Genetic and environmental modulation of natural transposition enhances the adaptive potential of Arabidopsis thaliana

Pierre Baduel, Basile Leduque, José Gil, Olivier O. Loudet, Colot Vincent, Leandro Quadrana

Genetic and environmental modulation of natural transposition enhances the adaptive potential of *Arabidopsis thaliana*

Baduel P.[1], Leduque B.[1], Gil J.[1,3], Loudet O.[2], Colot V.[1*], Quadrana L.[1*],

[1] *Institut de Biologie de l'École Normale Supérieure, ENS, 46 rue d'Ulm, 75005 Paris, France*

[2] *Institut Jean-Pierre Bourgin, INRAE, Route de Saint-Cyr, 78026 Versailles, France*

[3] *Current Address: Institut Curie, 26 rue d'Ulm, 75005 Paris, France*

** correspondence to* colot@ens.psl.eu*;* guadrana@ens.psl.eu

Abstract

How species can adapt to rapid environmental changes, particularly in the absence of standing genetic variation, is poorly understood and a pressing question in the face of ongoing climate change. Here we leveraged multi-omics sequencing data available for >1,000 wild *A. thaliana* accessions to determine the contribution of transposable element (TE) mobilization to the creation of new variation. We show that overall, transposition-induced insertion substitutions occur almost at the same rate as SNPs. However, transpositional activity of individual TE families varies greatly between accessions, in association with genetic and environmental factors and we identified in addition important gene-environment interactions. Although the distribution of TE insertions across the genome is ultimately mainly shaped by purifying selection, numerous recent transposition events show signatures of positive selection in relation to local climates. Based on these findings, we generated mathematical models to forecast the impact of global warming on TE mobilization. Our models predict higher transposition activity in central parts of the species' range, which in turn will enhance the generation of potentially adaptive TE-containing alleles and may rescue native *A. thaliana* populations from extinction.

Introduction

Adaptation to rapidly changing environments in the absence of standing genetic variation is a long-standing genetic paradox (1,2). Indeed, mutations typically arise at low rates and

predominantly produce neutral variants. However, this picture ignores sequence alterations generated by the mobilization of transposable elements (TEs), which have many properties that distinguish them from “classical”, small-size mutations. First TEs constitute powerful endogenous mutagens because of their potential to mobilize across the genome and to disrupt or alter genes and their expression in multiple ways. In addition, because of their genomic dispersion, TEs provide substrates for ectopic homologous recombination, thus leading to chromosomal rearrangements (3,4). Eukaryotic TEs are divided into two broad classes: DNA transposons, that use a cut and paste mechanism for their mobilization, and retrotransposons, that move through an RNA intermediate (5). TEs can be further subdivided into superfamilies and families based on particular sequence features.

Population genomic surveys of TE insertion polymorphisms (TIPs) revealed that many TE families insert preferentially towards genes and that TE insertions are rapidly purged from gene-rich regions (6), which indicates that natural transposition generates alleles with major deleterious effects. However, epigenetic mechanisms, which include DNA methylation in plants and animals, have evolved to limit TE mobilization. In plants, DNA methylation of TE sequences encompasses the three cytosine contexts (CG, CHG, and CHH, where H is A, T, or C). In the reference plant *A. thaliana*, establishment of DNA methylation at TEs occurs in an RNA-dependent manner (RNA-directed DNA methylation or RdDM) and requires the activity of the *de novo* DNA methyltransferases DRM1/2 as well as of two plant-specific RNA Pol II derivatives, Pol IV and Pol V. TE methylation is then maintained through replication by the DNA methyltransferases CMT3 and MET1, which act respectively on CHGs and CGs, as well as by DRM1/2 and CMT2, which have mostly non-overlapping CHH targets (7).

Mutations affecting DNA methylation have varying effects on TE mobilization (8–13), indicating that additional factors control transposition. Indeed, mobilization of the LTR-retroelement *ATCOPIA78* was shown experimentally to require heat-shock in addition to impaired RdDM

activity (14), indicating that at least in this case both genetic and environmental determinants are decisive.

Although there is evidence in *A. thaliana* of significant transposition activity in nature (6,15), the factors involved are poorly known. Here, we leveraged the sampling depth of the 1001 Genomes project (1001genomes.org) to identify the major genetic and environmental determinants of recent TE mobilization in *A. thaliana*. Based on these findings, we use ecological modelling to explore the evolutionary trajectories resulting from past transposition and predict the consequences of global warming on future transposition and its capacity to generate major effect alleles available for adaptation.

Results

Recent TE mobilization at the species level

In order to evaluate recent transposition dynamics in *A. thaliana*, we used short-reads sequencing data available for 1047 Arabidopsis accessions of the 1001 Genomes project (1001genomes.org, Table S1) to search for transposable element insertion polymorphisms (TIPs). TIPs were identified using a bioinformatic pipeline (16) combining TEPID (15) and SPLITREADER (6), which efficiently detects in resequenced genomes absence of reference TE sequences and presence of non-reference TE sequences, respectively. Considering both types of variation together is essential to detect recent TE mobilization even in the Col-0 reference genome. After stringent filtering (see Methods), we recovered 23,331 high-confidence TIPs, including 21,707 non-reference TE presence variants. The latter were contributed almost entirely by the two superfamilies of LTR-retroelements *COPIA* and *GYPSY* (respectively 6,941 and 5,794) and the two superfamilies of DNA transposons *MuDR* and *hAT* (respectively 4,973 and 2,101, Fig. 1a). Of note, while presence variants are not efficiently detected by our pipeline for the DNA

transposon *HELITRON* superfamily (16) and were therefore not considered, absence variants are and together with those for *MuDR* and *hAT* make up over half of the 1,624 detected in total (Fig. 1a-S1a). TIPs are broadly distributed across the genome, with the notable exception of those produced by *GYPSY* LTR retroelements, which are enriched in pericentromeric regions. This broad distribution is in stark contrast with the relative paucity of reference TE sequences along the chromosome arms and their clustering around centromeres (Fig. 1b), which confirms previous observations obtained using a smaller number of non-reference genomes (6). Furthermore, the site frequency spectrum (SFS) of TIPs, calculated using the number of informative genomes at every site, is heavily skewed towards low values compared to bi-allelic SNPs (Fig. 1c). Remarkably, one third of TIPs are carried by less than 2 per 1000 informative genomes, >80% of which are missed in previous analyses based on around 200 genomes (6). This excess of low-frequency TIPs is stronger that the one observed for SNPs causing synonymous or intergenic SNPs even when TIPs are located further than 2kb away from the closest gene (Fig. 1d-e-S1b). TIPs within 250bp of a gene were also at much lower frequencies than missense SNPs but not as rare as nonsense (gain of premature stop codon) mutations. To estimate the deleterious effects of TIPs, we then computed the distribution of fitness effects (DFE) of each category of mutations by comparing their SFS with that of synonymous SNPs by controlling for recent demographic changes (DFE-alpha; (17), which can affect SFSs in ways that resemble selection. We find that respectively 48% and 83% of missense and nonsense SNPs are effectively deleterious ($N_e s>1$), but this proportion reaches >99% for TIPs within 250bp of a gene (Fig. 1f), indicating that almost all TE insertions within or nearby genes produce large deleterious effects.

We previously provided evidence that low-frequency TIPs reflect recent transposition events, not yet purged by natural selection (6). To demonstrate that this is the case, we considered all TE insertions shared by at least two genomes and estimated their age by first calculating for each insertion the number of SNPs accumulated within 70kb of it (see Methods). We then transformed

this number into a predicted age by applying the base mutation rate of 7E-9 per genome per year determined experimentally using mutation accumulation lines (18). Results revealed a strong positive (R2=0.4) correlation between TIP frequency and age (Fig. 1d, S1d) and indicated in addition that almost all (>99%) TIPs that are less than 1,000 years old segregate at frequencies below ≤1.5%.

Using closely related accessions (<500 SNPs, Fig. S1e), we estimated the substitution rate for TE insertions, which is almost a third (0.076 +/- 0.0012 per genome per year; Fig. 1e, see Methods) of that calculated for single bases (19). Conversely, the most divergent accessions differ by a maximum of ~730 TIPs, which is two orders of magnitude lower than expected if TIPs accumulate at the same rate as in closely-related accessions (Fig. 1f). Indeed, using synonymous SNPs to estimate the SNP substitution rate across all accessions (see Methods), we predict that >99.8% of TE insertions that occur in nature are eventually purged, a percentage higher than for missense as well as nonsense mutations (68.9% and 92.5%, respectively; Fig. S1g-h). Together, these results suggest that most TE insertions have large deleterious effects (see below) and they indicate in addition a substantial contribution (>25%) of TE mobilization to the generation of inherited mutations in nature.

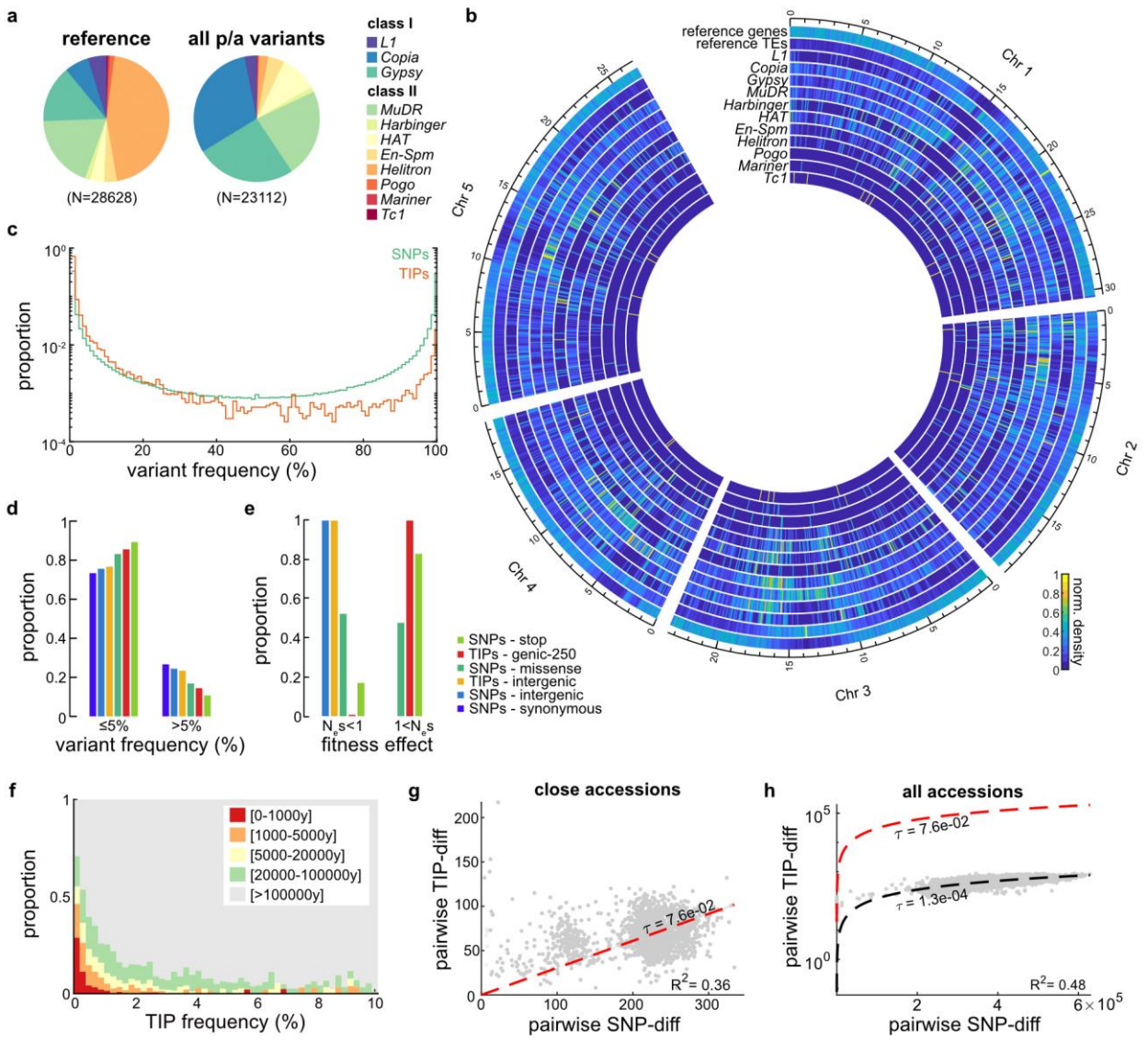

Figure 1.

(a) Contribution by superfamily to TEs annotated in the reference genome (TAIR10) and TE insertion polymorphisms (TIPs). (b) Density of the TIPs across the genome for the 11 major TE superfamilies compared to the distribution of genes or TEs annotated in the reference genome. (c) Folded frequency spectrum of TIPs and bi-allelic SNPs. (d) Proportion of each variant category at frequencies below 5% and above 5%. (e) Distribution of fitness effects of each variant category as effectively neutral (Nes<1) and deleterious (1<Nes). (f) Frequency distribution of non-private TIP by local haplotype age. (g) Pairwise differences in TIPs and SNPs for all accessions diverging by <500 SNPs. Regression line and confidence intervals are indicated in red and gray, respectively. (h) Pairwise differences in TIPs and SNPs between accessions. Regression lines between all and closely related accessions are shown in black and red, respectively.

Genetic basis of variable transposition

Principal component analysis indicated that the mutation pressure caused by recent TE mobilization (TIPs at frequencies ≤5%) and the TE families involved vary among accessions (Fig. 2a). In particular, Relicts, Asian and South-Sweden accessions present higher levels of mobilization across TE families (Fig. S2a) and suggested therefore that natural transposition is modulated by genetic and environmental factors.

To identify potential genetic modifiers of transposition activity, we carried out a genome-wide association study (GWAS) using as a proxy for recent transposition activity the quantitative trait the total number of most recent TIPs (less than 0.2% and <1000 years old or private) across all TE families (Fig. S2b-c). Remarkably, results revealed a single major peak of association (Fig. 2b) and hence a simple genetic architecture of global transposition activity. Indeed, the association peak spans the gene *NUCLEAR RNA POLYMERASE E1* (*NRPE1;* Fig. 2c), which encodes the largest subunit of RNA Pol V, essential to RdDM (20). The non-reference allele of *NRPE1,* called *NRPE1'*, was previously described to be associated with natural variation in CHH methylation (21), and is associated here with a 240% increase in transposition activity (Fig. 2d). Inspection of long-read sequencing data (22) from an accession (Sha) carrying the *NRPE1'* allele revealed additional polymorphisms beyond the SNPs and short-indels identified by the 1001 genomes project (23). Specifically, the single *NRPE1'* allele of Sha contains not only a 3bp in-frame deletion in the RNA polymerase domain, but also a 9-bp in-frame deletion in the 17aa-repeat domain (at 2:16721882) as well as three deletions in the QS tail of the C-terminal domain (CTD; 6bp at 2:16723152, 60bp at 2:16723235, and 9bp at 2:16723351 resp., Fig. 2e-S2g). These additional polymorphisms translate conceptually into a 27 amino-acids C-terminal truncation of NRPE1. In fact, the three QS deletions (delQS1-2-3) define a suballele of *NRPE1'*, which we named *NRPE1'_ΔQS* in contrast to *NRPE1'_Δrep* (Fig. 2e). Moreover, the two derived alleles of

NRPE1 resemble those produced experimentally in the reference accession Col-0 (24) and recapitulate their effect on CHH methylation, with a more pronounced loss when the QS domain is deleted (Fig. 2f-g-S2f, see Methods). Remarkably, the two truncated alleles explain on their own at least 17% of the variation in transposition at the species level (Fig. S2d, see below). Finally, GWAS performed at the TE superfamily level revealed that associations with *NRPE1* are strongest for *MuDR*, which are notably enriched among TEs directly targeted by NRPE1 but not by CMT2 (Fig. 2h, see Methods), effectively indicating causality (Fig. S2e).

To evaluate directly the impact of impaired RdDM on transposition genome-wide, we identified novel TE insertion events using TE-sequence capture (8) on pools of 1,000 seedlings derived from WT and *nrpd1* mutant parents of the Col-0 reference accession grown in control conditions. A total of 61 TE insertions were identified in the pool of WT seedlings but two third more (103) in the offspring from *nrpd1* mutant parents (Fig. 2i). The increase in transposition rate was most marked for *GYPSY* (x2.2) and *MuDR* (x1.7), notably *VANDAL* elements, while the mobilization of *COPIA* retro-elements was not affected (Fig. S2h).

Using a similar protocol, we then measured TE mobilization in the offspring of parent plants grown in control conditions (see Methods) from four natural accessions, one of them carrying the *NRPE1'_ΔQS* allele (Sha). Across the four accessions the transposition rate averaged 0.027 new insertion per genome per generation (Fig. 2j), which is remarkably close to the one we obtained based on closely related accessions growing in nature (Fig. 1g). More importantly, the highest levels of TE mobilization across superfamilies were observed in Sha (0.064 new insertion per genome per generation), confirming that the natural loss of mCHH due to the *NRPE1'_ΔQS* mutation is associated with higher TE mobilization in these accessions.

Taken together, our results demonstrate that mutations in the RdDM pathway trigger the transposition of several TEs, including *MuDR* DNA transposons, in link with the natural epigenetic

variation these mutations cause, while other environmental determinants may be necessary for the mobilization of others, notably *COPIA* retro-elements.

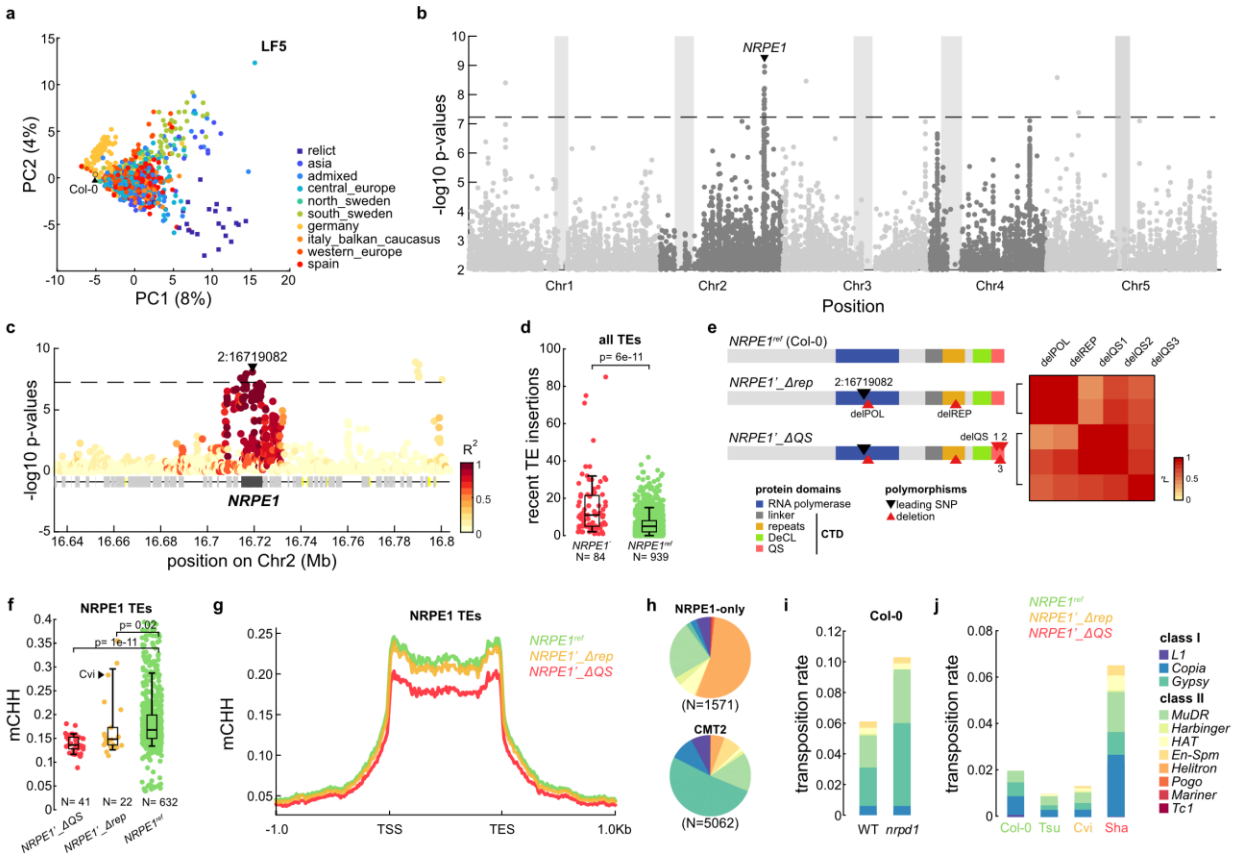

Figure 2.

(a) PCA of recent mobilome composition based on TIPs segregating at frequencies <=5% (LF5). Different genetic haplogroups are indicated in colors. (b) Manhattan plot of GWAS for very recent genome-wide TE mobilization. Dashed line represents the Bonferroni-corrected threshold for significance. (c) Detailed Manhattan plot within 80kb around *NRPE1* locus. Colors indicate the extent of linkage disequilibrium (r2) with the leading SNP (black triangle). (d) Boxplot of numbers of very recent TE insertions in carriers of the reference *NRPE1^{ref}* and derived *NRPE1'* alleles. (e) Alleles and polymorphisms at *NRPE1* locus and the linkage between their closest tagging SNPs. (f-g) Boxplot and metaplot of CHH methylation on *NRPE1*-dependent TEs within carriers of the derived *NRPE1'_ΔQS* allele, carriers of the derived *NRPE1'_Δrep* allele and a set of 100 randomly sampled carriers of the reference *NRPE1^{ref}* allele. (h) Composition by superfamily of *NRPE1- or CMT2*-specific TE sequences. (i) Transposition rates in 1,000 F1 plants derived from WT or *nrpd1* Col-0 parents grown in control conditions. (j) Transposition rates in 1,000 F1 plants derived from Cvi-0, Sha-0, Tsu-0 and Col-0 parents grown in control conditions.

Environmental modulation of TE mobilization

We next investigated potential environmental modulators of transposition activity using 19 climatic bio-variables measured between the years 1970 and 2000 and which describe local patterns of temperature and precipitation variations (Worldclim.org). We performed a stepwise selection of the most relevant bio-variables on the basis of their added explanatory power in a generalized linear model (GLM) of very recent transposition that also includes population structure and allelic variation at *NRPE1* (see Methods). Importantly, we also included in this model the possibility of GxE interactions involving *NRPE1*, based on previous observations that experimentally-induced mobilization of one *COPIA* family (*ATCOPIA78*) is only achieved by combining mutations in the RdDM pathway and heat shock (14). The GLM revealed that, while variation in very recent transposition is explained predominantly (27%) by genetic backgrounds and allelic variation at *NRPE1*, seasonality of precipitation and diurnal temperature range contribute another 9% of this variation (6.3% and 2.7%; Fig. 3a,d, S3a). Furthermore, GxE interactions between *NRPE1'* and temperature seasonality and precipitation of the coldest quarter (BIO04 and BIO19 respectively; Fig. 3a-b) are also significant contributors, which explain together an additional 4.2% of variation in TE mobilization. Remarkably, differential TE mobilization in association with these two bio-variables is only observed for accessions carrying *NRPE1'* alleles (Fig. 3c-d), extending in natural settings the experimental observation that mutations in the RdDM pathway can enhance the environmental sensitivity of TEs. Altogether, the different factors we considered account for at least 40% of variation in TE mobilization in nature. In addition, this analysis reveals an important contribution of environmental factors, notably in interaction with *NRPE1*.

To move beyond this global picture, we investigated environmental associations at the TE family level using a Mantel test, which also incorporates population structure (see Methods). Focusing on the 77 TE families with higher mobility and responsible for 89% of the very recent TIPs used

for the GWAS and GLM analyses, we detected for 57 TE families significant associations with at least one environmental variable (Fig. 3e). In line with the GLM, positive association with precipitation seasonality (BIO15) is the most prevalent at the individual TE family level (44 out of 57 TE families; Fig. 3e). Moreover, cluster analysis identified four groups of TE families that share similar environmental associations including one small group of four *COPIA* TE families, which stands out by presenting the strongest associations overall yet almost all with temperature bio-variables only. As expected from our previous work (25), *ATCOPIA78* belongs to this last group, hereafter called “temperature” group. Importantly, the association of *ATCOPIA78* mobility with temperature is only observed in the *NRPE1’* background (Fig. S3b-c), which mirrors the requirement established experimentally for impaired RdDM in combination with heat shock to induce the mobilization of this TE (14).

To evaluate directly the extent of this GxE interaction between impaired RdDM and environmental stress, we compared the levels of TE mobilization in offspring of *nrpd1* and WT Col-0 parents grown under control conditions, or exposed to heat-shock or flagellin, a bacterial peptide known for triggering plant biotic stress response (see Methods). Transposition was increased by 64% upon exposure to heat-shock in the *nrpd1* mutant, almost an 1.5-fold increase compared to the response of WT Col-0 (Fig. 3f). This increased mobilization was not restricted to *ATCOPIA78* (+700% increase, Fig. S3d) but could also be observed for some *GYPSY* notably *ATGP1* (+400%, Fig. S3d) or *MuDR* elements (+100% and +50% for *VANDAL2* and *VANDAL6* respectively, Fig. S3d). Such a trans-family sensitization by impaired RdDM could also be observed to a lesser extent with exposure to flagellin (+5% in *nrpd1* compared to -72% in WT Col-0). Even if some TE families responded to both stresses (*ATCOPIA78* +260% or *VANDAL2* +100%, *ATGP1* +105% Fig. S3d), others showed inhibition under one stress and activation under another (*META1* -100% upon HS and +63% upon flagellin, Fig. S3d). These results confirm the GxE interaction between impaired RdDM and family-specific transposition responses to biotic and abiotic stresses.

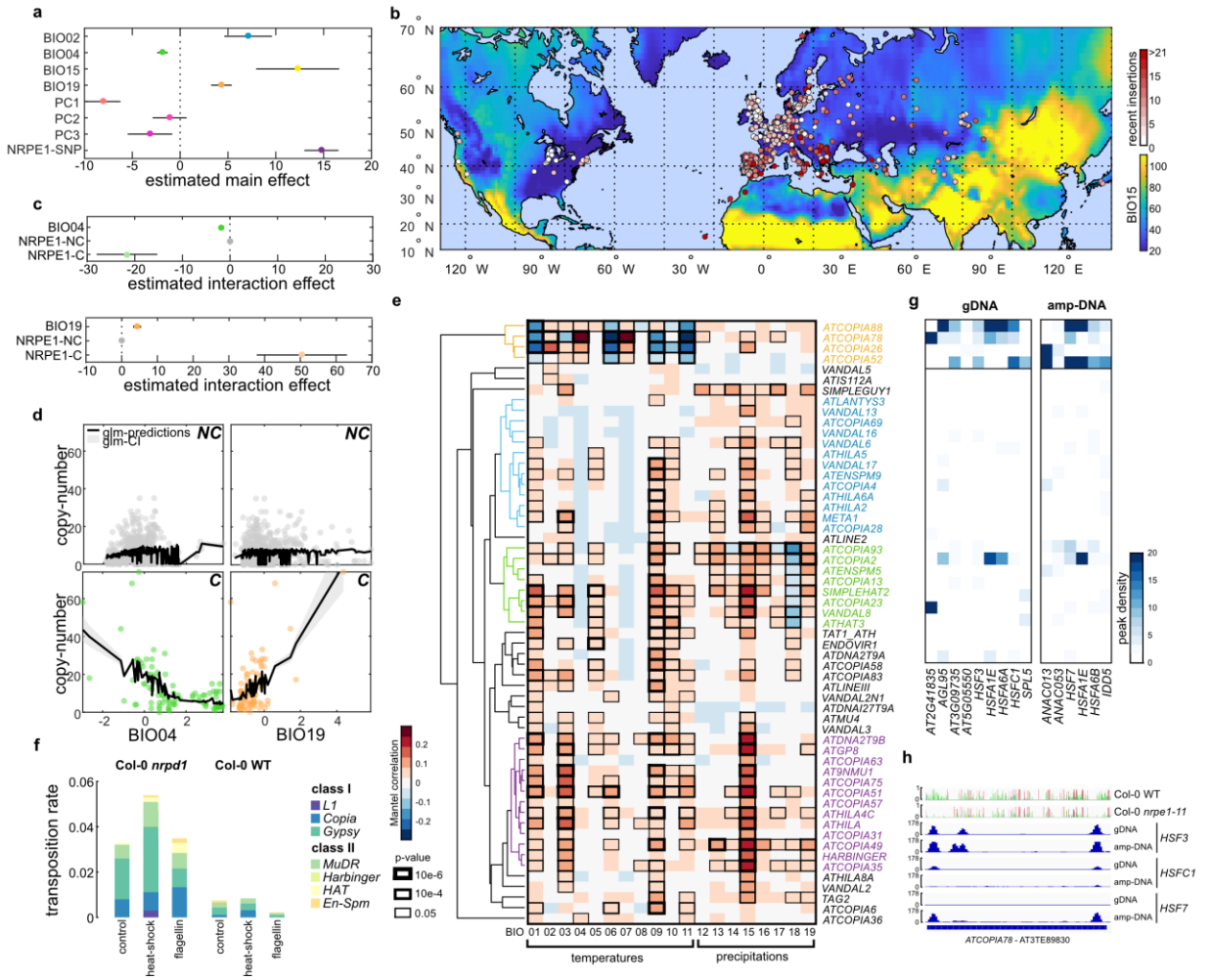

Figure 3.

(a) Estimated marginal effect at the mean of PC1-2-3, *NRPE1* allele, and BIO02, BIO04, BIO15, BIO19 in GLM of very recent transposition. (b) Number of very recent TE mobilizations in accessions distributed across the world. Precipitation seasonality (BIO15) is also shown. (c) Estimated interaction effect of BIO04 and *NRPE1* allele (upper) and BIO19 and *NRPE1* allele (bottom) in GLM of very recent transposition. (d) Scatter plot of very recent transposition against BIO04 (left) and BIO19 (right) in non-carriers (NC, up) and carriers (C, down) of derived *NRPE1'* allele. GLM predictions and confidence-intervals are indicated in black and grey, respectively. (e) Directional Mantel associations between very recent transposition of 77 TE families and 19 WorldClim bio-variables (1970-2000). Dendrogram of hierarchical clustering of coefficient correlations with four main clusters are indicated. (f) Transposition rates in 1,000 F1 plants derived from Col-0 *nrpd1* and WT parents grown in control conditions or exposed to heat-shock or flagellin. (g) Normalized peak density of in-vitro binding of TFs (DAP-seq) enriched over the "temperature" TE cluster in Col-0 gDNA and PCR-amplified DNA. (h) Tracks of DNA methylation (CG in red, CHG in blue, CHH in green) in Col-0 WT and *nrpe1_11* mutants and DAP-seq peaks of heat-shock factors *HSF3, HSFC1,* and *HSF7* in Col-0 gDNA and PCR-amplified DNA.

To investigate the molecular underpinnings of these environmental responses, we re-mapped, including over TE sequences, in vitro DNA affinity purification sequencing (DAPseq) datasets obtained using native or amplified (i.e. stripped of all DNA methylation) genomic DNA for 469 transcription factors (TFs) in *A. thaliana* (26). TE families in the three "precipitation" groups share few enrichments for sites bound by specific TFs (TFBSs; Fig. S4a), suggesting their environmental responsiveness, notably to biotic stress or drought for the cluster shared with *ATCOPIA93* (Fig. S4h-i), can be acquired through a diverse set of transcriptional wirings. In contrast, the four *COPIA* TE families belonging to the "temperature" group share enrichments in TFBSs for 14 TFs (Fig. 3g). These TFs include six known heat-shock factors (HSF3, HSF7, HSFC1, HSFA1E, HSFA6A, and HSFA6B*;* Fig. 3g-S4a-e) and another three TFs encoded by genes induced transcriptionally under different heat-shock treatments (ANAC013, ANAC053, SPL5; Fig. S4d-e). Transcriptome data for the reference accession Col-0 indicate also that three of the four *COPIA* families are transcriptionally up-regulated under heat-shock (Fig. S4f-g), most prominently *ATCOPIA78*. Moreover, comparison of binding data on native genomic DNA as well as amplified DNA, indicated that DNA methylation hinders the in vitro binding of HSF7, HSFA6B and ANAC013 at these sites (Fig. 3g-e-S4b-c), consistent with the sensitivity to DNA methylation reported for these TFs (26). Finally, we noted that *ATCOPIA26*, which is not transcriptionally up-regulated under heat-shock in Col-0, shows enrichment for the heat-responsive TF ANAC013 only when it is unmethylated (Fig. 3g). Together, these results point to an important role of environmentally responsive TFs and compromised DNA methylation in the increased mobilization of the "temperature" group of *COPIA* observed in accessions that carry the *NRPE1'* derived alleles and that are exposed to extreme seasonal shifts in temperature.

TE mobilization mainly generates highly deleterious genic mutations

In order to evaluate the phenotypic impact of TE mobilization and in particular the mutation load it generates, we measured the transcriptional impact of TIPs on neighboring genes. We ignored absence variants, as the presence of a TE annotation in the reference genome sequence at the corresponding position may have affected the annotation of the adjacent genes. In addition, we restricted our analysis to the rarest (first decile) TIPs present in one of at least 909 genomes, because collectively they provide the set of TIPs the least affected by the filter of natural selection. Of the 2,180 rarest non-reference TE presence variants (LF) retained for analysis, over 50% are located within genes, with exons being the most prevalent targets (66% of genic insertions, Fig. 4a) and as frequent as expected by chance. However, broad differences in insertion preferences can be observed across TE families with *GYPSYs* found typically within intergenic regions, *MuDRs* within promoters (<-250bp) and 5'-UTRs and *COPIAs* within exons (Fig. 4b). As a result, the vast majority (>70%) of exonic insertions are caused by *COPIAs* (Fig. 4c) and, consistent with experimental results (8), they affect preferentially environmentally responsive genes, especially those involved in defense response (Fig. 4d).

To assess the transcriptional impact of each of the 2,180 LF non-reference TE insertions, we used matched (mature leaf before bolting) transcriptomes available for 604 of the 1047 accessions (27) and compared the average transcript level of the nearest gene in the TE-carrying accessions (C) to that in non-carrier (NC) accessions. As expected, most TE insertions within exons are systematically (~75% of exonic TE insertions) associated with reduced transcript levels (Fig. 4e), mostly contributed by *COPIAs* (111 out of 124; Fig. S5a), and in almost 20% of cases, the TE-containing allele is an effective knock-out (Fig. S5b). Intronic TE insertions are also frequently associated with reduced gene expression, but this reduction is typically of smaller magnitude (Fig. S5c). Lastly, few TE insertions are associated with increased transcript levels,

and these tend to reside within the 5' UTR or the promoter regions of genes (Fig. 4e), in majority corresponding to *MuDRs* (Fig. S5a). Together, these results suggest that almost a quarter of mutations generated by TE mobilization in nature have major and mostly negative effects on gene expression.

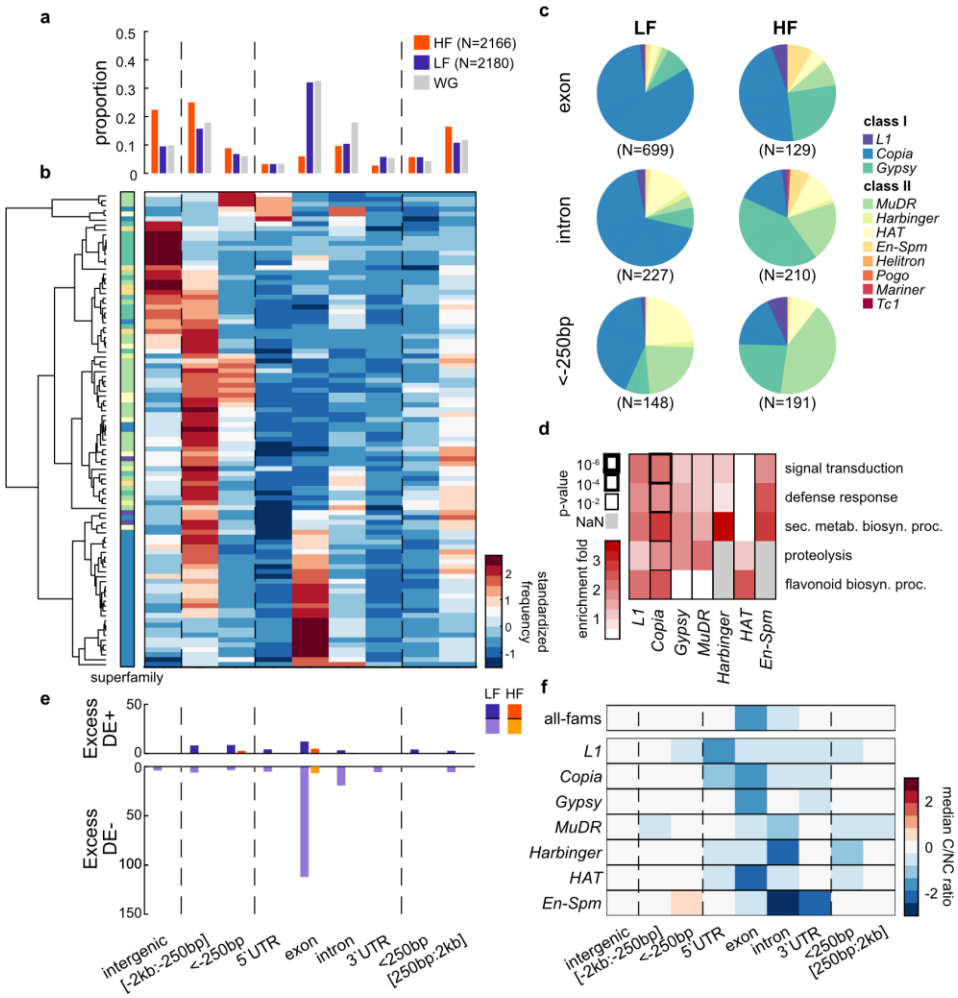

Figure 4.

(a) Fraction of low- and high-frequency TE presence variants overlapping genic annotations (exons, introns, 5′ or 3′ UTRs) or outside of genes (upstream or downstream within 250bp, or within 2kb) or intergenic (>2kb) compared to the whole-genome representation of each category. (b) Insertion frequencies across genomic categories for TE families with ≥50 TIPs. Rows are standardized and clustered based on correlation distance. (c) Contribution by TE superfamily to TE presence variants at low- and high-frequency in exons, introns, and close promoter (<-250bp). (d) GO enrichments of LF presence variants within genic annotations by TE superfamily. (e) Excess of extreme expression log ratios between carriers (C) over non-carriers (NC) by insertion category at low- and high-frequency (negative, DE-, bottom, and positive, DE+,

top) compared to random sampling of carriers and non-carriers . (f) Median transcriptomic impact (C/NC expression ratio) by TE family by insertion category.

To evaluate the evolutionary fate of mutations generated by TE mobilization in nature, we compared the genomic distribution of the 2,180 LF TE-containing alleles with that of the 2,166 most frequent (HF) ones (last decile, segregating at frequencies over 4.92%). In marked contrast to LF alleles, HF alleles are strongly biased away from genic sequences (~20% only vs ~60% expected based on the composition of genome, Fig. 4a). In particular, exons are strongly depleted (5.9% of HF vs 32% of LF variants) and transcriptome data indicates that knock-out alleles are totally absent at high-frequency (Fig. 4e-f). Similarly, no intronic TE insertions associated with major reduction in gene expression are observed at high-frequency (Fig. 4e S5). Conversely, intronic or promoter (<-250bp) variants associated with mild or positive transcriptomic impact are broadly retained at the species level (Fig. 4a,e,f, S5) as indicated by the fact that we recovered as many or more such variants at low and high-frequency (respectively 227 vs 210 for introns and 148 vs 191 for the promoters). Furthermore, the proportion of *COPIA* insertions within introns or promoters is strongly reduced among high frequency variants even though *COPIA* is the major contributor of such insertions at low frequency (68% for LF vs 16% for HF variants in introns; Fig. 4c). In contrast, *GYPSY* and *MuDR* insertions are enriched at high-frequency within introns or promoters, respectively (7% - 4% for LF vs 42% - 20% for HF, respectively; Fig. 4c), consistent with these insertions having less extreme transcriptional impacts (Fig. 4g,S5). Whether any of these high-frequency insertions are under positive selection remains to be determined. Together, these findings confirm that the majority of genic insertions, which are contributed by *COPIA,* are under strong purifying selection (Fig. 1e).

Contribution to local environmental adaptation

Consistent with this last conclusion, as well as the strong avoidance of essential genes in the case of *COPIA* insertions (8), the number of gene loci with TIPs is less than half that expected by chance (4078 vs 9090 +/- 45, see Methods) with a scarcity of essential genes (Fig. S6). Conversely, 566 gene loci are repeatedly visited by TE insertions, with three or more distinct TE-containing alleles, a proportion that this time is much higher than expected by chance (13.8% vs 7.0 +/- 0.2%, Fisher exact test p=5e-51). Given that most recurrent visits are low frequency insertions, they could result from preferential targeting, relaxed purifying selection, and/or diversifying selection. As the comparison with experimental transposition accumulation lines for the three TE families *ATCOPIA93, VANDAL21,* and *ATENSPM3* indicates a minimal overlap between gene loci visited in the lab and in nature (Fig. 5a,S6), we can rule out an important role for preferential targeting in generating recurrent visits, at least for these three TE families. Furthermore, the fact that pseudogenes are not strongly enriched in TE insertions (206 vs 167 expected by chance) indicate that multiple hits cannot solely result from relaxed purifying selection. Indeed, gene loci with TIPs do not appear to be functionally decaying, with 99% of pN/pS values under the upper 1% genome-wide threshold. In addition, the number of TIPs at a given gene correlates positively with pN/pS (Fig. 5c), suggesting that multiple visits may be the result of diversifying selection. That these visits are functionally relevant is also supported by the observation that in a quarter of cases (140 gene loci), all TE-containing alleles of a gene locus have similar effects on gene expression levels (Fig. 5a). This similarity is most striking at the *FLOWERING LOCUS C* (*FLC*) locus, which encodes a key repressor of flowering and is one of the main genetic determinants of natural variation in the onset of flowering (28). Specifically, we identified a total of 16 distinct TE-containing *FLC* alleles (Fig. 5b), twelve more than reported in our previous study based on 211 genomes (25). All 16 TE insertions reside in the first intron of the gene, which is essential to the environmental regulation of *FLC (29)*, and are associated with

reduced rather than complete loss of gene expression. Consistent with and expanding previous observations, all insertions are associated with earlier flowering (Fig. 5c) and lower expression (Fig. 5a), thus reinforcing the notion that TE insertions at *FLC* may underpin local adaptation (8,25).

As expected, genes with multiple TIPs are strongly depleted in essential genes (Fig. S6) and associated with core cellular processes, notably translation (Fig. 5e). Instead, genes with increasing numbers of TIPs are progressively enriched in GO terms linked to defense response, a category of genes under strong diversifying selection (30). To determine if local adaptation could explain beyond *FLC* the recurrence of TE insertions, we searched for potential associations between TE insertions and environmental variations. We summarized the 19 WorldClim bioclimatic variables into three climatic envelopes (CEs, Fig. 5f) that together explain >80% of the climate niche variations observed across the locations of the 1001 Genomes accessions (Fig. S6). CE1 increases with wetter winters (BIO12 and BIO19) and reduced temperature seasonality (BIO04 and BIO07); CE2 with hotter and drier summers (BIO05 and BIO18) and CE3 with increased temperature changes between winters and summers (BIO04, BIO05 and BIO06). Along each climatic envelope, we tested for each gene locus with multiple hits whether its TE-containing alleles were associated with an environmental shift using a logistic GLM that incorporates population structure (see Methods). Out of the 566 loci tested, 137 showed significant associations but none when the GLM was repeated using random permutations of the environmental variables. Associations are mainly with higher or lower CE2 values and/or with lower CE3 (Fig. 5g). Notably, TE-containing alleles of *FLC* are found preferentially in parts of the world characterized by milder winters (low CE3), which is expected given the role of *FLC* in repressing flowering before winter. Moreover, the population frequency of independent TE-containing alleles is higher when located within multiple-hit genes with environmental associations (Fig. 5h) and this observation also holds true for exonic TE insertions (Fig. 5i), despite their strong

effects on gene expression (Fig. 4d). These patterns are consistent with adaptation through recurrent mutation of environmental response genes combined with a local rise in frequency of the TE-containing alleles experiencing positive selection (31), and in line with recent evidence that repeated loss-of-function mutations can participate in recurrent local adaptation (32).

Finally, we evaluated whether the *NRPE1* locus itself, given its role in modulating genome-wide transposition levels, could be under positive selection in challenging environments. We quantified positive selection genome-wide using iHS and iHH12, two measures of haplotypic length whose extreme positive values are hallmarks of hard and soft sweeps respectively (ref). As expected given the wide distribution of the *NRPE1'* allele, no marks of hard sweep were detected at the *NRPE1* locus using iHS. Using iHH12 however, the *NRPE1* locus appears under positive selection in accessions located in extreme CE2 environments but not in those in moderate CE2 environments (Fig. 5j). Similarly, *NRPE1* bears marks of soft sweeps in low but not high CE3 environments (Fig. S6), in line with the high number of recurrent TE insertions associated with these low CE3 and extreme CE2 environments.

Together, these findings provide strong evidence that TE mobilization represents a powerful generator of locally adaptive alleles whose tuning by the RdDM machinery is itself under positive selection in challenging environments.

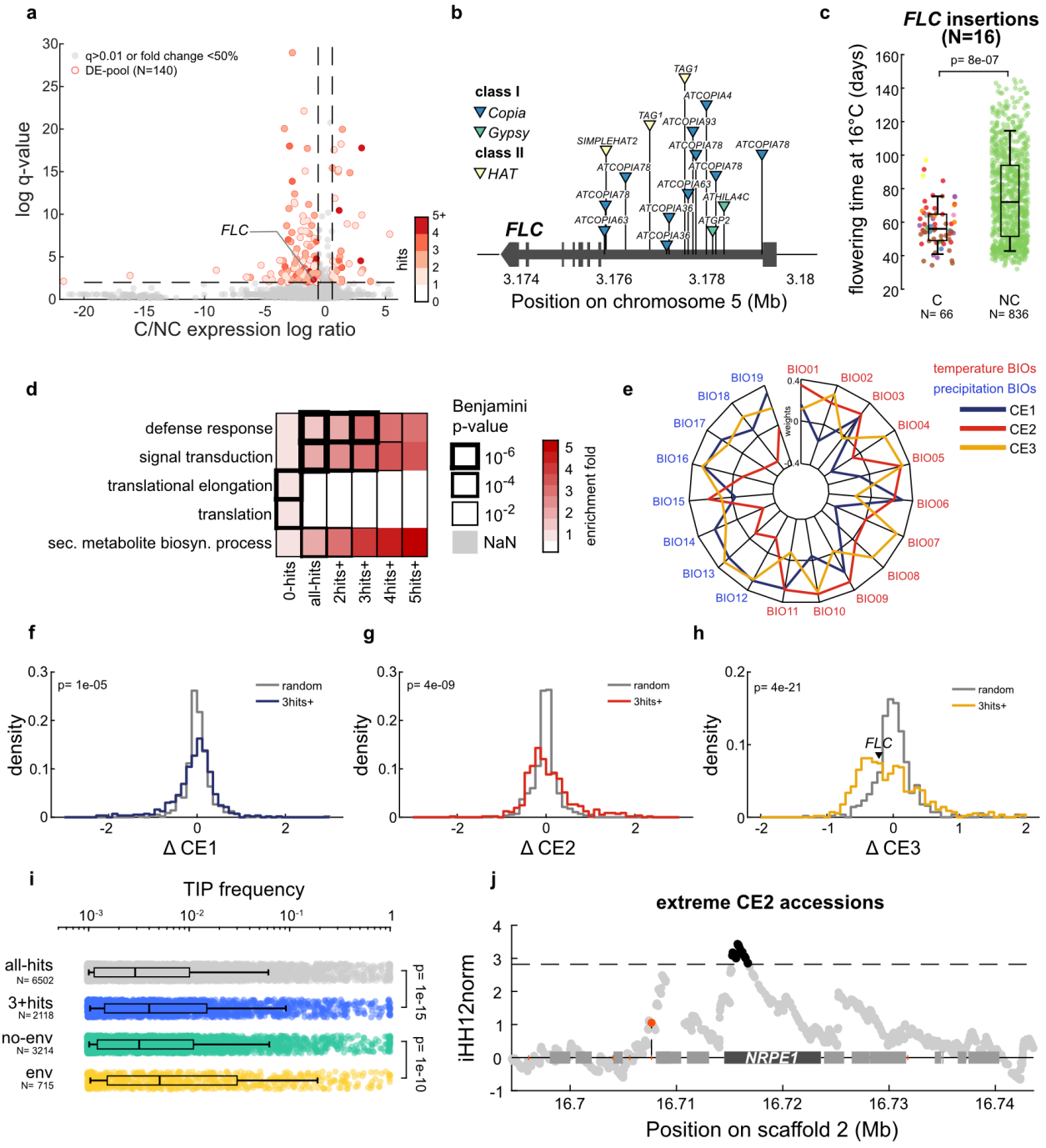

Figure 5.

(a) Significance against log ratio of combined transcriptomic effects of TE insertions within or near (<250bp) genes in carriers (C) compared to non-carriers (NC). The number of TE insertions found for each locus is indicated as a shade of red. (b) Location and identify of the 16 TE insertions detected within *FLC*. (c) Flowering time of accessions at 16°C carrying (C) or non-carrying (NC) an intronic insertion in *FLC.* The p-value of Wilcoxon test is indicated. (d) Top 5 GO enrichment terms across genes never visited or visited once or more. (e) Weights across 19 bio-variables of 3 first climatic envelopes (CEs) in PCA of 1047

accessions. (f-h) Distributions of climatic envelope shifts (ΔCEs) observed betweens carriers and non-carriers of TE insertions for each of the 566 genes hit 3 times or more compared to the distribution of ΔCEs with the same numbers of randomly selected carriers. The p-values of Kolmogorov-Smirnov comparisons between observed and random distributions are indicated. (i) Frequency of TE insertions found within or near genes visited (all-hits), visited 3 times or more (3hits+), in association with a CE shift (env) or not (no-env). The p-values of Wilcoxon tests between distributions are indicated. (j) iHH12 values in extreme CE2 accessions (upper and lower quartiles) across the *NRPE1* region with in black indicated values above the genome-wide 1% threshold (dashed line).

Increased TE mobilization under future climates

Given the strong environmental sensitivity of TE mobilization and the role of TE insertions in local adaptation to diverging environments, we can anticipate that the mutation pressure generated by transposition will be significantly affected by climate change, potentially with important consequences for the future of the species. To test this prediction, we first explored forecasts of climate change in the next 60-80 years (CMIP6, Eyring et al. 2016) for each of the ecological niches occupied by the *A. thaliana* accessions of the 1001 Genomes project. As forecasts differ at the local scale between global climate models (GCMs), we averaged the results obtained from four different GCMs (CNRM-CM6-1, IPSL-CM6A-LR, MIROC6, and MRI-ESM2-0; worldclim.org) under the most pessimistic gas emission scenario (SSP5-8.5; Eyring et al. 2016). Consistent with the expected global increase in the frequency of hotter and drier summers, CE2 was the most affected ecological niches's factor concerning accessions belonging to eight out of the ten haplogroups (Fig. 6a, S6).

We next used the full GLM of recent TE mobilization (Fig. 3e) to predict the extent to which the change in climate in 60-80 years from now will affect the number of recent TE insertions carried by individual genomes assuming that the genetic structure will remain unchanged over this short evolutionary time. To assess the predictive power of different GLMs, we carried out 100 bootstrappings of 100 accessions and compared for these testing sets the number of recent TE copies estimated by the GLMs based on today's climate to that observed currently (see Methods).

Compared to the GLMs based only on genetic variables (G: population structure and allelic variation at *NRPE1*) or including BIO15 (G+E), the full model with GxE interactions between bio-variables and *NRPE1'* reached higher predictive variance (R^2) on average (Fig. 6b). Estimates obtained using this last GxE model were consistently conservative, except for five accessions, which we did not consider further (Fig. S7). Estimates were then recalculated for the remaining 1042 accessions using this time the climate forecasted in 60-80 years. Comparisons between the number of recent TE insertions estimated under future and recent climates suggest rare decreases and frequent increases in the rate of TE mobilization during this very short time span. Furthermore, predicted increases in TE mobilization are most pronounced in Mediterranean populations (Fig. 6d), which are expected to experience higher precipitation seasonality (BIO15) (Fig. S6) and mean diurnal temperature range (BIO02) (Fig. 6c). Nonetheless, these climatic influences on TE mobilization should be modulated by the presence of *NRPE1',* as a result of the strong GxE interactions with temperature seasonality (BIO04), which is predicted to rise at lower latitudes (Fig. S6). Thus, Mediterranean populations carrying the derived *NRPE1'* alleles should experience reduced mutation pressures from TE mobilization compared to their *NRPE1* allele counterparts. Conversely in higher latitudes (i.e. Sweden haplogroups) the *NRPE1'* allele should result in a higher transposition pressure. Given that the *NRPE1* locus appears to be under positive selection in extreme CE2 accessions (Fig. 5j) but not in moderate CE2 accessions (Fig. S6), we can speculate that the GxE interactions with allelic variation of this gene will play an important role in the survival of *A. thaliana* in the face of climate change.

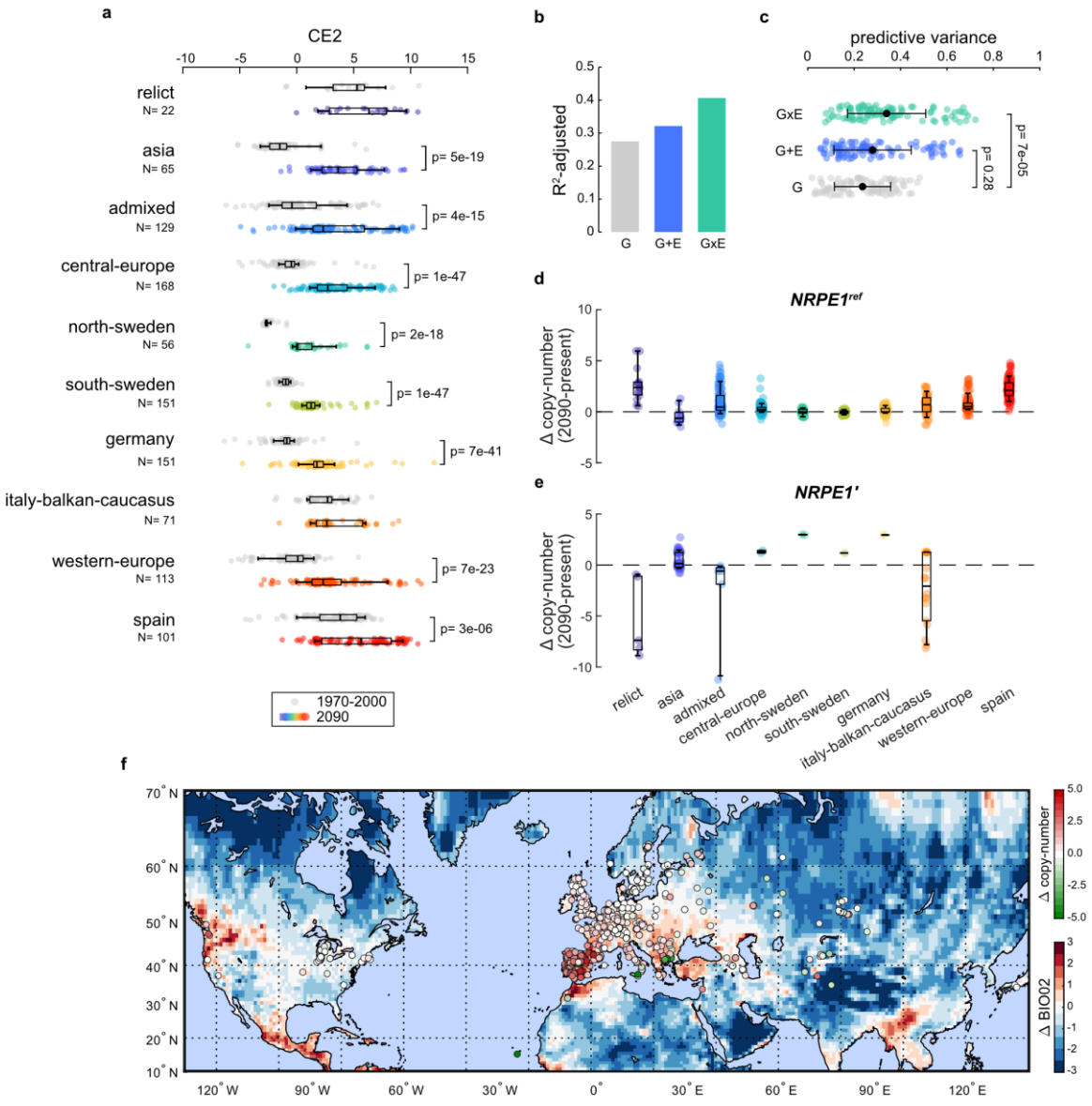

Figure 6.

(a) Forecasted change of climatic envelope CE2 by haplogroup under average CMIP6 GCM for 2081-2100 compared to recent climate (1970-2000). (b) Adjusted R^2 of GLMs based on population structure and allelic variation at NRPE1 (G), including BIO15 (G+E), or with interactions between bio-variables and NRPE1' (GxE) (c) Predictive variance of the numbers of recent copies on 100 random samples of 100 accessions extracted from the GLMs training sets. (d-e) Predicted change by haplogroups in copy-numbers using GxE GLM with future climate predictions for 2081-2100 for (d) carriers of the reference *NRPE1* allele and (e) carriers of the derived *NRPE1'* alleles. (f) Spatial variations in BIO02 predicted for 2081-2100 and their outcome on copy-number changes by accession.

Discussion

Understanding how organisms adapt to new environments is of major importance given that climate change is already leading to shifts in species range (33). Standing genetic variation is generally thought to be the main source of rapid adaptation to environmental changes (1). As a result, population genomic studies have been used to determine the potential of existing genetic diversity to support adaptation to future climates (34). However, the contribution of large effect alleles that TE mobilization typically generates have so far been ignored and so has the role of transposition as an important source of *de novo* variants.

Our comprehensive characterization of recent TE mobilization in *A. thaliana* revealed that the natural substitution rate of TE insertions is of the same order of magnitude as that of SNPs (Fig. 1) and close to the actual transposition rate that we measured experimentally (Fig. 2). Moreover, given the very stringent filters that we applied throughout our pipeline to detect TIPs, we likely underestimate the mutational pressure provided by TE mobilization. Our results indicate therefore that TE mobilization is a substantial contributor of inherited mutations in nature. Unlike SNPs, which are predominantly randomly distributed along the genome and neutral or mildly deleterious, most insertion mutations occur near or within genes, have large effects, and are eventually purged by natural selection (Fig. 1&4). Thus, the high transposition rate together with the relatively short lifetime of new TE insertions implies that TE mobilization is a major evolutionary force that is constantly reshaping the genome.

We identified *NRPE1* as a major genetic determinant of natural transposition. This gene is a key component of the RdDM pathway and the natural allelic series segregating at the *NRPE1* locus includes a hypomorphic truncated form that is associated with higher levels of recent transposition and causes lower methylation at TEs targeted by RdDM. Moreover, we demonstrated experimentally that loss of RdDM activity induces increased mobilization for several target TEs,

thus providing direct evidence that *NRPE1* variation is causal in modulating transposition in nature.

Materials and Methods

Detection and filtering of TE insertion polymorphisms (TIPs)

Paired-end short-read whole-genome sequencing data were obtained for 1047 *A. thaliana* accessions from 1001genomes.org and processed using a combined SPLITREADER and TEPID pipeline as described (16). Briefly, putative non-reference insertion sites detected at the individual level by the SPLITREADER were then intersected and filtered by TE family at the population level in order to merge compatible overlapping insertion sites where at least one individual presented enough supporting reads (DP filter = 3). For both presence and absence variants, local comparisons of the negative coverage were then used to reduce the rate of both false positives and false negatives. Indeed, a drop of coverage in the alignment to the reference genome is expected over true non-reference presence sites compared to surrounding regions (100bp up and down) and similarly at the edges of true non-reference absence variants. Following this step, high-specificity (low rates of false positives) was obtained across TE families, apart from *HELITRON* presence variants (Baduel et al. MMB 2020). Conversely, genomes with little coverage (neither supporting an insertion, i.e. positive coverage, or its absence, i.e. negative coverage) over the insertion site or over the reference TE sequence were classified as NA as they cannot be called by either pipeline. Sites with less than a 100 informative genomes were discarded as these bring little information on the frequency of the TIP across the 1047 genomes. Furthermore, we removed ~2,500 (2,474) non-reference insertion sites where the positive coverage is never higher than the negative coverage within a given carrier, as heterozygous non-reference insertions are not expected in a selfer like *A. thaliana* expect if they occured in the past one or two generations

which could represent transposition events that occurred in the lab. Within TE absence variants, 4,455 correspond to fragmented reference TE sequences (4,008) or ancestral reference TE sequences also found in the *A. lyrata* genome by a BLAST of the 200bp sequences bridging the two edges of the reference TE sequences (447). These absence calls were also removed as their absence most likely result from genomic rearrangements produced by unequal crossing-over events or non-homologous recombinations instead of recent TE mobilization events. Although some of absence variants likely reflect excision in non-reference genomes, a significant fraction segregate at frequency <20% and therefore likely represent recent insertions in Col-0.

Methylome analysis

Processed bisulfite sequencing (BS-seq) data of 779 of the 1047 genomes was obtained from 1001genomes.org (27). Methylation files of carriers of the derived *NRPE1'_Δrep* and of the derived *NRPE1'_ΔQS* alleles and 100 carriers of the reference *NRPE1*^{ref} allele were merged using the methylpy `merge-allc` option (35). Bigwigs were generated using methylpy `allc-to-bigwig`. Merged bigwigs were then processed and plotted in metaplots over all NRPE1-targeted TEs using deepTools (36) functions `computeMatrix` and `plotProfile`. BS-seq data from the experimental *nrpe1* allelic series were obtained from (24) and processed similarly. NRPE1-targeted TEs were defined as overlapping with DMRs identified in the *nrpe1_11* mutant line (24) while CMT2-targeted TEs were defined from *cmt2* DMRs (37).

Genomic analyses

The SNP vcf file was obtained from 1001genomes.org (23) and genome-wide pairwise divergences were calculated across all pairs of accessions using the `allvsall --sample-diff counts-only` option of PLINK2 (38) available download at https://www.cog-genomics.org/plink/2.0/. Pairwise SNP differences were then compared to pairwise TIP

differences within either only recently diverged accessions (diverging by less than 500 SNPs genome-wide) or 104,700 pairs of all accessions (100 random pairwise comparisons for each accession). A linear regression with no intercept was fit in both cases. The slope of the linear regression calculated over closely related accessions was used to derive the genome-wide TE insertion substitution rate from the one calculated for SNPs of 0.2511 per genome per generation (2.11E-9 per site per generation; (19). For all pairs of accessions, the substitution rate was rescaled to take into account the effect of selection on SNPs which we estimated using the putatively neutral synonymous SNPs (Fig. S1e).

Pairwise divergence were calculated within 70kb windows surrounding each TE insertion site between all carriers of the TE insertion using PLINK2 (38) .The age of TE insertions were then estimated based on the highest pairwise divergence observed within the 70kb window between any two carriers and divided by the mutation rate (7E-9) (18).

SNPs were annotated using snpEff (39), and sifted by functional effect using snpSift (40) (`"ANN[0].EFFECT has 'synonymous_variant'"` for synonymous, `"ANN[*].EFFECT has 'missense_variant'"` for missense, `"ANN[0].EFFECT has 'intergenic_region'"`, for intergenic, and `"ANN[*].EFFECT has 'stop_gained'"` for stop SNPs). Alternate and reference allele SFS for each SNP category were obtained using the `--freq` command of PLINK2 (38) then folded. The distribution of fitness effects (DFEs) of SNPs and TIPs were calculated from the folded site frequency spectrum (SFS) in 500 bins and compared to synonymous SNPs using DFE-alpha (17) with a two epochs model to take into account the recent population expansion of *A. thaliana (41)*. The time (`t2`) and the amplitude (`n2`) of the change of population were set for optimization by likelihood maximization (`search_n2` and `t2_variable` set to 1) starting from the initial `t2` value of 50. The mean effect of a deleterious mutation (`mean_s`) and the shape parameter (`beta`) of the gamma distribution of the DFE were also set to be optimized by likelihood maximization

(`mean_s_variable` and `beta_variable` set to 1) starting from the initial values of 0.1 and 0.5 respectively.

Estimates of recent TE mobilization were obtained genome-wide or by superfamily using 7,436 TIPs segregating at frequencies lower than 0.2% and private or younger than 1,000 years old, hereafter referred to as very recent TIPs. 89% of these very recent TIPs were contributed by 77 TE families with more than 20 TIPs species-wide. Genome-wide association study (GWAS) were run using EMMAX (42) using the 845,188 biallelic SNPs with minor allele frequencies >5% and missing genotyping rate <10% that have been identified across the 1001 Genomes (Alonso-blanco et al., 2016; 1001genomes.org) from which was calculated the recommended BN (Balding-Nichols) kinship matrix. Linkage between SNPs were calculated using PLINK (ref). Local scores were calculated using the R package from (ref). Generalized linear models (GLM) of the combined number of recent TE copies of the 77 most recently mobile TE families were fitted using the MATLAB function fitglm with a poisson distribution to estimate the percentage of variance explained (PVE) by the explanatory variables provided by the first three principal components (PCs) of the principal component analysis (PCA) of the IBS kinship matrix (which together represent 77.6% of the variation in kinship) with or without the *NRPE1-16719082* leading SNP. Including *NRPE1-16719082* improved the fit of the GLM to reach 27.5% of PVE compared to only 10.1% with only the 3 kinship-PCs (Table S2). For graphical purposes, marginal effects and 95% confidence intervals of each variable in a GLM (Fig. S2) were represented by approximating the GLM with a linear model and averaging the effect of all the other variables using the MATLAB function plotEffects.

TE-sequence capture

TE sequence capture was performed on exactly 1,000 F1 plants in all cases. Genomic DNA was extracted from seeds using the CTAB method, except in the case of Cvi for DNA was extracted

from germinated seeds. Libraries were prepared using 1 µg of DNA and KAPA HyperPrep Kit (Roche) following manufacturer instructions. Libraries were then amplified through 7 cycles of ligation-mediated PCR using the KAPA HiFi Hot Start Ready Mix and primers AATGATACGGCGACCACCGAGA and CAAGCAGAAGACGGCATACGAG at a final concentration of 2 µM. 1 µg of multiplexed libraries were then subjected to TE-sequence capture (8,25). Enrichment for captured TE sequences was confirmed by qPCR and estimated to be higher than 1000 fold. Pair-end sequencing was performed using one lane of Illumina NextSeq500 and 75 bp reads. Between 15 and 100 million paired reads were sequenced per library. After random downsampling (10 times) to 25 million paired reads of all samples with greater sequencing depth, reads were mapped to the TAIR10 reference genome using Bowtie2 v2.3.25 with the arguments –mp 13 –rdg 8,5 –rfg 8,5 –very-sensitive. An improved version of SPLITREADER (available at https://github.com/baduelp/public) was used to detect new TE insertions. Putative insertions supported by at least two and no more than 400 split-reads and/or discordant-reads at each side of the insertion sites were retained. Insertions spanning centromeric repeats or coordinates spanning the corresponding donor TE sequence were excluded. In addition, putative TE insertions detected in more than one library were excluded to retain only sample-specific TE insertions.

Environmental associations

Gridded weather and climate data at the 5’ resolution were obtained from WorldClim.org. Current climate for each accession was estimated from 19 bio-climatic variables summarizing monthly averages over the period 1970-2000 (WorldClim version 2.1) on the basis of their GPS coordinates (1001genomes.org) in the 5’ grid. After z-scoring, current bio-climatic variables were added sequentially to the GLM of numbers of recent TE copies on the basis of their contribution to the R^2 either as fixed effects or as interaction effects with the *NRPE1-16719082* SNP until no

added variable increased R^2 by more than 1% in order to prevent hyperinflation of the model (Table S3). For graphical purposes, marginal effects were represented as described above using a linear approximation of the GLM, and conditional effects were estimated for each pair of variables with a significant interaction term using the MATLAB function plotInteractions.

The Mantel test was performed using the MATLAB script RestrictedMantel (43) with 1000 permutations to test for associations between recent TE mobilization for each of the 77 most mobile TE families against each of the 19 current bio-variables after taking into account the IBS kinship matrix. TE families were then clustered by environmental associations using the MATLAB clustergram function based on the correlation distance (one minus the correlation between rows) between their 19 bio-variables association values.

To study the binding potential of transcription factors (TFs) on transposable element (TE) we reanalyzed DNA affinity purification and sequencing (DAP-seq) data obtained in *Arabidopsis thaliana* (O'Malley et al. 2016) for 529 TFs. We processed this data using a modified version of the bioinformatics pipeline implemented by (O'Malley et al. 2016) to consider, in addition to single-mapping reads, reads that map to multiple positions in the genome and that are often associated with identical TE copies present in multiple copies. Single-ended reads were mapped on the TAIR10 genome using Bowtie2 Bv.2.3.2, and PCR duplicates were removed using Picard. The detection of peaks for TF binding was performed with GEM (arguments --k_min 6 --kmax 20 --k_seqs 600 --k_neg_dinu_shuffle --t 5). Density of binding peaks (# peaks / kb) over each TE family were normalized by the genome-wide density of each TF to take into account differences between TFs. Preferential enrichment for a TF binding over a TE cluster were calculated using a Wilcoxon rank sum test of the normalized TF densities over the TE families of a cluster compared to the other recently mobile TE families (out of the 77).

Raw RNA-seq data were obtained from publicly available datasets (Supplementary table S1). Expression level was calculated by mapping reads using STAR v2.5.3a (44) on the *A. thaliana* reference genome (TAIR10) with the following arguments –outFilterMultimapNmax 50 --outFilterMatchNmin 30 --alignSJoverhangMin 3 --alignIntronMax 10000. Duplicated pairs were removed using picard MarkDuplicates. Read counts were calculated over annotated genes and TE sequences features and normalized between samples using DESeq2 (45).

Ecological niche modelling of the 1047 accessions was performed by PCA of the 19 bio-climatic variables which were summarized into three climatic enveloppes (CE1-3) which together explained 79.9% of the environmental variance. Association between the presence of a TE insertion within or near (250bp) recurrently hit genes (566 with 3 or more TIPs) and the three climatic enveloppes (CE1-3) was calculated using a binomial GLM (logit link function) using MATLAB fitglm function. P-values were then corrected using the Benjamini & Hochberg correction for false discovery rate (FDR) using MATLAB fdr_bh function. Random expectations were calculated by shuffling randomly the environment of all the accessions.

Random expectations of the number of genes or pseudogenes with TIPs located within 250bp were calculated by randomly distributing 23,331 TIPs across the genome. The average and standard deviation in the number of genes with random TIPs nearby were calculated over 10 replicates of the random distribution of TIPs.

Forecasting TE mobilization

Future climate forecasts were obtained by averaging CMIP6 downscaled future climate projections (calibrated on WorldClim v2.1 as baseline) for the 2081-2100 period with the most extreme Shared Socio-economic Pathway (SSP) 585 under four global climate models (GCMs): CNRM-CM6-1, IPSL-CM6A-LR, MIROC6, and MRI-ESM2-0 to take into account the

heterogeneity between different models. Future bio-variable values for each accession were then z-scored based on the mean and standard deviation of the current climate bio-variables in order to use them as input variables in the GLMs trained using current climate and estimate the future TE mobilization predicted by the model. To evaluate the predictive power of the model we extracted 100 random accessions (~10% testing set) and estimated the parameters of the full GLM (Table S2) using the remaining 947 accessions (training set). Using these parameters we then compared the number of recent TIPs predicted by the GLM for the 100 accessions of the testing set against the recent TIPs observed in these genomes and repeated the random sampling of a testing set a 100 times (Fig. 6). For each accession, we thus obtained ~10 estimates of the predicted number of recent TIPs from which we could derive a predictive accuracy (standard-deviation) and

Statistical analyses

All statistical analyses and graphics were realized using MATLAB R2020a, The MathWorks, Natick, 2020.

Acknowledgments

We acknowledge the World Climate Research Programme, which, through its Working Group on Coupled Modelling, coordinated and promoted CMIP6. We thank the climate modeling groups for producing and making available their model output, the Earth System Grid Federation (ESGF) for archiving the data and providing access, and the multiple funding agencies who support CMIP6 and ESGF. Support was from the Centre National de la Recherche Scientifique (MOMENTUM program, to L.Q.).

Supplementary Figures

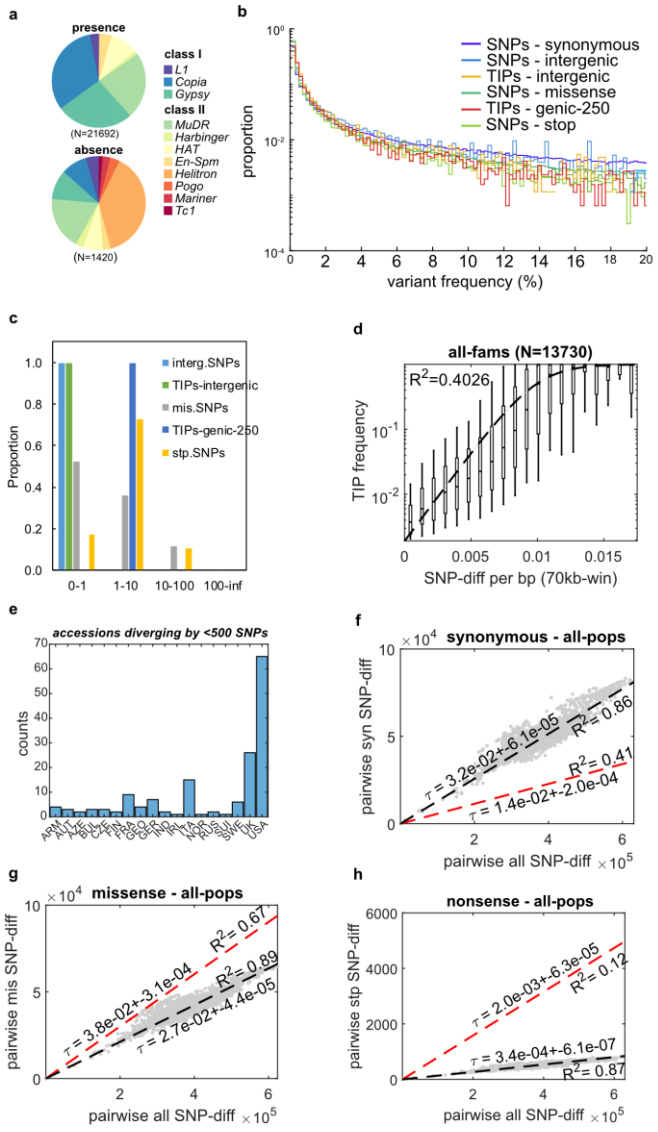

Figure S1.

(a) Contribution by TE superfamily to TE presence variants and TE absence variants. (b) Folded low-frequency spectrum (<20%) of TIPs further than 2kb from the nearest gene (intergenic), within 250bp of the nearest gene (genic-250) and bi-allelic SNPs by functional category (synonymous, intergenic, missense, and stop variants). (c) Distribution of fitness effects (DFE) of non synonymous variants as almost neutral (Nes<1), mildly deleterious (1<Nes<10), and strongly deleterious (Nes>10). (d) Correlation between TIP frequency and maximum pairwise SNP differences observed within 70kb between any two carriers for all non-private TIPs. (e) Distribution of closely related accessions by haplogroup.

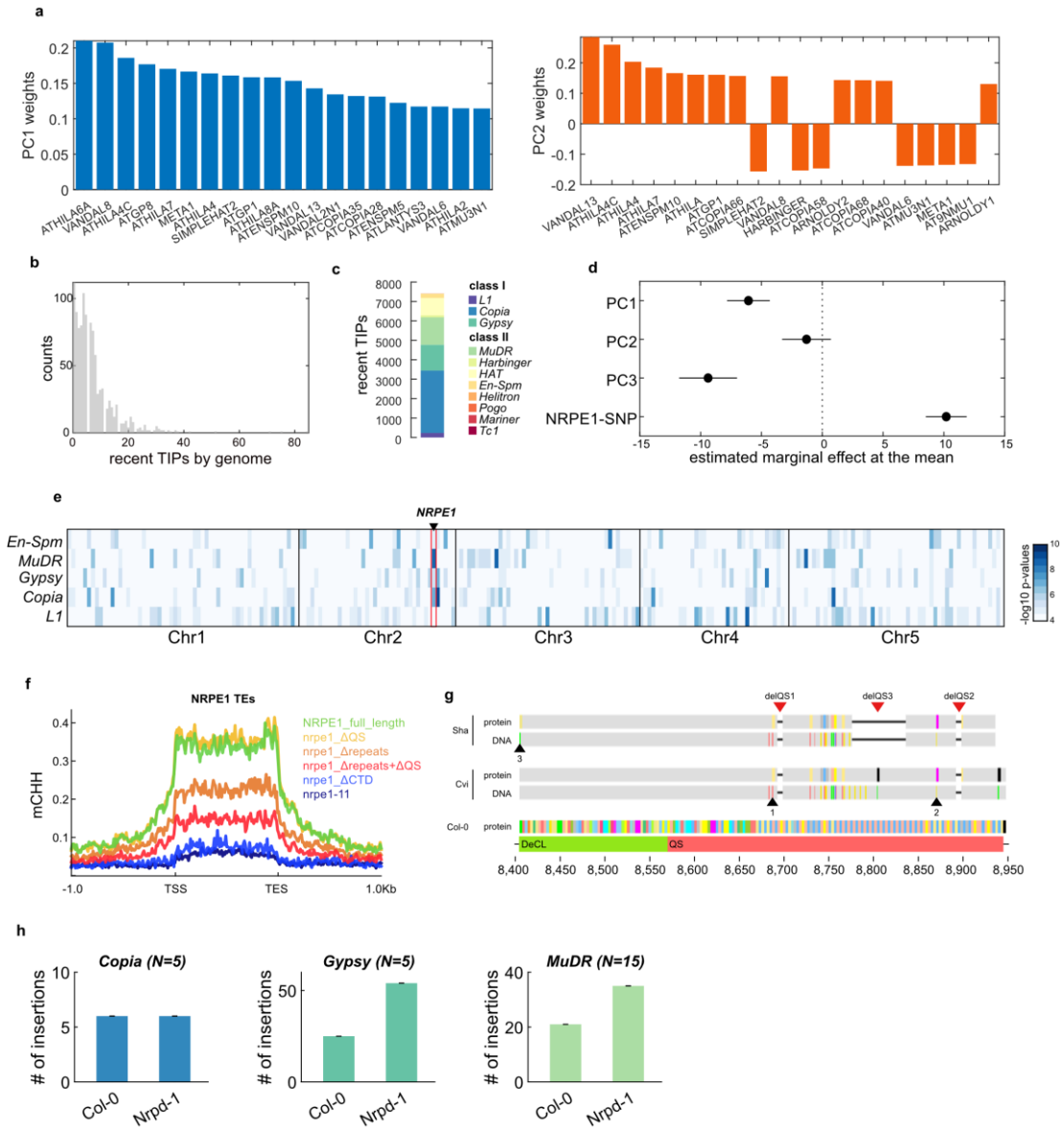

Figure S2.

(a) Weights of first 20 TE families in PC1 and PC2 of PCA of recent TE mobilization (TIP frequency <=5%, LF5) (b) Variations in numbers of very recent TIPs (frequency <=0.2% and private or <1,000 years old) across 1047 accessions (c) Contribution to very recent TIPs by superfamily. (d) Estimated marginal effect at the mean of PC1-2-3 of the kinship matrix and the *NRPE1* allele in GLM of genome-wide very recent TE mobilization. (e) Heatmap of maximal -log10 p-values in 500kb windows across the 5 chromosomes from the GWAS of very recent TE mobilization by superfamily against MAF005 SNPs. (f) Metaplot of CHH methylation over NRPE1-TEs in NRPE1 mutant constructs from Wendte et al. 2017.

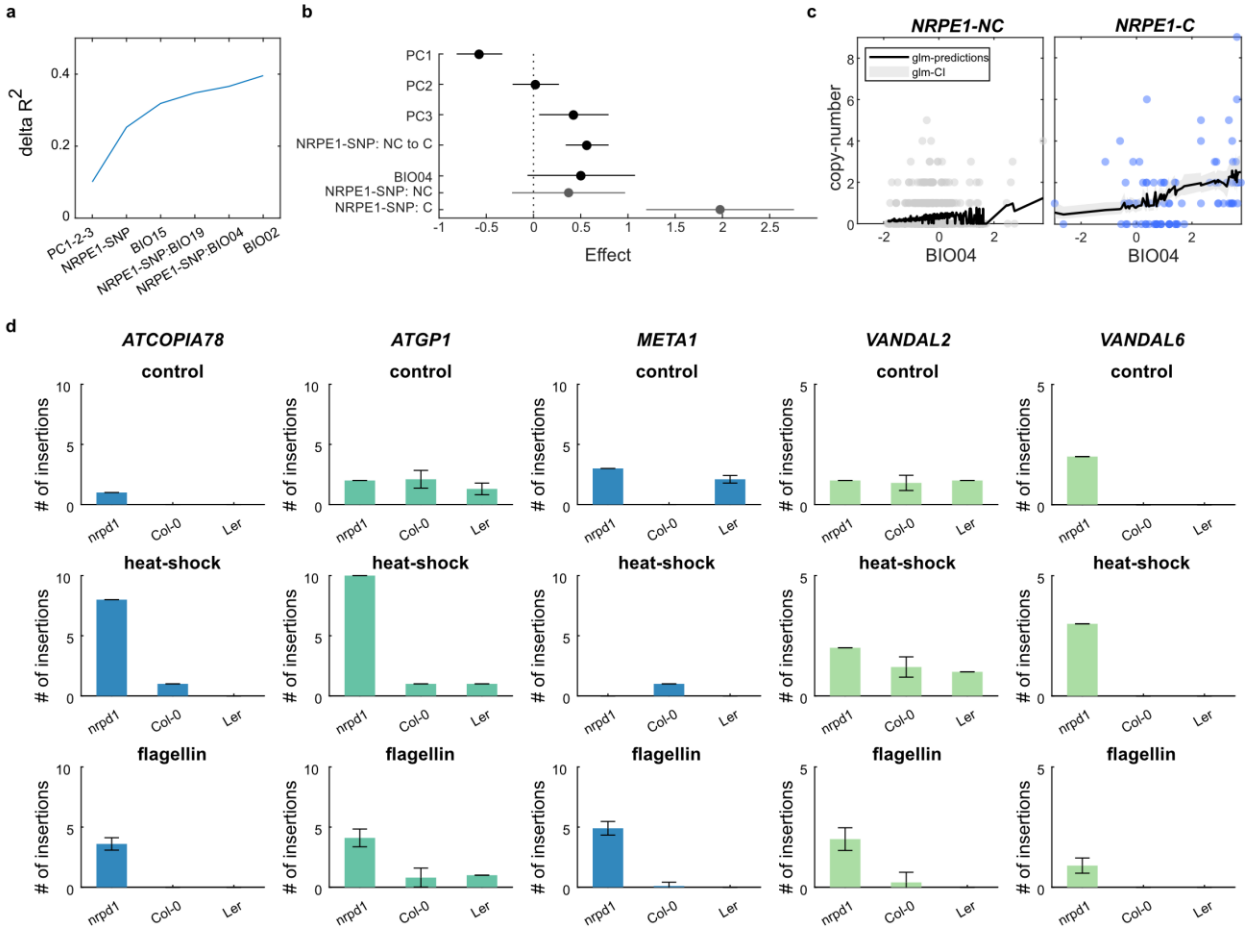

Figure S3.

(a) Cumulative fraction of variance explained (R^2) by successive addition of PC1-2-3, *NRPE1* allele, and BIO02, BIO04, BIO15, BIO19 in GLM with interaction effects of very recent transposition. (b) Marginal effect at the mean of PC1-2-3, *NRPE1* allele, and BIO04 and estimated interaction effect between BIO04 and *NRPE1* in GLM of very recent *ATCOPIA78* transposition. (c) Scatter plot of very recent *ATCOPIA78* transposition against BIO04 (left) and BIO19 (right) in non-carriers (NC, up) and carriers (C, down) of derived *NRPE1'* allele. GLM predictions and confidence-intervals are indicated in black and grey, respectively. (d) Numbers of new insertions by TE family detected in 1,000 seedlings derived from Col-0 *nrpd1* and WT and Ler WT parents grown in control conditions or exposed to heat-shock or flagellin.

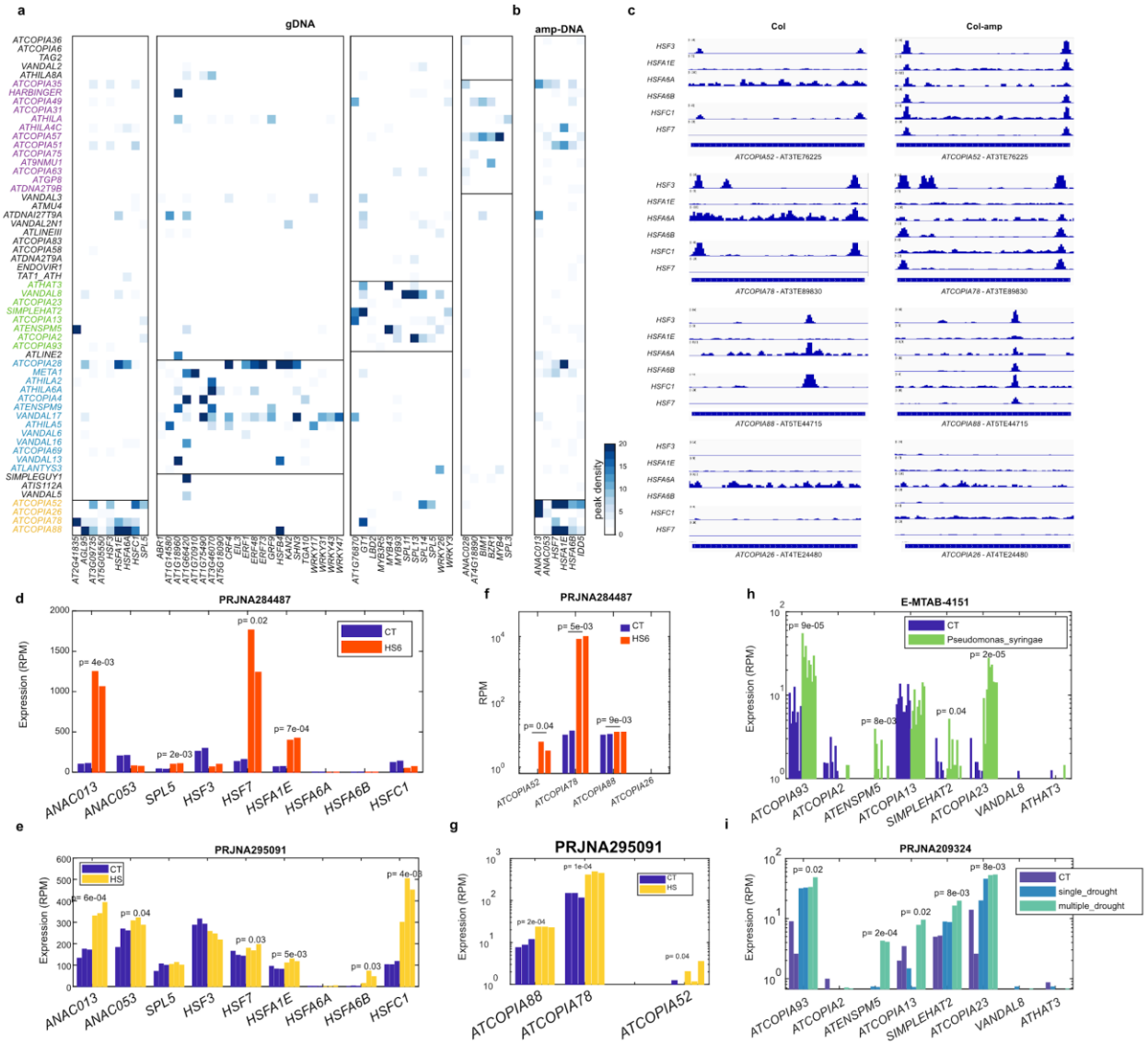

Figure S4.

(a) Normalized peak density of in-vitro binding of TFs (DAP-seq) enriched over each of the four TE clusters identified in Col-0 gDNA and (b) over the “temperature” TE cluster in Col-0 PCR-amplified DNA (c) Tracks of DAP-seq peaks of heat-shock factors *HSF3, HSFA1E, HSFA6A, HSFA6B, HSFC1,* and *HSF7* in Col-0 gDNA and PCR-amplified DNA over the four TEs composing the “temperature” cluster. (d-e) Normalized RNA-seq expression levels of TFs enriched over the “temperature” TE cluster in both Col-0 gDNA and PCR-amplified DNA under two heat-shock experiments. (f-g) Normalized RNA-seq expression levels of the four TEs composing the “temperature” TE cluster under two heat-shock experiments. (h-i) Normalized RNA-seq expression levels of the TEs composing one “precipitation” TE cluster under exposure to biotic stress (*P. syringae*) or drought.

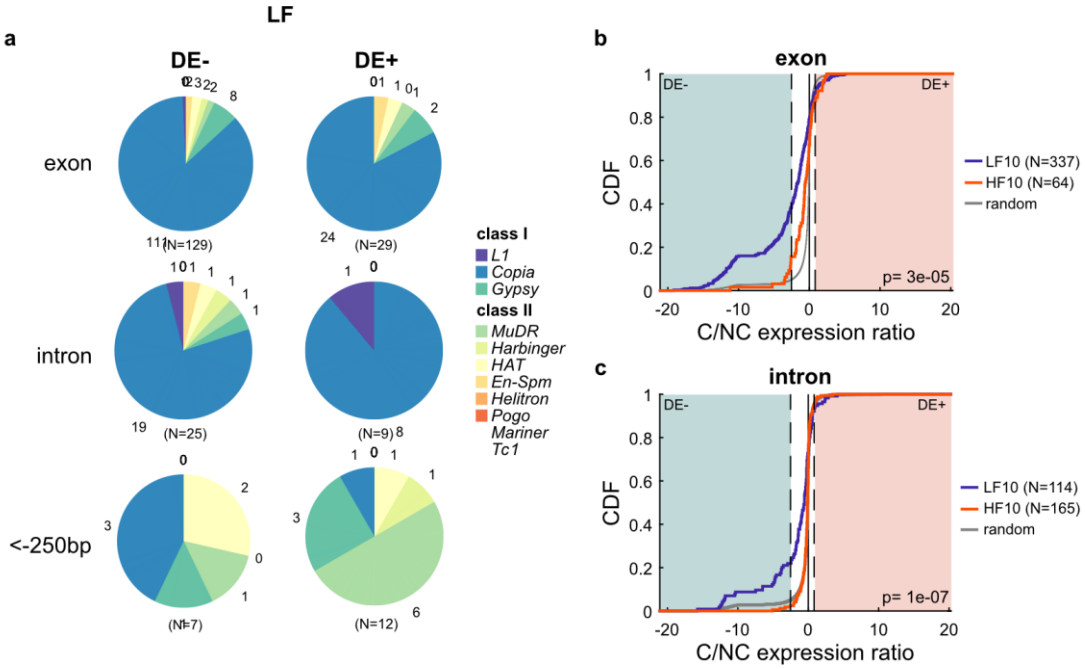

Figure S5.

(a) Contribution by TE superfamily to TE presence variants with negative (DE-) and positive (DE+) transcriptomic impacts in exons, introns, and close promoter (<-250bp).(b-c) Distribution of transcriptomic impacts (C over NC log ratios) for exonic and intronic TE presence variants at low-frequency (LF) vs high-frequency (HF) compared to random sampling of carriers and non-carriers. Dashed lines indicate top and bottom 5% values of random distribution, under and over which C/NC ratios are considered extreme.

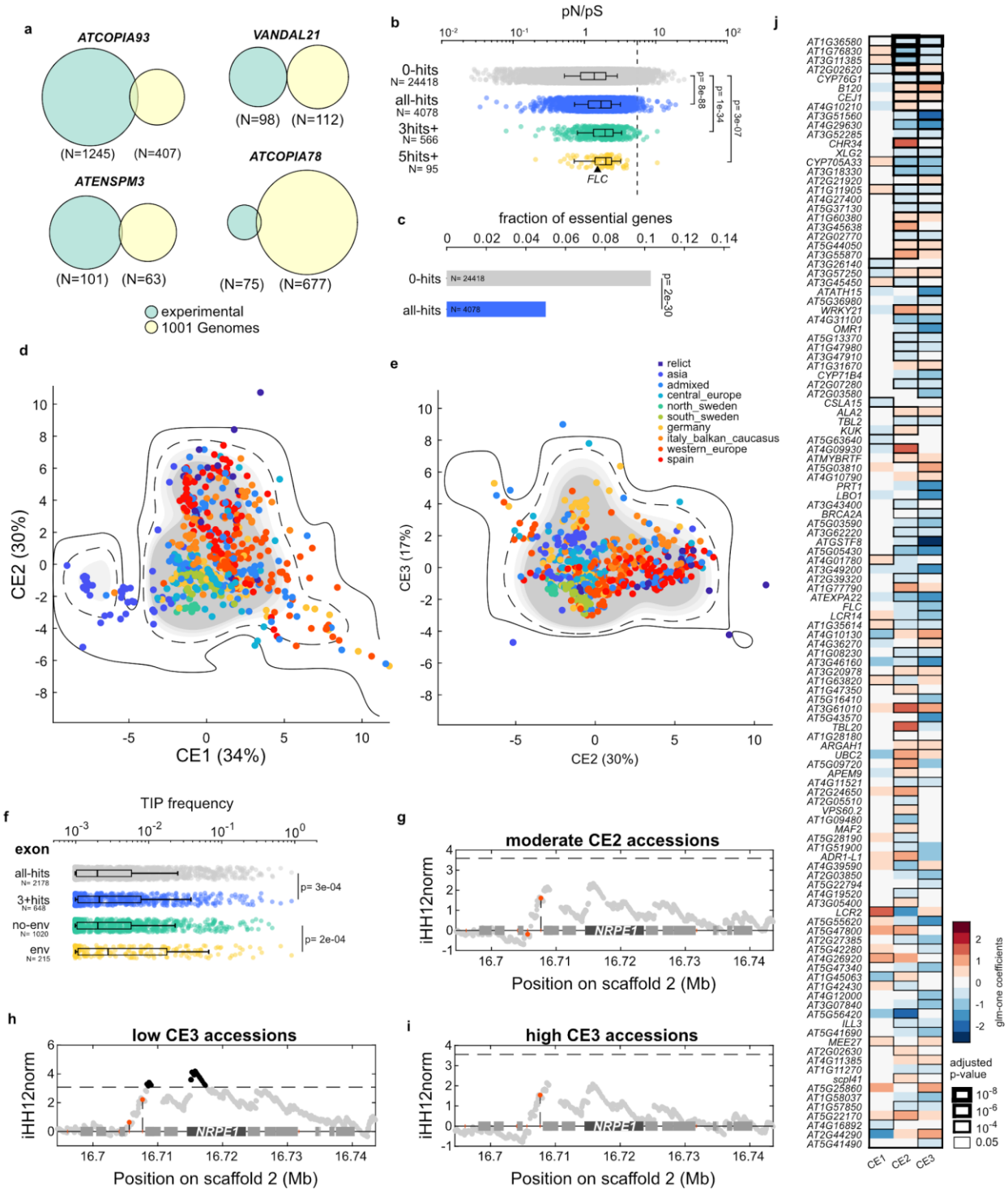

Figure S6.

(a) Genes visited within 250bp by *ATCOPIA93*, *ATCOPIA78*, *VANDAL21*, and *ATENSPM3* in experimental transposition accumulation lines (8) compared to that found in natural accessions. (b) Ratio of non-synonymous (missense, pN) against synonymous SNPs (pS) for genes never visited by TE insertions (0 hits), visited at least once (all-hits), at least thrice (3hits+), and 5 times or more (5hits+). P-values of

Wilcoxon test between distributions are indicated (c) Proportion of essential genes found within each category. P-values of Fisher exact tests between categories are indicated. (d-e) First three climatic envelopes (CE1-3) from principal component analysis of 19 BIO variables across 1047 accessions. (f) Frequency of TE insertions found within exons of genes visited at least once (all-hits), 3 times or more (3hits+), in association with a CE shift (env) or not (no-env). The p-values of Wilcoxon tests between distributions are indicated. (g) iHH12 values in moderate CE2 accessions (two middle quartiles) across the *NRPE1* region with in black indicated values above the genome-wide 1% threshold (dashed line). (h-i) iHH12 values in low and high CE3 accessions (two lower and two upper quartiles respectively) across the *NRPE1* region with in black indicated values above the genome-wide 1% threshold (dashed line). (j) Heatmap of associations in logistic GLM between presence of TE insertion within or near genes and the three climatic envelopes (CE1-3).

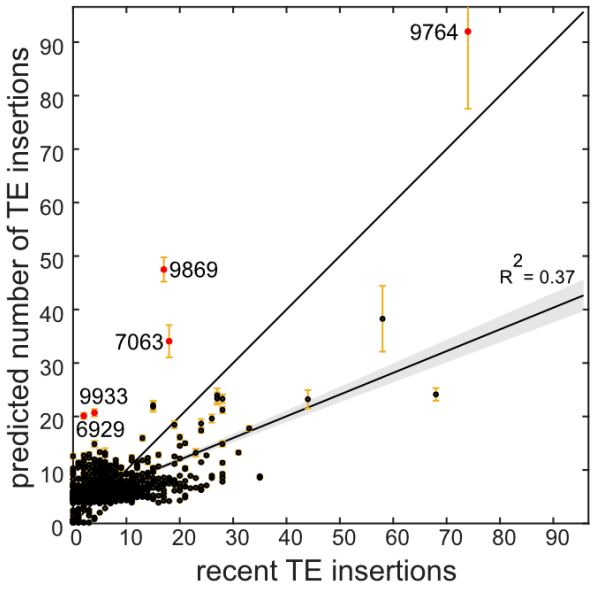

Figure S7.

(a) Average number of TE insertions predicted by accessions of 100 randomly sampled testing sets of 100 accessions of GLMs based on GxE interaction terms with coefficients trained over the remaining 947 accessions. Error bars represent standard error across predictions and shaded area represent 95% confidence intervals around linear regression.

References

1. Barrett R, Schluter D. Adaptation from standing genetic variation [Internet]. Vol. 23, Trends in Ecology & Evolution. 2008. p. 38–44. Available from: http://dx.doi.org/10.1016/j.tree.2007.09.008

2. Hermisson J, Pennings PS. Soft Sweeps [Internet]. Vol. 169, Genetics. 2005. p. 2335–52. Available from: http://dx.doi.org/10.1534/genetics.104.036947

3. Huang CRL, Burns KH, Boeke JD. Active transposition in genomes. Annu Rev Genet.

2012;46:651–75.

4. Friedli M, Trono D. The developmental control of transposable elements and the evolution of higher species. Annu Rev Cell Dev Biol. 2015 Sep 17;31:429–51.

5. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007 Dec;8(12):973–82.

6. Quadrana L, Silveira AB, Mayhew GF, LeBlanc C, Martienssen RA, Jeddeloh JA, et al. The Arabidopsis thaliana mobilome and its impact at the species level [Internet]. Vol. 5, eLife. 2016. Available from: http://dx.doi.org/10.7554/elife.15716

7. Zhang H, Lang Z, Zhu J-K. Dynamics and function of DNA methylation in plants. Nat Rev Mol Cell Biol. 2018 Aug;19(8):489–506.

8. Quadrana L, Etcheverry M, Gilly A, Caillieux E, Madoui M-A, Guy J, et al. Transposition favors the generation of large effect mutations that may facilitate rapid adaption. Nat Commun. 2019 Jul 31;10(1):3421.

9. Mirouze M, Reinders J, Bucher E, Nishimura T, Schneeberger K, Ossowski S, et al. Selective epigenetic control of retrotransposition in Arabidopsis. Nature. 2009 Sep 17;461(7262):427–30.

10. Tsukahara S, Kobayashi A, Kawabe A, Mathieu O, Miura A, Kakutani T. Bursts of retrotransposition reproduced in Arabidopsis. Nature. 2009 Sep 17;461(7262):423–6.

11. Miura A, Yonebayashi S, Watanabe K, Toyama T, Shimada H, Kakutani T. Mobilization of transposons by a mutation abolishing full DNA methylation in Arabidopsis. Nature. 2001 May 10;411(6834):212–4.

12. Singer T, Yordan C, Martienssen RA. Robertson's Mutator transposons in A. thaliana are regulated by the chromatin-remodeling gene Decrease in DNA Methylation (DDM1). Genes Dev. 2001 Mar 1;15(5):591–602.

13. Reinders J, Wulff BBH, Mirouze M, Marí-Ordóñez A, Dapp M, Rozhon W, et al. Compromised stability of DNA methylation and transposon immobilization in mosaic Arabidopsis epigenomes. Genes Dev. 2009 Apr 15;23(8):939–50.

14. Ito H, Gaubert H, Bucher E, Mirouze M, Vaillant I, Paszkowski J. An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress. Nature. 2011 Apr 7;472(7341):115–9.

15. Stuart T, Eichten SR, Cahn J, Karpievitch YV, Borevitz JO, Lister R. Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. Elife [Internet]. 2016 Dec 2;5. Available from: http://dx.doi.org/10.7554/eLife.20777

16. Baduel P, Quadrana L, Colot V. Efficient detection of transposable element insertion polymorphisms between genomes using short-read sequencing data [Internet]. Available from: http://dx.doi.org/10.1101/2020.06.09.142331

17. Keightley PD, Eyre-Walker A. Joint Inference of the Distribution of Fitness Effects of Deleterious Mutations and Population Demography Based on Nucleotide Polymorphism Frequencies [Internet]. Vol. 177, Genetics. 2007. p. 2251–61. Available from: http://dx.doi.org/10.1534/genetics.107.080663

18. Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, et al. The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. Science. 2010 Jan 1;327(5961):92–4.

19. Exposito-Alonso M, Becker C, Schuenemann VJ, Reiter E, Setzer C, Slovak R, et al. The rate and potential relevance of new mutations in a colonizing plant lineage [Internet]. Vol. 14, PLOS Genetics. 2018. p. e1007155. Available from: http://dx.doi.org/10.1371/journal.pgen.1007155

20. Matzke MA, Mosher RA. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity [Internet]. Vol. 15, Nature Reviews Genetics. 2014. p. 394–408. Available from: http://dx.doi.org/10.1038/nrg3683

21. Sasaki E, Kawakatsu T, Ecker JR, Nordborg M. Common alleles of CMT2 and NRPE1 are major determinants of CHH methylation variation in Arabidopsis thaliana. PLoS Genet. 2019 Dec;15(12):e1008492.

22. Jiao W-B, Schneeberger K. Chromosome-level assemblies of multiple Arabidopsis genomes reveal hotspots of rearrangements with altered evolutionary dynamics [Internet]. Vol. 11, Nature Communications. 2020. Available from: http://dx.doi.org/10.1038/s41467-020-14779-y

23. 1001 Genomes Consortium. Electronic address: magnus.nordborg@gmi.oeaw.ac.at, 1001 Genomes Consortium. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. Cell. 2016 Jul 14;166(2):481–91.

24. Wendte JM, Haag JR, Singh J, McKinlay A, Pontes OM, Pikaard CS. Functional Dissection of the Pol V Largest Subunit CTD in RNA-Directed DNA Methylation. Cell Rep. 2017 Jun 27;19(13):2796–808.

25. Quadrana L, Bortolini Silveira A, Mayhew GF, LeBlanc C, Martienssen RA, Jeddeloh JA, et al. The Arabidopsis thaliana mobilome and its impact at the species level. Elife [Internet]. 2016 Jun 3;5. Available from: http://dx.doi.org/10.7554/eLife.15716

26. O'Malley RC, Huang S-SC, Song L, Lewsey MG, Bartlett A, Nery JR, et al. Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape [Internet]. Vol. 165, Cell. 2016. p. 1280–92. Available from: http://dx.doi.org/10.1016/j.cell.2016.04.038

27. Kawakatsu T, Huang S-SC, Jupe F, Sasaki E, Schmitz RJ, Urich MA, et al. Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions. Cell. 2016 Jul 14;166(2):492–505.

28. Ietswaart R, Wu Z, Dean C. Flowering time control: another window to the connection between antisense RNA and chromatin. Trends Genet. 2012 Sep;28(9):445–53.

29. Whittaker C, Dean C. The FLC Locus: A Platform for Discoveries in Epigenetics and

Adaptation. Annu Rev Cell Dev Biol. 2017 Oct 6;33:555–75.

30. Van de Weyer A-L, Monteiro F, Furzer OJ, Nishimura MT, Cevik V, Witek K, et al. A Species-Wide Inventory of NLR Genes and Alleles in Arabidopsis thaliana. Cell. 2019 Aug 22;178(5):1260–72.e14.

31. Pennings PS, Hermisson J. Soft sweeps III: the signature of positive selection from recurrent mutation. PLoS Genet. 2006 Dec 15;2(12):e186.

32. Monroe JG, Powell T, Price N, Mullen JL, Howard A, Evans K, et al. Drought adaptation in by extensive genetic loss-of-function. Elife [Internet]. 2018 Dec 6;7. Available from: http://dx.doi.org/10.7554/eLife.41038

33. Chen I-C, Hill JK, Ohlemüller R, Roy DB, Thomas CD. Rapid range shifts of species associated with high levels of climate warming. Science. 2011 Aug 19;333(6045):1024–6.

34. Capblancq T, Fitzpatrick MC, Bay RA, Exposito-Alonso M, Keller SR. Genomic Prediction of (Mal)Adaptation Across Current and Future Climatic Landscapes [Internet]. Vol. 51, Annual Review of Ecology, Evolution, and Systematics. 2020. p. 245–69. Available from: http://dx.doi.org/10.1146/annurev-ecolsys-020720-042553

35. Schultz MD, He Y, Whitaker JW, Hariharan M, Mukamel EA, Leung D, et al. Human body epigenome maps reveal noncanonical DNA methylation variation. Nature. 2015 Jul 9;523(7559):212–6.

36. Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T. deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W187–91.

37. Stroud H, Greenberg MVC, Feng S, Bernatavichute YV, Jacobsen SE. Comprehensive analysis of silencing mutants reveals complex regulation of the Arabidopsis methylome. Cell. 2013 Jan 17;152(1-2):352–64.

38. Chang CC. Data Management and Summary Statistics with PLINK. Methods Mol Biol. 2020;2090:49–65.

39. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly . 2012 Apr;6(2):80–92.

40. Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, et al. Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. Front Genet. 2012 Mar 15;3:35.

41. Lee C-R, Svardal H, Farlow A, Exposito-Alonso M, Ding W, Novikova P, et al. On the post-glacial spread of human commensal Arabidopsis thaliana [Internet]. Vol. 8, Nature Communications. 2017. Available from: http://dx.doi.org/10.1038/ncomms14458

42. Kang HM, Sul JH, Susan K Service, Zaitlen NA, Kong S-Y, Freimer NB, et al. Variance component model to account for sample structure in genome-wide association studies

[Internet]. Vol. 42, Nature Genetics. 2010. p. 348–54. Available from: http://dx.doi.org/10.1038/ng.548

43. Prunier JG, Kaufmann B, Fenet S, Picard D, Pompanon F, Joly P, et al. Optimizing the trade-off between spatial and genetic sampling efforts in patchy populations: towards a better assessment of functional connectivity using an individual-based sampling scheme [Internet]. Vol. 22, Molecular Ecology. 2013. p. 5516–30. Available from: http://dx.doi.org/10.1111/mec.12499

44. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21.

45. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.