that takes gene size and clustering into account using a permutation-based approach to assess significance (Kofler et al. 2012).

We calculated unfolded SFS for each population and genomic class (whole genome, intergenic, genic, and defense genes). To produce the frequency spectrum for each population, we calculated proportion of variants that were present once in the population (i.e., singletons), twice (i.e., doubletons), etc. and these are plotted in figure 4 and supplementary figures S7 and S8, Supplementary Material online. The frequency spectra were polarized to a consensus ancestral genome. This consensus ancestral genome was created from five accessions chosen to represent distinct diverged *A. thaliana* lineages including (Lebanon [Qar-8a], Italy [Etna-0], Madeira [12761], North Middle Atlas [22000], and High Atlas [18511]) based on analyses in previous study (Durvasula et al. 2017). This set represents the major "relict" lineages of *A. thaliana*. We limited the set to five samples because only a single "relict" sample was available from the Levant (Lebanon) region. For each PAV, we calculated allele frequency among these five accessions. The highest frequency allele for each locus was used as the ancestral state. NA was assigned to the ancestral state where data were missing for more than one individual or when the allele frequency is equal to 0.5.

The samples that we retrieved from the 1001 Genomes project were separated into populations based on their admixture groups (Alonso-Blanco et al. 2016). Other samples were grouped based on their geographical origins including Madeira, China (Yangtze River Basin and North-Western China), and Morocco (High Atlas, South Middle Atlas, North Middle Atlas, and Riff).

We found a clear shift to the intermediate frequencies for PAVs on defense response genes. To test whether there was a statistically significant excess of intermediate frequency PAVs in defense response genes relative to the genome-wide distribution, we calculated the Tajima's *D* statistic for the set of defense genes in each population and compared this with the Tajima's *D* statistic for the entire genome. We found that for each population, Tajima's *D* was more positive for defense genes compared with the total set of genes. To assess significance for this result, we randomly subsampled genic PAVs to match the number of defense response PAVs 100,000 times and calculated an empirical *P* value based on the distribution of Tajima's *D* in the resampled data sets. We further examined evidence for balancing selection using BetaScan, which relies on a signal of high variation at the haplotype level (Siewert and Voight 2017, 2020). After employment of BetaScan, we extracted 5% extreme tail of highest β scored SNPs and performed enrichment analysis with FET. We found significant enrichment of defense response genes for all groups.

Correlation with Environmental Variables (GWAS)

We used the linear mixed model association method, GEMMA (v0.98.1) (Zhou and Stephens 2012), to assess correlation between PAVs and SNPs with environmental variables. Environmental variables were obtained as geoTiff files from WorldClim2 (Fick and Hijmans 2017). We extracted data for four environmental variables (geoTiff files) using the raster package in R. Then we used geo-referencing information to compute the values for each accession with the "extract" function from the raster package. These computed values were treated as phenotypes for association mapping. Individual samples that are genetically and environmentally divergent from the bulk of samples violate assumptions of the linear mixed model approach. Thus, the following divergent samples were removed before conducting GWAS (6911, 9762, 9764, 10024, 35520, 12761, 12672, 12763, 12908, 22017, 22019, 22022, 22638, and 27153) to avoid effects of outliers in the LMM analysis. To prepare files for GEMMA (v0.98.1), we converted the VCF file to plink file with VCFTOOLS (v0.16) (Danecek et al. 2011) (–plink) and then ran PLINK (v2.0) (Purcell et al. 2007) to create a bed file (–make-bed). Next, we used GEMMA (v0.98.1) to calculate the kinship matrix (-gk 1, -maf 0.1) and ran GEMMA (v0.98.1) under the linear mixed model with a minor allele frequency cut-off of 10% (-lmm 2, -maf 0.1). The minor allele frequency cutoff acts to remove outliers that could otherwise drive signals in the analysis. To prioritize candidate functional PAVs, we focused on genes within 10 kb of a given climate-associated PAV. In order to examine how well SNPs could represent climate-correlated PAVs, we estimated linkage disequilibrium (LD) between all SNPs within 10 kb of a PAVs with a GWAS *P* value <0.001 using PLINK (v2.0) with the command –r2 –ld-window-kb 10 –ld-window 999999 and filtered for the PAVs of interest for each environmental GWAS. For each PAV, we identified the SNP with the highest r^2, created histograms for these, and used the distribution to estimate the mean and SD of r^2 between these PAVs and the most correlated SNPs.

Supplementary Material

Acknowledgments

Data Availability

Raw PacBio data for Cvi-0 uploaded in FASTQ format to NCBI (PRJNA638240). Structural variants and indel calls for the 1,301 samples have been uploaded to European Variation Archive (PRJEB38975) and all SV calls are available in the ENSEMBL genome browser. Code used for analysis is available at https://github.com/HancockLab/SVs_in_1301_A.thaliana_lines.

References

Abel HJ, Larson DE, Regier AA, Chiang C, Das I, Kanchi KL, Layer RM, Neale BM, Salerno WJ, Reeves C, et al. 2020. Mapping and characterization of structural variation in 17,795 human genomes. *Nature* 583(7814):83–89.

Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, Suresh H, Ramakrishnan S, Maumus F, Ciren D, et al. 2020. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell* 182(1):145–161.e23.

Alonso-Blanco C, Andrade J, Becker C, Bemm F, Bergelson J, Borgwardt KMM, Cao J, Chae E, Dezwaan TMM, Ding W, et al. 2016. 1,135 Genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* 166(2):481–491.

Arunkumar R, Ness RW, Wright SI, Barrett SCH. 2015. The evolution of selfing is accompanied by reduced efficacy of selection and purging of deleterious mutations. *Genetics* 199(3):817–829.

Bagniewska-Zadworna A, Barakat A, Łakomy P, Smoliński DJ, Zadworny M. 2014. Lignin and lignans in plant defence: insight from expression profiling of cinnamyl alcohol dehydrogenase genes during development and following fungal infection in *Populus*. *Plant Sci.* 229:111–121.

Bergelson J, Kreitman M, Stahl EA, Tian D. 2001. Evolutionary dynamics of plant R-genes. *Science* 292(5525):2281–2285.

Berger S, Schlather M, de los Campos G, Weigend S, Preisinger R, Erbe M, Simianer H. 2015. A scale-corrected comparison of linkage disequilibrium levels between genic and non-genic regions. *PLoS One* 10(10):e0141216–e0141219.

Bomblies K, Yant L, Laitinen RA, Kim ST, Hollister JD, Warthmann N, Fitz J, Weigel D. 2010. Local-scale patterns of genetic variability, outcrossing, and spatial structure in natural stands of *Arabidopsis thaliana*. *PLoS Genet.* 6(3):e1000890–e1000914.

Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C, et al. 2011. Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat Genet.* 43(10):956–965.

Carbon S, Douglass E, Dunn N, Good B, Harris NL, Lewis SE, Mungall CJ, Basu S, Chisholm RL, Dodson RJ, et al. 2019. The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.* 47:D330–D338.

Chae E, Bomblies K, Kim ST, Karelina D, Zaidem M, Ossowski S, Martín-Pizarro C, Laitinen RAE, Rowan BA, Tenenboim H, et al. 2014. Species-wide genetic incompatibility analysis identifies immune genes as hot spots of deleterious epistasis. *Cell* 159(6):1341–1351.

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* 134(4):1289–1303.

Charlesworth D, Morgan MT, Charlesworth B. 1990. Inbreeding depression, genetic load and the evolution of outcrossing rates in a multi-locus system with no linakge. *Evolution* 44(6):1469–1489.

Charlesworth D, Willis JH. 2009. The genetics of inbreeding depression. *Nat Rev Genet.* 10(11):783–796.

Charlesworth D, Wright SI. 2001. Breeding systems and genome evolution. *Curr Opin Genet Dev.* 11(6):685–690.

Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, Marth GT, Quinlan AR, Hall IM. 2015. SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat Methods.* 12(10):966–968.

Choi JY, Lye ZN, Groen SC, Dai X, Rughani P, Zaaijer S, Harrington ED, Juul S, Purugganan MD. 2020. Nanopore sequencing-based genome assembly and evolutionary genomics of circum-basmati rice. *Genome Biol.* 21(1):27.

Coley PD, Bryant JP, Chapin FS. 1985. Resource availability and plant antiherbivore defense. *Science* 230(4728):895–899.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. *Bioinformatics* 27(15):2156–2158.

Durmaz E, Benson C, Kapun M, Schmidt P, Flatt T. 2018. An inversion supergene in *Drosophila* underpins latitudinal clines in survival traits. *J Evol Biol.* 31(9):1354–1364.

Durvasula A, Fulgione A, Gutaker RM, Alacakaptan SI, Flood PJ, Neto C, Tsuchimatsu T, Burbano H. A, Picó FX, Alonso-Blanco C, et al. 2017. African genomes illuminate the early history and transition to selfing in *Arabidopsis thaliana*. *Proc Natl Acad Sci U S A.* 114(20):5213–5218.

Duverger O, Morasso MI. 2008. Role of homeobox genes in the patterning, specification, and differentiation of ectodermal appendages in mammals. *J Cell Physiol.* 216(2):337–346.

Eisfeldt J, Pettersson M, Vezzi F, Wincent J, Käller M, Gruselius J, Nilsson D, Syk Lundberg E, Carvalho CMB, Lindstrand A. 2019. Comprehensive structural variation genome map of individuals carrying complex chromosomal rearrangements. *PLoS Genet.* 15(2):e1007858–e1007926.

Fick SE, Hijmans RJ. 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int J Climatol.* 37(12):4302–4315.

Flor HH. 1971. Current status of the gene-for-gene concept. *Annu Rev Phytopathol.* 9(1):275–296.

Fransz P, Linc G, Lee C-R, Aflitos SA, Lasky JR, Toomajian C, Hoda A, Peters J, van Dam P, Ji X, et al. 2016. Molecular, genetic and evolutionary analysis of a paracentric inversion in *Arabidopsis thaliana*. *Plant J.* 88(2):159–178.

Fuentes RR, Chebotarov D, Duitama J, Smith S, De la Hoz JF, Mohiyuddin M, Wing RA, McNally KL, Tatarinova T, Grigoriev A, et al. 2019. Structural variants in 3000 rice genomes. *Genome Res.* 29(5):870–880.

Fulgione A, Koornneef M, Roux F, Hermisson J, Hancock AM. 2018. Madeiran *Arabidopsis thaliana* reveals ancient long-range colonization and clarifies demography in Eurasia. *Mol Biol Evol.* 35(3):564–574.

Gao L, Roux F, Bergelson J. 2009. Quantitative fitness effects of infection in a gene-for-gene system. *New Phytol.* 184(2):485–494.

Glémin S. 2007. Mating systems and the efficacy of selection at the molecular level. *Genetics* 177(2):905–916.

González J, Karasov TL, Messer PW, Petrov DA. 2010. Genome-wide patterns of adaptation to temperate environments associated with transposable elements in Drosophila. PLoS Genet. 6(4):e1000905–e1000935.

Gu X, Le C, Wang Y, Li Z, Jiang D, Wang Y, He Y. 2013. Arabidopsis FLC clade members form flowering-repressor complexes coordinating responses to endogenous and environmental cues. Nat Commun. 4:1947.

Ho SS, Urban AE, Mills RE. 2020. Structural variation in the sequencing era. Nat Rev Genet. 21(3):171–189.

Hollister JD, Smith LM, Guo YL, Ott F, Weigel D, Gaut BS. 2011. Transposable elements and small RNAs contribute to gene expression divergence between Arabidopsis thaliana and Arabidopsis lyrata. Proc Natl Acad Sci U S A. 108(6):2322–2327.

Huang DW, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 4(1):44–57.

Huard-chauveau C, Perchepied L, Debieu M, Rivas S, Kroj T, Kars I, Bergelson J, Roux F, Roby D. 2013. An atypical kinase under balancing selection confers broad-spectrum disease resistance in Arabidopsis. PLoS Genet. 9(9):e1003766.

Huber CD, Nordborg M, Hermisson J, Hellmann I. 2014. Keeping it local: evidence for positive selection in Swedish Arabidopsis thaliana. Mol Biol Evol. 31(11):3026–3039.

Huot B, Yao J, Montgomery BL, He SY. 2014. Growth-defense tradeoffs in plants: a balancing act to optimize fitness. Mol Plant. 7(8):1267–1287.

Isah T. 2019. Stress and defense responses in plant secondary metabolites production. Biol Res. 52(1):39.

Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, Balloux F, Dessimoz C, Bähler J, Sedlazeck FJ. 2017. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. Nat. Commun. 8:1–11.

Kapun M, Fabian DK, Goudet J, Flatt T. 2016. Genomic evidence for adaptive inversion clines in Drosophila melanogaster. Mol Biol Evol. 33(5):1317–1336.

Karasov TL, Kniskern JM, Gao L, Deyoung BJ, Ding J, Dubiella U, Lastra RO, Nallu S, Roux F, Innes RW, et al. 2014. The long-term maintenance of a resistance polymorphism through diffuse interactions. Nature 512(7515):436–440.

Katz E, Bagaza C, Holden S, Angelovici R, Kliebenstein DJ. 2020. Genetic variation, environment and demography intersect to shape Arabidopsis defense metabolite variation across Europe. bioRxiv.

Kliebenstein DJ. 2004. Secondary metabolites and plant/environment interactions: a view through Arabidopsis thaliana tinged glasses. Plant Cell Environ. 27(6):675–684.

Kliebenstein DJ, Kroymann J, Brown P, Figuth A, Pedersen D, Gershenzon J, Mitchell-Olds T. 2001. Genetic control of natural variation in Arabidopsis glucosinolate accumulation. Plant Physiol. 126(2):811–825.

Koenig D, Hagmann J, Li R, Bemm F, Slotte T, Neuffer B, Wright SI, Weigel D. 2019. Long-term balancing selection drives evolution of immunity genes in Capsella. Elife 8:e43606.

Kofler R, Schlötterer C, Populationsgenetik I, Vienna V, Wien A. 2012. Gowinda: unbiased analysis of gene set enrichment for genome-wide association studies. Bioinformatics 28(15):2084–2085.

Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. 2019. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. Genome Biol. 20(1):8–11.

Labarrere B, Prinzing A, Dorey T, Chesneau E, Hennion F. 2019. Variations of secondary metabolites among natural populations of sub-Antarctic ranunculus species suggest functional redundancy and versatility. Plants 8(7):234–223.

Laenen B, Tedder A, Nowak MD, Toräng P, Wunder J, Wötzel S, Steige KA, Kourmpetis Y, Odong T, Drouzas AD, et al. 2018. Demography and mating system shape the genome-wide impact of purifying selection in Arabis alpina. Proc Natl Acad Sci U S A. 115(4):816–821.

Laflamme B, Dillon MM, Martel A, Almeida RND, Desveaux D, Guttman DS. 2020. The pan-genome effector-triggered immunity landscape of a host-pathogen interaction. Science 367(6479):763–768.

Layer RM, Chiang C, Quinlan AR, Hall IM. 2014. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 15(6):R84–R19.

Le Corre V, Roux F, Reboud X. 2002. DNA polymorphism at the FRIGIDA gene in Arabidopsis thaliana: extensive nonsynonymous variation is consistent with local selection for flowering time. Mol Biol Evol. 19(8):1261–1271.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14):1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25(16):2078–2079.

Lowry DB, Willis JH. 2010. A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. PLoS Biol. 8(9):e1000500–e1000514.

Lye ZN, Purugganan MD. 2019. Copy number variation in domestication. Trends Plant Sci. 24(4):352–365.

Monroe JG, Powell T, Price N, Mullen JL, Howard A, Evans K, Lovell JT, McKay JK. 2018. Drought adaptation in Arabidopsis thaliana by extensive genetic loss-of-function. Elife 7:1–18.

Morgan AP, Gatti DM, Najarian ML, Keane TM, Galante RJ, Pack AI, Mott R, Churchill GA, de Villena FP-M. 2017. Structural variation shapes the landscape of recombination in mouse. Genetics 206(2):603–619.

Nielsen R. 2005. Molecular signatures of natural selection. Annu Rev Genet. 39(1):197–218.

Paradis E, Schliep K. 2019. Ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35(3):526–528.

Pichersky E, Gang DR. 2000. Genetics and biochemistry of secondary metabolites in plants: an evolutionary perspective. Trends Plant Sci. 5(10):439–445.

Prinzenberg AE, Campos-Dominguez L, Kruijer W, Harbinson J, Aarts MGM. 2020. Natural variation of photosynthetic efficiency in Arabidopsis thaliana accessions under low temperature conditions. Plant Cell Environ. 43(8):2000–2013.

Provart NJ, Alonso J, Assmann SM, Bergmann D, Brady SM, Brkljacic J, Browse J, Chapple C, Colot V, Cutler S, et al. 2016. 50 years of Arabidopsis research: highlights and future directions. New Phytol. 209(3):921–944.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, De Bakker PIW, Daly MJ, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 81(3):559–575.

Ratcliffe OJ, Kumimoto RW, Wong BJ, Riechmann JL. 2003. Analysis of the Arabidopsis MADS AFFECTING FLOWERING gene family: MAF2 prevents vernalization by short periods of cold. Plant Cell 15(5):1159–1169.

Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 28(18):i333–i339.

Rifkind RA, Marks PA, Bank A, Terada M, Maniatis GM, Reuben R, Fibach E. 1976. Erythroid differentiation and the cell cycle: some implications from murine foetal and erythroleukemic cells. Ann Immunol. 127(6):887–893.

Rowan BA, Heavens D, Feuerborn TR, Tock AJ, Henderson IR, Weigel D. 2019. An ultra high-density Arabidopsis thaliana crossover map that refines the influences of structural variation and epigenetic features. Genetics 213(3):771–787.

Schemske DW, Lande R. 1985. The evolution of self-fertilization and inbreeding depression in plants. II. Empirical observations. Evolution 39(1):41–52.

Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, Von Haeseler A, Schatz MC. 2018. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods. 15(6):461–468.

Shields WM. 1982. Philopatry, inbreeding, and the evolution of sex. New York: State University of New York Press.

Shindo C, Aranzana MJ, Lister C, Baxter C, Nicholls C, Nordborg M, Dean C. 2005. Role of FRIGIDA and FLOWERING LOCUS C in determining variation in flowering time of Arabidopsis. *Plant Physiol.* 138(2):1163–1173.

Siewert KM, Voight BF. 2017. Detecting long-term balancing selection using allele frequency correlation. *Mol Biol Evol.* 34(11):2996–3005.

Siewert KM, Voight BF. 2020. BetaScan2: standardized statistics to detect balancing selection utilizing substitution data. *Genome Biol Evol.* 12(2):3873–3877.

Slotte T. 2014. The impact of linked selection on plant genomic variation. *Brief Funct Genomics Proteomics.* 13(4):268–275.

Slotte T, Hazzouri KM, Ågren JA, Koenig D, Maumus F, Guo YL, Steige K, Platts AE, Escobar JS, Newman LK, et al. 2013. The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. *Nat Genet.* 45(7):831–835.

Smith LM, Bomblies K, Weigel D. 2011. Complex evolutionary events at a tandem cluster of *Arabidopsis thaliana* genes resulting in a single-locus genetic incompatibility. *PLoS Genet.* 7(7):e1002164–e1002214.

Somerville C, Koornneef M. 2002. A fortunate choice: the history of *Arabidopsis* as a model plant. *Nat Rev Genet.* 3(11):883–889.

Soyk S, Lemmon ZH, Sedlazeck FJ, Jiménez-Gómez JM, Alonge M, Hutton SF, Van Eck J, Schatz MC, Lippman ZB. 2019. Duplication of a domestication locus neutralized a cryptic variant that caused a breeding barrier in tomato. *Nat Plants.* 5(5):471–479.

Springer NM, Ying K, Fu Y, Ji T, Yeh CT, Jia Y, Wu W, Richmond T, Kitzman J, Rosenbaum H, et al. 2009. Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genet.* 5(11):e1000734.

Stebbins CLJ. 1950. Variation and evolution in plants. London: Oxford University Press.

Stinchcombe JR, Weinig C, Ungerer M, Olsen KM, Mays C, Halldorsdottir SS, Purugganan MD, Schmitt J. 2004. A latitudinal cline in flowering time in *Arabidopsis thaliana* modulated by the flowering time gene FRIGIDA. *Proc Natl Acad Sci U S A.* 101(13):4712–4717.

Studer A, Zhao Q, Ross-Ibarra J, Doebley J. 2011. Identification of a functional transposon insertion in the maize domestication gene *tb1. Nat Genet.* 43(11):1160–1163.

Sturtevant AH. 1926. A crossover reducer in *Drosophila melanogaster* due to inversion of a section of the third chromosome. *Biol Zbl.* 46:697–703.

Sun S, Zhou Y, Chen J, Shi J, Zhao H, Zhao H, Song W, Zhang M, Cui Y, Dong X, et al. 2018. Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. *Nat Genet.* 50(9):1289–1295.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123(3):585–595.

Tian D, Araki H, Stahl E, Bergelson J, Kreitman M. 2002. Signature of balancing selection in *Arabidopsis. Proc Natl Acad Sci U S A.* 99(17):11525–11530.

Todesco M, Balasubramanian S, Hu TT, Traw MB, Horton M, Epple P, Kuhns C, Sureshkumar S, Schwartz C, Lanz C, et al. 2010. Natural allelic variation underlying a major fitness trade-off in *Arabidopsis thaliana. Nature* 465(7298):632–636.

Tran DTN, Chung E-H, Habring-Müller A, Demar M, Schwab R, Dangl JL, Weigel D, Chae E. 2017. Activation of a plant NLR complex through heteromeric association with an autoimmune risk variant of another NLR. *Curr Biol.* 27(8):1148–1160.

Tripathy BC, Oelmüller R. 2012. Reactive oxygen species generation and signaling in plants. *Plant Signal Behav.* 7(12):1621–1633.

Van de Weyer AL, Monteiro F, Furzer OJ, Nishimura MT, Cevik V, Witek K, Jones JDG, Dangl JL, Weigel D, Bemm F. 2019. A species-wide inventory of NLR genes and alleles in *Arabidopsis thaliana. Cell* 178(5):1260–1272.e14.

Walling LL. 2009. Adaptive defense responses to pathogens and insects. *Adv Bot Res.* 51:551–612.

Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, Li M, Zheng T, Fuentes RR, Zhang F, et al. 2018. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* 557(7703):43–49.

Weischenfeldt J, Symmons O, Spitz F, Korbel JO. 2013. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat Rev Genet.* 14(2):125–138.

Xing T, Laroche A. 2011. Revealing plant defense signaling getting more sophisticated with phosphoproteomics. *Plant Signal Behav.* 6(10):1469–1474.

Xu G, Ma H, Nei M, Kong H. 2009. Evolution of F-box genes in plants: different modes of sequence divergence and their relationships with functional diversification. *Proc Natl Acad Sci U S A.* 106(3):835–840.

Xu Y-C, Niu X-M, Li X-X, He W, Chen J-F, Zou Y-P, Wu Q, Zhang YE, Busch W, Guo Y-L. 2019. Adaptation and phenotypic diversification in *Arabidopsis* through loss-of-function mutations in protein-coding genes. *Plant Cell* 31(5):1012–1025.

Yang S, Li J, Zhang X, Zhang Q, Huang J, Chen JQ, Hartl DL, Tian D. 2013. Rapidly evolving R genes in diverse grass species confer resistance to rice blast disease. *Proc Natl Acad Sci U S A.* 110(46):18572–18577.

Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 25(21):2865–2871.

Yeats TH, Rose JKC. 2008. The biochemistry and biology of extracellular plant lipid-transfer proteins (LTPs). *Protein Sci.* 17(2):191–198.

Zhang L, Jiménez-Gómez JM. 2020. Functional analysis of FRIGIDA using naturally occurring variation in *Arabidopsis thaliana. Plant J.* 103(1):154–112.

Zhang L, Li WH. 2004. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Mol Biol Evol.* 21(2):236–239.

Zhou X, Stephens M. 2012. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet.* 44(7):821–824.

Zhou Y, Minio A, Massonnet M, Solares E, Lv Y, Beridze T, Cantu D, Gaut BS. 2019. The population genetics of structural variants in grapevine domestication. *Nat Plants.* 5(9):965–979.

Zmienko A, Marszalek-Zenczak M, Wojciechowski P, Samelak-Czajka A, Luczak M, Kozlowski P, Karlowski WM, Figlerowicz M. 2020. AthCNV: a map of DNA copy number variations in the *Arabidopsis* genome. *Plant Cell* 32(6):1797–1819.

Zou Y-P, Hou X-H, Wu Q, Chen J-F, Li Z-W, Han T-S, Niu X-M, Yang L, Xu Y-C, Zhang J, et al. 2017. Adaptation of *Arabidopsis thaliana* to the Yangtze River Basin. *Genome Biol.* 18(1):239.