Transposon dynamics in the emerging oilseed crop *Thlaspi arvense*

Adrián Contreras-Garrido[1,+], Dario Galanti[2,+], Andrea Movilli[1], Claude Becker[3], Oliver Bossdorf[2], Hajk-Georg Drost[4,*], Detlef Weigel[1,*]

[1]Department of Molecular Biology, Max Planck Institute for Biology Tübingen, 72076 Tübingen, Germany
[2]Plant Evolutionary Ecology, University of Tübingen, 72076 Tübingen, Germany
[3]LMU Biocenter, Faculty of Biology, Ludwig Maximilians University Munich, 82152 Martinsried, Germany
[4]Computational Biology Group, Max Planck Institute for Biology Tübingen, 72076 Tübingen, Germany
[+]These authors contributed equally to this study

*corresponding authors: drost@tue.mpg.de (H.-G.D.), weigel@tue.mpg.de (D.W.)

Abstract

Genome evolution is partly driven by the mobility of transposable elements (TEs) which often leads to deleterious effects, but their activity can also facilitate genetic novelty and catalyze local adaptation. We explored how the intraspecific diversity of TE polymorphisms is shaping the broad geographic success and adaptation capacity of the emerging oil crop *Thlaspi arvense*. We achieved this by classifying the TE inventory of this species based on a high-quality genome assembly, age estimation of retrotransposon TE families and a comprehensive assessment of their mobilization potential. Our survey of TE insertion polymorphisms (TIPs) captured 280 accessions from 12 regions across the Northern hemisphere. We quantified over 90,000 TIPs, with their distribution mirroring genetic differentiation as measured by single nucleotide polymorphisms (SNPs). The number and types of mobile TE families vary substantially across populations, but there are also shared patterns common to all accessions. We found that Ty3/Athila elements are the main drivers of TE diversity in *T. arvense* populations, while a single Ty1/Alesia lineage might be particularly important for molding transcriptome divergence. We further observed that the number of retrotransposon TIPs is associated with variation at genes related to epigenetic regulation while DNA transposons are associated with variation at a Heat Shock Protein (HSP19). We propose that the high rate of mobilization activity can be harnessed for targeted gene expression diversification, which may ultimately present a toolbox for the potential use of transposition in breeding and domestication of *T. arvense*.

Introduction

Transposable elements (TEs) are often neglected, mobile genetic elements that make up large fractions of most eukaryotic genomes (1). In plants with large genomes, such as wheat, TEs can account for up to 85% of the entire genome (2, 3). Due to their mobility, TEs can significantly shape genome dynamics and thus both long- and short-term genome evolution across the eukaryotic tree of life. TEs are typically present in multiple copies per genome and they are broadly classified based on their replication mechanisms, as copy-and-paste (class I or retrotransposons) or cut-and-paste (class II or DNA transposons) elements. The two categories can be broken down into superfamilies based on the arrangement and function of their open reading frames (4). Further distinctions can be made based on the phylogenetic relatedness of the TE encoded proteins (5, 6). To minimize the mutagenic effects of TE mobilization, host genomes tightly regulate TE load through an array of epigenetic repressive marks that suppress TE activity (7–9).

While epigenetic silencing of TEs is important for the maintenance of genome integrity and species-specific gene expression, TE mobilization can also generate substantial phenotypic variation through changing the expression of adjacent genes, either due to local epigenetic remodeling or direct effects on transcriptional regulation (10). Because TE activity is often responsive to environmental stress (11–13) and other environmental factors (14)(15)(16)(17), it has been proposed that it could be used for speed-breeding through externally controlled transposition activation (18).

Thlaspi arvense, field pennycress, yields large quantities of oil-rich seeds and is emerging as a new high-energy crop for biofuel production (19–21). As plant-derived biofuels can be a renewable source of energy (22), the past decade has seen efforts to domesticate this species and understand its underlying genetics in the context of seed development and oil production. *Thlaspi arvense* is particularly attractive as a crop because it can be grown as winter cover during the fallow period, protecting the soil from erosion (19). Natural accessions of *T. arvense* are either summer or winter annuals, with winter annuals being particularly useful as potential cover crop (23). Native to Eurasia, *T. arvense* was introduced and naturalized mainly in North America (24).

As a member of the Brassicaceae family, *T. arvense* is closely related to the oilseed crops *Brassica rapa* and *Brassica napus*, as well as the undomesticated model plant *Arabidopsis thaliana* (25). A large proportion of the *T. arvense* genome consists of TEs (26), and TE co-option has been proposed as a mechanism particularly for short-term adaptation and as a source of genetic novelty (27). As in many other species, differences in TE content is likely to be a major factor for epigenetic variation as well, especially through remodeling of DNA methylation (28).

Here, we use whole-genome resequencing data from 280 geographically diverse *T. arvense* accessions to characterize the inventory of mobile TEs (the ‘mobilome’), TE insertion patterns of class I and class II elements and their association with variation in the DNA

methylation landscape. We highlight a small TE family with preference for insertion near genes, which may be particularly useful for identifying new genetic alleles for *T. arvense* domestication.

Results

Phylogenetically distinct transposon lineages shape the genome of *T. arvense*

To be able to understand TE dynamics in *Thlaspi arvense*, we first reanalyzed its latest reference genome, MN106-Ref (26). In total, 423,251 transposable elements were categorized into 1984 unique families and grouped into 14 superfamilies (Table S1), together constituting 64% of the ~526 Mb MN106-Ref genome. Over half of the genome consists of LTR (Long Terminal Repeat)-TEs. Using the TE model of each LTR family previously generated by structural *de novo* prediction of TEs (26), we assigned 858 (~70%) of the 1,205 Ty1 and Ty3 LTR-TEs to known lineages based on the similarity of their reverse transcriptase domains (5) (Fig. 1A).

The most abundant LTR-TE lineage in *T. arvense* is Ty3 Athila (Table S2) with ~180,000 copies, 10-fold more than the next two most common lineages, Ty3 Tekay (~57,000) and Ty3 CRM (~30,000). The most abundant Ty1 elements belonged to the Ale lineage, with 108 families, while the Alesia and Angela lineages were represented only by one family each (Table S2).

Next, we compared the genomic distribution of lineages within the same TE superfamily (Fig. 1B). In the Ty1 superfamily, CRM showed a strong centromeric preference, whereas Athila was more common in the wider pericentromeric region. In the Ty3 superfamily, Ale elements were enriched in centromeric regions, whereas Alesia showed a preference for gene-rich regions.

Thlaspi arvense LTR retrotransposons present signatures of recent activity

To assess the potential and natural variation of TEs transposition across accessions, we used the complete set of protein domains identified for a respective TE model to classify each family as either potentially autonomous or non-autonomous (METHODS). About 60% of all TE families (1,260 out of 2,038) encoded at least one TE-related protein domain, but only about a quarter had all protein domains necessary for transposition, and we classified only these 537 families as autonomous. Autonomous TE families had on average more and longer copies than non-autonomous ones, although both contributed similarly to the total TE load in the genome (Fig S1). Next, we focussed on individual, intact LTR-TE copies, since they are often the source of ongoing mobilization activity (13)(18)(56). Overall, the 193 autonomous LTR-TE families had more members without apparent deletions than the 1,027 non-autonomous LTR-TE families (2,039 versus 339). Intact LTR-TEs from autonomous

families tended to be evolutionarily younger and more abundant than their non-autonomous counterparts (Fig.1C). As for lineages, Athila was the lineage with the most intact members, followed by Tekay and CRM (Fig. 1D), although estimates of insertion times revealed Ale and Alesia Ty1 lineages as actors of the most recent transposition bursts (Fig. 1E).

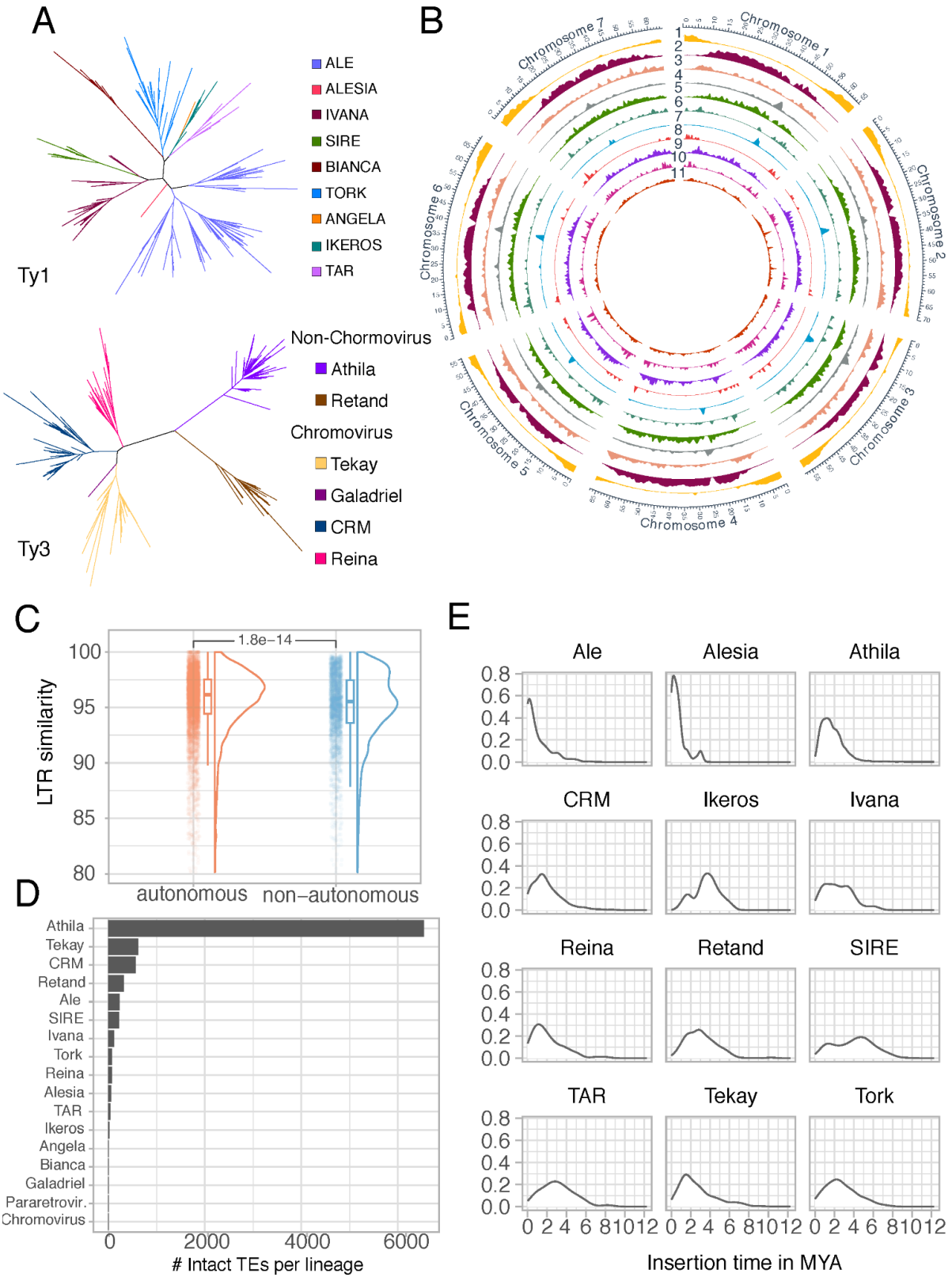


Figure 1. Genome-wide distribution and classification of TE families and superfamilies in the *T. arvense* reference genome MN106-Ref. A, Phylogenetic tree of LTR retrotransposons based on the reverse transcriptase domain. **B,** Genome-wide distribution of TE family and superfamily abundances. The tracks denote, from the outside to the inside, (1) protein-coding loci, (2) Athila, (3) Retand, (4) CRM, (5) Tekay, (6) Reina, (7) Ale, (8) Alesia, (9) Bianca, (10) Ivana, (11) all DNA TEs. **C,** Evolutionary age estimates of intact copies of autonomous versus non-autonomous TE families. P-value is computed based on performing a Wilcoxon Rank Sum test. **D,** Total number of intact TEs in different lineages. **E,** Distribution of insertion time estimates for intact LTR elements across different LTR TE lineages (shown if number of intact TEs was greater than 10).

TE polymorphisms in a collection of wild *T. arvense* populations

Our analysis of the MN106-Ref reference indicated that a substantial part of the genome consists of autonomous, likely still active, TE families. To learn how TE mobility has shaped genome variation at the species level, we surveyed differences in TE content in a large collection of natural accessions. We compiled whole-genome sequences of 280 accessions from different repositories (Table S3), covering twelve geographic regions, and much of the worldwide distribution of *T. arvense* in its native range and in regions where it has become naturalized (Fig. 2A).

We first characterized the population structure of this collection with a subset of high-confidence SNPs and short indels that we used to cluster the accessions by principal component analysis (PCA) (Fig. 2B) (Methods). We also constructed a maximum likelihood tree without considering migration flow for these populations, using the two sister species *Eutrema salsugineum* and *Schrenkiella parvula* as an outgroup (Fig. S2). North American accessions clustered together with European accessions, in support of *T. arvense* having been introduced to North America from Europe. Chinese accessions formed a separate cluster, but the most isolated cluster was composed of Armenian accessions, as it has been reported previously (20, 26).

Next, we screened our data for TE insertion polymorphisms (TIPs), *i.e.*, TEs not present in the reference genome assembly. This will in most cases be due to insertions that occurred on the phylogenetic branch leading to the non-reference accession, although it formally could also be the result of deletion or excision events of a shared TE on the branch leading to the reference accession.

We detected 18,961 unique insertions, which were unequally distributed among populations, with an excess of singletons (5,617 singletons) (Fig. 2C). The allele frequency of TIPs was on average lower than that of SNPs (Fig. 2C), with the caveat that detection of TIPs may incur more false negatives. Saturation analysis (Fig. 2D) indicated that we were far from sampling the total TE diversity in *T. arvense*, especially in Armenian and Chinese accessions. Taken at face value, the disparity in singleton frequencies between TIPs and SNPs would suggest either that TIPs are on average evolutionarily younger than SNPs, or that there is stronger selection pressure against TE insertions (29) (Fig. 2C). What speaks against this view is the higher TIP allele frequencies in the gene-rich fraction of the genome,

near the telomeres (Fig. 2E), while TIPs at the pericentromeric regions are more abundant, but have lower allele frequencies (see Fig. S3 for a statistical assessment).

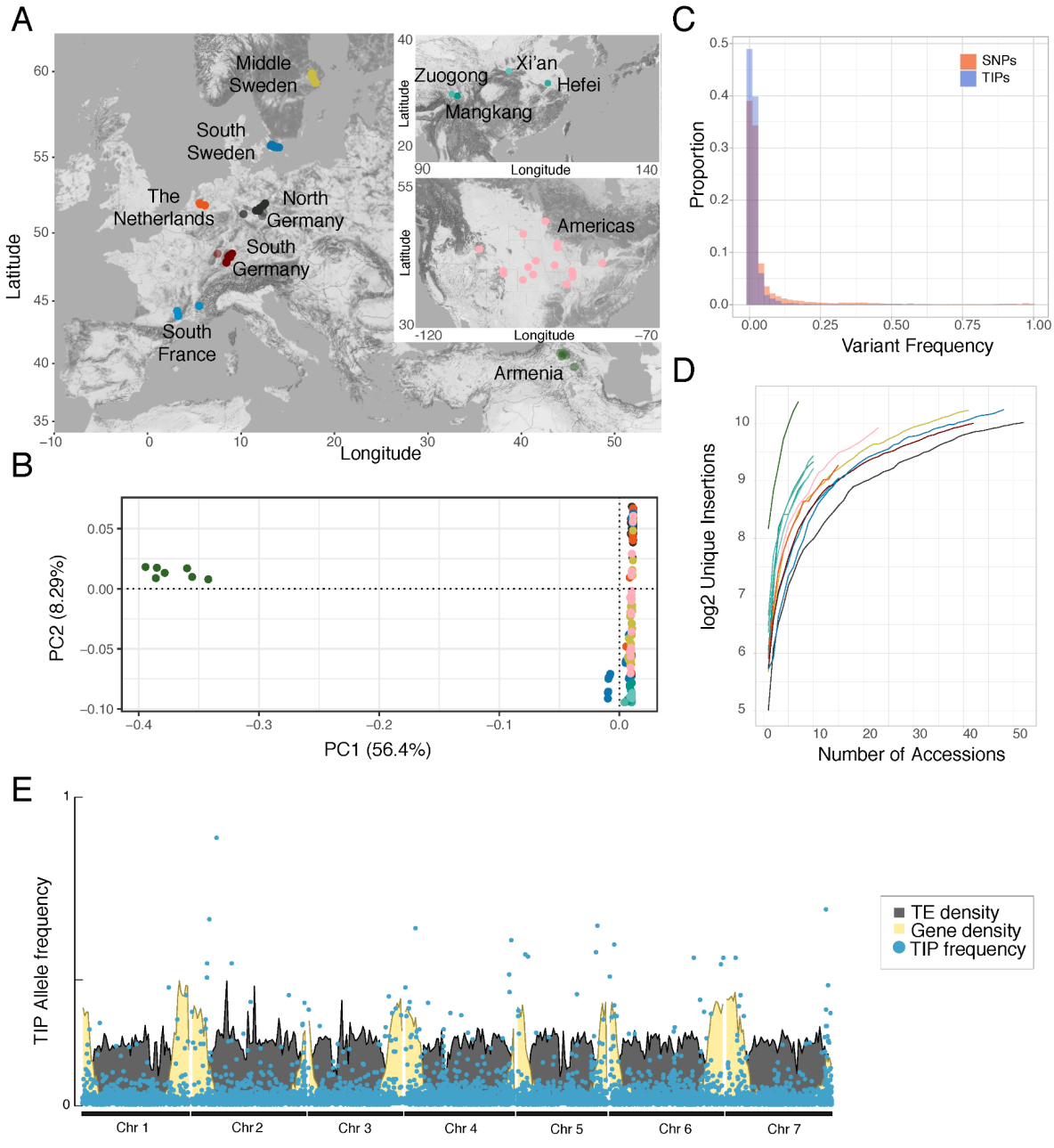


Figure 2. The genome-wide landscape of TE insertion polymorphisms in *T. arvense*.
A, Distribution of accessions across their native Eurasian and naturalized North American range in the Northern hemisphere (omitting a sample from Chile, included in the Americas group). **B**, A SNP-based principal component analysis (PCA) of all accessions, with color code as in (A). Due to the fact that the accessions contributing to the Armenian cluster are separated from the other geographic populations, we recalculated a PCA without the Armenian samples as shown in Fig. S4. **C**, Allele-frequency spectrum of TIPs (blue) and SNPs (red). **D**, Cumulative sums of unique insertions per region as a function of sampled accessions. **E**, TIP frequencies along the genome, compared to gene and TE densities, in 10 kb windows.

We complemented our analysis of TIPs with a corresponding analysis of TE absence polymorphisms (TAPs), which we define as TEs that are found in the reference assembly but

missing from other accessions. This could be due to insertions having occurred on the phylogenetic branch leading to the reference accession or excisions of DNA TEs by a cut-and-paste mechanism. TAPs were detected using a custom TAP annotation pipeline (METHODS).

Overall, a comparison of TIPs and TAPs distributions by PCA showed Armenian accessions to be clear outliers, with all other accessions clustering closely together (Fig. 2B, Fig. S5), indicating that most of the observed TE variation reflects the population structure observed with SNPs. As with SNPs, Armenian accessions harbor the largest number of both TIPs and TAPs. If we look at the impact of these polymorphisms on the genomic landscape (Fig. 3A), we find a major hotspot of TAPs in chromosome 4 for a subset of accessions from Southern Sweden. There also appears to have been major insertion activity in the clade leading to the reference accession, as indicated by the high density of reference insertions missing in all other populations at the ends of chromosomes 4 and 5. For both TIPs and TAPs, the major source of TE polymorphisms comes from activity of Ty3 LTRs (RLGs), especially Ty3 Athila (Fig. 3B). Many other TE families contributed to both TIPs and TAPs as well, with 1,203 families having at least one TIP, and 1,268 having at least one TAP. The more distant a population is geographically from the reference, the greater the contribution of non-autonomous families to the TIP load, with the exception of Northern Germany (Fig. 3C).

Across all populations, most TE activity was due to a small set of 25 TE families, with the Athila lineage standing out in particular (Fig. 3D). For highly active TE families, TIPs were more diverse than TAPs, as the latter were predominantly driven by LTR retrotransposons.

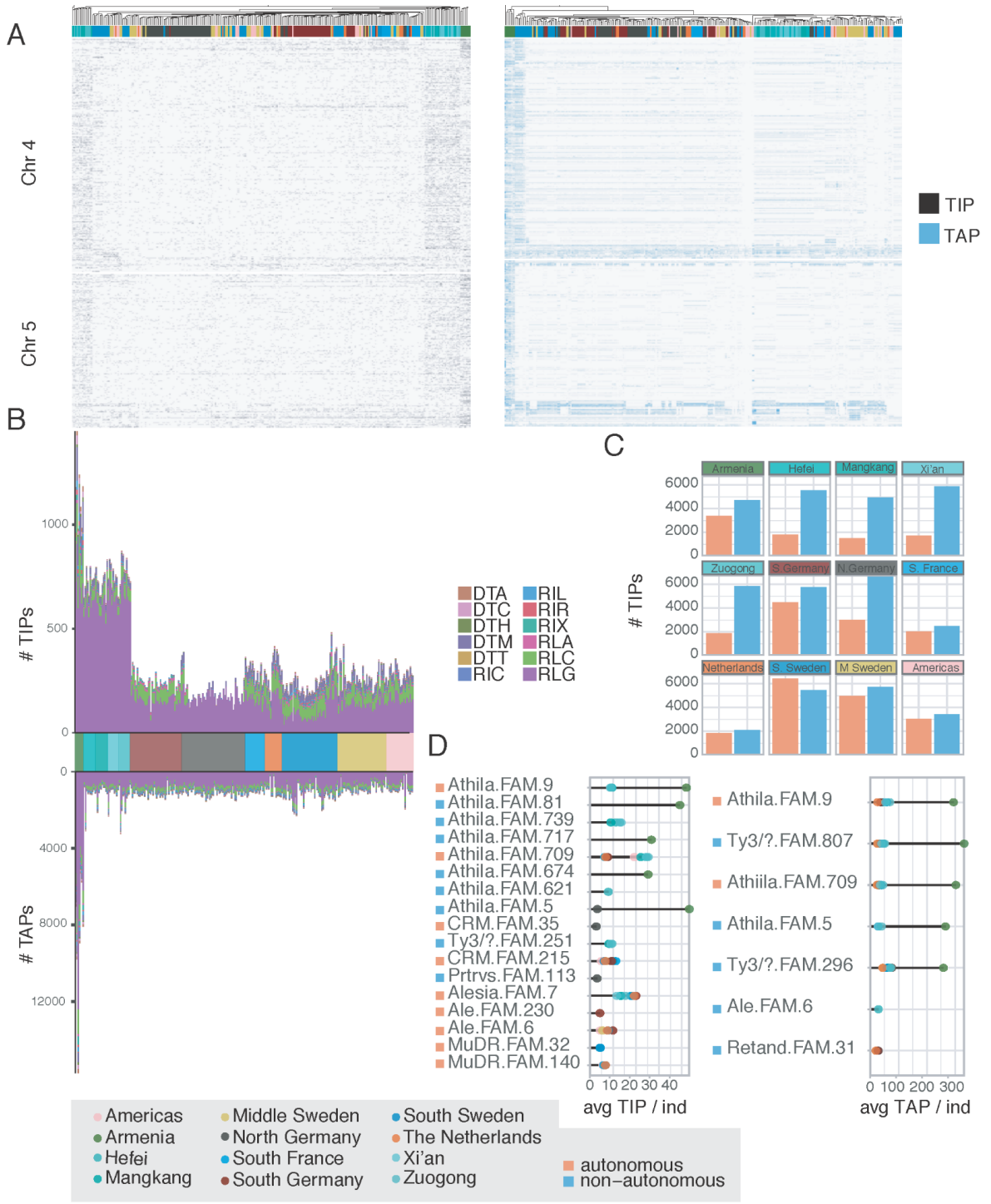
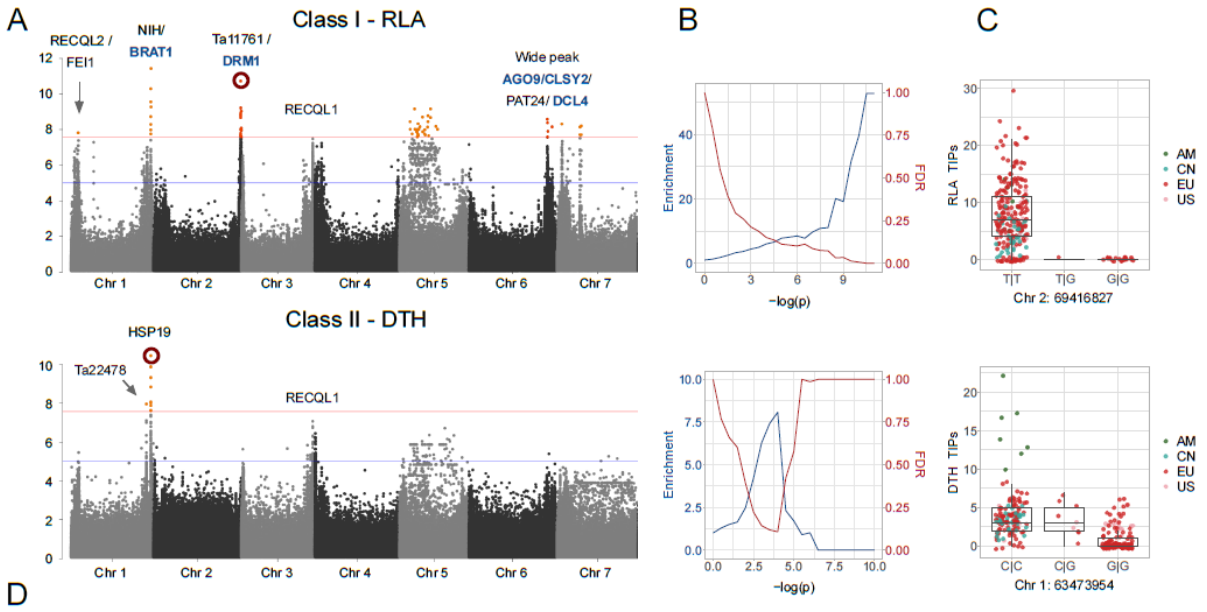


Figure 3. The *T. arvense* mobilome. **A**, Genomic distribution of TIPs and TAPs in chromosomes 4 and 5, where we observe major TIP/TAP hotspots. TIPs and TAPs along the other chromosomes are shown in Fig. S6. **B**, Contribution of different superfamilies to transposon insertion polymorphisms (TIPs) and transposon absence polymorphisms (TAPs). **C**, Frequencies of autonomous and non-autonomous TE-derived TIPs in different geographic regions. **D**, Average count of TIPs per individual for the five TE families with the highest contribution to either TIPs or TAPs in each geographic region. For all figure panels, the gray box illustrates the color scheme for the geographical populations and for autonomous/non-autonomous families.

Host control of TE mobility

In *A. thaliana*, natural genetic variation affects TE mobility and genome-wide patterns of TE distribution, driven by functional changes in key epigenetic regulators (14, 30–32). The rich inventory of TE polymorphisms in *T. arvense* offered an opportunity to investigate the genetic basis of TE mobility in a species with a more complex TE landscape. We tested for genome-wide association (GWA) between genetic variants (SNPs and short indels) and TIP load of different TE classes, TE orders and TE superfamilies (4). We found several GWA hits next to genes that are known to affect TE activity or are good candidates for being involved in TE regulation (Fig. 4A-D). The results differed strongly between class I and class II TEs: while class I TEs were associated with a wide range of genes encoding mostly components of the DNA methylation machinery (Fig. 4A-D), class II TEs were mostly associated with allelic variation at an ortholog of *O. sativa HEAT SHOCK PROTEIN 19* (*HSP19)*. This difference was consistent for most superfamilies that belonged to either class I or class II (Fig. S5). The most prominent hits for class I TIPs were near orthologues of *A. thaliana BROMODOMAIN AND ATPase DOMAIN-CONTAINING PROTEIN 1* (*BRAT1*), which prevents transcriptional silencing and promotes DNA demethylation (7), and components of the RNA-directed DNA methylation machinery such as *DOMAINS REARRANGED METHYLTRANSFERASE 1* (*DRM1*)*, ARGONAUTE PROTEIN 9* (*AGO9*) and *DICER LIKE PROTEIN 4* (*DCL4*) (33) (Fig. 4A-D, Fig. S7 and S8). Another category of genes that emerged in our GWA are genes encoding DNA and RNA helicases such as *RECQL1* and *2* (Fig. 4, and Fig. S8).

To further confirm the association between the DNA methylation pathway and class I TE polymorphisms, we used published bisulfite sequencing data to quantify methylation levels of the neighboring regions of TIPs (28). In all three epigenetic contexts (CG, CHG, CHH), we found a significant increase of methylation up to 1 kb around class I, but not around class II TE insertions (Fig. 4E). Taken together, we interpret these results such that class I TE mobility is primarily controlled by the DNA methylation machinery, leading to RdDM spreading around novel insertions, thus creating substantial epigenetic variation beyond TE loci.



D

Candidate gene	Gene ID	Ortholog	Putative function	Top SNP location
RECQL2	Tarvense_11198	AT1G31360	RECQ HELICASE L2 (ATP-dependent DNA helicase)	+ 5kb (downstream)
FEI1	Tarvense_11200	AT1G31420	Cell wall elongation	0 kb (intron)
Ta22478	Tarvense_22478	AT1G14990	Uncharacterized – transmembrane protein	0 kb (intron)
HSP19	Tarvense_21603	OsHSP19	class II heat shock protein	0 kb (intron)
NIH	Tarvense_21621	AT1G06670	NUCLEAR DEIH-BOXHELICASE (RNA/DNA helicase)	+ 0.1 kb (upstream)
BRAT1	Tarvense_21555	AT1G0510	acetylated histone-binding protein (prevention of transcriptional silencing and DNA demethylation)	0 kb (exon)
Ta11761	Tarvense_11761	AT1G63855	Putative methyltransferase family protein	0 kb (exon)
DRM1	Tarvense_11767	AT5G15380	DOMAINS REARRANGED METHYLTRANSFERASE (de novo and non-symmetric methylation maintenance)	+ 10 kb (downstream)
RECQL1	Tarvense_09699	AT3G05740	RECQ HELICASE L1 (DNA recombination and repair)	- 2 kb (downstream)
AGO9	Tarvense_05665	AT5G21150	ARGONAUTE 9 (siRNA silencing)	13 kb (downstream)
CLSY2	Tarvense_05713	AT5G20420	CLASSY 2 (Gene silencing by RdDM)	+ 17 kb (downstream)
PAT24	Tarvense_05719	AT5G20350	TIP GROWTH DEFECTIVE 1	- 1 kb (downstream)
DCL4	Tarvense_05722	AT5G20320	DICER-LIKE 4 (siRNA processing and RdDM)	- 12 kb (upstream)

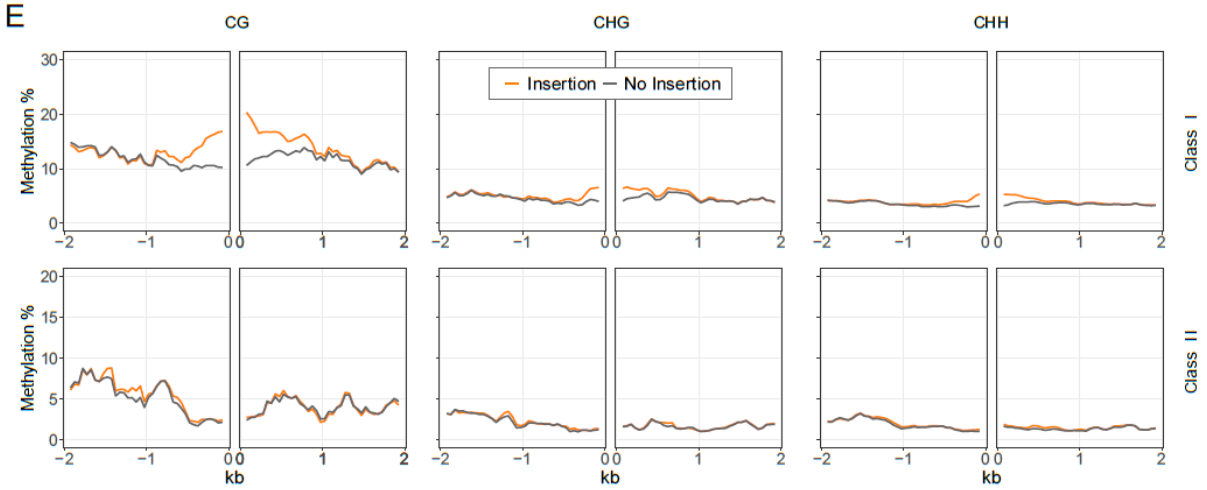


Figure 4. GWA analysis for TIP load of a class I and a class II TE superfamily. Results including all superfamilies are shown in Fig S5. **A**, Manhattan plots with candidate genes indicated next to neighboring variants. The red line corresponds to a genome-wide significance with full Bonferroni correction, the blue line to a more generous threshold of –log(p)=5. **B**, Enrichment and expected FDR of DNA methylation machinery genes, for stepwise significance thresholds (28, 34). **C**, Shown are the allelic effects of the red-circled variants from the corresponding Manhattan plots on the left. **D**, Shown are the candidate genes marked in **A**, their

putative functions and distances to the top variant of the neighboring peaks. Blue font denotes DNA methylation machinery genes included in the enrichment analyses. **E**, DNA methylation around class I and class II TIPs in carrier vs. non-carrier individuals.

An autonomous Alesia LTR family with insertion preference for specific genomic regions

Our characterization of the *T. arvense* mobilome revealed a strikingly uneven distribution of one autonomous LTR Ty1 family belonging to the Alesia lineage, Alesia.FAM.7. This family encompasses 144 elements in the reference genome, 51 of which are complete copies. Despite being a relatively small TE family, 44 copies are close to genes (< 1kb), of those, 8 copies are within genes (Table S3). Across all 4,215 Alesia.FAM.7 TIPs, that is insertions not present in the reference genome, we found a strong enrichment nearby and within genes, which was the case of ~75% of all insertions (Fig. 5A and 5B). The genes potentially affected by these insertions were involved in a wide range of functions, including metabolism and responses to biotic and abiotic factors (Fig. 5C). Reference insertions were rarely missing in other accessions, except an intronic reference insertion that was detected as absent in some Swedish accessions. The prevalence of Alesia.FAM.7 TIPs near genes suggests that the skewed distribution in the reference is not so much due to removal of insertions in other regions, but that it reflects an unusual insertion site preference of this family across all examined accessions.

Alesia.FAM.7 is highly similar to the Terestra TE family, first described in *A. lyrata* (35). The Terestra family, which has been reported in six Brassicaceae, is heat responsive due to a transcription factor binding motif also found in *A. thaliana* ONSEN, where it can be bound by heat shock factor A (HSFA2) via a cluster of four nGAAn motifs called heat responsive elements (HRE) (12). In Alesia.FAM.7, we found a similar four-nGAAn motif cluster in most copies in the 5' LTR portion of the elements (Fig. 5D). A search against the NCBI NT database (36) revealed the presence of this TE family, with an Alesia-diagnostic reverse transcriptase sequence signature, in several additional Brassicaceae (Fig. 5E), notably *B. rapa*, *B. napus*, *B. olaracea*, *Raphanus sativus*, and other *Arabidopsis* species, but not in *A. thaliana*. It is conceivable that this heat-responsive, euchromatophilic Alesia family rewires gene regulatory networks between and within Brassicaceae species. We conducted a similar search of a subset of TE families against the NCBI NT database (Fig. S9) and Alesia.FAM.7 was indeed the only deeply conserved TE family with evidence for recent activity.

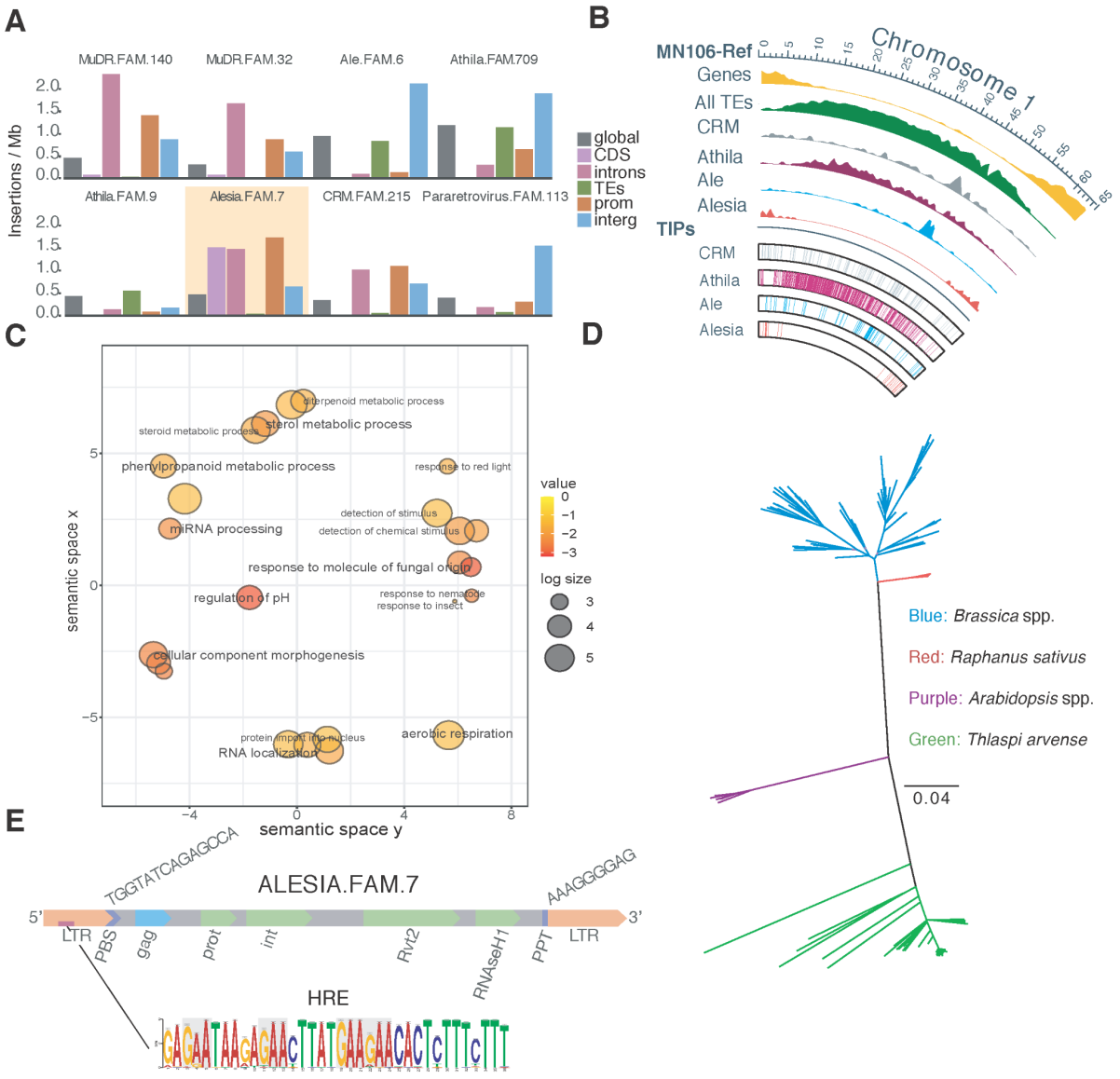


Figure 5. Summary statistics and characterization of the Alesia.FAM.7 family in *T. arvense* and other Brassicaceae. A, Distribution of several TE families across different genomic contexts in *T. arvense* accessions. While several other families, such as MuDR.FAM.140 or CRM.FAM.215, are also often found in introns, Alesia.FAM.7 is the only family that is commonly inserted in coding sequences. **B**, Distribution of several LTR lineages along chromosome 1 in MN106-Ref. **C**, GO enrichment of genes associated with Alesia.FAM.7 TIPs. **D**, Phylogenetic tree of Alesia.FAM.7 related copies across different Brassicaceae. **E**, Structure of the Alesia.FAM.7 model: 5' Long terminal repeat (LTR); primer binding site (PBS), a tRNA binding site, in this case complementary to *A. thaliana* methionine tRNA; *Gag* domain; *Pol* domains: Protease (Prot), Integrase (Int) and the two subdomains of the reverse transcriptase, the DNA polymerase subdomain (Rvt2) and the RNase H subdomain (RNAseH1); polypurine tract (PPT).The location of a putative heat responsive element (HRE) with the four-nGAAn motif in the LTR is indicated in by a purple segment.

Discussion

Although *A. thaliana* and *T. arvense* are close relatives, with evolutionary divergence estimates of 15-24 million years ago (27) and similar life histories in terms of demographic

dynamics, geographic expansion, and niche adaptation (25, 37), their genomes are very different, one key difference being the significantly higher TE load of the *T. arvense* genome. Exploring the diversity and dynamics of mobile elements in such TE-rich genomes enables a better understanding of the evolution of genome architecture. Here, we report how TEs drive genome variation in *T. arvense* by analyzing the diversity and phylogenetic relationships of TEs, as well as their autonomous status, ongoing activity, and contrasts between biogeographic populations.

Many recent studies have confirmed that several TE families do not insert randomly in the genome, and that their apparent enrichment in specific portions of the genome, such as centromeres, is not simply due to purifying selection (38). Many TEs have clear insertion site preference (39), both driven by primary DNA sequence and by epigenetic marks, e.g. Ty1 insertions in *A. thaliana* are biased towards regions enriched in H2A.Z (40). Our results confirm this view whereby the phylogenetic nature of an LTR element plays a role in the observable genome-wide insertion pattern in *T. arvense*. Within the Ty1 elements, Ales are preferentially centrophilic whereas Alesias are enriched in the genic regions of the genome. For the Ty3 elements, The Retand clade does not show any particular preference across the chromosome, while CRM are centrophilic and Athila insertions are often found in pericentromeric regions. Thus, a phylogenetic classification of TEs, alongside the classification into autonomous and non-autonomous elements, is key to understanding TE dynamics, especially in LTR retrotransposon-rich genomes.

We learned that one third of *T. arvense* genome consists of Ty3/Athila LTR-TEs, which is considerably more than in other Brassicaceae, such as *A. thaliana* and *Capsella rubella*, where Ty1/Ale elements are the most abundant TE lineage (41). This suggests that a single or multiple ancient Athila bursts may underlie genome size expansion in *T. arvense*. This is in line with the expansion of the Ty3 LTR-TE superfamily, to which Athila belongs, in *Eutrema salsugineum* (42), from which *T. arvense* diverged 10-15 million years ago (43). Similar Ty3 associated expansions have been reported, for example, for *Capsicum annuum* (hot pepper) (44).

Having established substantial variation in TE content among natural accessions, we asked whether there is also genetic variation for control of TE mobility, as is the case for *A. thaliana* (14, 30, 31). Perhaps not too surprisingly, the sets of genes associated with TE mobilization appear to depend on the nature of the TE transposition mechanism. While variation in retrotransposon insertions was strongly associated with several genes involved in the DNA methylation machinery, DNA transposon insertions were instead associated with a single *Heat Shock Protein 19* (*HSP19*) gene, and this was consistent across different class I (retrotransposon) and class II (DNA transposon) superfamilies. Although studies in *A. thaliana* have highlighted differences in the genetic control of methylation and mobility of the two classes of transposons, they are not as striking compared to the evidence we found here (14, 32, 45). In *A. thaliana*, GWA for CHH methylation of TE families did not produce very different signals for class I and II families (45). The same was true for TIP-counts of

different families and superfamilies as phenotypes (14, 32). Since *HSP19* is an ortholog of an *O. sativa* gene that is absent from the *A. thaliana* reference genome, it is possible that this gene is providing new functionality in *T. arvense*. What this functionality might be is difficult to answer with our data, but different types of HSPs are involved in DNA methylation-dependent silencing of genes and TEs in *A. thaliana* (46), and in controlling transposition in several other organisms (47–49). Our interpretation that natural genetic variation in *T. arvense* points to differences in the genetic control of silencing of class I and class II TEs is further supported by methylome evidence, where we found that DNA methylation spreads from class I TE insertions, but not from class II TE insertions.

The contrast between Alesia and Athila lineages suggests that TEs may be more than detrimental genome parasites. There are many examples from animals and plants of both TE proteins and TEs themselves having been domesticated and thereby enriching genome function (38, 50–52). While parasitic TEs may constitute the majority of TEs within a given species, there can be different life cycle strategies adopted by TEs (53). With respect to notable TE families in *T. arvense*, Alesia's gain of HREs might provide a unique selection advantage, allowing it to survive more easily in the genome, as long as copy numbers are low, in a relationship with the host that resembles other forms of symbiotic lifestyle. Further research of this enigmatic Alesia lineage, which is found in many angiosperms (41), could enhance our understanding of the different strategies used by TEs to persist over long evolutionary time scales.

Turning to more practical matters, it might be possible to exploit the preference of Alesia.FAM.7, which is conserved in several Brassicaceae species, for genic insertions as a source of fast genic novelty for crop improvement, either via gene disruption or modulation of gene expression via intronic insertion. It would therefore be useful to determine how easily Alesia.FAM.7 can be mobilized by heat in *T. arvense*, and conversely, whether heat responsiveness might also be a source of unwanted genetic variation in breeding programs.

Methods

Dataset summary

For the investigation of *T. arvense* natural genetic variation (TIPs, TAPs, and short variants), we leveraged Illumina short read data from three studies (26, 28, 43). The largest survey investigated both genetic and DNA methylation variation in 207 European accessions (13 from the Netherlands, 16 from the South of France, 42 from the South of Germany, 52 from the North of Germany, 48 from the South of Sweden and 40 from Middle Sweden). In addition, we used data from 39 Chinese accessions (10 each from Xi'an, Zuogong, and Hefei and 9 from MangKang) (43), 21 from the US, and one each from Chile and Canada (26). For most of the European accessions, Illumina whole-genome bisulfite-sequencing (BS-seq) data were available as well (28) (Table S3). We used as reference, the assembly generated in (26)), together with the gene and TE annotation also generated in that study.

We reinforced this dataset by sequencing 12 different accessions, 7 Armenian and 5 European, using Illumina paired-end 2x150 bp WGS (Table S3). Briefly, we grew plants in soil, collected fully developed rosette leaves, snap-froze them in liquid nitrogen and disrupted the tissue to frozen powder. We extracted genomic DNA and prepared Illumina libraries as described before (28). To validate our TIP analysis we also sequenced our samples using long read HiFi PacBio technology for a single Armenian accession (Ames32867/TA_AM_01_01_F3_CC0_M1_1). For the ancestry analysis, we used as outgroup species two assemblies for *Eutrema salsugineum* and *Schrenkiella parvula* (NCBI ID : PRJNA73205 ; Phytozome genome ID: 574 respectively).

TE analysis of the reference genome

To resolve phylogenetic relationships of the LTR-TEs in *T. arvense* using information from a collection of green plants (Viridiplantae) at REXdb (5), and to classify *T. arvense* LTR-TEs into lineages, we used the DANTE pipeline (https://github.com/kavonrtep/dante). We used a published *T. arvense* TE library (26) as query with default parameters except for "--interruptions", which we set to 10 to reflect the fact that we used as input the consensus TE models and therefore a likely increase in frameshifts and stop codons in the sequences.

After classification, we used the inferred amino acid sequences of the retrotranscriptase domains extracted from Ty3 and Ty1 elements identified by DANTE to produce two multiple sequence alignments using MAFFT with standard parameters (54). Using RAxML (55), we built a set of phylogenetic trees under a JTT + gamma model, with 100 rapid bootstraps to assess the branch reliability of the NJ tree.

Analysis of intact LTR-TEs analysis and estimates of LTR-TE age used LTRpred (56) against the reference genome with default parameters. We correlated the genomic positions of the *de novo* predicted LTR-TEs with those in the annotation using bedtools (57) intersecting with " -f 0.8 -r" parameters.

To analyze the extent of conservation of TE families larger than 2kb across Brassicaceae, we ran BLASTN (58) against the NCBI NT database (36), June 2022 release. Next, we filtered the result by requiring 80% identity and 80% alignment coverage of the query sequence. For Alesia.FAM.7 TE family filtered matches, we performed a multiple sequence alignment of the remaining matches using MAFFT (54) with default settings and constructed a tree with RaxML (55) with the parameters "-model JTT+G --bs-trees 100". To *de novo* discover nGAAn motifs in all the sequences of Alesia.FAM.7, we ran MEME (59) with the following parameters "-mod zoops -nmotifs 3 -minw 6 -maxw 50 -objfun classic -revcomp -markov_order 0".The *de novo* deemed HRE motif selected had 4 nGAAn clusters in the reverse strand: AAAGAAAGAGTGTTCTTCATAAGTTCTCTTATTCTC (E-value = 2.8e-33).

Short variant calling

We called variants with GATK4 (60), following best practices for germline short variant discovery

(https://gatk.broadinstitute.org/hc/en-us/sections/360007226651-Best-Practices-Workflows), as described in (28). Briefly, we trimmed reads, removed adaptors, and filtered low quality bases and short reads (≤25 bp) using cutadapt v2.6 (61). We aligned trimmed reads to the reference genome (26) with BWA-MEM v0.7.17 (62), marked duplicates with *MarkDuplicatesSpark* and ran *Haplotypecaller*, generating GVCF files for each accession. To combine GVCF files, we ran *GenomicsDBImport* and *GenotypeGVCFs* successively for each scaffold, and then merged files with *GatherVcfs*, to obtain a multisample VCF file. Based on quality parameters distributions, we removed low-quality variants using *VariantFiltration* with specific parameters for SNPs (QD < 2.0 || SOR > 4.0 || FS > 60.0 || MQ < 20.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0) and other variants (QD < 2.0 || QUAL < 30.0 || FS > 200.0 || ReadPosRankSum < -20.0). We filtered variants with *vcftools* v0.1.16 (63), retaining only biallelic variants with at most 10% missing genotype calls, and Minor Allele Frequency (MAF) > 0.01. Finally, we imputed missing genotype calls with *BEAGLE* 5.1 (64), obtaining a complete multisample VCF file. All the code for short variants calling, filtering and imputation can be found on GitHub (https://github.com/Dario-Galanti/BinAC_varcalling).

For calculating site frequency spectra, we used all biallelic SNPs with Minor Allele Count (MAC) of at least two. To assess the population structure of our dataset, we pruned variants in strong LD using *PLINK* (65) with the following parameters "--indep-pairwise 50 5 0.8" and then ran PCA analyses to assess the variance of natural variation. Due to the high divergence of the Armenian accessions from the rest, we ran separate PCAs with and without these accessions, to highlight the structure of the remaining populations (Fig. S4).

Lastly, we analyzed the genetic relatedness among accessions from different geographic regions constructing a maximum likelihood tree using *TREEMIX* (66) with 2,500 bootstrap replicates without considering migration flow and using as an outgroup two sister species, *Eutrema salsugineum* and *Schrenkiella parvula*. We merged all 2,500 independent *treemix* runs and generated a consensus tree with the *Phylip* "consense" command (https://evolution.genetics.washington.edu/phylip/).

TE polymorphism calling

To identify TE insertion polymorphisms (TIPs), we used *SPLITREADER* (32) as described in (67). We applied two custom steps (https://github.com/acontrerasg/Tarvense_transposon_dynamics). In short, we removed Helitron insertions, as they have been shown to have a high false positive ratio (32). Next, we mapped short reads from the reference accession MN106 to the reference genome, to identify regions of aberrant coverage. We marked regions corresponding to ~16% of the genome as aberrant and any TIP landing in these regions were excluded from the final dataset. Lastly, we removed TIPs with > 100 reads 500 bp upstream and/or downstream of the TIP, because this suggested aberrant structural variants in the sample, not reflected in the reference. To calculate the variant frequency spectra of TIPs, we classified TIPs as shared between two or more accessions if coordinates were identical.

To detect TIPs using *Splitreader*, a collection of TEs is required. We used a representative subset of the total number of TEs present in the *T. arvense* reference genome, generated with a custom script. As a selection criterion, we defined representatives according to the consensus TE sequence of each family and the five longest individual members of each family. If a family consisted of < 5 members, all members were used.

We visually inspected 2,790 TIPs spanning all analyzed TE superfamilies and all accessions using *IGV*. Over 70% of TIPs were deemed correct, which is in line with reports from other studies in *A. thaliana (32)* and tomato (68).

To further confirm our TIPs, we generated HiFi PacBio long reads for an Armenian accession (Ames32867/TA_AM_01_01_F3_CC0_M1_1). We stratified seeds at 4°C for one month and germinated them on soil. One month after germination, we subjected plants to 24h dark prior to harvesting. We extracted high molecular weight (HMW) DNA as described (69) using 600 µg of ground rosette material. Using a gTUBE (Covaris) we sheared 10 µg of HMW DNA to an average fragment size of of 24 kb and prepared two independent non-barcoded HiFi SMRTbell libraries using SMRTbell Express Template Prep Kit 2.0 (PacBio). We pooled the two libraries and performed size-selection with a BluePippin (SageScience) instrument with 10 kb cutoff in a 0.75% DF Marker S1 High Pass 6-10kb v3 gel cassette (Biozym). We sequenced the library on a single SMRT Cell (30 hours movie time) with the Sequel II system (PacBio) using the Binding Kit 2.0. Using PacBio CCS with "--all" mode (https://ccs.how/), we generated HiFi reads (sum = 31 Gb, n = 1,633,975, average = 19 kb). We called structural variants (SVs) against the reference using *Sniffles2* (70). 71% of the TIPs called in this accession using short reads had a PacBio HiFi-read supported SV within 200 bp, in line with our visual assessment of TIP quality.

Using paired-end short read Illumina data, we also screened for TE absence polymorphisms (TAPs). First, we calculated the GC-corrected median read depth (RD) in genome-wide 10 bp bins for short-read data sets from all accessions and from two reference controls. For every annotated TE ≥ 300 bp, we extracted its corresponding RD-bins for both the controls and a single sample and used a non-parametric test (Wilcoxon Rank Sum) to compare the bins of the focal sample with the bins of both controls. If i) the annotated TE showed a significant difference in coverage between the focal accession and the mean of the controls, and ii) the median coverage of that TE showed at least a 10-fold reduction in the focal accession compared to the all accession median coverage, then such a TE was considered absent in the focal accession. To exclude the possibility that our TAP calls were the result of major rearrangements in the vicinity of the TAP call, we calculated the coverage of the flanking regions of the TAPs and removed those with < 5X or > 50X mean coverage.

Genome Wide Association between TE polymorphisms and genomic regions

To detect genetic variants associated with variation in TE content, we ran GWA using the number of TIPs of different classes, orders and superfamilies as phenotypes. We used

mixed models implemented in *GEMMA* (71), correcting for population structure with an Isolation-By-State (IBS) matrix. Starting from the complete VCF file obtained from variant calling, we used *PLINK* (65) to prune SNPs in strong LD (--indep-pairwise 50 5 0.8) and computed the IBS matrix. We tested for associations between TIP counts and all variants with MAF > 0.04 (SNPs and short INDELs). We log-transformed TIP counts to approximate a log-normal distribution of the phenotype. To quantify the potential effects of components of the epigenetic machinery on TE content, we calculated the enrichment of associations in the proximity of a custom list of genes with connections to epigenetic processes (28) for increasing cutoffs (34). Briefly, we assigned an "a-priori candidate" status to all variants within 20 kb of the genes from the list and calculated the expected frequency as the fraction between "a-priori candidate" and total variants. We calculated enrichment for -log(p) threshold increments, comparing the fraction of significant a-priori candidates (observed frequency) to the expected frequency. We further calculated the expected upper bound for the false discovery rate (FDR) as described in (34). The code to run GWA and the described enrichment analysis is available on GitHub (https://github.com/Dario-Galanti/multipheno_GWAS/tree/main/gemmaGWAS).

DNA methylation around insertions

To investigate cytosine methylation in the proximity of TIPs, we leveraged Whole Genome Bisulfite Sequencing (WGBS) data from the European accessions, using multisample unionbed files (28). To reduce technical noise, we first excluded singleton TIPs and within 2 kb of another TIP or 1 kb to annotated TEs. We calculated average methylation of accessions with and without a focal TIP in 2 kb flanking regions. We then combined methylation values of all TIPs in 50 bp bins of the 2 kb flanking regions, averaging all positions within each bin. Finally, we calculated the moving average (arithmetic mean) of 3 bins to smoothen the curves. The workflow was based on custom bash and python scripts available at https://github.com/acontrerasg/Tarvense_transposon_dynamics.

Intersection with genomic features and Gene Ontology enrichment analysis

To investigate the targeting behavior of different TE families or superfamilies, we counted TIPs in different genomic features with *bedtools* (57) and divided them by the total genome space covered by each feature to obtain relative insertion density. We turned to gene ontology (GO) enrichment analysis to characterize genes potentially affected by insertions, using all genes located within 2 kb of an insertion. Briefly, we extracted GO terms from the *T. arvense* annotation and integrated them with the terms from *A. thaliana* orthologs identified by *OrthoFnder2* (72). We assessed enrichment with *clusterProfiler* (73) and piped all terms with p value < 0.05 to *REVIGO* (74), using default parameters.

Code availability

Source code for analysis and figures can be found at (https://github.com/acontrerasg/Tarvense_transposon_dynamics).

Data Availability

For this study, 12 accessions were sequenced using illumina WGS technology, of those one was also resequenced using PacBio HiFi technologie. Read sequencing data can be found at the European Nucleotide Archive (ENA) under accession number PRJEB62093. In addition, detailed description of the data can be found in Supplementary table S3.

Author Contributions

A.C.-G., D.G., O.B., H.-G.D. and D.W. conceived the study; A.C.-G. generated data; C.B. provided data; A.C.-G., D.G. and A.M. analyzed data; All authors interpreted the results; A.C.-G. and D.G. wrote the first draft of the manuscript; A.C.-G., D.G., A.M., C.B., O.B., H.-G.D. and D.W. edited the manuscript.

Competing Interest

D.W. holds equity in Computomics, which advises breeders. D.W. advises KWS SE, a plant breeder and seed producer. All other authors declare no competing or financial interests.

Acknowledgements

We thank Haim Ashkenazy, Wei Yuan and Gautam Shirsekar for technical advice, Christa Lanz for support during PacBio HiFi library preparation and Alejandra Duque-Jaramillo, Tess Renahan and Rebecca Schwab for comments on the manuscript. For computing, we acknowledge Prof. Peter Stadler at the University of Leipzig and David Langenberger from ecSeq, for hosting the EpiDiverse servers. The study was supported by the European Union's Horizon 2020 research and innovation programme via the Marie Skłodowska Curie ETN EpiDiverse (Grant Agreement No. 764965; C.B., O.B., D.W.), by the European Research Council (ERC; Grant Agreement No. 716823 "FEAR-SAP"; C.B.), the German Research Foundation (INST 37/935-1 FUGG), the Novo Nordisk Foundation (Novozyme Prize) and the Max Planck Society (D.W.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. J. N. Wells, C. Feschotte, A Field Guide to Eukaryotic Transposable Elements. *Annu. Rev. Genet.* (2020) https:/doi.org/10.1146/annurev-genet-040620-022145.

2. M. I. Tenaillon, J. D. Hollister, B. S. Gaut, A triptych of the evolution of plant transposable elements. *Trends Plant Sci.* **15**, 471–478 (2010).

3. T. Wicker, *et al.*, Impact of transposable elements on genome structure and evolution in bread wheat. *Genome Biol.* **19**, 103 (2018).

4. T. Wicker, *et al.*, A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973–982 (2007).

5. P. Neumann, P. Novák, N. Hoštáková, J. Macas, Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. *Mob. DNA* **10**, 1 (2019).

6. I. R. Arkhipova, Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories. *Mob. DNA* **8**, 19 (2017).

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

8. D. Bouyer, *et al.*, DNA methylation dynamics during early plant life. *Genome Biol.* **18**, 179 (2017).

9. M. J. Sigman, R. K. Slotkin, The First Rule of Plant Transposable Element Silencing: Location, Location, Location. *Plant Cell* **28**, 304–313 (2016).

10. T. Srikant, H.-G. Drost, How stress facilitates phenotypic innovation through epigenetic diversity. *Front. Plant Sci.* **11**, 606800 (2020).

11. A. Pecinka, *et al.*, Epigenetic regulation of repetitive elements is attenuated by prolonged heat stress in Arabidopsis. *Plant Cell* **22**, 3118–3129 (2010).

12. V. V. Cavrak, *et al.*, How a retrotransposon exploits the plant's heat stress response for its activation. *PLoS Genet.* **10**, e1004115 (2014).

13. H. Ito, *et al.*, An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress. *Nature* **472**, 115–119 (2011).

14. P. Baduel, *et al.*, Genetic and environmental modulation of transposition shapes the evolutionary potential of Arabidopsis thaliana. *Genome Biol.* **22**, 138 (2021).

15. S. Ou, *et al.*, Differences in activity and stability drive transposable element variation in tropical and temperate maize. *bioRxiv*, 2022.10.09.511471 (2022).

16. M. Benoit, *et al.*, Environmental and epigenetic regulation of Rider retrotransposons in tomato. *bioRxiv*, 517508 (2019).

17. S. Esposito, *et al.*, LTR-TEs abundance, timing and mobility in Solanum commersonii and S. tuberosum genomes following cold-stress conditions. *Planta* **250**, 1781–1787 (2019).

18. J. Paszkowski, Controlled activation of retrotransposition for plant breeding. *Curr. Opin. Biotechnol.* **32**, 200–206 (2015).

19. M. McGinn, *et al.*, Molecular tools enabling pennycress (Thlaspi arvense) as a model plant and oilseed cash cover crop. *Plant Biotechnol. J.* (2018) https:/doi.org/10.1111/pbi.13014.

20. T. García Navarrete, C. Arias, E. Mukundi, A. P. Alonso, E. Grotewold, Natural variation and improved genome annotation of the emerging biofuel crop field pennycress (Thlaspi arvense). *G3* (2022) https:/doi.org/10.1093/g3journal/jkac084.

21. K. M. Dorn, J. D. Fankhauser, D. L. Wyse, M. D. Marks, A draft genome of field

pennycress (Thlaspi arvense) provides tools for the domestication of a new winter biofuel crop. *DNA Res.* **22**, 121–131 (2015).

22. J. Hill, E. Nelson, D. Tilman, S. Polasky, D. Tiffany, Environmental, economic, and energetic costs and benefits of biodiesel and ethanol biofuels. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 11206–11210 (2006).

23. J. A. Cubins, *et al.*, Management of pennycress as a winter annual cash cover crop. A review. *Agron. Sustain. Dev.* **39**, 46 (2019).

24. K. Frels, *et al.*, Genetic Diversity of Field Pennycress (Thlaspi arvense) Reveals Untapped Variability and Paths Toward Selection for Domestication. *Agronomy* **9**, 302 (2019).

25. S. I. Warwick, A. Francis, D. J. Susko, The biology of Canadian weeds. 9. Thlaspi arvense L. (updated). *Can. J. Plant Sci.* **82**, 803–823 (2002).

26. A. Nunn, *et al.*, Chromosome-level Thlaspi arvense genome provides new tools for translational research and for a newly domesticated cash cover crop of the cooler climates. *Plant Biotechnol. J.* (2022) https:/doi.org/10.1111/pbi.13775.

27. Y. Hu, *et al.*, Rapid Genome Evolution and Adaptation of Thlaspi arvense Mediated by Recurrent RNA-Based and Tandem Gene Duplications. *Front. Plant Sci.* **12**, 772655 (2021).

28. D. Galanti, *et al.*, Genetic and environmental drivers of large-scale epigenetic variation in Thlaspi arvense. *PLoS Genet.* **18**, e1010452 (2022).

29. Y. Bourgeois, S. Boissinot, On the Population Dynamics of Junk: A Review on the Population Genomics of Transposable Elements. *Genes* **10** (2019).

30. M. J. Dubin, *et al.*, DNA methylation in Arabidopsis has a genetic basis and shows evidence of local adaptation. *Elife* **4**, e05255 (2015).

31. E. Sasaki, J. Gunis, I. Reichardt-Gomez, V. Nizhynska, M. Nordborg, Conditional GWAS of non-CG transposon methylation in Arabidopsis thaliana reveals major polymorphisms in five genes. *PLoS Genet.* **18**, e1010345 (2022).

32. L. Quadrana, *et al.*, The Arabidopsis thaliana mobilome and its impact at the species level. *Elife* **5**, e15716 (2016).

33. R. M. Erdmann, C. L. Picard, RNA-directed DNA Methylation. *PLoS Genet.* **16**, e1009034 (2020).

34. S. Atwell, *et al.*, Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature* **465**, 627–631 (2010).

35. B. Pietzenuk, *et al.*, Recurrent evolution of heat-responsiveness in Brassicaceae COPIA elements. *Genome Biol.* **17**, 209 (2016).

36. E. W. Sayers, *et al.*, Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **50**, D20–D26 (2022).

37. U. Krämer, Planting molecular functions in an ecological context with Arabidopsis thaliana. *Elife* **4** (2015).

38. G. Bourque, *et al.*, Ten things you should know about transposable elements. *Genome*

Biol. **19**, 199 (2018).

39. T. Sultana, A. Zamborlini, G. Cristofari, P. Lesage, Integration site selection by retroviruses and transposable elements in eukaryotes. *Nat. Rev. Genet.* **18**, 292–308 (2017).

40. L. Quadrana, *et al.*, Transposition favors the generation of large effect mutations that may facilitate rapid adaption. *Nat. Commun.* **10**, 3421 (2019).

41. C. Stritt, M. Thieme, A. C. Roulin, Rare transposable elements challenge the prevailing view of transposition dynamics in plants. *Am. J. Bot.* **108**, 1310–1314 (2021).

42. S.-J. Zhang, L. Liu, R. Yang, X. Wang, Genome Size Evolution Mediated by Gypsy Retrotransposons in Brassicaceae. *Genomics Proteomics Bioinformatics* **18**, 321–332 (2020).

43. Y. Geng, *et al.*, Genomic analysis of field pennycress (Thlaspi arvense) provides insights into mechanisms of adaptation to high elevation. *BMC Biol.* **19**, 143 (2021).

44. S. Kim, *et al.*, Genome sequence of the hot pepper provides insights into the evolution of pungency in Capsicum species. *Nat. Genet.* **46**, 270–278 (2014).

45. E. Sasaki, T. Kawakatsu, J. R. Ecker, M. Nordborg, Common alleles of CMT2 and NRPE1 are major determinants of CHH methylation variation in Arabidopsis thaliana. *PLoS Genet.* **15**, e1008492 (2019).

46. L. Ichino, *et al.*, MBD5 and MBD6 couple DNA methylation to gene silencing through the J-domain protein SILENZIO. *Science* (2021) https:/doi.org/10.1126/science.abg6130 (June 4, 2021).

47. V. Specchia, M. P. Bozzetti, The Role of HSP90 in Preserving the Integrity of Genomes Against Transposons Is Evolutionarily Conserved. *Cells* **10** (2021).

48. U. Cappucci, *et al.*, The Hsp70 chaperone is a major player in stress-induced transposable element activation. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 17943–17950 (2019).

49. V. Specchia, *et al.*, Hsp90 prevents phenotypic variation by suppressing the mutagenic activity of transposons. *Nature* **463**, 662–665 (2010).

50. J.-N. Volff, Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. *Bioessays* **28**, 913–922 (2006).

51. D. Jangam, C. Feschotte, E. Betrán, Transposable Element Domestication As an Adaptation to Evolutionary Conflicts. *Trends Genet.* **33**, 817–831 (2017).

52. M. V. Almeida, G. Vernaz, A. L. K. Putman, E. A. Miska, Taming transposable elements in vertebrates: from epigenetic silencing to domestication. *Trends Genet.* **38**, 529–553 (2022).

53. H.-G. Drost, D. H. Sanchez, Becoming a Selfish Clan: Recombination Associated to Reverse-Transcription in LTR Retrotransposons. *Genome Biol. Evol.* **11**, 3382–3392 (2019).

54. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

55. A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

56. H.-G. Drost, LTRpred: de novo annotation of intact retrotransposons. *JOSS* **5**, 2170 (2020).

57. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

58. C. Camacho, *et al.*, BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

59. T. L. Bailey, J. Johnson, C. E. Grant, W. S. Noble, The MEME Suite. *Nucleic Acids Res.* **43**, W39–49 (2015).

60. A. McKenna, *et al.*, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

61. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).

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

63. P. Danecek, *et al.*, The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

64. B. L. Browning, Y. Zhou, S. R. Browning, A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am. J. Hum. Genet.* **103**, 338–348 (2018).

65. S. Purcell, *et al.*, PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

66. J. K. Pickrell, J. K. Pritchard, Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8**, e1002967 (2012).

67. P. Baduel, L. Quadrana, V. Colot, “Efficient Detection of Transposable Element Insertion Polymorphisms Between Genomes Using Short-Read Sequencing Data” in *Plant Transposable Elements: Methods and Protocols*, J. Cho, Ed. (Springer US, 2021), pp. 157–169.

68. M. Domínguez, *et al.*, The impact of transposable elements on tomato diversity. *Nat. Commun.* **11**, 4058 (2020).

69. F. A. Rabanal, *et al.*, Pushing the limits of HiFi assemblies reveals centromere diversity between two Arabidopsis thaliana genomes. *Nucleic Acids Res.* **50**, 12309–12327 (2022).

70. F. J. Sedlazeck, *et al.*, Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).

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

72. D. M. Emms, S. Kelly, OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).

73. G. Yu, L.-G. Wang, Y. Han, Q.-Y. He, clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).

74. F. Supek, M. Bošnjak, N. Škunca, T. Šmuc, REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* **6**, e21800 (2011).

Supplementary Information

Contreras-Garrido et al.: Transposon dynamics in the emerging oilseed crop *Thlaspi arvense*

Table S1. Summary statistics of previously annotated TEs for the *T. arvense* reference genome MN106-Ref (26)**.**

Order	Superfamily	Key	Number of families	Number of copies	% gGenomic space
Helitron	Helitron	DHH	132	24,224	2.01
TIR	hAT	DTA	74	7,452	0.768
TIR	CACTA	DTC	103	12,093	1.30
TIR	Harbinger	DTH	46	6,204	0.435
TIR	MuLE	DTM	218	18,041	1.53
TIR	Mariner	DTT	3	708	0.02
LINE	NonLTR/L1	RIC	4	217	0.05
LINE	I	RII	2	83	0.01
LINE	L1	RIL	80	10,785	0.768
LINE	R2	RIR	26	8,892	0.42
LINE	Undefined	RIX	76	11,321	1.217
LTR	Undefined	RLA	15	3,301	1.12
LTR	Ty1	RLC	310	37,531	6.3
LTR	Ty3	RLG	895	28,2391	48.2

Table S2. Lineages of LTR-TEs in the *T. arvense* genome MN106-Ref.

Superfamily	LTR lineage	Number of families	Number of individuals	% genomic space
Ty3	non-chromovirus|OTA|Athila	267	179,544	33.8
Ty3	non-chromovirus|OTA|Tat|Retand	58	24,890	2.9
Ty3	chromovirus	1	40	0.007
Ty3	chromovirus|CRM	120	29,864	3.2
Ty3	chromovirus|Galadriel	4	178	0.04
Ty3	chromovirus|Reina	38	1,907	0.3
Ty3	chromovirus|Tekay	94	57,078	5.6
Ty3	pararetrovirus	7	1,074	0.1
Ty1	Ale	108	25,351	3.3
Ty1	Alesia	1	144	0.07
Ty1	Angela	1	557	0.06
Ty1	Bianca	32	9,986	1.0
Ty1	Ikeros	9	587	0.1
Ty1	Ivana	42	3,185	0.4
Ty1	SIRE	24	3,303	0.8
Ty1	TAR	21	3,223	0.3
Ty1	Tork	35	2,068	0.3

Table S3. Additional information.

Contreras 2023 SOM: Transposon dynamics in the oilseed crop Thlaspi arvense.

- S3A: Accession numbers of samples sequenced in this study.
- S3B: Metadata of all accessions used in this study.
- S3C: Association of TE family name and the inferred lineage.
- S3D: Complete list of TIPs discovered in this study.
- S3E: Complete list of TAPs discovered in this study.
- S3F: Distribution of Alesia.FAM.7 in the reference genome.
- S3G: Detailed GO enrichment results of genes located within 2kb of Alesia.FAM.7 detected TIPs.
- S3H: Filtered *blastn* results of querying all the nucleotide sequences of the *Thlaspi arvense* TE models used in this study (26) against the NCBI NT database as per June of 2022.

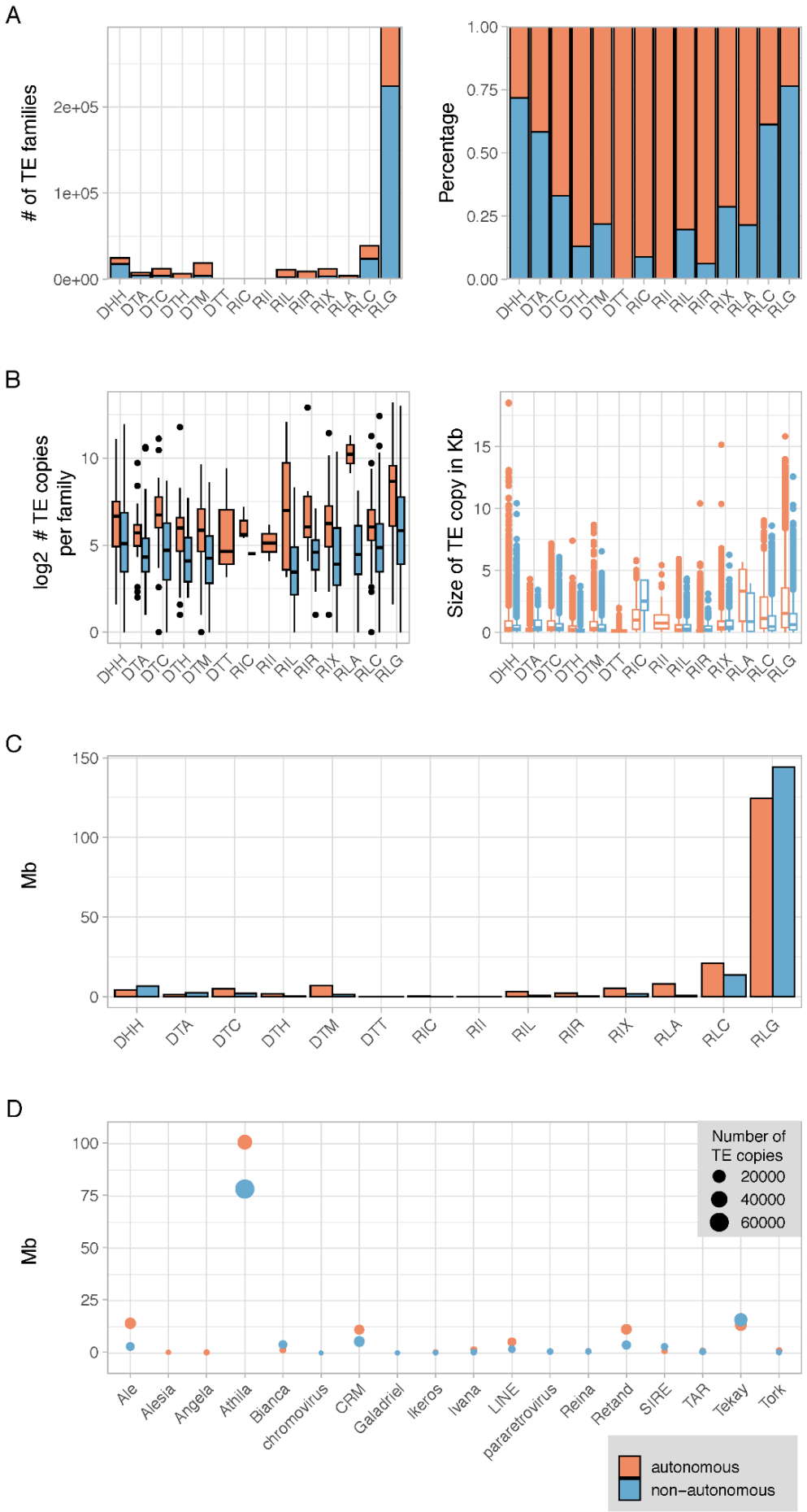


Figure S1. Comparison of autonomous and non-autonomous TE families in *T. arvense* MN120-Ref. A, Absolute (left) and relative (right) fraction of autonomous and non-autonomous elements in each TE superfamily. **B,** Comparison of the fraction of autonomous and non-autonomous elements in each TE superfamily (left). Size comparison of the TE copies according to their autonomy per superfamily (right). **C,** Contribution of each superfamily and their autonomous/non-autonomous fraction to total genome size in Mb. **D,** Distribution of size and copy number per LTR retrotransposon lineage.

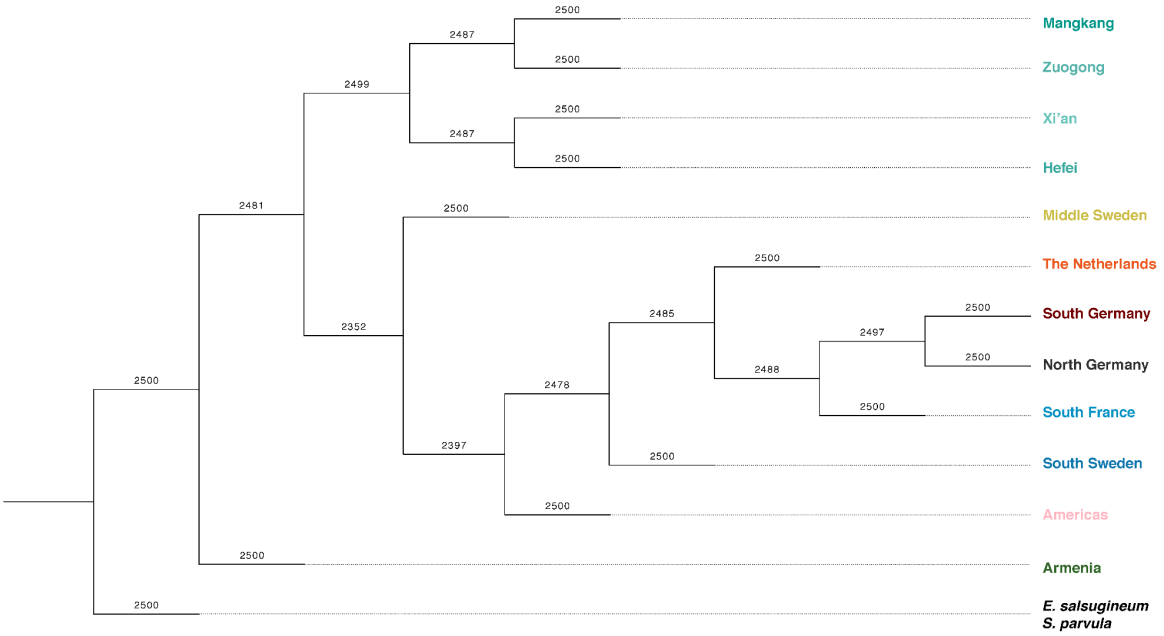


Figure S2. SNP-based maximum likelihood tree of *T. arvense* populations. Based on a model without migration, 2,500 bootstraps. Node weights represent bootstrap values. Outgroup species at the bottom.

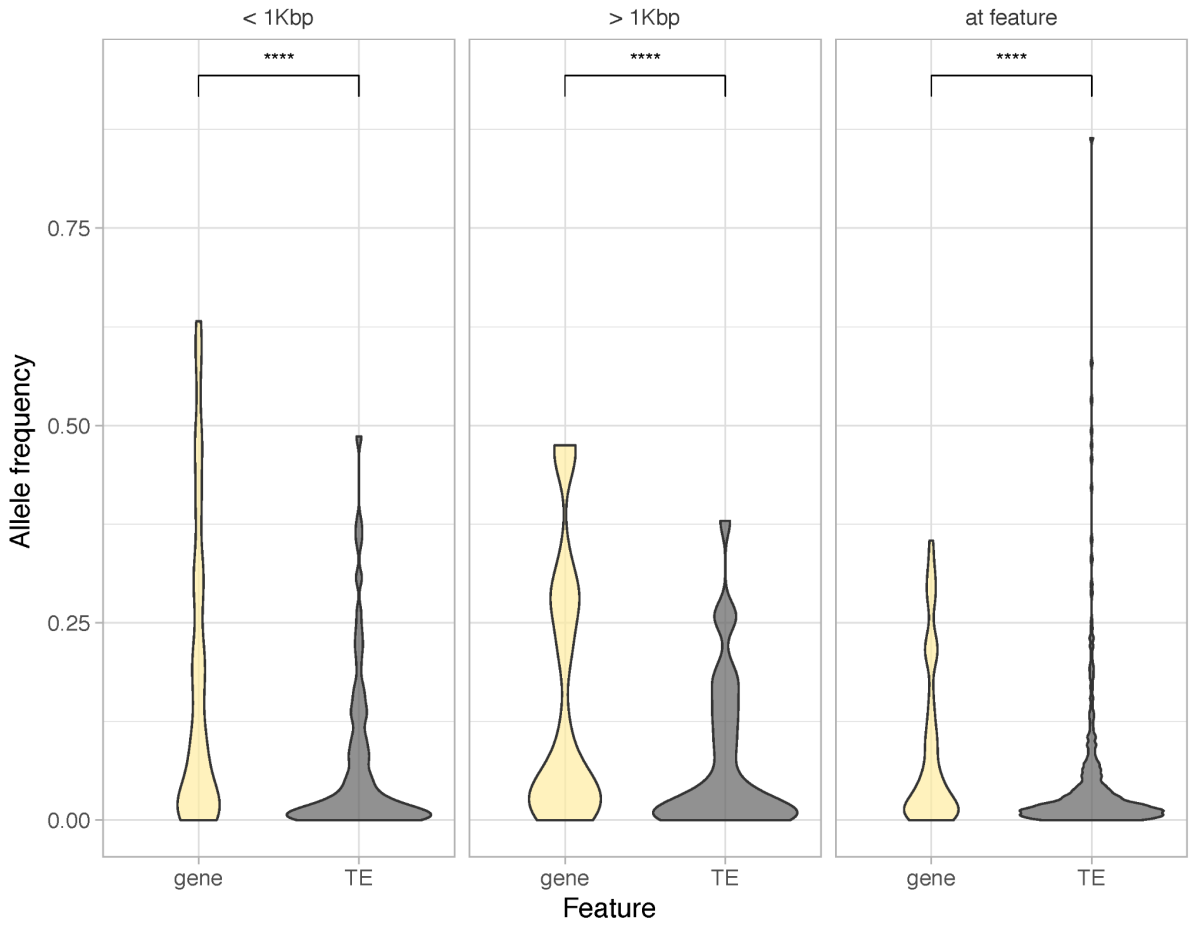


Figure S3. Frequency distribution of TIPs overlapping with annotated genes and TEs. TIP frequencies near other TEs are significantly lower than near genes (Wilcoxon Rank Sum test, $p < 2.22E^{-16}$).

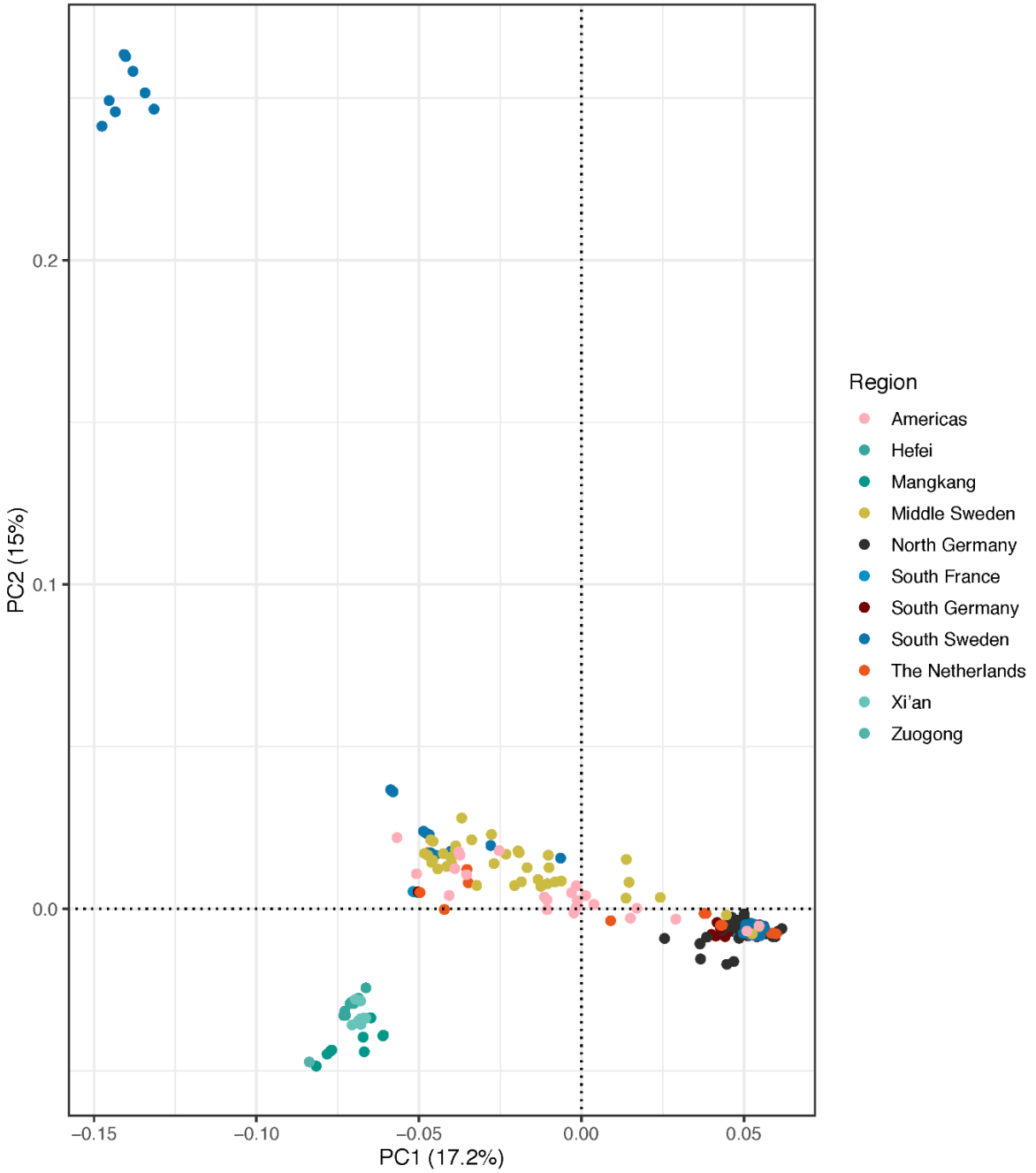


Figure S4. SNP-based PCA of a subset of *T. arvense* accessions. The Armenian accessions, which are outliers in the PCA using all accessions (Fig. 2), were excluded from this new PCA analysis, which shows how Chinese and European accessions cluster separately. We also observe part of the south Sweden accessions clustering far from the rest of the European accessions.

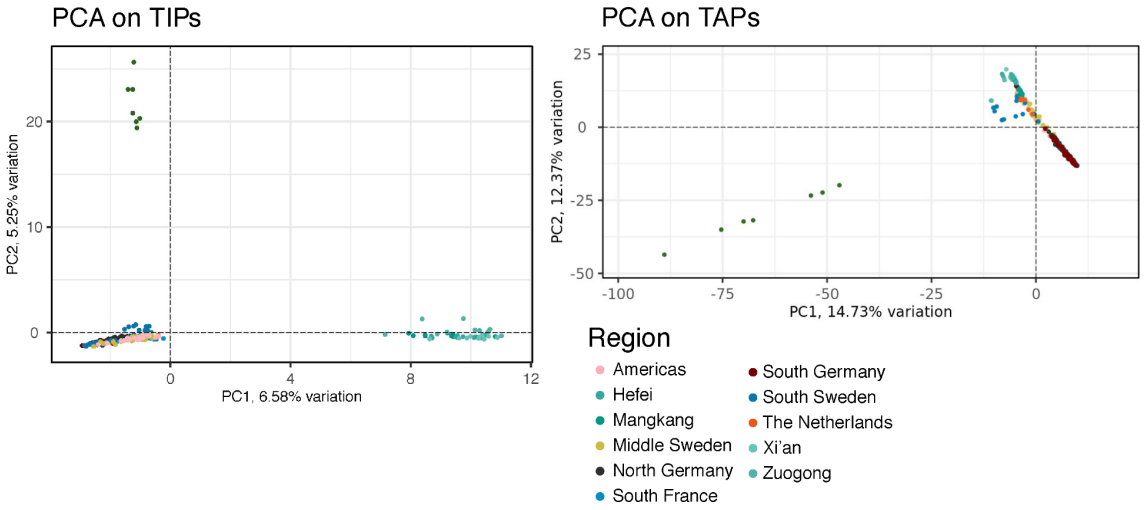


Figure S5. PCA analysis of 279 individuals of *T. arvense*. A presence/absence matrix of either TIPs (left) or TAPs, (right) was used as input to calculate PCA. This results recapitulates the clustering pattern observed with the SNP-PCA.

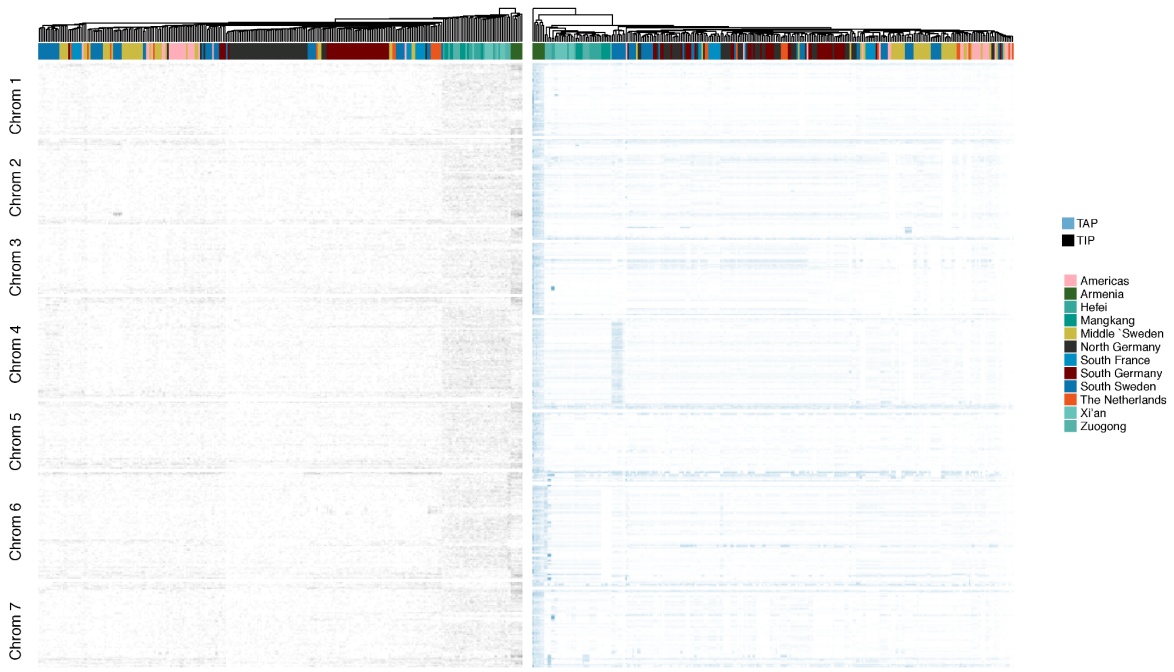
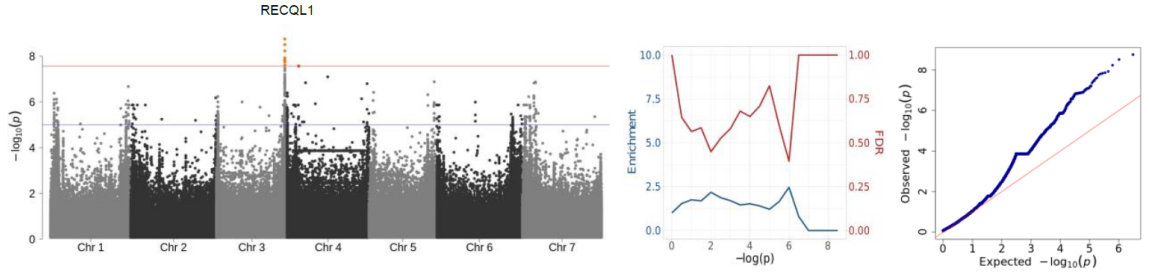
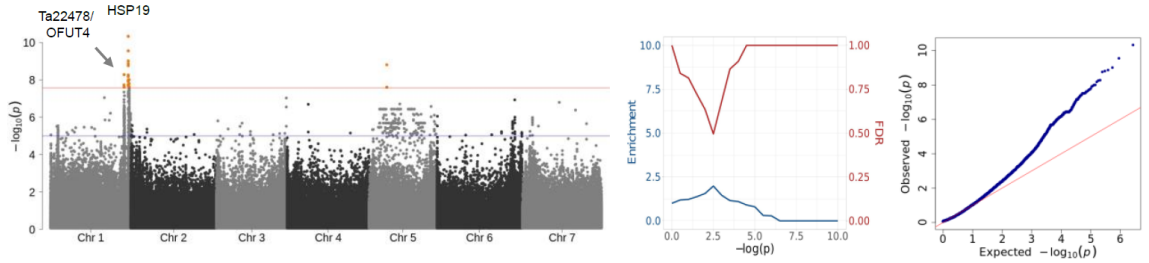


Figure S6. Genomic distribution of TIPs and TAPs along all seven chromosomes of *T. arvense*. Color columns indicate to which biogeographical population each accession belongs to.

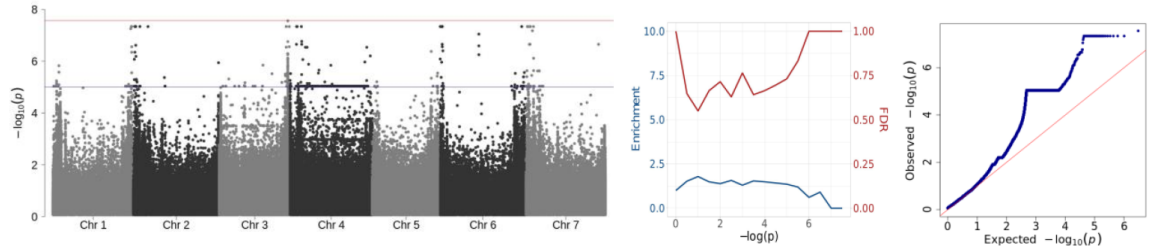
All TIPs



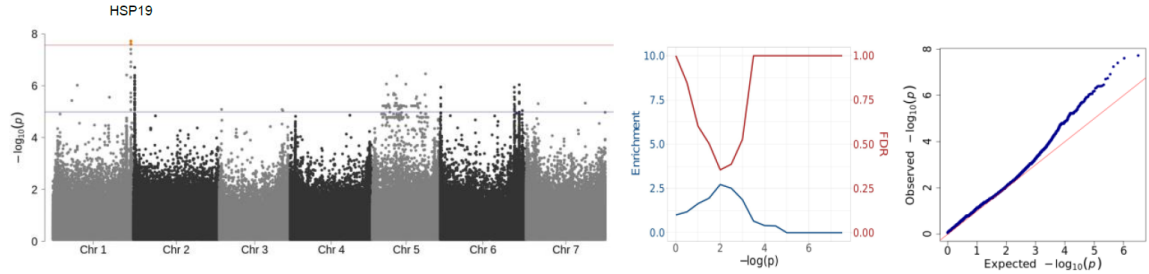
Class II



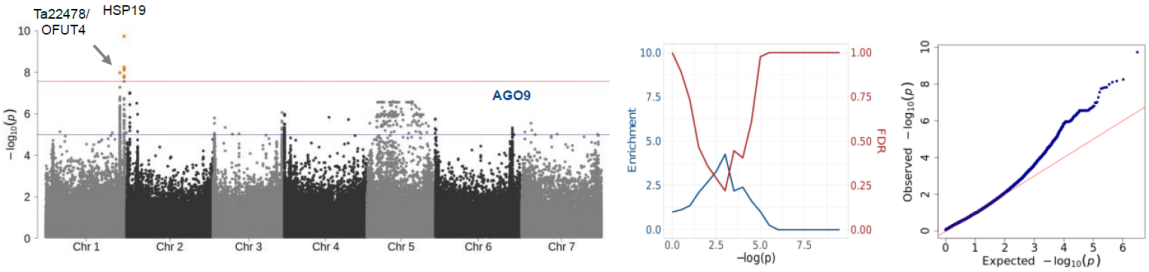
Class I



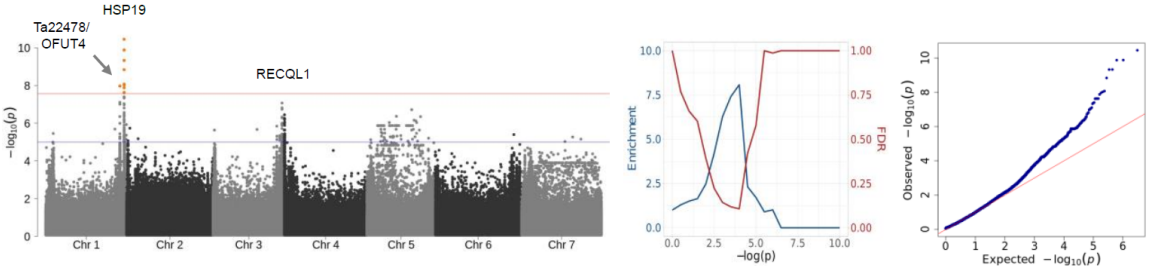
DTA



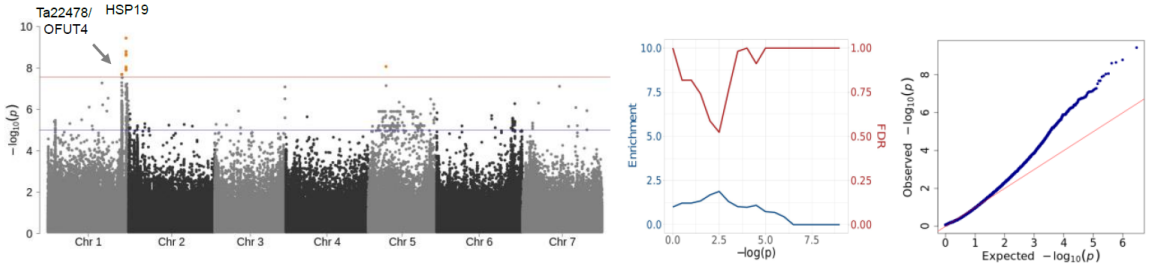
DTC



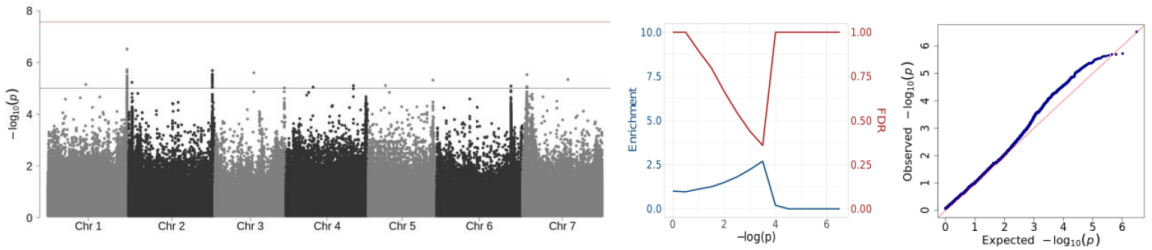
DTH



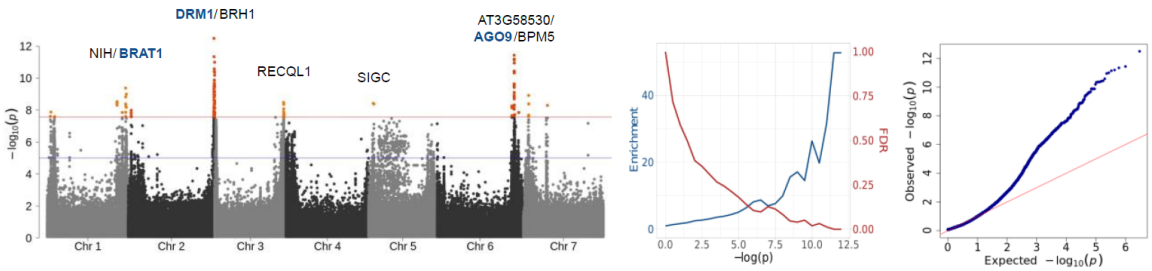
DTM



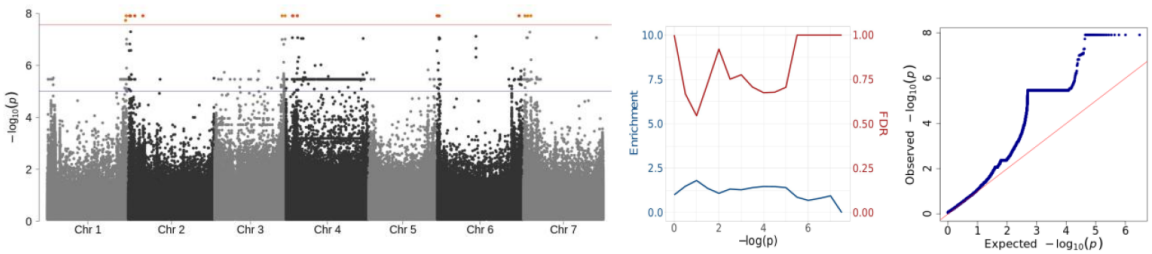
DTT



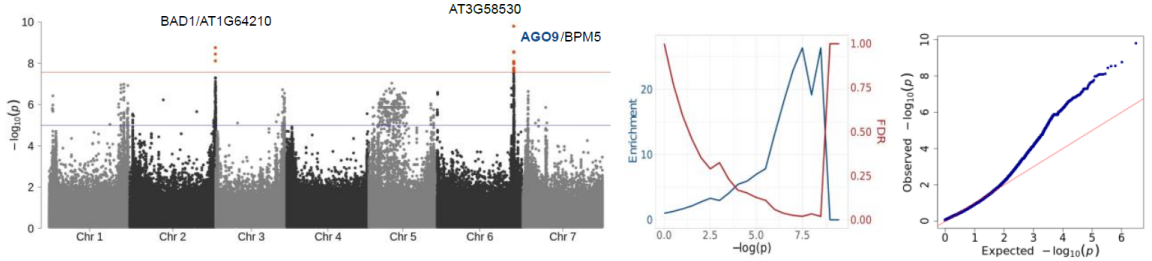
LINE



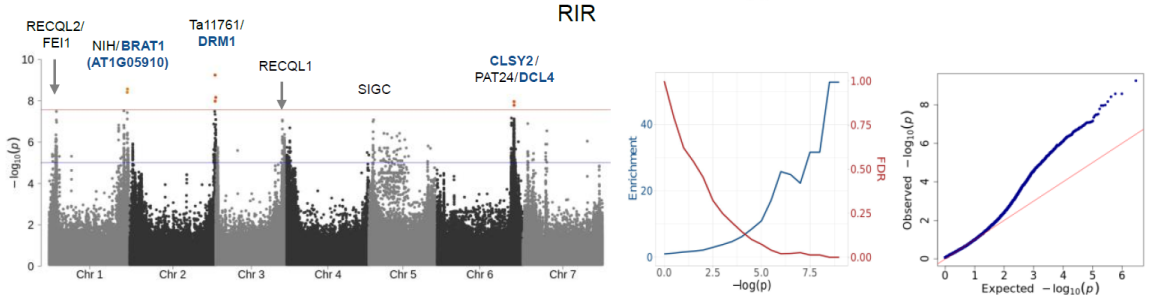
LTR



RIL



RIR



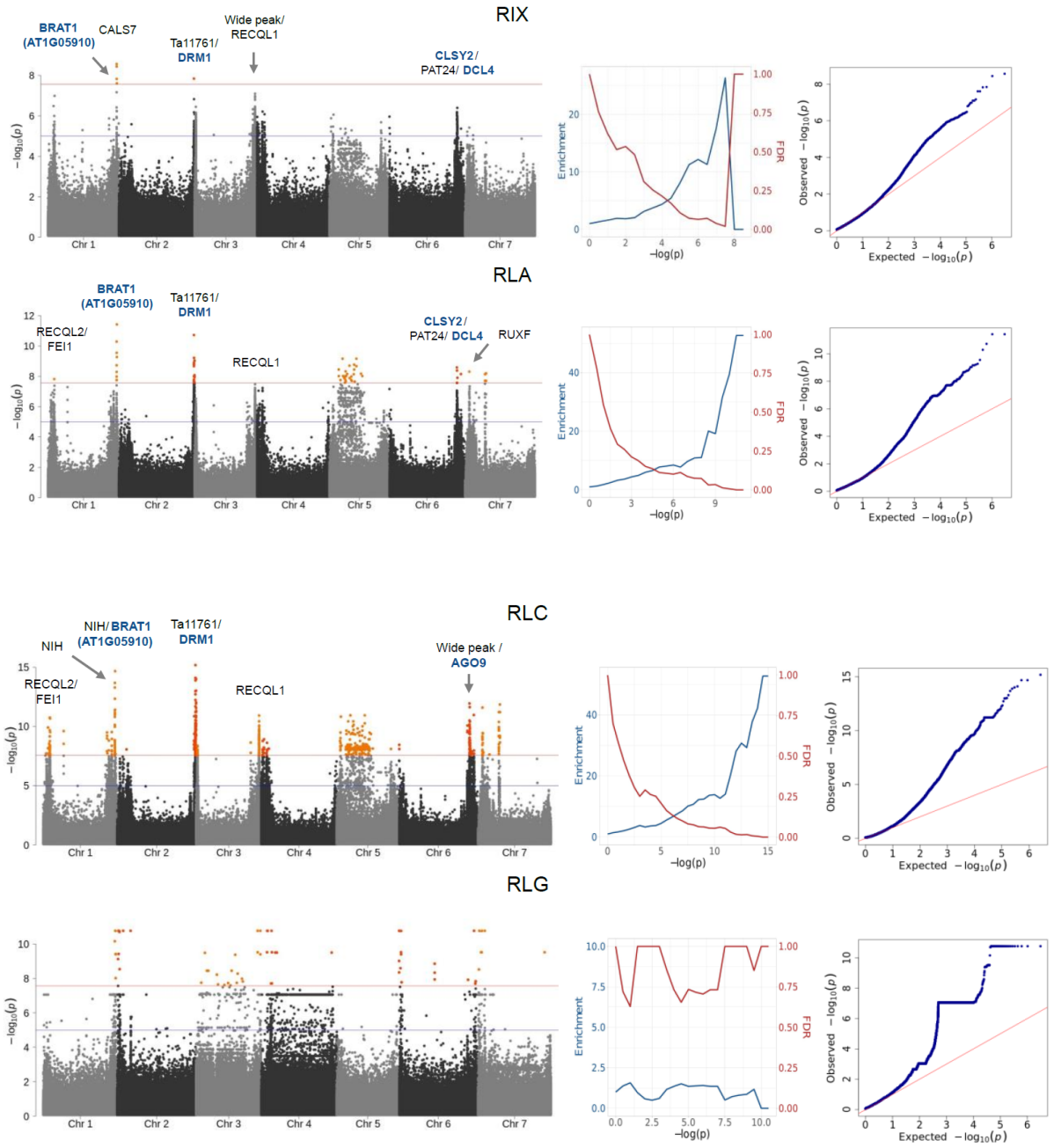


Figure S7. Complete GWA results for TIP load. Left: Manhattan plots for each TIP superfamily load. The genome-wide significance (red line) corresponds to a full Bonferroni correction, the suggestive line (blue) to a more generous hard threshold of –log(p)=5. Genes next to top variants are labeled with names, blue font indicates genes with link to DNA methylation included in the enrichment analyses. Middle: Enrichment and expected FDR of genes with link to DNA methylation, for significance threshold increments (28, 34). Right: QQplots of p-values.

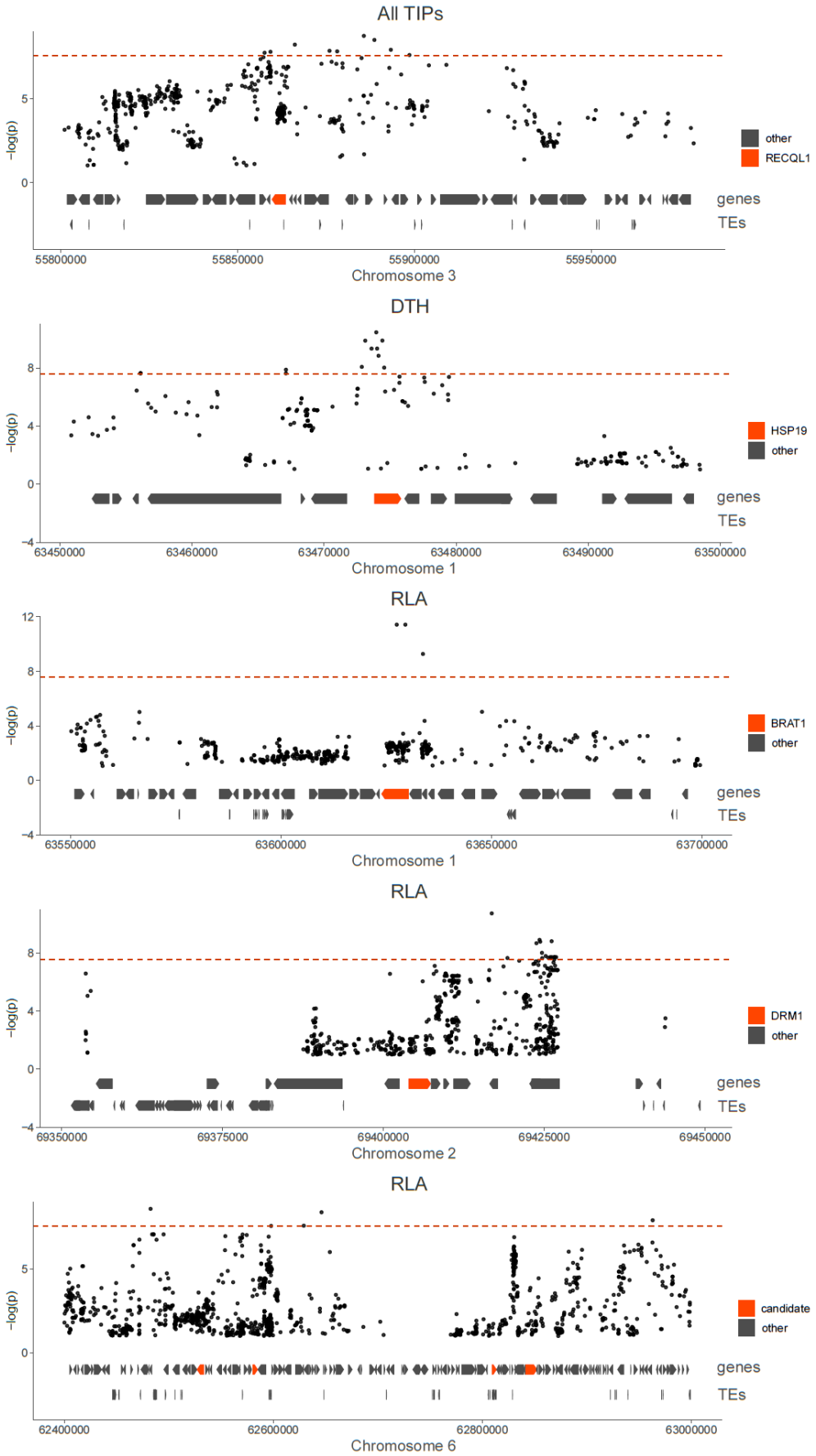


Figure S8. Zoom-in of GWA peaks with candidate genes highlighted in red. The genome-wide significance (dotted red line) corresponds to a full Bonferroni correction.

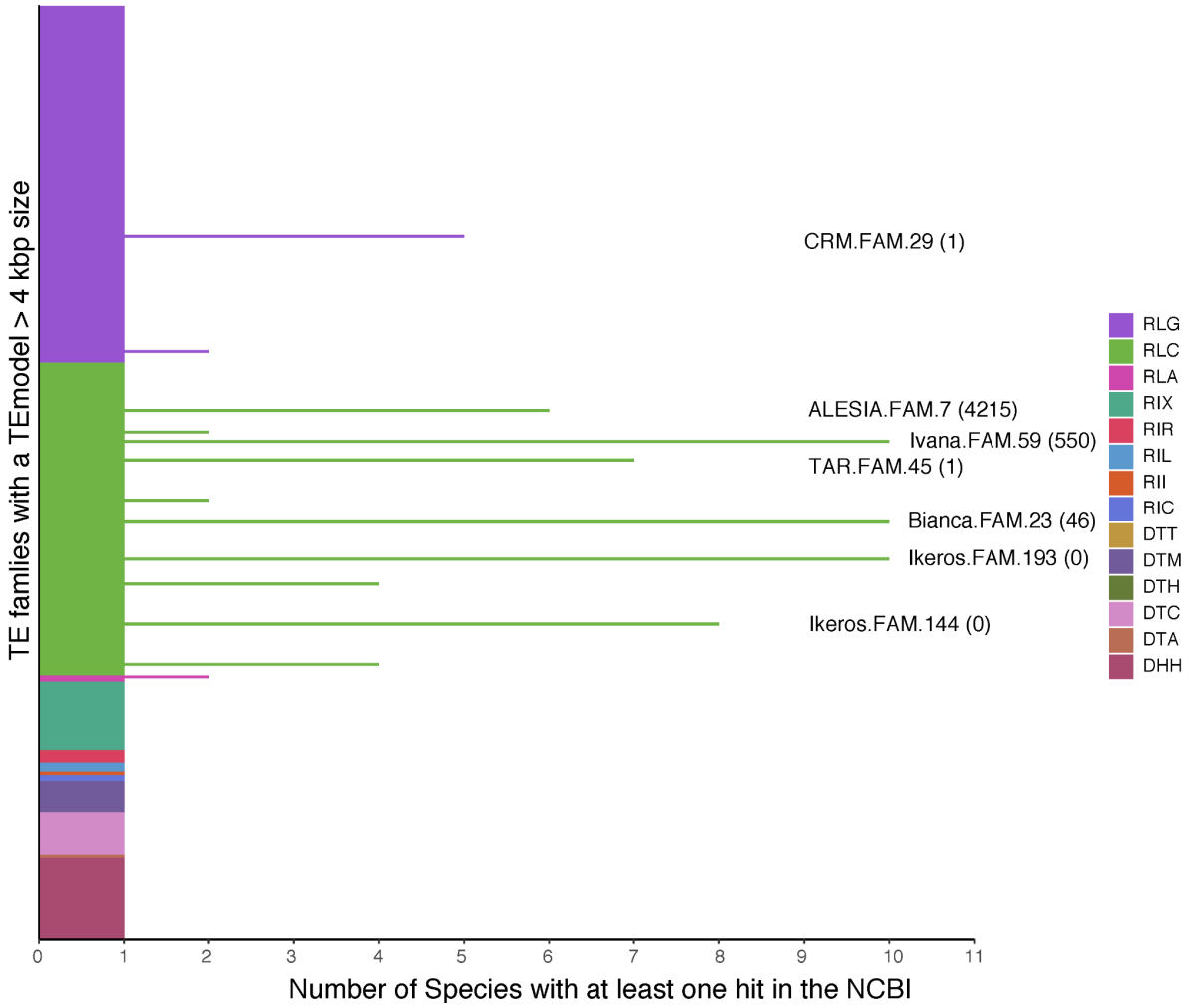


Figure S9. BLASTN hits of *T. arvense* TE families with model sizes > 4 kb against the NCBI NT database (36), **June 2022 release.** We filtered the matches using the 80/80/80 rule, and further constrained matches to fulfill > 2kb length criteria. The x-axis denotes the number of species with at least 1 hit. Each family has at least one hit, namely *T. arvense* itself. TE families with more than 5 hits are highlighted. The number of TIPs in *T. arvense* populations is shown in parentheses for the highlighted families to indicate that there is no obvious correlation between mobility in *T. arvense* and phylogenetic conservation.