Genetic and environmental modulation of transposition shapes the evolutionary potential of Arabidopsis thaliana

Pierre Baduel, Basile Leduque, Amandine Ignace, Isabelle Gy, José Gil, Olivier O. Loudet, Colot Vincent, Leandro Quadrana

Genetic and environmental modulation of transposition shapes the evolutionary potential of *Arabidopsis thaliana*

Baduel P.[1], Leduque B.[1], Ignace A.[2], Gy I.[2], 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, AgroParisTech, Université Paris-Saclay, 78000, Versailles, France*

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

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

Abstract

Background: How species can adapt to abrupt 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 and bio-climatic data available for >1,000 wild *A. thaliana* accessions to determine the rate of transposable element (TE) mobilization in natural settings and its potential to create adaptive variation.

Results: We show that TEs insert throughout the genome at almost the same rate as SNPs and we confirm this observation experimentally. Mobilization activity of individual TE families varies greatly between accessions, in association with genetic as well as environmental factors and we identified several gene-environment interactions. Although the distribution of TE insertions across the genome is ultimately shaped by purifying selection, reflecting their typically strong deleterious effects when located near or within genes, numerous recent TE-containing alleles show signatures of positive selection. Moreover, high transposition appears to be positively selected at the edge of the species' ecological niche. Based on these results and mathematical modeling, we predict higher transposition activity in Mediterranean regions within the next decades in response to global warming and thus an acceleration in the creation of large effect alleles.

Conclusions: Our study reveals that TE mobilization is a major generator of genetic variation in *A. thaliana* that is finely modulated by genetic and environmental factors. These findings and modeling indicate that TEs may be essential genomic players in the demise or rescue of native populations in times of climate crises.

Keywords: Transposable elements, genome evolution, population genetics, epigenomics, adaptation, climate change

Background

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 produce neutral variants predominantly. 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: through their mobilization, they can disrupt or alter genes as well as their expression in multiple ways and because of their dispersion across the genome, they provide many opportunities for the creation of chromosomal rearrangements through ectopic recombination (3,4).

Eukaryotic TEs belong to two broad classes: DNA transposons, which use a cut and paste mechanism for their mobilization, and retrotransposons, which move through an RNA intermediate (5). These two classes are further divided into TE superfamilies and families based on particular sequence features, such as the presence or absence of Long Terminal Repeats (LTRs) in the case of retrotransposons (5).

Population genomic surveys of TE insertion polymorphisms (TIPs) revealed that TEs from many families insert preferentially towards genes and that insertions are rapidly purged from gene-rich regions (6,7), suggesting that natural transposition tends to generate alleles with strong deleterious effects. 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 (8). DNA methylation deficiencies do not lead by themselves to widespread TE re-mobilization (9–14), indicating that additional factors control transposition. For instance, mobilization of the LTR-retroelement *ATCOPIA78* was shown experimentally to require heat-shock in addition to impaired RdDM activity (15), indicating that at least in this case, both genetic and environmental determinants are decisive.

Although there is evidence of sustained transposition activity in *A. thaliana* (6,16), a comprehensive understanding of the factors involved is missing. Here, we leveraged the sampling depth of the *A. thaliana* 1001 Genomes project (1001genomes.org) to identify the major factors associated with recent TE mobilization and to determine the impact of the thousands of insertions near or within genes it generated. We then use ecological modelling to explore the evolutionary trajectories resulting from this recent activity and to predict the consequences of global warming on the creation of genetic variation through transposition in the near future.

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 TE insertion polymorphisms (TIPs). TIPs were identified using a bioinformatic pipeline (17) combining SPLITREADER (6) and TEPID (16) , which efficiently detect in resequenced genomes presence of non-reference TE sequences and absence of reference TE sequences, respectively. Considering the latter is essential for the full characterization of recent TE mobilization 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. These were contributed almost entirely by the two superfamilies of LTR-retrotransposonss *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, presence variants for the DNA transposon *HELITRON* superfamily were ignored as they are not efficiently detected by our pipeline, unlike absence variants. Together with *MuDR* and *hAT, HELITRON* make up over half of the 1,624 absence variants detected in total (Fig. 1a-S1a). TIPs are broadly distributed across the genome, with the notable exception of those produced by the *GYPSY* superfamily of LTR retrotransposons, which are enriched in pericentromeric regions. The broad distribution of TIPs confirms previous observations obtained using a smaller number of non-reference genomes (6) and is in stark contrast with the relative paucity of reference TE sequences along the chromosome arms and their high density in pericentromeric regions (Fig. 1b). Furthermore, the site frequency spectrum (SFS) of TIPs, which we calculated using the number of informative genomes at each site, is heavily skewed towards low values compared to bi-allelic SNPs (Fig. 1c). Specifically, one third of TIPs have a minor allele frequency of less than 0.2% s and >80% of these were missed

in previous analyses based on ~200 genomes (6). The excess of low-frequency TIPs compared to SNPs suggests strong negative effects especially when genic, as only nonsense SNPs have lower frequencies than TIPs when these are located within 250bp of genes. (Fig. 1d-e-S1b). To estimate how deleterious TE insertions are, we computed the distribution of fitness effects (DFE) of each category of variants by comparing their SFS with that of synonymous SNPs to control for recent demographic changes (DFE-alpha; (18) as they can affect SFSs in ways that resemble selection. Using this approach, we estimate that >99% of TIPs within 250bp of a gene are deleterious ($N_e s>1$), compared to 48% and 83% of missense and nonsense SNPs, respectively (Fig. 1e). Thus, almost all TE insertions within or nearby genes produce sizable deleterious effects.

Low-frequency TIPs are thought to reflect recent transposition events, not yet purged by natural selection (6,19–21). To determine the relationship between age and TIP frequency, we considered all TIPs shared by at least two genomes and estimated their age by first calculating for each TE insertion the number of SNPs accumulated in its vicinity (35kb on either side; 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 (22). Using this approach, we found a positive correlation between predicted age and TIP frequency (R^2=0.4; Fig. S1c). However, this analysis indicates that TIP frequency is an imperfect proxy for age as only half of TIPs that segregate at frequencies below 1.5% are less than 5,000 years old (Fig. 1f).

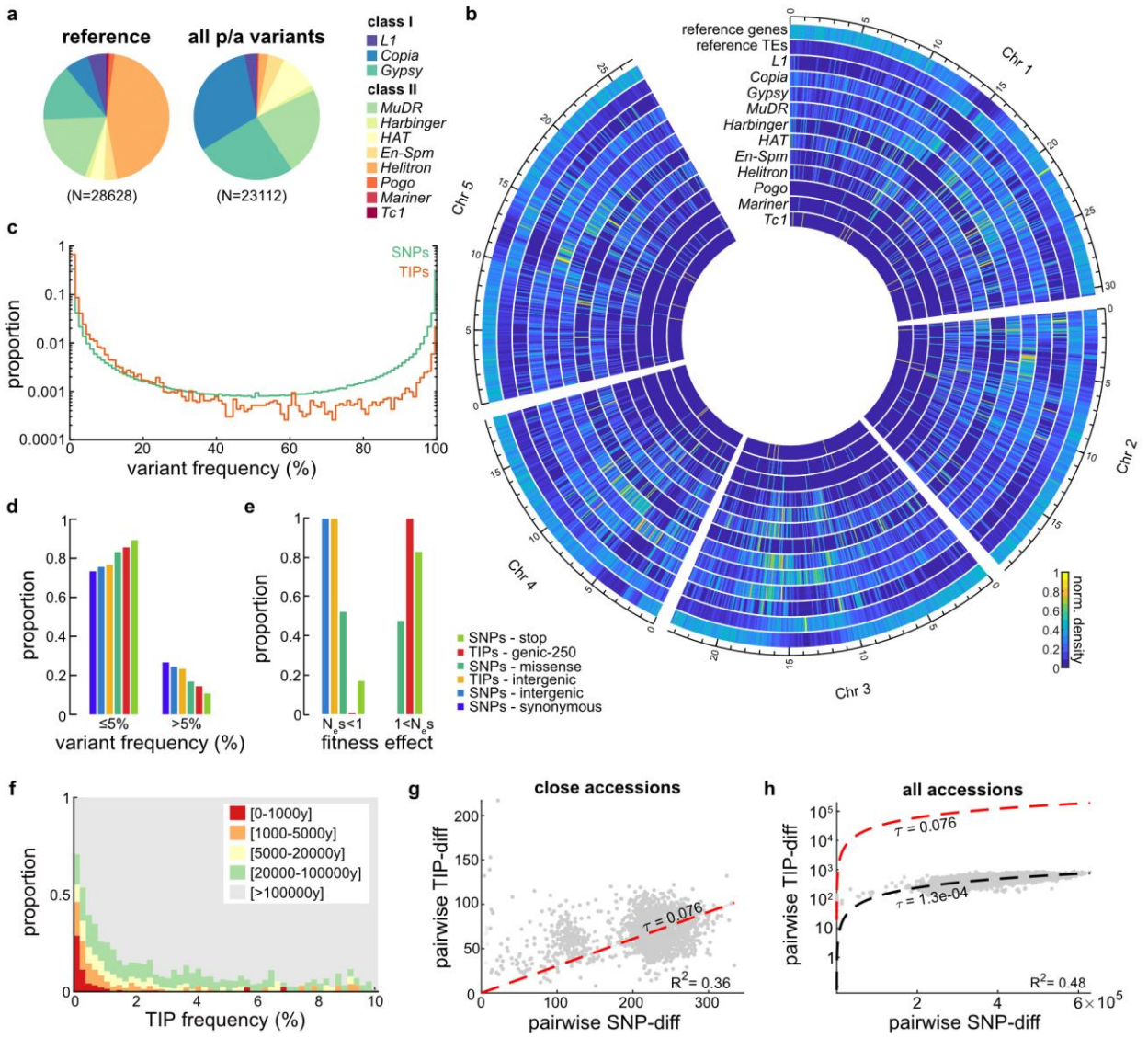

Figure 1. Recent TE mobilization at the species level

(a) Contribution by superfamily to TEs annotated in the reference genome (TAIR10) and TE insertion polymorphisms (TIPs). (b) Density of 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.

We next estimated the substitution rate for TE insertions using closely related accessions (i.e. accessions that differ by <500 SNPs, Fig. S1d) and found it to be almost a third (0.076 +/- 0.0012 per genome per generation; Fig. 1g, see Methods) of that calculated for single nucleotides (23). In contrast, 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. 1h). Indeed, we predict (see Methods) that >99.8% of TE insertions that occur in nature are eventually eliminated by natural selection, a percentage higher than for missense and even nonsense SNPs (68.9% and 92.5%, respectively; Fig. S1d-e). Moreover, TE insertion substitutions within or near genes occur at rates ten-fold higher than that of nonsense and of the same order to that of missense SN substitutions (0.025 vs 0.002 and 0.038 mutations per genome per generation, respectively; Fig. S2a,d,e). Together, these results indicate that TE mobilization is a major contributor of large-effect genetic variants in *A. thaliana*.

Genetic basis of variable transposition

To explore further the mutation pressure caused by TE mobilization in nature, we first carried out principal component analysis using the number of TIPs with a MAF ≤5% per TE family. Results revealed a significant structuration of overall transposition activity in relation to the 10 main genetic groups defined in *A. thaliana* (24), with Relicts, Asian and South-Sweden accessions being the most contrasted (Fig. 2a and S3a).

To identify potential genetic modifiers of transposition activity, we performed a genome-wide association study (GWAS) using as a quantitative trait the total number of most recent TE insertions per genome (MAF lower than 0.2% and <1000 years old or private, referred hereafter as very recent TIPs) across all TE families (Fig. S3b-c). GWAS revealed a single major peak (Fig. 2b), which suggests a simple genetic architecture of global transposition activity. This association peak spans the gene *NUCLEAR RNA POLYMERASE E1* (*NRPE1;* Fig. 2c), which encodes the

largest subunit of RNA Pol V, essential to RdDM (25) and which was previously identified in GWAS as a major determinant of CHH methylation at TEs targeted by RdDM (26). Moreover, the non-reference allele*,* called *NRPE1'*, which was linked to reduced CHH methylation, is associated here with a 240% increase in transposition activity (Fig. 2d), thus strongly supporting a causal role. GWAS performed at the TE superfamily level revealed in addition that associations with *NRPE1* are in fact strongest for *MuDR*, which have RdDM- rather than CMT2-dependent CHH methylation, (Fig. 2h, see Methods), effectively indicating causality (Fig. S3e).

Inspection of long-read sequencing data (27) from an accession (Sha) carrying the *NRPE1'* allele revealed extra polymorphisms beyond the SNPs and short-indels identified by the 1001 genomes project (24). Specifically, the *NRPE1'* allele of Sha contains also a 9-bp in-frame deletion in the 17aa-repeat domain and three deletions in the QS tail of the C-terminal domain (CTD; respectively 6bp, 60bp, and 9bp-long, Fig. 2e-S3g). In fact, the three QS deletions (delQS1-2-3) of the CTD define a suballele of *NRPE1'*, which we named *NRPE1'_ΔQS* in contrast to *NRPE1'_Δrep* that carries the repeat deletion and one or two of the three QS deletions at most (Fig. 2e). Moreover, the two derived *NRPE1* alleles resemble those produced experimentally in the reference accession Col-0 (28) and are associated with similar effects as those caused by these mutant alleles on CHH methylation of RdDM TE targets, with a more pronounced loss when the QS and repeat domains are deleted together (Fig. 2f-g-S3f). Remarkably, the two naturally truncated alleles explain by themselves at least 17% of the variation in transposition activity at the species level (Fig. S3d, see below).

To evaluate directly the impact of impaired RdDM on overall transposition activity, we carried out TE-sequence capture (9) on pools of 1,000 seedlings derived from WT and *nrpd1* mutant parents of the Col-0 reference accession grown under standard conditions. A total 99 novel TE insertions were detected in the *nrpd1* sample (Fig. 2i), a 80% increase compared to the WT. Higher transposition in *nrpd1* was most prominent for *GYPSYs* and *MuDRs*, consistent with most of their

CHH methylation being RdDM-dependent, unlike that of *COPIAs*, which are substantially targeted by CMT2 also (Fig. S3h). Furthermore, the rate of transposition determined experimentally is of the same order of the substitution rate for TE insertions we estimated at the species level (0.06 in WT vs 0.08 per genome per generation, respectively), thus providing direct experimental support for the latter.

Using the same approach, we also measured TE mobilization in three natural accessions, including Cvi and Sha, which respectively carry the *NRPE1'_Δrep* and *NRPE1'_ΔQS* alleles. Sha exhibited the highest transposition rate overall (Fig. 2j), which was mainly driven by *GYPSYs* and *MuDRs* (Fig. S3i), thus resembling in this respect the *nrpd1* mutant in Col-0. This observation further supports a causal role for *NRPE1'* in increased TE mobilization.

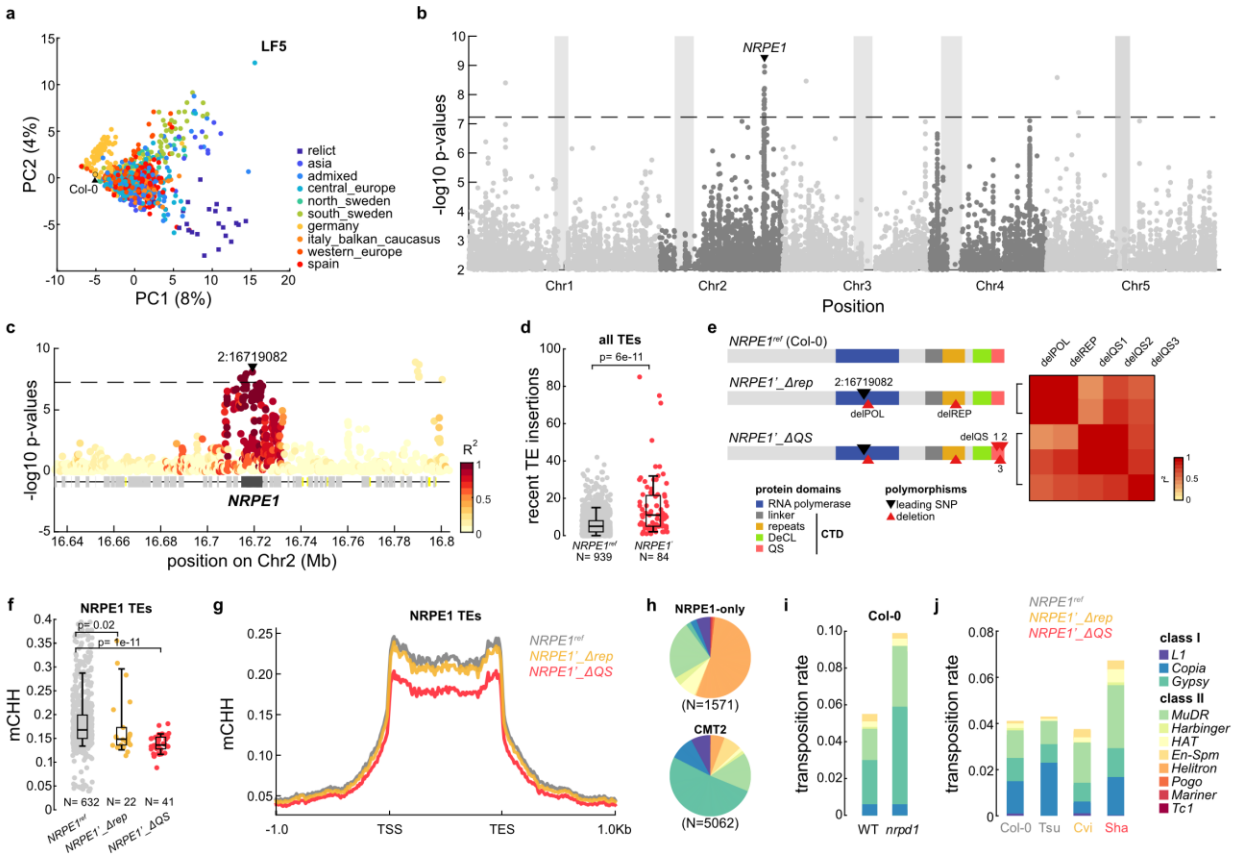

Figure 2. Genetic basis of variable transposition

(a) PCA of mobilome composition based on recent TIPs (MAF lower than 5%). Different genetic groups 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. The p-values of Wilcoxon tests between distributions are indicated. (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. The p-values of Wilcoxon tests between distributions are indicated. (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 includes population structure and allelic variation at *NRPE1* (see Methods). Importantly, we also considered the possibility of GxE interactions involving *NRPE1 (15)*. The GLM revealed that, while variation in transposition activity between accessions is explained predominantly (27%) by genetic backgrounds and allelic variation at *NRPE1*, seasonality of precipitation (BIO15) and diurnal temperature range (BIO02) contribute another 9% of this variation (6.3% and 2.7%; Fig. 3a,d, S4a). 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 overall explain an additional 4.2% of variation in TE mobilization. In fact, differential TE mobilization in association with these two bio-variables is only observed for accessions carrying *NRPE1'* alleles (Fig. 3c-d),

which extends to natural settings the experimental observation that mutations in the RdDM pathway modulate transposition in response to environmental changes.

To move beyond this global picture, we analysed 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 analysis, we detected for 57 TE families significant associations with at least one environmental variable (Fig. 3e). Consistent with the GLM results, positive association with precipitation seasonality (BIO15) is most prevalent at the individual TE family level (44 out of 57 TE families; Fig. 3e). Moreover, we identified four clusters of TE families that share similar environmental associations. One small cluster of four *COPIA* TE families stands out by exhibiting the strongest associations, all of which concern temperature bio-variables. Consistent with previous work (29), *ATCOPIA78* belongs to this last cluster. The present analysis reveals in addition that the association of *ATCOPIA78* mobility with temperature is only observed in the *NRPE1'* background (Fig. S4b-c), which mirrors the observation that *ATCOPIA78* transposition can only be induced following heat-shock in RdDM sensitized backgrounds (15).

To assess experimentally the extent of the interaction between RdDM and environmental stress, we compared transposition using TE-sequence capture in pools of 1,000 seedlings of *nrpd1* and WT Col-0 parents exposed this time to heat-shock or flagellin, a bacterial peptide known for triggering plant biotic stress response (see Methods). Transposition increased in the *nrpd1* mutant but not in WT following heat-shock (Fig. 3f) and this increase was not restricted to *ATCOPIA78* but concerned also most notably the *GYPSY* and *MuDR* families *ATGP1* and *VANDAL6*, respectively (Fig. S4d). Similar but weaker trans-family sensitization by impaired RdDM was also observed following exposure to flagellin and concerned the same families in many cases, with the COPIA family *META1* being one notable exception that showed increased transposition only following flagellin treatment (Fig. S4d).

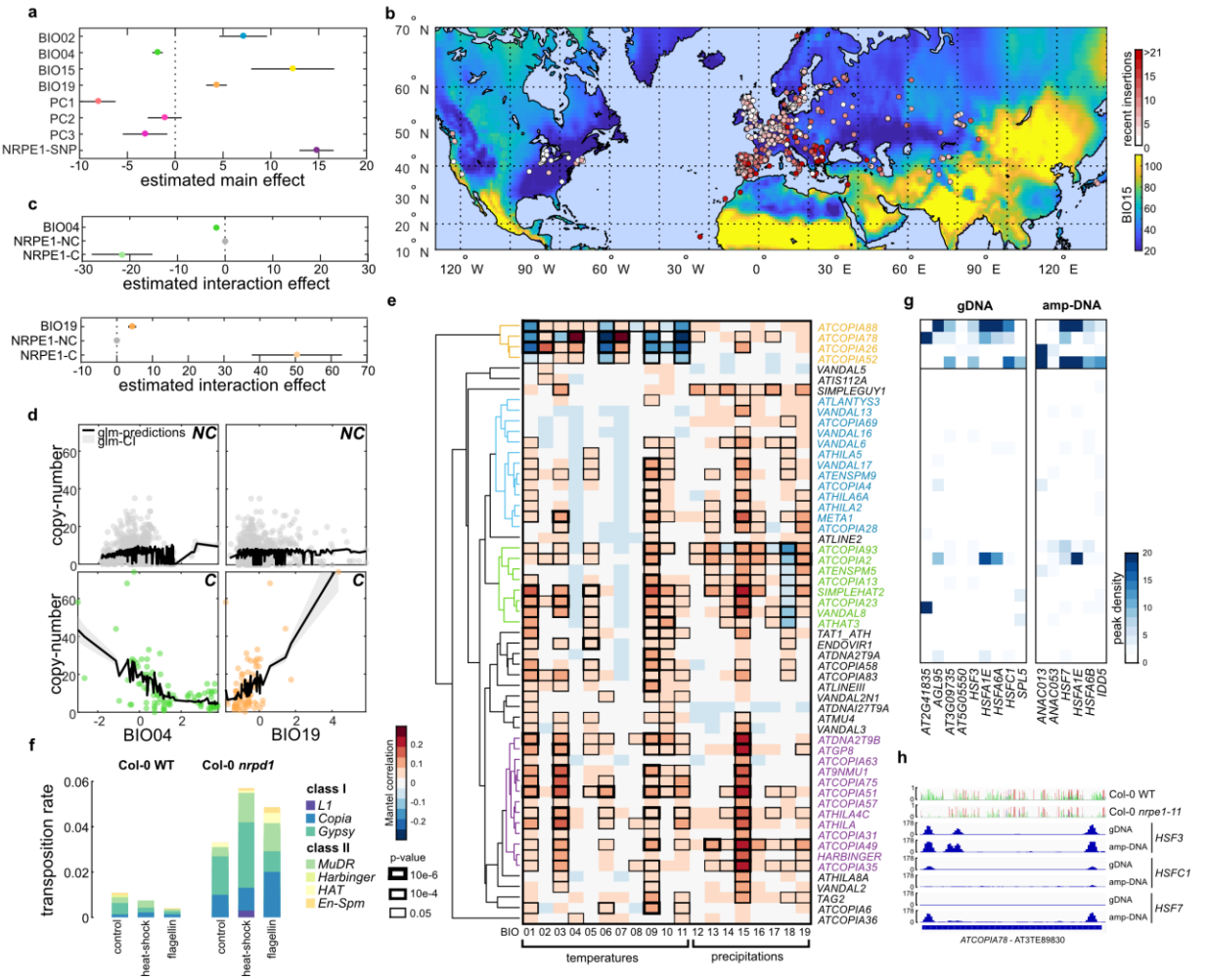

Figure 3. Environmental modulation of TE mobilization

(a) Marginal effect at the mean of each of the variables considered in the GLM of very recent transposition: the first three principal components of the kinship matrix (PC1-2-3), the *NRPE1* locus, and the BIO02, BIO04, BIO15, and BIO19 variables. (b) Number of very recent TE insertions detected across the world and levels of precipitation seasonality (BIO15). (c) Estimated interaction effect of BIO04 and *NRPE1* (upper) and BIO19 and *NRPE1* (bottom). (d) Scatter plot of very recent transposition against BIO04 (left) and BIO19 (right) in non-carriers (NC, up) and carriers (C, down) of the derived *NRPE1'* alleles. GLM predictions and confidence-intervals are indicated in black and grey, respectively. (e) Directional Mantel associations for 77 TE families between very recent transposition and 19 WorldClim bio-variables (1970-2000). Dendrogram of hierarchical clustering of coefficient correlations. The four main clusters are indicated (colors). (f) Transposition rates in 1,000 F1 plants derived from Col-0 WT and *nrpd1* parents grown in standard 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 in Col-0 using native or amplified (i.e. stripped of all DNA methylation) genomic DNA for 469 transcription factors (TFs) (30). TE families in the three "precipitation" clusters share few enrichments for sites bound by specific TFs (TFBSs; Fig. S5a), which suggests that their environmental responsiveness, notably to biotic stress or drought in the case of the cluster containing *ATCOPIA93* (Fig. S5h-i), can be acquired through a diverse set of transcriptional wirings. In contrast, the four *COPIA* TE families belonging to the "temperature" cluster 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-S5a-c) and another three TFs encoded by genes induced transcriptionally under different heat-shock treatments (ANAC013, ANAC053, SPL5; Fig. S5d-e). Transcriptome data for the reference accession Col-0 indicate also that three of the four *COPIA* families in this cluster are transcriptionally up-regulated under heat-shock (Fig. S5f-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 and S5b-c), consistent with the sensitivity to DNA methylation reported for these TFs (30). Finally, *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 *COPIAs* 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

To determine the mutation load generated by transposition, we measured the transcript levels of genes affected by the presence of TIPs near or within them. 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, thus complicating comparisons. 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 (9), 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 (31) 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 (~75%) TE insertions within exons are associated with reduced transcript levels (Fig. 4e) and almost all of these are contributed by *COPIAs* (111 out of 124; Fig. S6a). Furthermore, in ~20% of cases, the TE-containing allele is an effective knock-out (Fig. S6b). TE insertions in introns are also frequently associated with reduced gene expression, but the effects are typically of smaller magnitude (Fig. S6c). Despite these general trends, a few TE insertions are associated with increased transcript

levels, and these are contributed mainly by *MuDRs* and tend to reside within the 5' UTR or the promoter regions of genes (Fig. 4e and S6a), consistent with the insertion preferences exhibited by this TE superfamily (9). Altogether, these observations indicate that almost a quarter of mutations generated by TE mobilization in nature are likely to have major and mostly negative effects on gene transcript levels, with the remaining being either inconsequential or associated with increased expression in very rare cases.

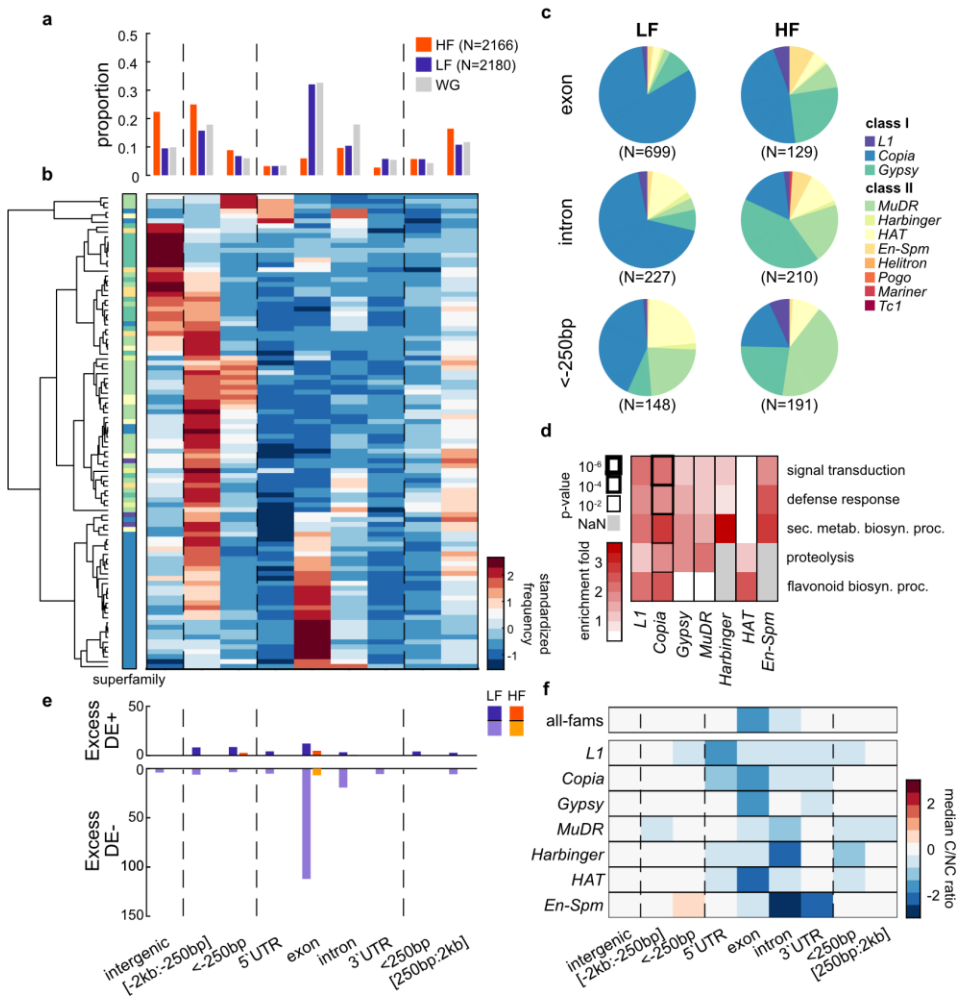

Figure 4. TE mobilization mainly generates highly deleterious genic mutations

(a) Fraction of low- and high-frequency TE presence variants overlapping genic annotations (exons, introns, 5′ or 3′ UTRs) or located near genes (upstream or downstream within 250bp, or within 2kb) or intergenic (>2kb away from nearest gene) compared to the genomic proportion 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) Distribution of low- and high-frequency TE presence variants in exons, introns, and promoter regions (<-250bp) for each TE superfamily. (d) GO enrichments of LF presence variants within genes. (e) Excess of extreme expression log ratios between carriers (C) and 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 determine 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). HF exonic insertions are particularly rare (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). Likewise, there are no HF intronic insertions associated with major reduction in gene expression (Fig. 4e and S6c). Conversely, we recovered as many or more intronic or promoter variants at low and high-frequency (respectively 227 vs 210 for introns and 148 vs 191 for the promoters; Fig. 4a,c), consistent with their minimal or positive transcriptomic impact, except in the case of *COPIA* insertions (Fig. 4e,f, S6a). Whether any of the high-frequency insertions are under positive selection remains to be determined. Together, these findings confirm that the majority of genic insertions are under strong purifying selection (Fig. 1e).

Contribution of recurrent genic insertions to local adaptation

Consistent with the conclusion that most genic insertions are deleterious and because of the marked insertion preference of *COPIAs* towards responsive genes and away from essential genes (9,29), the set of gene loci with TIPs is much smaller than expected by chance (4078 vs 9090 +/- 45, see Methods) and depleted in essential genes (Fig. S7c). Conversely, TIP-containing gene loci with at least three distinct TE-containing alleles are more abundant than expected by chance (566 vs 285 +/- 8, Fisher exact test p=5e-51). As these alleles tend to be low frequency

variants, they could reflect either recurrent targeting because of insertion preferences, relaxed purifying selection, and/or diversifying selection. We can rule out an important role of insertion preferences, given the minimal overlap between gene loci visited in the lab and in nature for four TE families most active in these two settings (Fig. 5a,S7a). Furthermore, the fact that pseudogenes are not strongly enriched in TE insertions (206 vs 167 expected by chance) indicates that multiple hits cannot solely result from relaxed purifying selection. Moreover, because 99% of gene loci with TIPs have pN/pS values under the upper 1% genome-wide threshold (Fig. S7b), they do not appear to be functionally decaying. In fact, the number of TIPs at a given gene locus correlates positively with pN/pS, suggesting instead that recurrent visits are functionally relevant and reflect diversifying selection. Consistent with this interpretation, we observed that for a quarter of the loci visited recurrently, associations between the different TE-containing alleles at the locus and gene expression are congruent (Fig. 5a). Congruence is most striking at *FLOWERING LOCUS C* (*FLC*), which encodes a key repressor of flowering and is one of the main genetic determinants of natural variation in the onset of flowering (32). Specifically, we identified 16 distinct TE-containing *FLC* alleles in total (Fig. 5a), each characterized by a unique insertion within the first intron. This intron is essential to the environmental regulation of *FLC* expression (33) and collectively, the 16 TE insertions are associated with lower expression and earlier flowering (Fig. 5a,c). Together with previous detailed analyses (9,29), these results indicate that recurrent TE mobilization within *FLC* may be a major contributor of local adaptation.

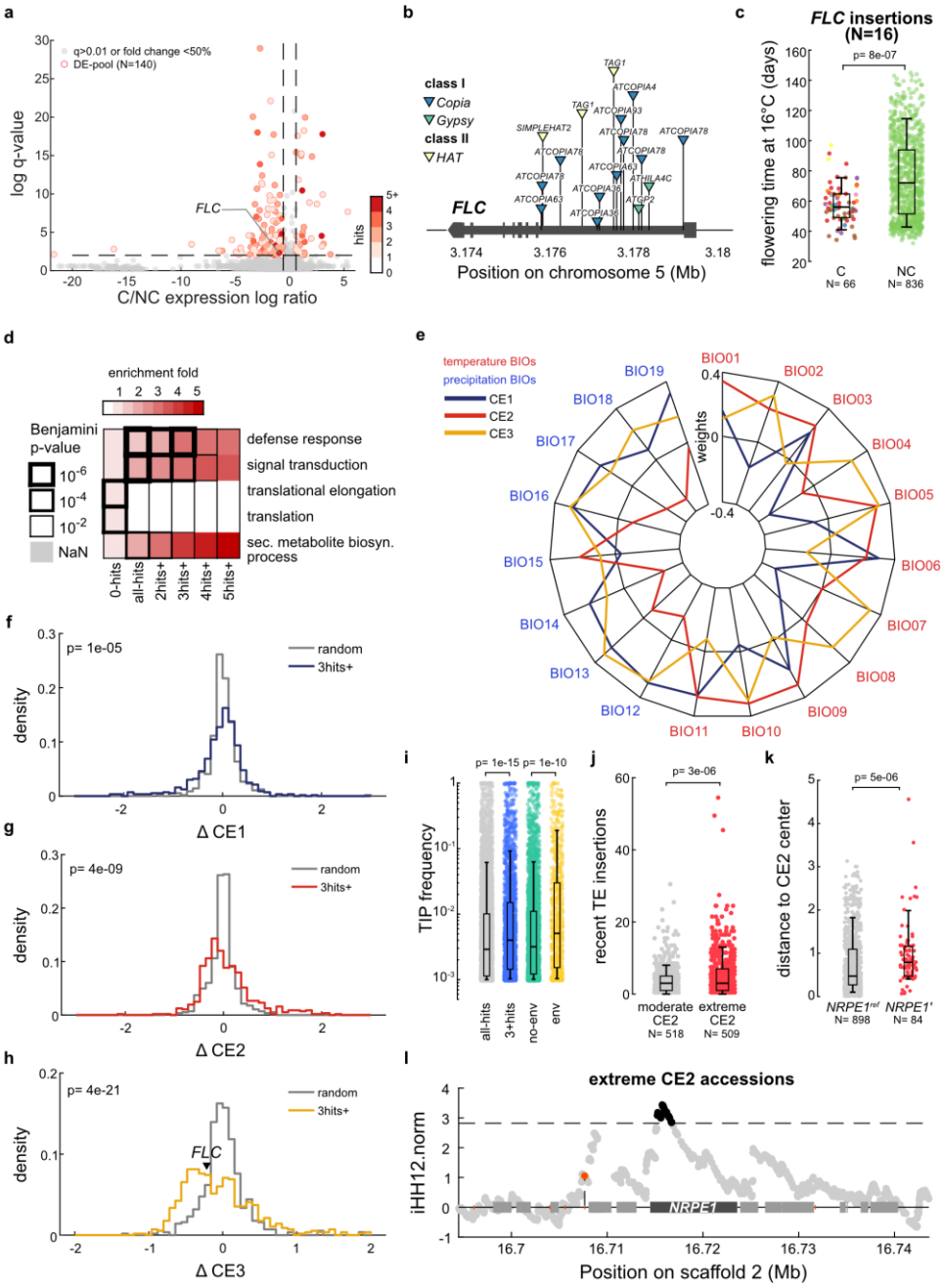

Figure 5. Contribution of transposition to local environmental adaptation

(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) Boxplot of numbers of very recent TE insertions in extreme or moderate CE2 accessions. The p-values of Wilcoxon tests between distributions are indicated. (k) Boxplot of distance to CE2 center (absolute zscored CE2) for carriers of the reference *NRPE1*^{ref} and derived *NRPE1'* alleles. The p-values of Wilcoxon tests between distributions are indicated. (l) iHH12 values in extreme CE2 accessions (upper and lower quartiles of CE2.z) across the *NRPE1* region with in black indicated values above the genome-wide 1% threshold (dashed line).

As expected, genes with multiple TIPs are strongly depleted in genes associated with core cellular processes, notably translation (Fig. 5d). 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 (34). To determine if the recurrence of TE insertions at loci other than *FLC* could also suggest a contribution to local adaptation, we searched for potential associations with environmental differences. We first summarized the 19 WorldClim bioclimatic variables into three climatic envelopes (CEs, Fig. 5e) that together explain >80% of the climate niche variations observed across the locations of the 1001 Genomes accessions (Fig. S7d-e). 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 then tested for each of the 566 multi-hit gene loci whether the TE-containing alleles are associated with an environmental shift using a logistic GLM that incorporates population structure (see Methods). In total, 137 gene loci showed significant associations, mainly with CE2 and/or CE3 (Fig. 5f-h). These associations are robust, given that none were identified when the GLM was repeated using random permutations of the environmental variables. Moreover, consistent with the notion that TE-containing alleles of *FLC* are locally adaptive (9,29), they are found preferentially in parts of the species range characterized by milder winters (low CE3), where they may enable flowering in the absence of vernalization thanks to their lower expression. More

generally, TE-containing alleles are systematically at higher frequency when multiple-hit genes show evidence of environmental associations (Fig. 5i), even in the case of exonic TE insertions (Fig. S7f). Together, these observations suggest that recurrent mutations via TE insertion are under positive selection (35) and contribute to local adaptation to divergent environments, in line with previous results indicating that repeated loss-of-function mutations are adaptive (36).

Recent transposition tends to be higher among accessions with extreme CE2 values (Fig. 5j), which are enriched for the *NRPE1'* allele (Fig. 5k). To explore the possibility that *NRPE1'* and hence increased TE mobilization is under positive selection in these environments, we quantified haplotype-length decay using iHS and iHH12, two measures used to identify hard and soft sweeps, respectively (37,38). Whereas iHS did not reveal any hard sweep, consistent with the wide distribution of the *NRPE1'* allele, iHH12 uncovered marks of soft sweep for accessions located in extreme but not moderate CE2 environments (Fig. 5j and S7g). Together, these findings suggest that transposition is a powerful generator of locally adaptive alleles in challenging environments, whose fine tuning by the RdDM machinery is itself the target of natural selection.

Increased TE mobilization under future climates

Given the strong environmental sensitivity of TE mobilization, the mutation pressure generated by transposition could be significantly affected by climate change. To investigate this possibility, we first considered forecasts of future climates under the most pessimistic gas emission scenario during the next 60-80 years (CMIP6 SSP5-8.5) for each of the locations occupied by *A. thaliana* accessions of the 1001 Genomes Project. Consistent with the expected global increase in the frequency of hotter and drier summers, CE2 was the most impacted environmental envelope for eight out of the ten genetic groups (Fig. 6a). We then evaluated the predictive power of three GLMs of recent TE mobilization that are based either on genetic variables alone (G), or in

combination with the most significant bio-variable (BIO15; G+E) or else with all four major bio-variables together with their GxE interactions (Fig. 3) in 100 random testing sets of 100 accessions each (see Methods). We confirmed that full GLMs provide the most robust estimates of the number of recent TE insertions (Fig. 6b). Applying the full model, we predict across most locations an increase in transposition activity, which is particularly pronounced in the Mediterranean region (Fig. 6c). Nonetheless, given the GxE interactions involving *NRPE1,* TE mobilization is expected to decrease in Mediterranean populations carrying the *NRPE1'* alleles. Conversely, these alleles should exacerbate transposition at higher latitudes, such as in Sweden. Given that *NRPE1'* appears to be under positive selection in extreme CE2 accessions (Fig. 5j), we anticipate that the GxE interactions involving these derived alleles will play an important role in the survival potential of native populations in the face of climate change.

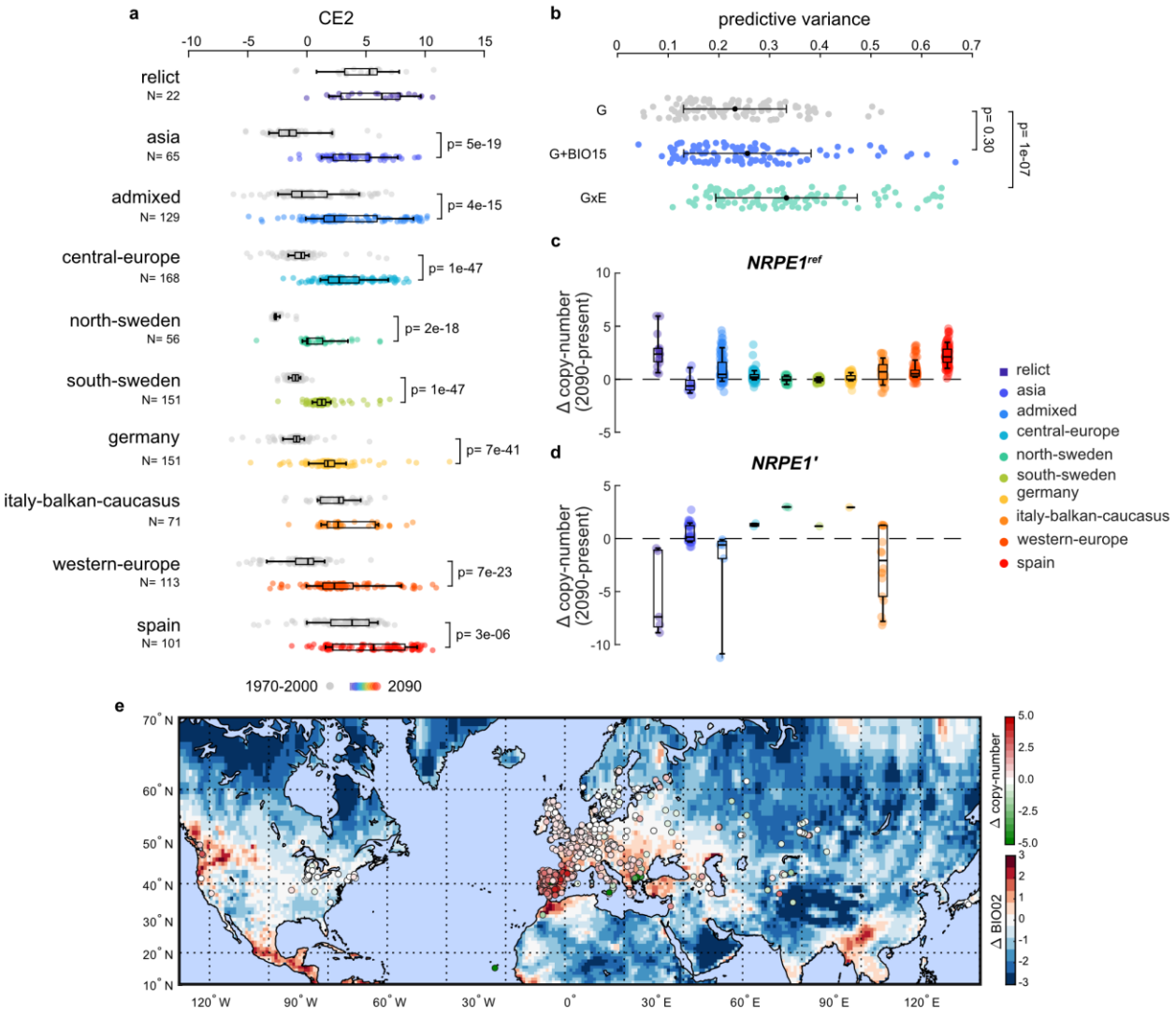

Figure 6. Increased TE mobilization under future climates

(a) Forecasted change of climatic envelope CE2 by genetic group under average CMIP6 GCM for 2081-2100 compared to recent climate (1970-2000). (b) Predictive variance of the numbers of recent copies on 100 random testing sets of 100 accessions for 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-e) Predicted change by genetic group in copy-numbers using GxE GLM with future climate predictions for 2081-2100 for (c) carriers of the reference *NRPE1* allele and (d) carriers of the derived *NRPE1'* alleles. (e) 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 ranges (39). Standing genetic variation is generally thought to be the main source of rapid adaptation to environmental changes (1) and genomic studies aimed at estimating the evolutionary potential of native populations in future climates have mostly focused on SNPs (40). Here, we set out to assess the contribution of the rare and typically large effect alleles created by TE insertions and to determine if transposition activity in nature can ensure a sustained supply of potentially adaptive *de novo* variants in response to the environment.

High transposition rate in nature and the evolutionary fate of new TE insertions

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 and 3). 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 new mutations in *A. thaliana*. Furthermore, unlike SNPs which are randomly distributed along the genome and predominantly neutral or mildly deleterious, the most transpositionally active TE families show strong insertion preferences towards genes. As a result, TIPs within or near genes occur at rates similar to that of missense SN substitutions. However, because TE-containing alleles typically have major functional impacts, they are more rapidly purged by natural selection than nonsense substitutions (Fig. 1, S2, and 4). Together, these findings indicate that TE mobilization is a major source of large-effect mutations and that sequenced genomes are mere snapshots of their rapid evolutionary turnover.

In spite of the major deleterious effects typically associated with TE insertions within or near genes, we identified over one hundred genes recurrently visited by TEs at the borders of the environmental niche of *A. thaliana* and with signatures of positive selection for the insertion alleles (Fig. 5). The power of transposition to generate adaptive variation is most remarkably illustrated by the multiple independent TE insertions we identified at *FLC,* specifically in accessions from parts of the species range with mild winters, consistent with its major role in the alignment of flowering time with seasons (32). More generally, our results suggest that leveraging TE mobilization to generate adaptive allelic variation *de novo* is faster than relying on introgression of standing genetic variation from distant populations. This scenario stands in sharp contrast with established views on rapid adaptation (1) and repeated evolution (41). However, most TE-containing alleles under positive selection are at relatively low-frequency, suggesting that they mainly contribute to local micro-evolutionary responses. This pattern is also consistent with observations from catalogues of adaptive loci (42) where large-effect mutations are important contributors of rapid adaptations (e.g. (43)) but not at longer evolutionary timescales, two scenarios reconciling the macro-mutationism of Goldsmith (44) with the infinitesimal model of Fisher (45).

Modulation of transposition in nature

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 we uncovered includes a truncated form that causes lower methylation at TEs targeted by RdDM and higher levels of transposition (Fig. 2). Previous work indicated that alleles of *CMT2* are also major determinants of CHH methylation variation in *A. thaliana* (26). However, we failed to detect any association between *CMT2* and recent TE mobilization, either because of insufficient resolution of our GWAS or else because of a more prevalent role of *NRPE1*-dependent CHH methylation in controlling TE

mobilization. Although this last hypothesis remains to be tested experimentally, the observation that the *CMT2'* and *NRPE1'* alleles rarely occur together in nature (26) is consistent with this allelic combination being selected against, possibly because a further increase in transposition activity is not sustainable in nature.

Our study revealed that the environment is also an important modulator of transposition activity in nature, which is potentiated further by allelic variation at *NRPE1* (Fig. 3). Moreover, we identified a hardwired network of TFs linked to the environmental responsiveness of several TEs (Fig. 3), including heat-shock factors known to target *ATCOPIA78* (43), whose binding over regions targeted by RdDM is enhanced *in vitro* in the absence of DNA methylation (Fig. 3). Also, we note that impaired RdDM is not sufficient in itself to trigger the mobilization of *COPIAs*, which are the main contributors of large-effect genic insertions (Fig. 2 and 4). Indeed, the high transpositional activity associated with the most severe truncation of *NRPE1'* is observed in extreme environments (Fig. 5). Together, these findings suggest that in these environments, *NRPE1'* is akin to mutator alleles described in bacteria. Such alleles are typically favored by selection in harsh environments when the advantage of beneficial mutations is greater than the cost of the higher mutation load they also generate (46–48). Consistent with this view, *NRPE1'* shows signatures of positive selection in extreme environments (Fig. 5). This last result also supports the notion that evolvability, defined as the ability of organisms to produce adaptive and heritable phenotypic variation, is subject to Darwinian selection (49).

Unlike classical mutator alleles, which are evolutionary transient because of the ever increasing mutation load they generate (50,51), *NRPE1'* is an ancestrally derived allele that is retained at low frequency across the species range. This long-term retention may be helped by the high selfing rate of *A. thaliana*, which decreases genetic heterogeneity especially at the borders of the species range, thus limiting the speed at which a mutator allele could be lost through outcrossing.

Determining the precise conditions allowing the evolution and persistence of *NRPE1'* will be key to understanding how universal such mutator alleles might be and their role in adaptation.

Forecasting transposition in future climates

Our mathematical modeling predicts a major role for transposition in shaping the mutational pressure in changing climates. Indeed, we forecast higher transposition rates in Mediterranean populations in response to global warming (Fig. 6) and thus an accelerated production of major-effect alleles. Some of the new alleles generated in this manner may rescue native *A. thaliana* populations from extinction, notably when they lack advantageous standing variation (52). However, the increased mutational pressure might also expose populations to a higher risk of extinction by mutational meltdown (53), which is expected to be more important in isolated small populations, where the efficiency of selection is limited. Yet, we showed that mutations generated by TE mobilization typically have strong fitness effects and are rapidly purged by natural selection (Fig. 1), consistent with theoretical predictions even for small, selfer populations (53,54)transposition is unlikely to lead to mutational meltdown. Further supporting this notion, we found active TE mobilization in North-American accessions, which were introduced on the continent on the continent during the 17th century from a handful of European individuals (23). Incidentally, the colonization of North-America by this population may help to solve the genetic paradox of invasive species, where despite the lack of genetic variation, colonizing individuals are able to adapt to the very environment they are invading (55,56). There, TE mobilization could be seen as a form of genetic bet-hedging strategy where, despite its strongly deleterious effects for a significant fraction of the offsprings, it provides unique opportunities to extensively explore the phenotypic landscape and thus reach adaptive optimas in divergent environments.

Conclusions

We demonstrate that TEs constitute a major and tunable source of large effect mutations in response to environmental challenges. Our findings as well as modeling provide a first indication that TEs may prevent the demise of native populations at evolutionary risk in the face of climate change, with broad implications for biodiversity.

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 (17). 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 they 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 segregates 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 (31). 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 (57). Bigwigs were generated using methylpy `allc-to-bigwig`. Merged bigwigs were then processed and plotted in metaplots over all NRPE1-targeted TEs using deepTools (58) functions `computeMatrix` and `plotProfile`. BS-seq data from the experimental *nrpe1* allelic series were obtained from (28) and processed similarly. NRPE1-targeted TEs were defined as overlapping with DMRs identified in the *nrpe1_11* mutant line (28) while CMT2-targeted TEs were defined from *cmt2* DMRs (59).

Genomic analyses

The SNP vcf file was obtained from 1001genomes.org (24) and genome-wide pairwise divergences were calculated across all pairs of accessions using the `allvsall --sample-diff counts-only` option of PLINK2 (60) 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 (0.2511 per genome per generation; 2.11E-9 per site per generation; (23). 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 synonymous SNPs, which are expected to be neutral. Indeed, these SNPs are overrepresented relative to all SNPs among distant accessions when compared to their respective proportions among close accessions (Fig. S2c). We used this discrepancy to estimate the scaling factor of the substitution rate that takes into account the average effect of selection on SNPs (Fig. 1h and S2c-e).

Pairwise divergence were calculated within 70kb windows surrounding each TE insertion site between all carriers of the TE insertion using PLINK2 (60) .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) (22).

SNPs were annotated using snpEff (61), and sifted by functional effect using snpSift (62) (`"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 (60) 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 (18) with a two epochs model to take into account the recent population expansion of *A. thaliana (63)*. 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.

Metrics of haplotypic decay (iHS and iHH12) were calculated using selscan (64) by chromosome after phasing biallelic SNPs with missing genotyping rates under 0.2 (plink option --geno) and MAF over 0.001 with shapeit (65). Chromosomal calculations were then normalized together using selscan's companion program norm (64).

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 (66) 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. S3) were represented by approximating the GLM with a linear model and averaging the effect of all the other variables using the MATLAB function plotEffects.

Plant growth

Seeds from four accessions (Col-0, Tsu-0, Cvi-0 and Shahdara) were grown in a controlled design aimed at amplifying successive generations under non-selective conditions (long days). Stratified seeds are first germinated in vitro under control conditions (½ MS media). At the fully developed cotyledons stage (5 days after sowing -DAS-), seedlings are transferred on a new plate containing ½ MS media supplemented with 1% sucrose. After 2 weeks, plants are then transferred to soil in individual pots for setting seeds in a growth room. 5 individuals (lines) are randomly selected for the next generation. This was repeated for 2 successive generations for each accession.

To study the effect of stress on transposition, plants were germinated at 23°C:19°C in long-days (16 h:8 h light:dark) on ½ MS plates then 2-weeks old seedlings were transferred to liquid ½ MS media (0.1% Agar) either pure (control and heat-shock) or containing 1µM of flagellin (flg-22). After 1 day, heat-shocked seedlings were transferred for 24h at 6°C then 24h at 37°C then returned to 23°C:19°C conditions. After 6 days, all seedlings were transferred to soil and plants were then grown to maturity in long-days at 24°C:22°C long-days.

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 (9,29). 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 15 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 (67) 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* (30) for 529 TFs. We processed this data using a modified version of the bioinformatics pipeline implemented by (30) 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 (Table S4). Expression level was calculated by mapping reads using STAR v2.5.3a (68) 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 (69).

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 inputs in the GLMs trained using current climate and estimate the future TE mobilization predicted by the model. Our model assumed that genetic structure will remain unchanged over this short evolutionary time. 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 evaluate the prediction behavior (over or underestimating) using a linear regression (Fig. S8). We considered as outliers the five accessions whose standardized residuals in the linear regression were greater than four, which we removed from further predictive analyses.

Statistical analyses

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

Availability of data and materials

The datasets generated and analysed during the current study are available in the European Nucleotide Archive (ENA) under project [PRJEBXXXXX].

Declarations

Author contributions

PB, VC and LQ conceived the project. PB performed all bioinformatic and statistical analyses. PB, BL and LQ performed TE sequence capture experiments. AI, IG, and OL, performed generational amplifications of natural accessions. BL analyzed Col-0 transcriptomic data. JG analyzed DAP-seq data. PB, VC and LQ interpreted the data and wrote the manuscript. All other authors reviewed the manuscript.

Funding

This work was supported by grants from the Centre National de la Recherche Scientifique (MOMENTUM program, to L.Q.) and the Agence Nationale de la Recherche (project MEMOSTRESS, grant no. ANR-12-ADAP-0020-01 to VC and OL). PB was supported by a postdoctoral fellowship (code SPF20170938626) from the Fondation pour la Recherche Médicale (FRM). Work in the Colot lab is supported by Investissements d'Avenir ANR-10-LABX-54 MEMO LIFE, 506 ANR-11-IDEX-0001-02 PSL* Research University. The IJPB benefits from the support of Saclay Plant Sciences-SPS (ANR-17-EUR-0007).

Competing interests

The authors declare that they have no competing interests

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.

Supplementary Figures

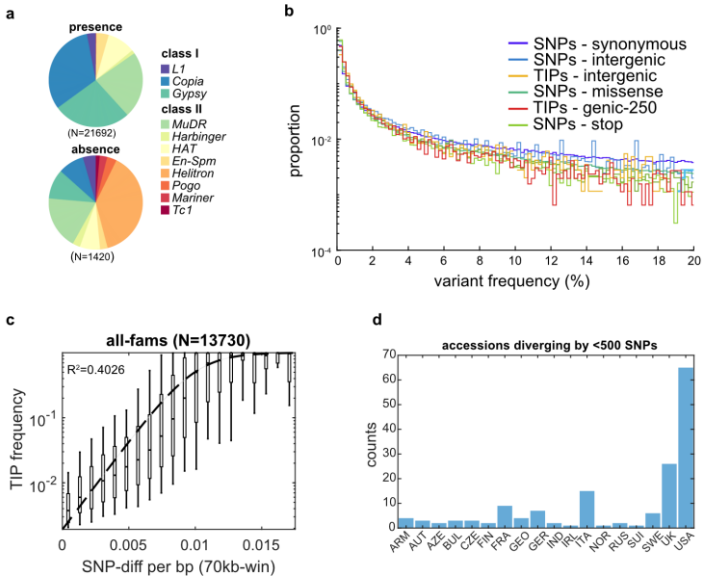

Figure S1.

(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) Correlation between TIP frequency and maximum pairwise SNP differences observed within 70kb between any two carriers for all non-private TIPs. (d) Distribution of closely related accessions by genetic group.

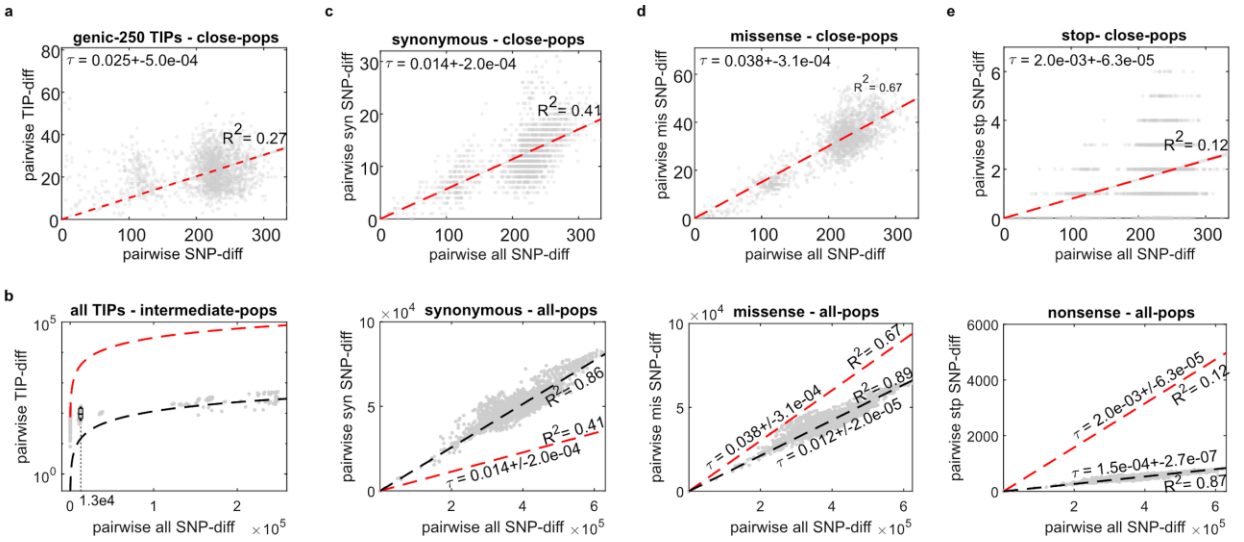

Figure S2.

(a) Pairwise differences in TIPs within and near genes (<250bp) and SNPs for all accessions diverging by <500 SNPs. Regression line and confidence intervals are indicated in red and gray, respectively. The substitution rate of this category of TIPs calculated from the substitution rate of 2.11e-9 per site per generation measured by Exposito-Alonso et al. (23) scaled by the regression slope is indicated with its 95% confidence intervals. (b) Pairwise differences in TIPs and SNPs between accessions. Regression lines between all and closely related accessions are shown in black and red, respectively. Boxplot of pairwise TIPs differences observed within 116 pairs of intermediate accessions (diverging by 10,000 to 15,000, on average 12,619 SNPs genome wide, i.e. ~50,000 generations). Mean TIPs pairwise differences in these accessions represents 2.75% of the expectation based on closely related accessions. (c-e) Pairwise differences in respectively synonymous, missense, and nonsense SNPs versus all SNPs between close (top) and all (bottom) accessions. Regression lines between all and closely related accessions are shown in black and red, respectively. Substitution rates for each category of SNPs calculated from the regression slopes and the substitution rate of all SNPs are indicated with their 95% confidence intervals. For closely related accessions we used the substitution rate of 2.11e-9 per site per generation measured by Exposito-Alonso et al. (23). For all accessions, we rescaled this substitution rate to take into account the effect of natural selection estimated from the putatively neutral synonymous SNPs (see Methods).

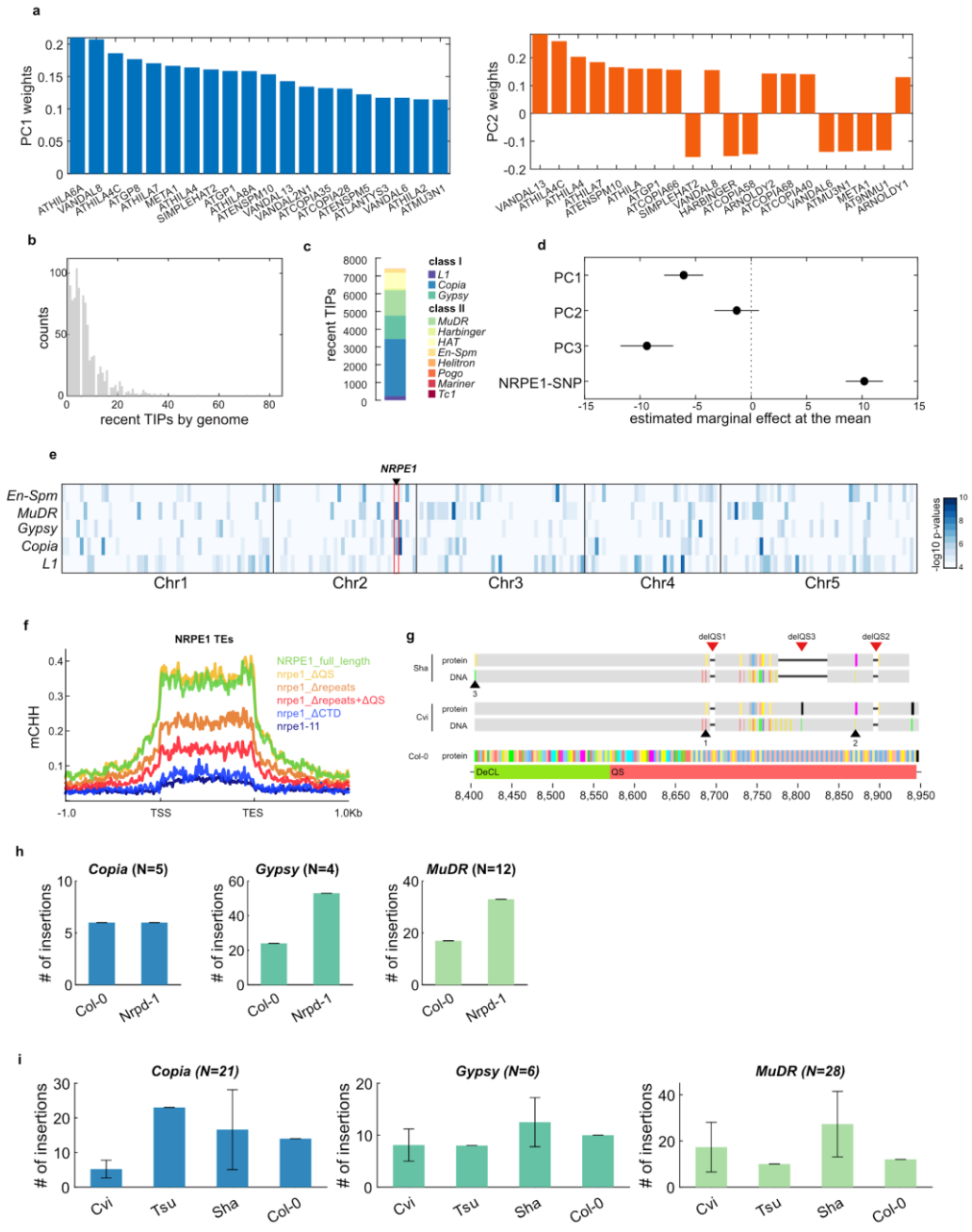

Figure S3.

(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. (g) Detailed alignment of DNA and protein sequences of the last exon of *NRPE1* in Col-0, Sha (27), and Cvi. The three deletions in the QS domains (6bp at 2:16723152, 60bp at 2:16723235, and 9bp at 2:16723351) are indicated with red arrows and their closest tagging SNPs with black arrows and corresponding numbers. (h) Transposition

rates of the three most mobile superfamilies (*COPIA, GYPSY,* and *MuDR*) in 1,000 F1 plants derived from WT or *nrpd1* Col-0 parents grown in control conditions. (i) Transposition rates of the three most mobile superfamilies (*COPIA, GYPSY,* and *MuDR*) in 1,000 F1 plants derived from Cvi, Tsu, Sha, and Col-0 parents grown in control conditions.

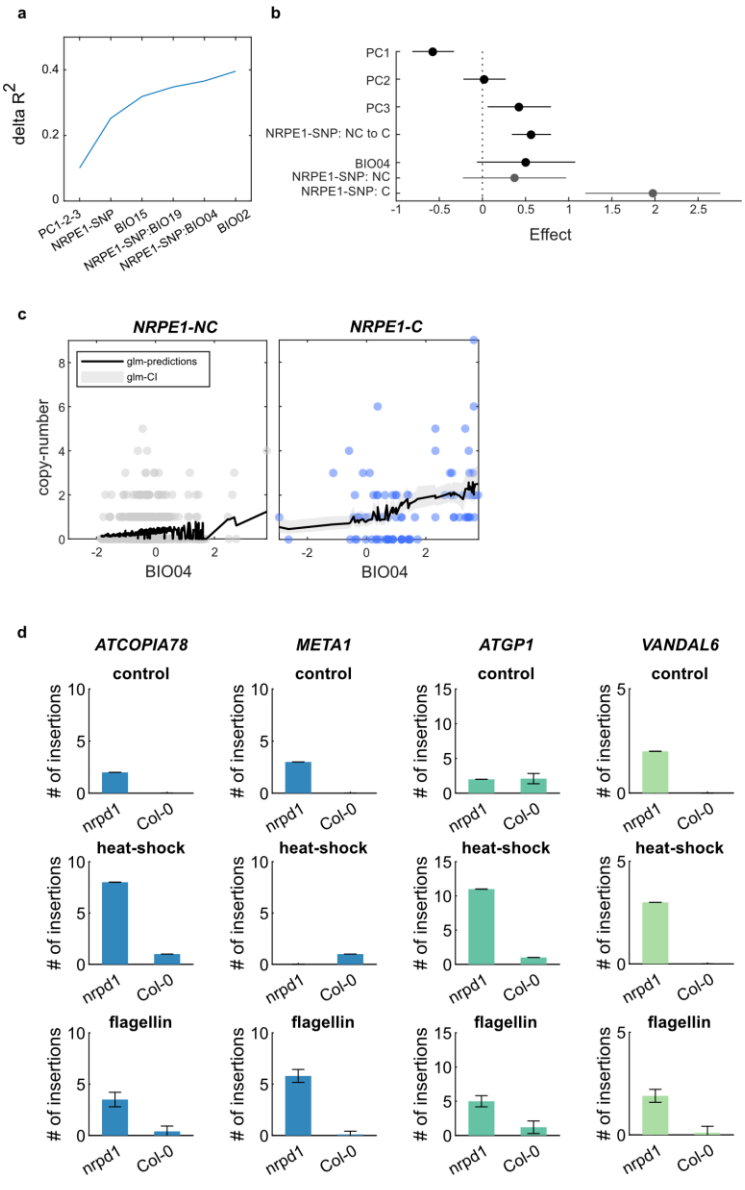

Figure S4.

(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.

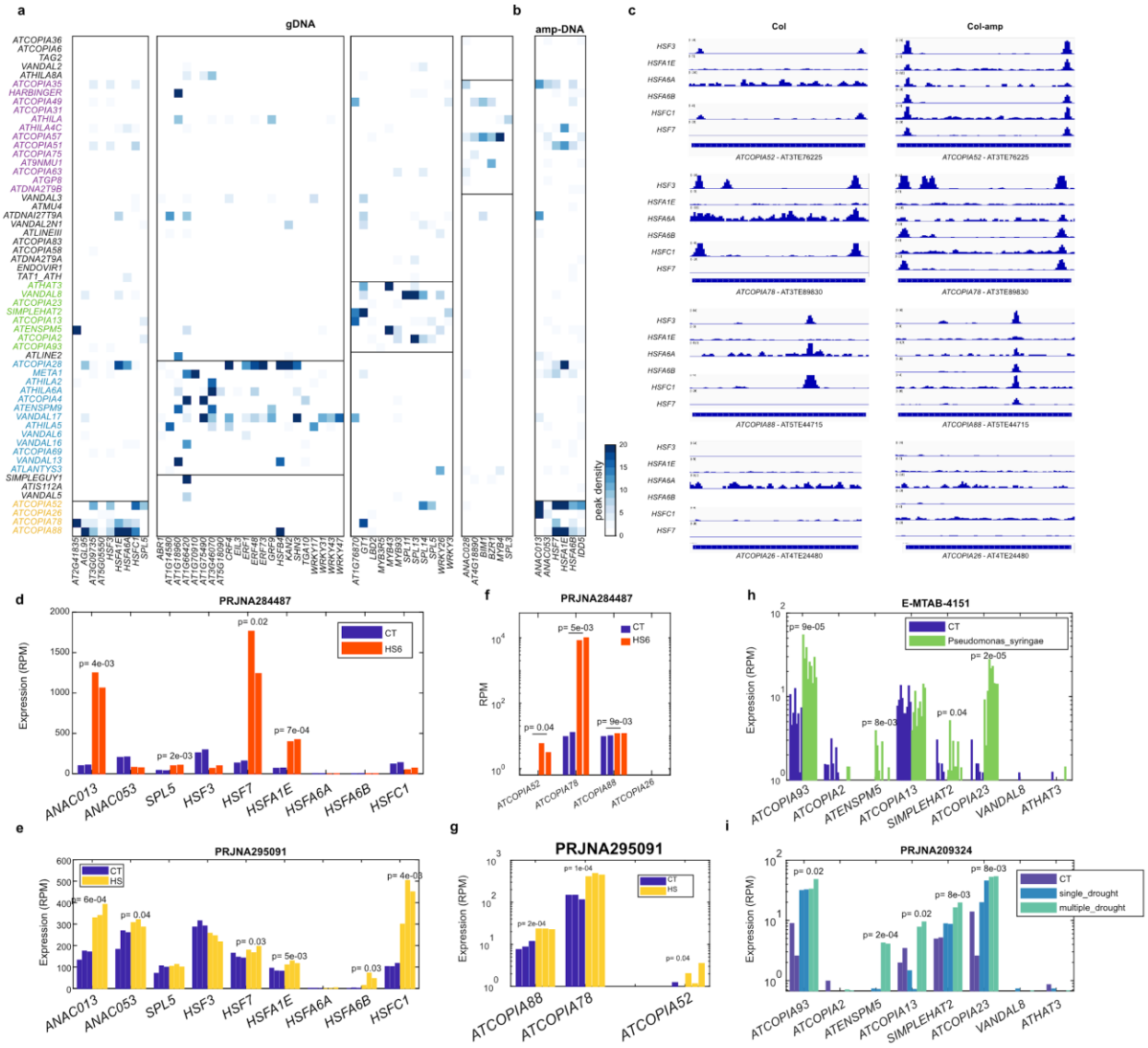

Figure S5.

(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.

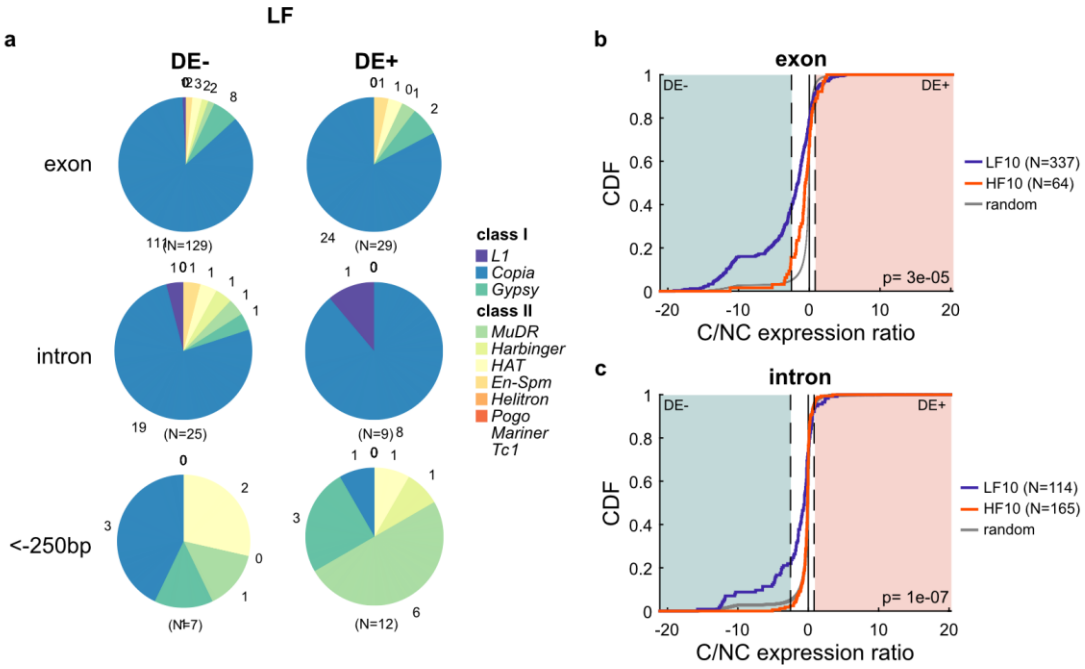

Figure S6.

(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.

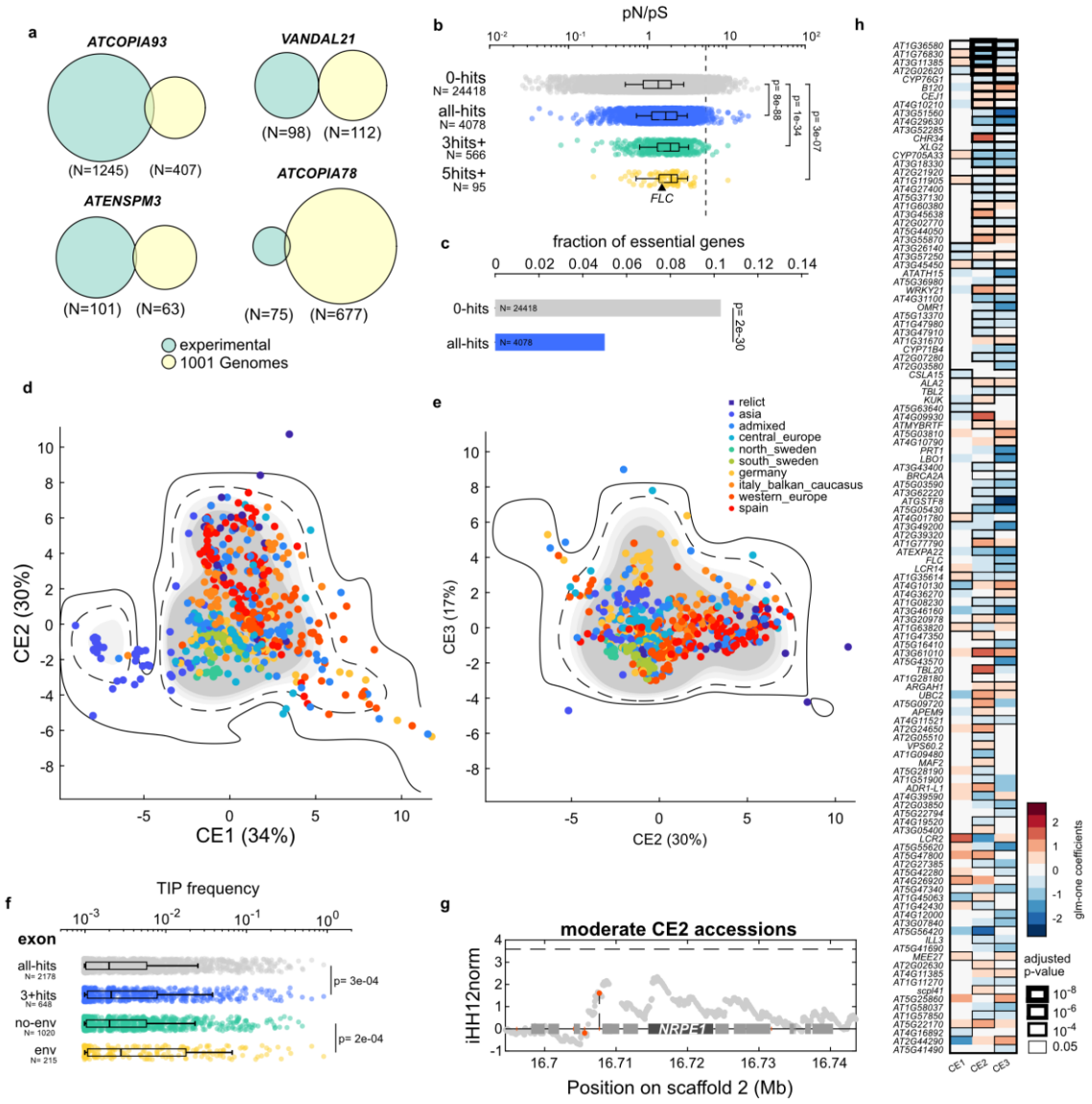

Figure S7.

(a) Genes visited within 250bp by *ATCOPIA93*, *ATCOPIA78*, *VANDAL21*, and *ATENSPM3* in experimental transposition accumulation lines (9) 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) Heatmap of associations in logistic GLM between presence of TE insertion within or near genes and the three climatic envelopes (CE1-3).

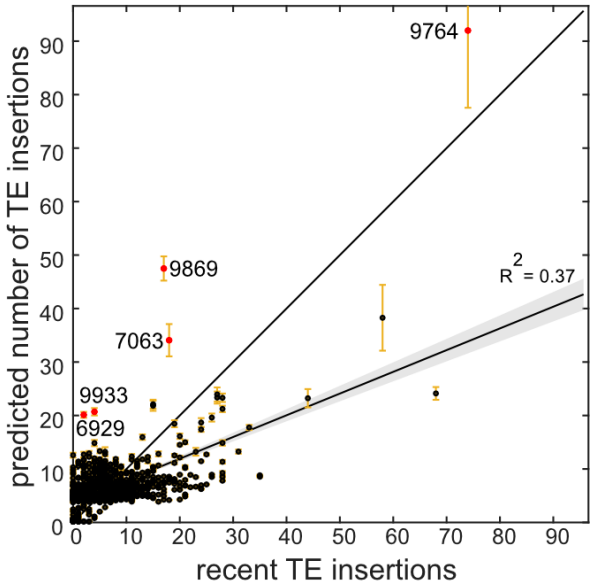

Figure S8.

(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. The five accessions showing large deviations between the predicted and observed number of TE insertions, and therefore excluded from our forecast models, are indicated in red.

Table S1. List and information of 1047 A. thaliana accessions included in the study

Table S2. Results of generalized linear models of recent transposition based on genetic factors

Table S3. Results of generalized linear models of recent transposition based on genetic and environmental factors

Table S4. List of publicly available RNA-seq datasets used for transcriptomic analyses

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. Kapun M, Barrón MG, Staubach F, Obbard DJ, Wiberg RAW, Vieira J, et al. Genomic Analysis of European Drosophila melanogaster Populations Reveals Longitudinal Structure, Continent-Wide Selection, and Previously Unknown DNA Viruses. Mol Biol Evol. 2020 Sep 1;37(9):2661–78.

8. 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.

9. 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.

10. 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.

11. 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.

12. 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.

13. 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.

14. 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.

15. 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.

16. 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

17. 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

18. 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

19. Carpentier M-C, Manfroi E, Wei F-J, Wu H-P, Lasserre E, Llauro C, et al. Retrotranspositional landscape of Asian rice revealed by 3000 genomes. Nat Commun. 2019 Jan 3;10(1):24.

20. Domínguez M, Dugas E, Benchouaia M, Leduque B, Jiménez-Gómez JM, Colot V, et al. The impact of transposable elements on tomato diversity. Nat Commun. 2020 Aug 13;11(1):4058.

21. Petrov DA, Fiston-Lavier A-S, Lipatov M, Lenkov K, Gonzalez J. Population Genomics of Transposable Elements in Drosophila melanogaster [Internet]. Vol. 28, Molecular Biology and Evolution. 2011. p. 1633–44. Available from: http://dx.doi.org/10.1093/molbev/msq337

22. 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.

23. 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

24. 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.

25. 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

26. 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.

27. 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

28. 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.

29. 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

30. 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

31. 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.

32. 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.

33. 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.

34. 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.

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

36. 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

37. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biol. 2006 Mar;4(3):e72.

38. Garud NR, Messer PW, Buzbas EO, Petrov DA. Recent selective sweeps in North American Drosophila melanogaster show signatures of soft sweeps. PLoS Genet. 2015 Feb;11(2):e1005004.

39. 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.

40. 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

41. Gompel N, Prud'homme B. The causes of repeated genetic evolution [Internet]. Vol. 332, Developmental Biology. 2009. p. 36–47. Available from: http://dx.doi.org/10.1016/j.ydbio.2009.04.040

42. Martin A, Orgogozo V. The Loci of repeated evolution: a catalog of genetic hotspots of phenotypic variation. Evolution. 2013 May;67(5):1235–50.

43. Hof AEV, van't Hof AE, Campagne P, Rigden DJ, Yung CJ, Lingley J, et al. The industrial

melanism mutation in British peppered moths is a transposable element [Internet]. Vol. 534, Nature. 2016. p. 102–5. Available from: http://dx.doi.org/10.1038/nature17951

44. The Material Basis of Evolution. Richard Goldschmidt [Internet]. Vol. 8, Philosophy of Science. 1941. p. 394–5. Available from: http://dx.doi.org/10.1086/286719

45. Fisher RA. The genetical theory of natural selection [Internet]. 1930. Available from: http://dx.doi.org/10.5962/bhl.title.27468

46. Healey KR, Zhao Y, Perez WB, Lockhart SR, Sobel JD, Farmakiotis D, et al. Prevalent mutator genotype identified in fungal pathogen Candida glabrata promotes multi-drug resistance. Nat Commun. 2016 Mar 29;7:11128.

47. Eshel I. Clone-selection and optimal rates of mutation [Internet]. Vol. 10, Journal of Applied Probability. 1973. p. 728–38. Available from: http://dx.doi.org/10.1017/s0021900200095930

48. Sniegowski PD, Gerrish PJ, Lenski RE. Evolution of high mutation rates in experimental populations of E. coli. Nature. 1997 Jun 12;387(6634):703–5.

49. Payne JL, Wagner A. The causes of evolvability and their evolution [Internet]. Vol. 20, Nature Reviews Genetics. 2019. p. 24–38. Available from: http://dx.doi.org/10.1038/s41576-018-0069-z

50. Lynch M, Ackerman MS, Gout J-F, Long H, Sung W, Thomas WK, et al. Genetic drift, selection and the evolution of the mutation rate. Nat Rev Genet. 2016 Oct 14;17(11):704–14.

51. Charlesworth D, Barton NH, Charlesworth B. The sources of adaptive variation. Proc Biol Sci [Internet]. 2017 May 31;284(1855). Available from: http://dx.doi.org/10.1098/rspb.2016.2864

52. Exposito-Alonso M, 500 Genomes Field Experiment Team, Burbano HA, Bossdorf O, Nielsen R, Weigel D. Natural selection on the Arabidopsis thaliana genome in present and future climates. Nature. 2019 Sep;573(7772):126–9.

53. Lynch M, Bürger R, Butcher D, Gabriel W. The Mutational Meltdown in Asexual Populations [Internet]. Vol. 84, Journal of Heredity. 1993. p. 339–44. Available from: http://dx.doi.org/10.1093/oxfordjournals.jhered.a111354

54. Caballero A, Bravo I, Wang J. Inbreeding load and purging: implications for the short-term survival and the conservation management of small populations. Heredity . 2017 Feb;118(2):177–85.

55. Stapley J, Santure AW, Dennis SR. Transposable elements as agents of rapid adaptation may explain the genetic paradox of invasive species. Mol Ecol. 2015;24(9):2241–52.

56. Frankham R. Resolving the genetic paradox in invasive species. Heredity . 2005;94(4):385.

57. 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.

58. 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.

59. 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.

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

61. 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.

62. 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.

63. 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

64. Szpiech ZA, Hernandez RD. selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. Mol Biol Evol. 2014 Oct;31(10):2824–7.

65. Delaneau O, Coulonges C, Zagury J-F. Shape-IT: new rapid and accurate algorithm for haplotype inference [Internet]. Vol. 9, BMC Bioinformatics. 2008. p. 540. Available from: http://dx.doi.org/10.1186/1471-2105-9-540

66. 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

67. 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

68. 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.

69. 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.