The Epigenomic Impact of Transposable Elements in Natural Populations

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Adrián Contreras Garrido
aus Valencia, Spanien

Tübingen
2023

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	09.02.2024
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Detlef Weigel
2. Berichterstatter:	Prof. Dr. Oliver Bossdorf
3. Berichterstatter:	Dr. Leandro Quadrana

Acknowledgments

I would not have been able to write these pages without the opportunity provided by Detlef Weigel to work as a researcher in his laboratory and I am very grateful for it as it has shown me a world I have never thought existed. I am also grateful for the time and freedom that was given to me to find the little corner of research I truly enjoy, it wasn't easy!

I would also like to thank my thesis advisors: Oliver Bossdorf, Koen Verhoeven and Hajk-Georg Drost for their time and guidance through this journey.

I am very thankful to Rebecca Schwab for sharing her experience in both research and academia. I will never stop learning from her. PS: The best advice always comes after 8 pm. Anjar Wibowo and Thanvi (T)Srikant had immense patience with me at the bench and showed me two great approaches for bench work. They have been the foundations of this thesis work,you will not see it, but they hold the entire thing.

I had the fortune to be part of a big consortium group, Epidiverse, that allowed me to meet tons of people and have access to cool science. In particular, I would like to thank, both as colleagues and friends, Adam Nunn for teaching me the beginning of python, Panpan Zhang for teaching me the tools of TE annotation and Dario Galanti for the fruitful discussions and projects we have done together.

Back at the lab, I am thankful to all WeigelWorld people for creating an amazing lab environment in which I feel fortunate to walk to every day. I would like to thank old lab members: Sergio Latorre and Cristina Barragan for the Blauer times and parties, they showed me the night side of Tubingen; and to Pablo Carbonell for the homemade Paellas. To present lab members: Ale, Miriam, Andrea and Natalie for bearing with my bad jokes and my science sermons, as if I know anything. To my office mates: Haim, Fernando and Lukas, for the scientific and not-so-scientific conversations, suggestions and feedback. Haim, we still need to finish that book!

I would like to thank Luisa, Kevin and, specially, Gautam for giving me the chance to work in a super collaborative project and for all the discussions about pop-gen, alignments and phylogenies. It has been a pleasure!

Finally, I would like to thank my mum, Pilar, and my family. It hasn't been easy to be abroad, and I have missed you a lot. Thanks for always supporting me and bearing with my mood changes all these years.

Table of Contents

Summary

Transposable elements (TEs) are a heterogeneous collection of DNA sequences characterised by their ability to relocate to new sites of the genome, employing either a cut-and-paste mechanism or an RNA intermediate. As a result, TEs are able to quickly replicate through the genome to such a degree that, often, they constitute the majority of the genome in plant species. Their replicative activity is usually harmful for their host genomes, thus, TEs tend to be considered genomic parasites and their activity is constantly suppressed by the host genome. In plants, TE silencing is achieved via a combination of the three epigenetic marks that maintain genome stability. One of such marks, small RNAs (sRNAs), consist of single stranded RNA molecules ranging from 20 to 24 nucleotides in size generated from a longer double stranded precursor. These sRNAs guide the epigenetic machinery to a particular region of the genome using a combination of DNA methylation and histone modifications that change the chromatin conformation of the region, modifying its expression. Hence, vast parts of the plant genome are epigenetically silenced, leading to important phenotypic consequences. Despite this, reference genomes are often generated with a comprehensive annotation of protein coding genes and the mRNAs they produce, but this offers only a partial view of genome functions, many of which involve epigenetic mechanisms.

In this thesis, we generated a new reference genome for the emerging oilseed crop *Thlaspi arvense* (field pennycress) with a special focus on the *de novo* annotation of TEs and small RNA loci (sRNA). We annotated 423,249 individual TEs, which together constitute 61% of the *T. arvense* genome, most of it were long terminal retrotransposons (LTR). To understand how TE activity is regulated, we complemented our TE annotation with sRNA data. Applying a custom pipeline to data from leaf, root, inflorescence and pollen, we identified 19,288 distinct sRNA loci, of which 72 were microRNAs. Then, I examined the dynamics of TE variation in a geographically diverse sample of this species. By surveying almost 300 wild accessions ranging from America to Eurasia, I discovered over 90,000 polymorphic TE insertions and ten times more TE deletions.

In parallel, I also conducted a *de novo* TE annotation of a set of *Arabidopsis thaliana* genomes to investigate to what extent TE activity shapes the plant repertoire of a set of disease resistance genes, Nucleotide-binding site leucine-rich repeat (NLR) genes. Comparing NLR loci with their genomic background, I showed that these NLR loci contained a higher proportion of young LTR TEs and a higher proportion of solo LTRs> Comparisons of these NLR loci between accessions also revealed the high prevalence of intraspecific TE variability at these loci. Altogether, this work contributes to the understanding of the nature of the genomic processes that generate the necessary NLR diversity needed to survive in a constantly changing, pathogen-loaded environment.

Zusammenfassung

Mobile genetische Elemente (TEs) sind eine heterogene Gruppe von DNA-Sequenzen, die ihre Position im Genom verändern können - entweder mittels Cut-and-Paste Mechanismen oder über ein RNA-Zwischenprodukt. Dies hat zur Folge, dass in einigen Pflanzenarten das Genom hauptsächlich aus TEs besteht. Die TE Replikation ist in der Regel schädlich für das Wirtsgenom, weshalb TEs als genomische Parasiten angesehen werden und ihre Aktivität normalerweise unterdrückt wird. In Pflanzen gibt es epigenetische Mechanismen, die zu dieser Stillegung beitragen: einzelsträngige kleine RNAs (sRNAs) von 20 bis 24 Nukleotiden Länge, generiert aus einem längeren doppelsträngigen Vorläufer, bestimmen die Region im Genom, an der über DNA-Methylierung und Histon-Modifikationen die Chromatin-Konformation in einen epigenetisch stillgelegten Bereich transformiert wird. Dies hat erhebliche phänotypische Auswirkungen. Trotzdem liegt der Fokus bei der Annotation von Genomen auf proteinkodierenden Genen und nicht auf TEs.

In dieser Arbeit habe ich ein neues Referenzgenom für die Ölsamenpflanze Thlaspi arvense L. (Acker-Hellerkraut) erstellt, mit einem Schwerpunkt auf de novo Annotation von TEs und kleinen RNA-Loci (sRNA). Die identifizierten 423,249 TEs machen 61% des Genoms aus, wobei “long terminal retrotransposons” (LTR) die Mehrheit darstellen. Um zu verstehen, wie die TE-Aktivität reguliert wird, habe ich zusätzlich kleine RNA-Expressions-Daten generiert, wofür RNA aus Blättern, Wurzeln, Blütenständen und Pollen isoliert wurde. Mittels einer selbsterstellten Pipeline konnte ich 19,288 Loci identifizieren, von denen 72 microRNAs-Loci waren. Mit diesem Wissen habe ich untersucht, wie stark sich TEs über eine Pflanzenpopulation geografisch diversen Ursprungs unterscheiden: in fast 300 Akzessionen mit Ursprung von Amerika bis Eurasien habe ich über 90,000 polymorphe TE-Insertionen und zehnmal mehr TE-Deletionen entdeckt.

Auch habe ich in einer Reihe von Arabidopsis thaliana Genomen TEs annotiert, um zu untersuchen, inwieweit TE-Aktivität das Repertoir von Resistenzgenen der “Nucleotide-binding site leucine-rich repeat”-Klasse (NLR) geprägt hat. Es hat sich gezeigt, dass NLR-Loci einen höheren Anteil an jungen LTR-TEs und an Solo-LTRs haben im Vergleich zum restlichen Genom. Auch zeigte sich hier eine hohe TE-Variabilität zwischen den Akzessionen. Dies erlaubt Rückschlüsse auf die Evolution der NLR-Loci, deren Diversität in einer Umwelt mit vielen Krankheitserregern bedeutsam ist.

Introduction

1. Epigenetic mechanisms in plants.

Evolutionary epigenetics is the study of heritable phenotypic and gene expression differences that do not involve changes in the nucleotide sequence of the genome. It is often considered a fine-tuning mechanism that allows organisms to modulate genome activity for specific developmental needs and to cope with constant environmental changes. It can allow for the generation of phenotypic variation among progeny in a reversible manner without making extensive and permanent changes to the genetic blueprint [1]. Under this broad definition, a wide spectrum of molecular mechanisms such as short interfering RNAs (siRNAs) and other small RNAs (sRNAs), DNA methylation, and chromatin modifications have been integrated in the field of epigenetics [2].

Epigenetic differences lead to changes in nucleosome composition and arrangement that in turn affect the accessibility of the local DNA and its transcriptional activity. Thus, the epigenetic makeup of a genome controls the activity of genes and regulatory elements, and provides the basis of genome stability by silencing the repetitive content of the genome, including transposable elements (TEs) and other repeats. Other roles of epigenetic marks in the genome are the maintenance of the genetic content by modulating the establishment of homologous recombination, which is the basis for many mutational events, that follows a double strand break [3]; the interplay between DNA methylation and DNA repair pathways in plants [4] and the delimitation of the centromere [5]. In fact,having a centromere determined epigenetically instead of by the specificity of its sequence , could be advantageous, as it helps maintain the function of the centromere in the separation of sister chromatids even in the event of mutational damage to the centromere locus [6,7].

Despite these critical roles as a homeostatic force of genome stability, epigenetic marks are sensitive to environmental cues. These alterations are usually viewed as a response of the organisms to the environmental stress, inducing a genome-wide transcriptional reprogramming to a stress-resilient state [8]. Still, epigenetic marks can, in some cases, be stable enough to be inherited between generations. It has been shown experimentally that induced differential epigenetic states in a population can be stably inherited in following generations, creating epigenetic variation that can contribute to selection [9]. Widely studied phenomena that seem not to follow Mendelian laws of inheritance such as epialleles [10], maize paramutations [11] or genomic imprinting [12] can be attributed to epigenetically-driven inheritance, although how many of these phenomena are exclusively independent of trans-acting DNA variants is difficult to determine [1].

1.1. DNA methylation and histone modifications.

DNA methylation refers to the addition of a methyl group to one of the DNA bases, most often to cytosine to form 5-methylcytosine. DNA methylation is conserved in both animals and plants, although plants have an increased complexity and redundancy in their epigenetic mechanisms as discussed below. This epigenetic mark is involved in several biological processes. In the model plant *Arabidopsis thaliana*, disruption of methylation results in late flowering, reduced plant size and slow growth [13]. In *Solanum lycopersicum* (tomato), active DNA demethylation mediates fruit ripening [14], while disruption of *de novo* deposition of DNA methylation results in strong developmental defects in *Zea mays* (maize) [15] and *Oryza sativa* (rice) [16]. Besides its role in plant development, DNA methylation changes during biotic [17] and abiotic stresses [18] alter chromatin structure and global gene expression; these changes are hypothesised to be a systemic response to such stresses. For example, methylation-impaired *A. thaliana* mutants have lower survival during salt stress than wild-type plants [19].

Generally speaking, DNA methylation acts via modification of gene expression. Usually, methylation in the promoter of a gene inhibits its transcription, but not always [20]. As with genes, DNA methylation also represses transcription of TEs. Indeed, silencing of genomic repeats, including TEs, is essential for promoting genome stability.

Cytosine methylation occurs in three different genetic context in plants: CG, CHG and CHH (where C stands for cytosine, G stands for guanine, and H stands for any nucleotide but guanine), in mammals DNA methylation occurs only in the CG context [21]. Mammals a use the enzyme DNA methyltransferase 3 (DNMT3) to catalyse the *de novo* addition of a methyl group in the CG context, whereas In plants, this is done by an homologous of DNMT3, DOMAINS REARRANGED METHYLTRANSFERASE 2 (DRM2) [20]. To maintain methylation patterns, mammals only use one enzyme, DNA methyltransferase 1 (DNMT1), while plants' methylation patterns are maintained by three pathways [21]. In the symmetrical contexts CG and CHG, methylation maintenance takes place during DNA replication via a semiconservative mechanism. CG methylation is maintained by DNA METHYLTRANSFERASE 1 (MET1), a homolog of DNMT1, while CHG methylation is maintained by CHROMOMETHYLASE 3 (CMT3), a plant-specific DNA methyltransferase. However, because CHH methylation is asymmetric, its maintenance is achieved through persistent *de novo* methylation by DRM2 [22].

The *de novo* CHH methylation is catalysed by various enzymes that are targeted to specific genomic regions by specific sets of small RNAs, in a pathway called RNA-directed DNA methylation (RdDM) [23]. This pathway plays a pivotal role in the repression of TEs and other repetitive DNA sequences that are found not only in heterochromatin regions but also in euchromatic chromosome arms [24]. The use of small noncoding RNAs to guide the silencing of TEs and other foreign genetic elements via the deposition of repressive epigenetic marks is a highly conserved strategy that most eukaryotic organisms rely on to combat genomic parasites [25].

In addition to DNA methylation, another well-studied epigenetic mark is the modification of histones. Histones are the key components of the nucleosome, the basic repeat subunit of the chromatin. Nucleosomes are composed of a 147 bp stretch of DNA and a histone octamer with two molecules of each of the four core histones H2A, H2B, H3, and H4. DNA is wrapped around this histone octamer forming a beads-on-a-string structure, which is heavily conserved across eukaryotes [26]. Alongside these canonical histones, histone variants are replication- and deposition-independent proteins that alter nucleosome composition and behaviour. Some of these variants are conserved across eukaryotes, like H2A.Z, while others are lineage specific, such as H2A.W, which is specific to flowering plants [27]. Histone variants alter the accessibility of the DNA by changing the conformation of the nucleosome. This has an impact in all DNA- and transcriptional-related processes, including post-transcriptional DNA modifications and DNA methylation deposition [27]. Specific histone variants are also associated with the major chromatin types: in plants, CenH3 histone variant is located at the centromeres, H2A.W colocalizes with pericentromeric regions, and H3.3, H2A.Z and H2A.X are associated with the euchromatin regions [27].

Importantly, epigenetic marks are heavily interconnected. In *A. thaliana* for example, the H3.3 histone variant stimulates DNA methylation at gene bodies [28], whereas deposition of DNA methylation marks causes H2A.Z depletion [29].

1.2. Molecular Mechanisms of TE Epigenetic Regulation.

The importance of the epigenetic silencing of TEs for the stability of the genome can be easily understood if we consider that most of the extensively methylated fraction of the genomes (up to 80%) is composed of inactive, heterochromatic regions including clusters of tandem, inverted and interspersed repetitive elements and TEs. In *A. thaliana*, TEs alone comprise around 20% of the total genome [30] and are frequently located within or around the centromeric regions [31] but they can also be found on the euchromatic regions of the chromosome.

In plants, epigenetic silencing of TEs is achieved via a combination of the three epigenetic marks discussed earlier: DNA methylation, siRNAs and histone deposition. TE silencing can be divided in three different stages: initiation of silencing, establishment of the silencing, and stable maintenance of silencing. Of these three stages, both establishment and maintenance of silencing are well understood.

Maintenance of TE silencing relies on two mechanisms, one is the symmetrical methylation of DNA at CG and CHG sites across cell divisions and the other is the continuous positive feedback loop between DNA methylation and histone lysine 9 dimethylation (H3K9me2) that reinforce each other to create stable heterochromatic state [20].

Establishment of epigenetic TE silencing is achieved through the action of small RNA pathways that induce methylation in all three sequence contexts, CG, CHG and CHH. These pathways are the canonical and non-canonical RNA mediated DNA methylation (RdDM) pathway and they are discussed in the following section.

The first stage of TE silencing, *de novo* establishment of TE repression, has been less studied in plants [32]. In *A. thaliana*, a study found that exogenous TEs with no homology to genomic TEs are targeted by multiple siRNA producing pathways in an expression-dependent manner, whereas newly introduced copies of TEs already present in the genome are targeted by RdDM produced siRNAs in a ´manner that does not require expression of the introduced TE sequences [33]. Thus, epigenetic silencing by DNA methylation and histone modification of TEs requires the crucial triggering and action of small RNAs in the plant to establish epigenetic silencing.

The role of small interfering RNAs.

Canonical small interfering RNAs (siRNAs) are 24 nucleotide long non-coding RNA molecules that participate in the post-transcriptional gene silencing pathway (PTGS) [34]. In plants, they are easily distinguishable from the other type of common small RNAs, microRNAs (miRNAs), which are 20–22 nucleotides long, although there are also non-canonical siRNAs in this size range. These small molecules are generated by long double stranded RNA precursors by the action of RNase III endonucleases of the DICER family [35] and then loaded into the AGO proteins which, broadly speaking, guided by these siNRAs, will either cleave a mRNA (PTGS) or direct the methylation of particular loci (RdDM) [36]. The PTGS pathway is mainly implicated in the silencing of transgenes and viruses, but it can also silence TEs [37]. PTGS and RdDM are evolutionarily related and, as we will see, share some common features [38].

siRNAs emerge from intergenic and repetitive genome regions and are in turn used by the RdDM machinery to guide them to these regions. Thus, these siRNAs mainly target the regions where they originate, but due to mismatches in the machinery they can also act in *trans* on sequence-related regions.

Defective production of siRNA can lead to an increased insertion of retrotransposons in the *A. thaliana* genome after plants experience abiotic stresses [39,40]. RdDM is also involved in biotic stresses. For example, infiltration of *A. thaliana* leaves with the bacterial flagellin-derived peptide flg22 triggers suppression of RdDM factors, which is correlated with DNA demethylation of some RdDM targets, including TEs and promoters of some immune-responsive genes, and thus can lead to transcriptional activation [41]. This suggests additional roles for the RdDM machinery.

For these reasons, constant siRNA production to feed the RdDM pathway is crucial to ensure heterochromatin formation and genome stability [25] during the *A. thaliana* life cycle.

Details of the RNA-Directed DNA Methylation (RdDM) pathway.

RdDM is initiated when two polymerases, the plant-specific RNA POLYMERASE IV (POL-IV) and RNA-DEPENDENT RNA POLYMERASE 2 (RDR2) collaborate to produce double-stranded RNA (dsRNA). These dsRNAs are then cleaved into 24-nucleotide siRNAs by DICER-LIKE 3 (DCL3), loaded into ARGONAUTE proteins (AGO4 and AGO6), and paired with complementary scaffold RNAs produced by the plant-specific RNA POLYMERASE V (POL-V). The AGO-siRNA complex then recruits the methyltransferase protein DOMAINS REARRANGED METHYLTRANSFERASE 2 (DRM2) to catalyse DNA methylation at POL-V-transcribed regions [38]. See Figure 1 for a summary.

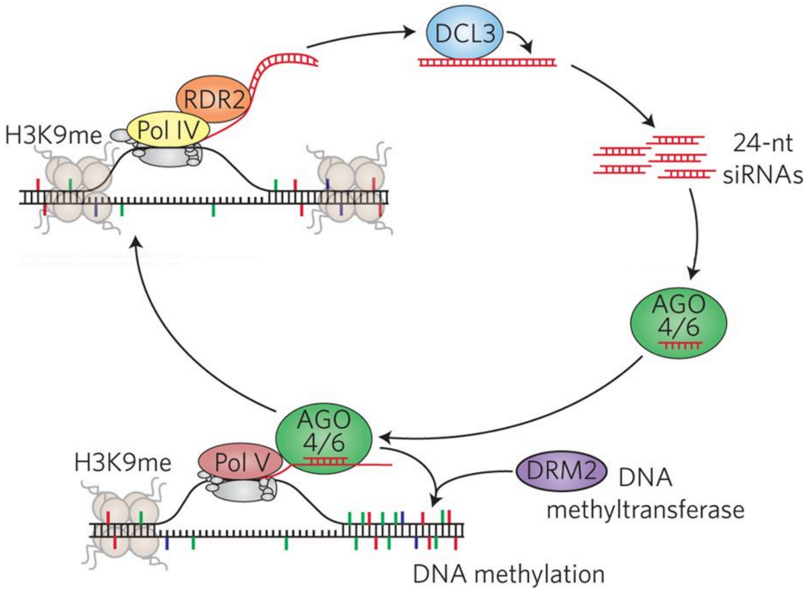


Figure 1 | Overview of the RdDM pathway. During the initial phase of siRNA generation, Pol IV and RDR2 produce a double-stranded RNA (dsRNA), which is subsequently cleaved into 24-nucleotide siRNAs by DCL3. These siRNAs guide AGO4 or AGO6 proteins to chromatin-bound transcripts generated by Pol V. In the next phase, the interaction between AGO4/6 and the Pol V transcript leads to the recruitment and activation of the DNA methyltransferase DRM2. In the diagram, black strands represent DNA, while red strands represent RNA. The colours used to represent DNA methylation indicate different sequence contexts, with CG in red, CHG in blue, and CHH in green, where H can be A, C, or T. Modified from [42]*).*

The most accepted hypothesis is that both POL IV and POL V are recruited to participate in the canonical RdDM through previously-established heterochromatic marks at the target loci. Thus, RdDM cannot, by itself, explain how heterochromatic marks are initiated at a target locus. POL IV is recruited to regions of the genome that are associated with CG methylation, histone deacetylation and dimethylated histone H3 lysine (9H3K9me) marks through an intermediate binding protein SAWADEE HOMEODOMAIN HOMOLOGUE 1 (SHH1) [43], while Pol V is recruited to regions of the genome associated with DNA methylation through two redundant histone methyltransferases, SET DOMAIN-CONTAINING PROTEIN SUVH2 and SUVH9 [44].

Besides the production of 24-nt siRNAs by the Pol IV–RDR2–DCL3 complex, there are other non-canonical small RNA pathways that can direct RdDM [42]. These non-canonical RdDM pathways rely on components of the PTGS pathway. PTGS can feed the RdDM pathway either by producing secondary 21-22-nucleotide siRNAs that are processed by DICER-LIKE 4 or DICER-LIKE 2 from double stranded mRNAs, then are directly loaded into AGO6. If Pol-II transcribes any RNA from an inverted repeat, microRNA, TE or any other non-coding RNA that, because of their repetitive nature can fold itself in a hairpin-like structure and form dsRNAs, this dsRNA can be processed by DCL3 and therefore skip the Pol-IV transcription step to feed into the RdDM machinery [45] [46].

It has been proposed that these non-canonical RdDM pathways act redundantly, both acting guided by this production of secondary siRNAs and with the dsRNA templates transcribed by POLYMERASE II (Pol II, Figure 2), in order to efficiently initiate repressive heterochromatic marks on a particular TE, virus or transgene [47]. It has been postulated that RdDM is recruited through these marks to reinforce targeted silencing [42].

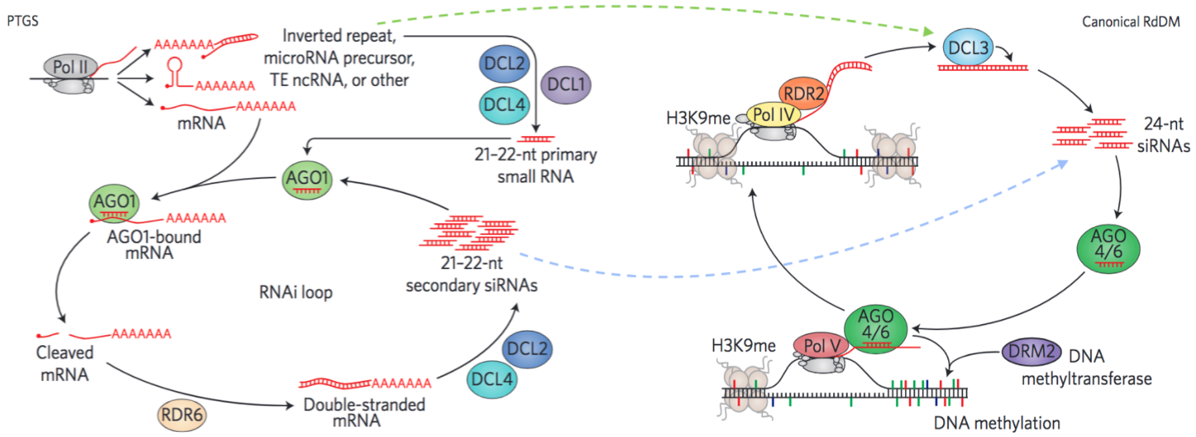


Figure 2 | Right: Plant PTGS (Post-Transcriptional Gene Silencing) begins with the transcription of specific regions, such as TEs or microRNA precursors, by RNA polymerase II (Pol II), resulting in the generation of primary siRNAs. The primary siRNAs, which are produced independently of RNA-dependent RNA polymerase (RDR)

activity, can in turn target Pol II transcripts for cleavage. Cleaved mRNAs can then serve as substrates for the generation of secondary siRNAs through the action of RDR6, Dicer-like 2 (DCL2), and Dicer-like 4 (DCL4), which facilitate the formation of double-stranded RNA (dsRNA). Secondary siRNAs can subsequently target additional mRNA copies, leading to their cleavage and the production of further siRNAs, thus perpetuating the cycle of RNA interference (RNAi). Right, canonical RdDM pathway as in Figure 1. Green and blue arrow are the postulated interactions between pathways. Modified from [42]).

Natural variation of epigenetic regulation and TEs.

Arabidopsis thaliana is a native species of the Eurasian continent, and its distribution range spans Europe to North Africa, Central Asia, and Southeast Asia. Human activity has led to its introduction to other regions, including North America, Japan, and East Australia [48]. The centre of origin of *A. thaliana* remains unknown, but it is speculated to be in Europe/North Africa or central Asia/Caucasus regions [49]. The study of natural genetic variation in this model plant has rapidly developed in the past decades, culminating in the genome sequencing and analysis of 1,135 naturally inbred lines of *A. thaliana* [50].

Many genes underlying phenotypic variation have been identified by genetic mapping, often using methods of Quantitative Trait Locus (QTL) mapping [51]. Although the majority of the phenotypic variation accounted for by all *A. thaliana* QTL described to date is caused by variation at the level of nucleotide sequence, it has been shown that natural epigenetic variation might also be responsible for some heritable trait variation [9] and several epialleles have been described in the literature [10].

In *A. thaliana*, approximately one-third of genes are at least partially methylated; moreover, while DNA methylation in TEs and DNA repeats is very similar in different accessions, 50% of methylated sequences in genes change between different ecotypes [52]. This was corroborated by the considerable amount of natural variation found in the methylomes of accessions from the 1001 Genomes Project, some of which has been linked to differences in expression of nearby genes [53]. They are responsible for the establishment and reversal of TE-like methylation of genes that harbour near TEs. siRNA populations have also been shown to differ among accessions by earlier studies [54] and have also been correlated with TE transposition [55].

1.3. Evolutionary relevance of epigenetic marks.

As we have seen, the epigenetic marks, histone modifications, histone variants, small RNAs and DNA methylation, function in plants as master regulators of DNA accessibility and therefore affect a myriad of activities and events related to DNA. Eukaryotes employ all these distinct epigenetic marks in a similar manner, and most of the key molecules in the epigenetics pathways share close orthologs across kingdoms. For example, as discussed above, DNM1 in mammals and MET1 in plants, are orthologous that participate in DNA methylation. Within the plant kingdom, most of the identified genes responsible for methylation deposition, maintenance, and demethylation evolved early and are shared from chlorophyte (green algae) to angiosperms (flowering plants) [56]. Epigenetic mechanisms are ancient in origin, and they are also found in prokaryotic systems. There is recent evidence of bacterial DNA methylation [57] as well as of histone-like mediated DNA packing [58] and the use of small RNAs to regulate gene expression post-transcriptionally [59].

However, the appearance in eukaryotes of pathways that link sRNAs to DNA methylation and histone modification (RdDM pathway and its non-canonical connections) and thus, RNA-directed silencing mechanisms, meant that *de novo* DNA silencing can be inducible, sequence-specific and heritable [60].

Epigenetics and genome size.

When comparing genomes of prokaryotes with those of eukaryotes, the most striking difference is the larger genome size of eukaryotes [61]. Within eukaryotes, variation in genome size is not explained in terms of gene content, but by TE content [62]. Prokaryotic genomes also harbour TEs, but why have TEs been so much more successful in parasitizing eukaryotic genomes?

There are two alternative hypotheses to explain the differences in TE success between prokaryotes and eukaryotes. Fedoroff [60] proposes that the innovations of the eukaryotic epigenetic machinery, which are mainly focused on TE repression and the suppression of homologous recombination, allow TEs and the proteins they encode to proliferate within eukaryotic genomes, fueling in turn their capacity to evolve. In her words: "*I argue that transposable elements accumulate in eukaryotic genomes because of, not despite, epigenetic silencing mechanisms that control homology-dependent recombination.*" Her hypothesis is based on several observations:

- Eukaryotes do seem to have several redundancies in their epigenetic machinery to ensure their robustness.

- Eukaryotes make use of sRNAs to interconnect their silencing pathways, which increases the robustness of the epigenetic machinery.

- Prokaryotes engage in localised genome duplication through homology, but such duplications are often quickly lost.

- Eukaryotes are able to retain duplicated sequences, and this may be a critical step in the evolution of multicellularity, as higher retention of duplicates offers more opportunities for such duplicates to diversify before being lost and thus acquire a new functionality niche.

A second hypothesis was laid down by Lynch and Conery [63]. Their hypothesis extends beyond the accumulation of TEs in eukaryotes, attempting to explain many features that distinguish the prokaryotic and the eukaryotic genome, such as the emergence of intronic sequences and the pervasiveness of gene duplications in eukaryotes. They argue that these differentiating genomic features arose as a consequence of dramatic changes in population size. In their words: *“We argue that many of these modifications emerged passively in response to the long-term population-size reductions that accompanied increases in organism size.”* [63].

The process that generated these genomic complexities in the first place is non-adaptive in nature, and a consequence of neutral evolution. However, this situation would allow the later development of mechanisms to cope with such complexities, such as epigenetic silencing, and this will be subject of adaptive evolution.

These two hypotheses swap the order of factors that gave rise to the increase of size of the eukaryotic genome. In Fedoroff’s view, the eukaryotic refinement of the epigenetic machinery is what allows eukaryotes to harbour more TEs, in turn allowing the eukaryotic genome to “evoke rapid genome restructuring, which is at the heart of eukaryotic evolvability” [60]. Lynch and Conerym [63], however, argue that it was the reductions of population size that provoked these increases in genome size and complexity due to TE overpopulation, and as a consequence, only eukaryotic lineages that evolve ways to cope with these constraints such as the refinement of the epigenetic machinery, were successful.

Underlying both hypotheses, however, is the strong impact of TEs on shaping the entire makeup of eukaryotic genomes. Thus, one important key for understanding the complex genomic ecosystem, which is the physical basis of phenotypic traits, is understanding the nature of TEs and how they have impacted eukaryotic genomes.

2. Transposable elements (TEs).

Transposable Elements (TEs), also known as a class of mobile genetic elements (MGEs) in prokaryotes, are discrete segments of DNA that populate the genomes of most organisms. Their main characteristic is the ability to propagate through the genome, increasing in number, due to the enzymatic proteins they encode [64]. TEs vary enormously in length, from less than one hundred base pairs to over dozens of kilobase pairs.

TEs were first discovered by Barbara McClintock [65] through a series of observations and experiments in maize during the mid-20th century. McClintock described them as agents that are able to modify gene expression, change their location due to stress, and provoke chromosomal double-strand breaks in the genome. The revolution of sequencing technologies accelerated their discovery, and their presence has been generalised to most, if not all, current organisms. Initially viewed as selfish genomic parasites [66] because of the mutagenic nature of transposition, TEs have been progressively considered an important asset of genomes, providing raw material for evolution [67].

A main characteristic of TEs is that they are maintained by vertical transmissions as copies integrated in the chromosomes of their hosts. This is a key characteristic, as it distinguishes them from viruses, phages and integrative conjugative elements which, although have similar features to TEs, are not considered TEs since they are able to move between hosts independently [68].

TEs exhibit a varying status within a host's genome, influenced by their age and activity levels. Autonomous TEs encode enzymes enabling their mobility, while non-autonomous TEs rely on enzymes from related autonomous elements [64]. These enzymes are quite specific to the encoding TEs, and because of it, have been subject to repeated domestication events by their hosts [69], TEs accumulate mutations over time, often rendering them inactive for further mobilization. Most genomes contain mainly inactive TEs and only a few active TEs. An important part of intergenic sequences is composed of fossilised TE sequences, which are hard to identify with common annotation approaches. For example, in the *A. thaliana* genome, an analysis focused on ancient remains identified over a third more of TEs than previously annotated [70]. These old TE remnants were probably generated early in the evolution of the family to which *A. thaliana* belongs, the Brassicaceae, more than 40 Myr ago [30], and are still detectable in the current *A. thaliana* genome, albeit in a very degenerate state. Another study claimed to find evidence of fossilised TE proteins in vertebrate genomes that originated more than 250 million years ago [71] and are still present in current genomes as relics.

2.1 Classification and brief evolutionary history of eukaryotic TEs.

A systematic classification of any biological entity helps to catalogue the desired elements by any chosen properties. Classifications are useful because they allow to quickly integrate new knowledge into described systems, provide a common framework of communications between researchers and help find connections between elements of groups in the literature. For TEs, efforts to achieve a systematic classification have been made since 1989 [72]. Nowadays the most commonly used classification system for eukaryotic TEs is the one proposed by Wicker in 2007 [64]. Wicker proposal is an update of that first proposed by Finnegan, and maintains its fundamental scheme. It applies mechanistic and enzymatic criteria to divide TEs in a hierarchical order of the following levels: Class, Subclass, Order, Superfamily and Family.

TEs are classified into Class I and Class II, based on the presence or absence of an RNA molecule as intermediate. Class I includes elements that transpose through an RNA intermediate, further divided into five orders based on mechanistic features and organization. Class II TEs transpose without an RNA intermediate and are split into subclass 1, characterised by Terminal Inverted Repeats (TIRs) and specific recombinases, and subclass 2, including Helitrons and Polintons/Mavericks, with unclear mobility mechanisms, but distinct of those from subclass 1.

Superfamilies, the next classification level, group TEs based on shared replication strategies and large-scale features. They differ in target site duplication (TSD) and element size, with minimal DNA sequence conservation and limited protein-level similarities. Superfamilies are subdivided into families, identified by DNA sequence conservation. Wicker's proposal sets 80% similarity in 80% of aligned sequences as the threshold for family classification, ensuring consistency across different tools used for annotation. To avoid misclassification due to short homologous segments, a minimum cutoff of 80 base pairs is applied, known as the 80/80/80 rule in TE classification. A schematic classification is shown in Figure 3.

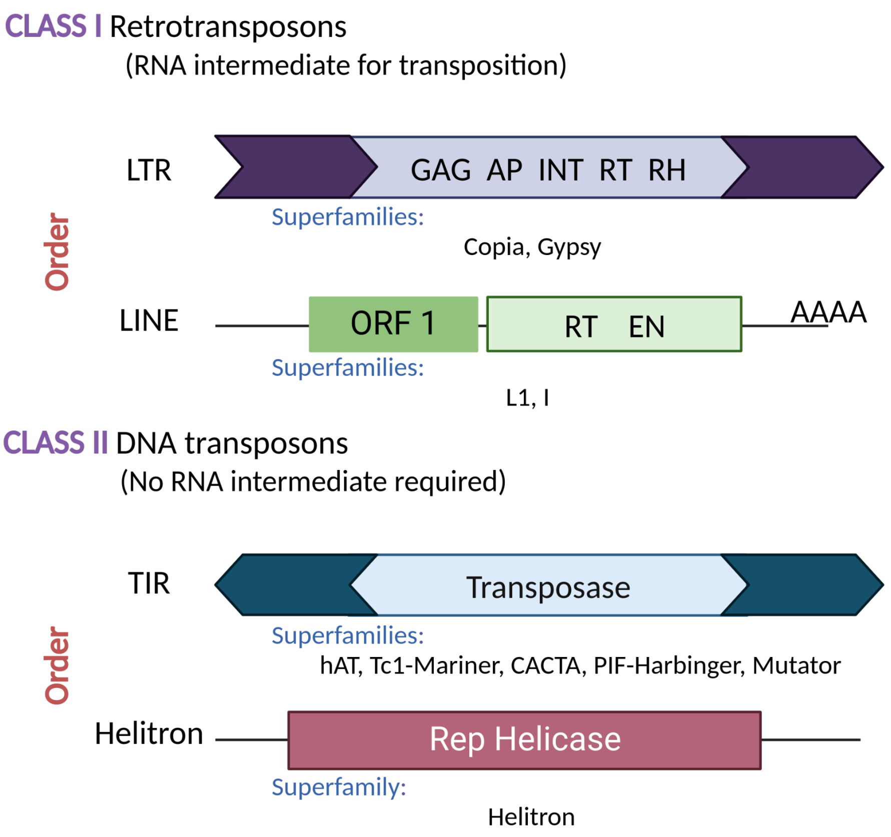


Figure 3 | Schematic representation of a subset of transposable elements (TEs) classified as proposed by Wicker. The subset shown here includes the different TE superfamilies found so far in plants. A complete classification can be found in [64]. LTR: long terminal repeat; LINE: long interspersed nuclear element; TIR: terminal inverted repeat. Dark lines show structural repeats of both LTRs and TIRs. Protein coding domains: GAG, Capsid protein; AP, Aspartic proteinase; INT, Integrase; RT, Reverse transcriptase; RH, RNase H; EN, endonuclease; AAAA, poly(A) tail; RP, Replication protein; HEL, Helicase. ORF 1, LINE opening reading frame.

Although the Wicker classification and nomenclature system aided researchers by simplifying annotation, and allowing structural, functional and evolutionary TE analyses to act within a common framework, several authors have highlighted weaknesses of this classification system, as reviewed by [68].

Two weaknesses of the Wicker proposal are of main relevance. First, it ignores the connection between prokaryotic TEs and eukaryotic TEs, which is important because most of the understanding of the enzymatic machinery encoded by the TEs has been developed for prokaryotic TEs (mainly for the DDE transposases of TIR elements) [73]. Second, it ignores the phylogenetic relationships between groups [74].

Indeed, although all TEs are grouped together as mobile elements, TEs are polyphyletic in origin, and their evolution shares similarities with viruses, to the point that some TEs have direct evolutionary connections to specific virus clades. For example, phylogenetic trees based on the conserved domains of the transcriptase, integrase and RNAse H proteins cluster together with Ty3/Gypsy elements and retroviruses [75]. Some authors have even hypothesised that retroviruses originated from Gypsy retrotransposons by the acquisition of an envelope gene, allowing them to move extracellularly [76].

It has been asserted that "the wide diversity of retrotransposons compared to the limited diversity of vertebrate retroviruses suggests the ancestral forms were retrotransposons." [77], and later phylogenetic work [78] led to the proposal that Ty1/Copia and the Ty3/Gypsy families represent the oldest groups among examined elements, with Bel/Pao, Retroviridae and Caulimoviridae viral families having emerged from Ty3/Gypsy radiations at different evolutionary times. In the same line, recent insights from the analysis of *Actiniaria* (sea anemone) species suggest a close evolutionary relationship between retrotransposons and retroviruses [79].

Evolutionary ties with viruses are not limited to retroelements. Since they were first discovered, Maverick/Polintons were thought to be related to adenoviruses [80,81]. Some authors refer to them as Polintonviruses [82] because they encode major and minor icosahedral virus capsid proteins. These authors also propose that, contrary to the retrotransposon case, these Class II TEs originated from an integration of a bacteriophage that entered the proto-eukaryote via the mitochondria [83], with a posterior integration to the nucleus. More recently, these Maverick/Polinton elements have been the major actors in a case of horizontal gene transfer between distant nematodes [84]. To this date, Maverick/Polinton TEs have been described in metazoan and fungi, but not in plants.

Connections between Class II TEs and viruses are also found in plants. Soon after the discovery of helitrons [85], due to the resemblance of their encoded protein domains (Rep domain) and their proposed rolling circle mechanism of transposition, some authors pointed out helitrons as the missing link between Geminiviruses and prokaryotic replicons [86,87]. Some authors, however, dispute this assertion [88].

The constant uncovery of ties between eukaryotic TEs, prokaryotic MGEs and viruses points to the often shared, mixed and convoluted evolutionary past among these genetic elements. Hence, there is a need for an integrated classification system for all mobile DNA elements that reflects their commonalities in mechanistic terms, with an emphasis on their phylogenetic relationships as already proposed by several authors [68,73].

2.2 Transposition mechanisms and structural motifs of TEs.

As we have seen, the unifying feature of all TEs is their ability to encode the enzymatic machinery necessary for their movement alongside the genome. This set of enzymes includes the reverse transcriptase (used by retrotransposons); DDE/D transposase (this enzyme is at the core of all plant DNA TEs); Tyrosine-recombinase (for Cryptons, eukaryotic TEs not present in plants) and Rep/Helicase (a HUH nuclease fused to a helicase utilized by Helitrons).

In the following sections, I will provide an overview of the enzymatic machinery and known mechanisms of transposition involving plant TEs. Note that other ways of transposition have been described for prokaryotic MGEs and eukaryotic TEs. Because their replication mechanisms are tightly tied to their structural conformation, I will also provide an overview of those features.

DNA TEs and the DDE transposase.

DNA transposases are enzymes that move discrete segments of DNA from one location in the genome to a different one without the involvement of a RNA intermediate. There are four major types of DNA transposases depending on their catalytic folds: RNAse-H-like or DDE/H; HUH; Tyrosine recombinase and Serine recombinase [89]. Notably, eukaryotic TIR DNA TEs harbour a DDE type of DNA transposase, which does not have homology requirements for the integration site [90], whereas the others are present in bacterial or viral systems. I will therefore focus on the DNA transposases with a catalytic DDE structure [91]. The invariant DDE residues are located in three separate regions in the catalytic domain surrounded by less conserved residues, but it was shown through mutagenesis studies that it is this triad that confers the essential role of transposition [90]. Interestingly, the position of the invariant DDE residues is remarkably similar in both transposases and retroviral (and retrotransposon) integrases [90].

DDE DNA TEs, also called TIR TEs, are the most diverse and widespread group of DNA TEs in eukaryotic genomes [92]. TIR TEs, are simple genetic elements that consist of a single opening reading frame that contains the transposase gene, containing the catalytic DDE triad, and two terminal inverted repeats (TIR) in the ends of the TEs that act as the transposase binding sites [64]. Miniature inverted-repeat transposable elements (MITEs) make up a heterogeneous group of non-autonomous TEs that consist of only a few hundred base pairs in size. Because in several species it has been demonstrated that enzymes codified by TIR elements are responsible for the replication of these elements [93,94], they tend to be lumped together with TIR TEs.

In plants, four superfamilies of DNA TEs have been reported to act through DDE transposases. The EnSpm/CACTA superfamily, sharing some residues between the second conserved D and E, and the superfamilies Tc1/Mariner, MuDR and hAT, these last two superfamilies sharing the motif C/DxxH between the second conserved D and E [95].

Biochemically, the DDE active site in transposases facilitates two types of reactions: a hydrolysis attack that results in the cleavage of the DNA bond, and a transesterification that allows the enzyme to simultaneously cleave one DNA strand and the bonding to another DNA strand [96]. The second reaction is utilized to create a hairpin structure by joining complementary DNA strands, linking two TE ends to form a circular intermediate, or to integrate a TE into a new site. Transposases employ these reactions in various combinations and on different DNA strands, leading to diverse pathways for transposition [91]. For plant DNA transposases, the transesterification step leaves a staggered strand, generally of 2-9 nucleotides (which varies for each superfamily of DNA TEs). Repair of these strands leaves a target site duplication, a hallmark of DDE transposition [90].

During transposition, the DNA transposase must bind specifically to the element ends and form a nucleoprotein complex, the "transposome", with the necessary configuration to catalyse the DNA break and join reactions [97]. Once a DNA TE has been successfully mobilised, there are two loci at the genome that have to be repaired, the empty donor site from which the TE has been excised, and any nick or gap resulted from the integration of the TE. In eukaryotes, these donor sites are repaired through homologous recombination [91] (Figure 4).

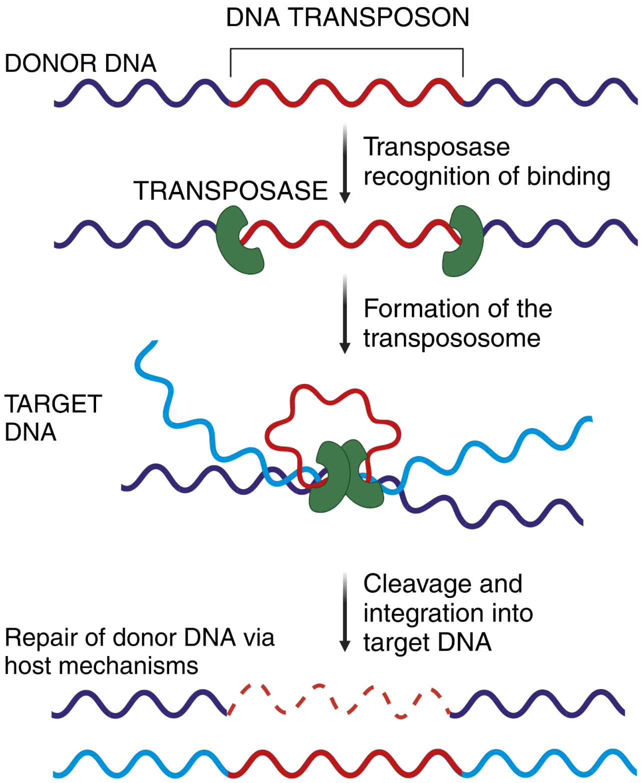

DNA TRANSPOSON
DONOR DNA
Transposase recognition of binding
TRANSPOSASE
Formation of the transpososome
TARGET DNA
Cleavage and integration into target DNA
Repair of donor DNA via host mechanisms

Figure 4 | "Cut-and-paste" model of transposition. The process begins with the specific recognition and binding of the DDE transposase to the TIR ends of the DNA

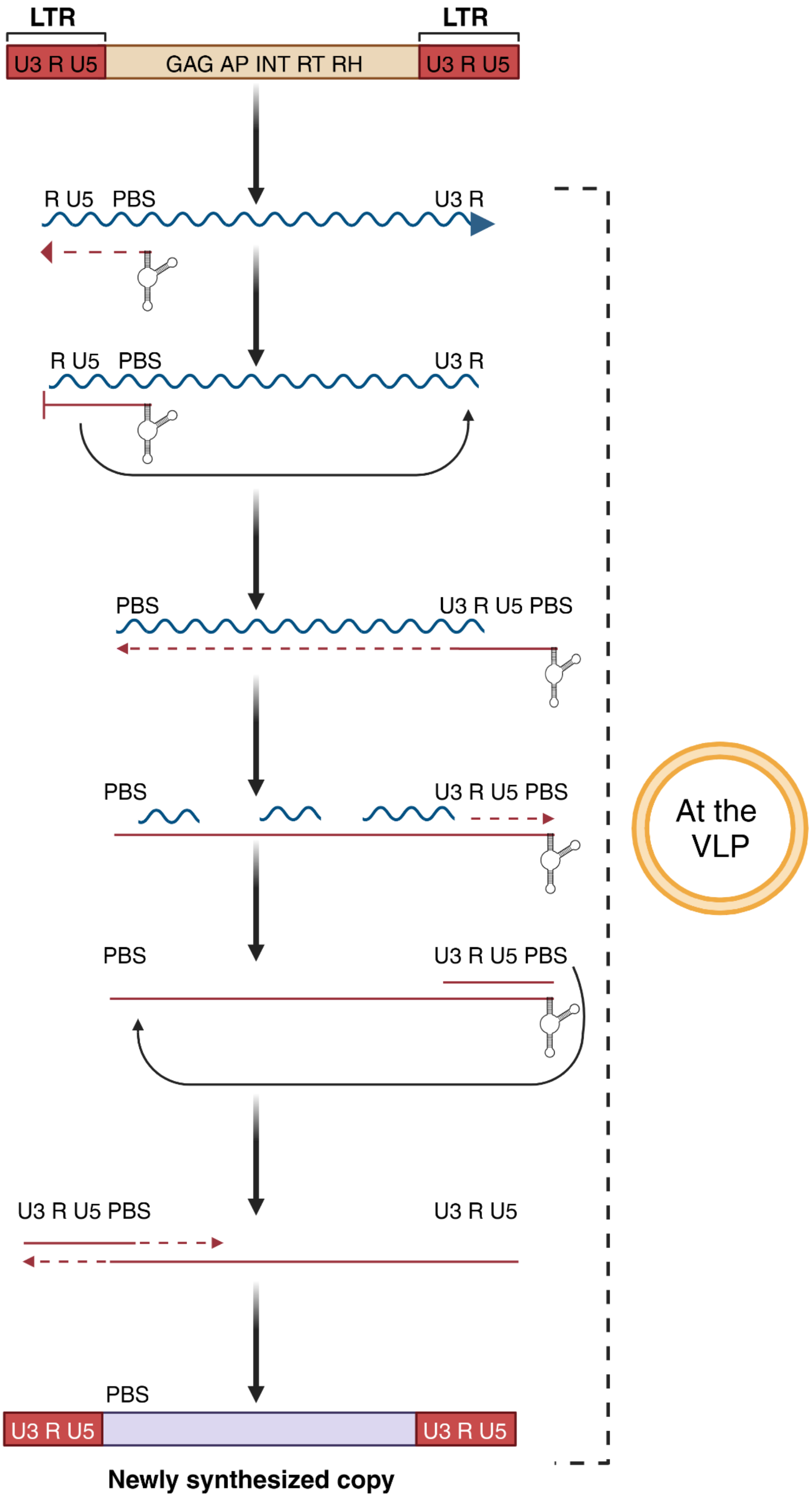
TE. Then, the dimerization of the transposase results in the formation of the transpososome, bringing together the TE ends at a target site. Simultaneous cleavage occurs at both the donor and target sites, followed by the integration of the excised TE into the target site and removal of the empty donor site, that will need to be restored by host repair mechanisms [98].

LTR replication cycle.

Long terminal repeat (LTR) retrotransposons transpose via an mRNA intermediate. For their replication, their sequence is transcribed into an mRNA from a genomic locus; this mRNA serves both as a template for replication as well as a production of a self-encoded retrotranscriptase and posterior integration in the genome. LTR retrotransposons are major contributors to genome size in plants [99], but they are also present in the rest of eukaryotes. LTR elements vary widely in size, ranging from a few hundred basepairs, like Terminal repeat retrotransposon in miniature (TriM) [100], up to 25 kilobase pairs (kb) long. The LTRs flanking these elements also exhibit considerable variation, ranging from a few hundred basepairs to over 5 kb in length. Upon integration, LTR retrotransposons generate a Target Site Duplication (TSD), typically of 4 to 6 bp. Autonomous LTR retrotransposons contain Open Reading Frames (ORFs) for GAG, a structural protein for virus-like particles, and for POL polymerase. The POL ORF encodes various enzymatic domains: aspartic proteinase (AP), reverse transcriptase (RT), RNase H (RH), and DDE integrase (INT). LTR retrotransposons also possess specific signals for packaging, dimerization, reverse transcription, and integration of the mRNA and cDNA intermediates. In the two primary superfamilies Ty3/Gypsy and Ty1/Copia, the arrangement of RT and INT in the POL region differs.

This mode of replication of LTR-retrotransposons (LTR-TEs) [64] is similar to that of retroviruses, as evident by the common structural arrangements and enzymatic capabilities. In fact, much of what we know about LTR retrotransposons is based on detailed genetic studies of the Ty3 elements on *Saccharomyces cerevisiae* and those of HIV [101]. The replication cycle starts when POL II begins to transcribe from the 5’ LTR, which contains a POL II promoter, and terminates within the 3’ LTR before its end, forming a pool of transcripts. Some of them will be translated to proteins and others serve as a template for reverse transcription [102]. Due to the position of the POL II promoters and terminators, RNA transcripts lack complete LTRs that must be restored to produce a functional cDNA copy of the LTR.

This is accomplished by a complex multistep process that takes part in the cytoplasm, where some of the transcripts are processed by the ribosome to produce both the GAG capsid protein and the polyprotein. The polyprotein contains the key reverse transcriptase (RT), an integrase (INT), and RNAseH (RH) and an aspartic protease (AP). AP cleavages this polyprotein into endoproteolytic fragments. Some of these products are then encapsulated together with retrotransposon RNAs forming a virus-like particle (VLP). This is done by Gag, which binds two plus stranded RNA molecules together with the RT-RNAseH and the IN. The process is detailed in Figure 5.



Newly synthesized copy

Figure 5 | In the virus-like particle (VLP), reverse transcription takes place. It starts at one of the plus-stranded RNA molecules loaded by GAG, at a region proximal to the 5' end denominated PBS, primer-binding site. From there, the RT initiates the minus strand DNA synthesis using as a template a host-derived tRNA hybridized to the PBS and continues until the 5'-end of the RNA molecule. Here, the RT-associated RNase H hydrolyzes the 5'-terminal repeat (R) and a unique 5' sequence, U5, this enables the transfer of the nascent minus single-stranded DNA segment to the 3'-terminal segment of the sister plus-stranded RNA molecule that was captured by GAG, which contains the same R repeat and a unique 3' sequence (U3). After this template switching, made possible due to the hybridization of the common R domain, the RT-RNase H complex continues translating, to a minus-strand DNA molecule, and digesting this RNA molecule until it reaches the 5'-terminal PBS site. However, shortly after it started, 5' upstream of the U3 site, there is an indigestible region of the RNA molecule, the polypurine tract PPT, fragments of which are used as primers to initiate the synthesis of the complementary plus stranded DNA, forming a complete copy of the fragment up to the 5' PBS region. Both PBS domains hybridise, allowing the continuation of the synthesis of the cDNA in both directions, as each strand uses the other as template. These two strand jumps homogenise the 5' and 3' LTR segment of the molecule, making them identical [103].

Once the VLP has completed the transcription of the two molecules of single stranded RNA to a double stranded cDNA, the newly created retrotransposon must integrate back to the genome. For that, it traverses the nucleus membrane via a signal which is presumed to be similar to that found in retroviral proteins, but it has an unclear nature in plant retrotransposons. After entering the nucleus, the integration on the genome is mediated by the integrase INT enzyme, which catalyses the hydrolysis of the target sites producing sticky ends, despite the blunt-end nature of the LTR molecule, this sticky ends will be then repaired via host processes and will be the source of the target site duplication (TSD) motif observed in the host genome after successful transposition.

LTR encapsulation and packaging has been extensively investigated in the HIV virus [104] but has also been documented for some cases in plants [101], most notably for barley *BARE1* elements [105]. As for the reverse transcription and LTR extension, it has been well studied in Ty3 yeast elements [106]. However, due to structural conservation of plant LTR motifs (PBS, PPT, U3,U5 and R) and some evidence of the cDNA complex formation, it is widely assumed to be very similar. It is notable that plant LTR promoters are activated by a wide array of biotic and abiotic stresses, including hormone treatments and tissue culture. [107,108]. Another thing to note is that the life cycle of the LTR replication is pseudodiploid, as it uses two LTR-derived, single stranded RNAs that generates a single double stranded DNA molecule. This fact has prompted researchers to argue about the possible benefits of recombination of the LTR retrotransposons, which, coupled with single site mutations, increases the retrotransposon evolvability increasing its rate of adaptation [109]. Evidence for such recombination events taking place extrachromosomically in VLPs was first noted for HIV [110] and later observed in *A. thaliana* with the ONSEN LTR retrotransposon [111].

Non-LTR retroelement replication cycle.

Much like LTR elements, non-LTR retroelements' key replication component is an RNA-based intermediate, but obviously they lack Long Terminal Repeats (LTRs). Non-LTR elements can extend for several kilobases, and are present in all eukaryotic systems. They are categorised into five major LINE (long interspersed elements) superfamilies: R2, L1, RTE, I, and Jockey. In plant genomes, only two superfamilies have been described, most of them from the L1superfamily and few from the RTE superfamily [112]. Contrary to LTRs, non-LTRs are predominant over LTR retrotransposons in many animals. Autonomous non-LTRs contain a Reverse Transcriptase (RT) and an endonuclease (EN) for transposition. Many LINEs in plants contain an opening reading frame of unknown function. Although non-LTRs typically create Target Site Duplications (TSDs) upon insertion, truncated 5' ends make them challenging to detect. These truncations likely result from the premature termination of reverse transcription. At their 3' end, LINEs may exhibit a poly(A) tail, a tandem repeat, or simply an A-rich region [64].

Together with LINEs, the second major non-LTR element in plants are short interspersed elements (SINEs). SINEs form a diverse and polyphyletic group of non-autonomous elements. These elements can be mobilised and propagated by enzymes associated with LINEs. SINEs consist of various transcripts, including tRNA, rRNA, and other polymerase III transcripts, ranging from ~70 to ~700 bp in size [113]. SINEs are widespread in plants and are generally present in the hundreds to low thousands. The replication of the tRNA-related SINE S1 in *A. thaliana* appears to be very similar to the mammalian SINE Alu [114].

Although earlier studies had already reported that non-LTR retrotransposons must undergo retrotransposition because intron sequences placed at the element were removed in posterior insertions [115]. Non-LTR retroelement replication was first described in insects using as model a non-LTR retrotransposon, R2, that shows an specific insertion pattern at 28S rRNA genes [116] through a process denominated target-primed reverse transcription (TPRT). Luan and colleagues showed that the R2 element protein, a reverse transcriptase, associates with the RNA transcript near the 3' end at the target site, nicks the target site and uses this exposed end to prime the reverse transcription on the insertion target. After reverse transcription, R2 protein cleaves the opposite DNA strand but this does not prime a second-strand synthesis. Instead, they hypothesise that second strand synthesis may occur via host-mediated DNA repair mechanisms (Figure 6).

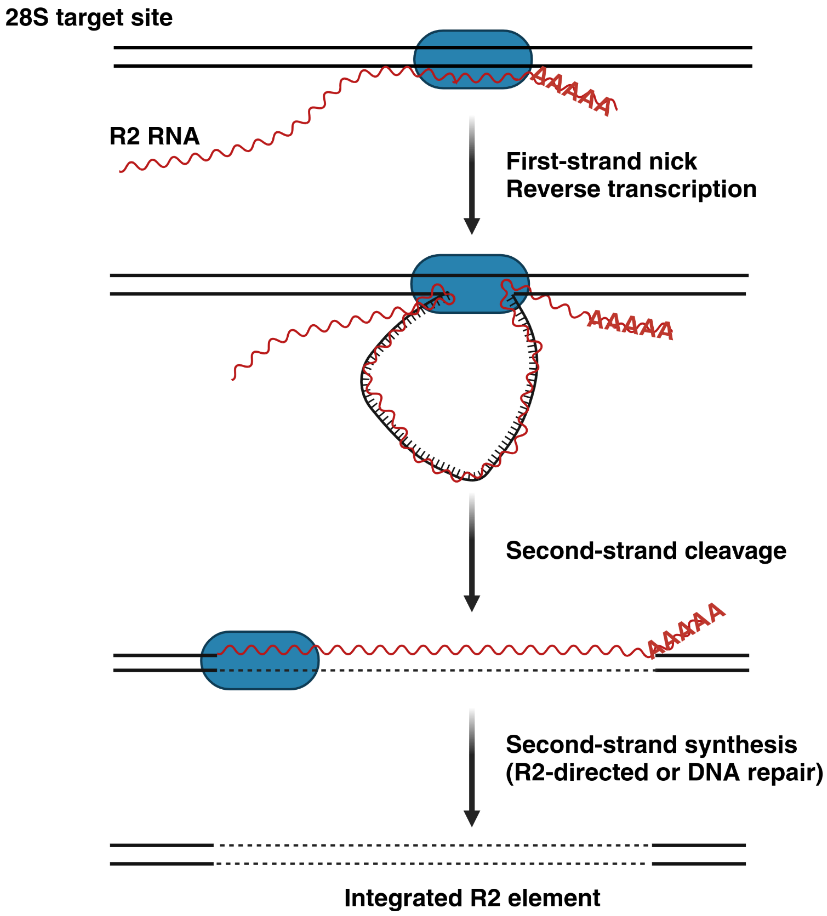


Figure 6 | Model of non-LTR retrotransposition as depicted in [116]. Solid lines, genomic DNA strand; spiked line, newly transcribed cDNA; wavy line, retrotransposon RNA template; dotted line, newly created cDNA. The R2 protein is found near the 3' end of the R2 transcript, including part of the poly(A) tail. Then, it creates a nick in the lower strand of the 28S gene target site, using the exposed 3' end for reverse transcription. The protein stays bound to the 28S gene, mostly upstream of the target site. After reverse transcription, cleavage of the upper DNA strand occurs, but in the described reaction, it doesn't prime second-strand synthesis. This synthesis might be facilitated by cellular enzymatic activities or DNA repair mechanisms, because the R2 protein does not encode that function.

Although first to be described, TPRT is the last step of the full replicative cycle of the non-LTR retroelement. First, the element has to be transcribed; many functional non-LTR TEs contain 5'UTRs with promoter activity, but the nature of it greatly varies across species [117,118]. These promoters kickstart the transcription of the element by RNA polymerase II. After transcription, the TE mRNA is exported to the cytoplasm, where the encoding ORFs are translated and assembled into a ribonucleoprotein particle (RNP) complex. In mammals, this complex has been associated with stress granules and P bodies in the cytoplasm [119]. As reviewed in [120], it is still very unclear what is the role of the RNP complex and the steps that lead to the transportation of this complex to the nucleus to undergo TPRT. Once in the nucleus, TPRT takes place as mentioned before, however, due to the eclectic nature of the different non-LTR elements, variations to the TPRT pathway have been described, especially after the second strand cleavage occurs. It has been suggested that the acquisition of different ORFs by lineage specific non-LTR elements may aid in this step. For example, in the *Drosophila suboscura* LINE-like *bilbo* element, [121], an ORF that codes for an RNAse H domain is thought to be responsible to eliminate the mRNA leftovers after the first cDNA synthesis occurs. Elements that lack RNase H-encoding ORFs may take advantage of the second strand synthesis to displace the mRNA [122]. Interestingly, as mentioned in [120], the final product of the replication cycle of a non-LTR element can produce a recognizable TSD if the second strand nick is produced (as it does in the case of the R2 element, but not always) due to gap filling by presumably host processes of DNA repair. Alternatively, in the case of an upstream second strand nick, non-homologous DNA flaps are deleted, again presumably by host processes, and no TSD is produced. For human non-LTR elements, it has been shown that LINE-1 requires host DNA repair mechanisms for successful integration [123].

Because no described non-LTR element has been shown to encode all the protein domains necessary for all the enzymatic activities required to complete the replication cycle, it is reasonable to assume that all non-LTR elements rely to some degree on host factors to complete their replication cycle.

Helitrons: Rolling circle replication.

Helitrons were the first major class of TEs discovered using bioinformatic approaches [85]. Helitrons are a subclass of DNA TEs that contain a single, homonim, superfamily. Helitron structural signature is very faint. It consists of a 5' TC and a 3' CTRR (R stands for purine, either a guanine or adenine) end. 15-20 base pairs upstream of the 3' end, helitrons contain a GC rich hairpin at insert at between AT dinucleotides [85]. Since their discovery, Y2-type tyrosine recombinase (also known as HUH nuclease) was found in recently active helitrons that was described in bacterial IS91 rolling-circle TEs, together with a helicase domain and, in plants, it also contains a ssDNA binding replication protein A (RPA) domain [85]. Because of the protein motifs they encode, helitrons were thought to replicate via a similar rolling circle replication (RCR) mechanism of that of the bacterial IS91 [124]. More recently, a computational reconstruction of Helibat1, a helitron from *Myotis lucifugus* (brown bats), to an autonomous form (Helraiser), followed by a *in vitro* and *in vivo* characterization of their mobility [125] provided better insights into helitron replication (summarised in Figure 7).

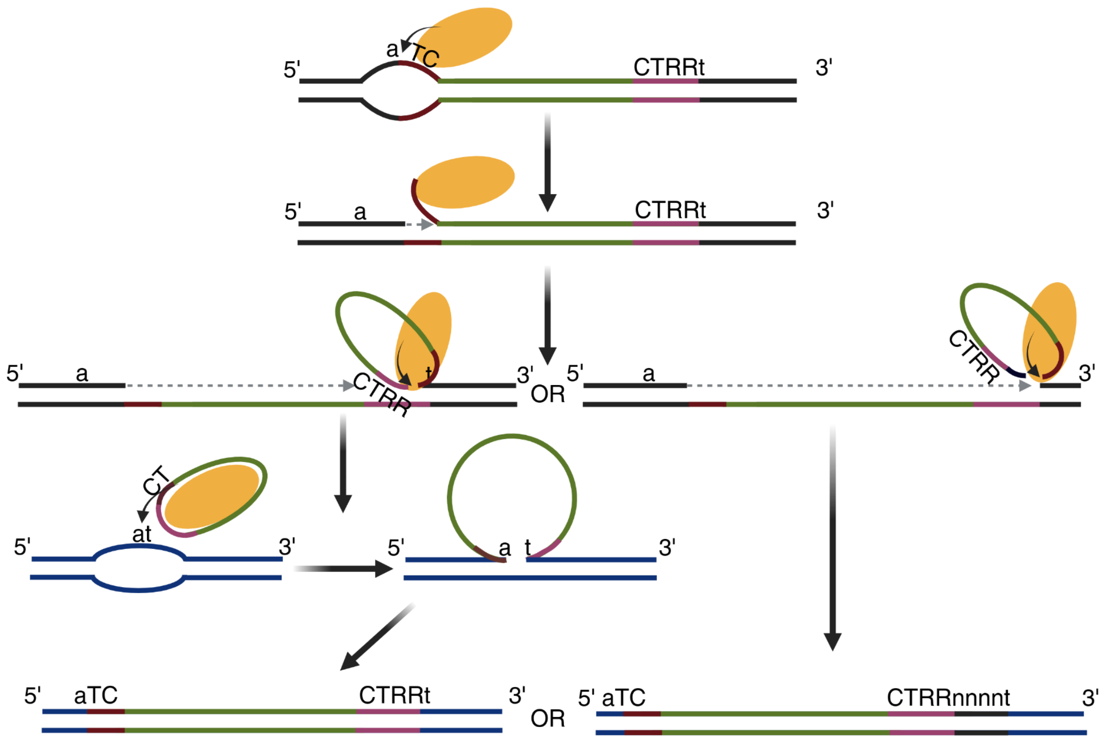


Figure 7 | Rolling circle mechanisms of transposition. As described by [125]. A tyrosine of the helitron transposase, depicted as a yellow oval, interacts with the 5' TC motif (in red) and nicks the single-strand of DNA (ssDNA) donor site, creating a 5'-phosphotyrosine intermediate between the transposase and the end of the transposon. This process kickstarts the DNA repair of the donor site, while the helicase domain unwinds the double-stranded DNA and displaces the TE DNA during replacement-strand synthesis. The hairpin structure at the 3' end (pink) causes the helicase to pause, facilitating the recognition and cleavage of the CTRR 3' tetrad by the second tyrosine of the HUH domain of the transposase. This action generates a free 3'-OH group that interacts with the 5'-phosphotyrosine intermediate and generates a free ssDNA circle. Alternatively, the transposase can read through the hairpin structure, skipping the CTRR and mobilising host DNA flanking the helitron, generating a new 3' end. The integration of the generated ssDNA circle (either with or without host DNA) is performed when one of the tyrosines interacts with the dinucleotide AT in the target DNA and the other tyrosine with the ssDNA circle. The single-stranded TE DNA, covalently bound to the target, is passively replicated and converted into its double-stranded form during the DNA synthesis phase of the cell cycle.

This RCR model explains why there are no TSDs during helitron integration. The *in vitro* experiments performed by Grabundzija and colleagues [125] revealed a 'read-through' mechanism that captures DNA sequences adjacent to the TE. In cases where the hairpin structure is absent, heavily mutated, or goes unnoticed by the transposition machinery, the transposase skips over the 3' end of the TE and locates an alternative transposition terminator sequence further downstream. This process leads to the incorporation of the flanking host DNA. Although Helraiser is a vertebrate helitron, there is evidence in plants of helitrons capturing host gene fragments [126], in line with these molecular insights.

Site selection preferences of TE insertion.

Understanding the replication mechanisms of TEs is useful for understanding the dynamics that have been shaping eukaryotic genomes. Similarly, understanding the molecular basis of the insertion preferences of TEs is important to predict how TEs are going to keep shaping the eukaryotic genomes.

One naïve approach to do so is to simply collect extant genomes and analyse TE variation patterns in those genomes. However, given that most of the insertions of TEs are rare and deleterious in nature, these patterns will be affected by selective pressures over evolutionary times, resulting in patterns that reflect not only the action of purifying selection, but also the drift acting among those populations [127]. A second obscuring factor is the host epigenetic machinery, which may alter the available insertion sites across the genome for a given TE [128], albeit one could argue that this particular host strategy minimises the damage TE may cause is part of the natural dynamics of site selection preferences of TEs.

A strategy to reduce these biases is to take into account the age of the insertion, if known, under the assumption that *de novo* insertions will be less affected by purifying selection than fixed ones. An age estimation of the insertion itself becomes important because relying on site frequency spectrum alone can be misleading, since both new insertions and purifying selection have the same pattern of low frequencies in population estimations [127]. Despite these limitations, a population-scale analysis of TE insertions in *A. thaliana* revealed that recent insertions, i.e. insertions present in a single accession, were evenly distributed along chromosomes [129], despite the marked pericentromeric-rich TE pattern that complete genomes of this species present.

The best approach is to directly measure insertions that occur between parents and progeny. This has been done in *Saccharomyces cerevisiae* for the retrotransposon Ty1/Copia superfamily [130], revealing a preference for nucleosome-bound DNA, but avoiding coding sequences. The inclusion of histone-related mutants in this study also indicated that the epigenetic state of the host influences the target sites. In the same line, a recent study of Ty1/Copia TEs in *A. thaliana* found that the histone variant H2A.Z modulates the insertion preference of this superfamily [128], highlighting again the important role of the host to regulate insertion-site preferences of TEs. Another study in *S. cerevisiae* but for the Ty3/Gypsy superfamily found that this superfamily had insertion site preferences for RNAP III-transcribed genes including tDNAs and rDNA loci [131]. This is interesting because the same superfamily in *A. thaliana* shows instead accumulation in the pericentromeric regions [30]. Indeed, all these studies support the tethering model of transposition first proposed in 1990 [132]. Drawing data from *Drosophila melanogaster* and *S. cerevisiae,* the tethering model proposes that host chromosomal DNA, proteins, and nuclear functions are part of the targeting process of retrotransposons and retroviruses.

This tethering model seems to not be applicable to DNA TEs. Mutagenesis analysis of *de novo* insertions in *Mus musculus* (mouse) embryonic cells using the Tc1/Mariner *Sleeping Beauty* TE have shown fewer biases than retrotransposons [133]. However, mutagenesis analysis in engineered *A. thaliana* lines with the Ds and dSpm system from *Zea mays* (maize) show that these CACTA elements had preferences near the translation start codon of genes [134].

One of the best studied DNA TEs, the *Drosophila* P element, has been invading natural populations of *D. melanogaster* since the 1950s [135]. *D. melanogaster* strains isolated from natural populations earlier than the decade of the 1950s contain no presence of this P element. However, the P element is found in strains isolated since the 1950s worldwide [135]. This P element is now invading another *Drosophila* species, *D. simulans*, and this process has been caught in the act [136]. Examining the genomic locations of this recent waves of insertion in natural *D. simulanis* strains has revealed that the P element in *D. simulans* is found in similar genomic location as those in *D. melanogaster,* probably a result that in both *Drosophila* species this genomic invasion is very recent. In both species, P elements have preference for the promoter regions of genes and for sequences targeted by the origin recognition complex [136].

In *A. thaliana*, examination of epigenetic recombinant inbred lines (epiRILs) have revealed that, when the host epigenetic machinery is defective, an induced condition that may reflect epigenetic conditions during stress, DNA TE insertions from the VANDAL21 and ATENSPM3 DNA TEs occurred [128]. Both TEs, belonging to the Mutator and CACTA superfamilies, respectively, have preferences nearby or within genes, and are associated with chromatin states. Interestingly, ATENSPM3 *de novo* insertions, derived from two non-autonomous elements, mobilised *in trans* by the single full length ATENSPM3 [128]. T*rans* transposition has also been recorded from the Ping/mPing system in *Oryza sativa* (rice) [137]. In the study, a similar strategy was used where recombinant inbred lines were generated by crossing a bursting accession with the reference accession. Comparative analysis of the progeny led to the recording of around 40 insertions per plant per generation for the non-autonomous MITE Ping (mPing) and very few insertions per plant of the autonomous Ping element. Interestingly, mPing insertions had preference for AT-rich target sites [137].

As we have seen, the insertion site preferences of TEs cannot be understood without integrating the host epigenetic mechanism, the enzymatic properties of a given TE superfamily and the natural history of the host population.

2.3 TEs and genomic innovation.

From the first studies of TEs by Barbara McClintck [65] it became clear that TEs are major agents of genomic plasticity, as it was first observed they could provoke chromosomal breaks and modify gene expression. Many studies since then have accounted for the different ways TE elements can fuel eukaryotic evolution. TEs can alter host genomes in several ways: creating genetic diversity, participating in the making of new genes, altering gene networks, being hotspots of large genomic rearrangements and providing machinist means of rapid evolutionary change [138].

Genetic diversity.

Due to imprecision in the excision of TEs, TE copies may carry host-derived DNA when transposing, including host exons within the vicinity of other host-genes and potentially creating new exon combinations and novel gene functions. In maize, evidence suggests that helitron activity is a major source of this type of gene variation [126], a process known as gene or exon shuffling. This has also been observed for Mutator [139] and CACTA elements [140].

TE insertions within genes often lead to total or partial disruption of gene function. TE insertion within introns or exons alter the conformation of a gene, modifying the behaviour of the host splicing machinery causing novel splice sites, exon skipping, premature stop codons or frameshifts. The effects of TE insertions altering gene function and phenotypes have been extensively documented for animal colouring variations [141], probably because colouring is an easy phenotype to notice and because colouring alterations likely cause neutral phenotypes, but not always. In *Biston betularia* (a moth), a TE insertion within the first intron of a gene lead to the accumulation of melanism in this species [142] that acted as an advantageous camouflage phenotype in the heavily carbon-polluted environment of the industrial revolution.

The interplay between the epigenetic silencing machinery of the host and TE insertions within genes may also increase genetic diversity, as is the case of the Karma epiallele [143]. This epiallele is generated due to the methylation of an intronic retrotransposon insertion; loss of the methylation leads to an alternative splicing and premature termination of the allele, causing a dramatic change in phenotype.

Evolution of new genes.

TE-derived proteins have been repeatedly co-opted, or domesticated, by their host organisms across eukaryotic organisms. For example, in *A. thaliana* the MAIL1 and MAIN genes encode proteins that appear to have evolutionary origins within a subset of Ty3/Gypsy retrotransposons [144]. These proteins have been repurposed by the host as part of a still unknown silencing pathway. Because TE-related proteins have DNA-binding proteins, they have also been reused by the host to bind DNA and participate in transcription modulation. The hAT derived transposase DAYSLEEPER gene in *A. thaliana* [145] or the FHY3 and FAR1 genes, derived from Mutator transposases, also in *A. thaliana [146]*, are examples of this repurposing. Perhaps the most famous example of co-option is the evolution of the antigen receptor gene assembly by V(D)J in jawed vertebrates, where two crucial proteins, RAG1 and RAG2, arose from the DNA TE protoRAG [147].

Another source for the evolution of new genes is the intronless pseudoduplication of genes, leading to the appearance of retrogenes. Retrogenes can arise when the enzymatic machinery of a retrotransposon randomly recognizes the polyadenylated end of host-derived mRNAs, and this has happened repeatedly in both eukaryotic and prokaryotic genomes [148]. These newly arising retrogenes are initially functionless, likely because they are separated from their original, *cis*-regulatory sequences; thus, subsequent evolutionary events will determine if these copies are retained or lost [148]. Some of these events can confer novel traits to the young retrogene so it can become functional, as for example has happened in the angiosperm *Thlaspi arvense,* in which recurrent retrogenization contributed to the expansion of gene families involved in development and stress responses *[149]*. Retrogenization lies also at the foundation of a novel phenolic pathway involving the cytochrome P450 enzyme in Brassicaceae [150].

Modification of transcriptional networks.

Regulatory gene networks coordinate the activity of multiple genes that belong to the same pathways to carry on complex biological functions in the organisms. Due to their mobility, TEs can rewire these networks by shuffling around and multiplying regulatory motifs such as enhancers, repressors, or transcription factor-binding motifs. Because TE activity relies on host machinery to propagate, any successful TE would have evolved *cis*-regulatory sequences that mimic endogenous host promoters. Thus, host organisms can co-opt these elements to rewire their regulatory networks, reaching novel configurations [138]. This theoretical role attributed to TEs has support in studies on maize [151] and wheat [152] that show a strong correlation between the presence of certain TE families upstream of stress-response genes, affecting their expression.

Genome Rearrangements.

TEs can contribute to genomic rearrangements as a by-product of transposition events. However, TEs can also promote structural variation due to recombination events between their homologous copies dispersed throughout the genome. This may result in large-scale inversions, duplications or deletions. Comparative studies of full-length assemblies of plant genomes are revealing the real extent of these large structural variants [153]. Small-scale genomic variants can lead to the transposition (and duplication) of host-derived sequences caught between the borders of the variants [154]. Derepression of TEs releases endonuclease activity of the TE-encoded machinery, leading to genomic instability by the increase of double strand breaks, as documented in mammals [155], and for the As/Ds system in *Z. mays* [156]. These authors traced the output of different transposition events in a mutant screening. They observed a variety of complex rearrangements in the *p1* gene locus, consequence of different transposition events among different termini of Ac elements lead to [156].

To summarise, TEs are an important source of genome variation, which is the main component of evolution, both neutral and adaptive. Adaptive evolution also requires “free” or unconstrained sequence space, and TEs are a great mechanism to duplicate and expand the genomic space. The true extent of TEs’ contribution to genome evolution remains largely undiscovered, but is rapidly being filled by ongoing efforts to dramatically expand genome sequencing across the Tree of Life [157].

3. Opportunities for making new discoveries in epigenetics with non-model species.

Many advances in understanding the molecular mechanisms of epigenetic control have been made in the model plant *A. thaliana* [20], with additional evidence from other species that pathways of epigenetic regulation can be surprisingly plastic. For example, in both *Eutrema salsugineum* and *Conringia planisiliqua*, the loss of a key enzyme for DNA methylation has resulted in the loss of gene body methylation [158], challenging many views about an essential role of this type of methylation in angiosperms. In addition, it has been proposed that the evolution of epigenetic silencing mechanisms has been a major force for the transition to multicellular life from the ancestral unicellular state [159]. Therefore a larger diversity of epigenetic control in unicellular, or early branching lineages, of the plant kingdom is expected. In the lycophyte (clubmoss) *Selaginella moellendorffii* and in the bryophyte (moss) *Physcomitrium patens,* methylation of TEs is similar to *A. thaliana*, but genes are completely depleted of methylation [160]. In contrast, in the chlorophyte (green alga) *Volvox carteri*, gene bodies are heavily methylated [160]. Albeit less studied than DNA methylation, presence/absence of several histone modification marks also varies across the plant kingdom [161]. A comparative study of histone deacetylase and histone acetyltransferase proteins revealed a complex presence/absence variance of these gene families and their expansion in specific lineages [162].

Comparative studies of differences in sRNA profiles across land plants, including algae and non-model vascular plants, have revealed distinct patterns between major plant lineages [163]. As in *A. thaliana*, typical plant sRNA size distribution profiles peak at 21 and 24 size nucleotides, which correspond to miRNAs and siRNAs, respectively. There is depletion of the 24 sRNA peak in the two green algae *Chlamydomonas reinhardtii* and *Volvox carteri*, but this peak is found in several gymnosperms and ferns, suggesting that the epigenetic machinery that generates 24-nucleotide sRNAs, RdDM (as discussed before), evolved before the gymnosperms [163]. It also suggests that early branching photosynthetic organisms may lack this pathway, and different strategies for epigenetic silencing may have been adopted. Analysing parts of the genetic component of RdDM, AGO proteins, revealed a parallel diversification of these key components in land plants [164].

Setting aside major differences across eukaryotic lineages in the epigenetic machinery itself, DNA methylation content and thus, heterochromatin states, vary among angiosperms [165]. These variations in DNA methylation can be stochastic or environmentally-induced and thus, these patterns can be shaped by evolution [165]. Indeed, differences in epigenetic states between species or natural populations can reveal different adaptation patterns of organisms to the environmental conditions, and explain some of the phenotypic variation not explained by differences in DNA sequence [166].

However, epigenetic variation across natural populations may have a genetic basis. In *A. thaliana,* CHH methylation variation is associated with specific *CMT2* and *NRPE1* alleles

[167]. But it can also happen that epigenetic changes influence genetic variation, as relaxation in the epigenetic silencing of TEs due to stress or other events can lead to genetic changes that, in turn, also affect the epigenetic states [168]. This genomic feedback makes it difficult to disentangle cause and effect.

The extent of epigenetic variation in natural populations has been mostly studied in *A. thaliana,* which is at the lower end of the total DNA methylation range for flowering plants [169]. Hence, an argument can be made that in other plant species with higher total DNA methylation, interspecific variation in DNA methylation may be more significant and have more impact at the phenotypic level. Finally, because the reconstruction of the genomic and epigenomic, evolutionary past relies on the comparison of today's extant genomes, the increase in studied genomes will increase our understanding of how the epigenetic and genomic variation evolved to shape today's biological diversity.

In conclusion, studying epigenetic mechanisms and its landscape in species with larger and more complex genomes will increase our understanding of the role epigenetic variation plays in adapting to natural environments over both short and long periods [168].

3.1 The biology of *Thlaspi arvense*, a member of the Brassicaceae.

One system in which one can study the epigenetic landscape because of its larger and more complex genome than *A. thaliana* is *Thlaspi arvense.* This system also has public genomic resources of natural populations [170], making it suitable to study TE dynamics.

T. arvense, field pennycress, is a common weed from the Brassicaceae family [171]. *Thlaspi arvense* diverged from their sister species around four million years ago [172]. Like many Brassicaceae, *T. arvense* is self-fertilising and has an annual life cycle, with both early and late flowering strains reflecting contrasting strategies for overwintering, namely either as seeds or vegetative rosettes [173]. Native from Eurasia, this species has been widely naturalised in both the Northern and Southern hemispheres, from temperate to subarctic regions, demonstrating a high capability for adaptation to different climates (Figure 8). As a ruderal species, *T. arvense* is found mainly in disturbed soil, like roadsides, railways or waste lands, but also in meadows or crop fields. Because of its prevalence in crop fields, it was first considered as an undesirable weed [171]; but *T. arvense* has more recently begun to be appreciated as a potential crop, since the natural oil content of its seed, together with its considerable seed size, makes it a desirable candidate for biofuel production [174].

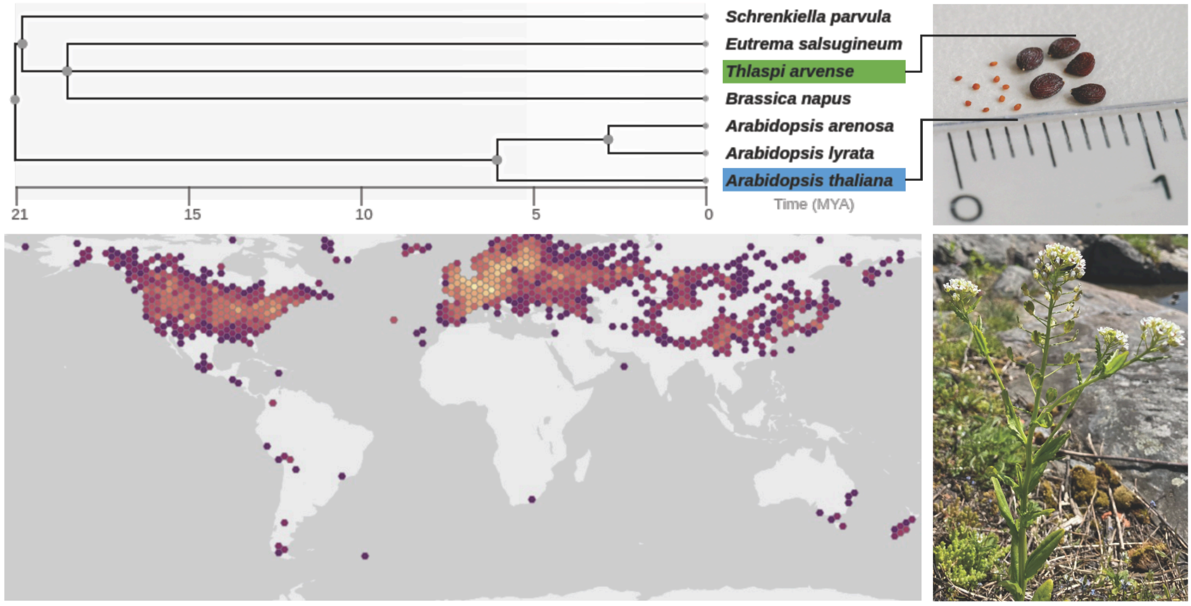


Figure 8 | Upper left: Phylogenetic tree that shows divergence time between a selected group of Brassiciaceae. It was constructed using TimeTree [175]. Upper right: Seed size comparison between *A. thaliana* and *T .arvense*. Bottom left: Global distribution of T*hlaspi arvense L.* Data points represent density of georeferenced human observations. Retrieved from GBIF Secretariat. GBIF Backbone Taxonomy. Checklist dataset https://doi.org/10.15468/39omei. Bottom right: Image of *T. arvense*. Photo credits: Natalie Betz and Detlef Weigel, respectively.

T. arvense is a diploid organism whose 539 Mb genome is distributed in 7 chromosomes. Due to selfing, wild *T. arvense* individuals present a high level of homozygosity [176]. The first draft genome of *T. arvense [176]* focused mainly in the gene space of this species.This showed a high degree of gene homology with the model plant *A. thaliana*, with ~86% *T. arvense* genes having an *A. thaliana* homolog, although the genome of *T. arvense* is around four times larger. This begs the question what the composition of the non-genic rest of the genome in *T. arvense* is.

This first assembly [176] consisted of only 343 Mb of sequencing, leaving around 40% of the genome unassembled. Likely, most of the unassembled portion of the genome of *T. arvense* is composed of TEs and other repetitive content. The advance of long-read sequencing technologies in the last years makes it possible to access these complex regions of this species and gives us the opportunity of examining the full extent of the TE content and its epigenetic regulation in this species.

4. Overview of doctoral research

Reference genomes are often generated with a comprehensive annotation of protein coding genes and the mRNA isoforms they produce, but this offers only a partial view of genome functions, many of which involve epigenetic mechanisms. Thus, the first two chapters of this thesis aimed to understand the dynamics of TEs in the non-model species *Thlaspi arvense*.

In the first chapter, I improved the annotation of a new reference genome of the potential oil crop *T. arvense* [177] by providing a detailed TE annotation together with the annotation of one of the main agents of TE repression: small RNA producing loci. This resulted in a resource for future studies involving comparative genomics of TE evolution in Brassicaceae, and for the *T. arvense* molecular breeding community.

Next, I explored the extent of the intraspecific TE variation in *T. arvense* in several populations covering much of the range of this species. In the second chapter, I identified TE families that have been differentially active across the different *T. arvense* populations, and I analysed the impact of this activity at the methylation level. To achieve this, I annotated TE insertion polymorphisms (TIPs) and TE absence polymorphisms (TAPs). Finally, I looked into whether the observed TE variation on *T. arvense* has any genetic basis.

As mentioned before, long read technologies and algorithms have enabled the cost-effective sequencing and assembly of complex genetic regions, including repetitive and TE-rich regions. The last chapter of this thesis takes advantage of the improved full genomic resolution provided by long-read sequencing to study interspecific variation at complex genomic loci in *thaliana*, namely nucleotide-binding site leucine-rich repeat (NLR) loci. NLR proteins are encoded by a highly diverse gene family that forms the core of the plant defence system mechanisms [178]. These NLRs are clustered in structurally complex regions within the genome, often replete with TEs (TEs). As TEs might shape the diversity of NLR clusters, a comprehensive and uniform annotation of these elements is crucial to understand the local genomic environments the NLRs reside in. Hence, I constructed a *de novo* “panTEome” annotation of 18 *A. thaliana* accessions, which will help in the understanding of the nature of the processes that generate the necessary NLR diversity.

Chapter One

Chromosome-level *Thlaspi arvense* genome provides new tools for translational research and for a newly domesticated cash cover crop of the cooler climates.

Citation: Nunn, Adam, Isaac Rodríguez-Arévalo, Zenith Tandukar, Katherine Frels, Adrián Contreras-Garrido, Pablo Carbonell-Bejerano, Panpan Zhang, et al. 2022. “Chromosome-Level *Thlaspi Arvense* Genome Provides New Tools for Translational Research and for a Newly Domesticated Cash Cover Crop of the Cooler Climates.” Plant Biotechnology Journal, January. doi.org/10.1111/pbi.13775.

Abstract: *Thlaspi arvense* (field pennycress) is being domesticated as a winter annual oilseed crop capable of improving ecosystems and intensifying agricultural productivity without increasing land use. It is a selfing diploid with a short life cycle and is amenable to genetic manipulations, making it an accessible field-based model species for genetics and epigenetics. The availability of a high-quality reference genome is vital for understanding pennycress physiology and for clarifying its evolutionary history within the Brassicaceae. Here, we present a chromosome-level genome assembly of var. MN106-Ref with improved gene annotation and use it to investigate gene structure differences between two accessions (MN108 and Spring32-10) that are highly amenable to genetic transformation. We describe non-coding RNAs, pseudogenes and transposable elements, and highlight tissue-specific expression and methylation patterns. Resequencing of forty wild accessions provided insights into genome-wide genetic variation, and QTL regions were identified for a seedling colour phenotype. Altogether, these data will serve as a tool for pennycress improvement in general and for translational research across the Brassicaceae.

Author contributions: RC, AN, KF and CB conceived the study. RC and AN led the genome assembly and evaluation, assisted by IRA and PCB. IRA performed the comparative genomics analysis of synteny during genome re-scaffolding and in the final evaluation. AN led the genome annotation and performed analysis for protein-coding genes, non-coding genes (tRNA, rRNA, snoRNA) and pseudogenes. ACG performed small RNA library sequencing, annotation and analysis, supervised by DW. PZ and ACG performed the transposable element annotation, supervised by DW and MM. AN performed the gene expression analysis and evaluation of tissue specificity. CB and KJ provided PCR-free libraries. RC performed k-mer analysis for genome estimation. KF and RC provided the CCS libraries, which were prepared by the UMGC. CB and IRA provided the DNA methylation libraries and analysis. RC, ZT, MDM and KF developed linkage mapping populations, designed primers, performed genotyping and built genetic maps. KF, RC and KD generated resources for Hi-C, Bionano and resequencing of accessions. KF and ZT phenotyped resequenced accessions. RC performed SNP analysis of resequenced datasets. ZT performed the linkage disequilibrium decay analysis. AB performed population genomics. RC and BJ prepared samples for Iso-seq libraries. RC and ZT performed gene structure variation analysis. RC and MDM performed bulk-segregant analysis. The PacBio CLR library was prepared and sequenced by PCB and AN under the guidance of CL. DR prepared and sequenced mRNA-seq libraries. RC, AN, CB, ACG, IRA and ZT wrote the manuscript. All authors reviewed and approved the manuscript.

Status in publication process: Published in Journal.

Chapter Two

Transposon dynamics in the emerging oilseed crop *Thlaspi arvense.*

Citation: Contreras-Garrido, Adrián, Dario Galanti, Andrea Movilli, Claude Becker, Oliver Bossdorf, Hajk-Georg Drost, and Detlef Weigel. 2023. “Transposon Dynamics in the Emerging Oilseed Crop *Thlaspi Arvense*.” bioRxiv. doi.org/10.1101/2023.05.24.542068.

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 catalyse 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 moulding 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*.

Author contributions: ACG, DG, OB, HGD and DW conceived the study. ACG generated data. CB. provided data. ACG, DG and AM analysed the data. All authors interpreted the results. ACG and DG wrote the first draft of the manuscript. ACG, DG, AM, CB, OB, HGD and DW revised the manuscript.

Status in publication process: Published as a pre-print in biorXiv.

Chapter Three

Impact of TE activity on *A. thaliana* NLR cluster evolution.

1. NLRs, plant disease resistant genes

Plant disease and the ZigZag model of plant immunity

Despite being constantly challenged by pathogens, one of the characteristics of plants is the lack of adaptive immune system compared to vertebrates, possibly due to the lack of a circulatory system. Instead, every cell of the plant is capable of defence, even though they can coordinate locally and systemically [179].

Plant cells use innate immune receptors to detect and initiate pathogen defences. A subset of these receptors are present in the surface of cells, surveying for the presence in the environment of a set of conserved molecules known as pathogen-associated molecular patterns (PAMPs). Upon recognition, these receptors will trigger a cell-wide response, known as PAMP-triggered immunity (PTI). A second layer of recognition takes place inside the cell, where intracellular immune receptors scan through the cytoplasm the presence of pathogen molecules, or the distress caused by them, and activate the effector-triggered immunity (ETI) defence response [180],

PTI normally prevents nonadapted microbes from infecting and is therefore an important barrier against disease [181]. Successfully adapted pathogens, however, have bypassed this first layer of immunity by creating a diverse set of factors that block PTI from starting. This set of factors are a subset of a larger collection of virulent factors that the pathogen uses to successfully infect the cell known as effectors, provoking the plant effector-triggered susceptibility (ETS). These effectors, in turn, can be recognized by the second layer of plant defence (ETI) through resistance (R) genes, which results in an enhanced PTI response, shielding the plant of any susceptibility by the pathogen. Pathogens that lose the recognized effector and still are able to infect trigger again ETS. Host fitness is reduced until one of the new effectors is recognized by new, emerging, R alleles of the host. This back and forth between ETS and ETI is known as the zigzag model [182] and represents a constant cross-kingdom co-evolution between pathogens and their host targets, between effectors and R genes.

NLRs: function and classification.

Nucleotide-binding Leucine-rich repeat Receptors (NLRs) are the keystone of plant immune ETI defence response, as most of the known R genes are NLRs [183]. NLRs act as intracellular sensors, detecting the pathogen-produced effector proteins. These pathogens range from viruses to bacterias, oomycetes, fungi, herbivores or even other parasitic plants [184]. NLRs achieve this by mixing and matching combinations of different protein domains at the variable N terminal domain. Downstream of this N terminal domain, plant NLRs carry a domain present also in animal genes APAF-1 and CED-4 called NB-ARC [185]. This domain consists of a nucleotide binding domain (NB) and several ARC motifs, which are proposed to act as a molecular switch, cycling between the inactivated, ADP, and the activated, ATP, bound forms [186]. At their C-termini, NLRs show a collection of leucine-rich-repeat (LRR) domains, which are involved in the recognition of effectors [187]. We can recognize three types of NLRs by their differences in their N-terminal domain: the coiled-coil domain NLRs (CNLs or CCNLs), the Toll/interleukin-1 (TIRs) receptor domain NLRs (TNLs) and the RPW8-like (resistance to powdery mildew 8)-like coiled-coil domain NLRs (RNLs). [184,188]. Not all NLRs present this canonical structure; it has been shown that some TIR-only or RPW8-like-only NLRs are sufficient to trigger ETI [189,190]. It has also been reported that NLRs may show additional domains that resemble pathogen targets (integrated domains, NLR-IDs), which has been postulated to act as molecular decoys [191].

NLRs are classified by their role in the immunity pathway as either sensor NLRs or helper NLRs. Sensor NLRs mainly act either by direct recognition of pathogen effector molecules or by monitoring the alteration of host immune components by the effectors. As pathogen effectors are more diverse than what plant NLRs can possibly be, indirect recognition of perturbations on effector's common targets may allow single NLRs to confer resistance to multiple effectors of multiple pathogens [180].

A pioneer work [192], demonstrated that RPS5 monitors the presence of the cleavage product of a host protein PBS1. This cleavage was produced by the pathogen effector AvrPphB. Although firstly thought to be a NLR decoy system in play, the deep conservation of PBS1 in flowering plants and the fact that it was present even in plants with no RPS5 [193] suggested an important role within the plant other than a simple decoy. In fact, it was later shown that PBS1 recognized pathogen flagellin and it had a role in PTI [194]. This case beautifully illustrates the validity of the zigzag model. A member of the PTI pathway (PBS1) is recognized by at least a known pseudomonas effector (AvrPphB), but the enzymatic activity of this effector neutralising PBS1 is monitored by a guard NLR (RPS5) that recognizes PBS1 cleavage products and triggers ETI.

Although some NLRs alone can both recognize a pathogen attack and start the downstream signalling cascade to activate the immune system, it appears that several sensor NLRs need the aid of other NLRs, helper NLRs, to activate the immune response

[195]. This decoupling of functions between the pathogen sensing role and the initiation of the downstream signalling cascade role, allows some NLRs to freely expand and evolve to keep up with the effectors, sensor NLRs, whereas other NLRs conserve the pathway activation role, the denominated helper NLRs [196]. This expansion and diversification of sensor NLRs with helper NLRs acting as signalling hubs makes the plant immune system act as a network gaining in resilience to mutation and pathogen suppression due to redundancy while maintaining core immune components in homeostasis [197]. This new conceptual framework of plant immunity moves on from the initial gene-gene interaction model proposed by Flor in 1942 to a network model with complex topology [197].

Once a pathogenic molecule is detected by the NLR immunity network, a signalling cascade initiates dramatic changes in the cell, raising an array of cellular defence responses, mainly through hormone production and transcriptional reprogramming [184]. This includes an increase of reactive oxygen species (ROS) and hypersensitive cell death (HR) [198]. Followed by the activation of the salicylic acid (SA)-dependent signalling pathway and expression of SA-dependent defence genes [199]. This increases the resilience of the plant to be infected by pathogens, a phenomenon called systematic acquired resistance (SAR) [200]. Lastly, in order to prevent autoimmunity, NLRs should be produced and activated only when needed through conformational changes and through a tight regulation of the NLRs at all levels, from expression, mRNA stability and splicing to protein localization and turnover [201,202]**.**

Mechanisms of NLR variation.

As we have seen previously, the configuration of plant immunity as an interconnected network in which sensor NLRs occupy the external nodes, allows for a high degree of diversity and variation within NLRs across and within species. But how is this variation generated in the first place?

An early model suggested that NLR genes were mainly evolving due to divergent evolution at individual loci, similar to proposals for mammalian immune genes [179]. This model suggests that the main mechanisms of NLR evolution were interallelic recombination and intragenic gene conversion events, as it has been also proposed for other multigenic families [203]. It also predicted that variation will be generated predominantly by interallelic recombination as this would result in high levels of protein diversification without the need of a local higher rate of point mutations. In this model repeated DNA stretches surrounding NLRs could provide the starting point of unequal crossing events, duplicating the genetic material between them. Point mutations however, will increase local divergences in such a way that meiotic mispairing leading to gene conversion will be reduced, fixing variants in the haplotype.

Due to their genomic self-propagation nature,TEs can play several roles in the evolution of NLR genes. Recurrent insertions of elements of the same family in the vicinity of NLR genes can provide the substrate needed for unequal crossing-over and duplication of an NLR, but at the same time, because TEs are often deleterious, they tend to accumulate more mutations than their neighbouring genes, which are more constrained due to purifying selection. Thus, these homology hotspots caused by recent TE insertions will disappear over evolutionary time and become the main source of divergence between loci, reducing the chances of future recombination events. Indeed, although with notable exceptions, there is a negative association between TEs and meiotic recombination rates [204].

TE insertions at NLR loci will also modify the NLR loci's epigenetic composition. Because epigenetic mechanisms mainly target and silence TE elements [205], this conformational change in the heterochromatin will affect neighbouring regions. Furthermore, this alteration can be beneficial and being co-opted by the host plant, for example *A. thaliana*'s RPP7 gene has an LTR-copia element inserted in the first intron of the gene [206]. Another example is *Oryza sativa* (rice) PigmS gene expression. PigmS is controlled in a tissue-specific manner due to the presence of MITEs in the promoter region [207]. In *A. thaliana* an LTR solo fragment of ATCOPIA93 (EVADE) cis-regulates RPP4 expression during bacterial infection [208]. Ultimately, because the epigenetic landscape is altered during biotic stress, TE derepression can affect the regulation of NLR genes [17,209].

Another way that TEs can bring to NLR evolution can increase allelic variability of NLRs by gene capture processes, intronic insertion leading to splicing alteration and transcription factor shuffling, resulting in regulation network rewiring [126,152]. In particular for NLRs, a recent study suggests a significant role for TE transposition to the formation of ID-NLRs [210].

NLRs are organized in clusters in the genome.

Possibly, one of the most well known genomic characteristics of NLRs is that they are often organised in clusters. This was first observed in *Lactuca sativa* (lettuce) [211] and this finding led to the above referred model of NLR evolution. Gene clusters, regions of the genome where genes of the same family are physically close, are a rare phenomena in eukaryotes [212]. However, for plant NLRs, cluster confirmation is quite common from mosses to angiosperms [213]. Therefore, it is plausible to consider the maintenance of NLR cluster arrangements to respond to evolutionary pressures.

We can distinguish two classes of clusters, head-to-head gene pairs and large clusters. NLRs can be found in a head-to-head orientation with one of the members carrying an integrated domain (ID-NLR), suggesting coordination of the sensing functions between the canonical NLR and the paired ID-NLR [213]. Coexpression of these NLR pairs has been observed In *A. thaliana* for SOC3– CHS1–TN2 [214], but this is not a general pattern for head-to-head NLR pairs.

Much of the NLR genes appear to be in larger clusters, and this tendency has been closely examined in few angiosperms [211,215,216]. These examined NLR clusters appear to be products of small, gene size, chromosome rearrangements, due to gene conversion, tandem duplication events and unequal crossing over. A detailed analysis of crossovers between two commonly used lab accessions of *A. thaliana* [217] revealed that NLRs in tandem, inverted and singleton confirmation had a high proportion of crossovers. Although if the structure of the NLR loci was too divergent between the two accessions, then the chances of crossover were reduced.

The beneficial effects of increase in copy numbers are obvious, after a duplication event, one of the copies can experience a more relaxed selection due to redundancy [218] . Also, an increase of copy numbers increases the chances of point mutations occurring in any of the copies for natural selection to act upon.

NLR diversity. Ubiquity beyond the plant Kingdom.

As we have seen previously, the configuration of plant immunity as an interconnected network in which sensor NLRs occupy the external nodes, allows for a high degree of diversity and variation within NLRs across and within species. Furthermore, cluster configuration of these NLRs favours increase of diversity of these genes through cluster expansions and contractions. But how much variation within species exists? What is the real degree of diversity across species?

NLR genes originated quite early, before land colonisation by plants. A comparative genomic analysis pinpointed the origin of those genes in green algae, Charophytes [219]. Their analysis used the NB-ARC conserved domain as bait to search for NLR in basal species. They found that NLRs were present in basal embryophytes (land plants) like liverworts (*Marchantia polymorpha*); mosses (*Physcomitrella patents*); and lycophytes (*Selaginella moellendorffii*). Although they found NB-ARC domain also in rodopythes (red algae), they were discarded as not true NLR genes due to their lack of other related domains. They narrowed down the emergence and diversification of NLRs in charophytes as they found both TIR NLRs and nonTIR NLRs. The presence of NB-ARC domains in non-NLR proteins is not surprising as this domain is also present in animals [185] as a motif involved in the regulation of cell death. Other studies [220] have examined the complete loss of TIR-NLs on multiple monocot lineages, illustrating also the contraction of this gene lineages in some branches.

A recent review [221], compiled NLR number estimates from a high variety of organisms from algae, mosses, grasses and trees. It showed that, although the number of NLRs weakly correlates with the total number of genes, there are some significant contractions in *Zea mays* (maize) and *Brassica napus* (oilseed rape), as well as expansions in *Medicago truncatula* (barrelclover) and *Triticum aestivum* (wheat). Indicating that although domestication can be an explanation for lower numbers of NLRs, especially compared with wild relatives, is not always the case.

Early empirical studies on *A. thaliana* natural populations that looked at variation of NLR genes [222] found different trends at NLR loci, suggesting different selective pressures and life histories acting at each loci. For example, they found RPP13 to be highly polymorphic, and likely under balancing selection, whereas other NLR loci were quite invariant. Although early studies are hampered by the impossibility to examine the genomic architecture of the examined loci, which can have a high influence determining variation at a given NLR gene. Recently [223] examined a highly diverse set of NLR genes, more than 13,167 genes in 64 *A. thaliana* accessions. Their study achieved near-complete saturation of the species' NLR space, due to the strategic selection of accessions based on prior knowledge of the species' diversity. In contrast, a study in *Solanum pennellii [224]*, a tomato wild relative, detected a reduced SNP diversity in NLRs, probably due to the reduced overall diversity of the examined collection. This shows that part of the NLR diversity loosely correlates with the general population structure of the species. Although these genes were compared in isolation, without knowledge of their accession specific genomic background, which may impair the correct assessment of NLR diversity.

Another recent study [225] illustrates the importance of genomic background in understanding NLR variation. They focus on the RPP8 locus in 37 accessions of *A. thaliana*. This loci contains three paralogs of RPP8 (two as a tandem duplication and the third at 2 megabases of distance) arranged in different chromosomal configurations including one with an absence of one of the tandem paralogs. The variation and shared diversity between these paralogs prompted the researchers to elaborate a complex three-locus gene-exchange circuit explanation where both gene conversion events and rare allele advantage take place to ensure a constant generation of novel alleles. This study demonstrates the importance that the local genomic architecture has on NLR variation and evolution. Outside of A*. thaliana*, intraspecific NLR variation has been examined intensely in crop cultivars such as pepper [226], maize [227] or wheat [228] to name a few. Understanding the level of variance of intraspecific NLRs is of foremost importance in breeding programs to combat diseases and overall crop improvement.

2. Chapter aim, data sources and contributors

Background and motivation

As we have seen in the introduction to this chapter, NLR evolution can be influenced by the local composition of TEs in the genome. Moreover, TEs may influence the function of some NLRs, as described before with the role of ATCOPIA93 in the expression of *RPP7* [208] in Col-0. Although different studies have hinted at a great amount of intraspecific diversity of these NLR genes, these have been limited mostly to the level of individual genes, highlighting the variation within NLR genes but lacking the genomic context where those variants are found [178]. Nonetheless, observations of presence/absence variance at the TE level [53] nearby NLR genes have been done.

Advances in genome sequencing and assembly have made it much easier to explore genomic diversity within species at reasonable costs. Some authors have proposed the concept of species pangenomes to be the new standard [229]. In fact, a recent study of seven *A. thaliana* full-length genome assemblies analysed the level of synteny within the species and found regions enriched in NLRs as hotspots of rearrangements between the assemblies, although the genomic features and evolutionary forces that may have caused such rearrangements were not explored [230].

Aim of the chapter

To better grasp the extent of the role of TEs in generating diversity of NLRs in *A. thaliana*, and, moreover, to better understand the evolutionary dynamics that generate and maintain this diversity, a comprehensive and uniform annotation of TEs becomes necessary.

In this chapter I will first describe what we know about TE diversity in the *A. thaliana* Col-0 reference genome and then I will apply this knowledge to the annotation of a collection of 18 diverse accessions across the *A. thaliana* native range (Figure 1). Then, I will describe the variation of TEs across these 18 accessions, followed by a composition assessment and an overview of TE dynamics within these accessions. Lastly, I will focus on the TE configuration around and within NLRs.

Main data sources

The *A. thaliana* accessions, known as Differential Lines, used in this chapter were selected by Gautam Shirsekar based on previous analyses [231] that best represented the genetic diversity of the species in Europe. These accessions were sequenced using PACBIO HiFi technology and assembled using hifiasm (v0.15.4-r343) by Maximilian Collenberg (Figure 2). Luisa Teasdale and Leon van Ess produced the transcriptome data used here. It was obtained using long-read PacBio IsoSeq RNA technologies using pools of 10-day old seedlings of each of the 18 accessions with three isolates of the oomycete specialist pathogen *Hyaloperonospora arabidopsidis*.

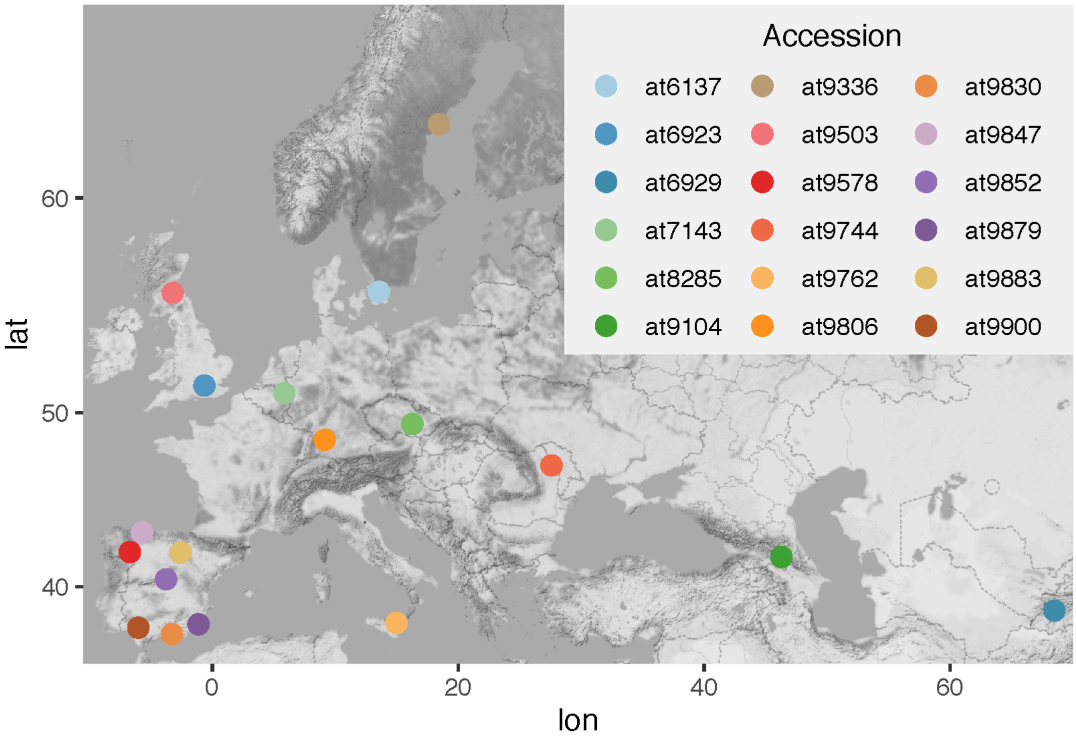


Figure 1 | Accessions chosen as part of the differential lines.

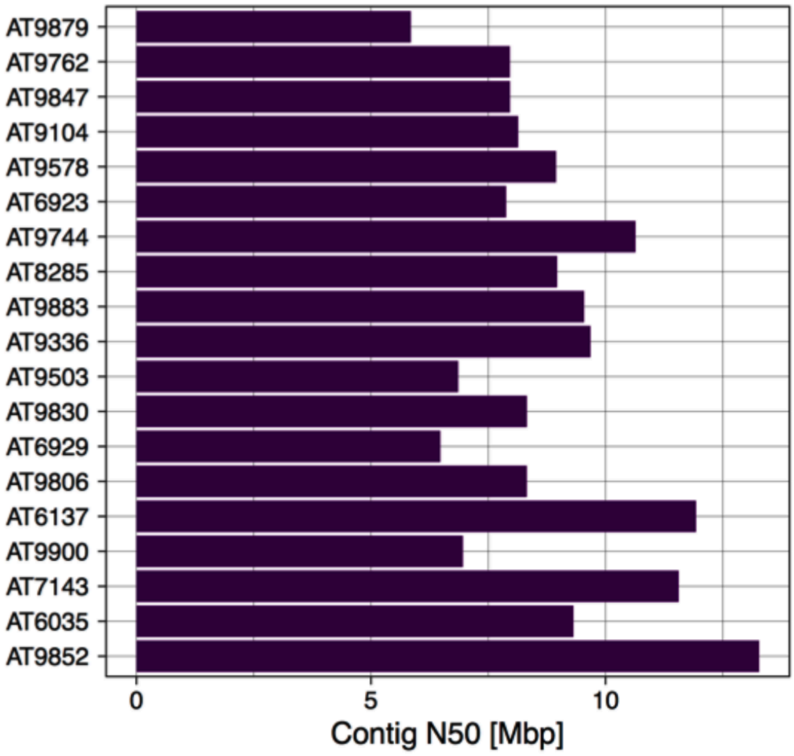


Figure 2 | N50 quality metric of 18 de novo genome assemblies.

Contributors

Since I made use of data produced, collected and analysed by other members of the group, Table 1 describes the roles and/or tasks of the different contributors.

Role/task	Contributors
HMW-DNA extraction and sequencing.	Theresa Schlegel and Christa Lanz.
Genome assemblies and *de novo* annotation.	Max Collenberg and Kevin Murray.
Manual curation and liftover of gene annotation.	Luisa Teasdale, Justina Juettner, and Kevin Murray.
Transposable element and repeat annotation.	Adrian Contreras Garrido.
Transcriptome generation and processing.	Luisa Teasdale, Kevin Murray, Gautam Shirsekar, Anette Habring, and Christa Lanz.
Orthogroup assignments and evidence-based annotation.	Luisa Teasdale, Leon van Ess, and Kevin Murray.
Synteny analysis.	Kevin Murray, Luisa Teasdale, and Adrian Contreras-Garrido.
Pseudogene identification.	Luisa Teasdale, Gautam Shirsekar.
Segmental duplication and gene conversion.	Gautam Shirsekar.
Methylation inference.	Gautam Shirsekar. (Adrian Contreras-Garrido and Regina Mencia: Bisulfite library preparation).
Population-scale recombination rate.	Gautam Shirsekar.
Webapollo server.	Joffrey Fitz.
Supervision.	Gautam Shirsekar, Luisa Teasdale, Kevin Murray, Hajk-Georg Drost, and Detlef Weigel.
Conceptualization and coordination.	Gautam Shirsekar, Detlef Weigel.

Table 1. List of contributors that participated in the differential lines project, their tasks and roles.

3. Methods

To annotate repeats in each genome, I first used EDTA's (v.1.9.7) [232] raw module to annotate: LTRs with LTRharvest (v1.5.10) [233], LTR_FINDER_parallel (v1.0) [234] and LTR_retriever [235]; Helitrons with HelitronScanner (v1.0) [236]; and TIR elements with TIR-Learner (v1.23) [237] and MITE-Hunter (v1.0) [238]. I then merged all 90 chromosomes from the 18 accessions and proceeded with the downstream automatic curation by EDTA (v.1.9.7) with a curated library obtained from the Araport11 annotation [239]. I added to this library additional rDNA, telomeric, and centromeric repeats sequences obtained from [240].

To refine these automated annotations and to detect novel repeat families, I conducted several additional steps: Independent and *de novo* satellite annotation combining TRASH and TRF and removed EDTA-annotated TEs overlapping >20% with these newly annotated tandem repeats. I removed repeats not assigned to a known repeat superfamily, as they were predominantly either artefacts of the joint analysis of all 18 genomes, or unidentified satellite repeats. To increase the fidelity of helitron annotations, we characterised which helitron families have at least one intact member that contained a Rep/Hel domain. Additionally, we used EAhelitron [241] to reannotate Helitrons, and characterize which helitrons are identified by EAhelitron and EDTA. Solo intact TEs identified by EDTA but not grouped in any family were reclassified them using blast [242] and following the 80/80/80 rule (TEs are considered to belong to a certain family if they are longer than 80 base pairs, and share at least 80% sequence identity over 80% of their length with the TE model of the family). The remaining unclassified TEs were clustered together in families using the same 80/80/80 rule with CD-hit [243]. Finally, we used TEsorter [244] for the obtained LTR families to phylogenetically classify them.

For downstream analysis we used the bedtools suite [245] to manipulate and perform different genomic arithmetic analysis among the TE annotation and NLR annotation. We also used R to conduct statistical and visualisation analysis [246] and figures were produced using the package ggplot2 [247]. To calculate the LTR insertion age, we used the *A. thaliana* mutation rate given by [248] and we calculated the DNA distance in R using functions in the ape package. [249]. To classify LTR TEs as "solo" LTRs we used a custom method based on the alignment of the most intact long terminal repeat of each LTR TE family against the rest of the family members using minimap2 [250], those elements that mapped to the long terminal repeat with a high score and were of similar length were classified as "solo". To calculate the enrichment of overlaps between TE families and NLR neighbourhoods, we used the BEDTools suite [245] fisher function to perform Fisher's exact test within the pangenome. Borrowing it from the ecological field, we measured the levels of TE diversity, understood as the different number of TE copies for each TE family, of each NLR neighbourhood for each accession, with the Shannon Entropy Index.

4. Results

4.1 TE annotation of a diverse set of *A. thaliana* accessions.

TE landscape in *A. thaliana* Col-0.

It has been estimated that 21% of the 125 Mb long *Arabidopsis thaliana* Col-0 TAIR10 reference genome (which excludes the centromeres) consists of TEs, based on the official TE annotation given by TAIR10 [251]. According to this, the Col-0 genome contains 34,856 transposon fragments, 31,189 full length transposable elements and 3,903 transposable element genes. In the TAIR10 genome, the most abundant TE type is RC/Helitrons, accounting for more than 6% of the total genome coverage, followed by LTR/Ty3 and TIR TEs. Notably, non-ltr retrotransposons, represent a small fraction of the TE content in TAIR10 (Figure 3). Remarkably, compared with other plants ([252], the TAIR10 genome is relatively rich in Helitrons compared with LTRs. This is probably due to a lack of LTR expansion compared with other plants that have larger genomes [253].

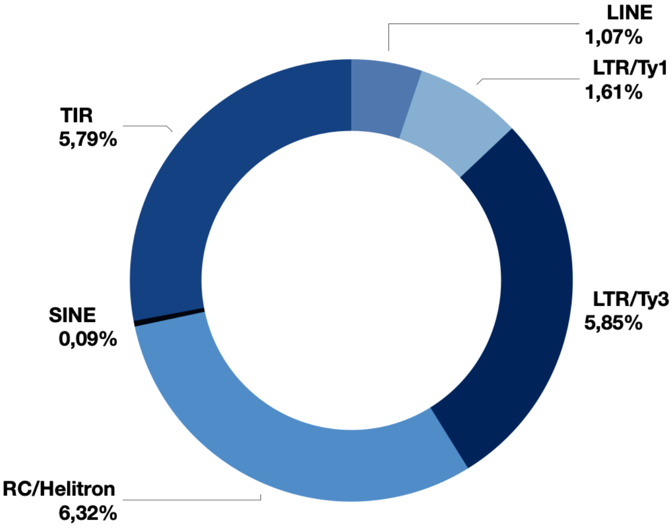


Figure 3 | Relative abundance of the major TE types in the *A. thaliana* Col-0 TAIR10 genome.

These 31,189 full length elements are divided into 319 families of different sizes, plus a catc-hall family term called “Unassigned”, with 113 elements, formed of unclassified TEs. There are, however, some inaccuracies in this annotation. For example the SHADU TE family is thought to be a member of the SINE superfamily [254], yet is categorized as Unassigned. Another inaccuracy in TAIR10 is not categorizing the families RathE1_cons RathE2_cons and RathE3_cons as the SINE superfamily as shown in [30] but rather as their own superfamily.

The distribution of TE copies per family presents a skewed profile with a long tail: few families are very abundant whereas many families are composed of few individuals, a pattern that is typical for the composition of ecological communities [255] (Figure 4.A). The DR1 family has only one member, and at the other end, ATREP3 has 1,439 members. As for abundant families, the most abundant ones are: ATREP3, HELITRONY3, ATREP10D, which belong to the Helitron superfamily. For LTR TEs, the ATHILA and ATCOPIA families stand out. A snippet of the most abundant families can be seen in Figure 4.B. Moreover, there are seven transposable element families whose representatives are shorter than 300 bp (ATDNATA1, ATHATN8, ATHATN9, ATTIRTA1, DRL1, RathE1_cons, and RathE3_cons).

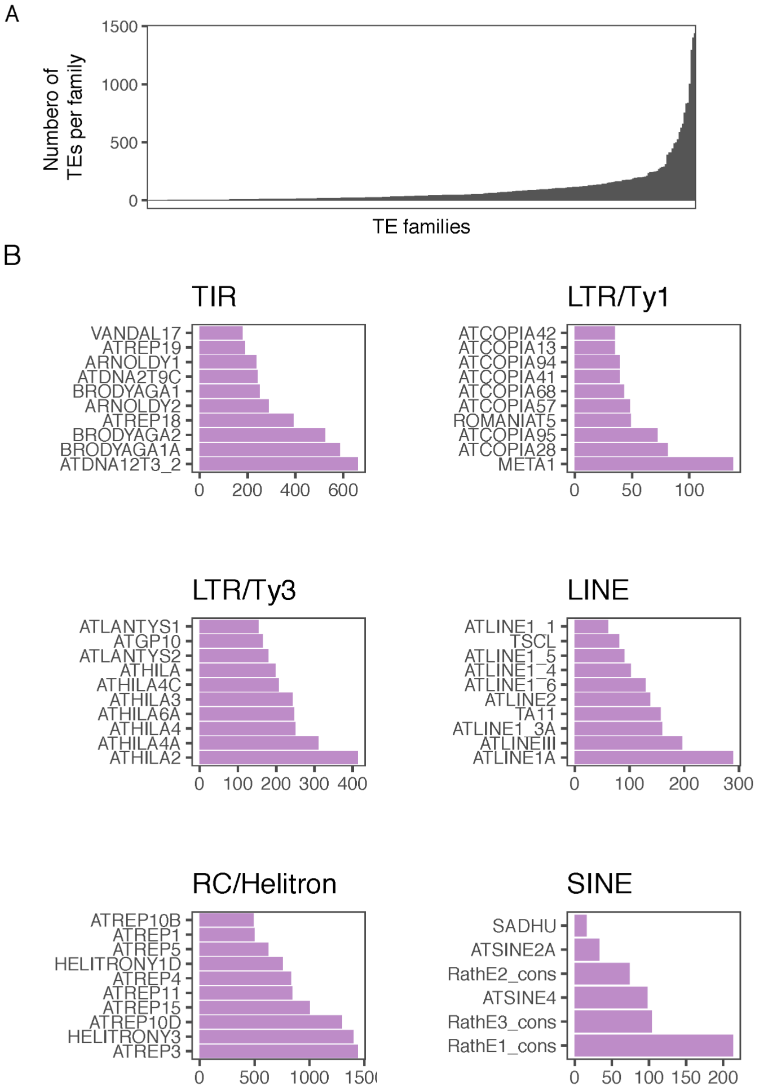


Figure 4 | A) Copy number profile of TAIR10 TE families. B) For annotated full length TEs, 10 most abundant TE families per TE type. X axis represents the total number of full length TEs.

TE distribution is a function of both target site preference of the different TE types and distinct rates of purifying selection in different genomic regions [256]. Thus, TEs are unequally distributed across the chromosomes in a distinctive fashion relative to genes (Figure 5). In *A. thaliana* Col-0 TAIR10, gene content is equally distributed across the chromosome arms, with a progressive, steady decline towards the pericentromeric regions. In contrast, TEs present different profiles depending on their type, Helitrons are mostly enriched in the pericentromeric regions, as are LTR/Ty1 TEs, whereas LTR/Ty3 are enriched in the near-centromeric portion, as early studies have shown [257]. In a similar way as retrotransposons, TIR TEs are mostly located in pericentromeric areas, with a few hotspots in the chromosome arms for chromosome 1. This peak may be caused by the result of two ancestral chromosomes fusions, as comparative genomics and ancestral reconstruction have shown [258], Lastly, SINES and LINES show a similar distribution as TIR TEs, although with less of a bias for the pericentromeric regions.

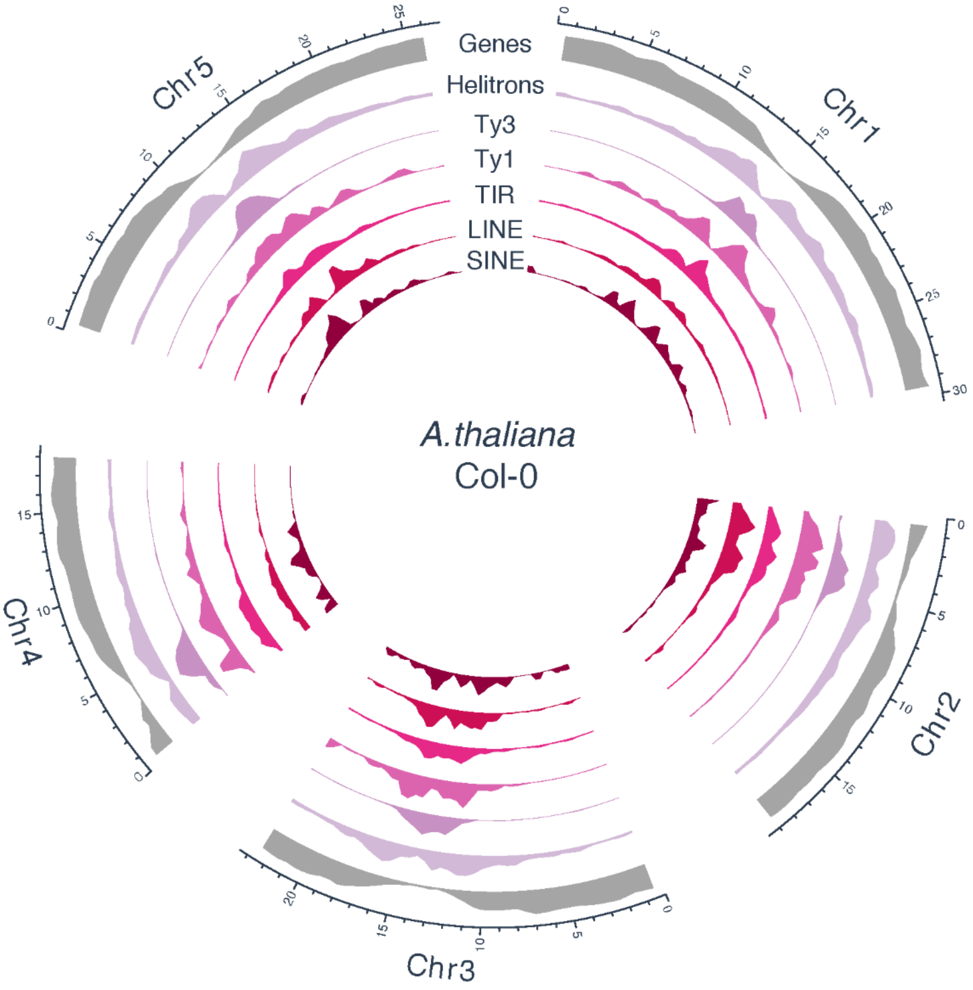


Figure 5 | Genomic distribution of gene and TE content for *A. thaliana* Col-0 strain as annotated by TAIR10.

Using the TAIR10 TE library for TE annotation and the pangenomic approach.

There is a high quality TE annotation available for *A. thaliana* Col-0 TAIR10 genome assembly. Moreover, much of the current knowledge in the field of NLRs and disease resistance in general has been done on the basis of the TAIR10 annotation. Thus, one of the priorities of my TE annotation was to carry over as much as possible this information to the new assemblies in order to facilitate possible comparisons across NLR clusters and NLR genes between Col-0 TAIR10 and the Differential Lines.

Another reason is that, generally, *de novo* TE annotations tend to yield imperfect TE models and manual curation is required to produce a high quality TE annotation [259,260]. Thus, by using the TAIR10 TE library I can take advantage of high-quality TE information to aid in our annotation. I combined the sequences of the TAIR10 TE library with other satellite repeat sequences, to also annotate rDNA gene clusters, telomeres and centromeres [240] based on homology. A breakdown of the number and nature of the sequences can be found in Table 2. I also supplied to this pipeline a file of coding sequences (CDS) covering all genes in Araport11 (which improves the gene models of TAIR10 based on transcriptome data) [239].

I first applied the EDTA pipeline [232] to each of the accessions from the *A. thaliana* Differential Lines resource individually. EDTA is a pipeline that combines several *de novo* annotation tools for the major types of TEs (LTR TIR and Helitrons), with several filtering steps and options to include a TE library for homology annotation. This pipeline is currently heavily used for *de novo* TE annotation of plant genomes. Examples are [240,261–264].

TE Order	**Number of TE copies**
TIR	10188
Helitron	12945
LTR	5962
LINE	1447
SINE	522
Satellite repeats	**Number of templates**
rDNA	4
Telomere	1
Centromere	6

Table 2 | Content of the library of TE sequences used in the annotation for homology based TE annotation.

Unfortunately, this approach of independent annotations yielded unsatisfactory results. Upon examination, TE content was extremely divergent across assemblies, which was unexpected given all accessions belong to the same species. The identity levels of copies from *de novo* TE families was much lower than those TE copies from TAIR10 TE families. This discrepancy suggested issues in the conformation of de novo TE families, as we expected average values to be similar (Figure 6.A). Moreover, any *de novo* TE model recovered was independent for a single accession and not shared, which would impede comparisons among accessions for *de novo* TE families.

To address these issues, I decided to concatenate all chromosomes of the 18 assemblies in a single pangenome, indicating in each chromosome the accession origin.By uniting them in a single fasta file to run EDTA, I could annotate all the *de novo* TEs families with shared denominations avoiding the issues mentioned before. Also, in theory, combining assemblies and thus chromosomes, could get the *de novo* TE identification software to recover full length TEs and then better determine the borders of a TE model. Although this reasoning was not fully tested, the overall performance of the tool when comparing it with the independent runs was encouraging. I called this approach "panTEome" and it performed overall better, achieving a more uniform distribution of total number of TEs, annotating a higher number of elements in total. It achieved a higher percentage of identity of the *de novo* families than those from the independent run (Figure 6). It also assigned some TE copies to *de novo* TE families instead to the TAIR10 TE families, likely due to the better distribution of copies among *de novo* families from different accessions. Although the main purpose was to produce a common TE library for all accessions to annotate the TEs of the assemblies under a common framework. A similar approach has also been recently used [265], where the authors combined taxon wide TE models into a single, common TE library in order to annotate genome assemblies from a taxon wide collection of species.

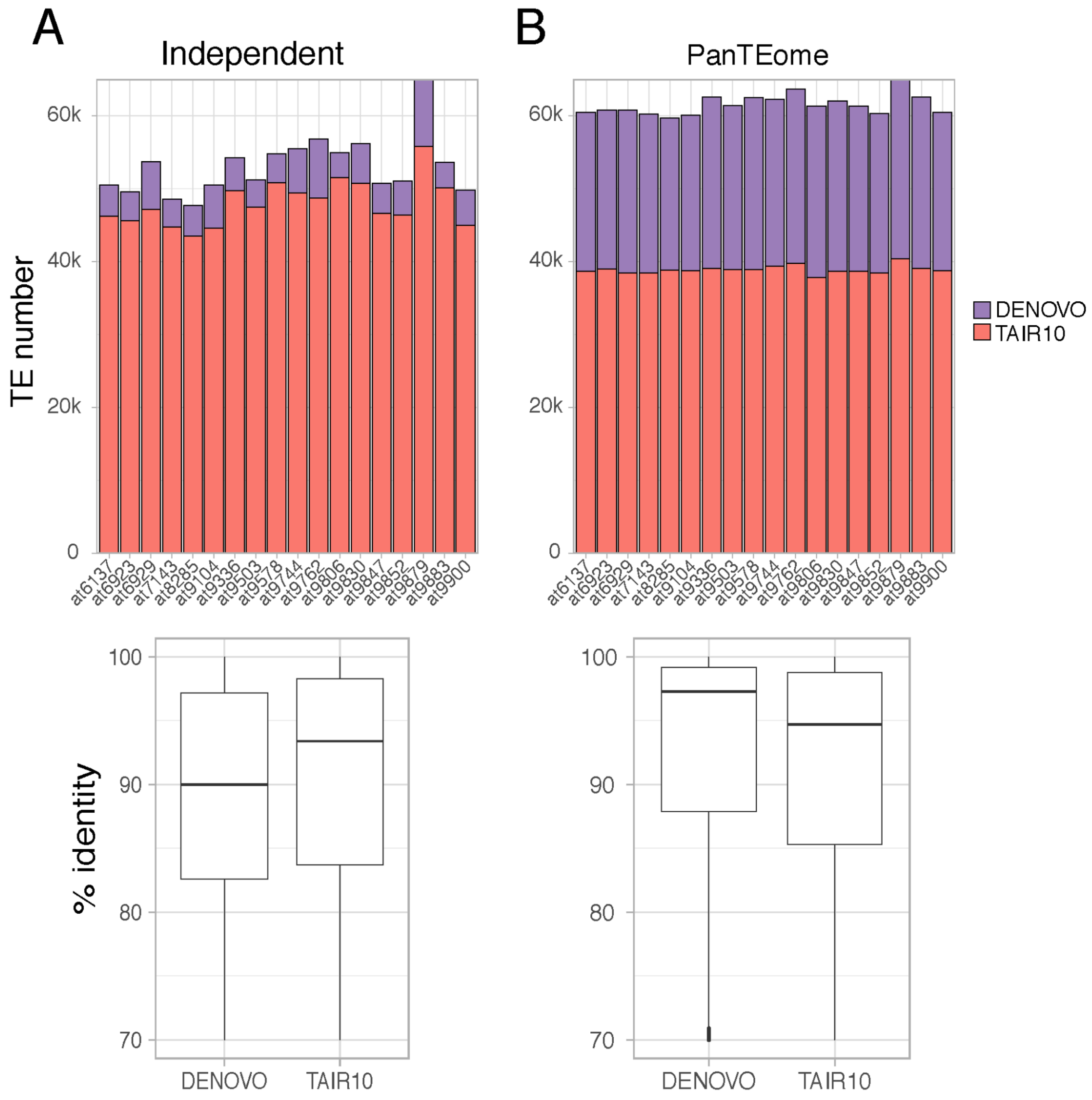
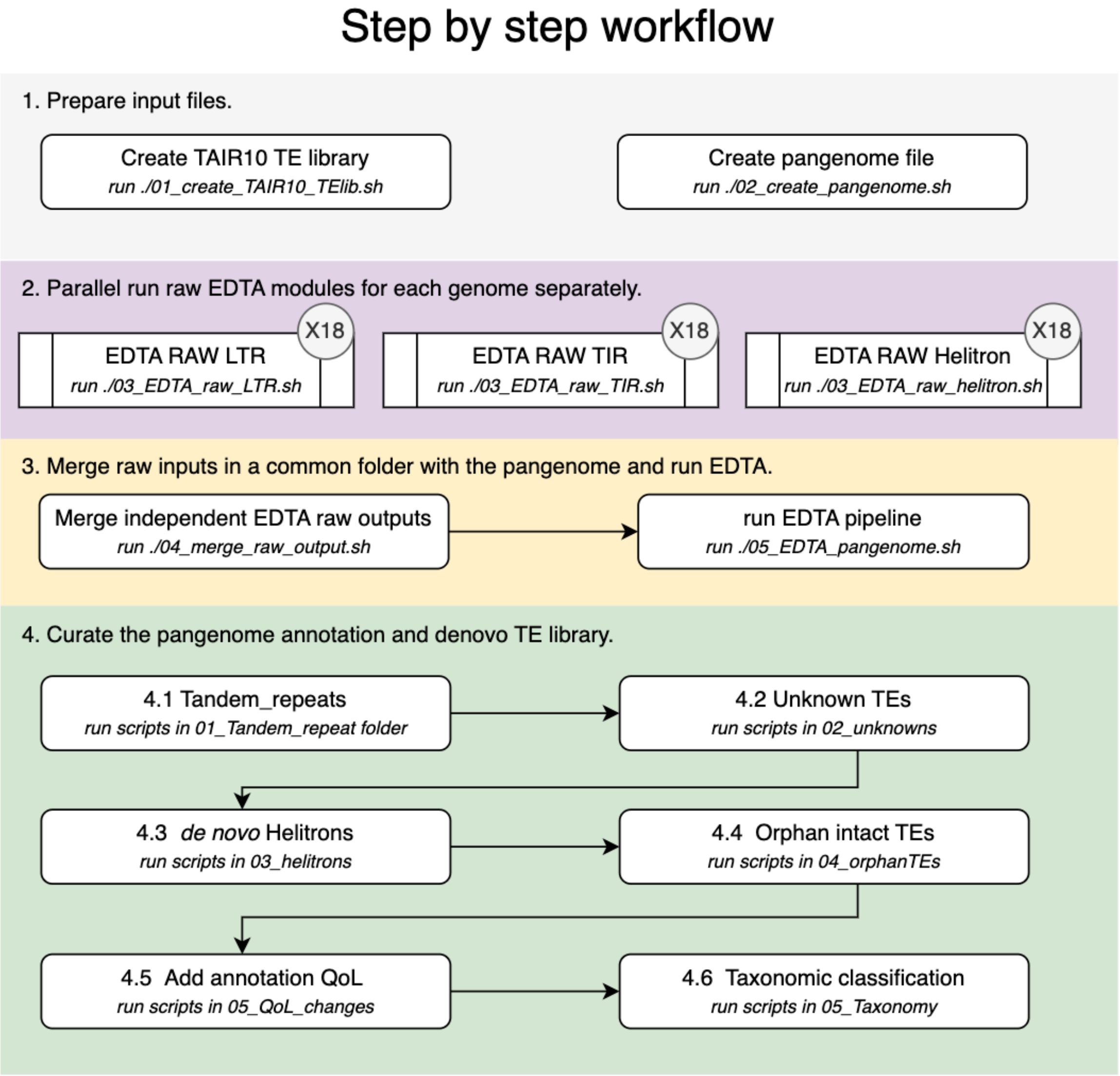


Figure 6 | Comparison of (A) different EDTA runs for 18 independent runs, and (B) the combined approach, which yielded a "panTEome". Upper panel: comparison of overall TE copies annotated using homology with the TAIR10 library and *de novo annotated* TEs. Lower panel: Percentage of identity of each TE copy relative to their corresponding TE model in either the TAIR10 or *de novo* TE library.

Despite the better performance of the panTEome approach, it is still an automated TE discovery pipeline, so a more detailed examination of the TE models produced was desirable. Overall, compared with the 320 TE families presented in Col-0, this approach annotated 2,369 *de novo* TE families, averaging 131 new families per genome. In total, these 2,369 new families contributed less than half to the total TE number per accession (Figure 6). This is likely due to the lower quality of the *de novo* TE models, possibly due to their fragmentation level. Thus I decided to address some of these issues with a manual inspection of the annotation and implemented those changes in an automated manner via a collection of scripts deposited together with the scripts I used to produce the automated annotation in the first place in github https://github.com/acontrerasg/PanTEome (Figure 7). I will describe the reasoning behind those steps in the following sections of the chapter.



Step by step workflow
1. Prepare input files.
Create TAIR10 TE library
run ./01_create_TAIR10_TElib.sh
Create pangenome file
run ./02_create_pangenome.sh
2. Parallel run raw EDTA modules for each genome separately.
X18
EDTA RAW LTR
run ./03_EDTA_raw_LTR.sh
X18
EDTA RAW TIR
run ./03_EDTA_raw_TIR.sh
X18
EDTA RAW Helitron
run ./03_EDTA_raw_helitron.sh
3. Merge raw inputs in a common folder with the pangenome and run EDTA.
Merge independent EDTA raw outputs
run ./04_merge_raw_output.sh
run EDTA pipeline
run ./05_EDTA_pangenome.sh
4. Curate the pangenome annotation and denovo TE library.
4.1 Tandem_repeats
run scripts in 01_Tandem_repeat folder
4.2 Unknown TEs
run scripts in 02_unknowns
4.3 de novo Helitrons
run scripts in 03_helitrons
4.4 Orphan intact TEs
run scripts in 04_orphanTEs
4.5 Add annotation QoL
run scripts in 05_QoL_changes
4.6 Taxonomic classification
run scripts in 05_Taxonomy

Figure 7 | Workflow of the TE annotation and evaluation of the *A. thaliana* Differential Lines.

Curation of the TE annotation: Satellite TEs.

One of the first issues I observed in the raw annotation was the existence of 110 putative TEs of 0.5 Mbp or larger, all belonging to the ATREP18 family, with 3 to 10 copies per assembly. The ATREP18 family has been already reported to contain a canonical telomere repeat "AAACCCTAA" [266] and in this annotation is precisely present in telomeres. ATREP18 also contains degenerate centromere repeats, and those are responsible for the erroneous annotation of centromere sequences as TEs of more than 0.5 Mbp in size (Figure 8). I suspect this issue with ATREP18 is not detected in the TAIR10 annotation because it is based on a genome assembly in which centromeres remain unassembled. This issue was also present in a *de novo* TE family “TE_00000583”, which also had degenerate centromere repeat sequences within its sequence model and thus, gets annotated in the centromere as extremely large TEs (0.1 Mbp).

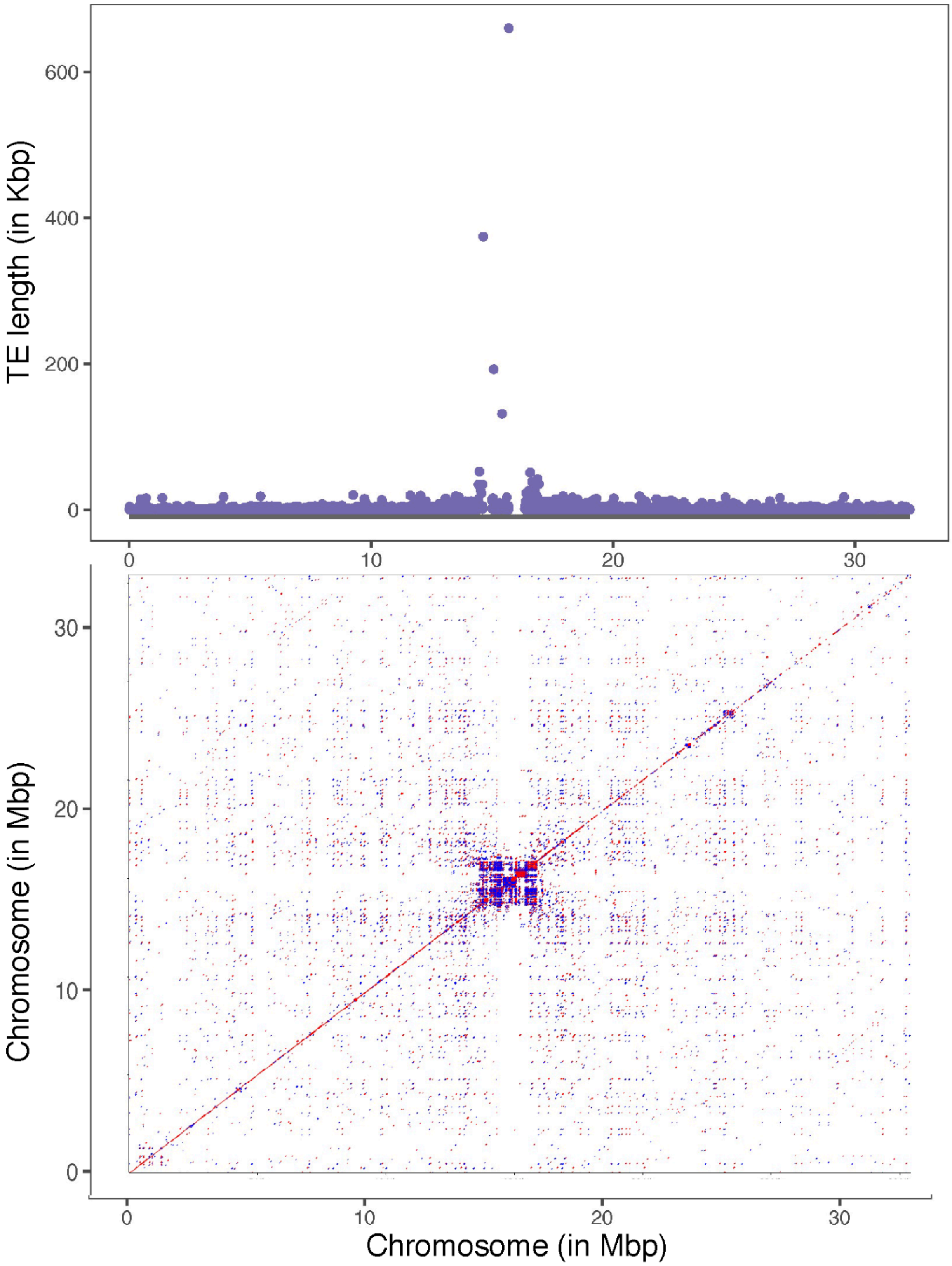


Figure 8 | Colocalization of the large TEs with the centromeric space. Upper panel: annotated TE sizes across the first chromosome of at6137. Bottom panel: Dot plot of

chromosome 1 of accession at6137 aligned against itself; high degree of local sequence repetitiveness indicates position of centromere roughly in the centre of chromosome 1.

Therefore, the underlying issue is the presence of fully assembled centromeres and telomeres in the fasta files used for the annotation, which confounds the software used for the annotation, specially RepeatMasker (https://www.repeatmasker.org/), in spite of having included centromere and telomere *A. thaliana* canonical repeats in the curated library used for the homology based annotation. The issue could be compounded by the repeats being based on the Col-0 *A. thaliana* accession, since a recent study has shown a high degree of centromeric diversity among accessions [267]. To tackle this, I decided to *de novo* annotate satellite repeats in the pangenome to complement the TE annotation and, importantly, remove any annotated TE that are within those satellite repeats.

I used a combination of two tools, TRASH and TRF, to annotate satellite repeats in the pangenome. I used TRASH because it is very sensitive for the centromeres and it is able to detect gaps in centromeric satellite repeats, which tend to be populated by TEs [267]. Because TRASH was not developed to detect telomeric repeats or interstitial satellite repeats, I complemented it with TRF [268], which is able to fully annotate telomeres, but does not detect breaks in the centromeres. A snapshot of how both tools perform in telomeres and centromeres can be seen in (Figure 9).

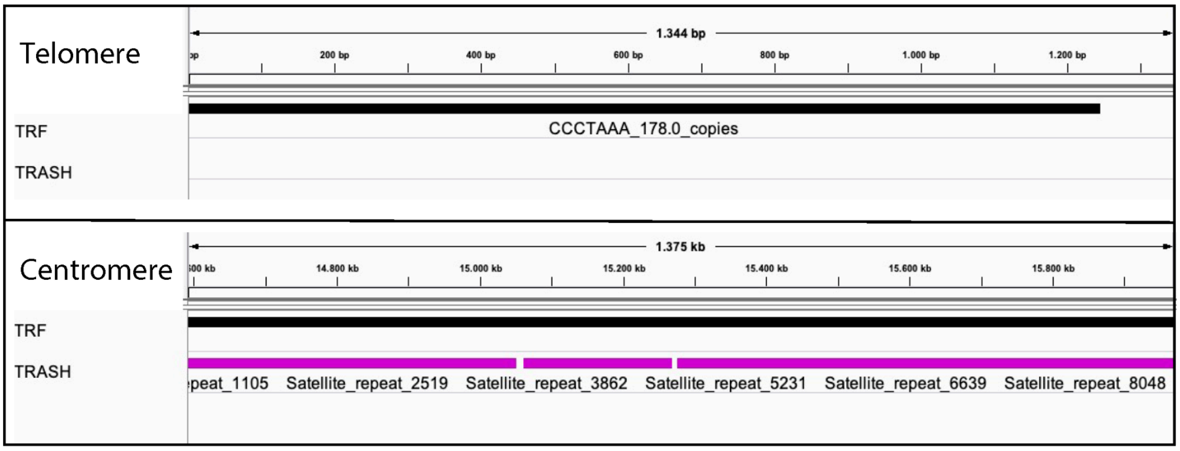


Figure 9 | IGV views of telomere (Upper panel) and centromere (bottom panel) regions of the chromosome 1 of at6137. Notice the breaks in the centromere satellite repeats that TRASH detects; these are populated by *bona fide* TEs.

Curation of the TE annotation: Unknown TEs.

A considerable number of *de novo* TE families was classified as "Unknown" (189 TE families, totaling 61,010 TE copies). This discovery was surprising, given that TEs in *A. thaliana*, and plants in general, have been thoroughly characterised. To investigate further, I traced the origin of these "Unknown" repeats and found that they resulted from running RepeatModeler as an optional step in the EDTA pipeline. RepeatModeler employs RECON, which uses self-comparison approaches to identify interspersed repeats [260]. However, a closer examination of these families using BLAST [269] against the NCBI database revealed that the majority of these "Unknown" TEs were actually multigene copy families. These false positives were a consequence of concatenating the different assemblies into a single pangenome, leading to erroneous detection of interspersed repetitive elements, despite my attempt to mask the coding sequences using Araport11 CDS files. As an anecdote, many of these false positives were NLRs. Consequently, I removed all occurrences of "Unknown" TEs from the annotations and regarded them as false positives.

Curation of the TE annotation: *de novo* Helitrons.

EDTA uses HelitronScanner [236] to detect *de novo* Helitron TE families. It annotated 193,875 TEs as Helitrons, divided into 4,471 TE families in the pangenome, accounting for most of the annotated *de novo* TE fraction (Figure 10). In order to increase my level of confidence on the Helitron *de novo* calls, I opted to add several layers of confidence in the annotation in three different ways. First, I added to the annotation whether a given family has at least one intact TE as a member of the family, as this will increase the confidence in the family. Only 87 TE families had no intact TE, most of them had only 1 intact TEs, and 35 TE families had at least 15 intact TEs (maximum, 23). Second, I searched for Rep/Hel protein domains in the sequences of these Helitrons. The Rep/Hel domain is a hallmark of autonomous Helitrons, as they are presumably the proteins necessary for their transposition [270]. Only 135 de novo TE families had a copy with a signature of this protein domain. However, although presence of the Rep/Hel domain is a signature of a correctly identified member of the helitron family, absence of this protein domains is not a sign of erroneous assignment, as many helitron families are known to be non-autonomous [271].

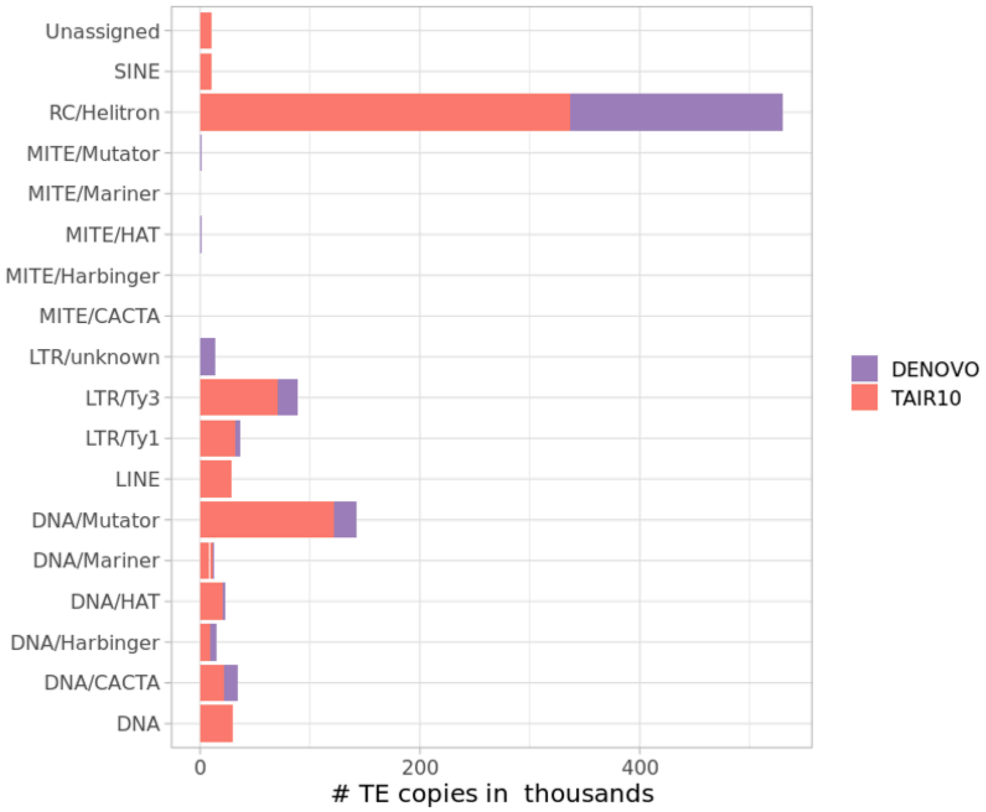


Figure 10 | Distribution of annotated TEs in the pangenome as either TAIR10 homology-based or *de novo* annotation.

Finally, I ran EAhelitron [241] against the pangenome. EAhelitron is the only available alternative to HelitronScanner for *de novo* annotating Helitrons [260]. EAhelitron detected 11,150 Helitron sequences, of which only 454 elements are in full agreement with the original EDTA annotation (Intersecting both using the command “bedtools intersect -r -f 0.9”). If I relax the overlapping parameters to an 80% of overlap required, “bedtools intersect -f 0.8”, the number increases to 2,121. This exercise illustrates the poor agreement between tools of defining Helitron boundaries, and therefore any Helitron annotation should be treated with caution. Nonetheless, I included 288 Helitrons found only by EAhelitron to the final TE annotation. These are distributed evenly among accessions (11 to 18 per accession). For the 2,121 helitrons annotated by EDTA with an 80% overlap with EAhelitron, I added to the annotation file the fact that they are recalled by both tools, increasing the confidence in them.

Curation of the TE annotation: Intact orphan TEs.

In the TE annotation, 10,605 TEs were structurally intact but not included in any TE family, likely because I ran the raw EDTA module independently for each assembly. To address this issue, I implemented two corrective measures. First, I clustered these TEs into known TE families detected by EDTA, following the 80/80/80 rule. My reasoning was that some of these TEs are orphans in one accession, but other accessions may contain similar copies and together they constitute families. Indeed, 4,288 orphan TEs were assigned to a known

TE family in this manner. I clustered the remaining TE copies that do not show correspondence with any known TE family. 818 clusters comprised two or more copies and were designated as new TE families. In total I was able to classify 5,032 additional orphans into new TE families. I assigned new family names to these clusters and incorporated the corresponding TE models into the TE library of the pangenome. 1,285 TE elements remained as orphans in the final annotation.

Curation of the TE annotation: Phylogenetic classification of TEs.

After all the curation steps, I wanted to enhance the structure-based classification of TEs provided by EDTA, which classifies TEs up to the superfamily level, with a phylogenetic classification of the TE families to the clade level, which will group these families in different taxa that better reflect their phylogenetic relationships. This classification can be done solely in the LTR fraction of the TE library [272], as TIR TEs and Helitron TEs lack a proper database with a phylogenetic classification.

For *de novo* annotated LTR TE families, I used their corresponding TE models as produced by EDTA. For TAIR10 TEs, I first prepared model sequences for each TAIR10 TE family creating a consensus with the TE copies of the family. Then I used TEsorter to detect TE-related protein domains [244] and to classify them using a published database [272]. The results can be seen in Table 3. I was able to classify phylogenetically 135 out of 141 TAIR10 LTR TE families. For the *de novo* TE families, I could classify only 321 out of 604 LTR TE families. This reflects the discrepancy in quality between a highly curated TE library (TAIR10) and a automatically generated one, despite some degree of manual curation.

TE Order/Superfamily	TE clade	TAIR10 TE families	*de novo* TE families
LTR/Ty1	Ale	62	59
LTR/Ty1	Alesia	-	3
LTR/Ty1	Angela	-	1
LTR/Ty1	Bianca	6	1
LTR/Ty1	Ikeros	-	3
LTR/Ty1	Invana	16	10
LTR/Ty1	SIRE	5	1
LTR/Ty1	TAR	1	2
LTR/Ty1	Tork	18	7
LTR/Ty1	Unknown	-	40
LTR/Ty3	Athila	13	1
LTR/Ty3	CRM	3	31
LTR/Ty3	Reina	2	53
LTR/Ty3	Retand	3	17
LTR/Ty3	Tekay	5	3
LTR/Ty3	Unknown	-	88

Table 3 | Taxonomic classification of LTR families using TEsorter.

4.2 TE landscape of 18 *A. thaliana* accessions.

One of the aims of the Differential Lines project was to capture and describe much of the diversity in the *A. thaliana* species with the resolution that long-read technologies bring to genome assemblies. This diversity is, in part, the result of the variability of the TE fraction of each genome. This section describes TE variation in the 18 accessions.

Overview of TE composition in the 18 Differential Lines.

Overall, the TE composition of the 18 accessions is remarkably similar (with a mean of 33.62 Mb). at9806 is the accession with the lowest TE content with 32.3 Mb, and the accession with the highest TE content, 34.6 Mb, was at9879. Also, the TE contribution to genome size is weakly correlated with the assembly size (Pearson's correlation R =0.43, p= 0.075, Figure 11).

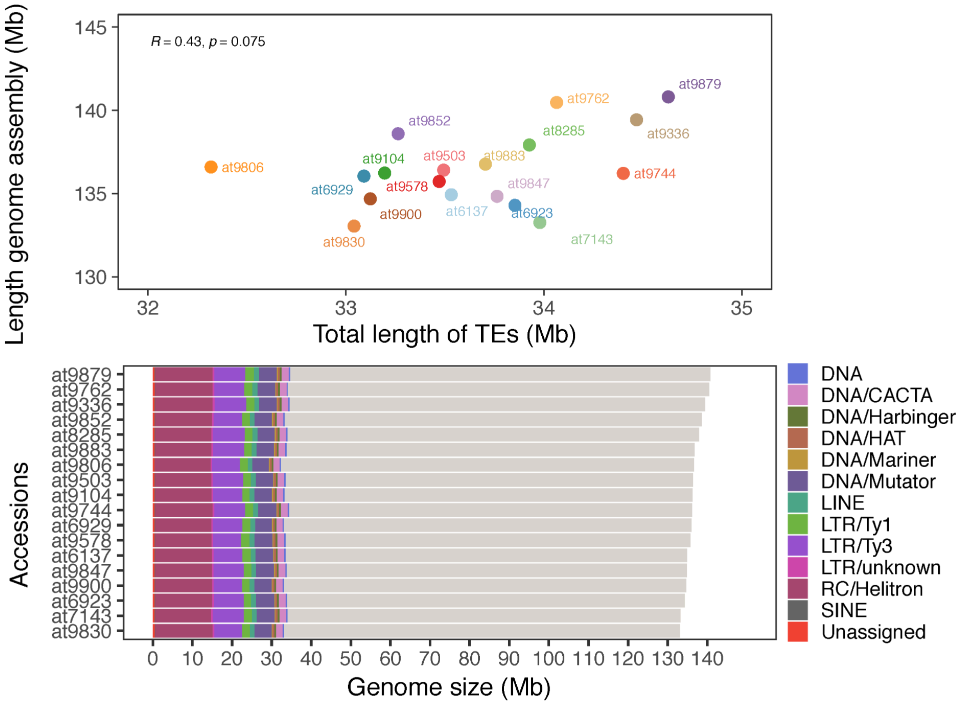


Figure 11 | Upper panel: Pearson correlation between the total length of assemblies and TE fraction. Bottom panel: Total TE content and its composition for each assembly.

Helitrons are the most prominent contributors to total genome size followed by LTR/Ty3 and DNA/Mutators, although the relative contribution of each TE superfamily is similar between accessions (Figure 11). Despite this, LTR/Ty3 presented the most variable fraction among accessions (Figure 12.A) compared with any other superfamily. After curation, I retrieved a set of 2,997 of *de novo* TE families, 818 of which were manually added after or clustering of orphan TEs; this is a significant increase compared with the set of original 319 TE families present in Col-0 Table 4.

Order	Superfamily	TE families
TIR	DTA	124
TIR	DTC	344
TIR	DTH	109
TIR	DTM	502
TIR	DTT	48
Helitron	Helitron	1262
Line	Unknown	4
LTR	Ty1	175
LTR	Ty3	303
LTR	Unknown	126

Table 4 | Classification of the *de novo* TE families.

Given that the 319 TAIR10 TE families were found in each accession, I asked how widely the newly defined TE families are shared. As we can observe in Figure 12.B, the distribution of new TE families presents a skewed tail, with 1,817 being shared across all accessions, and only 83 being private for a single accession.The fact that 316 TE families are shared among 2 or 3 accessions could suggest a recent admixture event between a subset of accessions. We can also observe that larger families tend to be more likely to be present in all accessions.

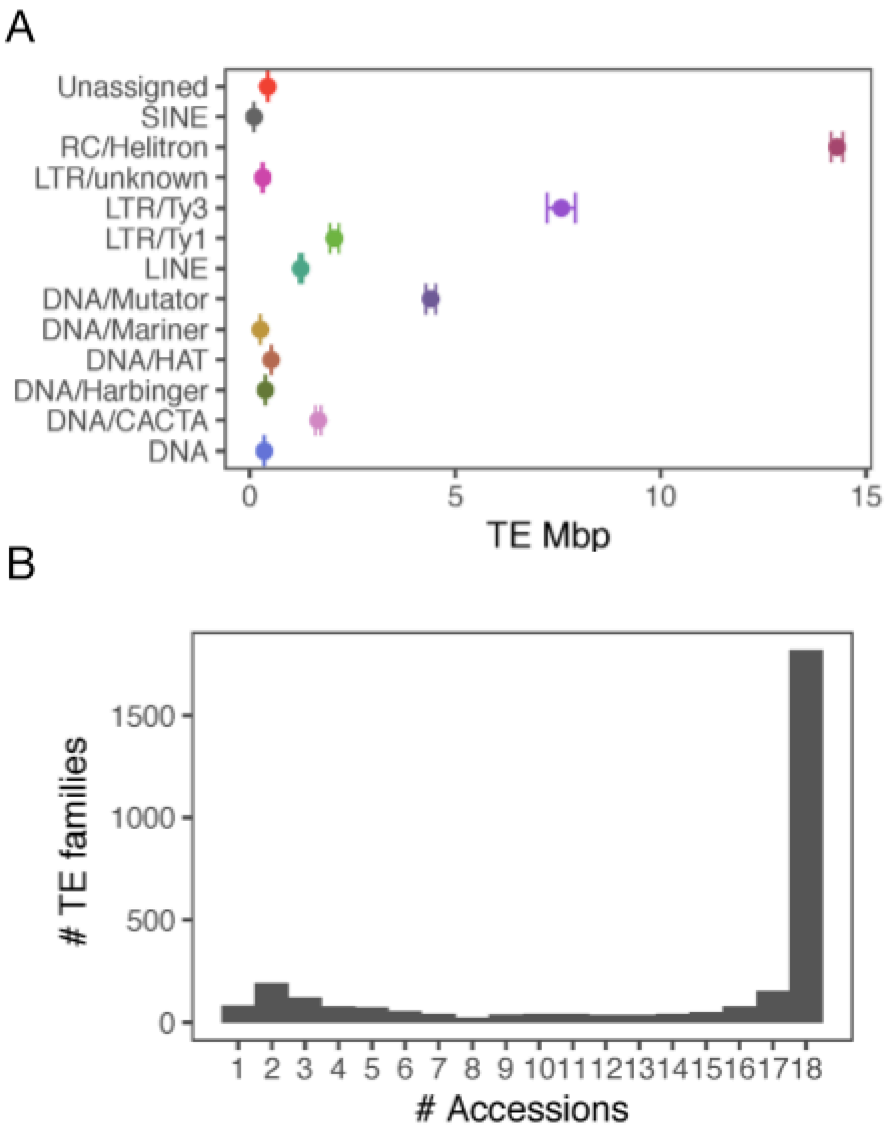


Figure 12 | (A) Average TE content of each superfamily in the 18 Differential Lines, with their corresponding standard deviation. (B) Histogram of number of TE families per accession.

Accession divergence driven by TE families.

Next, I wanted to compare the most abundant TE families in the Differential Lines with those in Col-0 TAIR10. The general tendency is that TE families with many copies in Col-0 are also abundant in this set. However, there are a few changes in the rank order (Figure 13). Although the number of copies is similar across accessions for most TE types, Helitrons turned out to be much more common in some accessions. For example, ATREP10D is the most common Helitron in the Differential Lines, with an average of 2,497 copies, compared to 1,295 in Col-0. The most numerous Helitron family in Col-0, ATREP3, with 1,439 copies, has on average 2,350 copies in the Differential Lines

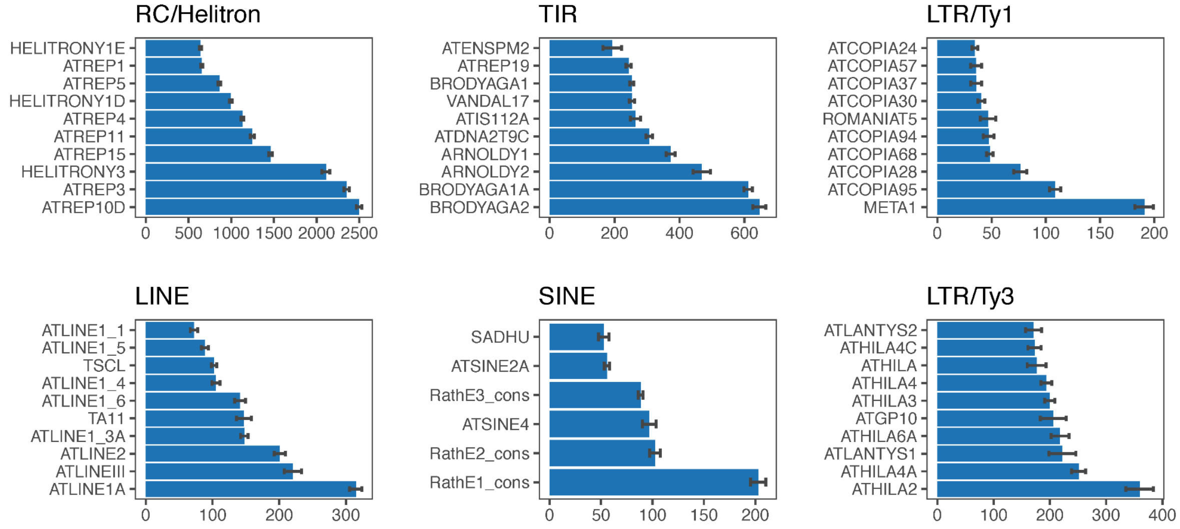


Figure 13 | Most abundant TE families per TE type in the Differential Lines. Error bars represent standard deviation.

I conducted a principal component (PC) analysis to better understand which TE families contribute the most to differences in TE content among the Differential Lines and which accessions are most distinct at the TE level (Figure 14.A). As we can see, accession at8285 has the most distinct TE content composition. Notably, the divergence between the Differential Lines was predominantly driven by two major superfamilies, LTR/Ty3 and TIR/Mutator. Within these superfamilies, specific TE families played significant roles in generating differences between accessions (Figure 14.B):

```
Top 1% PCA loadings:
Ty3/ATHILA2, DTM/VANDAL3, Ty3/ATHILA, Ty3/ATLANTYS1, Ty3/ATHILA6A, DTM/VANDAL2,
and Ty3/ATHILA3.
```

While none of these families was a *de novo* TE family, I wanted to know how the newly identified TE families drive differences in TE content among accessions. I therefore used again PCA, but this time subsetting to *de novo* TE families only (Figure 14.C). As expected, the differentiation was weaker (PC1 drives 8.3% of the variance compared with 20.1% in the PCA of all TEs), with the most divergent accessions for *de novo* TE families being at9879. Many *de novo* TE families that drive differentiation among accessions are RC/Helitrons (Figure 14.D):

```
Top 1% PCA loadings:
Helitron/TE_00001459, Helitron/TE_00000914, Helitron/TE_00001389,
Helitron/TE_cluster_381, Helitron/TE_00000955, Helitron/TE_00000902,
Helitron/TE_cluster_451,Helitron/TE_00002077, Helitron/TE_00000796,
Helitron/TE_cluster_446, DTM/TE_00004190
```

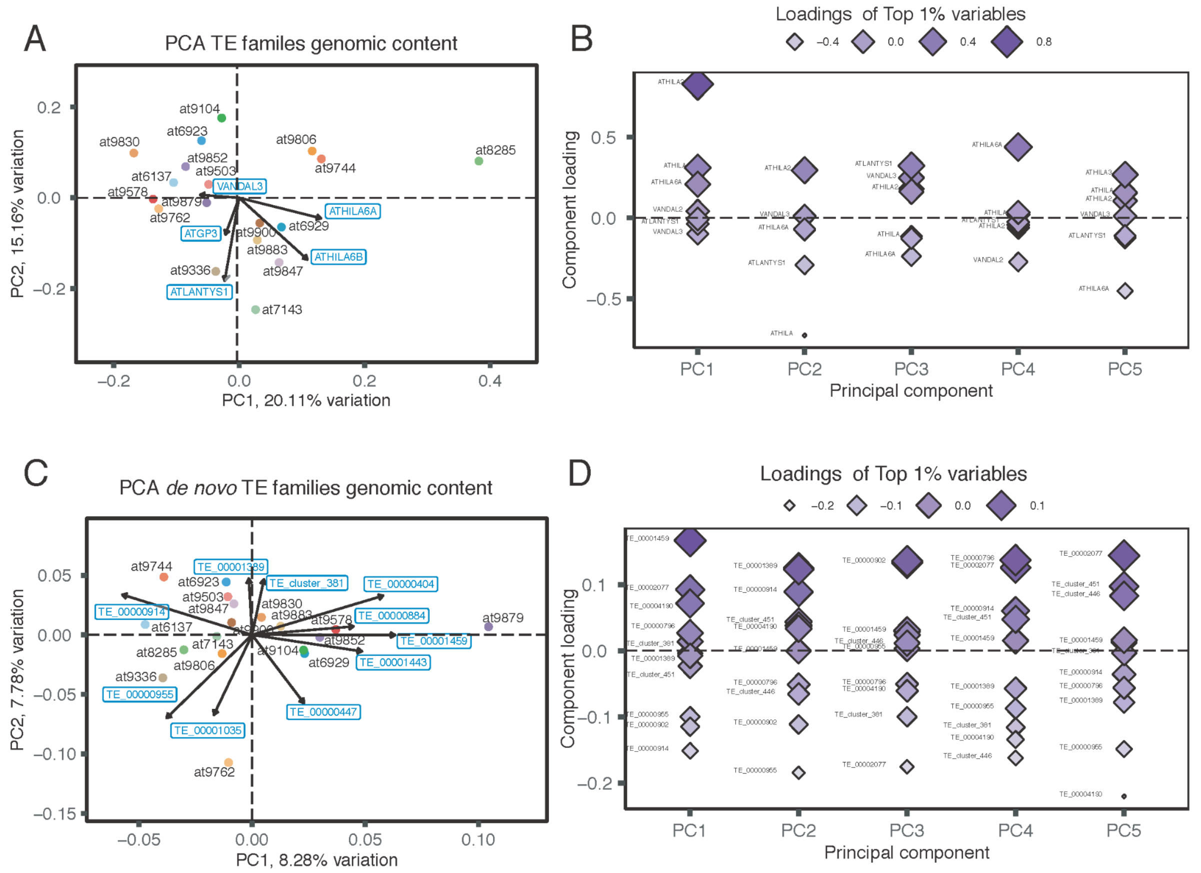


Figure 14 | Principal Component Analysis (PCA) of the contribution of each TE family to genome sequence of Differential Lines. The TE family matrix was computed based on the total TE content of each family across all 18 Differential Lines. (A) PCA biplot illustrating the two PCs representing the genomic content of all TE families. (B) Loading plot for the Top 1% TE families contributing to each PC. (C) and (D) PCA biplot and the corresponding loading plot for *de novo* TE families only.

Lastly, I wanted to assess the contribution to the genome sequence made by the different clades of Ty1 and Ty3 superfamilies. The most significant contribution comes from the TE families of the Ty3/Athila clade, followed by Ty3/Retand and Ty3/Tekay. For accessions at9104, at9762, and at9883, the ranking of the last two differs, with Ty3/Tekay ranking before Ty3/Retand. Shifting focus to the Ty1 superfamily, Ty1/Ale emerged as the most prominent contributor to the genome, ranking as the fourth most prominent globally, followed by Ty1/Bianca (Figure 15).

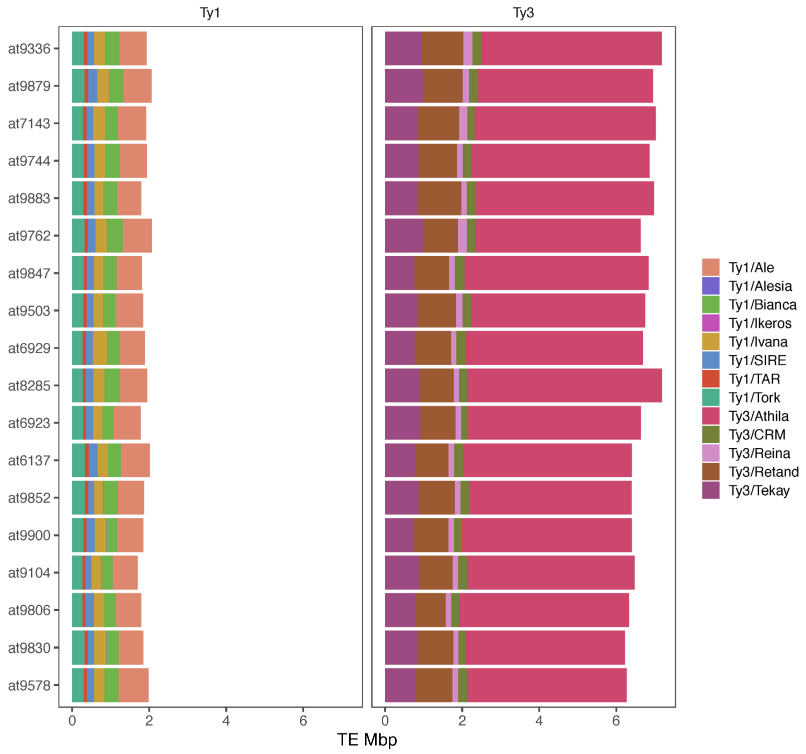


Figure 15 | Contribution of the different LTR clades to the genomes of the 18 Differential Lines.

The active TE fraction of *A. thaliana*

The genome assemblies contain both intact or fragmented TEs. TE fragments are considered relics or remnants of TEs that have lost capacity for transposition. Intact TEs refer to structurally complete elements that often are able to mobilize, either autonomously or non-autonomously. These intact TEs are considered to be the result of recent mobilization events, with insufficient time for decay or removal due to selection. As a result, intact TEs represent the most recent and likely active fraction of the genome. Thus, their analysis provides insights into ongoing TE dynamics in *A. thaliana*.

Autonomous TEs are TEs that contain opening reading frames with all the protein domains needed for their transposition, whereas non-autonomous intact TEs lack these domains and rely on hijacking autonomous TEs machinery to transpose. For an intact LTR TE, being autonomous means having an opening reading frame with GAG, PROT, INT, RT and RH protein domains. For intact TIR TEs, being autonomous means harbouring the TPase protein domain. Despite constituting a high fraction of the TE landscape in *A. thaliana*, I excluded Helitrons from these analyses as our understanding of their transposition mechanism is limited [273] despite some advances [125]. Thus, it is challenging to perform an *in silico* evaluation of their mobility.

In general, a remarkably similar number of TE copies, around 1,000, are classified as intact in the Differential Lines. The accession with the lowest intact TE fraction is at9104 with

1,007 copies. at6137 and at9744 have the highest intact fraction with 1,088 copies. Helitrons constitute the largest fraction of intact TEs in all accessions, followed by TIR TEs and LTR TEs (Figure 16.A), which is remarkable considering that RC/Helitrons are the most abundant TEs followed by LTR and TIR TEs (Figure 11).

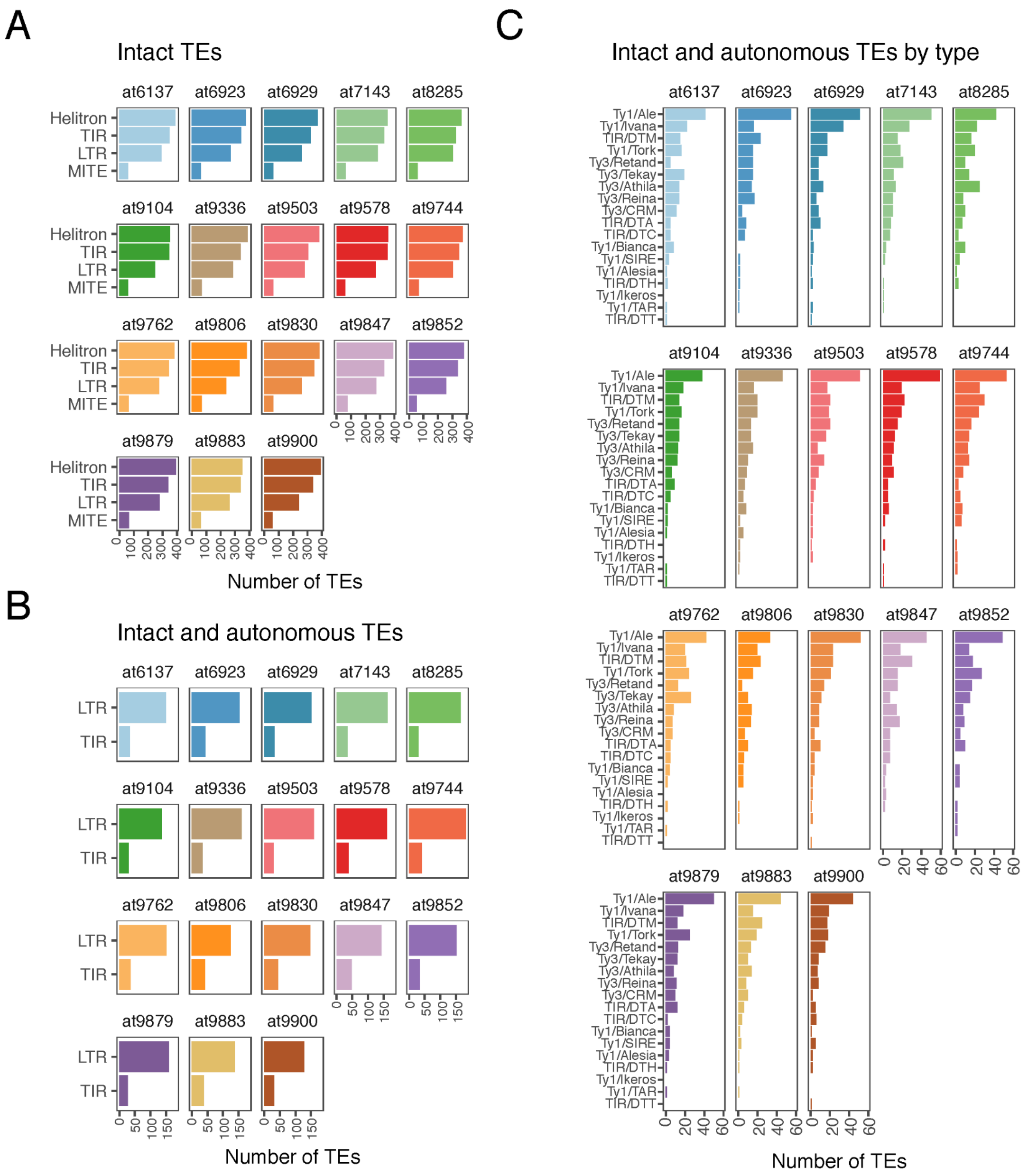


Figure 16 | Overview of the active TE landscape in 18 Differential Lines. (A) Intact TEs by order (with a differentiation between full TIR TEs and MITEs). (B) Autonomous copies for LTR TEs and TIR TEs. (C) Detailed breakdown of autonomous TEs by clade and superfamily.

However, when considering only a fraction of autonomous TEs, the picture changes, and autonomous LTR TEs are more abundant than autonomous TIR TEs. This is surprising, as

LTR TEs are more complex in nature when considering the number of protein domains they require to transpose, i.e., they should be more easily inactivated by mutation (Figure 16.B). Around half of intact LTR TEs are autonomous, whereas for TIR TEs only around a tenth of intact TEs are considered autonomous. Breaking down these autonomous TEs by superfamily in the case of TIR TEs and by superfamily and clade for LTR TEs, we see how the highest fraction of autonomous TEs are Ty1 (Ale, Ivana and Tork) followed by Mutator-like elements of the TIR/DTM superfamily (Figure 16.C).

This analysis reveals an interesting contrast between Ty3/Athila and Ty1/Ale TEs. While Ty3/Athila TEs contribute more to the overall genome (Figure 15), it is evident that most of the Ty3/Athila elements are degraded or remnants. In contrast, the Ty1/Ale clade stands out as the most active TE group across *A. thaliana* accessions, yet its relative contribution to the total genome sequence is much lower (Figure S1). These divergent patterns might be attributed to the distinct insertion and targeting mechanisms employed by each LTR clade. A direct comparison between the number of autonomous TE copies and total contribution to the genome is given in Figure 17. We can observe that in general, Ty3 elements contribute disproportionately to the genome. Ty3/Reina and Ty3/CRM appear to have a more Ty1-like pattern of expansion.

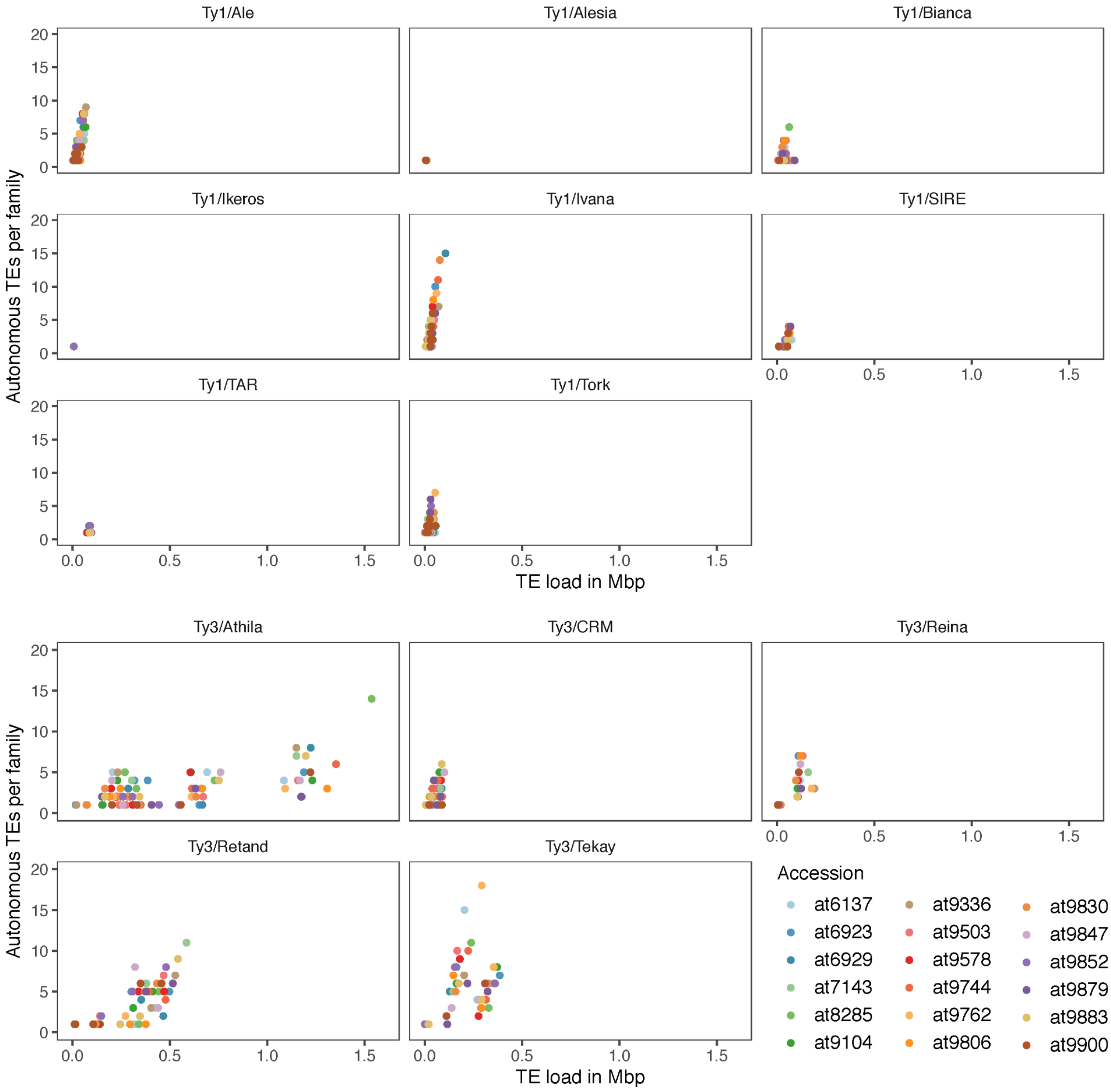


Figure 17 | TE load of LTR clades in the Differential Lines and its relationship to the number of autonomous TEs for each clade.

To evaluate patterns of expansion and contractions of the different autonomous TEs among accessions, I calculated a z-score for each type of autonomous TE (Figure 17). We can see how in general, accessions tend to have unique profiles, reflecting their general divergence. at9744 stands out with above-average copy numbers of all LTR TE clades, and at9104 stands out with below-average copy numbers for almost all LTR (Figure 18.A) and TIR TE clades (Figure 18.B). For the three Ty1 clades that lack any copy in the TAIR10 annotation (Table 3), Ty1/Angela had no single autonomous copy, and Ty1/Alesia and Ty1/Ikeros were rare. Both Ty1/Alesia and Ty1/Ikeros were overrepresented in at9336.

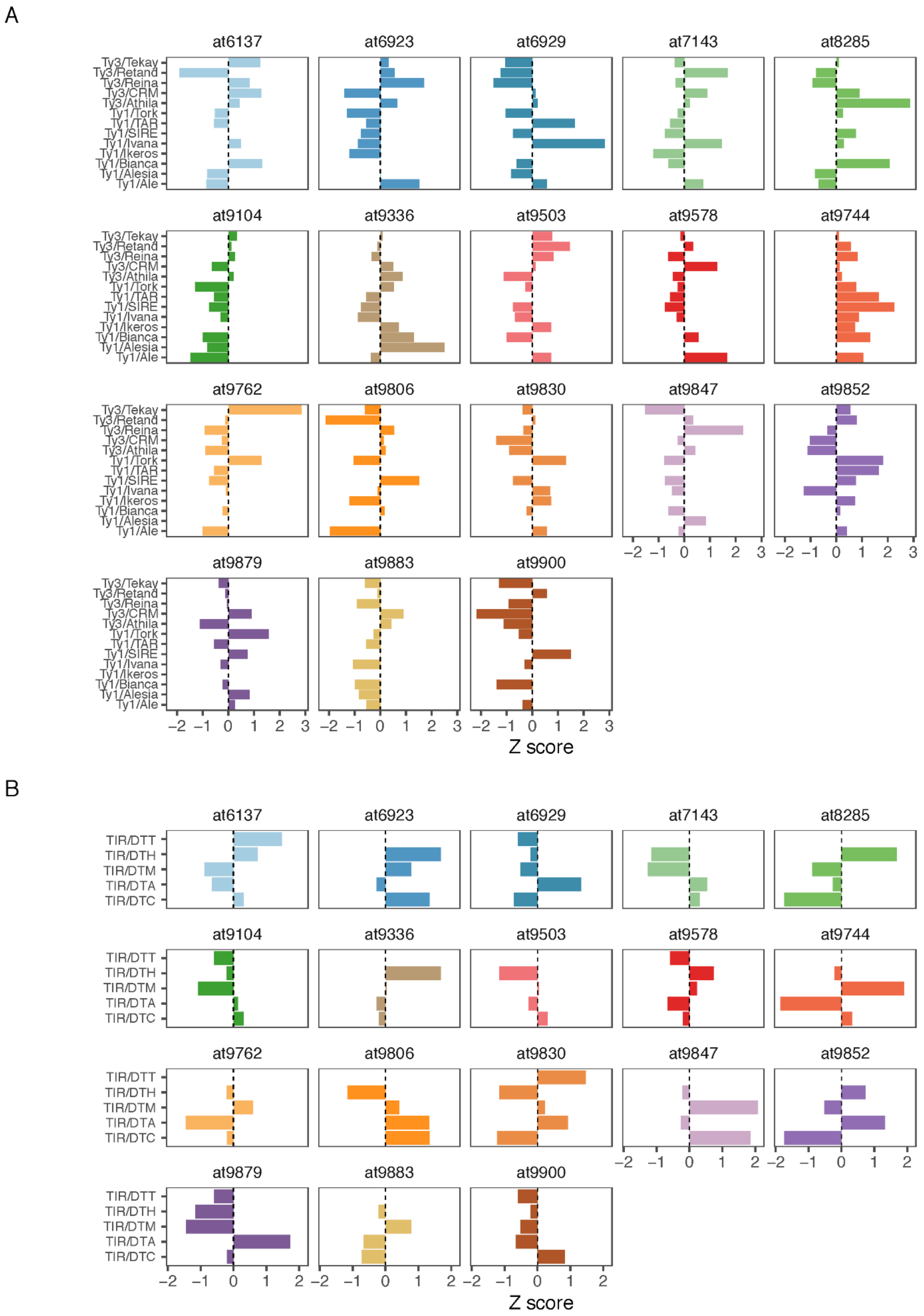


Figure 18 | Expansions and contractions of autonomous TEs in the 18 differential lines. (A), z-scores of copy numbers for the different LTR TE clades. (B) z-scores of copy numbers for the different TIR TE superfamilies.

Age of TE landscape in *A. thaliana.*

In most genomes, the majority of TEs is old and riddled with mutations, with only a few, occasionally active copies. Assessing the insertion time, even with rough estimates, can inform on which aspects of genomic location patterns are due to purifying selection, which acts to remove the TE insertions, and which aspects are due to TE insertion preferences.

To estimate LTR age of *A. thaliana* accessions, we used the formula T = K/2 × r, where T = time of divergence, K = divergence of the LTR region and r = substitution rate [274]. For the mutation rate of *A. thaliana*, we used the estimate of $7x10^{-9}$ base substitutions per site per generation [248]. This is currently the most accurate way to calculate the age of a given TE copy, but it is only applicable to a subset of LTR TEs with identifiable LTR regions [275].

I was able to calculate the LTR age of 4,866 intact LTR TEs (2075 Ty3 elements, 2,054 Ty1 elements, and 747 unknown LTR TEs) (Figure 19.A). As we can see in the figure, most accessions have primarily evidence for recent transposition for Ty1 elements, but Ty3 is not very different. The trends observed for Ty1 and Ty3 superfamilies are, mostly, a reflection of the insertion times of the most common superfamilies, Ty1/Ale and Ty3/Athila (Figure 19.B).

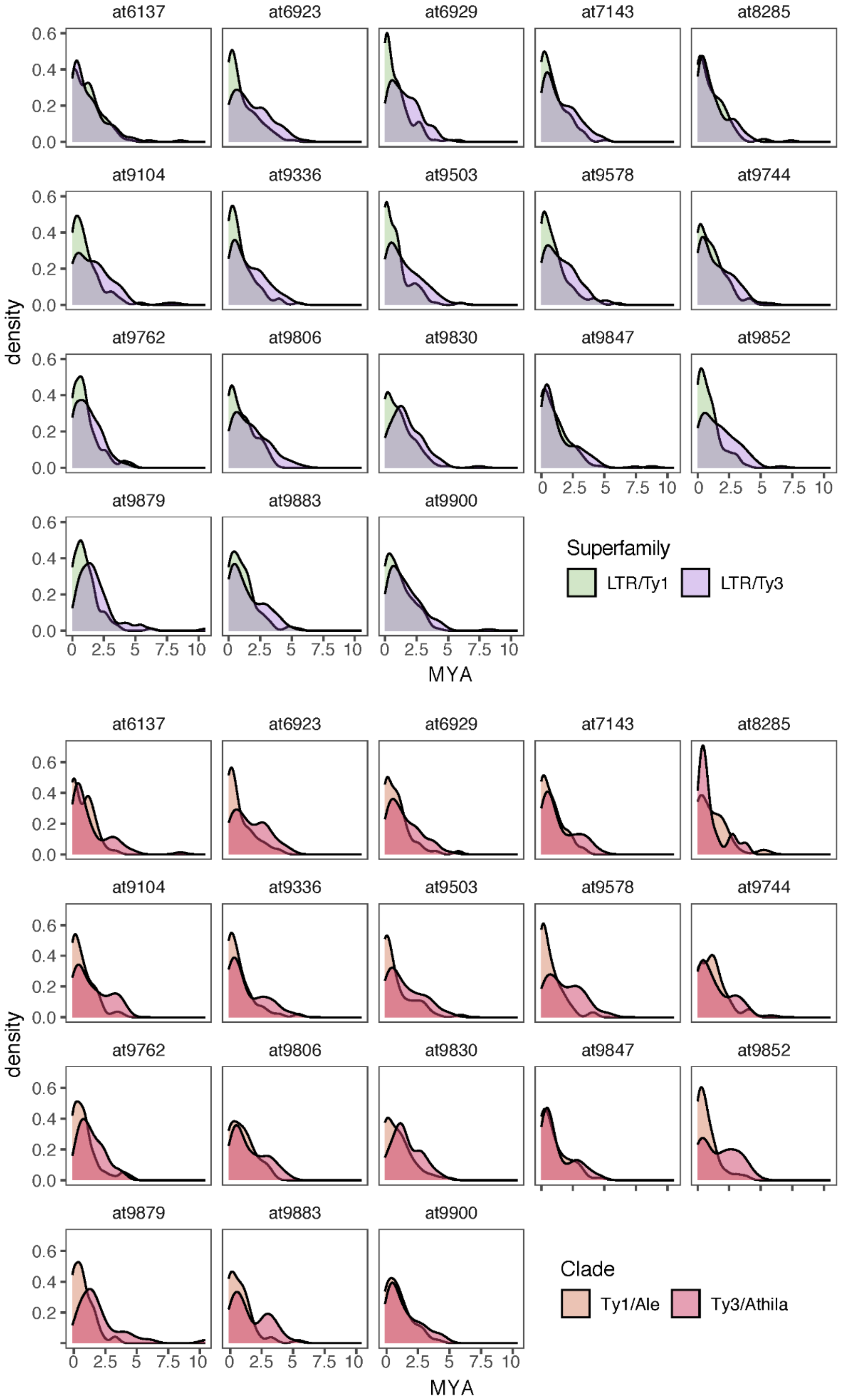


Figure 19 | LTR age estimation. (A) For the complete Ty1 and Ty3 LTR superfamilies. (B) Only for Ty1/Ale and Ty3/Athila clades.

For incomplete LTR TE copies and for TIR TE copies, I used the percentage of identity to the TE model to estimate their age [30,265,276]. The reasoning is that higher identity scores imply fewer mutations among TE elements within the same family, given that TEs are expected to be identical copies of their parent right after insertion, with fewer mutations indicating less accumulated time of divergence from their ancestor. Thus, in order to calculate an estimation of the age of a TE copy, we resort to the Kimura 2-parameter distances (K-values) [277].

I had estimates for both percentage identity to the respective TE model and age of the copy for 4,608 LTR TEs, and found them to be strongly correlated (R=-0.98, p< 2.2e^{-16}) (Figure 20). This supports the assumption that identity with the respective TE model is a good proxy for TE age.

There were, however, a few outliers. One reason could be that when comparing the two LTRs of a TE copy, we only account for single nucleotide polymorphisms and ignore any indels, thus underestimating the true age of the copy. Another related reason could be that the TE model we compared a given copy with, lacks the polymorphic sites used for LTR age estimation, overestimating the similarity.

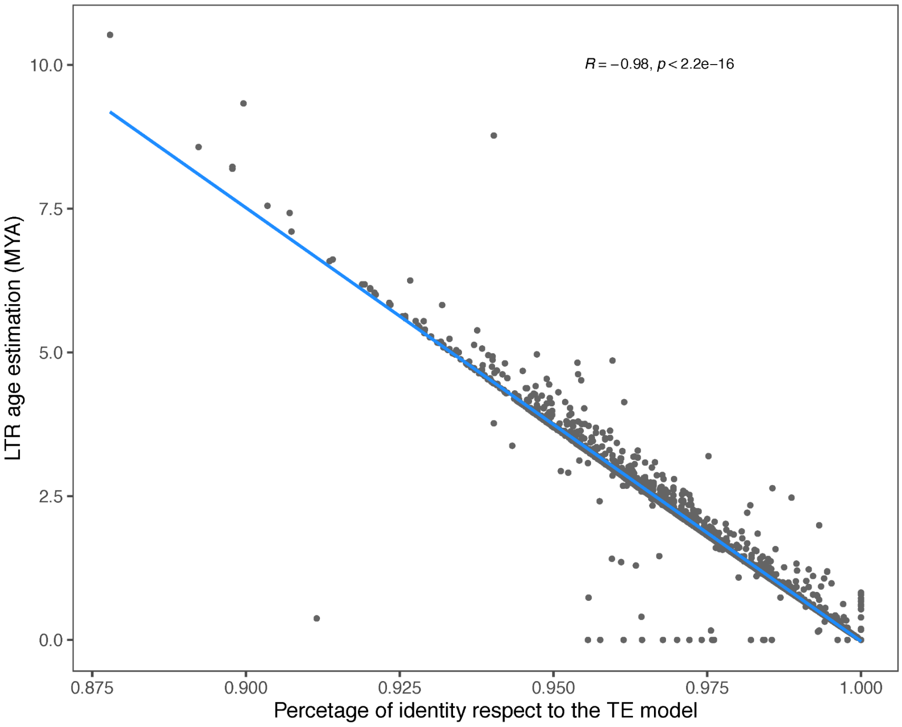


Figure 20 | Correlation between the percentage of identity to the TE model of each complete LTR copy with the age estimate based on LTR-LTR distance for 4,608 complete LTR TE copies.

The TE age landscape in the 18 accessions presents a skewed pattern (Figure 21.A and Figure S2) towards recent events, with some peaks at 0.9 and 0.85 identity. This could have been a product of biases in the generation of TE models, but, when comparing identity percentages from copies annotated with the TAIR10 library and with *de novo* TE models, I observed a similar pattern (Figure 21.B).

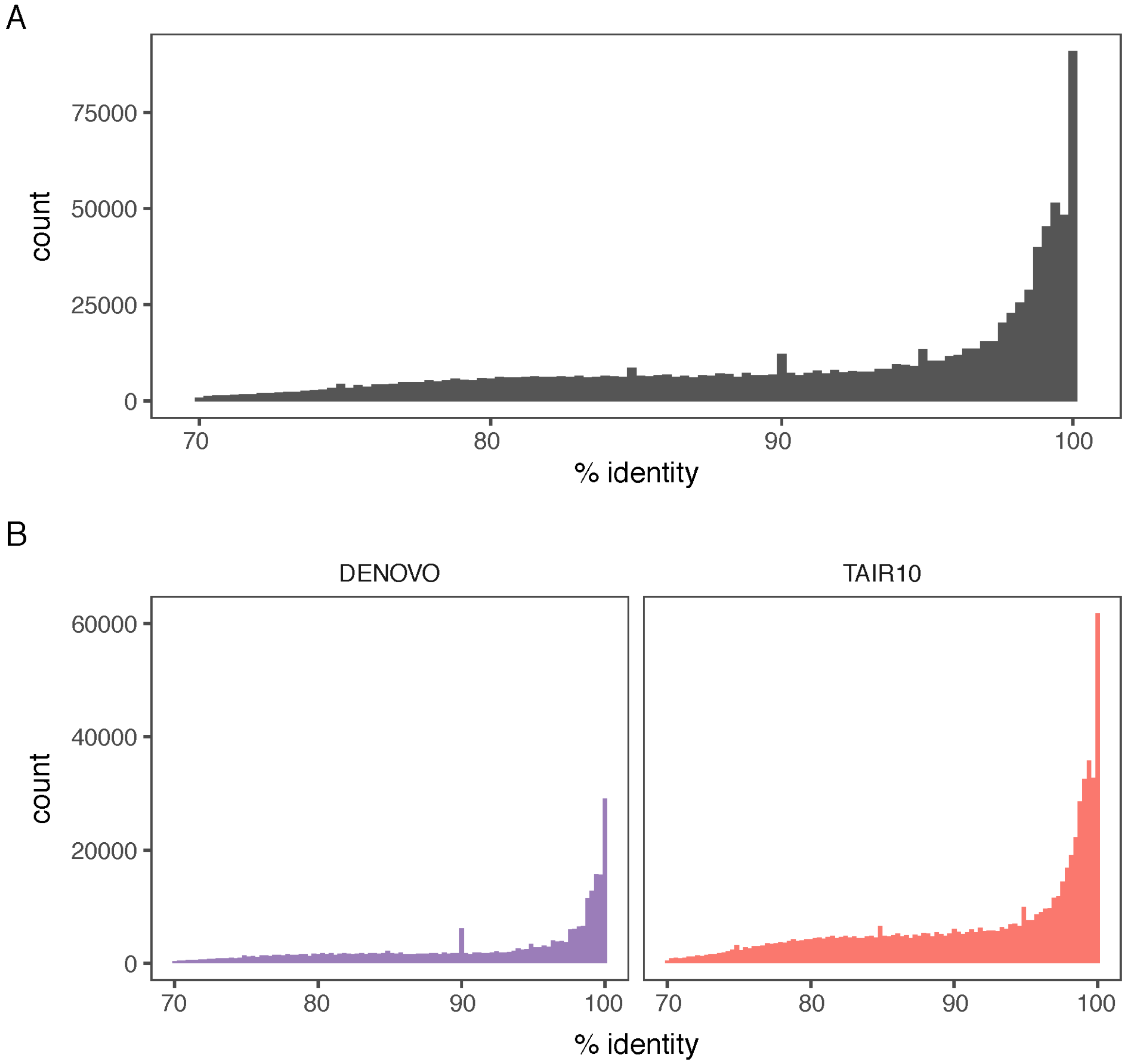


Figure 21 | TE identity landscape in the Differential Lines. (A) Distribution of model identity for all TE copies across the different *A. thaliana* accessions. (B) Distribution of model identity for all TE copies split between *de novo* and TAIR10 annotated copies.

To better understand the age patterns, I decided to classify each TE family into recent, young, moderate, and old based on the average percentage of identity within the family (Table 5).

Category	Average identity	Number of TE families	Number of TE copies
Recent	0.99-1	849	11844
Young	0.95-0.99	1099	117544
Moderate	0.85-0.9	1855	786718
Old	0.7-0.85	250	41390

Table 5 | Binned TE families in different age categories based on the family average identity.

The composition of these age categories was uneven between TE types. Unassigned TE families were classified entirely as “Old” together with SINE families, while Helitron TE families were evenly distributed, albeit their proportion increased in the “Recent” category. Families belonging to both TIR/Mutators and LTR/Ty3 were mostly of intermediate age. Ty1/Copia families were enriched in the “Young” category, which is in line with most potentially active TEs belonging to Ty1/Copia. TIR/CACTA families had an unusual pattern, being enriched in “Recent” and “Old” categories but depleted in the “Young” category, which separates the former two categories. This may pinpoint to two different events of CACTA bursts (Figure 22).

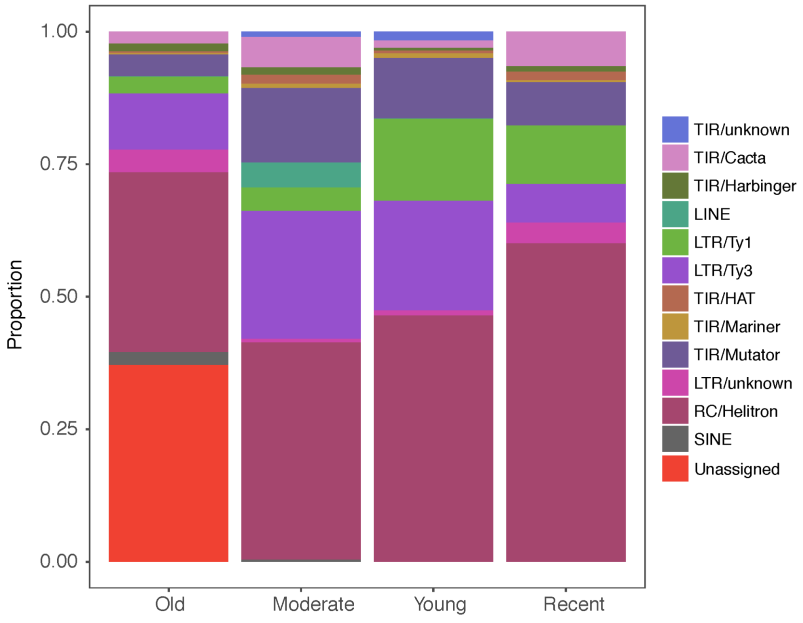


Figure 22 | Contribution of TE types to different age groups, based on the total length of each TE family.

To gain more insight into how the age of the different TE families might be related to the overall diversity within the Differential Lines, I conducted several principal components analyses, one per age category (Figure 23). Drivers of diversity in “Young” and “Moderate” intermediate categories are TE families also present in TAIR10 Col-0, whereas differences between accessions in “Old” and “Recent” categories are *de novo* TE families.

```
Top 1% PCA loadings for “Old” category (number of accessions present):
Helitron/TE_00000902 (18), DTC/TE_00000404 (5), Helitron/TE_00001443 (18),
Helitron/TE_00002077 (18), Helitron/TE_0000055 (18), Helitron/TE_00003044 (18),
Helitron/TE_00001035 (18).
```

```
Top 1% PCA loadings for “Recent” category (number of accessions present):
DTM/TE_cluster_443 (1), Helitron/TE_00001460 (18), Helitron/TE_00001535 (17),
Helitron/TE_00001119 (6), Helitron/TE_00000663 (18), LTR/unknown/TE_00001917 (13),
Helitron/TE_00000428 (18), Helitron/TE_00001296 (18), Helitron/TE_00001 (18).
```

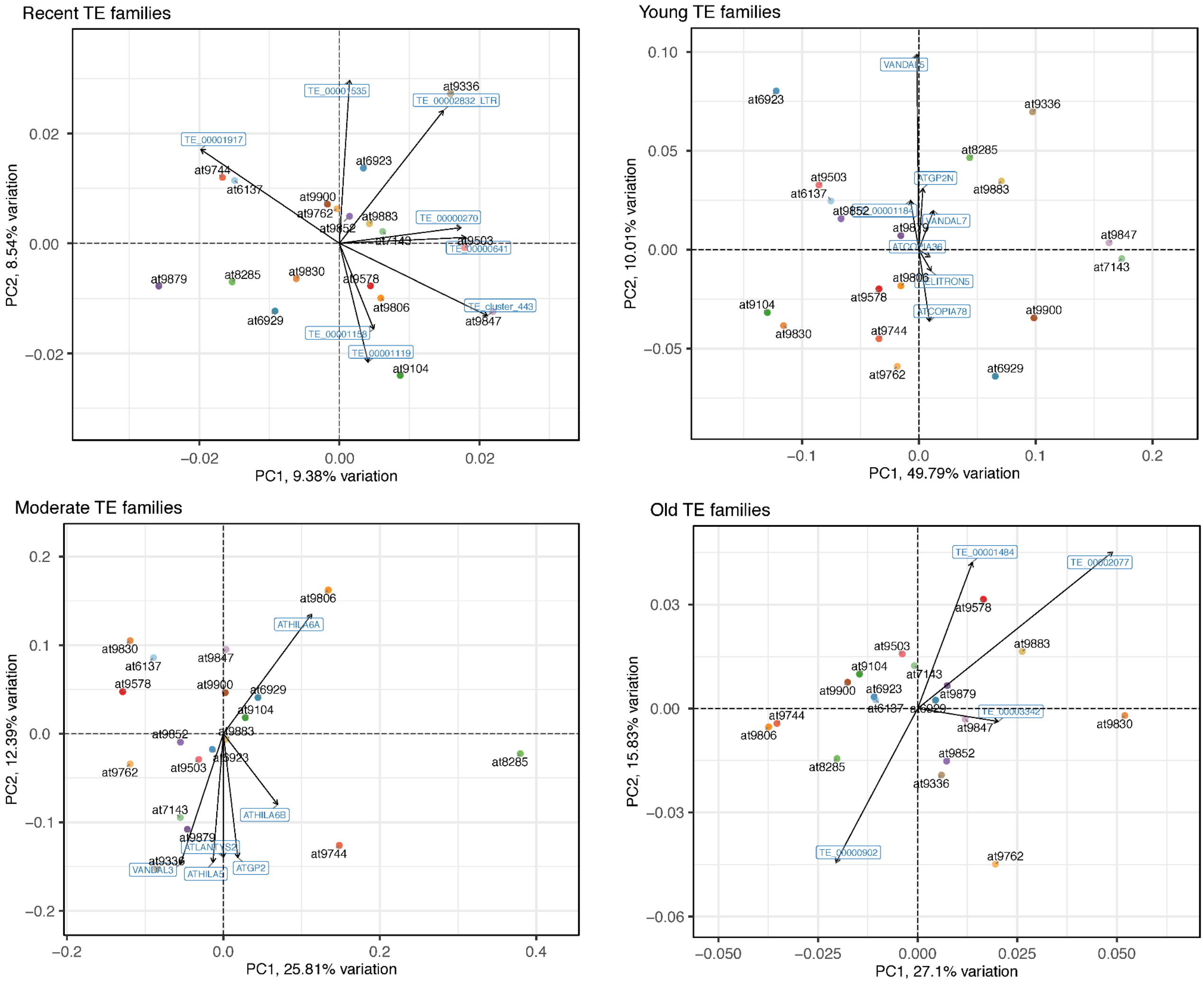


Figure 23 | Contribution by different TE age groups to the genomes of Differential Lines.

A closer examination of the “Recent” TE families revealed that most copies were small fragments of TE remnants. Long TE copies were present at much lower frequencies across accessions. Thus, an explanation for this observation is that differences in TE load for TEs of the “Recent” category are driven not by the unequal activity of different TE families, but by the unequal removal of TE fragments from the genomes.

To summarise, I find that in the Differential Lines, LTR activity peaked 0.5-1 Million years ago (MYA). This peak was mainly driven by LTR/Ty1 Ale elements. LTR/Ty3 Athila elements also peaked at the same time but generally in a less pronounced manner. However, most accessions show more activity at prior ages for LTR/Ty3 Athila, with some peaks around 2.5 MYA. Taking at face value the correlation between absolute LTR age and identity (Figure 20), these peaks correspond to “Young” and “Moderate” TE families, where LTR/Ty1 and LTR/Ty3 are the drivers of TE load differentiation (Figure 23), and few families explain more of the TE load differences among accessions, as seen by the strength of the PCs in Figure 23. In the last 500,000 years, there appears to have been a steep decrease in LTR TE activity and variation among accessions for “Recent” TEs is more spread out among TE families as seen again by the PC loads.

4.3 TE dynamics within NLR clusters.

TE-driven NLR Evolvability.

As we have mentioned in the introduction of this chapter, NLR evolution can be influenced by the local composition of TEs in the genome. Inspecting the spatial relationships between NLR genes and TEs and comparing them to other genes and TEs, we can formulate hypotheses of the processes responsible for the observed relationships between NLR genes and TEs. First, we need to define those NLR gene clusters across each accession and define the syntenic relationships between common NLR gene clusters among accessions. NLR gene clusters often differ greatly in copy number and arrangement of NLR genes [178].

To enable comparison of NLR gene clusters, Dr. Shirsekar built a pangenome graph of the Differential Liness using the PGGB pipeline [278]. In a pangenome graph, nodes represent sequences, and sequence variation is represented as a network of nodes. He defined NLR neighbourhoods as regions in the pangenome that contained at least one NLR gene and were anchored by nodes on either 5’ and 3’ side of the NLR/s that (i) contained at least 100 bp sequence, and (ii) were non-repetitively present across all 18 accessions. This method of anchoring with conserved, shared nodes identifies broad-range synteny across NLR neighbourhoods and it does not require that every accession has an NLR gene in a specific NLR neighbourhood. This way, we can capture the evolutionary processes that resulted in the presence/absence of NLR genes in a given accession (For an example, Fig. 3.24). Dr. Luisa Tesadale conducted an annotation of NLR genes using iso-seq transcriptomic data in

17 out of our 18 genomes, thus, from now own I will focus only in the 17 genomes with a NLR annotation, dropping the genome of at6137.

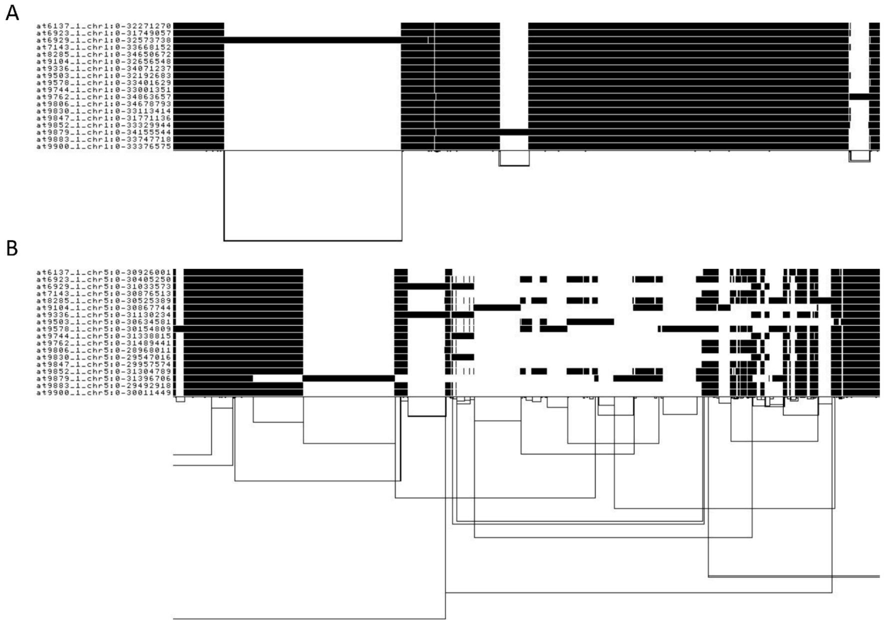


Figure 24 | Two NLR neighbourhoods as defined by the genome graph. (A) Highly syntenic neighbourhood, with a single long insertion for at6923 and two other small insertions for at9879 and at9762, respectively. (B) Highly fragmented neighbourhood with few common syntenic regions.

In total, 126 NLR neighbourhoods were characterised, mostly with a single loci per accession. Upon inspecting the size of these neighbourhoods, the number of total genes and NLRs, we saw that the neighbourhoods on chromosome 2 were much larger than elsewhere and contained relatively few NLR and other genes for their size (Figure 25). A closer look at these loci revealed that most of the makeup of the neighbourhood "chr2_e_r1_nh03" was centromeric in nature. Due to the poor synteny across centromeres, the neighbourhood definition had caused this specific neighbourhood to cover most of the centromere until it reached a syntenic node on the other side. Thus, we decided to remove this neighbourhood from the analysis to avoid skewing the overall results.

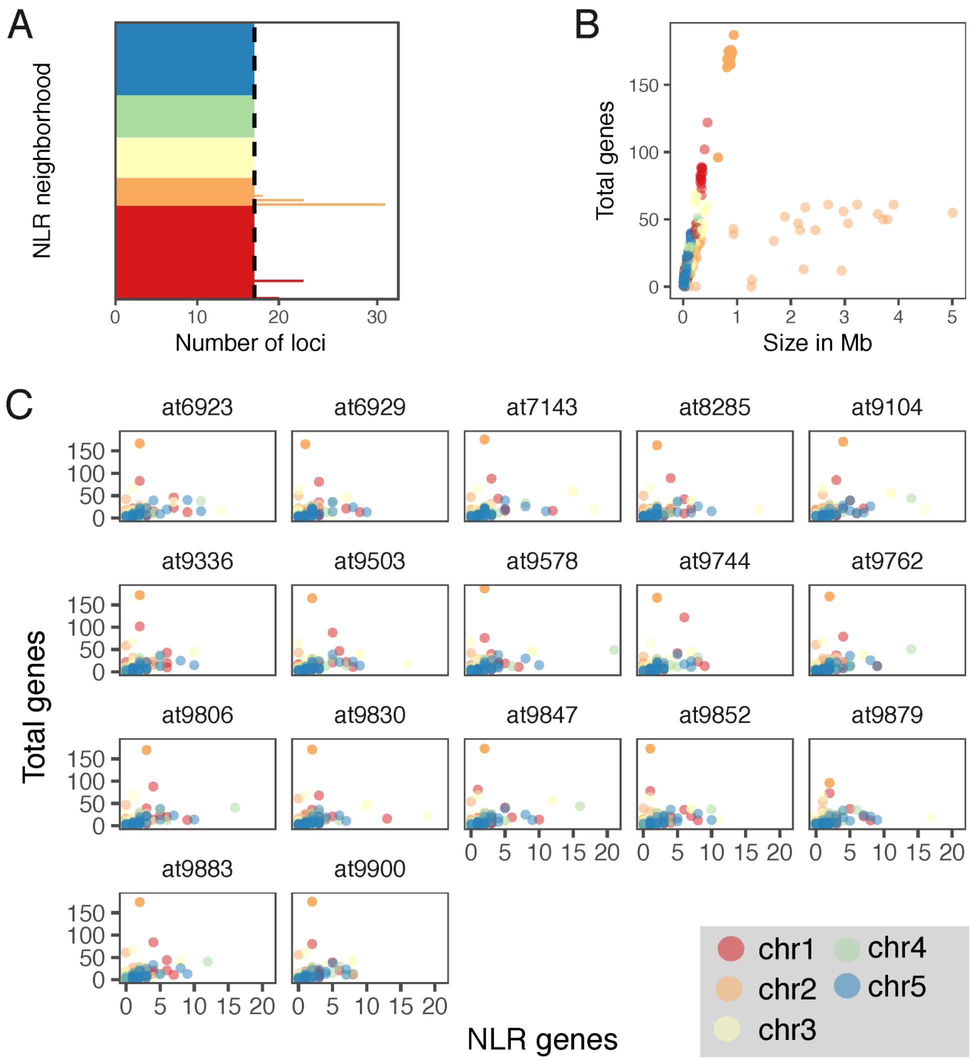


Figure 25 | Overview of the NLR neighbourhoods per chromosome across the pangenome. Colours represent which chromosome the neighbourhood belongs to. (A) Number of loci per NLR neighbourhood across the pangenome, black dashed line marks n=17, the number of accessions in the pangenome. (B) Comparison of the number of genes in a neighbourhood and neighbourhood size. (C) Comparison of the number of total genes and the number of NLR genes in these neighbourhoods.

TE dynamics at NLR neighbourhoods.

The TE composition of NLR neighbourhoods is highly variable, but in general, the number of TE copies correlates with the total size of the neighbourhoods (correlation coefficient: 0.73), albeit only relatively weakly when compared to the correlation between the number of genes and neighbourhood size in bp (correlation coefficient 0.98), (Figure S3).

Looking at the type of these TEs, it becomes evident that neighbourhoods on chromosome 2 harbour a significantly larger quantity of TEs. Additionally, their composition closely mirrors the overall TE distribution in the pangenome. Nevertheless, as the number of TEs within neighbourhoods decreases, distinct dominant TE superfamilies emerge, particularly CACTA and LINEs (Figure 26.A), in some neighbourhoods. The average TE diversity, measured using the Shannon entropy index, at the NLR neighbourhoods is lower than in the rest of the chromosomal arms (Figure S4). This suggests that NLR neighbourhoods are preferred by certain TE families.

To formally test this observation, we used Fisher's exact test for each TE family with copies within the neighbourhoods. In total, we tested 1,036 TE families for 2,375 neighbourhoods. After correcting for multiple testing and with a p-value of 0.005, we retrieved 52 TE families enriched at these loci, most of them Helitron families. Notably, none of the three most mobile TE families in *A. thaliana* (ATCOPIA93 aka Evadé; ATENSPM3; and VANDAL21) [129] were identified as enriched for the NLR neighbourhoods. However, many of the enriched TE families were enriched across multiple neighbourhoods (Figure 26.B). Additionally, most of the neighbourhoods exhibiting TE family enrichment were found to be enriched in more than one distinct TE family. Finally, out of the 126 total NLR neighbourhoods, only 37 showed at least one TE family enriched.

```
NLR neighbourhoods that show enrichment in at least one TE family:
chr1_nh03, chr1_e_r4_nh01, chr1_nh09, chr1_e_nh01, chr1_nh18, chr1_nh19,
chr1_nh20, chr1_nh23, chr1_nh24, chr1_nh25, chr1_nh31, chr1_nh32, chr2_e_r1_nh01,
chr2_e_r1_nh02, chr2_e_r2_nh01, chr2_nh01, chr2_nh02, chr2_nh04, chr3_nh09,
chr3_nh10, chr3_nh11, chr3_nh12, chr4_nh01, chr4_nh05, chr4_nh06, chr4_nh07,
chr4_nh12, chr5_nh06, chr5_nh11, chr5_nh14, chr5_nh16, chr5_nh18, chr5_nh21,
chr5_nh22, chr5_nh25, chr5_nh26, chr5_nh33.
```

```
TE families that show enrichment in at least an NLR neighbourhood:
ARNOLD3, ARNOLDY1, ARNOLDY2, ATDNA2T9C, ATENSPM1A, ATENSPM2, ATGP10, ATHILA,
ATHILA0_I, ATHILA2, ATHILA3, ATHILA4, ATHILA4A, ATHILA4C, ATHILA6A, ATHILA8A,
ATHILA8B, ATHPOGON1, ATLANTYS1, ATLANTYS2, ATLANTYS3, ATLINE1A, ATREP1, ATREP10A,
ATREP10B, ATREP10D, ATREP11, ATREP11A, ATREP15, ATREP19, ATREP3, ATREP4, ATREP5,
ATREP6, ATREP9, BRODYAGA1A, BRODYAGA2, HELITRON1, HELITRON2, HELITRON4,
HELITRONY1A, HELITRONY1B, HELITRONY1D, HELITRONY1E, HELITRONY3, RathE1_cons,
TE_00002236_INT, TNAT1A, VANDAL17, VANDAL4, VANDAL6, VANDAL9.
```

NLR neighbourhoods that showed enrichment of TE families were markedly different than those that did not. These NLR neighbourhoods tended to be more variable in terms of size, as well as harbouring more TE copies and more NLR genes (Figure S5). However, a subset of non-TE family enriched NLR neighbourhoods from chromosome 3 showed a different behaviour: they were small, but had a high density of NLR genes.

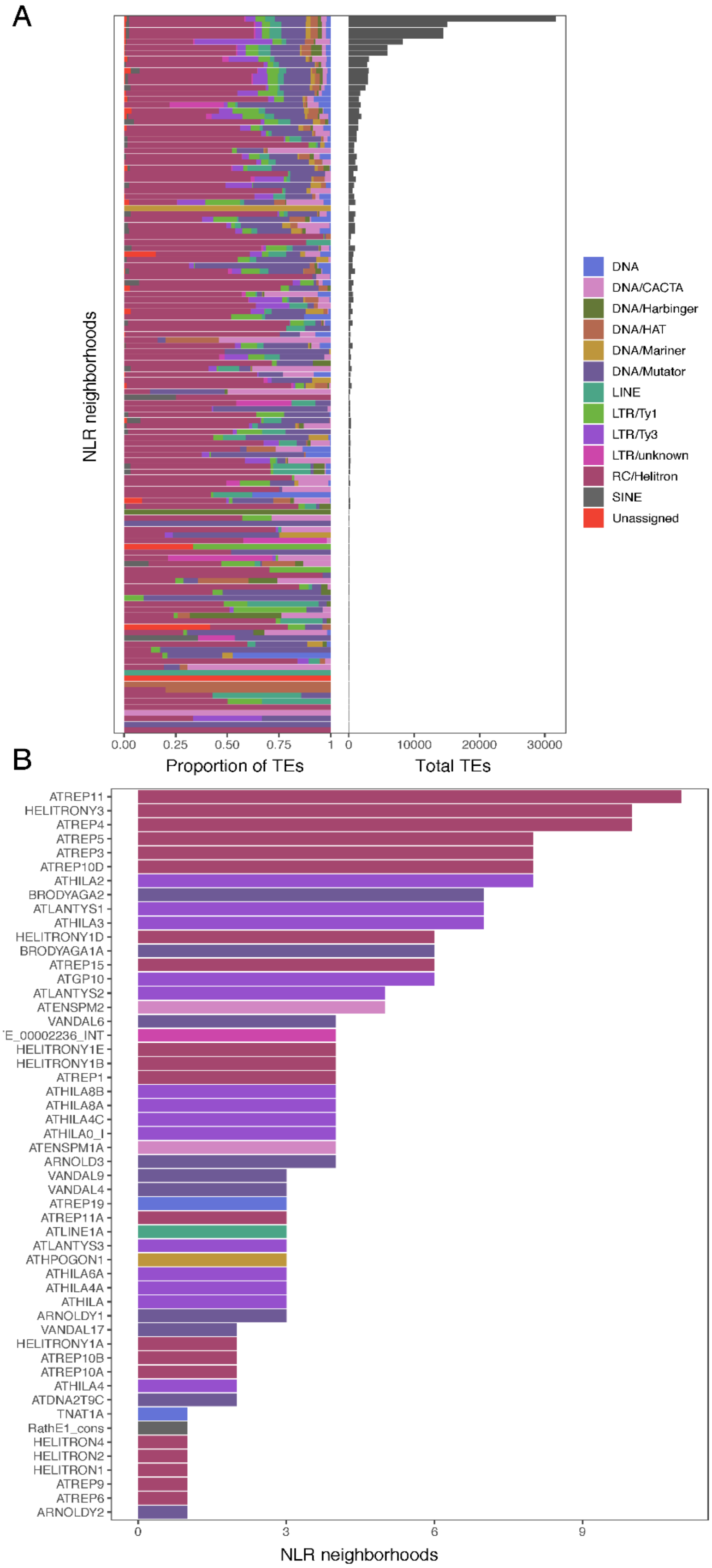


Figure 26 | TE profiling at NLR neighbourhoods. (A) Left, proportional TE composition of NLR neighbourhoods; Right, total number of TE copies at NLR neighbourhoods. (B) TE families that show enrichment in several NLR neighbourhoods.

Finally, we looked at the age of intact LTR TEs in the NLR clusters of all accessions and compared those with the age of LTR TEs in the rest of the chromosome arms. 1,057 LTR TEs reside in the NLR clusters, with a combined size of around 112.18 Mbp. The number of TEs in the rest of the chromosome arms, which span a total of 2313.98 Mbp, is about ten times higher, 11,861. Therefore, there is not an apparent enrichment of the number of intact LTR TEs at the NLR clusters. However, when we compare the age profile of NLR neighbourhoods against chromosome arms, we find that intact LTR TEs in NLR neighbourhoods are younger than in the rest of the genome (Figure 27.A), and this difference appears to hold for all accessions except at7143, at9762 and at9900 (Figure 3.27.B).

Is this apparent increase in younger, intact LTR TEs a consequence of an elevated insertion rate at these loci or a byproduct of the more rapid removal of LTRs by illegitimate recombination?

To answer this question, we used a custom method to retrieve “solo” LTR elements and measured the density of these solo LTRs across the pangenome, as these solo LTRs are proposed to be a product of illegitimate recombination events [279,280]. In total we retrieved 11,363 elements across the pangenome, with 871 ATHILA2 elements (Figure S6). When we compared the accumulation of solo LTRs in NLR neighbourhoods against the chromosome arms (Figure 28), we found more solo LTRs in the NLR neighbourhoods. Thus, it is possible that in the NLR neighbourhoods, illegitimate recombination rates are higher than the rest of the chromosome arms, producing a rapid loss of intact LTR TEs compared with the rest of the chromosome, which could also explain the lack of intact LTR TEs older than 5.5 MYA (Figure 3.27) in the NLR neighbourhoods. This suggests again that in *A. thaliana*, recent TE dynamics are dominated by host-driven TE removal.

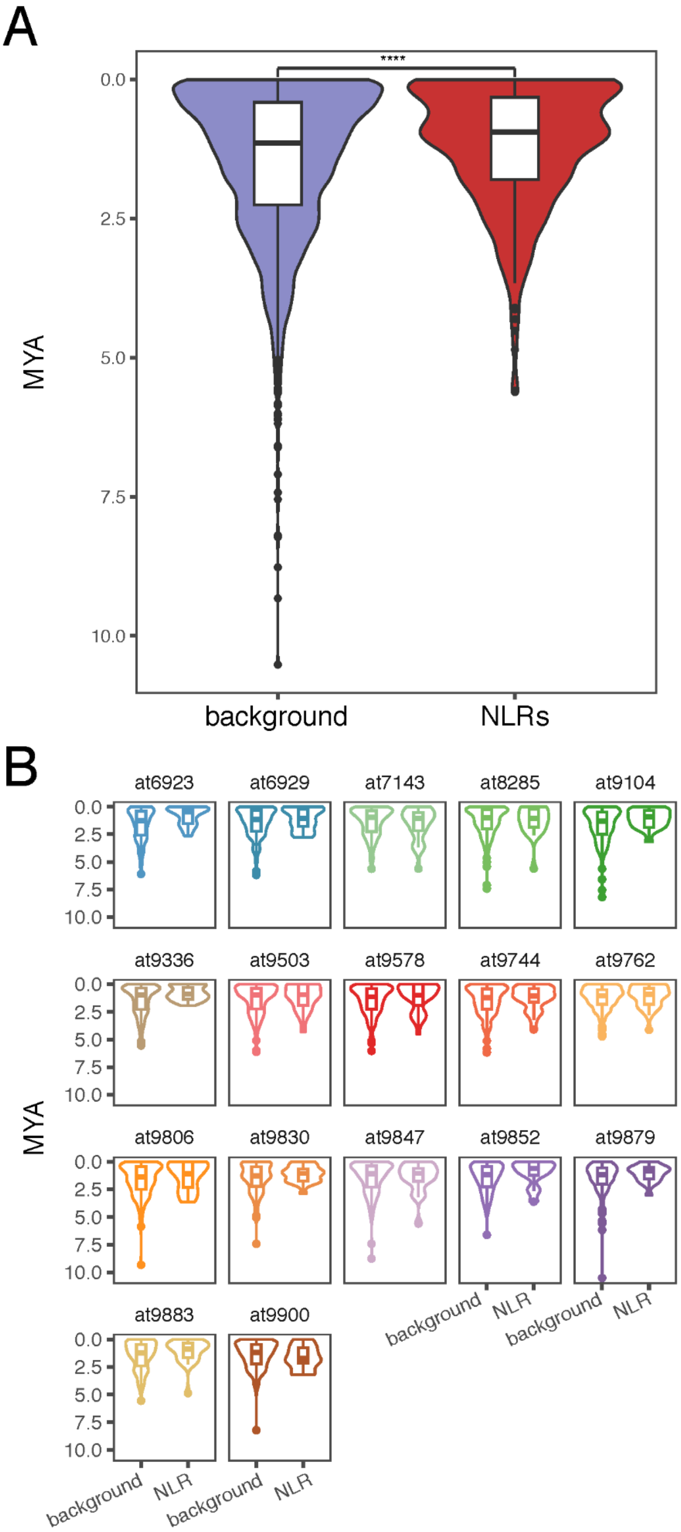


Figure 27 | LTR age comparison between NLR neighbourhoods and genomic background. (A) LTR age comparison between intact LTR TEs in background chromosome arms and in NLR neighbourhoods. Statistical comparison was done using a t-test (p-value = 2.38e-14). Age mean in chromosomes 1.451; Age mean in NLR neighbourhoods 1.178. B) LTR age visualisation across all the 18 different accessions.

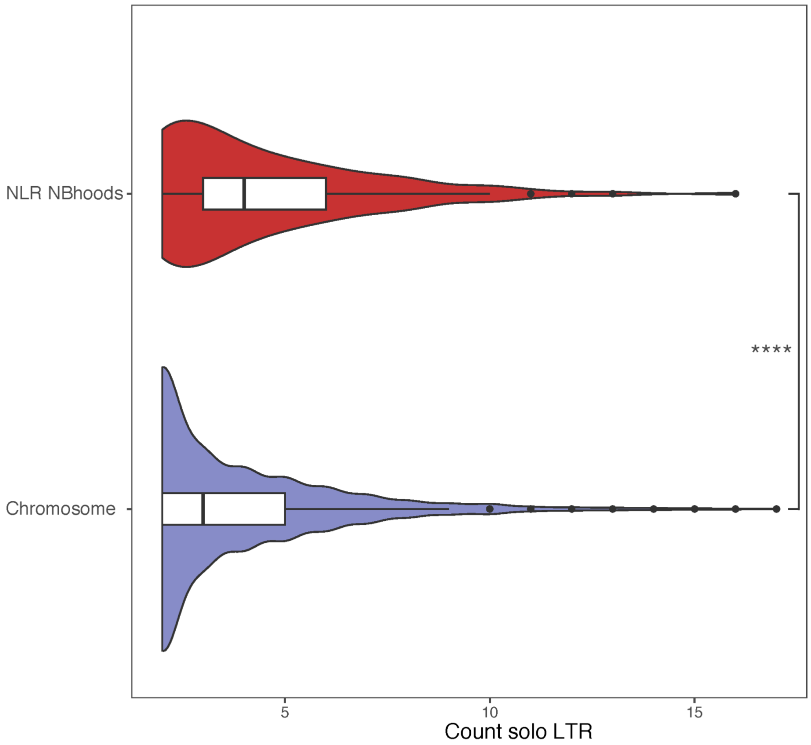


Figure 28 | Solo LTRs in NLR neighbourhoods and in background regions. Student's t-test, p-value = 6.921e-05.

TE dynamics at NLR genes.

We have described in the last section how NLR neighbourhoods are shaped by TE dynamics. In this section, we will focus on the direct action of TEs on the NLR genes themselves. First we will look at the differential TE loads between NLR and other genes (Figure 39). We find that, overall, TEs insert more often into NLR genes than into non-NLR genes. 63% of NLRs present at least one TE insertion, whereas for other genes, this percentage drops to just 30%. Non-NLR TE insertions are very similar across the different TE superfamilies. In comparison, UTRs and intronic regions of NLR genes are enriched in Harbinger, LTR/Ty3 and LTR/unknown superfamilies (Figure 39). Earlier reports pointed at the Helitrons superfamily as a source of variability in NLR clusters, but these inferences were based on short read data [53]. Most of the TE insertions into NLR genes appear to be TE remnants, judging by their size (Figure S7) or, alternatively, MITEs in the case of DNA TEs.

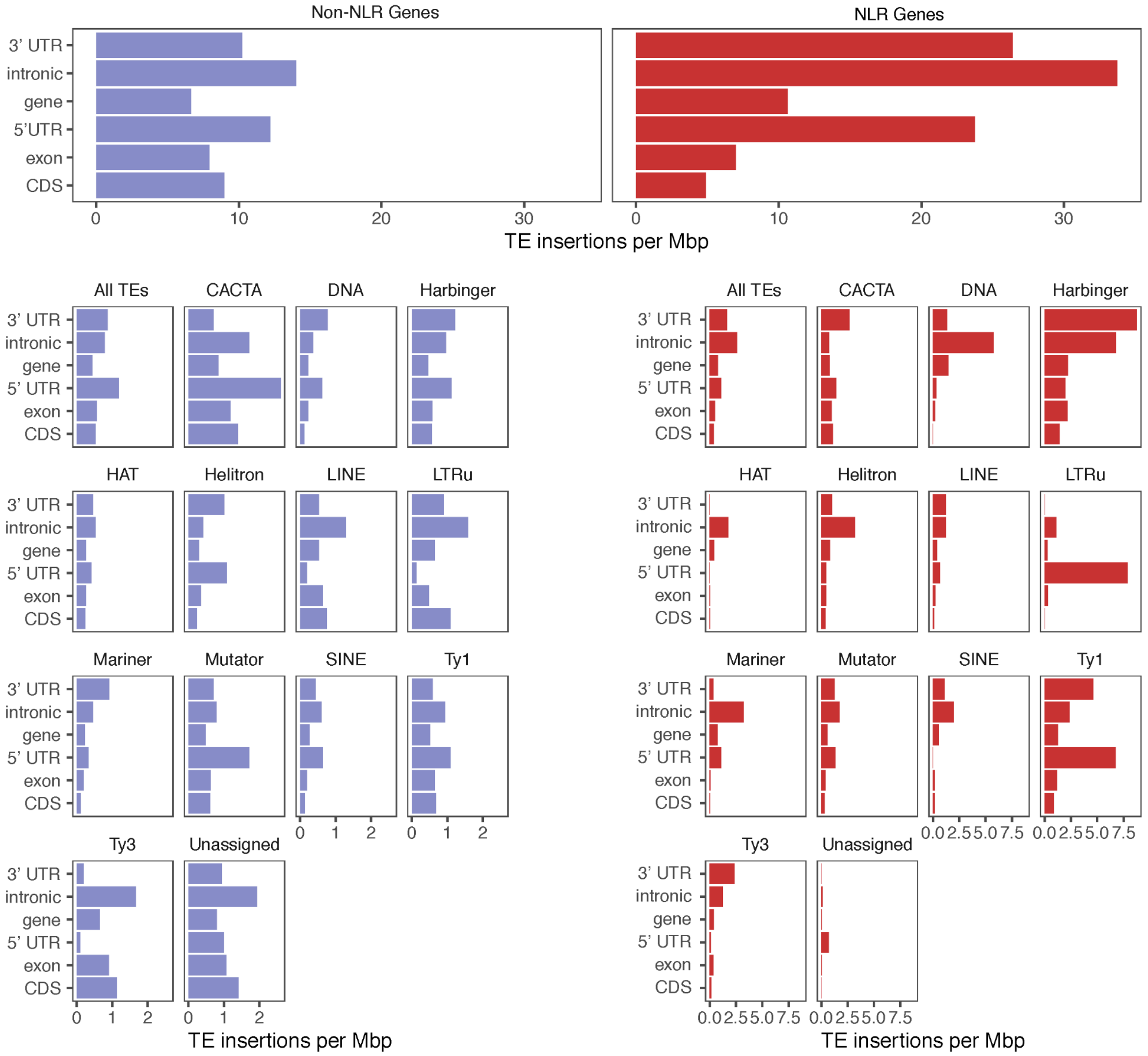


Figure 29 | Normalised TE load in NLR and other genes across the pangenome. TE count overlapping with different genomic features was normalised by the genome space they cover.

Finally, we would like to take in an exemplary fashion a look at the variability of a single NLR gene across the pangenome. *RPS5* (AT1G12220 in TAIR10) is a NLR gene that has been identified as a resistance gene involved in the defence against a natural *A. thaliana* pathogen, *Pseudomonas syringae,* through the recognition of the effector AvrPphB2 [281]. *RPS5* presents intermediate frequencies in allele polymorphisms globally. This has been suggested first as signatures of balancing selection between the effector-resistance gene pair [281]. More recent work, however, suggested that this global polymorphism is maintained through a complex relationship between multiple resistance alleles in *A. thaliana* populations and multiple effectors in different species of pathogens [282]. The coexistence of different functional alleles for *RPS5* species-wide makes it a very compelling case to examine the overall genomic context in which different alleles exist in our dataset.

We identified a total of 12 *RPS5* orthologs (Figure 30) and looked at its genomic environment within 10 kb upstream and downstream. We found that in three accessions the downstream end of the gene and second and last exon of *RPS5* is disrupted by a young intact LTR/Ty3 from the ATGP3 family. We estimate the insertions to have occurred 1.06 MYA for all 3 accessions, moreover, sequence alignment of these three TEs revealed only three polymorphic sites between them. Altogether, this suggests that a single event took place and that the insertion was maintained in only three accessions. The distribution of synteny blocks around *RPS5* reveal the downstream gene to be very similar to *RSP5*, having likely arisen from a tandem duplication (Figure 30 and Figure S8), but subsequently lost in many accessions.

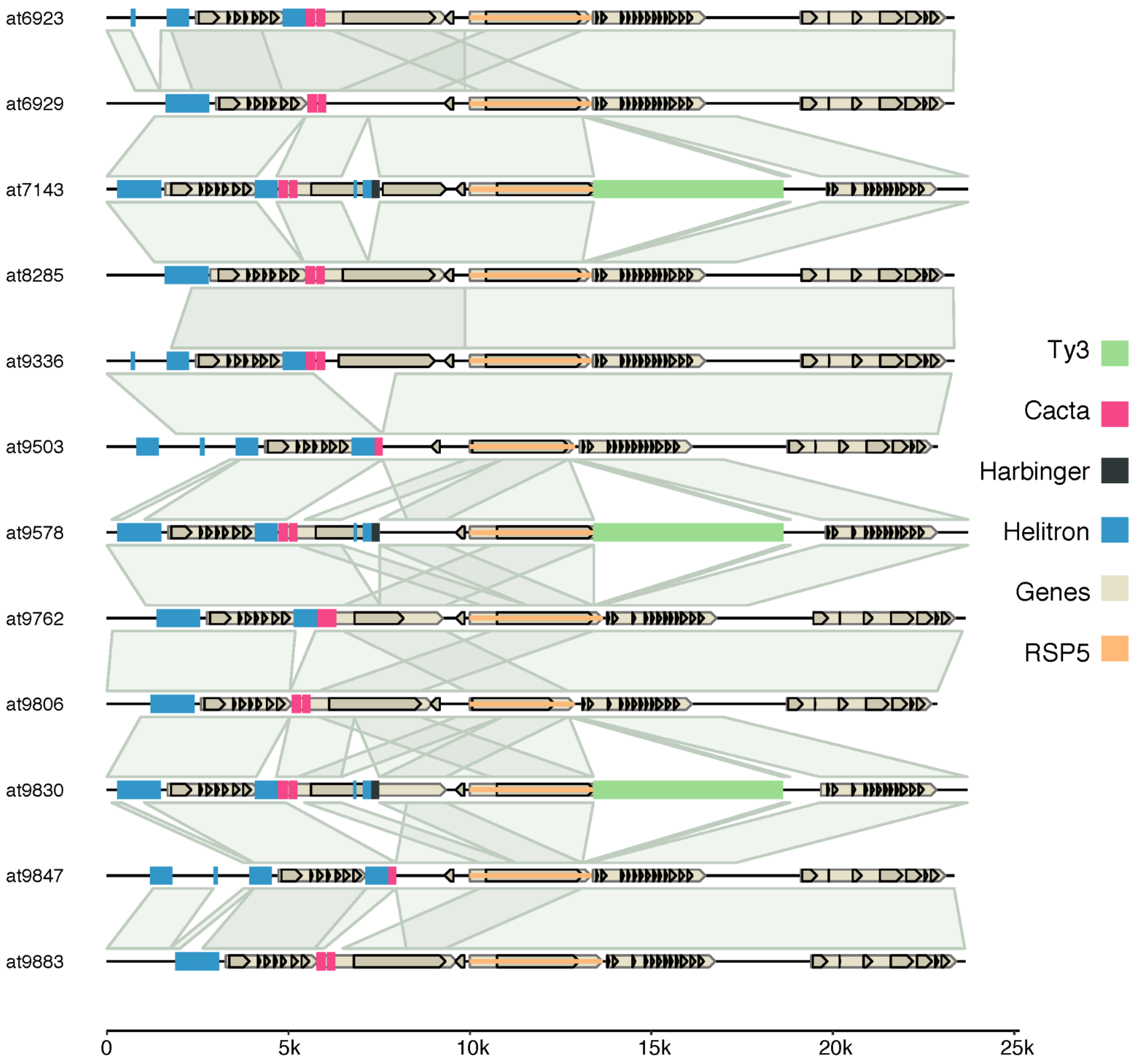


Figure 30 | The *RSP5* neighbourhood at the 12 accessions in which *RPS5, orange gene model,* is present. Green bars represent ATGP3 insertions. Synteny blocks between accessions are marked as semi-transparent grey ribbons.

5. Outlook

With this chapter, we provided a complete annotation of the TE landscape in a set of diverse *A. thaliana* accessions. The additional examination of enriched TE families in NLR neighbourhoods together with the identification of TE families directly affecting NLR genes will further aid the identification of such TEs that may have participated in processes of NLR diversification.

As we have seen at the beginning of the chapter, several processes linked to TE activity have been associated with the diversification of NLRs. A possible distinction can be made between genetic processes that are a direct consequence of TE activity, TE-driven processes, and those caused by subsequent mutations or selection (Table 6).

	Process	TIR	RC/Helitron	LTR	Publications
TE-driven	Gene capture	Yes	Yes	No	[140,270,283]
TE-driven	Gene fusion	Yes	Yes	No	[139]
TE-driven	Retrogenization	No	No	Yes	[216]
TE-driven	Gene disruption	Yes	Yes	Yes	[138]
TE-mediated	Non-allelic homologous recombination (NAHR)	Yes	No	Yes	[284]
TE-mediated	Exon splicing	Yes	Yes	Yes	[206]

Table 6. TE related mechanisms that may play a role in the diversification of NLRs.

The processes listed in Table 6 focus on single events that affect one, or part, of a single NLR gene. There are also TE-driven processes that affect multiple genes, often called segmental duplications.

Segmental duplications have been in general proposed as the main driving force of sequence diversification of plant NLR genes, similar to other major protein families such as cytochrome P450 genes, UDPG-glycosyltransferase genes or receptor-like kinase genes [285]. However, the molecular mechanisms that allow NLR genes to increase in number by segmental duplications remain uncharacterized. Indeed, duplications can be formed by TEs [286], either by their transposition, or by providing regions of sequence homology in non-syntenic regions, facilitating illegitimate recombination. For example, the maize Ac/fAc system causes complex chromosomal rearrangements including tandem or segmental duplications [286].

A major downside of studying this set of diverse accessions has been that specific alleles are likely separated by hundreds of thousands of years of evolution, making it difficult to reconstruct the mutational events that have led to the observed patterns. Increasing the number of accessions at intermediate levels of divergence will help clarify the mechanisms that give rise to the variation described in this chapter.

Supplementary figures

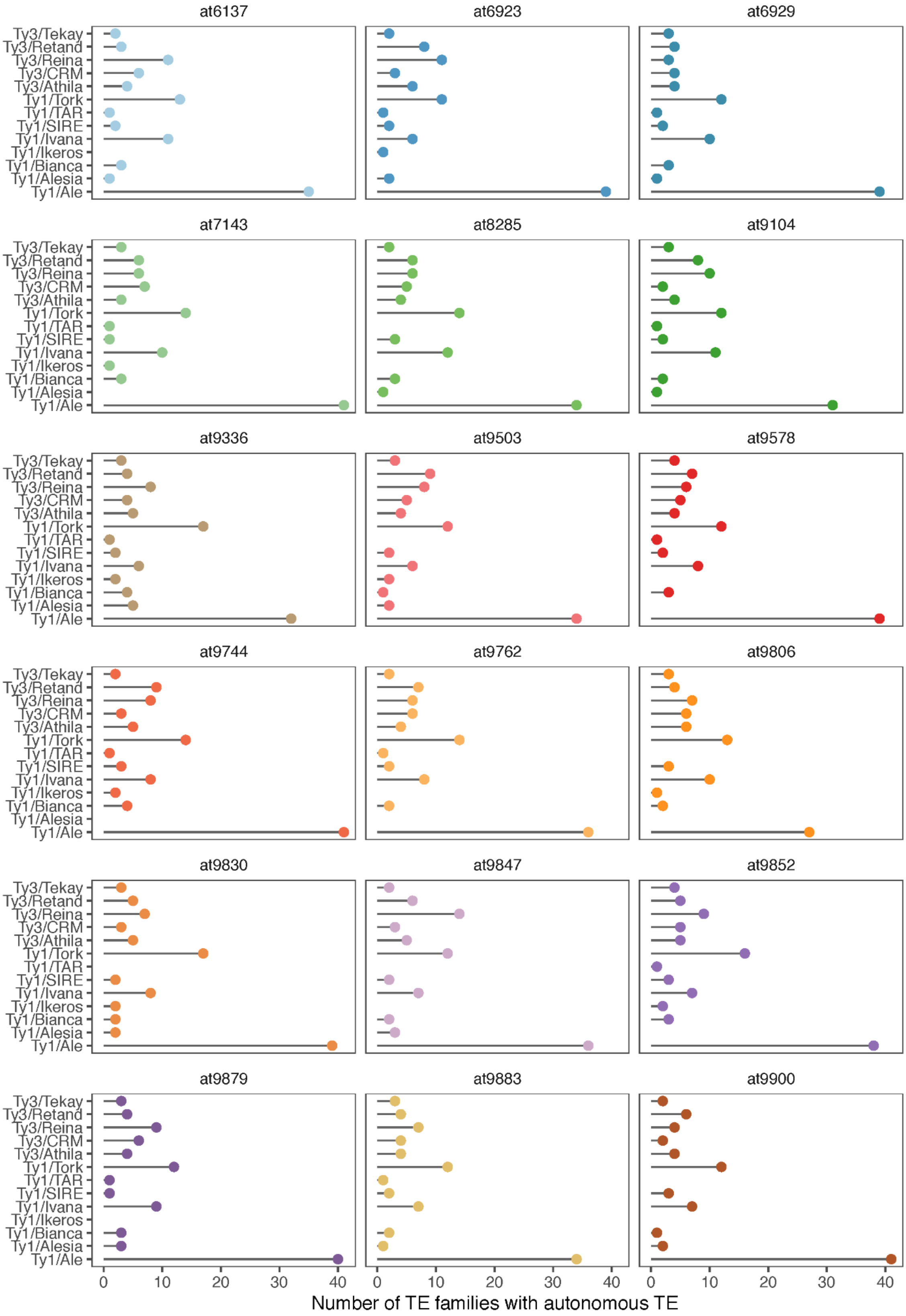


Figure S1 | Number of TE families with autonomous members for each LTR TE clade.

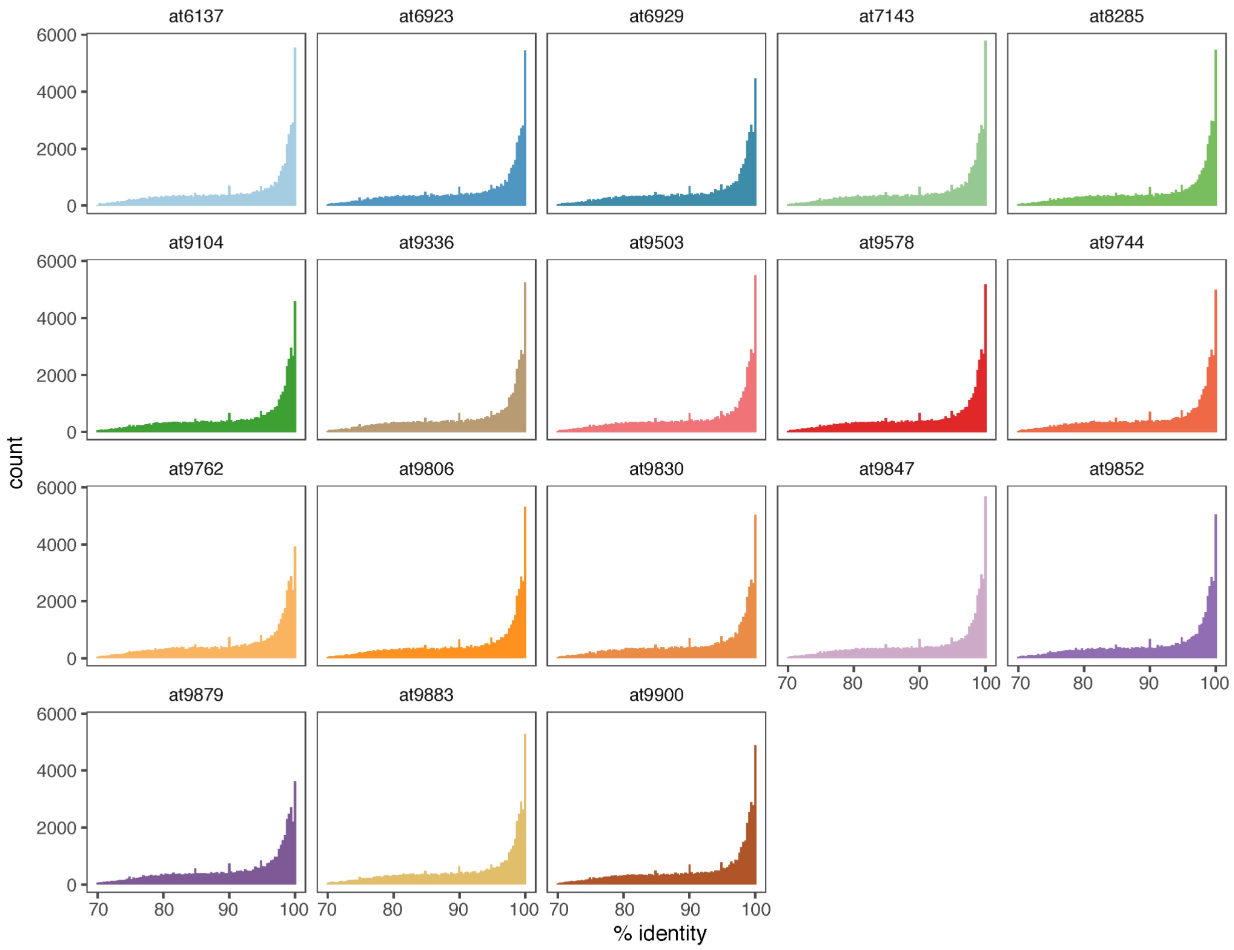


Figure S2 | Histogram of the number of the percentage identity of all the TE copies towards their corresponding TE model across the set of 18 different accessions of *A. thaliana.*

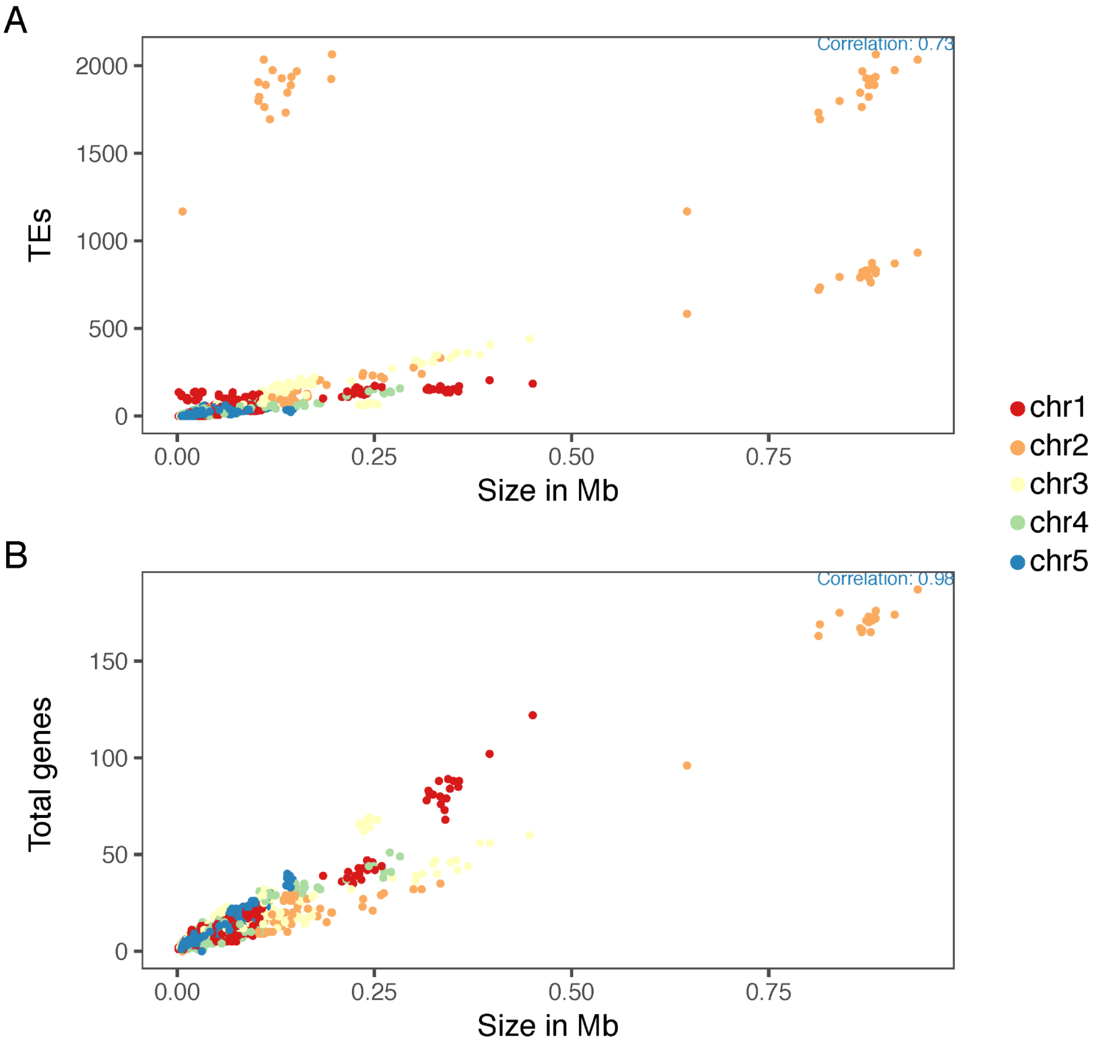


Figure S3 | Correlation of NLR neighbourhood size with (A) number of TEs and (B) number of genes.

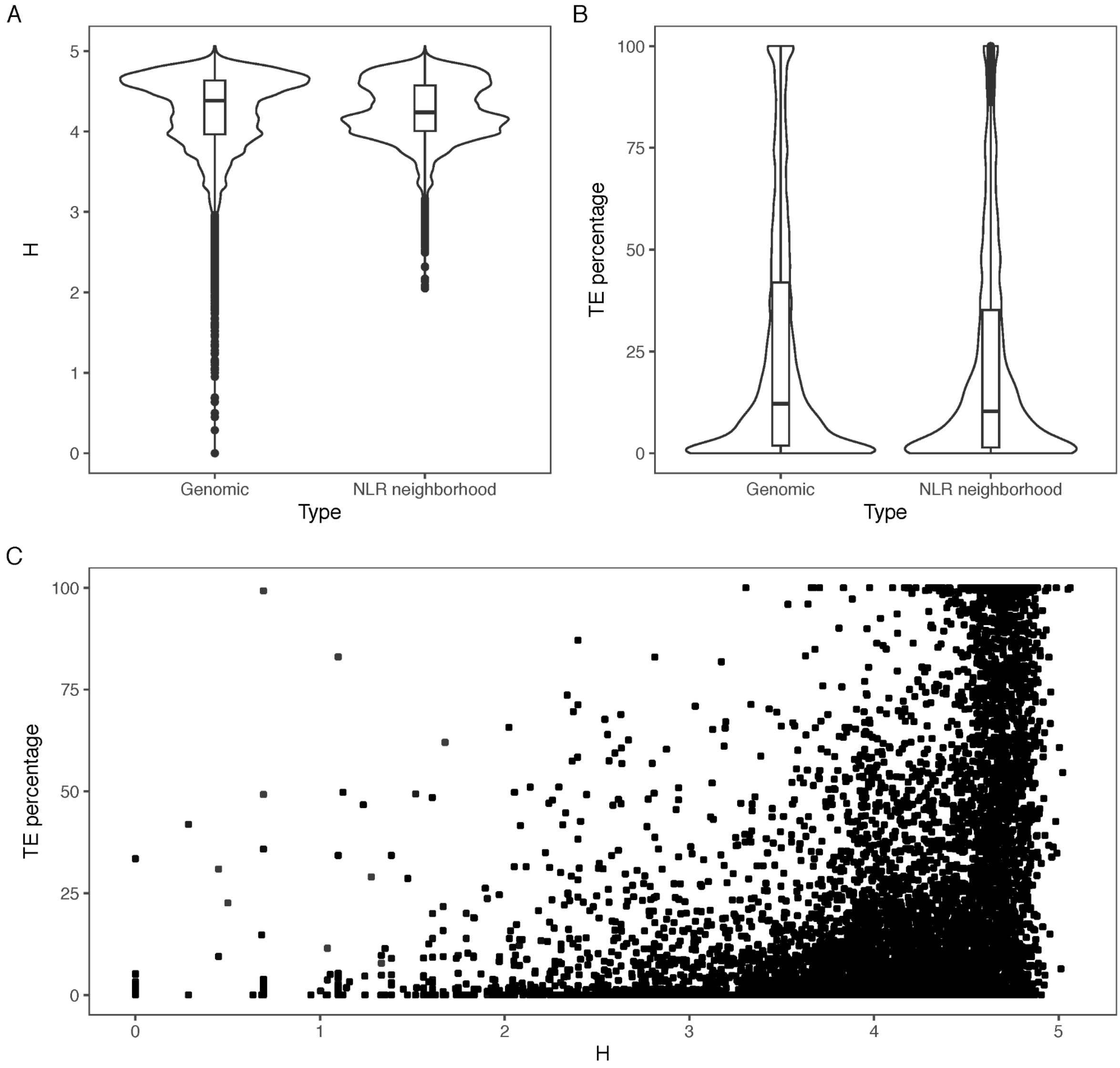


Figure S4 | Genome wide TE diversity. We calculated the Shannon diversity index (H) of (A) TE composition and (B) TE coverage in 10 kb sliding windows (step size 5 kb) for chromosome arms and then divide them either in genomic intervals or NLR neighbourhood intervals. (C) Correlation between TE diversity (H) and TE coverage for each one of the intervals.

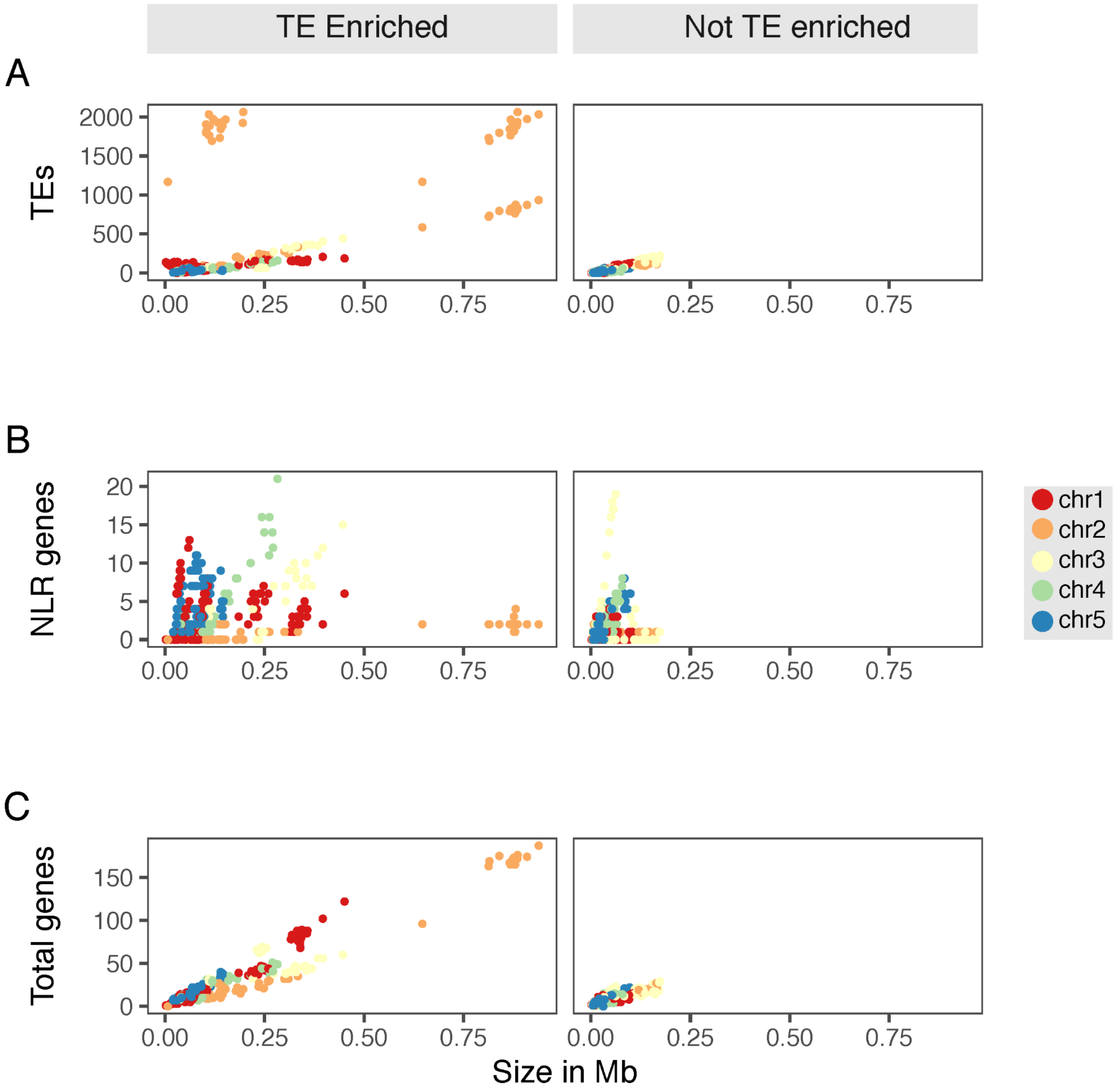


Figure S5 | Comparison of NLR neighbourhoods with at least one enriched TE family (left) or without any enriched TE family (right) for total number of (A) TE copies, (B) NLR genes, (C) other genes as a function of neighbourhood size.

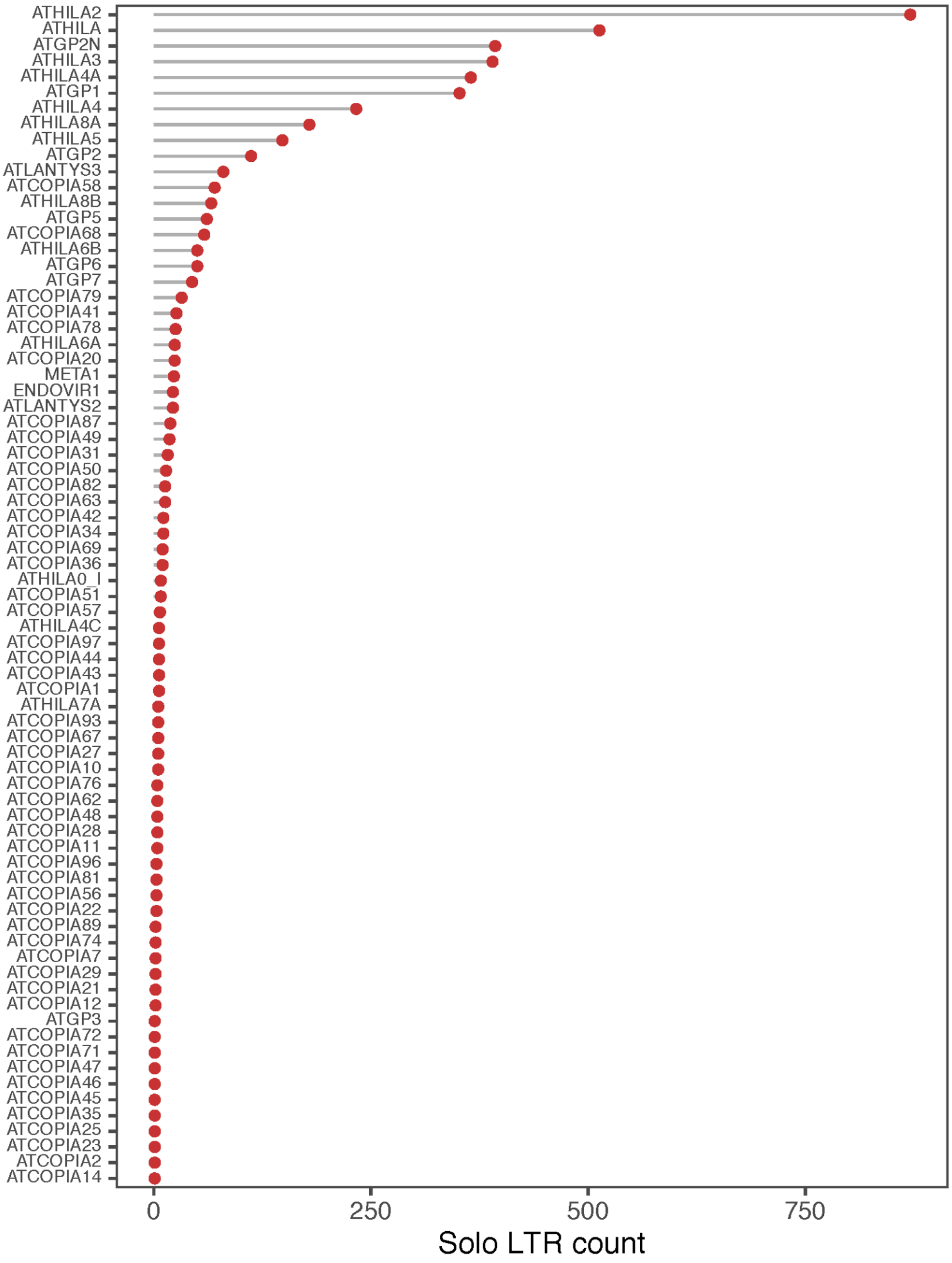


Figure S6 | Counts of solo LTRs by TE family across the pangenome.

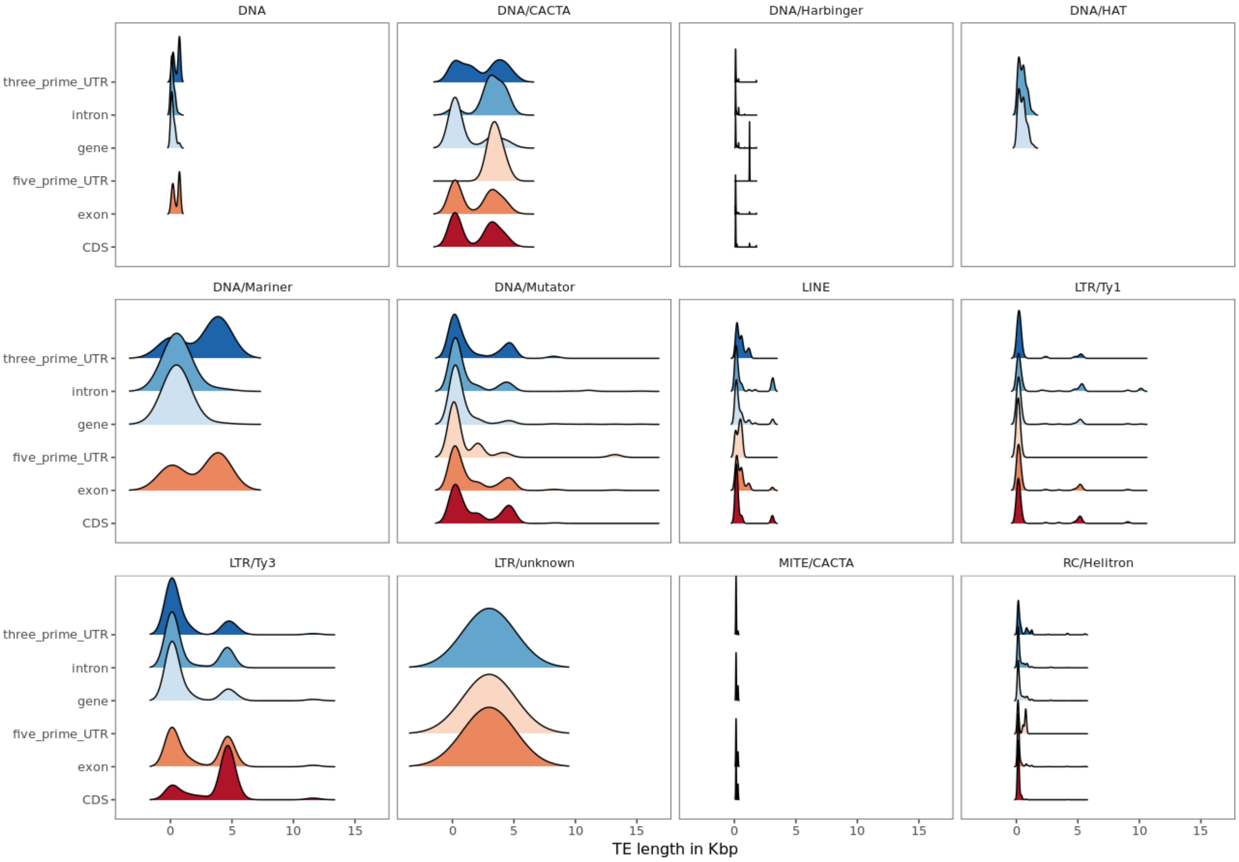


Figure S7 | Sizes of TE insertions at NLR genes.

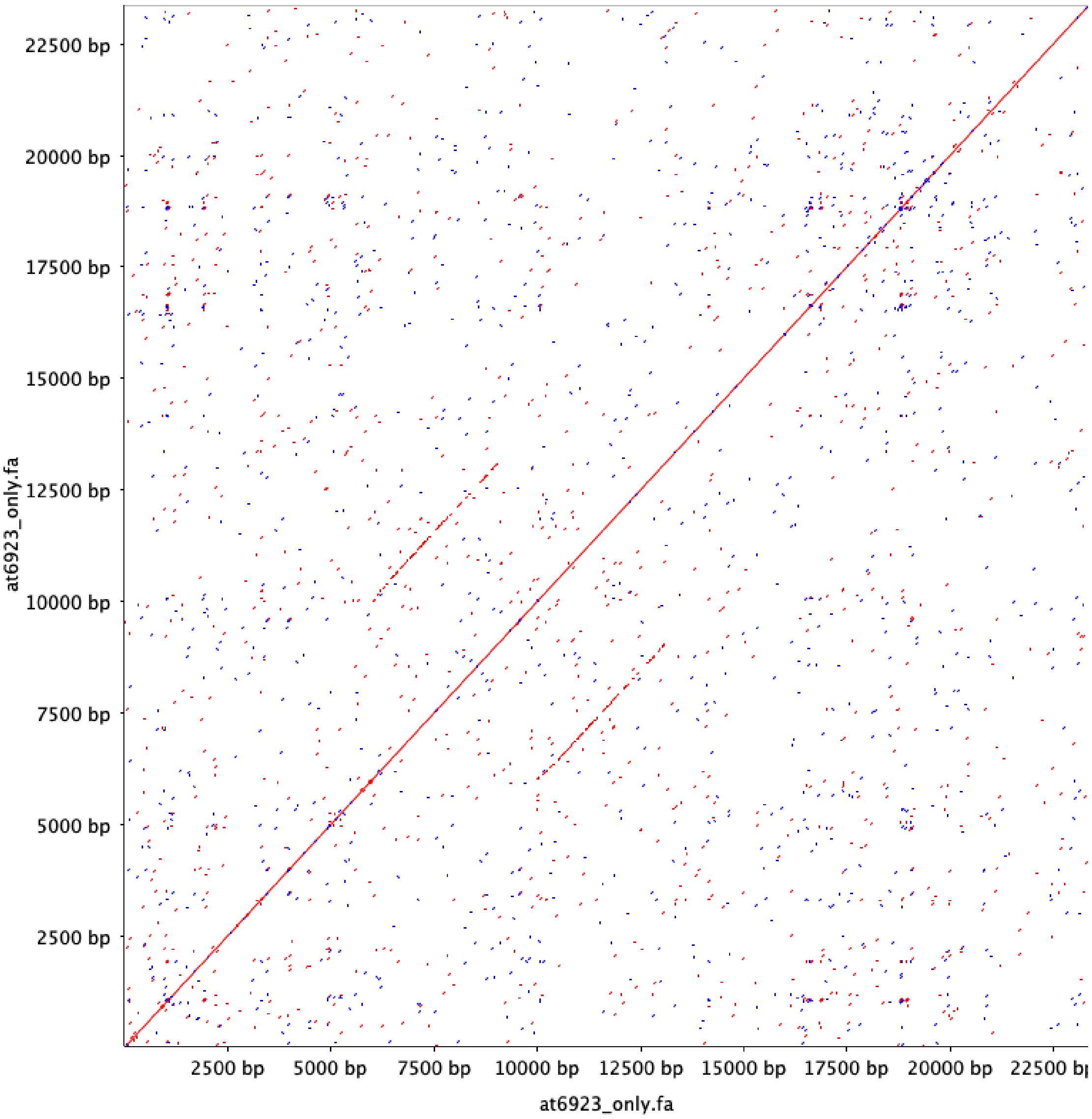


Figure S8 | Self alignment of the at6923 *RPS5* neighbourhood indicates the apparent duplication (coordinates: at6923_1_chr1:4128869-415220).

Discussion

Despite being recognized as “controlling elements” early on in the pioneering work of Barbara McClintock [287], there are many challenges for studying TE diversity across populations: there are often many copies that are highly related in sequence, limiting the usefulness of short read resequencing data; deletions are frequent, making alignments difficult; and they are often nested, with TEs inserted inside TEs. Nevertheless, short-read sequencing of TEs across populations has confirmed that TEs constitute some of the most variable components of genomes [288].

More recently, the rapid drop in costs for long-read sequencing has enabled the high-quality *de novo* assembly of many individual genomes of the same species. In addition, there has been a shift from the notion of a single reference genome of a species to a pan-genomic approach, where representatives of different populations within species are used to gain a more holistic view of genome composition [229]. To make these full-length, high-quality genomes meaningful for biological research, accurately identifying the locations of features such as genes and TEs within genome assemblies is crucial, as their interaction can structure and inform function of the genome.

This thesis mirrors these advances. I started with a single reference of the non-model plant *Thlaspi arvense* to annotate small RNA-producing loci and the TE landscape. I followed with an assessment of the capacity of mobilization of different TE families found in this reference and their phylogenetic classification and dating of a major subset of the TE complement. I analysed 280 globally-distributed whole genome sequenced (WGS) accessions of *T. arvense*. Using the depth and breadth that this WGS set, I aimed to identify TE polymorphisms relative to the reference, which enabled me to identify highly mobile TE families at the species level, obtaining a broad picture – albeit blurry due to the limitations of the technology – of the TE population dynamics of *T. arvense*. To finalise, I had access to a small but diverse set of 18 high-quality long-read genome assemblies of the model plant *Arabidopsis thaliana*. I annotated TEs in these 18 genomes, facing the technical challenges that translating a single reference annotation to a pan-genomic annotation implies. As a major aim of the collaborative project was to investigate the relationship between NLR clusters and transposable element variation, I identified TE families enriched for these clusters and also which TE elements are in direct interaction with NLR genes.

1. A phylogenetic approach for the study of TE dynamics.

Our TE annotation of *T. arvense* revealed that 60% of the entire genome consisted of LTR TEs, with a major superfamily (LTR/Ty3) corresponding to 54% of the genome and a second one (LTR/Ty1), to 6% of the genome. Both superfamilies are further divided into more than one thousand TE families, the vast majority of which are private for *T. arvense*. This situation is, by no means, unique to *T. arvense*, as many plant genomes have LTR TEs as their major component [289].

We are currently still in a situation in which a high proportion of plant genomes is accounted for by two distinct TE categories (superfamilies) that are consistent across taxa, followed by private categories (families) that are mostly species-specific and therefore fail to capture distant past relationships. Thus, many researchers have been pushing for the need of a consistent LTR classification that is not only more exhaustive but, importantly, also more robust across taxa [68]. One valuable approach is to create an intermediate classification between the superfamily level and the TE family level by using the enzymatic machinery encoded by the TEs as markers of the phylogenetic relationships among these groups of LTRs [272].

Using such a phylogenetically-informed reference dataset I was able to fill the classification gap between the TE superfamily and family classification. This new intermediate level, the clade, revealed interesting patterns within the *T. arvense* genome. My clade classification revealed distinct dynamics for different lineages of LTR TEs in terms of age and position within the genome (Chapter 2). Thanks to this new classification, I was able to compare clade profiles between my study and previous studies using a similar approach [290], highlighting the three-fold strength of this clade classification. First, it facilitates comparative genomic analysis between heterogeneous TE annotation approaches performed by different researchers. Second, it establishes a common nomenclature framework that enables literature comparisons among TE datasets without the need to perform additional analysis. Third, it adds nuance to within-species dynamics by increasing the range of items within the classification, in this case going from a single Ty1/Copia superfamily to 13 different clades and from the Ty3/Gypsy superfamily to 14 clades.

The immediate question that arises is: What about DNA transposons? Indeed, given the strengths of the LTR TE classification, one has to wonder why there have been so few efforts to provide to the community with a unified, phylogenetically-informed DNA TE phylogenetic classification. This seems especially surprising given that all DNA TEs encode a DDE/D transposase, which greatly facilitates DNA TE identification. Although at least one research group has been providing the community with broad descriptions and categorizations of DNA TE clades based on DDE/D transposase genes [291–293], these have not yet been integrated into any database or adopted by any TE classification tool.

Broadly adopting this system of LTR classification, and preferably also one for DNA TEs, as quickly as possible will greatly simplify the task of understanding TE evolution and TE-host coevolution dynamics. Ideally, this will happen before the further explosion of genomic datasets and TE studies that are expected to be fueled by the improvements of long-read sequencing technologies and their continuously decreasing costs [294].

2. Transposable elements as hotspots of genetic and epigenetic diversity.

Despite the limitations that short-read technologies have in the study of TEs, their lower cost compared with long-read technologies allows researchers to increase sample breadth, which is important particularly in non-model species, where the extent of the diversity within and between populations is generally unknown and a broad profiling is recommendable.

Following this reasoning, I exploited publicly available data of both whole genome re-sequencing and DNA methylation profiling for 280 *T. arvense* accessions to assess the TE diversity in this species. With the help of collaborators, I was able to pinpoint several genetic determinants of TE variation within our dataset using a GWAS approach that takes TE polymorphisms as a phenotype. Inter alia, allelic variation at a gene that encodes an ortholog of rice HEAT SHOCK PROTEIN 19 (HSP19) suggests a novel player in TE silencing, either as part of previously defined pathways or, albeit less likely, even pointing at an alternative path of TE silencing to those described in model species such as *A. thaliana*. Either way, this exciting result shows the value that non-model species may bring to the broad understanding of epigenetic regulation.

I also examined the impact of TE insertion polymorphisms on local epigenetic states. Comparing individuals carrying TE insertions against those which did not, I found an increase in methylation up to 1 kb away from the insertion location for retrotransposons but not for DNA TEs. This observation, together with other observations [295], underscore that TEs are major determinants of epigenetic diversity in *T. arvense*.

3. Induced transposition as a potential breeding tool.

A particularly exciting finding of this thesis is the identification of an active TE family in *T. arvense* that is highly enriched nearby and within genes in most of the sampled wild individuals. I highlighted in chapter two the potential of this TE family to be used as a breeding tool, and others have already speculated about this [296,297]. That *T. arvense* is already being bred as a biodiesel crop increases the appeal of this possibility. What are the necessary steps to achieve this?

First, it will be necessary to evaluate the transcriptional activation of this retrotransposon family. In Chapter two, I described a putative motif linked to heat stress activation. Thus, heat stress is a likely environmental cue for the activation of this TE family, as is the case in other Brassica TEs harbouring similar motifs [298]. To evaluate the transcriptional activation of this TE family, one could induce heat stress to a batch of samples of the reference accession, MN106, and measure the transcriptional activity by quantitative PCR or other methods. If the bioinformatic predictions are correct, it should be possible to detect a differential increase in the transcription of TE copies of this TE family under heat stress. This may be followed by experiments to define the best stress conditions for TE activation, to increase the number of accessions susceptible to the activation of the retrotransposon family and to investigate whether this increase of expression is also detected in both pollen and inflorescence tissue and the effect of the stress in seed viability and plant recovery.

An alternative approach (if heat stress fails to induce TE activation) could be the use of drugs such as 5-azacytidine and zebularine, which have been shown to transiently reduce DNA methylation levels throughout the genome [299].

Regardless of the method, evaluation of transcriptional activation in germline tissue will be most interesting, as the second step will be detection of successful transposition events. In order to establish this TE family as a breeding tool, detecting high transposition rates in the progeny of stressed plants will be desirable. This can be initially done using targeted long-read sequencing to examine which copies are most likely to mobilize, if any. Once the mobilised TE instances are known, it will be relatively easy to develop molecular markers to evaluate the transposition rates of these TE copies in a cost-effective manner.

Once this has been achieved, stress-mediated transposition in bulk can be performed followed by screening for trait selection (Figure 1). Moreover, one of the advantages of using TEs as mutagenesis agents, in comparison to chemical mutagens, is the automatic labelling of the insertion site with a known sequence. This greatly simplifies the identification of the genes controlling the altered phenotypic traits using a combination of TE-specific and degenerate primers in PCR assays [297].

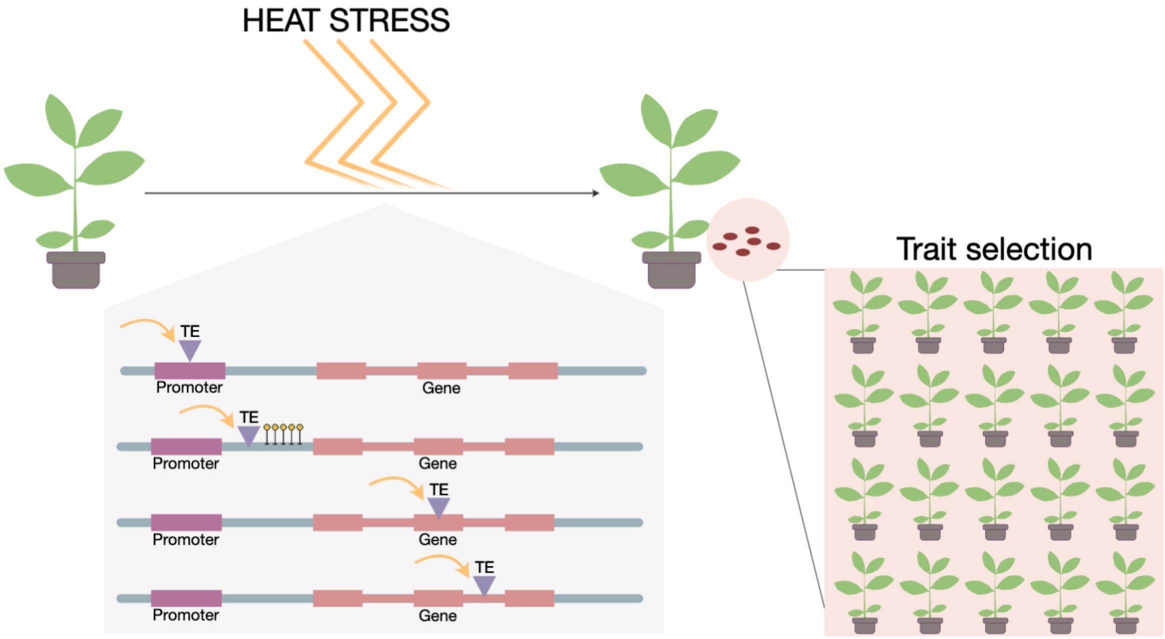


Figure 1 | Schematic representation of a plant breeding strategy that uses heat stress as activation cue for transposition-mediated mutagenesis.

FInally, we have also identified the presence of this TE family in several other Brassicaceae species, several of which are commercially used, such as *Raphanus sativus* (radish), *Brassica oleracea* (cabbage, broccoli, cauliflower) and *B. napus* (oilseed rape). Evaluation of the potential of this TE family as a breeding tool could also help efforts in already established crops, as cultivar requirements may change due to the climate emergency [300].

4. Opportunities and challenges in understanding the role of TEs in the plant immune system evolution.

Long-read sequencing has allowed researchers to interrogate the complex and repetitive features of plant genomes, improving our understanding, for example, of the centromeric diversity in *A. thaliana [267]*. It has also allowed for integration of several accessions of the same species, revealing large structural variants and hotspots of complexity, barely accessible with previous sequencing methods, allowing researchers to uncover hidden genetic variation in agronomically important species such as maize, tomato, rice, soybean or potato [301–304], and part of this variation has been pinpointed in resistance genes [305].

We have seen in Chapter three, that the plant immune system relies on a set of genes (NLRs) to detect pathogens and trigger immunity. As the Red Queen hypothesis predicts, there is a need to constantly generate diversity and novelty at these genes to maintain immunity. In selfing flowering plants such as *A. thaliana*, absence of outcrossing impedes the increase of diversity through sexual allele interchange. Despite this, there is no evidence, at least in the short term, that *A. thaliana* has a reduced suite of NLR genes in comparison with its outcrosser relative *A. lyrata [306]*. Other events, such as ectopic duplication, segmental duplications and chromosomal rearrangements have also been proposed to increase diversity at NLR genes [285]. These events will increase the diversity of these genes at the expense of a reduction of genome colinearity. Indeed, recent studies have shown that there is a reduced level of collinearity in *A. thaliana* genomes in certain genomic hotspots enriched in biotic stress related genes such as NLRs [230]. What is the mechanistic basis that allows for these mentioned major chromosomal rearrangements that break collinearity, but only at NLR clusters?

TEs have already been proposed to act as one of the genomic mechanisms that possibilities this in conjugation with DNA repair mechanisms [284]. TE insertions may provide homologous duplicates for ectopic recombination to act upon, creating tandem repeats. TE-produced enzymes such as transposases may act at distant degenerated TEs to produce large chromosome rearrangements. Cleavage at the donor site of a DNA TE may create a blunt DNA end that may be miss handled by the host DNA repair mechanisms provoking an alteration in the DNA sequence. Studies of the reference accession in *A. thaliana [208]* or in a few selected accessions [206] already demonstrated the importance that TEs have in the regulation of the plant immune system. As we have learned from Chapter three, these examples are likely not a singular event. The three main findings of Chapter three are:

- LTR TEs are younger near NLR loci.
- Solo LTRs are more common near NLR loci.
- NLR neighbourhoods are enriched in TE insertions.

These three findings revealed the high prevalence of TE variability at NLR genes and neighbouring loci and how much of this variability is recent, and therefore, unique among accessions. Moreover, because our evaluation was done comparing NLR neighbourhoods against the genome (excluding centromeres and other satellite regions) we can argue that this variability is not just due to the underlying population structure, but a particular feature of NLR diversity. Can we establish a causal link between the increase of TE variability at NLR clusters with the need of the plant host to increase diversity at these genes?

As previously discussed [230], relying only on shuffling existing variability to combat the ever-evolving pathogen challenge may not be sufficient to achieve immunity. Thus, is it possible that the plant host allows TE to thrive and act almost freely only at these clusters as a way to increase variability? Epigenetic mechanisms that control chromatin condensation such as histone modifications can provide the tools for the organisms for a “controlled burst” of TE activity but only in specific regions of the genome. For example, it has been proposed that the histone variant H2A.Z guides the insertion of certain transposons in the vicinity of stress related genes [128].

Another, more parsimonious, hypothesis for the observed phenomena is that TEs act equally along the chromosome arms, but that the observed insertion patterns are heavily influenced by purifying selection in most regions. However, because genes related to environmental responses are by definition only necessary in certain conditions, the fitness cost of alterations in a subset of these genes may be null for a long time, creating a more permissive context for TEs to act freely upon such loci and increasing diversity at the population level. Once the environmental conditions change, these loci become essential for the plant fitness and a selective sweep may occur at those loci, reducing diversity, but retaining “successful” new configurations of these loci. In other words, the cycle of discontinued selection pressures is what creates the observed signatures of increase TE activity at NLR loci.

In order to test which one of these hypotheses is correct, or to fully reject them, an increase of the collection of genomes is needed to conduct broad population studies at the genomic level, together with a better understanding at the molecular level of the mechanisms behind TE site selection insertion preferences.

References

1. Quadrana L, Colot V. Plant Transgenerational Epigenetics. Annu Rev Genet. 2016;50: 467–491.

2. Allis CD, Jenuwein T. The molecular hallmarks of epigenetic control. Nat Rev Genet. 2016;17: 487–500.

3. Choi K, Zhao X, Tock AJ, Lambing C, Underwood CJ, Hardcastle TJ, et al. Nucleosomes and DNA methylation shape meiotic DSB frequency in Arabidopsis thaliana transposons and gene regulatory regions. Genome Res. 2018;28: 532–546.

4. Schalk C, Drevensek S, Kramdi A, Kassam M, Ahmed I, Cognat V, et al. DNA DAMAGE BINDING PROTEIN 2 (DDB2) Shapes the DNA Methylation Landscape. Plant Cell. 2016. doi:10.1105/tpc.16.00474

5. Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, et al. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biol. 2013;14: R10.

6. Allshire RC, Karpen GH. Epigenetic regulation of centromeric chromatin: old dogs, new tricks? Nat Rev Genet. 2008;9: 923–937.

7. Naish M, Alonge M, Wlodzimierz P, Tock AJ, Abramson BW, Schmücker A, et al. The genetic and epigenetic landscape of the *Arabidopsis* centromeres. Science. 2021;374: eabi7489.

8. Annacondia ML, Magerøy MH, Martinez G. Stress response regulation by epigenetic mechanisms: changing of the guards. Physiol Plant. 2018;162: 239–250.

9. Cortijo S, Wardenaar R, Colomé-Tatché M, Gilly A, Etcheverry M, Labadie K, et al. Mapping the epigenetic basis of complex traits. Science. 2014;343: 1145–1148.

10. Weigel D, Colot V. Epialleles in plant evolution. Genome Biol. 2012;13: 249.

11. Arteaga-Vazquez MA, Chandler VL. Paramutation in maize: RNA mediated trans-generational gene silencing. Curr Opin Genet Dev. 2010;20: 156–163.

12. Bai F, Settles AM. Imprinting in plants as a mechanism to generate seed phenotypic diversity. Front Plant Sci. 2014;5: 780.

13. Zhang X, Jacobsen SE. Genetic analyses of DNA methyltransferases in Arabidopsis thaliana. Cold Spring Harb Symp Quant Biol. 2006;71: 439–447.

14. Liu R, How-Kit A, Stammitti L, Teyssier E, Rolin D, Mortain-Bertrand A, et al. A DEMETER-like DNA demethylase governs tomato fruit ripening. Proc Natl Acad Sci U S A. 2015;112: 10804–10809.

15. Erhard KF Jr, Stonaker JL, Parkinson SE, Lim JP, Hale CJ, Hollick JB. RNA polymerase IV functions in paramutation in Zea mays. Science. 2009;323: 1201–1205.

16. Moritoh S, Eun C-H, Ono A, Asao H, Okano Y, Yamaguchi K, et al. Targeted disruption of an orthologue of DOMAINS REARRANGED METHYLASE 2, OsDRM2, impairs the growth of rice plants by abnormal DNA methylation. Plant J. 2012;71: 85–98.

17. Dowen RH, Pelizzola M, Schmitz RJ, Lister R, Dowen JM, Nery JR, et al. Widespread dynamic DNA methylation in response to biotic stress. Proc Natl Acad Sci U S A. 2012;109: E2183–E2191.

18. Liu J, He Z. Small DNA Methylation, Big Player in Plant Abiotic Stress Responses and Memory. Front Plant Sci. 2020;11: 595603.

19. Baek D, Jiang J, Chung J-S, Wang B, Chen J, Xin Z, et al. Regulated AtHKT1 Gene Expression by a Distal Enhancer Element and DNA Methylation in the Promoter Plays an Important Role in Salt Tolerance. Plant Cell Physiol. 2010;52: 149–161.

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

21. Law JA, Jacobsen SE. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet. 2010;11: 204–220.

22. He X-J, Chen T, Zhu J-K. Regulation and function of DNA methylation in plants and animals. Cell Res. 2011;21: 442–465.

23. Movahedi A, Sun W, Zhang J, Wu X, Mousavi M, Mohammadi K, et al. RNA-directed DNA methylation in plants. Plant Cell Rep. 2015;34: 1857–1862.

24. Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SW-L, Chen H, et al. Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. Cell. 2006;126: 1189–1201.

25. Castel SE, Martienssen RA. RNA interference in the nucleus: roles for small RNAs in transcription, epigenetics and beyond. Nat Rev Genet. 2013;14: 100–112.

26. Cutter AR, Hayes JJ. A brief review of nucleosome structure. FEBS Lett. 2015;589: 2914–2922.

27. Foroozani M, Holder DH, Deal RB. Histone Variants in the Specialization of Plant Chromatin. Annu Rev Plant Biol. 2022;73: 149–172.

28. Wollmann H, Stroud H, Yelagandula R, Tarutani Y, Jiang D, Jing L, et al. The histone H3 variant H3.3 regulates gene body DNA methylation in Arabidopsis thaliana. Genome Biol. 2017;18: 94.

29. Zilberman D, Coleman-Derr D, Ballinger T, Henikoff S. Histone H2A.Z and DNA methylation are mutually antagonistic chromatin marks. Nature. 2008;456: 125–129.

30. Quesneville H. Twenty years of transposable element analysis in the Arabidopsis thaliana genome. Mob DNA. 2020;11: 28.

31. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, et al. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell. 2008;133: 523–536.

32. Liu P, Cuerda-Gil D, Shahid S, Slotkin RK. The Epigenetic Control of the Transposable Element Life Cycle in Plant Genomes and Beyond. Annu Rev Genet. 2022;56: 63–87.

33. Fultz D, Slotkin RK. Exogenous Transposable Elements Circumvent Identity-Based Silencing, Permitting the Dissection of Expression-Dependent Silencing. Plant Cell. 2017;29: 360–376.

34. Wilson RC, Doudna JA. Molecular mechanisms of RNA interference. Annu Rev Biophys. 2013;42: 217–239.

35. Axtell MJ, Snyder JA, Bartel DP. Common functions for diverse small RNAs of land plants. Plant Cell. 2007;19: 1750–1769.

36. Fang X, Qi Y. RNAi in Plants: An Argonaute-Centered View. Plant Cell. 2016;28: 272–285.

37. Creasey KM, Zhai J, Borges F, Van Ex F, Regulski M, Meyers BC, et al. miRNAs trigger widespread epigenetically activated siRNAs from transposons in Arabidopsis. Nature. 2014;508: 411–415.

38. Matzke MA, Kanno T, Matzke AJM. RNA-Directed DNA Methylation: The Evolution of a Complex Epigenetic Pathway in Flowering Plants. Annu Rev Plant Biol. 2015;66: 243–267.

39. Ito H, Kim J-M, Matsunaga W, Saze H, Matsui A, Endo TA, et al. A Stress-Activated Transposon in Arabidopsis Induces Transgenerational Abscisic Acid Insensitivity. Sci Rep. 2016;6: 23181.

40. 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;472: 115–119.

41. Yu A, Lepère G, Jay F, Wang J, Bapaume L, Wang Y, et al. Dynamics and biological relevance of DNA demethylation in Arabidopsis antibacterial defense. Proc Natl Acad Sci U S A. 2013;110: 2389–2394.

42. Cuerda-Gil D, Slotkin RK. Non-canonical RNA-directed DNA methylation. Nature Plants. 2016;2: 16163.

43. Law JA, Du J, Hale CJ, Feng S, Krajewski K, Palanca AMS, et al. Polymerase IV occupancy at RNA-directed DNA methylation sites requires SHH1. Nature. 2013;498: 385.

44. Liu Z-W, Shao C-R, Zhang C-J, Zhou J-X, Zhang S-W, Li L, et al. The SET domain proteins SUVH2 and SUVH9 are required for Pol V occupancy at RNA-directed DNA methylation loci. PLoS Genet. 2014;10: e1003948.

45. Marí-Ordóñez A, Marchais A, Etcheverry M, Martin A, Colot V, Voinnet O.

Reconstructing de novo silencing of an active plant retrotransposon. Nat Genet. 2013;45: 1029–1039.

46. Ye R, Chen Z, Lian B, Rowley MJ, Xia N, Chai J, et al. A Dicer-Independent Route for Biogenesis of siRNAs that Direct DNA Methylation in Arabidopsis. Mol Cell. 2016;61: 222–235.

47. Fultz D, Choudury SG, Slotkin RK. Silencing of active transposable elements in plants. Curr Opin Plant Biol. 2015;27: 67–76.

48. Krämer U. Planting molecular functions in an ecological context with Arabidopsis thaliana. Elife. 2015;4. doi:10.7554/eLife.06100

49. Hoffmann MH. Evolution of the realized climatic niche in the genus Arabidopsis (Brassicaceae). Evolution. 2005;59: 1425–1436.

50. 1001 Genomes Consortium. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. Cell. 2016;166: 481–491.

51. Weigel D. Natural variation in Arabidopsis: from molecular genetics to ecological genomics. Plant Physiol. 2012;158: 2–22.

52. Vaughn MW, Tanurdzić M, Lippman Z, Jiang H, Carrasquillo R, Rabinowicz PD, et al. Epigenetic natural variation in Arabidopsis thaliana. PLoS Biol. 2007;5: e174.

53. 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;166: 492–505.

54. Zhai J, Liu J, Liu B, Li P, Meyers BC, Chen X, et al. Small RNA-directed epigenetic natural variation in Arabidopsis thaliana. PLoS Genet. 2008;4: e1000056.

55. Lang Z, Xie S, Zhu J-K. The 1001 Arabidopsis DNA Methylomes: An Important Resource for Studying Natural Genetic, Epigenetic, and Phenotypic Variation. Trends Plant Sci. 2016;21: 906–908.

56. Pei L, Zhang L, Li J, Shen C, Qiu P, Tu L, et al. Tracing the origin and evolution history of methylation-related genes in plants. BMC Plant Biol. 2019;19: 307.

57. Blow MJ, Clark TA, Daum CG, Deutschbauer AM, Fomenkov A, Fries R, et al. The Epigenomic Landscape of Prokaryotes. PLoS Genet. 2016;12: e1005854.

58. Stojkova P, Spidlova P, Stulik J. Nucleoid-Associated Protein HU: A Lilliputian in Gene Regulation of Bacterial Virulence. Front Cell Infect Microbiol. 2019;9: 159.

59. Quendera AP, Seixas AF, Dos Santos RF, Santos I, Silva JPN, Arraiano CM, et al. RNA-Binding Proteins Driving the Regulatory Activity of Small Non-coding RNAs in Bacteria. Front Mol Biosci. 2020;7: 78.

60. Fedoroff NV. Presidential address. Transposable elements, epigenetics, and genome evolution. Science. 2012;338: 758–767.

61. Elliott TA, Gregory TR. What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. Philos Trans R Soc Lond B Biol Sci. 2015;370: 20140331.

62. Kidwell MG. Transposable elements and the evolution of genome size in eukaryotes. Genetica. 2002;115: 49–63.

63. Lynch M, Conery JS. The origins of genome complexity. Science. 2003;302: 1401–1404.

64. 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;8: 973–982.

65. McCLINTOCK B. The origin and behavior of mutable loci in maize. Proc Natl Acad Sci U S A. 1950;36: 344–355.

66. Orgel LE, Crick FH. Selfish DNA: the ultimate parasite. Nature. 1980;284: 604–607.

67. Lisch D. How important are transposons for plant evolution? Nat Rev Genet. 2013;14: 49–61.

68. Piégu B, Bire S, Arensburger P, Bigot Y. A survey of transposable element classification systems--a call for a fundamental update to meet the challenge of their diversity and complexity. Mol Phylogenet Evol. 2015;86: 90–109.

69. Jangam D, Feschotte C, Betrán E. Transposable Element Domestication As an Adaptation to Evolutionary Conflicts. Trends Genet. 2017;33: 817–831.

70. Maumus F, Quesneville H. Ancestral repeats have shaped epigenome and genome composition for millions of years in Arabidopsis thaliana. Nat Commun. 2014;5: 4104.

71. Frith MC. Paleozoic Protein Fossils Illuminate the Evolution of Vertebrate Genomes and Transposable Elements. Mol Biol Evol. 2022;39. doi:10.1093/molbev/msac068

72. Finnegan DJ. Eukaryotic transposable elements and genome evolution. Trends Genet. 1989;5: 103–107.

73. Arkhipova IR. Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories. Mob DNA. 2017;8: 19.

74. Seberg O, Petersen G. A unified classification system for eukaryotic transposable elements should reflect their phylogeny. Nature reviews. Genetics. 2009. p. 276.

75. Vázquez-Manrique RP, Hernández M, Martínez-Sebastián MJ, de Frutos R. Evolution of gypsy endogenous retrovirus in the Drosophila obscura species group. Mol Biol Evol. 2000;17: 1185–1193.

76. Pélisson A, Teysset L, Chalvet F, Kim A, Prud'homme N, Terzian C, et al. About the origin of retroviruses and the co-evolution of the gypsy retrovirus with the Drosophila flamenco host gene. Genetica. 1997;100: 29–37.

77. Eickbush TH, Jamburuthugoda VK. The diversity of retrotransposons and the properties of their reverse transcriptases. Virus Res. 2008;134: 221–234.

78. Llorens C, Muñoz-Pomer A, Bernad L, Botella H, Moya A. Network dynamics of eukaryotic LTR retroelements beyond phylogenetic trees. Biol Direct. 2009;4: 41.

79. Wang J, Han G-Z. A Missing Link between Retrotransposons and Retroviruses. MBio. 2022;13: e0018722.

80. Kapitonov VV, Jurka J. Self-synthesizing DNA transposons in eukaryotes. Proc Natl Acad Sci U S A. 2006;103: 4540–4545.

81. Pritham EJ, Putliwala T, Feschotte C. Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. Gene. 2007;390: 3–17.

82. Krupovic M, Bamford DH, Koonin EV. Conservation of major and minor jelly-roll capsid proteins in Polinton (Maverick) transposons suggests that they are bona fide viruses. Biol Direct. 2014;9: 6.

83. Krupovic M, Koonin EV. Polintons: a hotbed of eukaryotic virus, transposon and plasmid evolution. Nat Rev Microbiol. 2015;13: 105–115.

84. Widen SA, Bes IC, Koreshova A, Pliota P, Krogull D, Burga A. Virus-like transposons cross the species barrier and drive the evolution of genetic incompatibilities. Science. 2023;380: eade0705.

85. Kapitonov VV, Jurka J. Rolling-circle transposons in eukaryotes. Proc Natl Acad Sci U S A. 2001;98: 8714–8719.

86. Murad L, Bielawski JP, Matyasek R, Kovarík A, Nichols RA, Leitch AR, et al. The origin and evolution of geminivirus-related DNA sequences in Nicotiana. Heredity . 2004;92: 352–358.

87. Feschotte C, Wessler SR. Treasures in the attic: rolling circle transposons discovered in eukaryotic genomes. Proceedings of the National Academy of Sciences of the United States of America. 2001. pp. 8923–8924.

88. Heringer P, Kuhn GCS. Exploring the Remote Ties between Helitron Transposases and Other Rolling-Circle Replication Proteins. Int J Mol Sci. 2018;19. doi:10.3390/ijms19103079

89. Yang W. Nucleases: diversity of structure, function and mechanism. Q Rev Biophys. 2011;44: 1–93.

90. Curcio MJ, Derbyshire KM. The outs and ins of transposition: from mu to kangaroo. Nat Rev Mol Cell Biol. 2003;4: 865–877.

91. Hickman AB, Dyda F. Mechanisms of DNA Transposition. Microbiol Spectr. 2015;3: MDNA3–0034–2014.

92. Feschotte C, Pritham EJ. DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet. 2007;41: 331–368.

93. Jurka J, Kapitonov VV. PIFs meet Tourists and Harbingers: a superfamily reunion. Proceedings of the National Academy of Sciences of the United States of America. 2001. pp. 12315–12316.

94. Macko-Podgórni A, Stelmach K, Kwolek K, Grzebelus D. Stowaway miniature inverted repeat transposable elements are important agents driving recent genomic diversity in wild and cultivated carrot. Mob DNA. 2019;10: 47.

95. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015;6: 11.

96. Kennedy AK, Haniford DB, Mizuuchi K. Single active site catalysis of the successive phosphoryl transfer steps by DNA transposases: insights from phosphorothioate stereoselectivity. Cell. 2000;101: 295–305.

97. Dyda F, Chandler M, Hickman AB. The emerging diversity of transpososome architectures. Q Rev Biophys. 2012;45: 493–521.

98. Changela A, Perry K, Taneja B, Mondragón A. DNA manipulators: caught in the act. Curr Opin Struct Biol. 2003;13: 15–22.

99. Wang D, Zheng Z, Li Y, Hu H, Wang Z, Du X, et al. Which factors contribute most to genome size variation within angiosperms? Ecol Evol. 2021;11: 2660–2668.

100. Zhou Y, Cahan SH. A novel family of terminal-repeat retrotransposon in miniature (TRIM) in the genome of the red harvester ant, Pogonomyrmex barbatus. PLoS One. 2012;7: e53401.

101. Sabot F, Schulman AH. Parasitism and the retrotransposon life cycle in plants: a hitchhiker's guide to the genome. Heredity . 2006;97: 381–388.

102. Chang W, Jääskeläinen M, Li S-P, Schulman AH. BARE retrotransposons are translated and replicated via distinct RNA pools. PLoS One. 2013;8: e72270.

103. Cullen H, Schorn AJ. Endogenous Retroviruses Walk a Fine Line between Priming and Silencing. Viruses. 2020;12. doi:10.3390/v12080792

104. Lee S-K, Potempa M, Swanstrom R. The Choreography of HIV-1 Proteolytic Processing and Virion Assembly*. J Biol Chem. 2012;287: 40867–40874.

105. Jaaskelainen M, Mykkanen AH, Arna T, Vicient CM, Suoniemi A, Kalendar R, et al. Retrotransposon BARE-1: expression of encoded proteins and formation of virus-like particles in barley cells. Plant J. 1999;20: 413–422.

106. Rausch JW, Miller JT, Le Grice SFJ. Reverse Transcription in the Saccharomyces cerevisiae Long-Terminal Repeat Retrotransposon Ty3. Viruses. 2017;9. doi:10.3390/v9030044

107. Masuta Y, Nozawa K, Takagi H, Yaegashi H, Tanaka K, Ito T, et al. Inducible Transposition of a Heat-Activated Retrotransposon in Tissue Culture. Plant Cell Physiol. 2017;58: 375–384.

108. Hirochika H. Activation of tobacco retrotransposons during tissue culture. EMBO J. 1993;12: 2521–2528.

109. Drost H-G, Sanchez DH. Becoming a Selfish Clan: Recombination Associated to Reverse-Transcription in LTR Retrotransposons. Genome Biol Evol. 2019;11: 3382–3392.

110. Onafuwa-Nuga Adewunmi, Telesnitsky Alice. The Remarkable Frequency of Human Immunodeficiency Virus Type 1 Genetic Recombination. Microbiol Mol Biol Rev. 2009;73: 451–480.

111. Sanchez DH, Gaubert H, Drost H-G, Zabet NR, Paszkowski J. High-frequency recombination between members of an LTR retrotransposon family during transposition bursts. Nat Commun. 2017;8: 1283.

112. Heitkam T, Holtgräwe D, Dohm JC, Minoche AE, Himmelbauer H, Weisshaar B, et al. Profiling of extensively diversified plant LINEs reveals distinct plant-specific subclades. Plant J. 2014;79: 385–397.

113. Wicker T, Yahiaoui N, Keller B. Illegitimate recombination is a major evolutionary mechanism for initiating size variation in plant resistance genes. Plant J. 2007;51: 631–641.

114. Pélissier T, Bousquet-Antonelli C, Lavie L, Deragon J-M. Synthesis and processing of tRNA-related SINE transcripts in Arabidopsis thaliana. Nucleic Acids Res. 2004;32: 3957–3966.

115. Evans JP, Palmiter RD. Retrotransposition of a mouse L1 element. Proc Natl Acad Sci U S A. 1991;88: 8792–8795.

116. Luan DD, Korman MH, Jakubczak JL, Eickbush TH. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell. 1993;72: 595–605.

117. Haas NB, Grabowski JM, North J, Moran JV, Kazazian HH, Burch JB. Subfamilies of CR1 non-LTR retrotransposons have different 5’UTR sequences but are otherwise conserved. Gene. 2001;265: 175–183.

118. Schumann G, Zündorf I, Hofmann J, Marschalek R, Dingermann T. Internally located and oppositely oriented polymerase II promoters direct convergent transcription of a LINE-like retroelement, the Dictyostelium repetitive element, from Dictyostelium discoideum. Mol Cell Biol. 1994;14: 3074–3084.

119. Goodier JL, Zhang L, Vetter MR, Kazazian HH Jr. LINE-1 ORF1 protein localizes in stress granules with other RNA-binding proteins, including components of RNA interference RNA-induced silencing complex. Mol Cell Biol. 2007;27: 6469–6483.

120. Han JS. Non-long terminal repeat (non-LTR) retrotransposons: mechanisms, recent developments, and unanswered questions. Mob DNA. 2010;1: 15.

121. Blesa D, Martínez-Sebastián MJ. bilbo, a non-LTR retrotransposon of Drosophila subobscura: a clue to the evolution of LINE-like elements in Drosophila. Mol Biol Evol.

1997;14: 1145–1153.

122. Kurzynska-Kokorniak A, Jamburuthugoda VK, Bibillo A, Eickbush TH. DNA-directed DNA polymerase and strand displacement activity of the reverse transcriptase encoded by the R2 retrotransposon. J Mol Biol. 2007;374: 322–333.

123. Gasior SL, Wakeman TP, Xu B, Deininger PL. The human LINE-1 retrotransposon creates DNA double-strand breaks. J Mol Biol. 2006;357: 1383–1393.

124. Mendiola MV, Bernales I, de la Cruz F. Differential roles of the transposon termini in IS91 transposition. Proc Natl Acad Sci U S A. 1994;91: 1922–1926.

125. Grabundzija I, Messing SA, Thomas J, Cosby RL, Bilic I, Miskey C, et al. A Helitron transposon reconstructed from bats reveals a novel mechanism of genome shuffling in eukaryotes. Nat Commun. 2016;7: 10716.

126. Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nat Genet. 2005;37: 997–1002.

127. Bourgeois Y, Boissinot S. On the Population Dynamics of Junk: A Review on the Population Genomics of Transposable Elements. Genes . 2019;10. doi:10.3390/genes10060419

128. 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;10: 3421.

129. 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. 2016;5: e15716.

130. Baller JA, Gao J, Stamenova R, Curcio MJ, Voytas DF. A nucleosomal surface defines an integration hotspot for the Saccharomyces cerevisiae Ty1 retrotransposon. Genome Res. 2012;22: 704–713.

131. Sandmeyer S, Patterson K, Bilanchone V. Ty3, a Position-specific Retrotransposon in Budding Yeast. Microbiol Spectr. 2015;3: MDNA3–0057–2014.

132. Sandmeyer SB, Hansen LJ, Chalker DL. Integration specificity of retrotransposons and retroviruses. Annu Rev Genet. 1990;24: 491–518.

133. Dupuy AJ, Akagi K, Largaespada DA, Copeland NG, Jenkins NA. Mammalian mutagenesis using a highly mobile somatic Sleeping Beauty transposon system. Nature. 2005;436: 221–226.

134. Pan X, Li Y, Stein L. Site preferences of insertional mutagenesis agents in Arabidopsis. Plant Physiol. 2005;137: 168–175.

135. Anxolabéhère D, Kidwell MG, Periquet G. Molecular characteristics of diverse populations are consistent with the hypothesis of a recent invasion of Drosophila melanogaster by mobile P elements. Mol Biol Evol. 1988;5: 252–269.

136. Kofler R, Hill T, Nolte V, Betancourt AJ, Schlötterer C. The recent invasion of natural *Drosophila simulans* populations by the P-element. Proceedings of the National Academy of Sciences. 2015;112: 6659–6663.

137. Chen J, Lu L, Robb SMC, Collin M, Okumoto Y, Stajich JE, et al. Genomic diversity generated by a transposable element burst in a rice recombinant inbred population. Proc Natl Acad Sci U S A. 2020;117: 26288–26297.

138. Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, et al. Ten things you should know about transposable elements. Genome Biol. 2018;19: 199.

139. Jiang N, Ferguson AA, Slotkin RK, Lisch D. Pack-Mutator-like transposable elements (Pack-MULEs) induce directional modification of genes through biased insertion and DNA acquisition. Proc Natl Acad Sci U S A. 2011;108: 1537–1542.

140. Catoni M, Jonesman T, Cerruti E, Paszkowski J. Mobilization of Pack-CACTA transposons in Arabidopsis suggests the mechanism of gene shuffling. Nucleic Acids Res. 2019;47: 1311–1320.

141. Galbraith JD, Hayward A. The influence of transposable elements on animal colouration. Trends Genet. 2023. doi:10.1016/j.tig.2023.04.005

142. Van't Hof AE, Campagne P, Rigden DJ, Yung CJ, Lingley J, Quail MA, et al. The industrial melanism mutation in British peppered moths is a transposable element. Nature. 2016;534: 102–105.

143. Ong-Abdullah M, Ordway JM, Jiang N, Ooi S-E, Kok S-Y, Sarpan N, et al. Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm. Nature. 2015;525: 533–537.

144. Ikeda Y, Pélissier T, Bourguet P, Becker C, Pouch-Pélissier M-N, Pogorelcnik R, et al. Arabidopsis proteins with a transposon-related domain act in gene silencing. Nat Commun. 2017;8: 15122.

145. Knip M, Hiemstra S, Sietsma A, Castelein M, de Pater S, Hooykaas P. DAYSLEEPER: a nuclear and vesicular-localized protein that is expressed in proliferating tissues. BMC Plant Biol. 2013;13: 211.

146. Hudson ME, Lisch DR, Quail PH. The FHY3 and FAR1 genes encode transposase-related proteins involved in regulation of gene expression by the phytochrome A-signaling pathway. Plant J. 2003;34: 453–471.

147. Huang S, Tao X, Yuan S, Zhang Y, Li P, Beilinson HA, et al. Discovery of an Active RAG Transposon Illuminates the Origins of V(D)J Recombination. Cell. 2016;166: 102–114.

148. Schrader L, Schmitz J. The impact of transposable elements in adaptive evolution. Mol Ecol. 2019;28: 1537–1549.

149. Hu Y, Wu X, Jin G, Peng J, Leng R, Li L, et al. Rapid Genome Evolution and Adaptation of Thlaspi arvense Mediated by Recurrent RNA-Based and Tandem Gene Duplications. Front Plant Sci. 2021;12: 772655.

150. Matsuno M, Compagnon V, Schoch GA, Schmitt M, Debayle D, Bassard J-E, et al. Evolution of a novel phenolic pathway for pollen development. Science. 2009;325: 1688–1692.

151. Makarevitch I, Waters AJ, West PT, Stitzer M, Hirsch CN, Ross-Ibarra J, et al. Transposable elements contribute to activation of maize genes in response to abiotic stress. PLoS Genet. 2015;11: e1004915.

152. Zhang Y, Li Z, Zhang Y ’e, Lin K, Peng Y, Ye L, et al. Evolutionary rewiring of the wheat transcriptional regulatory network by lineage-specific transposable elements. Genome Res. 2021;31: 2276–2289.

153. Zhao Q, Feng Q, Lu H, Li Y, Wang A, Tian Q, et al. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. Nat Genet. 2018;50: 278–284.

154. Woodhouse MR, Pedersen B, Freeling M. Transposed genes in Arabidopsis are often associated with flanking repeats. PLoS Genet. 2010;6: e1000949.

155. Hedges DJ, Deininger PL. Inviting instability: Transposable elements, double-strand breaks, and the maintenance of genome integrity. Mutat Res. 2007;616: 46–59.

156. Zhang J, Peterson T. A segmental deletion series generated by sister-chromatid transposition of Ac transposable elements in maize. Genetics. 2005;171: 333–344.

157. of Life Project Consortium TDT, Blaxter M, Mieszkowska N, Palma FD, Holland P, Durbin R, et al. Sequence locally, think globally: The Darwin Tree of Life Project. Proceedings of the National Academy of Sciences. 2022;119: e2115642118.

158. Bewick AJ, Ji L, Niederhuth CE, Willing E-M, Hofmeister BT, Shi X, et al. On the origin and evolutionary consequences of gene body DNA methylation. Proc Natl Acad Sci U S A. 2016;113: 9111–9116.

159. Gombar S, MacCarthy T, Bergman A. Epigenetics decouples mutational from environmental robustness. Did it also facilitate multicellularity? PLoS Comput Biol. 2014;10: e1003450.

160. Zemach A, McDaniel IE, Silva P, Zilberman D. Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science. 2010;328: 916–919.

161. Vigneau J, Borg M. The epigenetic origin of life history transitions in plants and algae. Plant Reprod. 2021;34: 267–285.

162. Pandey R, Müller A, Napoli CA, Selinger DA, Pikaard CS, Richards EJ, et al. Analysis of histone acetyltransferase and histone deacetylase families of Arabidopsis thaliana suggests functional diversification of chromatin modification among multicellular eukaryotes. Nucleic Acids Res. 2002;30: 5036–5055.

163. Chávez Montes RA, de Fátima Rosas-Cárdenas F, De Paoli E, Accerbi M, Rymarquis LA, Mahalingam G, et al. Sample sequencing of vascular plants demonstrates widespread conservation and divergence of microRNAs. Nat Commun. 2014;5: 3722.

164. You C, Cui J, Wang H, Qi X, Kuo L-Y, Ma H, et al. Conservation and divergence of small RNA pathways and microRNAs in land plants. Genome Biol. 2017;18: 158.

165. Niederhuth CE, Bewick AJ, Ji L, Alabady MS, Kim KD, Li Q, et al. Widespread natural variation of DNA methylation within angiosperms. Genome Biol. 2016;17: 194.

166. Bossdorf O, Richards CL, Pigliucci M. Epigenetics for ecologists. Ecol Lett. 2008;11: 106–115.

167. 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;15: e1008492.

168. Richards CL, Alonso C, Becker C, Bossdorf O, Bucher E, Colomé-Tatché M, et al. Ecological plant epigenetics: Evidence from model and non-model species, and the way forward. Ecol Lett. 2017;20: 1576–1590.

169. Alonso C, Pérez R, Bazaga P, Herrera CM. Global DNA cytosine methylation as an evolving trait: phylogenetic signal and correlated evolution with genome size in angiosperms. Front Genet. 2015;6: 4.

170. Dorn KM, Johnson EB, Daniels EC, Wyse DL, Marks MD. Spring flowering habit in field pennycress (Thlaspi arvense) has arisen multiple independent times. Plant Direct. 2018;2: e00097.

171. Warwick SI, Francis A, Susko DJ. The biology of Canadian weeds. 9. Thlaspi arvense L. (updated). Can J Plant Sci. 2002;82: 803–823.

172. Bayat S, Lysak MA, Mandáková T. Genome structure and evolution in the cruciferous tribe Thlaspideae (Brassicaceae). Plant J. 2021;108: 1768–1785.

173. G. I. Mc Intyre, Best KF. Studies on the Flowering of Thlaspi arvense L. IV. Genetic and Ecological Differences between Early- and Late-Flowering Strains. Bot Gaz. 1978;139: 190–195.

174. Dorn KM, Fankhauser JD, Wyse DL, Marks MD. De novo assembly of the pennycress (Thlaspi arvense) transcriptome provides tools for the development of a winter cover crop and biodiesel feedstock. Plant J. 2013;75: 1028–1038.

175. Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol Biol Evol. 2017;34: 1812–1819.

176. Dorn KM, Fankhauser JD, Wyse DL, Marks MD. A draft genome of field pennycress (Thlaspi arvense) provides tools for the domestication of a new winter biofuel crop. DNA Res. 2015;22: 121–131.

177. Cubins JA, Wells MS, Frels K, Ott MA, Forcella F, Johnson GA, et al. Management of pennycress as a winter annual cash cover crop. A review. Agron Sustain Dev. 2019;39: 46.

178. Barragan AC, Weigel D. Plant NLR diversity: the known unknowns of pan-NLRomes. Plant Cell. 2021;33: 814–831.

179. Michelmore RW, Meyers BC. Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. Genome Res. 1998;8: 1113–1130.

180. Cui H, Tsuda K, Parker JE. Effector-triggered immunity: from pathogen perception to robust defense. Annu Rev Plant Biol. 2015;66: 487–511.

181. Zipfel C. Plant pattern-recognition receptors. Trends Immunol. 2014;35: 345–351.

182. Jones JDG, Dangl JL. The plant immune system. Nature. 2006;444: 323–329.

183. Dangl JL, Jones JD. Plant pathogens and integrated defence responses to infection. Nature. 2001;411: 826–833.

184. Ngou BPM, Ding P, Jones JDG. Thirty years of resistance: Zig-zag through the plant immune system. Plant Cell. 2022;34: 1447–1478.

185. van der Biezen EA, Jones JD. The NB-ARC domain: a novel signalling motif shared by plant resistance gene products and regulators of cell death in animals. Curr Biol. 1998;8: R226–7.

186. Steele JFC, Hughes RK, Banfield MJ. Structural and biochemical studies of an NB-ARC domain from a plant NLR immune receptor. PLoS One. 2019;14: e0221226.

187. Krasileva KV, Dahlbeck D, Staskawicz BJ. Activation of an Arabidopsis resistance protein is specified by the in planta association of its leucine-rich repeat domain with the cognate oomycete effector. Plant Cell. 2010;22: 2444–2458.

188. Maruta N, Burdett H, Lim BYJ, Hu X, Desa S, Manik MK, et al. Structural basis of NLR activation and innate immune signalling in plants. Immunogenetics. 2022;74: 5–26.

189. Nishimura MT, Anderson RG, Cherkis KA, Law TF, Liu QL, Machius M, et al. TIR-only protein RBA1 recognizes a pathogen effector to regulate cell death in Arabidopsis. Proc Natl Acad Sci U S A. 2017;114: E2053–E2062.

190. Wang W, Devoto A, Turner JG, Xiao S. Expression of the membrane-associated resistance protein RPW8 enhances basal defense against biotrophic pathogens. Mol Plant Microbe Interact. 2007;20: 966–976.

191. Kroj T, Chanclud E, Michel-Romiti C, Grand X, Morel J-B. Integration of decoy domains derived from protein targets of pathogen effectors into plant immune receptors is widespread. New Phytol. 2016;210: 618–626.

192. Shao F, Golstein C, Ade J, Stoutemyer M, Dixon JE, Innes RW. Cleavage of Arabidopsis PBS1 by a bacterial type III effector. Science. 2003;301: 1230–1233.

193. Caldwell KS, Michelmore RW. Arabidopsis thaliana genes encoding defense signaling and recognition proteins exhibit contrasting evolutionary dynamics. Genetics. 2009;181: 671–684.

194. Zhang J, Li W, Xiang T, Liu Z, Laluk K, Ding X, et al. Receptor-like cytoplasmic kinases integrate signaling from multiple plant immune receptors and are targeted by a

Pseudomonas syringae effector. Cell Host Microbe. 2010;7: 290–301.

195. Bonardi V, Tang S, Stallmann A, Roberts M, Cherkis K, Dangl JL. Expanded functions for a family of plant intracellular immune receptors beyond specific recognition of pathogen effectors. Proc Natl Acad Sci U S A. 2011;108: 16463–16468.

196. Wu C-H, Abd-El-Haliem A, Bozkurt TO, Belhaj K, Terauchi R, Vossen JH, et al. NLR network mediates immunity to diverse plant pathogens. Proc Natl Acad Sci U S A. 2017;114: 8113–8118.

197. Wu C-H, Derevnina L, Kamoun S. Receptor networks underpin plant immunity. Science. 2018;360: 1300–1301.

198. Veronese P, Ruiz MT, Coca MA, Hernandez-Lopez A, Lee H, Ibeas JI, et al. In Defense against Pathogens. Both Plant Sentinels and Foot Soldiers Need to Know the Enemy,. Plant Physiol. 2003;131: 1580–1590.

199. Glazebrook J. Genes controlling expression of defense responses in Arabidopsis--2001 status. Curr Opin Plant Biol. 2001;4: 301–308.

200. Durrant WE, Dong X. Systemic acquired resistance. Annu Rev Phytopathol. 2004;42: 185–209.

201. Staiger D, Korneli C, Lummer M, Navarro L. Emerging role for RNA-based regulation in plant immunity. New Phytol. 2013;197: 394–404.

202. Gloggnitzer J, Akimcheva S, Srinivasan A, Kusenda B, Riehs N, Stampfl H, et al. Nonsense-mediated mRNA decay modulates immune receptor levels to regulate plant antibacterial defense. Cell Host Microbe. 2014;16: 376–390.

203. Ohta T. On the evolution of multigene families. Theor Popul Biol. 1983;23: 216–240.

204. Kent TV, Uzunović J, Wright SI. Coevolution between transposable elements and recombination. Philos Trans R Soc Lond B Biol Sci. 2017;372: 20160458.

205. Slotkin RK, Martienssen R. Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet. 2007;8: 272–285.

206. Tsuchiya T, Eulgem T. An alternative polyadenylation mechanism coopted to the *Arabidopsis RPP7* gene through intronic retrotransposon domestication. Proc Natl Acad Sci U S A. 2013;110: E3535–43.

207. Deng Y, Zhai K, Xie Z, Yang D, Zhu X, Liu J, et al. Epigenetic regulation of antagonistic receptors confers rice blast resistance with yield balance. Science. 2017;355: 962–965.

208. Zervudacki J, Yu A, Amesefe D, Wang J, Drouaud J, Navarro L, et al. Transcriptional control and exploitation of an immune-responsive family of plant retrotransposons. EMBO J. 2018;37. doi:10.15252/embj.201798482

209. López Sánchez A, Stassen JHM, Furci L, Smith LM, Ton J. The role of DNA (de)methylation in immune responsiveness of Arabidopsis. Plant J. 2016;88: 361–374.

210. Bailey PC, Schudoma C, Jackson W, Baggs E, Dagdas G, Haerty W, et al. Dominant integration locus drives continuous diversification of plant immune receptors with exogenous domain fusions. Genome Biol. 2018;19: 23.

211. Meyers BC, Chin DB, Shen KA, Sivaramakrishnan S, Lavelle DO, Zhang Z, et al. The major resistance gene cluster in lettuce is highly duplicated and spans several megabases. Plant Cell. 1998;10: 1817–1832.

212. Lee JM, Sonnhammer ELL. Genomic gene clustering analysis of pathways in eukaryotes. Genome Res. 2003;13: 875–882.

213. van Wersch S, Li X. Stronger When Together: Clustering of Plant NLR Disease resistance Genes. Trends Plant Sci. 2019;24: 688–699.

214. Liang W, van Wersch S, Tong M, Li X. TIR-NB-LRR immune receptor SOC3 pairs with truncated TIR-NB protein CHS1 or TN2 to monitor the homeostasis of E3 ligase SAUL1. New Phytol. 2019;221: 2054–2066.

215. Ratnaparkhe MB, Wang X, Li J, Compton RO, Rainville LK, Lemke C, et al. Comparative analysis of peanut NBS-LRR gene clusters suggests evolutionary innovation among duplicated domains and erosion of gene microsynteny. New Phytol. 2011;192: 164–178.

216. Kim S, Park J, Yeom S-I, Kim Y-M, Seo E, Kim K-T, et al. New reference genome sequences of hot pepper reveal the massive evolution of plant disease-resistance genes by retroduplication. Genome Biol. 2017;18: 210.

217. Rowan BA, Heavens D, Feuerborn TR, Tock AJ, Henderson IR, Weigel D. An Ultra High-Density Arabidopsis thaliana Crossover Map That Refines the Influences of Structural Variation and Epigenetic Features. Genetics. 2019;213: 771–787.

218. Jiang X, Assis R. Natural Selection Drives Rapid Functional Evolution of Young Drosophila Duplicate Genes. Mol Biol Evol. 2017;34: 3089–3098.

219. Gao Y, Wang W, Zhang T, Gong Z, Zhao H, Han G-Z. Out of Water: The Origin and Early Diversification of Plant R-Genes. Plant Physiol. 2018;177: 82–89.

220. Tarr DEK, Alexander HM. TIR-NBS-LRR genes are rare in monocots: evidence from diverse monocot orders. BMC Res Notes. 2009;2: 197.

221. Baggs E, Dagdas G, Krasileva KV. NLR diversity, helpers and integrated domains: making sense of the NLR IDentity. Curr Opin Plant Biol. 2017;38: 59–67.

222. Bakker EG, Toomajian C, Kreitman M, Bergelson J. A genome-wide survey of R gene polymorphisms in Arabidopsis. Plant Cell. 2006;18: 1803–1818.

223. 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;178: 1260–1272.e14.

224. Stam R, Scheikl D, Tellier A. Pooled Enrichment Sequencing Identifies Diversity and Evolutionary Pressures at NLR Resistance Genes within a Wild Tomato Population.

Genome Biol Evol. 2016;8: 1501–1515.

225. MacQueen A, Tian D, Chang W, Holub E, Kreitman M, Bergelson J. Population Genetics of the Highly Polymorphic RPP8 Gene Family. Genes . 2019;10. doi:10.3390/genes10090691

226. Kim M-S, Chae GY, Oh S, Kim J, Mang H, Kim S, et al. Comparative analysis of de novo genomes reveals dynamic intra-species divergence of NLRs in pepper. BMC Plant Biol. 2021;21: 247.

227. Thatcher S, Jung M, Panangipalli G, Fengler K, Sanyal A, Li B, et al. The NLRomes of Zea mays NAM founder lines and Zea luxurians display presence-absence variation, integrated domain diversity, and mobility. Mol Plant Pathol. 2023;24: 742–757.

228. Walkowiak S, Gao L, Monat C, Haberer G, Kassa MT, Brinton J, et al. Multiple wheat genomes reveal global variation in modern breeding. Nature. 2020;588: 277–283.

229. Bayer PE, Golicz AA, Scheben A, Batley J, Edwards D. Plant pan-genomes are the new reference. Nat Plants. 2020;6: 914–920.

230. Jiao W-B, Schneeberger K. Chromosome-level assemblies of multiple Arabidopsis genomes reveal hotspots of rearrangements with altered evolutionary dynamics. Nat Commun. 2020;11: 989.

231. Shirsekar G, Devos J, Latorre SM, Blaha A, Queiroz Dias M, González Hernando A, et al. Multiple Sources of Introduction of North American Arabidopsis thaliana from across Eurasia. Mol Biol Evol. 2021;38: 5328–5344.

232. Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. 2019;20: 275.

233. Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics. 2008;9: 18.

234. Ou S, Jiang N. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. Mob DNA. 2019;10: 48.

235. Ou S, Jiang N. LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. Plant Physiol. 2018;176: 1410–1422.

236. Xiong W, He L, Lai J, Dooner HK, Du C. HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. Proc Natl Acad Sci U S A. 2014;111: 10263–10268.

237. Su W, Gu X, Peterson T. TIR-Learner, a New Ensemble Method for TIR Transposable Element Annotation, Provides Evidence for Abundant New Transposable Elements in the Maize Genome. Mol Plant. 2019;12: 447–460.

238. Han Y, Wessler SR. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res.

2010;38: e199.

239. Cheng C-Y, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, Town CD. Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. Plant J. 2017;89: 789–804.

240. Rabanal FA, Gräff M, Lanz C, Fritschi K, Llaca V, Lang M, et al. Pushing the limits of HiFi assemblies reveals centromere diversity between two Arabidopsis thaliana genomes. Nucleic Acids Res. 2022;50: 12309–12327.

241. Hu K, Xu K, Wen J, Yi B, Shen J, Ma C, et al. Helitron distribution in Brassicaceae and whole Genome Helitron density as a character for distinguishing plant species. BMC Bioinformatics. 2019;20: 354.

242. Kapitonov VV, Tempel S, Jurka J. Simple and fast classification of non-LTR retrotransposons based on phylogeny of their RT domain protein sequences. Gene. 2009;448: 207–213.

243. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22: 1658–1659.

244. Zhang R-G, Li G-Y, Wang X-L, Dainat J, Wang Z-X, Ou S, et al. TEsorter: an accurate and fast method to classify LTR-retrotransposons in plant genomes. Hortic Res. 2022;9. doi:10.1093/hr/uhac017

245. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26: 841–842.

246. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2022. Available: https://www.R-project.org/.

247. Wickham H. ggplot2. Springer New York;

248. 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;327: 92–94.

249. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2019;35: 526–528.

250. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34: 3094–3100.

251. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, et al. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. 2012;40: D1202–10.

252. Feschotte C, Jiang N, Wessler SR. Plant transposable elements: where genetics meets genomics. Nat Rev Genet. 2002;3: 329–341.

253. Cossu RM, Casola C, Giacomello S, Vidalis A, Scofield DG, Zuccolo A. LTR

Retrotransposons Show Low Levels of Unequal Recombination and High Rates of Intraelement Gene Conversion in Large Plant Genomes. Genome Biol Evol. 2017;9: 3449–3462.

254. Rangwala SH, Elumalai R, Vanier C, Ozkan H, Galbraith DW, Richards EJ. Meiotically stable natural epialleles of Sadhu, a novel Arabidopsis retroposon. PLoS Genet. 2006;2: e36.

255. Magurran AE, Henderson PA. Explaining the excess of rare species in natural species abundance distributions. Nature. 2003;422: 714–716.

256. Yang L, Bennetzen JL. Distribution, diversity, evolution, and survival of *Helitrons* in the maize genome. Proc Natl Acad Sci U S A. 2009;106: 19922–19927.

257. Pereira V. Insertion bias and purifying selection of retrotransposons in the Arabidopsis thaliana genome. Genome Biol. 2004;5: R79.

258. Murat F, Louis A, Maumus F, Armero A, Cooke R, Quesneville H, et al. Understanding Brassicaceae evolution through ancestral genome reconstruction. Genome Biol. 2015;16: 262.

259. Goubert C, Craig RJ, Bilat AF, Peona V, Vogan AA, Protasio AV. A beginner's guide to manual curation of transposable elements. Mob DNA. 2022;13: 7.

260. Storer JM, Hubley R, Rosen J, Smit AFA. Curation Guidelines for de novo Generated Transposable Element Families. Curr Protoc. 2021;1: e154.

261. Yan N, Yang T, Yu X-T, Shang L-G, Guo D-P, Zhang Y, et al. Chromosome-level genome assembly of Zizania latifolia provides insights into its seed shattering and phytocassane biosynthesis. Commun Biol. 2022;5: 36.

262. Jiang R, Chen X, Liao X, Peng D, Han X, Zhu C, et al. A Chromosome-Level Genome of the Camphor Tree and the Underlying Genetic and Climatic Factors for Its Top-Geoherbalism. Front Plant Sci. 2022;13: 827890.

263. Usha T, Middha SK, Babu D, Goyal AK, Das AJ, Saini D, et al. Hybrid Assembly and Annotation of the Genome of the Indian Punica granatum, a Superfood. Front Genet. 2022;13: 786825.

264. Huang K, Ostevik KL, Elphinstone C, Todesco M, Bercovich N, Owens GL, et al. Mutation Load in Sunflower Inversions Is Negatively Correlated with Inversion Heterozygosity. Mol Biol Evol. 2022;39. doi:10.1093/molbev/msac101

265. Osmanski AB, Paulat NS, Korstian J, Grimshaw JR, Halsey M, Sullivan KAM, et al. Insights into mammalian TE diversity through the curation of 248 genome assemblies. Science. 2023;380: eabn1430.

266. Vergara Z, Sequeira-Mendes J, Morata J, Peiró R, Hénaff E, Costas C, et al. Retrotransposons are specified as DNA replication origins in the gene-poor regions of Arabidopsis heterochromatin. Nucleic Acids Res. 2017;45: 8358–8368.

267. Wlodzimierz P, Rabanal FA, Burns R, Naish M, Primetis E, Scott A, et al. Cycles of

satellite and transposon evolution in Arabidopsis centromeres. Nature. 2023;618: 557–565.

268. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27: 573–580.

269. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10: 421.

270. Dong Y, Lu X, Song W, Shi L, Zhang M, Zhao H, et al. Structural characterization of helitrons and their stepwise capturing of gene fragments in the maize genome. BMC Genomics. 2011;12: 609.

271. Roffler S, Menardo F, Wicker T. The making of a genomic parasite - the Mothra family sheds light on the evolution of Helitrons in plants. Mob DNA. 2015;6: 23.

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

273. Thomas J, Pritham EJ. Helitrons, the Eukaryotic Rolling-circle Transposable Elements. Microbiol Spectr. 2015;3. doi:10.1128/microbiolspec.MDNA3-0049-2014

274. Bowen NJ, McDonald JF. Drosophila euchromatic LTR retrotransposons are much younger than the host species in which they reside. Genome Res. 2001;11: 1527–1540.

275. Wicker T, Keller B. Genome-wide comparative analysis of copia retrotransposons in Triticeae, rice, and Arabidopsis reveals conserved ancient evolutionary lineages and distinct dynamics of individual copia families. Genome Res. 2007;17: 1072–1081.

276. Kapitonov V, Jurka J. The age of Alu subfamilies. J Mol Evol. 1996;42: 59–65.

277. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 1980;16: 111–120.

278. Garrison E, Guarracino A, Heumos S, Villani F, Bao Z, Tattini L, et al. Building pangenome graphs. bioRxiv. 2023. doi:10.1101/2023.04.05.535718

279. Devos KM, Brown JKM, Bennetzen JL. Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. Genome Res. 2002;12: 1075–1079.

280. Ma J, Devos KM, Bennetzen JL. Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. Genome Res. 2004;14: 860–869.

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

282. Karasov TL, Kniskern JM, Gao L, DeYoung BJ, Ding J, Dubiella U, et al. The long-term maintenance of a resistance polymorphism through diffuse interactions. Nature. 2014;512: 436–440.

283. Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR. Pack-MULE transposable elements mediate gene evolution in plants. Nature. 2004;431: 569–573.

284. Krasileva KV. The role of transposable elements and DNA damage repair mechanisms in gene duplications and gene fusions in plant genomes. Curr Opin Plant Biol. 2019;48: 18–25.

285. Leister D. Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance genes. Trends Genet. 2004;20: 116–122.

286. Reams AB, Roth JR. Mechanisms of gene duplication and amplification. Cold Spring Harb Perspect Biol. 2015;7: a016592.

287. Feschotte C. Transposable elements: McClintock's legacy revisited. Nat Rev Genet. 2023. doi:10.1038/s41576-023-00652-3

288. Goerner-Potvin P, Bourque G. Computational tools to unmask transposable elements. Nat Rev Genet. 2018;19: 688–704.

289. Mokhtar MM, Abd-Elhalim HM, El Allali A. A large-scale assessment of the quality of plant genome assemblies using the LTR assembly index. AoB Plants. 2023;15: lad015.

290. Stritt C, Thieme M, Roulin AC. Rare transposable elements challenge the prevailing view of transposition dynamics in plants. Am J Bot. 2021;108: 1310–1314.

291. Dupeyron M, Singh KS, Bass C, Hayward A. Evolution of Mutator transposable elements across eukaryotic diversity. Mob DNA. 2019;10: 12.

292. Dupeyron M, Baril T, Bass C, Hayward A. Phylogenetic analysis of the Tc1/mariner superfamily reveals the unexplored diversity of pogo-like elements. Mob DNA. 2020;11: 21.

293. Dupeyron M, Baril T, Hayward A. Broadscale evolutionary analysis of eukaryotic DDE transposons. bioRxiv. 2021. p. 2021.09.26.461848. doi:10.1101/2021.09.26.461848

294. Marx V. Method of the year: long-read sequencing. Nat Methods. 2023;20: 6–11.

295. Galanti D, Ramos-Cruz D, Nunn A, Rodríguez-Arévalo I, Scheepens JF, Becker C, et al. Genetic and environmental drivers of large-scale epigenetic variation in Thlaspi arvense. PLoS Genet. 2022;18: e1010452.

296. Paszkowski J. Controlled activation of retrotransposition for plant breeding. Curr Opin Biotechnol. 2015;32: 200–206.

297. Thieme, Bucher. Transposable elements as tool for crop improvement. Adv Bot Res. 2018. Available: https://www.sciencedirect.com/science/article/pii/S0065229618300454

298. Pietzenuk B, Markus C, Gaubert H, Bagwan N, Merotto A, Bucher E, et al. Recurrent evolution of heat-responsiveness in Brassicaceae COPIA elements. Genome Biol. 2016;17: 209.

299. Griffin PT, Niederhuth CE, Schmitz RJ. A Comparative Analysis of 5-Azacytidine- and Zebularine-Induced DNA Demethylation. G3 . 2016;6: 2773–2780.

300. Anderson R, Bayer PE, Edwards D. Climate change and the need for agricultural adaptation. Curr Opin Plant Biol. 2020;56: 197–202.

301. Shang L, Li X, He H, Yuan Q, Song Y, Wei Z, et al. A super pan-genomic landscape of rice. Cell Res. 2022;32: 878–896.

302. Haberer G, Kamal N, Bauer E, Gundlach H, Fischer I, Seidel MA, et al. European maize genomes highlight intraspecies variation in repeat and gene content. Nat Genet. 2020;52: 950–957.

303. Liu Y, Du H, Li P, Shen Y, Peng H, Liu S, et al. Pan-Genome of Wild and Cultivated Soybeans. Cell. 2020;182: 162–176.e13.

304. Gao L, Gonda I, Sun H, Ma Q, Bao K, Tieman DM, et al. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. Nat Genet. 2019;51: 1044–1051.

305. Tang D, Jia Y, Zhang J, Li H, Cheng L, Wang P, et al. Genome evolution and diversity of wild and cultivated potatoes. Nature. 2022;606: 535–541.

306. Guo Y-L, Fitz J, Schneeberger K, Ossowski S, Cao J, Weigel D. Genome-wide comparison of nucleotide-binding site-leucine-rich repeat-encoding genes in Arabidopsis. Plant Physiol. 2011;157: 757–769.

Appendix

Plant Biotechnology Journal Open Access

Plant Biotechnology Journal (2022) **20**, pp. 944–963

doi: 10.1111/pbi.13775

Chromosome-level *Thlaspi arvense* genome provides new tools for translational research and for a newly domesticated cash cover crop of the cooler climates

Adam Nunn[1,2], Isaac Rodríguez-Arévalo[3,4], Zenith Tandukar[5], Katherine Frels[5,6], Adrián Contreras-Garrido[7], Pablo Carbonell-Bejerano[7], Panpan Zhang[8,9], Daniela Ramos Cruz[3,4], Katharina Jandrasits[3,4], Christa Lanz[7], Anthony Brusa[5], Marie Mirouze[8,9], Kevin Dorn[10,11], David W Galbraith[12], Brice A. Jarvis[13], John C. Sedbrook[13], Donald L. Wyse[5], Christian Otto[1], David Langenberger[1], Peter F. Stadler[2,14], Detlef Weigel[7], M. David Marks[10], James A. Anderson[5], Claude Becker[3,4,*] and Ratan Chopra[5,10,*]

[1] *ecSeq Bioinformatics GmbH, Leipzig, Germany*

[2] *Department of Computer Science, Leipzig University, Leipzig, Germany*

[3] *Genetics, Faculty of Biology, Ludwig Maximilians University, Martinsried, Germany*

[4] *Gregor Mendel Institute of Molecular Plant Biology GmbH, Austrian Academy of Sciences (ÖAW), Vienna BioCenter (VBC), Vienna, Austria*

[5] *Department of Agronomy and Plant Genetics, University of Minnesota, Saint Paul, MN, USA*

[6] *Department of Agronomy and Horticulture, University of Nebraska, Lincoln, NE, USA*

[7] *Department of Molecular Biology, Max Planck Institute for Developmental Biology, Tübingen, Germany*

[8] *Institut de Recherche pour le Développement, UMR232 DIADE, Montpellier, France*

[9] *Laboratory of Plant Genome and Development, University of Perpignan, Perpignan, France*

[10] *Department of Plant and Microbial Biology, University of Minnesota, Saint Paul, MN, USA*

[11] *USDA-ARS, Soil Management and Sugarbeet Research, Fort Collins, CO, USA*

[12] *BIO5 Institute, Arizona Cancer Center, Department of Biomedical Engineering, University of Arizona, School of Plant Sciences, Tucson, AZ, USA*

[13] *School of Biological Sciences, Illinois State University, Normal, IL, USA*

[14] *Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany*

Received 17 August 2021;
revised 28 November 2021;
accepted 23 December 2021.

**Correspondence* (Tel +49 89 2180 74740; fax +49 89 2180 74702; email: claude.becker@biologie.uni-muenchen.de (CB); Tel +1 551-689-5299; fax +1 612-625-1268; email: rchopra@umn.edu (RC))

Keywords: pennycress, genome assembly, genome annotations, genetic mapping, comparative genomics.

Summary

Thlaspi arvense (field pennycress) is being domesticated as a winter annual oilseed crop capable of improving ecosystems and intensifying agricultural productivity without increasing land use. It is a selfing diploid with a short life cycle and is amenable to genetic manipulations, making it an accessible field-based model species for genetics and epigenetics. The availability of a high-quality reference genome is vital for understanding pennycress physiology and for clarifying its evolutionary history within the Brassicaceae. Here, we present a chromosome-level genome assembly of var. MN106-Ref with improved gene annotation and use it to investigate gene structure differences between two accessions (MN108 and Spring32-10) that are highly amenable to genetic transformation. We describe non-coding RNAs, pseudogenes and transposable elements, and highlight tissue-specific expression and methylation patterns. Resequencing of forty wild accessions provided insights into genome-wide genetic variation, and QTL regions were identified for a seedling colour phenotype. Altogether, these data will serve as a tool for pennycress improvement in general and for translational research across the Brassicaceae.

Introduction

Native to Eurasia, field pennycress (*Thlaspi arvense* L.) is a member of the Brassicaceae family and is closely related to the oilseed crop species rapeseed (*Brassica rapa* and *Brassica napus* L.), camelina (*Camelina sativa* L.) and the wild plant *Arabidopsis thaliana* (Beilstein *et al*., 2010; Warwick *et al*., 2002). It is an emerging oil feedstock species with the potential to improve sustainability of cold climate cropping systems through use as a cash cover crop (Boateng *et al*., 2010; Chopra *et al*., 2018; Sedbrook *et al*., 2014). Pennycress is extremely winter hardy (Warwick *et al*., 2002) and can be planted in traditional fallow periods following summer annuals such as wheat, maize or soya bean (Cubins *et al*., 2019; Johnson *et al*., 2015; Ott *et al*., 2019; Phippen and Phippen, 2012). By providing a protective living cover from the harvest of the previous summer annual crop through early spring, pennycress prevents soil erosion and nutrient loss, which in turn protects surface and below-ground water sources, suppresses early-season weed growth, and provides a food source for pollinators (Del Gatto *et al*., 2015; Johnson *et al*., 2015; Weyers *et al*., 2019, 2021). The short life cycle allows for harvest in May or June in temperate regions, with reported seed yields ranging from 750 to 2400 kg/ha (Cubins *et al*., 2019; Moore *et al*., 2020). Following harvest, an additional crop of summer annuals can be grown in a double-crop system that provides increased total seed yields and beneficial ecosystem services (Johnson *et al*., 2015; Phippen and Phippen, 2012; Thomas *et al*., 2017). The pennycress seed contains an average of 30%–35% oil, and the fatty acid profile is conducive to producing biofuels (Fan *et al*., 2013; Moser, 2012; Moser *et al*., 2009). Seed oil also has the potential to be converted into an edible oil and protein source (Chopra *et al*., 2020b; Claver *et al*., 2017; McGinn *et al*., 2019).

Thlaspi arvense is a homozygous diploid species ($2n = 2x = 14$) (Mulligan, 1957) and is predominantly self-pollinating (Mulligan and Kevan, 1973), suggesting that breeding efforts could proceed with relative ease and speed. It is amenable to genetic transformation using the floral dip method (McGinn *et al*., 2019), and its diploid nature with many one-to-one gene correspondence with *A. thaliana* (Chopra *et al*., 2018) could provide an avenue for gene discovery followed by field-based phenotypic validation. Indeed, several agronomic and biochemical traits have already been identified in pennycress using this translational approach, including traits crucial for *de novo* domestication of *T. arvense* such as transparent testa phenotypes (Chopra *et al*., 2018), early flowering (Chopra *et al*., 2020b), reduced shatter (Chopra *et al*., 2020b) and seed oil composition traits (Chopra *et al*., 2020b; Esfahanian *et al*., 2021; Jarvis *et al*., 2021; McGinn *et al*., 2019). Field pennycress could thus serve as a *de novo*-domesticated oilseed crop for the cooler climates of the world and at the same time as a new dicotyledonous model for functional genetics studies. Its amenability for translational research constitutes a clear advantage vis-a-vis *A. thaliana*. However, to establish *T. arvense* as a genetic model and a crop, it is important to develop genomic resources that will help explore the spectrum of genetic diversity, the extent and patterns of gene expression, genetic structure and untapped genetic potential for crop improvement.

Here, we describe a set of new resources developed for research and breeding communities, including a high-quality, chromosome-level genome assembly of *T. arvense* var. MN106-Ref, representing ~97.5% of the estimated genome size of 539 Mbp. We provide robust annotations of both protein-coding and non-coding genes, including putative transfer RNA (tRNA), ribosomal RNA (rRNA) and small nucleolar RNA (snoRNA) predictions, alongside small RNA-producing loci, transposable element (TE) families and predicted pseudogenes. From transcriptome data based on a panel of eleven different tissues and life stages, we built a gene expression atlas. In combination with whole-genome DNA methylation profiles of both roots and shoots, this provides a basis for exploring gene regulatory and/or epigenetic mechanisms within pennycress. A comprehensive analysis of forty resequenced pennycress accessions highlights the nucleotide diversity in these collections, alongside gene variants and population structure. Finally, by means of modified bulked-segregant analysis (BSA), we identified quantitative trait loci (QTL) associated with seedling colour phenotype, exemplifying the usefulness of this resource. The genome and resequencing information presented in this study will increase the value of pennycress as a model and as tool for translational research and accelerate pennycress breeding through the discovery of genes affecting important agronomic traits.

Results

An improved reference genome sequence

The genome of *T. arvense* var. MN106-Ref was assembled *de novo* from 476X (256 Gb) depth PacBio Sequel II continuous long reads (CLRs) (38 kb N50). The initial assembly attempts exceeded the genome size by ~53% with respect to the range of 459–540 Mbp total size estimated from flow cytometry and k-mer analyses (Table S1). Reducing the duplicated fraction, polishing and scaffolding/rescaffolding using several approaches resulted in a final assembly of ~526 Mbp, corresponding to ~97.5% of the upper limit of the flow cytometry-based estimate and representing an improvement of ~20% relative to the original assembly size. Scaffolding/rescaffolding of the genome assembly was achieved using Bionano optical, Hi-C contact, genetic linkage and comparative synteny maps. The final genome contains 964 scaffolds, with ~83.6% of the total estimated size represented by seven large scaffolds, in agreement with the haploid chromosome number, demonstrating a vast improvement in overall contiguity and bringing the assembly to chromosome level. The coding space is 98.7% complete on the basis of conserved core eukaryotic single-copy genes (BUSCO), with 92.1% being single copy and 6.6% duplicated. Full descriptive statistics of the final version in comparison with T_arvense_v1 are given in Table 1; intermediary versions are summarized in Table S2.

The seven largest scaffolds are all characterized by high gene density towards both telomeres and a high density of repeats and TEs in the pericentromeric and centromeric regions (Figure 1, Figure S1). While the protein-coding gene fraction of the genome is similar in size to other closely related Brassicaceae (Wang *et al*., 2011), the large repetitive fraction suggests an increased genome size driven by TE expansion (Beric *et al*., 2021). In addition, the spatial distribution of sRNA loci followed the gene density but was concentrated predominantly at the boundary between genes and TEs.

In addition to the duplicate-containing contigs, alignments of the raw CLR reads to the new genome revealed the presence of what appeared to be a small number of collapsed repeats in scaffolds 1, 3, 5 and 7, which were typically larger than 25 kbp and indicative of misassembly in these loci (Figure S2). Further investigation revealed an overlap with tandem repeat clusters of 18S and 28S rRNA annotations at those loci on scaffolds 3 and 5, and a large supersatellite of 5S rRNA on scaffold 1. In addition,

Table 1 Full descriptive statistics comparing the previously published T_arvense_v1 assembly with the present version T_arvense_v2

Assembly category	T_arvense_v1	T_arvense_v2
No. of contigs	44 109	4714
Largest contig	–	41.6 Mbp
contig N50	0.02 Mbp	13.3 Mbp
No. of scaffolds	6768	964
No. of scaffolds (≥50 000 bp)	1807	607
Largest scaffold	2.4 Mbp	70.0 Mbp
Total length	343 Mbp	526 Mbp
Total length (≥50 000 bp)	276 Mbp	514 Mbp
GC (%)	37.99	38.39
N50	0.14 Mbp	64.9 Mbp
NG50	0.05 Mbp	64.9 Mbp
N75	0.06 Mbp	61.0 Mbp
NG75	–	55.2 Mbp
L50	561	4
LG50	1678	4
L75	1469	6
LG75	–	7
No. of Ns per 100 kbp	5165.00	0.51

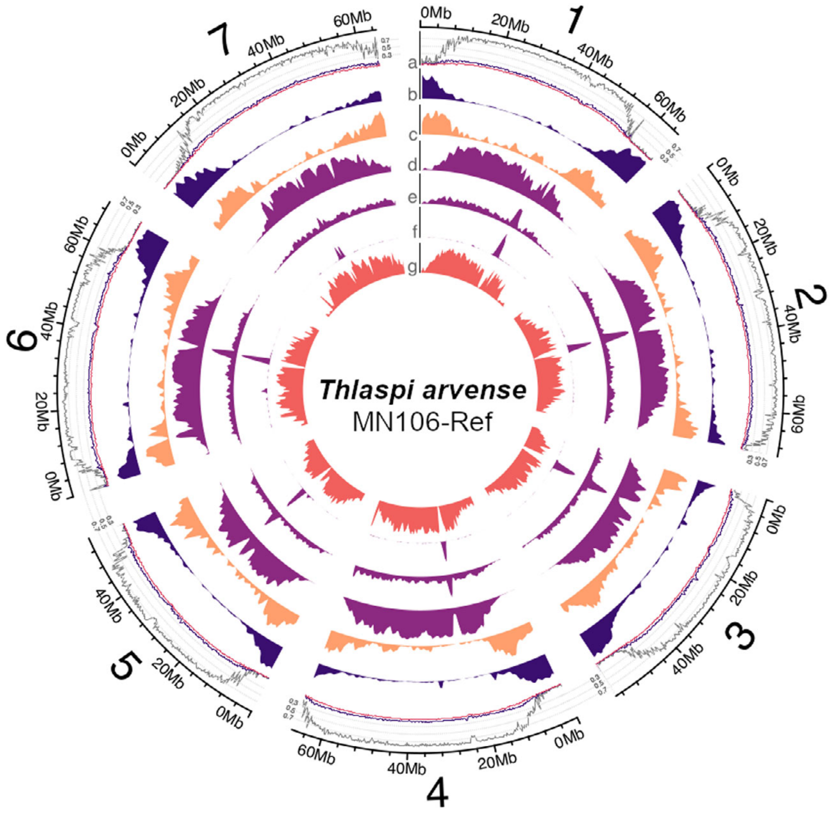


Figure 1 Overview of the seven largest scaffolds representing chromosomes in *T. arvense* var. MN106-Ref. The tracks denote (a) DNA methylation level in shoot tissue (CG: grey; CHG: black; CHH: pink; 200 kbp window size), and density distributions (1 Mbp window size) of (b) protein-coding loci, (c) sRNA loci, (d) Gypsy retrotransposons, (e) Copia retrotransposons, (f) LTR retrotransposons and (g) pseudogenes.

there were corresponding genes associated with organellar DNA at those loci on scaffolds 3 and 7, indicating either erroneous incorporation of plastome sequence during assembly or genuine nuclear integrations of plastid DNA (NUPTs) (Michalovova *et al*., 2013).

Comparative genomics

Exploiting information from the genome of *Eutrema salsugineum* (Yang *et al*., 2013), a closely related species (Franzke *et al*., 2011) with a much smaller genome (241 Mbp) but the same karyotype (n = 7), aided during rescaffolding (see methods; Figure S3) and confirmed synteny of the seven largest scaffolds in the two species (Figure S4). There is a large-scale synteny between the two genomes, with the exception of some regions on scaffolds 2, 3, 6 and 7. This could be due to the low gene density observed in the *T. arvense* genome towards the centre of each chromosome and/or the high presence of dispersed repeats in those regions.

Chromosome evolution in the Brassicaceae has been studied through chromosome painting techniques, and 24 chromosome blocks (A-X) have been defined from an ancestral karyotype of n = 8 (Murat *et al*., 2015; Schranz *et al*., 2006). We identified the 24 blocks in *T. arvense* based on gene homology and synteny between *T. arvense* and *A. thaliana* (Figure 2). While in general the distribution of the chromosomal blocks resembles that in the close relatives *E. salsugineum* and *S. parvula*, some blocks are rearranged in a small section at the end of the scaffold representing chromosome 1 and at the beginning of chromosome 6. The first case involves the transposition of a small part of block C in between A and B, while chromosome 6 has a possible inversion between the blocks O and W when compared to *E. salsugineum* and *S. parvula*. Overall, despite having an increase in genome size compared with *E. salsugineum* and *S. parvula*, *T. arvense* conserves all the ancestral Brassicaceae karyotype blocks. The synteny analysis also revealed intra-chromosomal rearrangements, but no obvious inter-chromosomal rearrangements.

Genome annotation

Transcriptome assembly

We sequenced total cDNA with strand-specific RNA-seq from eleven tissues, including rosette leaves, cauline leaves, inflorescences, open flowers, young green siliques, old green siliques, green seeds, mature seeds, seed pods, roots of 1-week-old seedlings and shoots of 1-week-old seedlings (Table S3). Reads from each tissue sample were aligned to the genome with unique mapping rates between 76% and 91%, with the exception of old green silique (19%), green seed (59%) and mature seed (12%). The majority of unmapped reads in each case were due to insufficient high-quality read lengths. We constructed independent tissue-specific transcriptome assemblies and combined them into a multi-sample *de novo* assembly, yielding 30 650 consensus transcripts. These were further refined by prioritizing isoforms supported by Iso-seq data, resulting in 22 124 high-quality consensus transcripts to inform gene models.

Protein-coding genes

In addition to the expression data, gene models were informed by protein homology using a combined database of Viridiplantae from UniProtKB/Swiss-Prot (Boutet *et al*., 2007) and selected Brassicaceae from RefSeq (Pruitt *et al*., 2012). Following initial

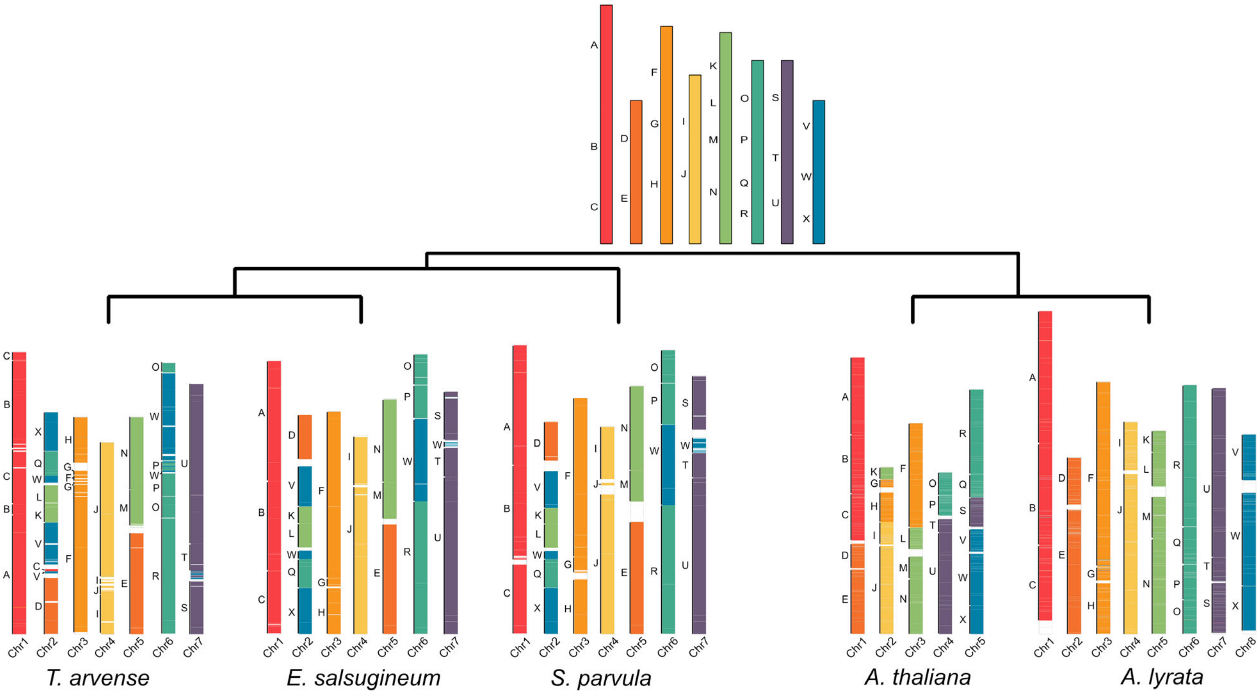


Figure 2 Distribution of ancestral genomic blocks (top panel) along the seven largest scaffolds of *T. arvense* MN106-Ref (T_arvense_v2), and a comparison of these genomic blocks with *Eutrema salsugineum*, *Schrenkiella parvula, Arabidopsis thaliana* and *Arabidopsis lyrata*.

training and annotation by *ab initio* gene predictors, protein-coding loci were further annotated with InterPro to provide PFAM domains, which were combined with a BLAST search to the UniProtKB/Swiss-Prot Viridiplantae database to infer gene ontology (GO) terms. In accordance with MAKER-P recommendations (Campbell *et al*., 2014), the final set of 27 128 protein-coding loci was obtained by filtering out those with an annotation edit distance (AED) score of 1 unless they also contained a PFAM domain. Approximately 95% of loci had an AED score <0.5 (Figure S5), demonstrating a high level of support with the available evidence, and 21 171 (~78%) were annotated with a PFAM domain. Analysis of gene orthologs and paralogs among related Brassicaceae confirmed the close relationship with *E. salsugineum*, with the protein-coding fraction occupying a genome space comparable to related species (Figure 3a). A total of 4433 gene duplication events were recorded with OrthoFinder, comparable to *E. salsugineum* (5108), but fewer than in *B. rapa* (11 513), for example.

The full descriptive statistics are given in Table 2, in comparison with the original T_arvense_v1 annotation (Dorn *et al*., 2015) lifted over to the new genome with Liftoff v1.5.2 (Shumate and Salzberg, 2020), where applicable. Gene feature distributions are comparable between T_arvense_v1 and the present assembly of MN106-Ref (hereafter referred to as T_arvense_v2; Figure S6). Unique genes that were successfully lifted over from the previous version were included as a separate fraction in the final annotation (source: T_arvense_v1), resulting in 32 010 annotated genes in total. Up to ~95.2% completeness can be obtained by combining the full set of both the current and previous annotations according to a BUSCO evaluation of 2121 conserved, single-copy orthologs. The improved contiguity of the genome space allowed for the resolution of genes such as the tandem duplicated *MYB29* and *MYB76*, which were concatenated in the previous version (Figure 3b).

Non-coding loci

In addition to the protein-coding gene annotations, we annotated non-coding RNA (ncRNA) genes, pseudogenes, and TEs. Descriptive annotation statistics are summarized in Table 2. While many of these annotation features in *T. arvense* were similar to those found in other plant species, we observed several unique patterns, which we will describe in detail below. ncRNA annotations were inferred from sequence motifs (tRNA, rRNA, snoRNA) or from sequencing data (siRNA, miRNA). We predicted clusters of both 5S rRNA and tandem repeat units of 18S and 28S rRNA with RNAmmer (Lagesen *et al*., 2007), often in relative proximity to loci identified with Tandem Repeats Finder v4.09.1 (Benson, 1999) and putatively associated with centromeric repeat motifs (not shown). Of the largest seven scaffolds, only scaffolds 4 and 7 carried no such annotations. Notably, several large clusters of 5S rRNA genes were interspersed throughout the pericentromeric region of scaffold 1, whereas the remaining four scaffolds contained 18S and 28S rRNA gene annotations. Finally, we identified 243 homologs from 114 snoRNA families.

sRNA annotation

We identified 19 386 siRNA loci. More than 98% of these loci corresponded to heterochromatic 23- to 24-nt siRNA loci, with only 196 producing 20- to 22-nt siRNAs. The sRNA loci were expressed unevenly across tissues, as inferred from prediction with data from different tissues. Only 2938 loci were shared across all four tissues studied (rosette leaves, roots, inflorescences and pollen). Inflorescences were the major contributor with 6728 private loci. Despite these differences between tissues, we

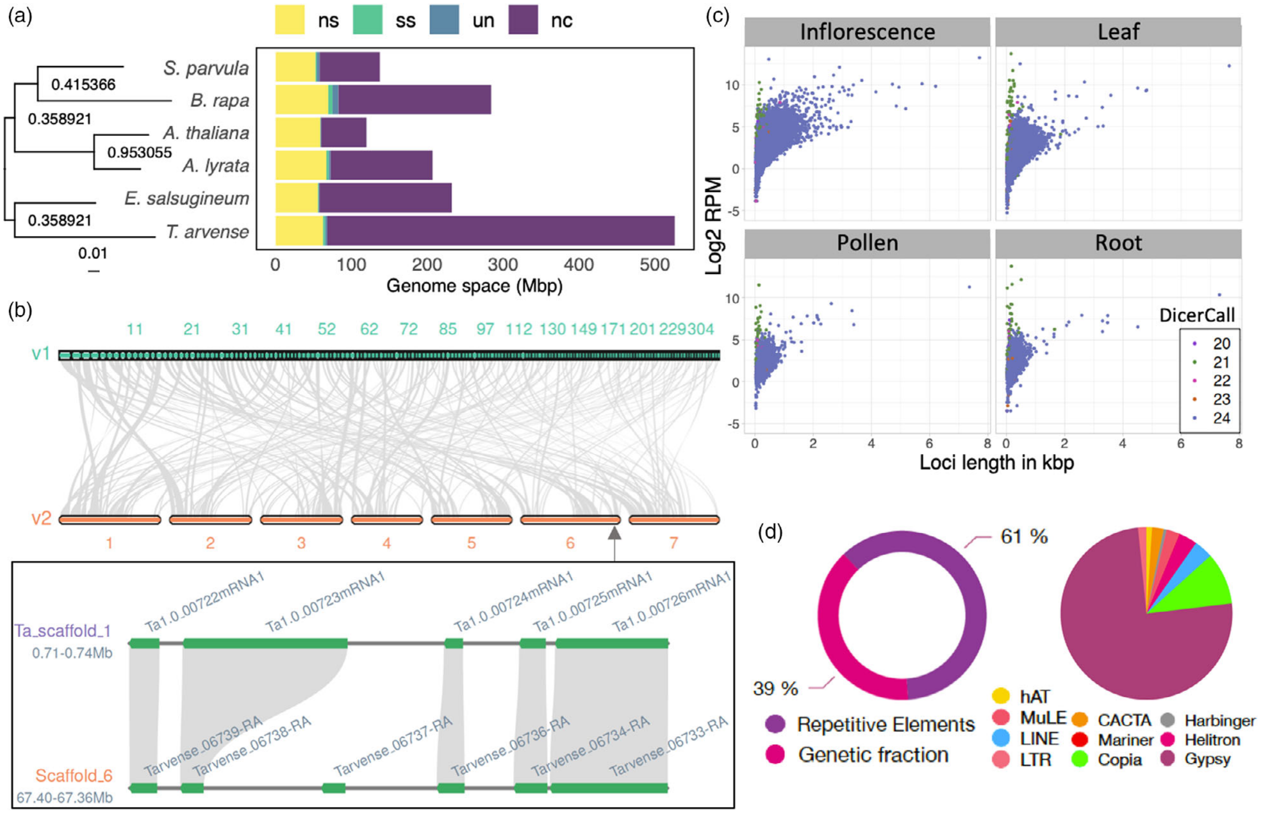


Figure 3 Feature annotations within *T. arvense* var MN106-Ref. (a) Rooted species tree inferred from all genes, denoting node support and branch length in substitutions per site, and horizontal stacked bar chart comparing the genetic fraction in pennycress with other *Brassicaceae* sp. (ns = nonspecific orthologs, ss = species-specific orthologs, un = unclassified genes, nc = non-coding/intergenic fraction). (b) Comparison of gene macrosynteny between v1 and v2 of the genome, and a microsynteny example of genes *MYB29* and *MYB76*, which are resolved in the v2 annotation. (c) Small RNA biogenesis locus length and expression values in each of four tissues. (d) Overall repetitive content in the genome as discovered by RepeatMasker2, and relative abundance of TEs within the fraction of repetitive elements.

observed similar overall patterns in terms of locus length, expression (Figure 3c) and complexity (Figure S7).

Altogether, sRNA loci accounted for ~8 Mbp or ~1.5% of the assembled genome. Of the seven largest scaffolds, where the majority of genes are located, the total coverage of siRNA loci ranged between 1.5% and 2% and the loci appeared to be preferentially concentrated at the boundary between TEs and the protein-coding gene fraction of the genome. To further explore this, we partitioned the seven largest scaffolds into gene-enriched and gene-depleted regions, based on a median of 14 genes per Mbp and a mean of 54.2 genes per Mbp. We defined gene-enriched loci as those above and gene-depleted loci as those below the mean. At the chromosomal level, sRNA loci correlated with gene-enriched regions and were scarce in regions with high TE content. This trend is in contrast to that observed in *A. thaliana* (Hardcastle *et al.*, 2018) but resembles what has been observed, for example, in maize (He *et al.*, 2013) and tomato (Tomato Genome Consortium, 2012).

Phased secondary siRNAs (phasiRNAs) are a class of secondary sRNAs that, due to the way they are processed, produce a distinct periodical pattern of accumulation (Axtell, 2013b). In the *T. arvense* genome, we observed 139 loci with such phased patterns. In contrast to the general notion that phasiRNAs are typically 21 nt long (Lunardon *et al.*, 2020), we found 24-nt siRNAs to be dominant in 133 of these loci.

MicroRNAs

MicroRNA (miRNA)-encoding genes were predicted using a combination of ShortStack and manual curation (see Methods). We identified 72 miRNA-producing loci, with 53 that were already known from other species, and 19 appeared to be species-specific. Most of the identified families were produced from only one or two loci, with miR156 and miR166 being produced by the most loci, with eight and five family members, respectively. A total of 21 out of 25 families in *T. arvense* are found in other rosids, and three (miR161, miR157 and miR165) only in other Brassicaceae. One family, miR817, is also present in rice. There is a strong preference for 5′-U at the start of both unique and conserved miRNAs (Figure S8), in line with previous reports (Voinnet, 2009). The expression level of both conserved and novel miRNA families was compared between tissues, showing that the ten most highly expressed across all tissues are conserved families, whereas novel miRNA demonstrates a marginal tendency to be more lowly expressed or with potential for differential expression (Figure S9).

sRNA loci

When we overlaid the sRNA loci with our annotated genomic features, most sRNAs localized to the intergenic space, but a substantial fraction, especially 20- to 22-nt sRNAs, were produced

Table 2 Summary of feature annotations in comparison with the original version T_arvense_v1

Type	T_arvense_v1	T_arvense_v2	diff.
(A) Protein-coding genes			
Total number of loci	27 390	27 128	-262
Total number of unique loci	4780	5034	+254
Total number of transcript isoforms	–	30 650	+30 650
Number of matching loci with changes in CDS	–	–	+14 102
Number of matching loci with changes in UTR(s)	–	–	+22 559
Loci containing one or more PFAM domain	–	21 171	+21 171
Loci annotated with one or more GO term	–	13 074	+13 074
(B) Non-coding genes			
tRNA	–	1148	+1148
rRNA clusters (<25 kbp)	–	63	+63
snoRNA	–	243	+243
Small interfering RNA (siRNA)	–	19 373	+19 373
MicroRNA (miRNA)	–	72	+72
(C) Other gene types			
Pseudogenes (set II Ψs)	–	44 490	+44 490
Transposable element genes	–	423 251	+423 251

from intronic sequences (Figure S10a). Helitrons make up only 1.5% of the genome space, yet more than 5% of sRNA biogenesis loci overlap with this type of TE. Most sRNA loci (93.0%) fell within 1.5 kbp of annotated genes or TEs (Figure S10b,c). As expected, 23- to 24-nt sRNAs were more frequently associated with TEs, whereas 20- to 22-nt sRNAs were more often produced by coding genes (Axtell, 2013a).

Pseudogenes

In accordance with the MAKER-P protocol, pseudogenes (Ψ) were predicted in intergenic DNA with the ShiuLab pseudogene pipeline (Zou *et al.*, 2009). A total of 44 490 set II pseudogenes were annotated, exceeding those in *A. thaliana* (~3700) or rice (~7900) by one order of magnitude. We identified 35 818 pseudogenes overlapping with TEs, and 8672 pseudogenes that were either concentrated in intergenic space or more towards the protein-coding gene complement of the genome, and thus perhaps less likely to have arisen from retrotransposition. Approximately 59.2% of these contained neither a non-sense nor a frameshift mutation, indicating either (i) that the regulatory sequences of the pseudogenes were silenced first, (ii) a pseudo-exon that may be linked to another non-functional exon, or (iii) a possible undiscovered gene.

Transposable elements

In total, we identified 423 251 TEs belonging to 10 superfamilies and covering ~61% of the genome (Figure 3d). Retrotransposons (75% of all TEs are Gypsy elements; 10% Copia; 4% LINE) by far outnumbered DNA transposons (3% Helitrons; 1% hAT; 2% CACTA; 1% Pif-Harbinger; 2% MuLE). A detailed breakdown of repeats is shown in Table S4. As the most abundant retrotransposon superfamily, Gypsy elements accounted for 46% of the total genome space, which is consistent with a high abundance observed in the pericentromeric heterochromatin of *E. salsugineum*, where centromere expansion is thought to have been caused by Gypsy proliferation (Zhang *et al.*, 2020). In addition, we identified 359 protein-coding genes located fully within TE bodies that could represent Pack-TYPE elements and contribute to gene shuffling (Catoni *et al.*, 2018). Among these elements, 153 were intersecting with mutator-like elements, suggesting they correspond to Pack-MULE loci. TEs were located primarily in low gene density regions, while the fraction of TE-contained genes was randomly distributed.

Expression atlas

With cDNA sequences from 11 different tissues or developmental stages, we could annotate tissue-specific expression patterns. The complete expression atlas is provided in Data S1. We evaluated the relative extent of tissue-specific gene expression using the Tau (τ) algorithm (Yanai *et al.*, 2005), from the normalized trimmed mean of *M*-value (TMM) counts in all tissues (Robinson and Oshlack, 2010). To preclude potential biases caused by substantial differences in library size, we excluded low-coverage samples from mature seeds and old green siliques. In total, 4045 genes had high or even complete tissue specificity (τ = 0.8–1.0), while 5938 genes had intermediate specificity (0.2–0.8) and 6107 had no or low specificity (0–0.2); the remaining genes were ignored due to missing data. The relative breakdown of each specificity fraction by tissue type is shown in Figure 4a, with 'roots', 'green seeds' and 'inflorescences' representing the tissues with the greatest proportion of high or complete specificity genes. The relative $\log_2$(TMM) expression values of the top 30 most highly expressed genes in each tissue, given a high or complete specificity score, are plotted in Figure 4b with respect to the overall mean expression per gene across all included tissues. These include, for example, genes with homology to *EXTENSIN 2* (*EXT2; A. thaliana*) in 'roots', *CRUCIFERIN* (*BnC1; B. napus*) in 'green seeds', and *PECTINESTERASE INHIBITOR 1* (*PMEI1; A. thaliana*) in 'inflorescences' and 'open flowers' (Data S2).

DNA methylation

Cytosine methylation (also commonly referred to as DNA methylation) is a prevalent epigenetic mark in plant genomes and is often associated with heterochromatin and transcriptional inactivation of TEs and promoters, but also with higher and more stable expression when present in gene bodies (Zhang *et al.*, 2018). In plants, DNA methylation occurs in three cytosine contexts, CG, CHG and CHH (where H is any base but G), with the combined presence of CG, CHG and CHH methylation usually indicative of heterochromatin formation and TE silencing, while gene body methylation consists only of CG methylation (Bewick and Schmitz, 2017). In the light of the high TE density in *T. arvense*, we analysed genome-wide DNA methylation by whole-genome bisulphite sequencing (WGBS) in shoots and roots of 2-week-old seedlings. Genome-wide, 70% of cytosines were methylated in the CG context, 47% in the CHG context and 33% in the CHH context. In line with findings in other Brassicaceae, methylation at CG sites was consistently higher than at CHG and CHH (Figure 1a; Figure S11). When we compared the WGBS data against the genome annotation, high levels of DNA methylation (mostly ^{m}CG) colocalized with regions of dispersed repeats and TEs in the centre of the chromosomes. Conversely, methylation was depleted in gene-rich regions (Figure 1a,b). In line with this, DNA methylation was consistently high along TEs, particularly in the CG context (Figure 4c). In contrast to *E. salsugineum* (Bewick

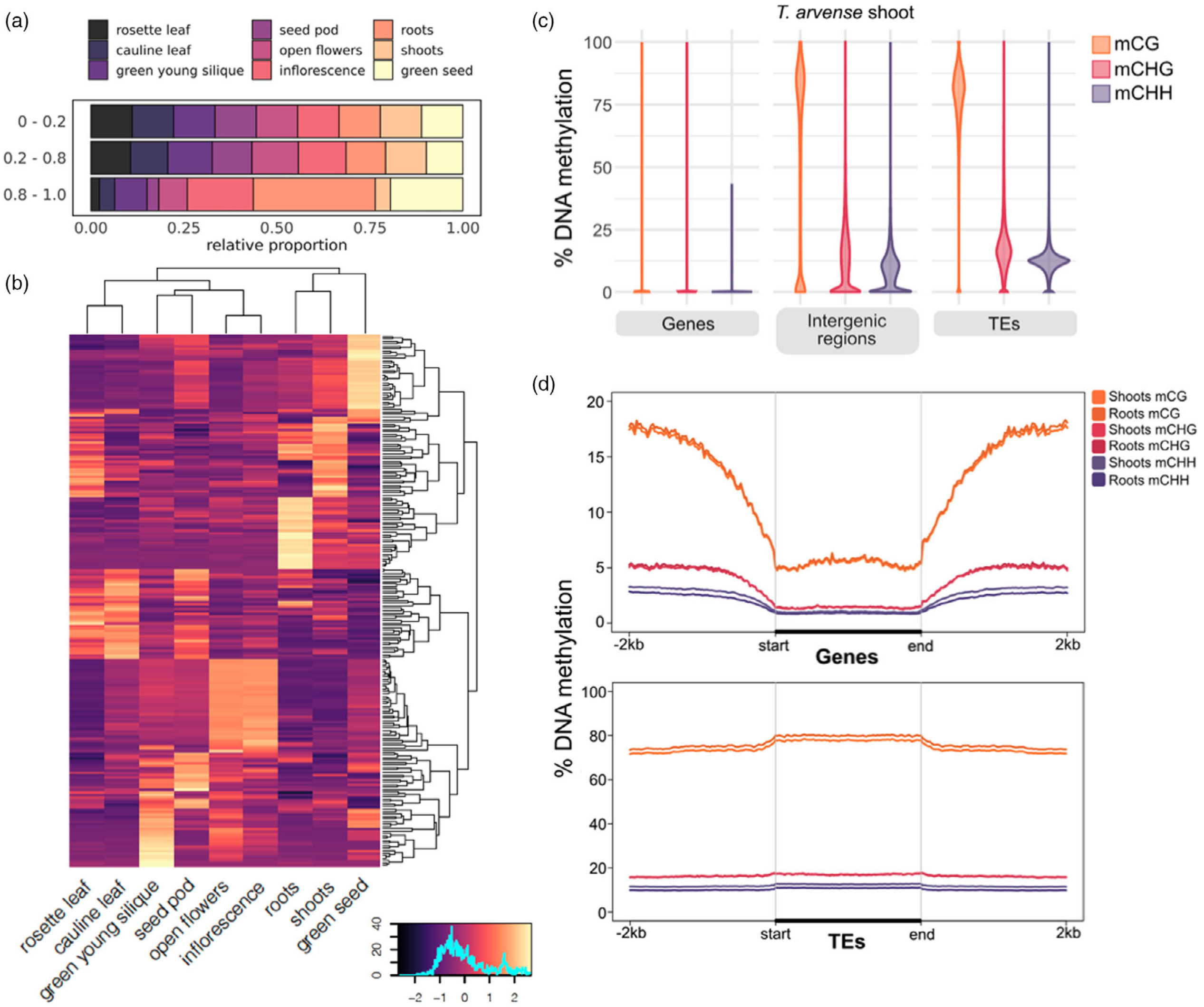


Figure 4 Regulatory dynamics in pennycress. (a) Relative fraction of genes in each tissue for low (0–0.2), intermediate (0.2–0.8) and high/absolute specificity (0.8–1.0) subsets. (b) Log_2(TMM) expression values of the top 30 most highly expressed genes in each tissue, relative to the mean across all tissues, from the subset of genes with a high/absolute tau specificity score. (c) Distribution of average DNA methylation for different genomic features, by cytosine sequence context. (d) DNA methylation along genes (top) and TEs (bottom), including a 2-kb flanking sequence upstream and downstream. DNA methylation was averaged in non-overlapping 25-bp windows.

et al., 2016; Niederhuth *et al*., 2016), DNA methylation dropped only slightly in regions flanking TEs, which might be related to the overall dense TE content in *T. arvense*.

In contrast to TE and promoter methylation, gene body methylation (gbM) is generally associated with medium-to-high gene expression levels (Zhang *et al*., 2006; Zilberman *et al*., 2006). gbM occurs in ~30% of protein-coding genes in *A. thaliana*, with DNA methylation increasing towards the 3′-end of the gene (Zhang *et al*., 2006). The *T. arvense* relative *E. salsugineum* lacks gbM (Bewick *et al*., 2016; Niederhuth *et al*., 2016). gbM was also largely absent in *T. arvense* (Figure 4d), suggesting that gbM was lost at the base of this clade.

Applications towards crop improvement

Genetic variation in a pennycress collection

Knowledge of genetic diversity within wild populations is an essential process for improvement and domestication of new crop species. We analysed a geographically broad sample of forty accessions (Figure S12) using whole-genome resequencing to characterize population structure and variation in germplasm available for breeding. We identified a total of 13 224 528 variants with QD value of ≥2000. Of these, 12 277 823 (92.8%) were SNPs, 426 115 (3.2%) were insertions, and 520 590 (3.9%) were deletions relative to the reference genome. Across all variants, 661 156 (2.9%) were in exons, with 340 132 synonymous, 314 075 nonsynonymous and 6949 non-sense changes. STRUCTURE analysis of both indel and SNP data sets resulted in optimal models of k = 3 populations (Figure S13). Both data sets assigned the three lines of Armenian descent, which were highly distinct and had the largest genetic distance to the other accessions, to a single discrete population with limited to no gene flow to the other populations. These results are consistent with previous reports in pennycress (Frels *et al*., 2019) and were further supported by whole-genome dendrograms (Figure 5a). We also calculated linkage disequilibrium (LD) among 2 518 379 genome-wide markers and chromosome-specific markers using TASSEL v5.2.75 (Bradbury *et al*., 2007) with a sliding window of

40 markers. The *r*-squared values were plotted against the physical distance with a LOESS curve fitted to the data to show LD decay (Figure S14). Genome-wide, LD decayed to an *r*-squared value (r^2) of 0.2 over 6.2 kbp (Hill and Weir, 1988), which is comparable to LD decay reported in related Brassica species at r^2 = 0.3, including *B. rapa* (2.1 kbp) (Wu *et al.*, 2019) and *B. napus* (12.1 kbp) (Lu *et al.*, 2019).

Gene structure variation in pennycress accessions

The natural variation present in germplasm is an important source of alleles to facilitate breeding efforts and presents an opportunity to understand the evolution of gene families and adaptation within a species. To understand these in a more targeted approach, we sequenced on the PacBio Sequel platform the transcriptomes of two accessions, MN108 and Spring32-10, that are amenable to transformation and gene editing (McGinn *et al.*, 2019), using RNA from leaves, roots, seeds, flowers and siliques. We constructed *de novo* reference transcriptomes using the Iso-seq3 pipeline, resulting in 25 296 and 26 571 accession-specific isoforms for MN108 and Spring32-10, respectively. These transcriptomes were then polished using the raw reads and processed through the SQANTI3 pipeline (Tardaguila *et al.*, 2018) to characterize the genes and isoforms identified in each of the accessions. We identified 212 of 220 unique genes and 3780 of 3857 unique isoforms for MN108 and Spring32-10 respectively compared with the new reference. Transcripts mapping to the known reference denoted by ‘Full Splice Match’ (FSM) and ‘Incomplete Splice Match’ (ISM) accounted for 28.7% and 30.6% of all transcript models in MN108 and Spring32-10, respectively (Figure 5b, c). Transcripts of the antisense, intergenic and genic intron categories collectively accounted for a total of 12.0% (MN108) and 11.2% (Spring32-10). About ~15% of all identified transcripts were novel isoforms when compared to the reference transcriptome for T_arvense_v2.

Mapping a pale seedling phenotype

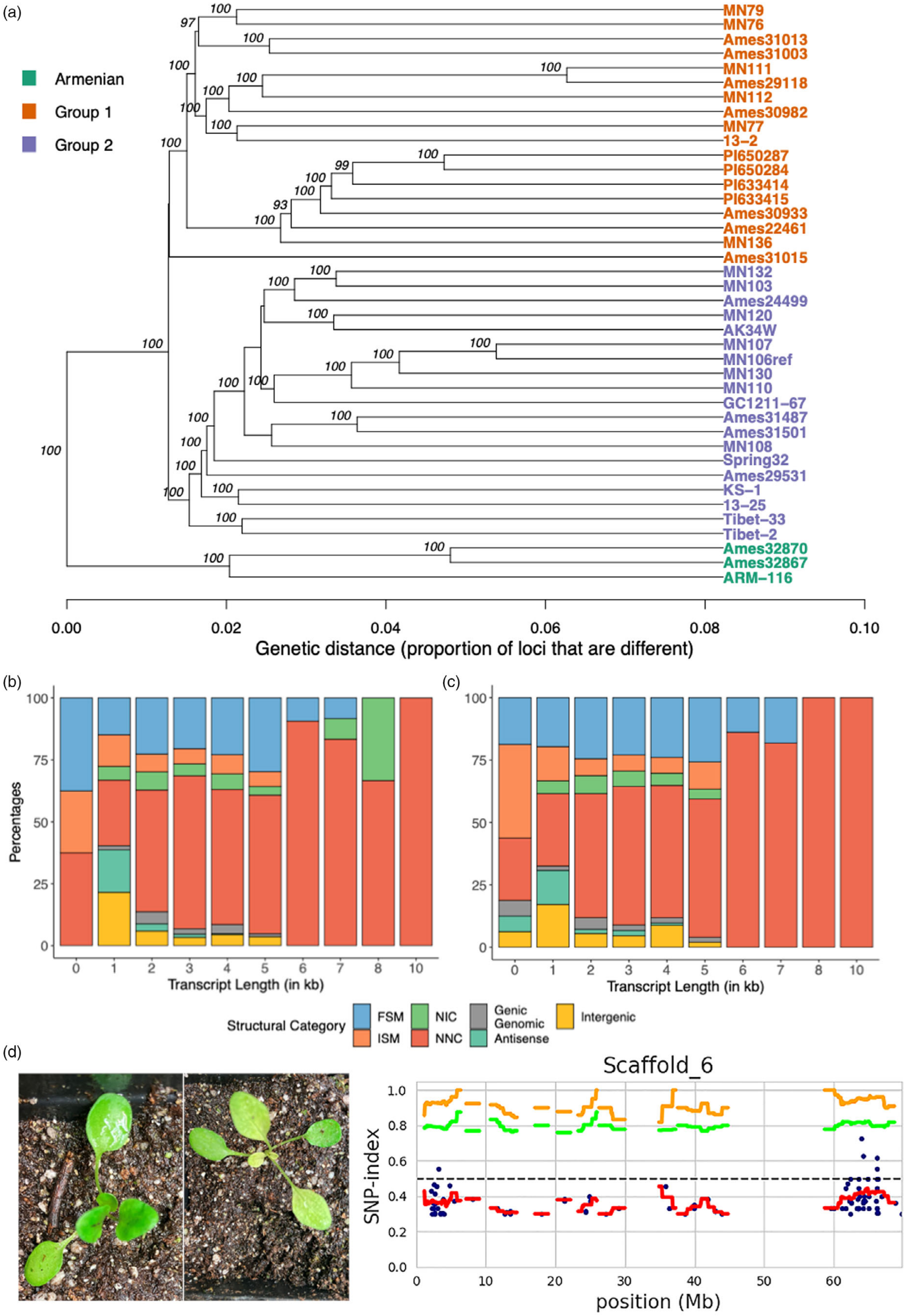
From a segregating population with a high oleic pennycress (*fae-1/rod1-1*) background (Chopra *et al.*, 2020b), we identified pale seedling lines (Figure 5d). This phenotype segregated in a Mendelian fashion. To determine the genetic control for this phenotype, we separately pooled genomic DNA from 20 wild-type and 20 pale plants. We processed sequence data obtained from each of these pools through the MutMap pipeline (Sugihara *et al.*, 2020) and discovered a putative genomic interval (63.85–63.95 Mbp) on scaffold 6 linked to the pale phenotype. SnpEff (Cingolani *et al.*, 2012) identified polymorphisms that might have deleterious effects on function of genes in this region (Table S5). The most obvious candidate is *MEX1*, encoding a maltose transporter located in the chloroplast, knockout of which causes a pale seedling phenotype in *A. thaliana* (Niittylä *et al.*, 2004).

Discussion

In this study, we report a high-quality reference genome assembly and annotation for *T. arvense* (var. MN106-Ref), a newly domesticated oilseed crop for the cooler climates of the world. The improved genome assembly, containing seven chromosome-level scaffolds, revealed two main features: a landscape characterized by a large repetitive fraction populated with TEs and pseudogenic loci in pericentromeric regions, and a gene complement similar in size to other Brassicaceae and densely concentrated towards the telomeres (Figure 1). Previous annotations were enriched with additional gene models for protein-coding loci, and now include non-coding genes for tRNAs, rRNAs, snoRNAs, siRNAs and miRNAs, alongside predicted pseudogenes and TEs (Table 2). These newly improved assembly features will allow for efficient combining of traits and help accelerate future breeding as it would provide knowledge about the gene localization and the linkage of genes of interest. For example, the improved genome assembly has revealed that multiple domestication syndrome genes (*ALKENYL HYDROXALKYL PRODUCING 2-like, TRANSPARENT TESTA 8, EARLY FLOWERING 6*) (Figure S1) are located on a single chromosome.

Improved genomic resources can facilitate general understanding of plant biology and evolutionary biology while aiding plant breeding and crop improvement (Scheben *et al.*, 2016). For example, pennycress and *Arabidopsis* share many key features that made *Arabidopsis* the most widely studied model plant system (Meinke *et al.*, 1998). The use of *Arabidopsis* for translational research and for identifying potential gene targets in *T. arvense* is possible and has been extensively validated (Chopra *et al.*, 2018, 2020a, 2020b; Jarvis *et al.*, 2021; McGinn *et al.*, 2019). Previous studies have suggested that over a thousand unique genes in *T. arvense* are represented by multiple genes in *Arabidopsis* and vice versa. Our comparative genomics by way of synteny with *E. salsugineum* (Yang *et al.*, 2013) revealed a high level of agreement, particularly between the protein-coding fraction of the genome, represented as conserved blocks in the largest seven scaffolds relative to the ancestral karyotype in Brassicaceae (Murat *et al.*, 2015; Figure 2). The detailed description of gene synteny between *T. arvense* and other Brassicaceae provides insights into the evolutionary relevance of *T. arvense* within lineage II of Brassicaceae. In addition, the difference in genome size between *T. arvense* and other species, despite the reduced level of gene duplication and the 1:1 gene relationship, can be explained by the large repetitive fractions present throughout both the centromeric and pericentromeric regions. In the absence of whole-genome duplication events, these repetitive fractions indicate that the increased genome size may be a consequence of active TE expansion. This is therefore suggestive of a mechanism by which deleterious retrotransposon insertions must be mitigated in *T. arvense*. This could be explained by the high proportion of *Gypsy* retrotransposons in this species, usually located in heterochromatic regions, or by integration site selection (Sultana *et al.*, 2017), or otherwise by silencing by small RNA activity and/or DNA methylation (Bucher *et al.*, 2012; Sigman and Slotkin, 2016). Given the relatively high error rate of PacBio CLR reads (~10% before correction) with respect to circular consensus sequencing (CCS), the repetitive fraction would also help to explain the initial overestimation of the assembly size as a result of duplicated contigs. We also detected several loci with highly overrepresented read coverage indicative of repeat collapsing during the assembly process, often intersecting with 5S, 18S and 28S rRNA annotations. Such regions are difficult even for current long read technologies due to the large size of the tandem repeat units.

With the availability of improved genomic resources, increasing interest has turned towards understanding tissue-specific gene regulation to reduce pleiotropic effects upon direct targeting of genes during crop improvement. In this study, we have generated a resource using mRNA-seq, sRNA-seq and WGBS to gain insights into genes and their associated regulatory landscape. These data sets help elucidate the extent of tissue specificity and provide useful


(a)
Armenian
Group 1
Group 2
MN79
MN76
Ames31013
Ames31003
MN111
Ames29118
MN112
Ames30982
MN77
13–2
PI650287
PI650284
PI633414
PI633415
Ames30933
Ames22461
MN136
Ames31015
MN132
MN103
Ames24499
MN120
AK34W
MN107
MN106ref
MN130
MN110
GC1211-67
Ames31487
Ames31501
MN108
Spring32
Ames29531
KS–1
13–25
Tibet–33
Tibet–2
Ames32870
Ames32867
ARM–116
100
97
93
99
0.00
0.02
0.04
0.06
0.08
0.10
Genetic distance (proportion of loci that are different)
(b)
(c)
Percentages
100
75
50
25
0
0 1 2 3 4 5 6 7 8 10
Transcript Length (in kb)
Structural Category
FSM
ISM
NIC
NNC
Genic Genomic
Antisense
Intergenic
(d)
Scaffold_6
SNP-index
1.0
0.8
0.6
0.4
0.2
0.0
0 10 20 30 40 50 60
position (Mb)

information for gene modification targets. For example, fatty acid desaturase 2 gene (FAD2; *Ta12495* – T_arvense_v1) is involved in the oil biosynthesis pathway and is expressed in many different tissues analysed in this study (Data S1). *FAD2* gene knockout should result in higher levels of oleic acid in the seed oil and provide an opportunity for pennycress oil to be used in food applications. It has been observed, however, that knockout mutants in pennycress display delayed growth and reduced seed yields in spring types (Jarvis *et al*., 2021), and reduced winter survival in the winter types (Chopra *et al*., 2019), as a purported consequence of its broad expression profile. Similarly, genes such as *AOP2-LIKE* (*Tarvense_05380* – T_arvense_v2) have been targeted to reduce glucosinolates in pennycress seed meal for food and animal feed applications (Chopra *et al*., 2020b). However, *AOP2-LIKE*, too, is expressed in many tissues during development, which might explain why knockout plants with reduced glucosinolate content are reportedly more susceptible to insect herbivores such as flea beetles feeding on rosette leaves and root tissues (Marks *et al*., 2021). Our tissue-specific expression data suggest that, to overcome this challenge, one could alternatively target genes such as *Glucosinolate Transporter 1* (GTR1; *Tarvense_14683*), which is expressed specifically in reproductive tissues (Data S1). This might achieve the desired reductions of seed glucosinolates while avoiding developmental defects. Such approaches have been effectively used in *Arabidopsis* and many *Brassica* species (Andersen and Halkier, 2014; Nour-Eldin *et al*., 2012).

Finally, the forty resequenced accessions described here provide a rich source of variants that reflect the genetic diversity and population structure of the species in the collection (Figure 5a). Further evaluations of transcriptome sequences showed ample variation in the transcripts from two separate lines – MN108 and Spring32-10 – that are highly amenable to transformation and highlighted the potential for developing pangenomes in the future. These genomic resources will facilitate genetic mapping studies in pennycress in both natural populations and mutant panels. We have identified genomic regions associated with a pale leaf mutant in pennycress seedlings using a modified BSA-Seq approach in this study (Figure 5d).

Over the last few years, significant efforts have been made towards the discovery of crucial traits and translational research in pennycress, centring on MN106-Ref and the gene space information generated by Dorn *et al*. (2013) and Dorn *et al*. (2015). In this study, we continued to generate genomic tools for this accession, with improved contiguity and high-quality annotations to make *T. arvense* var. MN106-Ref more accessible as a field-based model species for genetics and epigenetics studies and to provide tools for this new and extremely hardy winter annual cash cover crop. However, the assembly of additional accessions can only help to further enrich the resources available for the study of pennycress. In parallel to this study, a Chinese accession of *T. arvense* (YUN_Tarv_1.0) was assembled using Oxford Nanopore, Illumina HiSeq and Hi-C sequencing (Geng *et al*., 2021). This timely availability of an additional frame of reference opens the door to a pan-genomic approach in evolutionary research and allows for the better characterization of structural variants moving forward. Furthermore, the use of different sequencing technologies and assembly software provides an additional avenue to correct misassemblies and base calling errors in either case. The overall longer contigs assembled with PacBio CLR, for example, and the consideration of various genetic map data in addition to Hi-C provides a greater resolution of scaffolds particularly throughout the centromere and pericentromeric regions (Figure S15). The reduced error rate of PacBio CCS (used for polishing) is also reflected in the overall k-mer content, which is measured with a two-order magnitude higher consensus quality over scaffolds representing chromosomes and ~99% overall completeness for T_arvense_v2 (Tables S6-S8), indicative of high-quality, error-free sequences more appropriate for variant calling, for instance. Geng *et al*. (2021) also reported WGS analysis on forty Chinese accessions and reported an LD decay of 150 kbp at an *r*-squared value (r^2) of 0.6, which is considerably higher than the values determined on the forty accessions in this study, as well as those reported for related *Brassica* species (Lu *et al*., 2019; Wu *et al*., 2019). We believe the combination of resources will allow us to investigate the differences that might exist between accessions originating from different geographic locations around the world and help provide further insight into structural variations and evolutionary dynamics.

In conclusion, the T_arvense_v2 assembly offers new insights into the genome structure of this species and of lineage II of Brassicaceae more generally, and it provides new information and resources relevant for comparative genomic studies. The tools presented here provide a solid foundation for future studies in an alternative model species and an emerging crop.

Methods

Seeds for the reference genome development

Seeds from a small natural population of *T. arvense* L. were collected near Coates, MN by Dr. Wyse, and the accession number MN106 was assigned to this population. We propagated a single plant for ten generations from this population, and we refer to this line as MN106-Ref.

Sample collection, library preparation and DNA sequencing for assembly

PacBio CLR library

Plants were cultivated, sampled and prepared at the Max Planck Institute for Developmental Biology (Tübingen, Germany). Plant seeds were stratified in the dark at 4 °C for 4–6 day prior to planting on soil. Samples were collected from young rosette leaves of *T. arvense* var. MN106-Ref seedlings, cultivated for 2 weeks under growth chamber conditions of 16–23 °C, 65% relative humidity and a light/dark photoperiod of 16 h:8 h under 110–140 μmol/m^2/s light. High molecular weight (HMW) DNA was obtained following nucleus isolation and DNA extraction with the Circulomics Nanobind Plant Nuclei Big DNA Kit according to the protocol described in Workman *et al*. (2018) and (Workman *et al*., 2019). A total of 11 extractions from 1.5–2 g frozen leaves each were processed in that way, yielding a pooled sample with a total of 12 μg of DNA by Qubit® 2.0 fluorometer

Figure 5 (a) Dendrogram representing the forty wild accessions in our study showing three distinct subpopulations, inferred from STRUCTURE analysis (Figure S13). (b,c) Variation of transcript isoforms for MN108 (b) and Spring32-10 (c) accessions based on SQANTI3 analysis. (d) A pale phenotype segregating in an improved pennycress line (*fae-1-1/rod1-1*) was analysed with a modified bulked-segregant analysis, and the QTL region associated with this phenotype was mapped using the MutMap approach.

(Thermo Fisher Scientific, Waltham, MA) estimation, and high DNA purity with a mean absorbance ratio of 1.81 at 260/280 nm absorbance and 2.00 at 260/230 nm absorbance, as measured by NanoDrop 2000/2000c spectrophotometer (Thermo Fisher Scientific, Waltham, MA). HMW DNA was sheared by one pass through a 26G needle using a 1-mL syringe, resulting in an 85-kb peak size sample as estimated by FEMTO Pulse Analyzer (Agilent Technologies, Santa Clara, CA). A large insert gDNA library for PacBio Sequel II CLR sequencing was prepared using the SMRTbell® Express Template Preparation Kit 2.0. The library was size-selected for >30 kb using BluePippin with a 0.75% agarose cassette (Sage Science) and loaded into one Sequel II SMRT cell at a 32 pM concentration. This yielded a genome-wide sequencing depth of approximately 476X over ~6.9 million polymerase reads with a subread N50 of ~38 kbp.

PacBio CCS library

MN106-Ref plants were grown in growth chambers at the University of Minnesota. Individual plants were grown to form large rosettes for isolating DNA. Approximately 25 g of tissue was harvested and submitted to Intact Genomics (Saint Louis, MO) for high molecular weight DNA extraction. This yielded a pooled sample with a total of 269 ng of DNA by Qubit® (Thermo Fisher Scientific, Waltham, MA) estimation, and high DNA purity with a mean absorbance ratio of 1.87 at 260/280 nm and 2.37 at 260/230 nm, as measured by Nanodrop spectrophotometer (Thermo Fisher Scientific, Waltham, MA). To further clean up the high molecular weight DNA, we used Salt:Chloroform Wash protocol recommended by PacBio. This yielded a total of 12.1 ng/uL of high-quality DNA for library preparation. A large insert gDNA library was prepared, and 15 kb High Pass Size Selection on Pippin HT was performed at the University of Minnesota Genomics Center (Minneapolis, MN). These libraries were sequenced on 4 SMRT cells using PacBio Sequel II (Pacific Biosciences, Menlo Park).

Bionano library

High molecular weight DNA was isolated from young leaves and nicking endonuclease – BspQI was chosen to label high-quality HMW DNA molecules. The nicked DNA molecules were then stained as previously described (Lam *et al*., 2012). The stained and labelled DNA samples were loaded onto the NanoChannel array (Bionano Genomics, San Diego, CA) and automatically imaged by the Irys system (Bionano Genomics, San Diego, CA).

Hi-C library

The MN106-Ref plant tissue used for PacBio CCS was submitted to Phase Genomics (San Diego, CA). The Hi-C library was prepared following the proximo Hi-C plant protocol (Phase Genomics, San Diego, CA), and the libraries were sequenced to 116X depth on an Illumina platform with the paired-end mode and read length of 150 bp.

Illumina PCR-free library

Libraries for PCR-free short read sequencing were prepared from MN106-Ref genomic DNA using the TruSeq DNA PCR-Free Low Throughput Library Prep Kit (Illumina, San Diego, CA) in combination with TruSeq DNA Single Indexes Set A (Illumina, San Diego, CA) according to the manufacturer's protocol. We prepared two libraries, with average insert sizes of 350 bp and 550 bp, respectively. Samples were sequenced to 125X depth (~66 Gb) on an Illumina HiSeq 2500 (Illumina, San Diego, CA) instrument with 125-bp paired-end reads.

Genome assembly and construction of chromosome-level scaffolds

The initial assembly was performed using Canu v1.9 (Koren *et al*., 2017) with default options, aside from cluster runtime configuration and the settings `corOutCoverage=50`, `minReadLength=5000`, `minOverlapLength=4000`, `correctedErrorRate=0.04` and `genomeSize=539m`, which were selected based on the characteristics of the library. Canu performs consensus-based read correction and trimming, resulting in a curated set of reads that were taken forward for assembly (Figure S16).

The resulting assembly overestimated the genome size by approximately 53% (Table S2), which we surmised was likely due to uncorrected sequencing errors in the remaining fraction of reads, in which Canu was able to assemble into independent, duplicated contigs. Analysis of single-copy orthologs from the *Eudicotyledons odb10* database with BUSCO v3.0.2 (Simão *et al*., 2015) revealed a high completeness of 98.4% and a duplication level of 23.6% (Table S6). Subsequent alignment of the reads to the assembly using minimap2 v2.17 (Li, 2018) and purge_dups v1.0.1 (Guan *et al*., 2020) presented bimodal peaks in the read depth distribution, indicative of a large duplicated fraction within the assembly (Figure S17). As efforts to collapse this duplicated fraction using assembly parameters were unsuccessful, and purge_dups is intended to correct duplication arising from heterozygosity (which does not apply in *T. arvense*), the fraction was reduced by manual curation instead. Contigs starting from the left-hand side of the read depth distribution were consecutively removed until reaching an approximation of the estimated genome size, with any contigs containing non-duplicated predicted BUSCO genes kept preferentially in favour of discarding the next contig with lower read depth in the series.

The deduplicated assembly from Canu was polished with the PacBio Sequel II HiFi CCS reads using two iterations of RACON v1.4.3 (Vaser *et al*., 2017), prior to repeat reassembly. Bionano maps were used to build *de novo* scaffolds using the polished assembly; hybrid scaffolds were generated using the *de novo* Bionano maps and the assembly (https://bionanogenomics.com/support-page/data-analysis-documentation/). To further resolve repetitive regions and improve assembly contiguity, the bionano-scaffolded assembly was integrated into the HERA pipeline (Du and Liang, 2019). The Hi-C data were aligned with bwa-mem v0.7.17 (Li and Durbin, 2009), PCR duplicates were marked with picard tools v1.83 (http://broadinstitute.github.io/picard), and the quality was assessed with the `hic_qc.py` tool of Phase Genomics (https://github.com/phasegenomics/hic_qc). The assembly was then scaffolded with the Hi-C alignments using SALSA v2.2 (Ghurye *et al*., 2017) and subsequently polished with the PCR-free Illumina data using two iterations of PILON v1.23 (Walker *et al*., 2014). The final assembly was the result of a meta-assembly with quickmerge v0.3 (Chakraborty *et al*., 2016), which combined the current assembly with an earlier draft version assembled using Canu 1.8 (Koren *et al*., 2017) directly from the PacBio CCS reads and polished only with the Illumina PCR-free short-reads, following an almost identical workflow, in order to help address the possibility of misassembly arising from technical sources and improve overall contiguity. This resulting assembly was evaluated with BUSCO (Simão *et al*., 2015) and QUAST v5.0.2 (Gurevich *et al.*, 2013). Intermediate assembly statistics are given in comparison with (i) immediately after Canu, and (ii) the final version after rescaffolding (Table S2).

Genome size estimation using flow cytometry and k-mer-based approach

The nuclei of field pennycress line MN106-Ref, *Arabidopsis thaliana*, maize and tomato were stained with propidium iodide, and fluorescent signals were captured using a Becton-Dickinson FACSCanto flow cytometer (https://www.bdbiosciences.com/). DNA content for all four species that corresponded to $G_{0/1}$ nuclei is listed in Table S1. The genome size of Arabidopsis is 135 Mb, and therefore, the genome size of pennycress was calculated to be 501 ± 33 Mb. Using the Illumina HiSeq2500 platform, we obtained ~100× PCR-free reads, which were used for subsequent K-mer analysis using Jellyfish (Marçais and Kingsford, 2011). The 101-mer frequency distribution curve exhibited a peak at 22 k-mer, and analysis showed that the total number of K-mers was 11 403 836 319. Using the formula of genome size = total K-mer number/peak depth, the genome size of this sequencing sample was estimated to be 518 356 196 bp. Similarly, the single-copy content of the genome was estimated to reach 79%. Using both methods of genome size estimation, we found the pennycress genome ranged from 459 to 540 Mb.

Development of genetic maps for rescaffolding

To improve the contiguity and correct misassemblies, we developed two genetic linkage maps using F_2 populations. The first linkage map was derived from a cross between a wild Minnesota accession 'MN106-Ref' and a genetically distant Armenian accession 'Ames32867'. The resulting F_1 plants were allowed to self-fertilize, and seeds from a single plant were collected and propagated to the F_2 generation. Approximately 500 mg fresh tissue was collected from 94 individuals in the F_2 population. The tissue was desiccated using silica beads and pulverized using a TissueLyser. DNA was isolated with the BioSprint DNA Plant Kit (Qiagen, Valencia, CA). The F_2 population along with the two parental genotypes was genotyped with genotyping by sequencing at the University of Minnesota Genomics Center (Minneapolis, MN). Each sample was digested with the *BtgI_BgLII* restriction enzyme combination, barcoded and sequenced on the Illumina NovaSeq S1 (single-end 101 bp) yielding 1 237 890 mean reads per sample. The raw reads were demultiplexed based on the barcode information and aligned to the most recent iteration of the pennycress genome using bwa. Sequence-aligned files were processed through samtools v1.9 (Li *et al.*, 2009) and picard tools to sort the files and remove group identifiers. Variants were called using GATK HaplotypeCaller v3.3.0. SNPs identified among these 94 lines were used for the development of genetic maps. The second linkage map was derived from a cross between MN106-Ref and a mutant line '2019-M2-111'. To identify the variant alleles in 2019-M2-111, we performed whole-genome resequencing using paired-end reads on the Illumina Platform. SNPs were identified using a similar approach as described above. Sixty-seven SNP markers were designed using the biallelic information from resequence data. DNA was extracted from 48 samples from the mutant F_2 population using the Sigma-Aldrich ready extract method, allele-specific and flanking primers synthesized from IDT (Iowa, USA) for each of the alleles were mixed (Data S3), and genotyping was performed using the methods described in Chopra *et al.* (2020a).

A total of 35 436 SNPs were identified among the population used for the first linkage map, SNP sites were selected with no-missing data, QD > 1000, and the segregation of the markers was 1:2:1. A total of 743 high-quality SNPs were retained for further analysis. A genetic map for the population was constructed using JoinMap 5 (Stam, 1993). Only biallelic SNPs were used in the analysis, and genetic maps were constructed with regression mapping based on default parameters of recombination frequency of <0.4 with only the first two steps. The Kosambi mapping function was chosen for map distance estimation, and the Ripple function was deployed to confirm marker order within each of the seven linkage groups. A total of 319 markers were mapped to seven linkage groups (Data S4). Similarly, 67 markers were genotyped on 48 individuals from the second population of linkage and 52 markers were mapped to six linkage groups (Data S5). Both of these linkage maps were used for reordering and correcting the scaffolds as described below.

Rescaffolding

Initial exploration regarding gene and TE distributions and methylation patterns pointed to potential misassemblies in the assembled genome. Further investigation by way of synteny comparison with a closely related species, *Eutrema salsugineum* (Yang *et al.*, 2013), revealed that several of these likely occurred during scaffolding as orientation errors. Some of these errors could also be supported in comparison with the recent assembly of a Chinese accession (YUN_Tarv1.0) of *T. arvense*. Consequently, we manually introduced breakpoints at selected loci in the assembled genome where they were supported by at least two sources of data from whole-genome alignments to YUN_Tarv1.0, synteny maps to *E. salsugineum* (derived from reciprocal best blast), genetic linkage maps (wild-derived and EMS mutation based) and Hi-C contact maps. These were cross-examined with minimap2 alignments of PacBio CLR reads to the genome, an overview of corresponding gene distributions produced by Liftoff v1.5.2 (Shumate and Salzberg, 2020) and the resulting synteny analysis to *E. salsugineum*. The resulting contigs were then rescaffolded with ALLMAPS v1.1.5 (Tang *et al.*, 2015) to produce the final assembly, integrating both the synteny map and genetic map data and manually discounting contigs that were supported only by single markers. The final assembly statistics in comparison with previous intermediary stages are given in Table S2.

Comparative genomics

Genome sequences

Arabidopsis thaliana (Araport 11), *Schrenkiella parvula* (v2.2) and *Arabidopsis lyrata* (v2.1) genome sequences and gene annotation were downloaded from Phytozome (Goodstein *et al.*, 2012). The *Eutrema salsugineum* gene annotation was obtained from Phytozome and lifted over the assembly GenBank GCA_000325905.2.

Genome alignments and synteny analysis

The genome alignments between the different versions of the *T. arvense* assembly to *E. salsugineum* were done using MUMmer v4.0.0 (Marçais *et al.*, 2018) with a minimal length of 200 nt and followed by filtering for 1:1 matches and removing alignments smaller than 1000 bp. To identify the interspecies gene orthologs and syntenic relationships between *T. arvense* and other species, we used MCScan in the JCVI utility library (https://github.com/tanghaibao/jcvi; Tang *et al*., 2008). The ortholog relationships were obtained using the proteinic translation of the CDS and using the argument `--cscore=0.99`. To define the syntenic blocks and the corresponding genomic coordinates, we used the parameters `--minspan=15` and `--minsize=5`. The genomic coordinates from the syntenic blocks were parsed to draw the syntenic relationships using Circos v0.69-8 (Krzywinski *et al*., 2009).

To determine the different ancestral Brassicaceae chromosomal blocks (ABKs), we took the ortholog relationship between each gene in *T. arvense* and *A. thaliana* from the synteny analysis, and compared it with a gene list derived from Murat *et al*. (2015) where each ortholog gene of *A. thaliana* had an assigned ABK block (Murat *et al*., 2015).

Genome annotation

Tissue preparation for RNA sequencing

Thlaspi arvense var. MN106-Ref seeds were surface-sterilized with chlorine gas for 1 h and stratified for 3 day at 4 °C. For seedling-stage RNA extractions, seeds were plated on ½ MS medium supplemented with 1% plant agar and stratified for 3 day at 4 °C. For all other tissue collections, plants were sown on soil and grown in a climate-controlled growth chamber in long-day conditions (16/8-h light/dark at 21°/16 °C, light intensity 140 μE/m^2*s, with 60% relative humidity); plants were watered twice per week. Two weeks after germination, plants growing on soil were vernalized at 4 °C in the dark for 4 weeks, then moved back to the growth chamber. Samples were collected from 11 different tissues in three biological replicates (two in case of mature seeds); for each replicate, we pooled tissue from two individuals. Tissues included the following: one-week-old shoots (from plate culture), one-week-old roots (from plate culture), rosette leaves, cauline leaves, inflorescences, open flowers, young green siliques (about 0.5 × 0.5 cm), older green siliques (about 1 × 1 cm), seed pods, green seeds and mature seeds.

RNA extraction and sequencing

Total mRNA was extracted using the RNeasy Plant Kit (Qiagen, Valencia, CA) and treated with DNase I using the DNA-free Kit DNase Treatment and Removal Reagents (Ambion by Life Technologies, Carlsbad, CA), following the manufacturer's protocols. cDNA libraries were constructed using the NEBNext Ultra II Directional RNA Library Prep Kit (New England BioLabs, Ipswich, MA, USA Inc.) for Illumina following the manufacturer's protocol. Libraries were sequenced on a HiSeq 2500 instrument (Illumina, San Diego, CA) as 125-bp paired-end reads.

Transcriptome assembly

Following quality control and adapter clipping with cutadapt (Martin, 2011), biological replicates for each of eleven tissue types from Illumina mRNA-seq libraries were aligned independently using STAR v2.5.3a (Dobin *et al*., 2013), then merged according to tissue type, prior to assembly by a reference-based approach. Each assembly was performed using Ryuto v1.3m (Gatter and Stadler, 2019), and consensus reconstruction was then performed using TACO v0.7.3 (Niknafs *et al*., 2017) to merge tissue-specific transcriptome assemblies. PacBio Iso-seq libraries from MN106-Ref were refined, clustered and polished following the Iso-seq3 pipeline (https://github.com/PacificBiosciences/IsoSeq), prior to alignment with STARlong and isoform collapsing using the cDNA_Cupcake (https://github.com/Magdoll/cDNA_Cupcake) suite. The Iso-seq data were later leveraged together with the Illumina mRNA-seq data to prioritize convergent isoforms using custom in-house scripting.

Genome annotation

The final assembly was annotated using the MAKER-P v2.31.10 (Campbell *et al*., 2014, 2014) pipeline on the servers provided by the EpiDiverse project, at ecSeq Bioinformatics GmbH (Leipzig, Germany). Plant proteins were obtained from the *Viridiplantae* fraction of UniProtKB/Swiss-Prot and combined with RefSeq sequences derived from selected Brassicaceae: *Arabidopsis thaliana*, *Brassica napus*, *Brassica rapa*, *Camelina sativa* and *Raphanus sativus*. TEs were obtained from RepetDB (Amselem *et al*., 2019) for selected plant species: *Arabidopsis lyrata*, *Arabidopsis thaliana*, *Arabis alpina*, *Brassica rapa*, *Capsella rubella* and *Schrenkiella parvula (Eutrema parvulum)*. Repeat library construction was carried out using RepeatModeler v1.0.11 (Smit and Hubley, 2008) following basic recommendations from MAKER-P (Campbell *et al*., 2014). Putative gene fragments were filtered out following BLAST search to the combined Swiss-Prot + RefSeq protein plant database after exclusion of hits from RepetDB. The *de novo* library was combined with a manually curated library of plant sequences derived from repbase (Bao *et al*., 2015). Genome masking is performed with RepeatMasker v4.0.9 (Smit, 2004) as part of the MAKER-P pipeline. Protein-coding genes, non-coding RNAs and pseudogenes were annotated with the MAKER-P pipeline following two iterative rounds under default settings, using (i) transcript isoforms from Illumina mRNA-seq and PacBio Iso-seq data, (ii) protein homology evidence from the custom Swiss-Prot + RefSeq plant protein database and (iii) the repeat library and TE sequences for masking. The initial results were used to train gene models for *ab initio* predictors SNAP v2006-07-28 (Korf, 2004) and Augustus v3.3.3 (Stanke *et al*., 2006), which were fed back into the pipeline for the subsequent rounds. The final set of annotations was filtered based on Annotation Edit Distance (AED) < 1 except in cases with corresponding PFAM domains, as derived from InterProScan v5.45-80.0 (Jones *et al*., 2014). The tRNA annotation was performed with tRNAscan-SE v1.3.1 (Lowe and Eddy, 1997) and the rRNA annotation with RNAmmer v1.2 (Lagesen *et al*., 2007). The snoRNA homologs were derived using Infernal v1.1.4 (Nawrocki and Eddy, 2013) from plant snoRNA families described in Patra Bhattacharya *et al*. (2016). A small phylogeny based on gene orthologs and duplication events in comparison with selected Brassicaceae (*A. lyrata*, *A. thaliana, B. rapa*, *S. parvula* and *E. salsugineum*) was performed with OrthoFinder v2.5.2, and the resulting species tree is rooted using STRIDE (Emms and Kelly, 2017) and inferred from all genes using STAG (Emms and Kelly, 2018).

Transposable element annotation

Two *de novo* annotation tools, EDTA v1.7.0 (Ou *et al*., 2019) and RepeatModeler v2.0 (Flynn *et al*., 2020), were used to annotate TEs independently. For EDTA, the following parameters were used in addition to defaults: `--species others`, `--step all`, `--sensitive 1`, `--anno 1`, and `--evaluate 1`. For RepeatModeler2, the additional parameters were `-engine ncbi` and `-LTRStruct`. The outputs of both tools were evaluated by manual curation. First, we used tblastn to align each TE consensus with the transposase database obtained from repbase, and the retrotransposon domains (GAG, Pol, Env, etc.) were viewed one by one with dotter (Sonnhammer and Durbin, 1995). Sequences with multiple paralogs were mapped back to the genome and manually extended to determine the full-length boundary of each TE. A total of 107 full-length, representative *Copia* and *Gypsy* families were successfully evaluated. The TE consensus from RepeatModeler2 was selected as the most accurate model based on full-length paralogs. RepeatMasker was then used to construct the GFF3-like file from the FASTA file from RepeatModeler2, with the optional settings: `-e ncbi -q -no_is -norna -nolow -div 40 -cutoff 225`. The perl script `rmOutToGFF3.pl` was used to generate the final GFF3 file.

sRNA plant material

Seeds were sterilized by overnight incubation at −80 °C, followed by 4 h of bleach treatment at room temperature (seeds in open 2 mL tube in a desiccator containing a beaker with 40 mL chlorine-based bleach (<5%; DanKlorix, Colgate-Palmolive, New York, NY) and 1 mL HCl (32%; Carl Roth, Karlsruhe, Germany)). For rosette, inflorescence and pollen, seeds were stratified in the dark at 4 °C for six days prior to planting on soil, then cultivated under growth chamber conditions of 16–23 °C, 65% relative humidity and a light/dark photoperiod of 16 h:8 h under 110–140 μmol/m^2/s light. Rosette leaves were harvested after two weeks of growth. For inflorescence and pollen, 6-week-old plants were vernalized for 4 weeks at 4 °C in a light/dark photoperiod of 12 h:12 h under 110–140 μmol/m^2/s light. Two weeks after bolting, inflorescence and pollen were collected. Pollen grains were collected by vortexing open flowers in 18% sucrose for 5 min followed by centrifugation at 3000***g*** for 3 min in a swinging bucket rotor. For root samples, seeds were stratified for 6 days at 4 °C in the dark on ½ MS media. Plants were grown in 3–4 mL ½ MS medium plates in long day (16 h) at 16 °C. Root samples were collected 12–14 days after stratification.

sRNA extraction and library preparation

Total RNA was extracted by freezing collected samples with liquid nitrogen and grinding with a mortar and pestle with TRIzol reagent (Life Technologies, Carlsbad, CA). Then, total RNA (1 μg) was treated with DNase I (Thermo Fisher Scientific, Waltham, MA) and used for library preparation. Small RNA libraries were prepared as indicated by the TruSeq Small RNA Library Prep Kit (Illumina, San Diego, CA), using 1 μg of total RNA as input, as described by the TruSeq RNA sample prep V2 guide (Illumina, San Diego, CA). Size selection was performed using the BluePippin System (SAGE Science, Massachusetts). Single-end sequencing was performed on a HiSeq 3000 instrument (Illumina, San Diego, CA).

sRNA locus annotation

Raw FASTQ files were processed to remove the 3′-adapter and quality-controlled with trim_galore v0.6.6 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) using `trim_galore -q 30 --small_rna`. Read quality was checked with FastQC v0.11.9 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The reference annotation of sRNA loci was created following the steps indicated by Lunardon *et al*. (2020). In short, each library was aligned to the reference genome independently using ShortStack v3.8.5 (Axtell, 2013b), with default parameters, to identify clusters of sRNAs *de novo* with a minimum expression threshold of 2 reads per million (RPM). sRNA clusters from all libraries of the same tissue were intersected using BEDTools v2.26.0 `multiIntersectBed` (Quinlan and Hall, 2010) with default parameters, and only those loci present in at least three libraries were retained. For each tissue, sRNA clusters 25 nt apart were padded together with the `bedtools merge -d` option. sRNA loci whose expression was <0.5 RPM in all libraries of each tissue were also removed. Finally, sRNA loci for all different tissues were merged in a single file retaining tissue of origin information with `bedtools merge -o distinct` options. miRNAs predicted by the ShortStack tool were manually curated (Appendix S1) following the criteria of Axtell (2013b): maximum hairpin length of 300 nt; ≥75% of reads mapping to the hairpin must belong to the miRNA/miRNA* duplex; for the miRNA/miRNA* duplex, no internal loops allowed, two-nucleotide 3′ overhangs, maximum five mismatched bases and only three of which are nucleotides in asymmetric bulges; and mature miRNA sequence should be between 20 and 24 nt.

Expression atlas

Gene expression was measured from the same tissue-specific STAR alignments taken prior to merging biological replicates for transcript assembly, excluding coverage outliers 'mature seed' and 'green old silique'. A total of 27 samples from 9 tissues were therefore considered for gene expression analysis. Raw counts were generated using subread featureCounts v2.0.1 (Liao *et al*., 2014) and subsequently normalized using the trimmed mean of M-values (TMM) (Robinson and Oshlack, 2010) derived from edgeR v3.34 (Robinson *et al*., 2010). Averaged expression counts by group were taken for tissue specificity evaluation using the Tau (τ) algorithm (Yanai *et al*., 2005), as implemented in the R package tispec v0.99.0 (https://rdrr.io/github/roonysgalbi/tispec/), which provides a measure of τ in the range of 0 - 1, where 0 is non/low specificity, and 1 indicates high/absolute specificity.

DNA methylation

We extracted genomic DNA from roots and shoots of 2-week-old seedlings grown on ½ MS medium with 0.8% agar and 0.1% DMSO. Seedlings were grown vertically in 16-h/8-h light/dark cycle; at the time of sampling, roots were separated from shoot tissue with a razor blade and the plant tissue was flash-frozen in liquid nitrogen. Genomic DNA was extracted from ground tissue using the DNeasy Plant Mini Kit (Qiagen, Hilden, Germany). Libraries for WGBS were prepared using the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs). Adapter-ligated DNA was treated with sodium bisulphite using the EpiTect Plus Bisulfite Kit (Qiagen, Hilden, Germany) and amplified using the Kapa HiFi Uracil + ReadyMix (Roche, Basel, Switzerland) in 10 PCR cycles. WGBS libraries were sequenced on an Illumina HiSeq2500 instrument with 125-bp paired-end reads.

The WGBS libraries were processed using the nf-core/methylseq v1.5 pipeline (10.5281/zenodo.2555454) combining bwa-meth v0.2.2 (Pedersen *et al*., 2014) as an aligner and MethylDackel v0.5.0 (https://github.com/dpryan79/MethylDackel) for the methylation calling. The default parameters were used for the entire workflow with the exception of the methylation calling where the following arguments were used: `-D 1000 --maxVariantFrac 0.4 --minOppositeDepth 5 --CHG --CHH --nOT 3,3,3,3 --nOB 3,3,3,3 -d 3`. Only cytosines with a minimum coverage of 3x were kept for the subsequent analysis.

Further comparisons between the methylated cytosines and the genome annotation were performed using BEDtools v2.27.1 (Quinlan and Hall, 2010).

Population genomics

DNA from forty pennycress accessions was extracted from approximately 500 mg of leaf tissue pooled from five plants using a plant genomic DNA kit (Epoch Life Science). DNA was then subjected to whole-genome sequencing on an Illumina Novaseq sequencer (2 × 125 bp). Raw reads were then aligned to the new reference genome (T_arvense_v2) using bwa-mem (Li and Durbin, 2009). The aligned files were processed with Samtools and Picard tools, and variants were called using GATK HaplotypeCaller v3.3.0 (Ren *et al*., 2018). Variants were annotated using SnpEff 5.0e (Cingolani *et al*., 2012). Data sets for both Indel and SNP panels were trimmed based on LD prior to population genomic analysis using Plink v1.9 (Purcell *et al*., 2007) with the parameter `--indep-pairwise 1000 5 0.5`. Population structure for both SNP and indel data was then characterized using the admixture model and independent allele frequencies in STRUCTURE v2.3.4 (Pritchard *et al*., 2000). Dendrograms of both SNP and Indel data were generated under the UPGMA method using the R package poppr (Kamvar *et al*., 2014).

The forty accessions were planted in a three replication, randomized complete block design in a greenhouse maintained at 21/20 °C and 16 hour days. Ten seeds per replicate were planted in 13.3-cm^2 pots in Sungrow propagation potting mix. Seedlings were thinned to one plant per pot after emergence. Winter annual accessions require vernalization to induce flowering, so all winter accessions were placed in a growth chamber maintained at 4 °C with 16-h light for a period of 21 days about 4 weeks after emergence. Spring annual accessions were planted approximately five weeks after winter accessions. Data for days to flowering were collected on 34 accessions that germinated as the number of days that elapsed from the date of emergence to the appearance of the first flower. The vernalization requirement for winter accessions explains the large differences in mean number of days to flowering between spring and winter accessions. Additional phenotypes and data associated with these sequenced accessions are available in Data S6.

Structural variants using Iso-seq data

Single-molecule real-time (SMRT) isoform sequencing (Iso-seq) based on PacBio (Pacific Biosciences, Menlo Park, CA) generated reads was used to investigate unambiguous full-length isoforms for two pennycress wild accessions, MN108 and Spring32-10. Total RNA extraction was performed on the green seed, hypocotyl, seedling root and flower tissues from pennycress plants grown in a climate-controlled growth chamber maintained 21/20 °C during 16-h:8-h day–night setting. Approximately 250 ng of total RNA was obtained and subjected to the Iso-seq Express Library Workflow (Pacific Biosciences, Menlo Park, CA). cDNA is synthesized from full-length mRNA with the NEBNext Single Cell/ Low Input RNA Prep Kit (New England Biolabs, Ipswich, MA) followed by PCR amplification. The amplified cDNA is converted into SMRTbell templates using the PacBio SMRTbell Express Template Prep Kit 2.0 for sequencing on the Sequel System. Sequencing was performed at the University of Minnesota Genomics Center Facility (Minneapolis, MN).

The polished high-quality FASTA file obtained from Iso-seq3 was aligned to pennycress version 2 (T_arvense_v2) with minimap2 (Li, 2018). The resulting SAM file was sorted and collapsed using the cDNA_Cupcake package to obtain an input GFF file such that each transcript has exactly one alignment and at most one ORF prediction. `Sqanti3_qc.py`, part of the SQANTI3 package (Tardaguila *et al*., 2018), was deployed on the resulting GFF file along with the reference genome in the FASTA format and a GTF annotation file. This returned a reference corrected transcriptome, transcript-level and junction-level files with structural and quality descriptors, and a QC graphical report. Among the splice junction sites, SQANTI3 defines canonical junctions such as AT-AC, GC-AG and GT-AG, whereas all others are classified as non-canonical splice junctions.

Linkage disequilibrium analysis

Linkage disequilibrium (LD) among genome-wide markers and chromosome-specific markers was calculated with TASSEL v5.2.75 (Bradbury *et al*., 2007) with a sliding window size of 40 markers with 100 734 460 total comparisons. The *r*-squared values obtained via the linkage disequilibrium function in TASSEL were plotted against the physical distance with a LOESS curve fitted to the data to show LD decay (Figure S14).

Bulked-segregation sequencing and MutMap analysis

Bulked-segregant analysis (BSA) (Michelmore *et al*., 1991) coupled with whole-genome sequencing (BSA-Seq) was performed to locate genomic region harbouring the gene responsible for the pale mutant phenotype in pennycress (Figure 5d). Two pools were created with one pool containing leaf tissue from 20 individual pale mutants and the other pool consisting of wild-type individuals that did not exhibit the pale phenotype. DNA was extracted from fresh pennycress leaves using the DNeasy Plant Mini Kit (Qiagen, Valencia, CA). Both pools were sequenced on an Illumina HiSeq 2000 instrument using 2 × 125 base-paired reads at the University of Minnesota Genomics Center (Minneapolis, MN). The reads were analysed using the MutMap pipeline (Sugihara *et al*., 2020), and the QTL region was surveyed for candidate genes.

Comparison with YUN_Tarv_1.0

Synteny between T_arvense_v2 and YUN_Tarv_1.0 was assessed with minimap2 alignments and the resulting dotplot generated with the R package dotPlotly (https://github.com/tpoorten/dotPlotly). The k-mer analysis of quality and completeness was carried out for each assembly with Merqury v1.3 (Rhie *et al*., 2020; Table S7, S8 and S9), using both the PCR-free Illumina HiSeq reads generated in this study and those obtained from Geng *et al*. (2021) under the accession SRR14757813 in the NCBI Sequence Read Archive.

Acknowledgements

We thank Win Phippen, Thomas Gatter, Prabin Bajgain, Korbinian Schneeberger, Raúl Wijfjes, MPI DB Genome Center, Vienna Biocenter Core Facilities (VBCF), University of Minnesota Genomics Center (UMGC), Minnesota Supercomputing Institute (MSI), and all EpiDiverse network members and beneficiaries. We acknowledge the hard work of many who contributed to this study including Brett Heim, Krishan Rai, Nicole Folstad, Matthew A. Ott, Shweta Jain and many others.

Funding

This material is based upon work that is supported by the Minnesota Department of Agriculture (J.A., K.F., R.C.) and by the National Institute of Food and Agriculture, U.S. Department of Agriculture, under award numbers 2018-67009-27374 (J.A., R.C., K.F.), and 2019-67009-29004 (M.D.M, J.S.) and the Agriculture and Food Research Initiative Competitive Grant No. 2019-69012-29851 (M.D.M, R.C., J.S.). This research was supported by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics Science Program grant no. DE-SC0021286 (M.D.M, R.C.). This work was further funded by the Austrian Academy of Sciences (C.B., I.R.A., K.J., D.R.C.); the Max Planck Society (D.W., A.C.G., P.C.B., C.L.); the European Union's Horizon 2020 research and innovation programme by the European Research Council (ERC), Grant Agreement No. 716823 'FEAR-SAP' (I.R.A., C.B.), by the Marie Sklodowska-Curie ETN 'EpiDiverse', Grant Agreement No. 764965 (D.R.C., C.B.) and by Marie Sklodowska-Curie, Grant Agreement MSCA-IF No 797460 (P.C.B.); and the German Federal Ministry of Education and Research BMBF, Grant No. 031A538A, de.NBI-RBC (A.N., P.F.S.).

Conflicts of interest

The authors declare potential competing interests as intellectual property applications have been submitted on some of the genes discussed in this study.

Author contributions

RC, AN, KF and CB conceived the study. RC and AN led the genome assembly and evaluation, assisted by IRA and PCB. IRA performed the comparative genomics analysis of synteny during genome rescaffolding and in the final evaluation. AN led the genome annotation and performed analysis for protein-coding genes, non-coding genes (tRNA, rRNA, snoRNA) and pseudogenes. ACG performed small RNA library sequencing, annotation and analysis, supervised by DW. PZ and ACG performed the transposable element annotation, supervised by DW and MM. AN performed the gene expression analysis and evaluation of tissue specificity. CB and KJ provided PCR-free libraries. RC performed k-mer analysis for genome estimation. KF and RC provided the CCS libraries, which were prepared by the UMGC. CB and IRA provided the DNA methylation libraries and analysis. RC, ZT, MDM and KF developed linkage mapping populations, designed primers, performed genotyping and built genetic maps. KF, RC and KD generated resources for Hi-C, Bionano and resequencing of accessions. KF and ZT phenotyped resequenced accessions. RC performed SNP analysis of resequenced datasets. ZT performed the linkage disequilibrium decay analysis. AB performed population genomics. RC and BJ prepared samples for Iso-seq libraries. RC and ZT performed gene structure variation analysis. RC and MDM performed bulk-segregant analysis. The PacBio CLR library was prepared and sequenced by PCB and AN under the guidance of CL. DR prepared and sequenced mRNA-seq libraries. RC, AN, CB, ACG, IRA and ZT wrote the manuscript. All authors reviewed and approved the manuscript.

Data availability statement

The assembly and all NGS-based raw data are deposited in the ENA Sequence Read Archive Repository (www.ebi.ac.uk/ena/) under study accession number PRJEB46635. The summary of data provided by each institute and corresponding application is described in Table S10.

References

Amselem, J., Cornut, G., Choisne, N., Alaux, M., Alfama-Depauw, F., Jamilloux, V. and Maumus, F. (2019) RepetDB: a unified resource for transposable element references. *Mob. DNA*, **10**, 6.

Andersen, T.G. and Halkier, B.A. (2014) Upon bolting the GTR1 and GTR2 transporters mediate transport of glucosinolates to the inflorescence rather than roots. *Plant Signal. Behav.* **9**, e27740.

Axtell, M.J. (2013a) Classification and comparison of small RNAs from plants. *Annu. Rev. Plant Biol.* **64**, 137–159.

Axtell, M.J. (2013b) ShortStack: comprehensive annotation and quantification of small RNA genes. *RNA*, **19**, 740–751.

Bao, W., Kojima, K.K. and Kohany, O. (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA*, **6**, 11.

Beilstein, M.A., Nagalingum, N.S., Clements, M.D., Manchester, S.R. and Mathews, S. (2010) Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA*, **107**, 18724–18728.

Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580.

Beric, A., Mabry, M.E., Harkess, A.E., Brose, J., Schranz, M.E., Conant, G.C., Edger, P.P. *et al*. (2021) Comparative phylogenetics of repetitive elements in a diverse order of flowering plants (Brassicales). *G3 Genes|genomes|genetics*, **11**(7), https://doi.org/10.1093/g3journal/jkab140

Bewick, A.J., Ji, L., Niederhuth, C.E., Willing, E.-M., Hofmeister, B.T., Shi, X., Wang, L. *et al*. (2016) On the origin and evolutionary consequences of gene body DNA methylation. *Proc. Natl. Acad. Sci. USA*, **113**, 9111–9116.

Bewick, A.J. and Schmitz, R.J. (2017) Gene body DNA methylation in plants. *Curr. Opin. Plant Biol.* **36**, 103–110.

Boateng, A.A., Mullen, C.A. and Goldberg, N.M. (2010) Producing stable pyrolysis liquids from the oil-seed Presscakes of mustard family plants: pennycress (*Thlaspi arvense* L.) and Camelina (*Camelina sativa*). *Energy Fuels*, **24**, 6624–6632.

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M. and Bairoch, A. (2007) UniProtKB/Swiss-Prot. *Methods Mol. Biol.* **406**, 89–112.

Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y. and Buckler, E.S. (2007) TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, **23**, 2633–2635.

Bucher, E., Reinders, J. and Mirouze, M. (2012) Epigenetic control of transposon transcription and mobility in Arabidopsis. *Curr. Opin. Plant Biol.* **15**, 503–510.

Campbell, M.S., Holt, C., Moore, B. and Yandell, M. (2014) Genome annotation and curation using MAKER and MAKER-P. *Curr. Protoc. Bioinform.* **48**, 4.11.1–39.

Campbell, M.S., MeiYee, L., Carson, H., Stein, J.C., Moghe, G.D., Hufnagel, D.E., Jikai, L. *et al*. (2014) MAKER-P: A tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol.* **164**(2), 513–524. https://doi.org/10.1104/pp.113.230144

Catoni, M., Jonesman, T., Cerruti, E. and Paszkowski, J. (2018) Mobilization of Pack-CACTA transposons in Arabidopsis suggests the mechanism of gene shuffling. *Nucleic Acids Res.* **47**, 1311–1320.

Chakraborty, M., Baldwin-Brown, J.G., Long, A.D. and Emerson, J.J. (2016) Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Res.* **44**, e147.

Chopra, R., Folstad, N., Lyons, J. and Ulmasov, T. (2019) The adaptable use of Brassica NIRS calibration equations to identify pennycress variants to facilitate the rapid domestication of a new winter oilseed crop. *Ind. Crops Prod.* **128**, 55–61.

Chopra, R., Folstad, N. and Marks, M.D. (2020a) Combined genotype and fatty-acid analysis of single small field pennycress (*Thlaspi arvense*) seeds increases the throughput for functional genomics and mutant line selection. *Ind. Crops Prod.* **156**, 112823.

Chopra, R., Johnson, E.B., Daniels, E., McGinn, M., Dorn, K.M., Esfahanian, M., Folstad, N. *et al*. (2018) Translational genomics using Arabidopsis as a model enables the characterization of pennycress genes through forward and reverse genetics. *Plant J.* **96**(6), 1093–1105. https://doi.org/10.1111/tpj.14147

Chopra, R., Johnson, E.B., Emenecker, R., Cahoon, E.B., Lyons, J., Kliebenstein, D.J., Daniels, E. *et al*. (2020) Identification and stacking of crucial traits required for the domestication of pennycress. *Nature Food*, **1**(1), 84–91. https://doi.org/10.1038/s43016-019-0007-z

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J. *et al*. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms. *SnpEff. Fly*, **6**(2), 80–92. https://doi.org/10.4161/fly.19695

Claver, A., Rey, R., López, M.V., Picorel, R. and Alfonso, M. (2017) Identification of target genes and processes involved in erucic acid accumulation during seed development in the biodiesel feedstock Pennycress (*Thlaspi arvense* L.). *J. Plant Physiol.* **208**, 7–16.

Cubins, J.A., Wells, M.S., Frels, K., Ott, M.A., Forcella, F., Johnson, G.A., Walia, M.K. *et al*. (2019) Management of pennycress as a winter annual cash cover crop. A review. *Agron. Sustain. Dev.*, **39**, 5. https://doi.org/10.1007/s13593-019-0592-0

Del Gatto, A., Melilli, M.G., Raccuia, S.A., Pieri, S., Mangoni, L., Pacifico, D., Signor, M. *et al* (2015) A comparative study of oilseed crops (*Brassica napus* L. subsp. oleifera and Brassica carinata A. Braun) in the biodiesel production chain and their adaptability to different Italian areas. *Ind. Crops Prod.* **75**, 98–107. https://doi.org/10.1016/j.indcrop.2015.04.029

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P. *et al*. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dorn, K.M., Fankhauser, J.D., Wyse, D.L. and Marks, M.D. De novo assembly of the pennycress (*Thlaspi arvense*) transcriptome provides tools for the development of a winter cover crop and biodiesel feedstock. *Plant J.* 2013;**75**(6):1028-1038. https://doi.org/10.1111/tpj.12267

Dorn, K.M., Fankhauser, J.D., Wyse, D.L. and Marks, M.D. (2015) A draft genome of field pennycress (*Thlaspi arvense*) provides tools for the domestication of a new winter biofuel crop. *DNA Res.* **22**, 121–131.

Du, H. and Liang, C. (2019) Assembly of chromosome-scale contigs by efficiently resolving repetitive sequences with long reads. *Nat. Commun.* **10**, 5360.

Emms, D.M. and Kelly, S. (2017) STRIDE: Species tree root inference from gene duplication events. *Mol. Biol. Evol.* **34**, 3267–3278.

Emms, D.M. and Kelly, S. (2018) STAG: Species tree inference from all genes. *bioRxiv*, 267914.

Esfahanian, M., Nazarenus, T.J., Freund, M.M., McIntosh, G., Phippen, W.B., Phippen, M.E., Durrett, T.P. *et al*. (2021) Generating Pennycress (*Thlaspi arvense*) seed triacylglycerols and acetyl-triacylglycerols containing medium-chain fatty acids. *Front. Energy Res.* **9**, https://doi.org/10.3389/fenrg.2021.620118

Fan, J., Shonnard, D.R., Kalnes, T.N., Johnsen, P.B. and Rao, S. (2013) A life cycle assessment of pennycress (*Thlaspi arvense* L.) -derived jet fuel and diesel. *Biomass Bioenergy*, **55**, 87–100.

Flynn, J.M., Hubley, R., Goubert, C., Rosen, J., Clark, A.G., Feschotte, C. and Smit, A.F. (2020) RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA*, **117**, 9451–9457.

Franzke, A., Lysak, M.A., Al-Shehbaz, I.A., Koch, M.A. and Mummenhoff, K. (2011) Cabbage family affairs: the evolutionary history of Brassicaceae. *Trends Plant Sci.* **16**, 108–116.

Frels, K., Chopra, R., Dorn, K.M., Wyse, D.L., Marks, M.D. and Anderson, J.A. (2019) Genetic diversity of field pennycress (*Thlaspi arvense*) reveals untapped variability and paths toward selection for domestication. *Agronomy*, **9**, 302.

Gatter, T. and Stadler, P.F. (2019) Ryūtō: network-flow based transcriptome reconstruction. *BMC Bioinform.* **20**, 190.

Geng, Y., Guan, Y., Qiong, L., Lu, S., An, M., Crabbe, M.J.C., Qi, J. *et al*. (2021) Genomic analysis of field pennycress (*Thlaspi arvense*) provides insights into mechanisms of adaptation to high elevation. *BMC Biol.* **19**(1), https://doi.org/10.1186/s12915-021-01079-0

Ghurye, J., Pop, M., Koren, S., Bickhart, D. and Chin, C.-S. (2017) Scaffolding of long read assemblies using long range contact information. *BMC Genom.* **18**, 527.

Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T. *et al*. (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* **40**(D1), D1178–D1186. https://doi.org/10.1093/nar/gkr944

Guan, D., McCarthy, S.A., Wood, J., Howe, K., Wang, Y. and Durbin, R. (2020) Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*, **36**, 2896–2898.

Gurevich, A., Saveliev, V., Vyahhi, N. and Tesler, G. (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, **29**, 1072–1075.

Hardcastle, T.J., Müller, S.Y. and Baulcombe, D.C. (2018) Towards annotating the plant epigenome: the *Arabidopsis thaliana* small RNA locus map. *Sci. Rep.* **8**, 6338.

He, G., Chen, B., Wang, X., Li, X., Li, J., He, H., Yang, M. *et al*. (2013) Conservation and divergence of transcriptomic and epigenomic variation in maize hybrids. *Genome Biol.* **14**(6), https://doi.org/10.1186/gb-2013-14-6-r57

Hill, W.G. and Weir, B.S. (1988) Variances and covariances of squared linkage disequilibria in finite populations. *Theor. Popul. Biol.* **33**, 54–78.

Jarvis, B.A., Romsdahl, T.B., McGinn, M.G., Nazarenus, T.J., Cahoon, E.B., Chapman, K.D. and Sedbrook, J.C. (2021) CRISPR/Cas9-induced fad2 and rod1 mutations stacked with fae1 confer high oleic acid seed oil in pennycress (*Thlaspi arvense* L.). *Front. Plant Sci.* **12**, 652319.

Johnson, G.A., Kantar, M.B., Betts, K.J. and Wyse, D.L. (2015) Field pennycress production and weed control in a double crop system with soybean in Minnesota. *Agron. J.* **107**, 532–540.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H. *et al*. (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**(9), 1236–1240. https://doi.org/10.1093/bioinformatics/btu031

Kamvar, Z.N., Tabima, J.F. and Grünwald, N.J. (2014) Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ*, **2**, e281.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H. and Phillippy, A.M. (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736.

Korf, I. (2004) Gene finding in novel genomes. *BMC Bioinform.* **5**, 59.

Krzywinski, M., Schein, J., Birol, İ., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J. *et al*. (2009) Circos: An information aesthetic for comparative genomics. *Genome Res.* **19**(9), 1639–1645. https://doi.org/10.1101/gr.092759.109

Lagesen, K., Hallin, P., Rødland, E.A., Staerfeldt, H.-H., Rognes, T. and Ussery, D.W. (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35**, 3100–3108.

Lam, E.T., Hastie, A., Lin, C., Ehrlich, D., Das, S.K., Austin, M.D., Deshpande, P. *et al*. (2012) Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat. Biotechnol.* **30**(8), 771–776. https://doi.org/10.1038/nbt.2303

Li, H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–3100.

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G. *et al*. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Liao, Y., Smyth, G.K. and Shi, W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.

Lowe, T.M. and Eddy, S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964.

Lu, K., Wei, L., Li, X., Wang, Y., Wu, J., Liu, M., Zhang, C. *et al*. (2019) Whole-genome resequencing reveals *Brassica napus* origin and genetic loci involved in its improvement. *Nat. Commun.* **10**(1), https://doi.org/10.1038/s41467-019-09134-9

Lunardon, A., Johnson, N.R., Hagerott, E., Phifer, T., Polydore, S., Coruh, C. and Axtell, M.J. (2020) Integrated annotations and analyses of small RNA–producing loci from 47 diverse plants. *Genome Res.* **30**, 497–513.

Marçais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L. and Zimin, A. (2018) MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol*, **14**, e1005944.

Marçais, G. and Kingsford, C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, **27**, 764–770.

Marks, M.D., Chopra, R. and Sedbrook, J.C. (2021) Technologies enabling rapid crop improvements for sustainable agriculture: example pennycress (*Thlaspi arvense* L.). *Emerg. Top Life Sci.* **5**, 325–335.

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

McGinn, M., Phippen, W.B., Chopra, R., Bansal, S., Jarvis, B.A., Phippen, M.E., Dorn, K.M. *et al.* (2019) Molecular tools enabling pennycress (*Thlaspi arvense*) as a model plant and oilseed cash cover crop. *Plant Biotechnol. J.* **17** (4), 776–788. https://doi.org/10.1111/pbi.13014

Meinke, D.W., Cherry, J.M., Dean, C., Rounsley, S.D. and Koornneef, M. (1998) *Arabidopsis thaliana*: a model plant for genome analysis. *Science*, **282**, 662, 679–82.

Michalovova, M., Vyskot, B. and Kejnovsky, E. (2013) Analysis of plastid and mitochondrial DNA insertions in the nucleus (NUPTs and NUMTs) of six plant species: size, relative age and chromosomal localization. *Heredity*, **111**, 314–320.

Michelmore, R.W., Paran, I. and Kesseli, R.V. (1991) Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. *Proc. Natl. Acad. Sci. USA*, **88**, 9828–9832.

Moore, S.A., Wells, M.S., Gesch, R.W., Becker, R.L., Rosen, C.J. and Wilson, M.L. (2020) Pennycress as a cash cover-crop: improving the sustainability of sweet corn production systems. *Agronomy*, **10**, 614.

Moser, B.R. (2012) Biodiesel from alternative oilseed feedstocks: camelina and field pennycress. *Biofuels*, **3**, 193–209.

Moser, B.R., Knothe, G., Vaughn, S.F. and Isbell, T.A. (2009) Production and evaluation of biodiesel from field pennycress (*Thlaspi arvense* L.) oil. *Energy Fuels*, **23**, 4149–4155.

Mulligan, G.A. (1957) Chromosome numbers of Canadian weeds. I. *Can. J. Bot.* **35**, 779–789.

Mulligan, G.A. and Kevan, P.G. (1973) Color, brightness, and other floral characteristics attracting insects to the blossoms of some Canadian weeds. *Can. J. Bot.* **51**, 1939–1952.

Murat, F., Louis, A., Maumus, F., Armero, A., Cooke, R., Quesneville, H., Crollius, H.R. *et al.* (2015) Understanding Brassicaceae evolution through ancestral genome reconstruction. *Genome Biol.* **16**(1), https://doi.org/10.1186/s13059-015-0814-y

Nawrocki, E.P. and Eddy, S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–2935.

Niederhuth, C.E., Bewick, A.J., Ji, L., Alabady, M.S., Kim, K.D., Li, Q., Rohr, N.A. *et al.* (2016) Widespread natural variation of DNA methylation within angiosperms. *Genome Biol.* **17**, 194.

Niittylä, T., Messerli, G., Trevisan, M., Chen, J., Smith, A.M. and Zeeman, S.C. (2004) A previously unknown maltose transporter essential for starch degradation in leaves. *Science*, **303**, 87–89.

Niknafs, Y.S., Pandian, B., Iyer, H.K., Chinnaiyan, A.M. and Iyer, M.K. (2017) TACO produces robust multisample transcriptome assemblies from RNA-seq. *Nat. Methods*, **14**, 68–70.

Nour-Eldin, H.H., Andersen, T.G., Burow, M., Madsen, S.R., Jorgensen, M.E., Olsen, C.E. and Dreyer, I. (2012) NRT/PTR transporters are essential for translocation of glucosinolate defence compounds to seeds. *Nature*, **488**, 531–534.

Ott, M.A., Eberle, C.A., Thom, M.D., Archer, D.W., Forcella, F., Gesch, R.W. and Wyse, D.L. (2019) Economics and agronomics of relay-cropping pennycress and Camelina with soybean in Minnesota. *Agron. J.* **111**, 1281–1292.

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J.R.A., Hellinga, A.J., Lugo, C.S.B. *et al.* (2019) Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**(1), https://doi.org/10.1186/s13059-019-1905-y

Patra Bhattacharya, D., Canzler, S., Kehr, S., Hertel, J., Grosse, I. and Stadler, P.F. (2016) Phylogenetic distribution of plant snoRNA families. *BMC Genom.* **17**, 969.

Pedersen, B.S., Eyring, K., De, S., Yang, I.V. and Schwartz, D.A. (2014) *Fast and accurate alignment of long bisulfite-seq reads. arXiv [q-bio.GN].*

Phippen, W.B. and Phippen, M.E. (2012) Soybean seed yield and quality as a response to field pennycress residue. *Crop Sci.* **52**, 2767–2773.

Pritchard, J.K., Stephens, M. and Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Pruitt, K.D., Tatusova, T., Brown, G.R. and Maglott, D.R. (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.* **40**, D130–D135.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J. *et al.* (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Human Genet.* **81**(3), 559–575. https://doi.org/10.1086/519795

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

Ren, S., Bertels, K. and Al-Ars, Z. (2018) Efficient acceleration of the Pair-HMMs forward algorithm for GATK HaplotypeCaller on graphics processing units. *Evol. Bioinform. Online*, **14**, 1176934318760543.

Rhie, A., Walenz, B.P., Koren, S. and Phillippy, A.M. (2020) Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245.

Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Robinson, M.D. and Oshlack, A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25.

Scheben, A., Yuan, Y. and Edwards, D. (2016) Advances in genomics for adapting crops to climate change. *Curr. Plant Biol.* **6**, 2–10.

Schranz, M.E., Lysak, M.A. and Mitchell-Olds, T. (2006) The ABC's of comparative genomics in the Brassicaceae: building blocks of crucifer genomes. *Trends Plant Sci.* **11**, 535–542.

Sedbrook, J.C., Phippen, W.B. and Marks, M.D. (2014) New approaches to facilitate rapid domestication of a wild plant to an oilseed crop: example pennycress (*Thlaspi arvense* L.). *Plant Sci*. **227**, 122–132.

Shumate, A. and Salzberg, S.L. (2020) Liftoff: accurate mapping of gene annotations. *Bioinformatics*, **37**, 1639–1643.

Sigman, M.J. and Slotkin, R.K. (2016) The first rule of plant transposable element silencing: location, location, location. *Plant Cell*, **28**, 304–313.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. and Zdobnov, E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.

Smit, A.F.A. (2004) *Repeat-Masker Open-3.0*. http://www.repeatmasker.org

Smit, A.F.A. and Hubley, R. (2008) *RepeatModeler Open-1.0*.

Sonnhammer, E.L. and Durbin, R. (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene*, **167**, GC1–10.

Stam, P. (1993) Construction of integrated genetic linkage maps by means of a new computer package: Join Map. *Plant J.* **3**, 739–744.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S. and Morgenstern, B. (2006) AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439.

Sugihara, Y., Young, L., Yaegashi, H., Natsume, S., Shea, D.J., Takagi, H., Booker, H. *et al.* (2020) *High-performance pipeline for MutMap and QTL-seq*. bioRxiv, 2020.06.28.176586.

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

Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M. and Paterson, A.H. (2008) Synteny and collinearity in plant genomes. *Science*, **320**, 486–488.

Tang, H., Zhang, X., Miao, C., Zhang, J., Ming, R., Schnable, J.C., Schnable, P.S. *et al.* (2015) ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biol.* **16**, 3.

Tardaguila, M., de la Fuente, L., Marti, C., Pereira, C., Pardo-Palacios, F.J., del Risco, H., Ferrell, M. *et al.* (2018) SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res.* **28**(3), 396–411. https://doi.org/10.1101/gr.222976.117

Thomas, J.B., Hampton, M.E., Dorn, K.M., David Marks, M. and Carter, C.J. (2017) The pennycress (*Thlaspi arvense* L.) nectary: structural and transcriptomic characterization. *BMC Plant Biol.* **17**, 201.

Tomato Genome Consortium. (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, **485**, 635–641.

Vaser, R., Sović, I., Nagarajan, N. and Šikić, M. (2017) Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746.

Voinnet, O. (2009) Origin, biogenesis, and activity of plant microRNAs. *Cell*, **136**, 669–687.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A. *et al*. (2014) Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, **9** (11), e112963. https://doi.org/10.1371/journal.pone.0112963

Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., Bai, Y. *et al*. (2011) The genome of the mesopolyploid crop species *Brassica rapa*. *Nat. Genet.* **43**(10), 1035–1039. https://doi.org/10.1038/ng.919

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

Weyers, S.L., Gesch, R.W., Forcella, F., Eberle, C.A., Thom, M.D., Matthees, H.L., Ott, M. *et al*. (2021) Surface runoff and nutrient dynamics in cover crop–soybean systems in the Upper Midwest. *J. Environ. Qual.* **50**(1), 158–171. https://doi.org/10.1002/jeq2.20135

Weyers, S., Thom, M., Forcella, F., Eberle, C., Matthees, H., Gesch, R., Ott, M. *et al*. (2019) Reduced potential for nitrogen loss in cover crop-soybean relay systems in a cold climate. *J. Environ. Qual.* **48**(3), 660–669. https://doi.org/10.2134/jeq2018.09.0350

Workman, R., Fedak, R., Kilburn, D., Hao, S., Liu, K. and Timp, W. (2019) *High molecular weight DNA extraction from recalcitrant plant species for third generation sequencing v1 (protocols.Io.4vbgw2n). protocols.io.*.

Workman, R.E., Myrka, A.M., Wong, G.W., Tseng, E., Welch, K.C. and Timp, W. (2018) Single-molecule, full-length transcript sequencing provides insight into the extreme metabolism of the ruby-throated hummingbird Archilochus colubris. *GigaScience*, **7**(3), giy009. https://doi.org/10.1093/gigascience/giy009

Wu, D., Liang, Z., Yan, T., Xu, Y., Xuan, L., Tang, J., Zhou, G. *et al*. (2019) Whole-genome resequencing of a worldwide collection of rapeseed accessions reveals the genetic basis of ecotype divergence. *Molecular Plant*, **12**(1), 30–43. https://doi.org/10.1016/j.molp.2018.11.007

Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., Bar-Even, A. *et al*. (2005) Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics*, **21**(5), 650–659. https://doi.org/10.1093/bioinformatics/bti042

Yang, R., Jarvis, D.E., Chen, H., Beilstein, M.A., Grimwood, J., Jenkins, J., Shu, S. *et al*. (2013) The reference genome of the halophytic plant *Eutrema salsugineum*. *Front. Plant Sci.* **4**, https://doi.org/10.3389/fpls.2013.00046

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

Zhang, S.-J., Liu, L., Yang, R. and Wang, X. (2020) Genome size evolution mediated by gypsy retrotransposons in Brassicaceae. *Genom. Proteom. Bioinform.* **18**, 321–332.

Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S.W.-L., Chen, H., Henderson, I.R. *et al*. (2006) Genome-wide high-resolution mapping and functional analysis of DNA methylation in Arabidopsis. *Cell*, **126**(6), 1189–1201. https://doi.org/10.1016/j.cell.2006.08.003

Zilberman, D., Gehring, M., Tran, R.K., Ballinger, T. and Henikoff, S. (2006) Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. *Nat. Genet.* **39**, 61–69.

Zou, C., Lehti-Shiu, M.D., Thibaud-Nissen, F., Prakash, T., Buell, C.R. and Shiu, S.-H. (2009) Evolutionary and expression signatures of pseudogenes in Arabidopsis and rice. *Plant Physiol.* **151**, 3–15.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure S1 Karyotype plot of the seven largest scaffolds representing chromosomes in *T. arvense* MN106-Ref (T_arvense_v2), alongside a concatenation of all minor scaffolds.

Figure S2 Integrative Genome Viewer (IGV) snapshot of PacBio read coverage (top track) over the largest seven scaffolds of the genome, including distributions of genes (middle track) and transposable elements (bottom track).

Figure S3 Sequence dot plots showing the largest seven scaffolds of the closely-related species *E. salsugineum* and their equivalent in *T. arvense* var. MN106-Ref (T_arvense_v2), comparing the difference both (a) before and (b) after rescaffolding.

Figure S4 Synteny analysis between the largest seven scaffolds of *T. arvense* var. MN106-Ref (Ta) and (a) their equivalent in the closely-related species *E. salsugineum* (Es), and (b) *A. thaliana* (At).

Figure S5 The cumulative distribution of annotation edit distance (AED) scores from the final set of protein-coding loci, denoting that ~95% of annotated genes are supported with a score ≤ 0.5 overall.

Figure S6 An overview of annotated genomic feature distributions in comparison to T_arvense_v1 for (a) gene lengths, (b) CDS lengths, (c) per gene exon number, and (d) intron lengths.

Figure S7 Small RNA (sRNA) annotation in the T_arvense_v2 genome assembly.

Figure S8 Predicted miRNAs in the T_arvense_v2 genome assembly.

Figure S9 Relative expression level of novel and conserved miRNA families between tissue types.

Figure S10 sRNA types and their association with different genomic features.

Figure S11 Methylation rate frequency distribution by sequence context in shoot and root tissues.

Figure S12 Map showing original sampling sites of pennycress accessions used for resequencing analysis in this study.

Figure S13 Structure plot showing inferred population membership for SNP data (top) and Indel data (bottom) at $k = 3$ for the resequenced accessions.

Figure S14 Genome-wide linkage disequilibrium decay plotted against physical distance for MN106-Ref (T_arvense_v2) at an r-squared value of 0.2 and chromosome level LD decay described in the right. Linkage disequilibrium (LD) was calculated using 2 518 379 genome-wide markers with a sliding window of 40 markers.

Figure S15 Synteny between T_arvense_v2 (*x*-axis) and YUN_-Tarv_1.0 (*y*-axis).

Figure S16 Read length distribution of trimmed PacBio Sequel II HiFi CLR reads taken forward for assembly with Canu v1.9.

Figure S17 Distribution of PacBio Sequel II HiFi CLR read mapping depth frequency over assembled contigs, with bimodal peaks due to contig regions with lower depth than the average indicating that they are duplicated.

Table S1 Estimation of the genome size of *T. arvense* using flow cytometry with *Arabidopsis thaliana*, tomato (*Solanum lycopersicum*), and maize (*Zea mays*) as references.

Table S2 Full descriptive statistics for intermediate versions of the assembly starting with correction, trimming and initial assembly of PacBio reads (Canu), further polishing and scaffolding using optical maps and contact maps (Bionano + HiC), and the final version following manual curation and rescaffolding with the help of genetic linkage and synteny maps (ALLMAPS).

Table S3 Alignment statistics of mRNA-seq reads prior to merging by tissue type.

Table S4 Detailed per-class statistics of the transposable element fraction of the *T. arvense* genome.

Table S5 Description of genes identified in the QTL region (Scaffold_6: 63.85–63.95 Mbp) of the BSA analysis of pale seedling phenotype in pennycress.

Table S6 BUSCO statistics on (a) initial assembly, immediately after CANU, and (b) final assembly. Both are derived from orthologs to the *Eudicotyledons odb10* database.

Table S7 Merqury k-mer (k = 21) analysis of Illumina HiSeq reads sequenced from the accession in YUN_Tarv_1.0, showing greater QV scores in T_arvense_v2 for the equivalent top 7 scaffolds based on k-mers found uniquely in each assembly and those shared with the read set.

Table S8 Merqury k-mer (k = 21) analysis of Illumina HiSeq reads (PCR-free) sequenced from the accession MN106-Ref, showing greater QV scores in T_arvense_v2 for the equivalent top 7 scaffolds based on k-mers found uniquely in each assembly and those shared with the read set.

Table S9 Merqury k-mer (k = 21) analysis of each total assembly showing relative completeness of k-mers present in each read set from Illumina HiSeq.

Table S10 Summary of data provided by each institute and corresponding application.

Appendix S1 Manual curation of predicted miRNAs.

Data S1 Normalized read counts for the genes expressed in each of the tissues analysed (See excel file). Tau values are incorporated in each of the genes to highlight the specificity.

Data S2 Top 30 most-expressed genes in each tissue, relative to the mean across all tissues, from the subset of genes with a high/absolute tau specificity score.

Data S3 Location of SNPs and the primers used in the genotyping of EMS-based population for development of linkage map.

Data S4 Genetic map developed using an F2 population derived from MN106 and Ames32867.

Data S5 Genetic map developed using an F2 population derived from MN106 and 2019-M2-111.

Data S6 Phenotypes, total reads, and coverage associated with the accessions used for GWAS.

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).

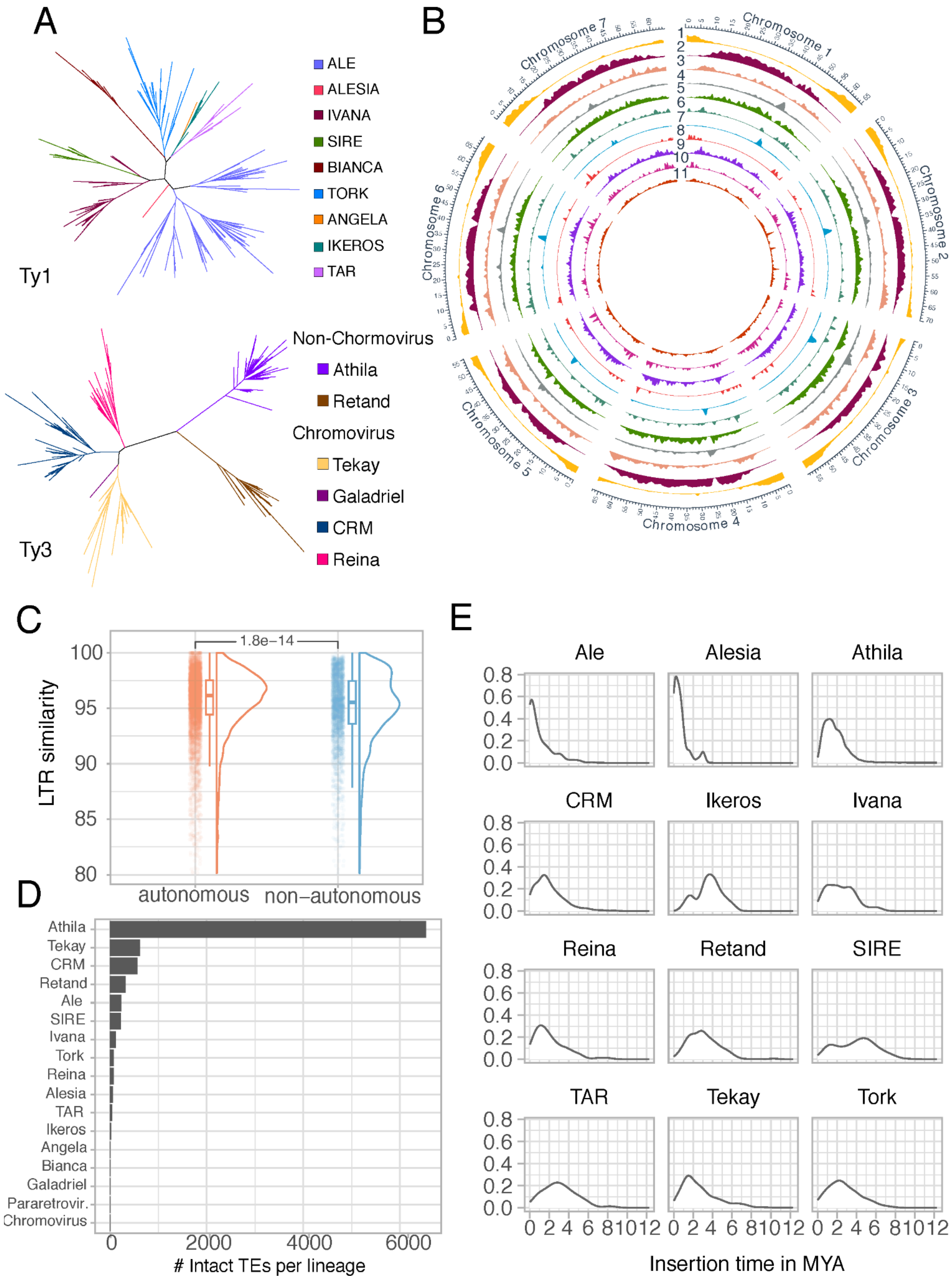


Figure 1. Genome-wide distribution and classification of TE families and superfamilies in the *T. arvense* reference genome MN106-Ref. A, Phylogenetic tree of LTR retrotransposons based on the reverse transcriptase domain. **B,** Genome-wide distribution of TE family and superfamily abundances. The tracks 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).

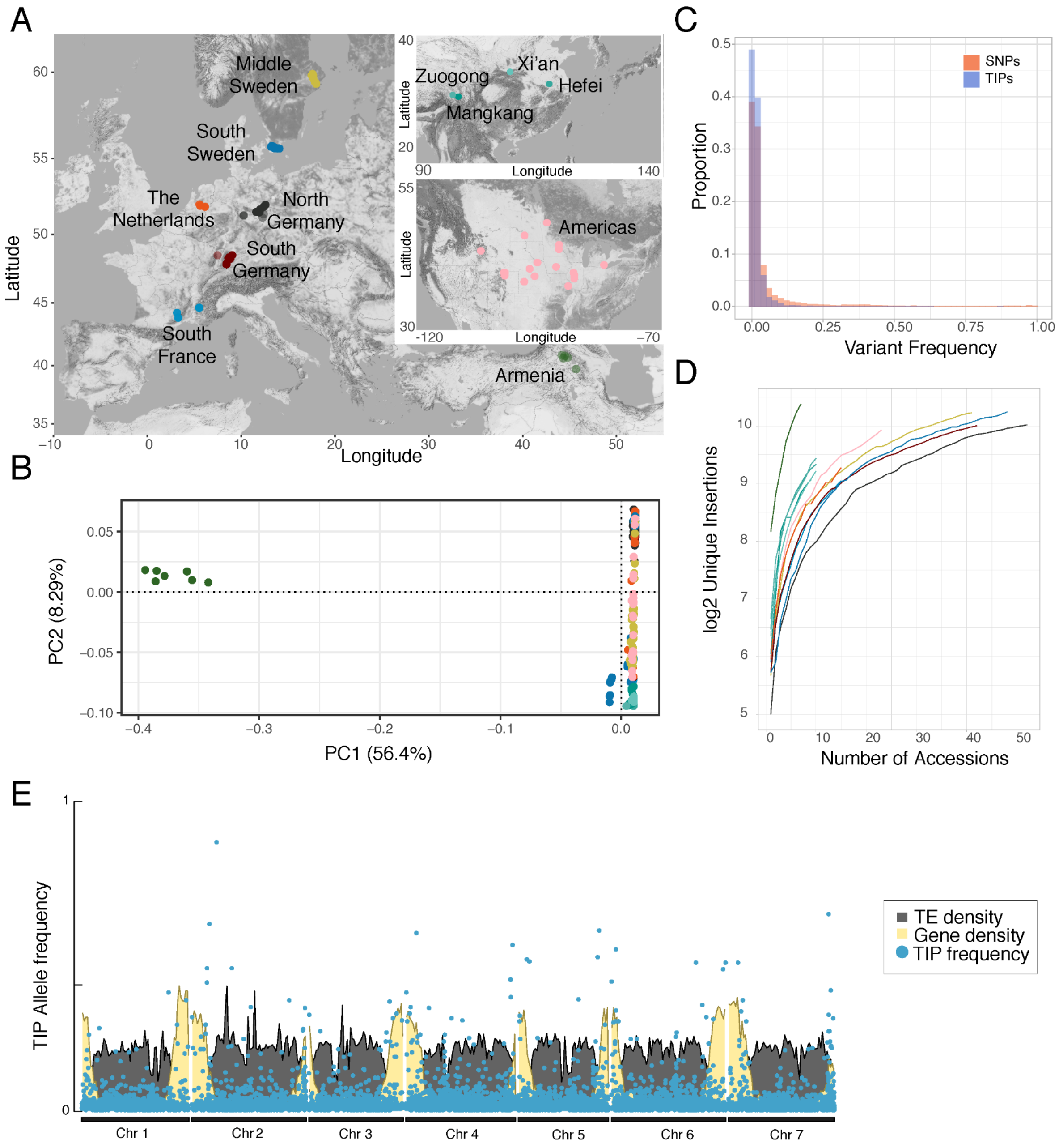


Figure 2. The genome-wide landscape of TE insertion polymorphisms in *T. arvense*.
A, Distribution of accessions across their native Eurasian and naturalized North American range in the Northern hemisphere (omitting a sample from Chile, included in the Americas group). **B**, A SNP-based principal component 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.

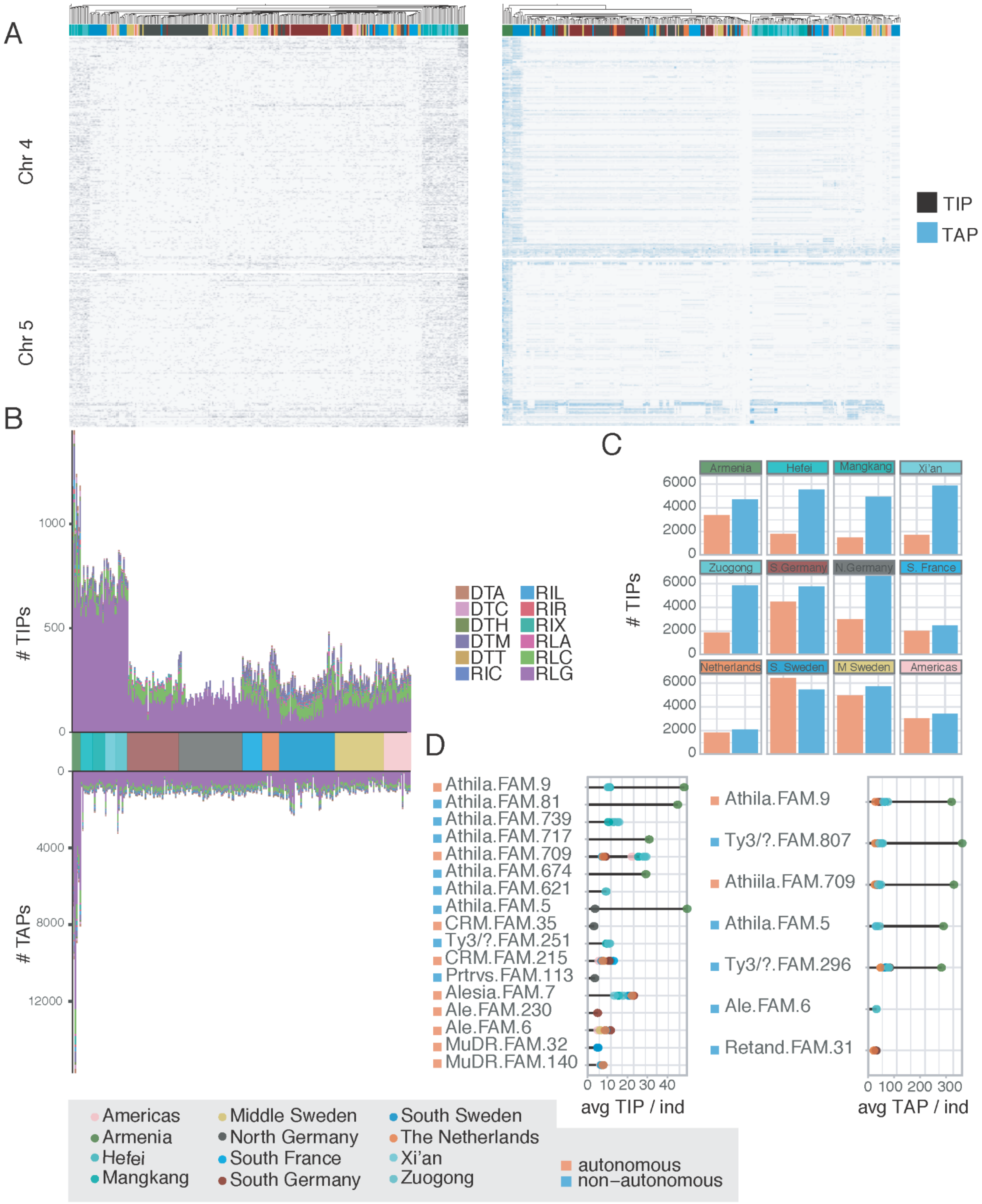
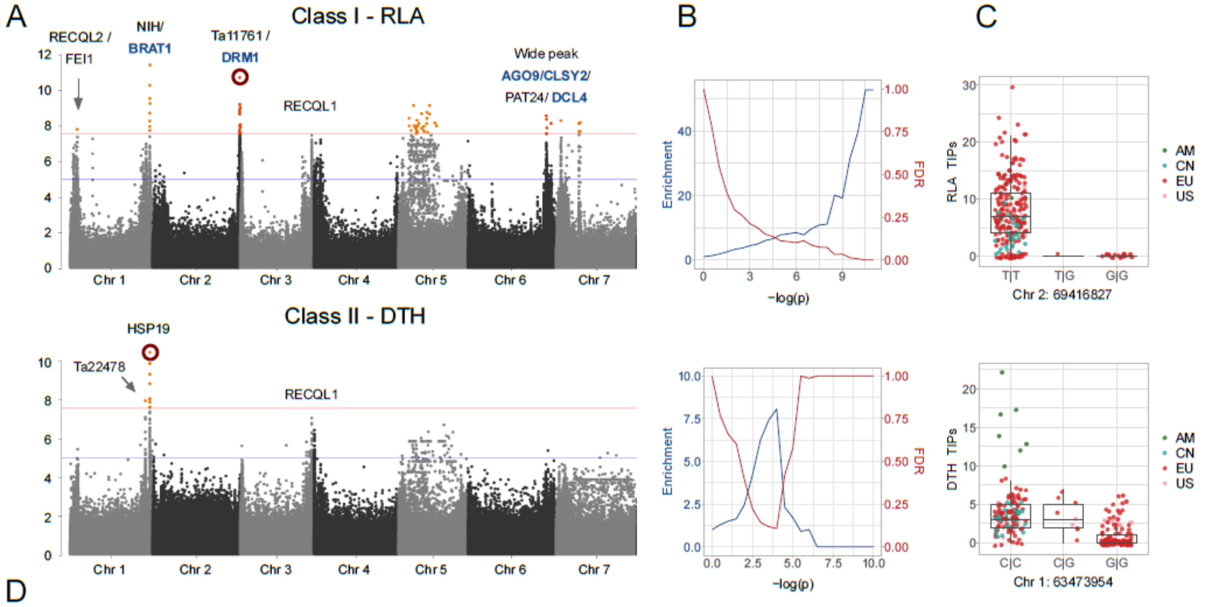


Figure 3. The *T. arvense* mobilome. A, Genomic distribution of TIPs and TAPs in chromosomes 4 and 5, where we observe major TIP/TAP hotspots. TIPs and TAPs along the other chromosomes are shown in Fig. S6. **B**, Contribution of different superfamilies to transposon insertion polymorphisms (TIPs) and 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.



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)

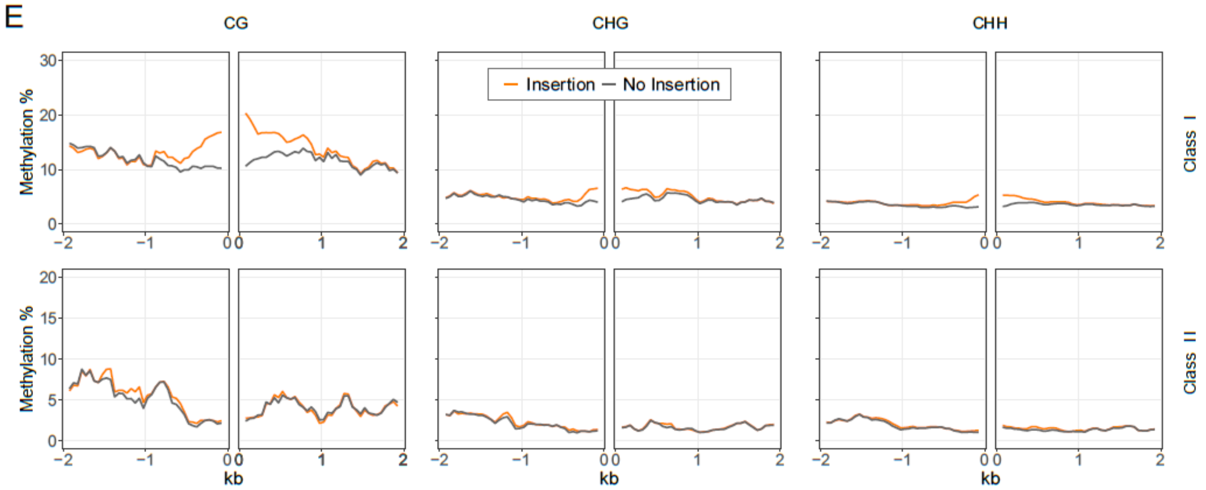


Figure 4. GWA analysis for TIP load of a class I and a class II TE superfamily. Results including all superfamilies are shown in Fig S5. **A**, Manhattan plots with candidate genes indicated next to neighboring variants. The red line corresponds to a genome-wide significance with full Bonferroni correction, 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.

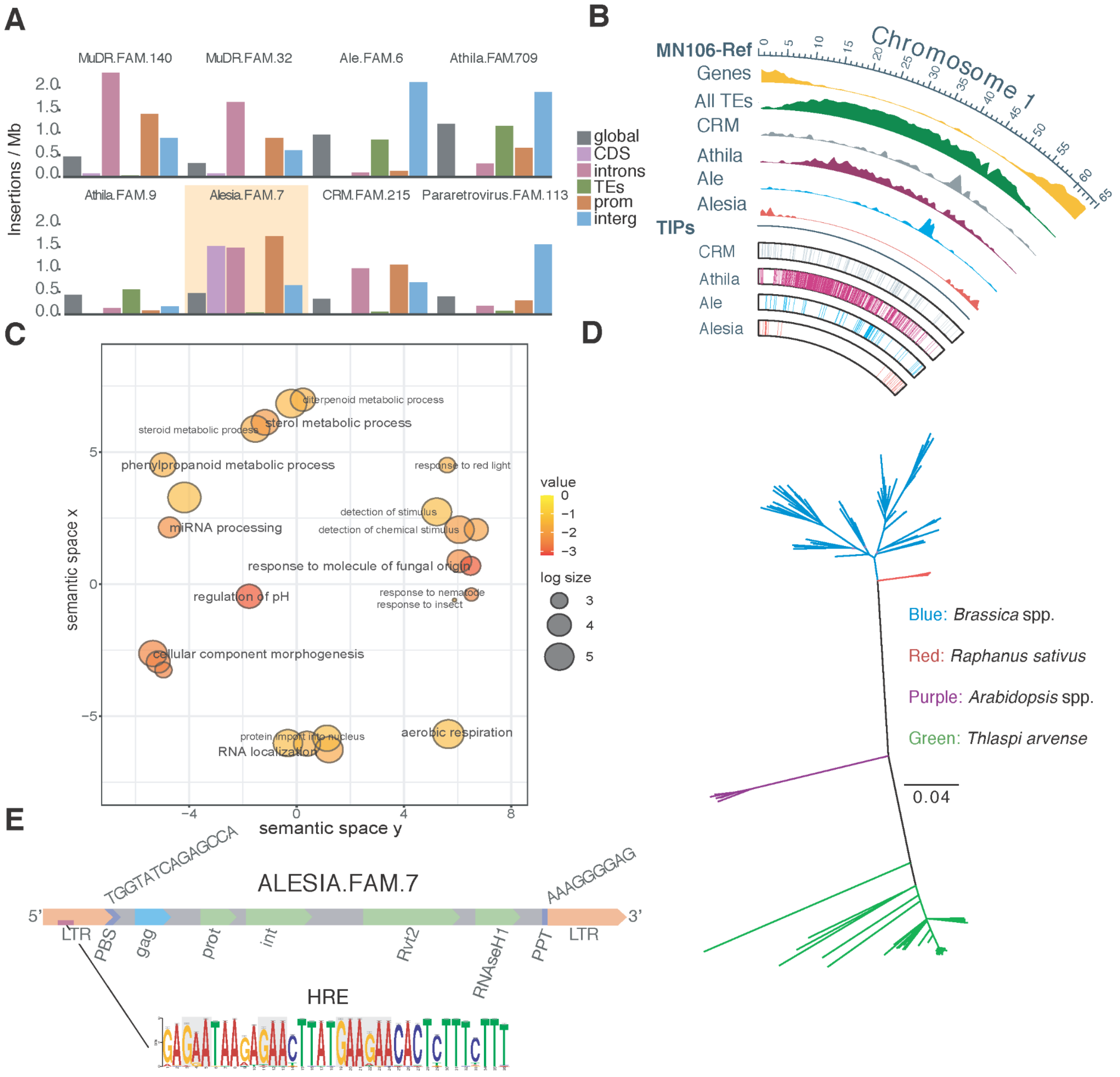


Figure 5. Summary statistics and characterization of the Alesia.FAM.7 family in *T. arvense* and other Brassicaceae. A, Distribution of several TE families across different genomic contexts in *T. arvense* accessions. While several other families, such as MuDR.FAM.140 or CRM.FAM.215, are also often found 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

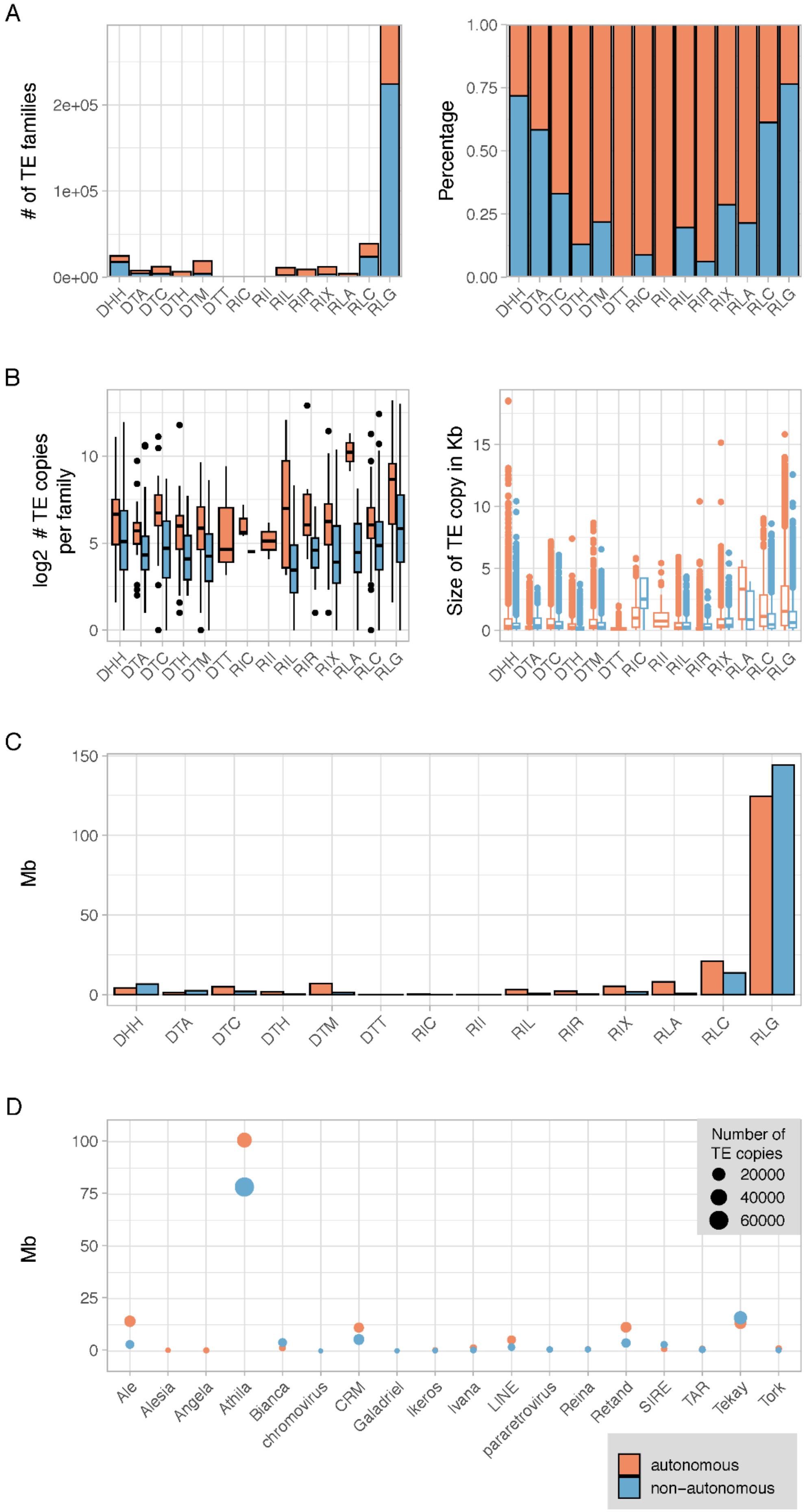
Table S2. Lineages of LTR-TEs in the *T. arvense* genome MN106-Ref.

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.


A
# of TE families
0e+00
1e+05
2e+05
Percentage
0.00
0.25
0.50
0.75
1.00
DHH
DTA
DTC
DTH
DTM
DTT
RIC
RII
RIL
RIR
RIX
RLA
RLC
RLG
B
log2 # TE copies per family
0
5
10
Size of TE copy in Kb
0
5
10
15
C
Mb
0
50
100
150
D
Mb
0
25
50
75
100
Number of TE copies
20000
40000
60000
Ale
Alesia
Angela
Athila
Bianca
chromovirus
CRM
Galadriel
Ikeros
Ivana
LINE
pararetrovirus
Peina
Retand
SIRE
TAR
Tekay
Tork
autonomous
non−autonomous

S4

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.

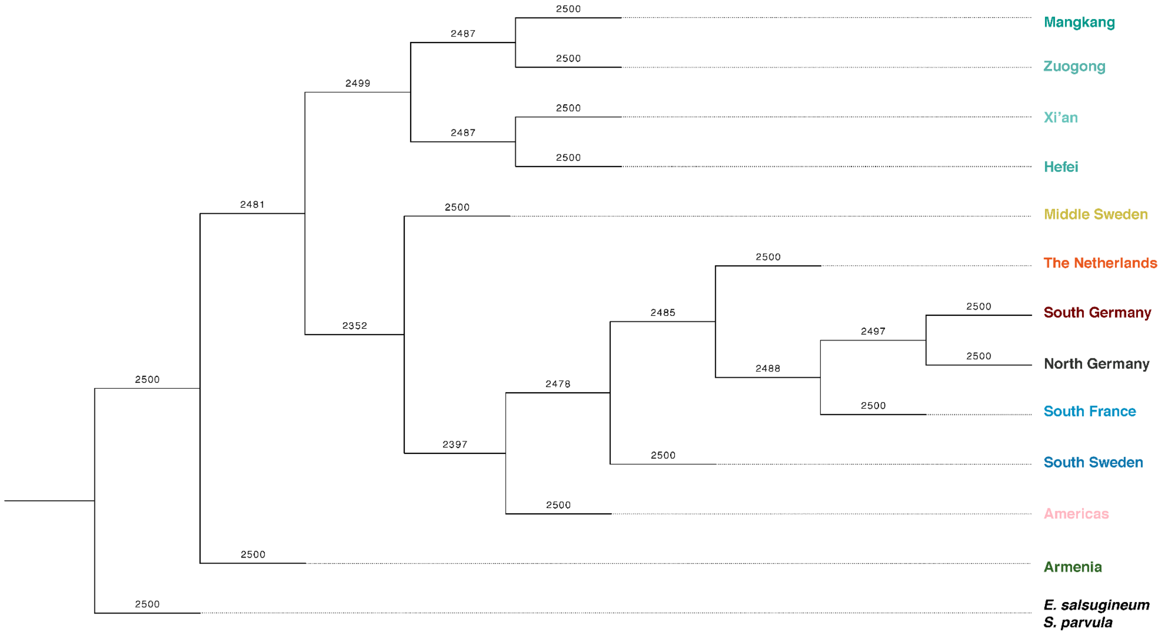


Figure S2. SNP-based maximum likelihood tree of *T. arvense* populations. Based on a model without migration, 2,500 bootstraps. Node weights represent bootstrap values. Outgroup species at the bottom.

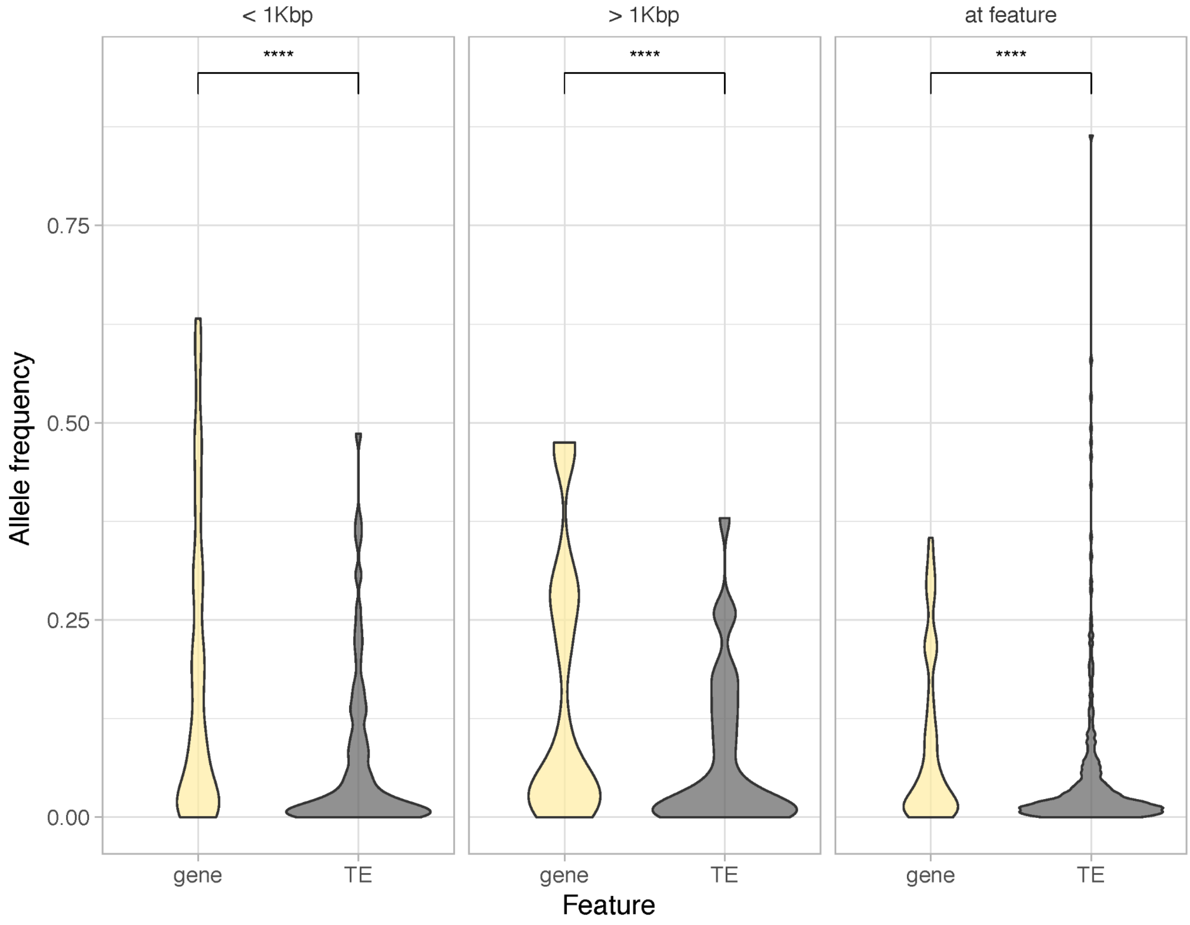


Figure S3. Frequency distribution of TIPs overlapping with annotated genes and TEs. TIP frequencies near other TEs are significantly lower than near genes (Wilcoxon Rank Sum test, $p < 2.22E^{-16}$).

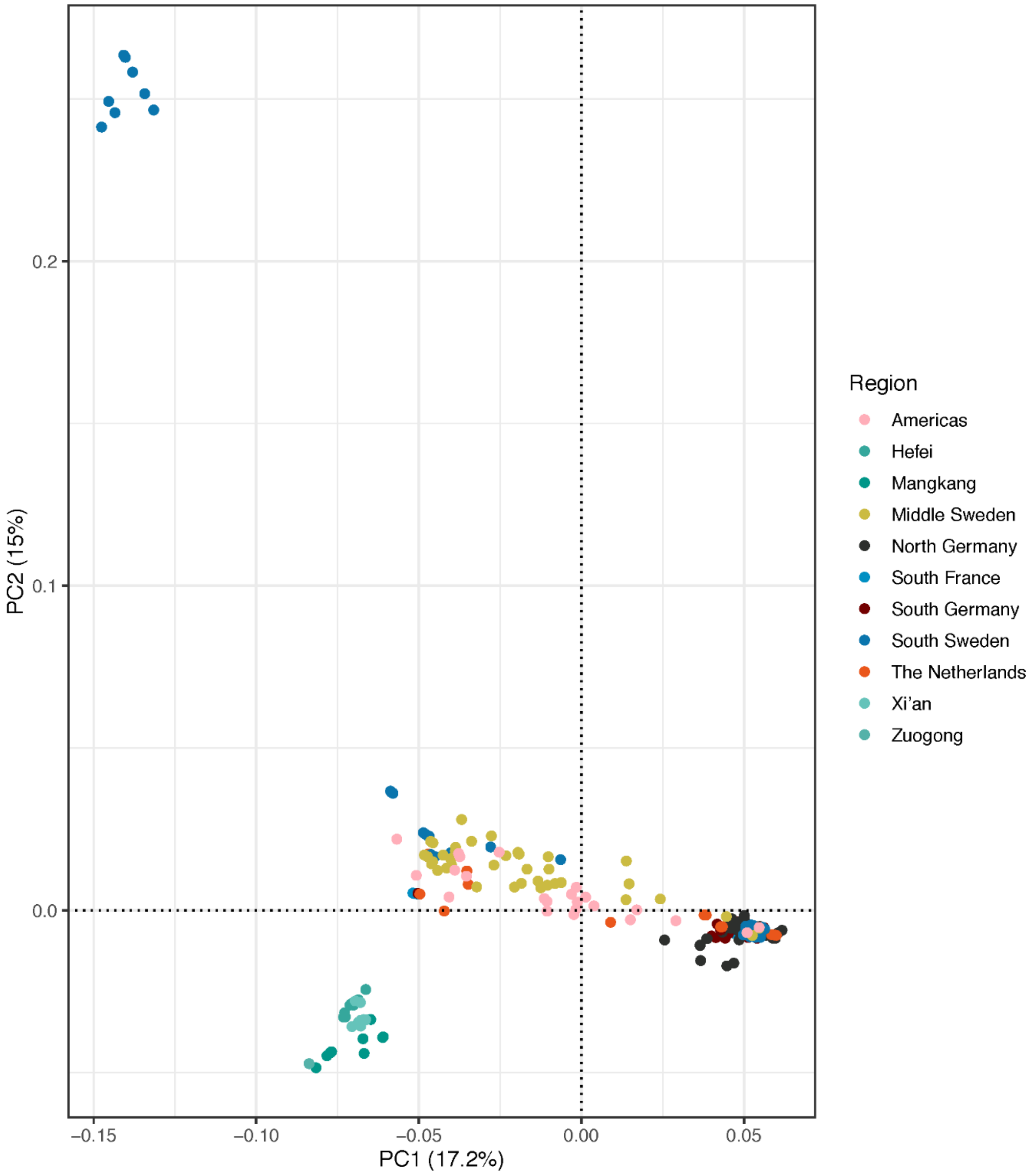


Figure S4. SNP-based PCA of a subset of *T. arvense* accessions. The Armenian accessions, which are outliers in the PCA using all accessions (Fig. 2), were excluded from this new PCA analysis, which shows how Chinese and European accessions cluster separately. We also observe part of the south Sweden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

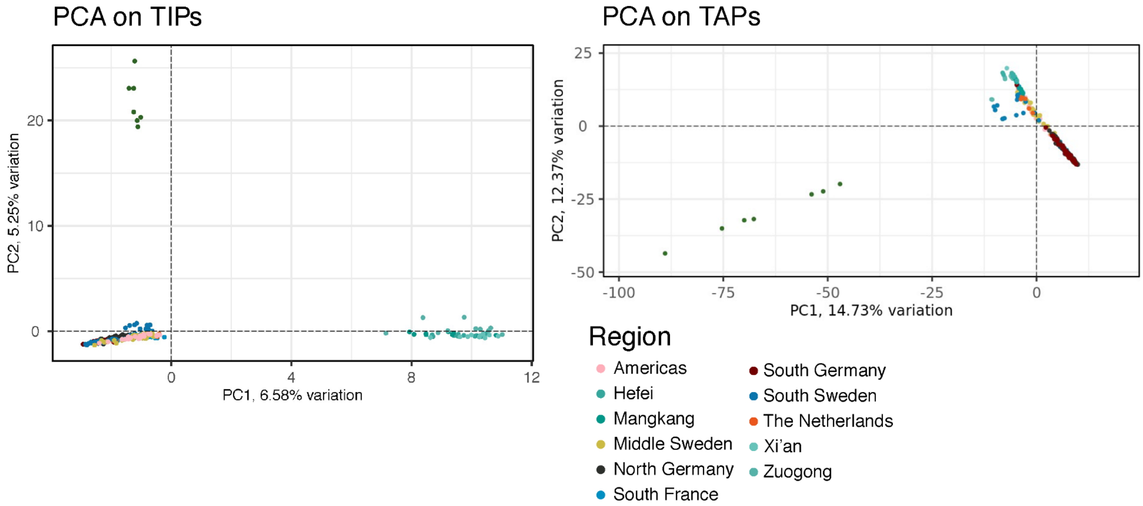


Figure S5. PCA analysis of 279 individuals of *T. arvense*. A presence/absence matrix of either TIPs (left) or TAPs, (right) was used as input to calculate PCA. This results recapitulates the clustering pattern observed with the SNP-PCA.

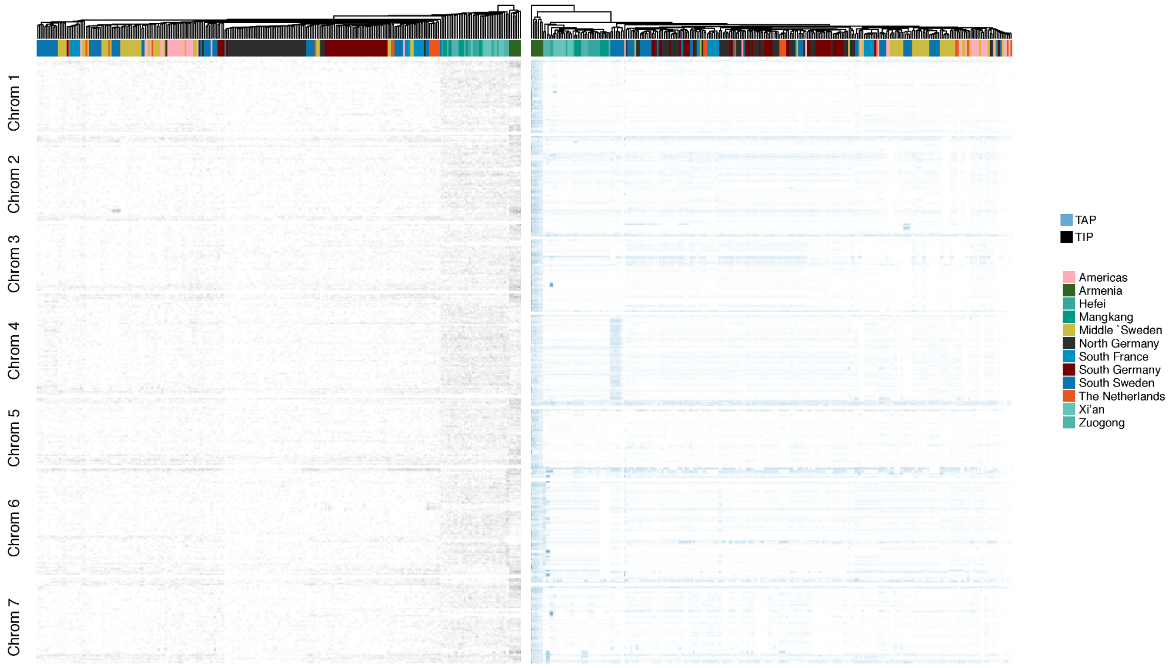
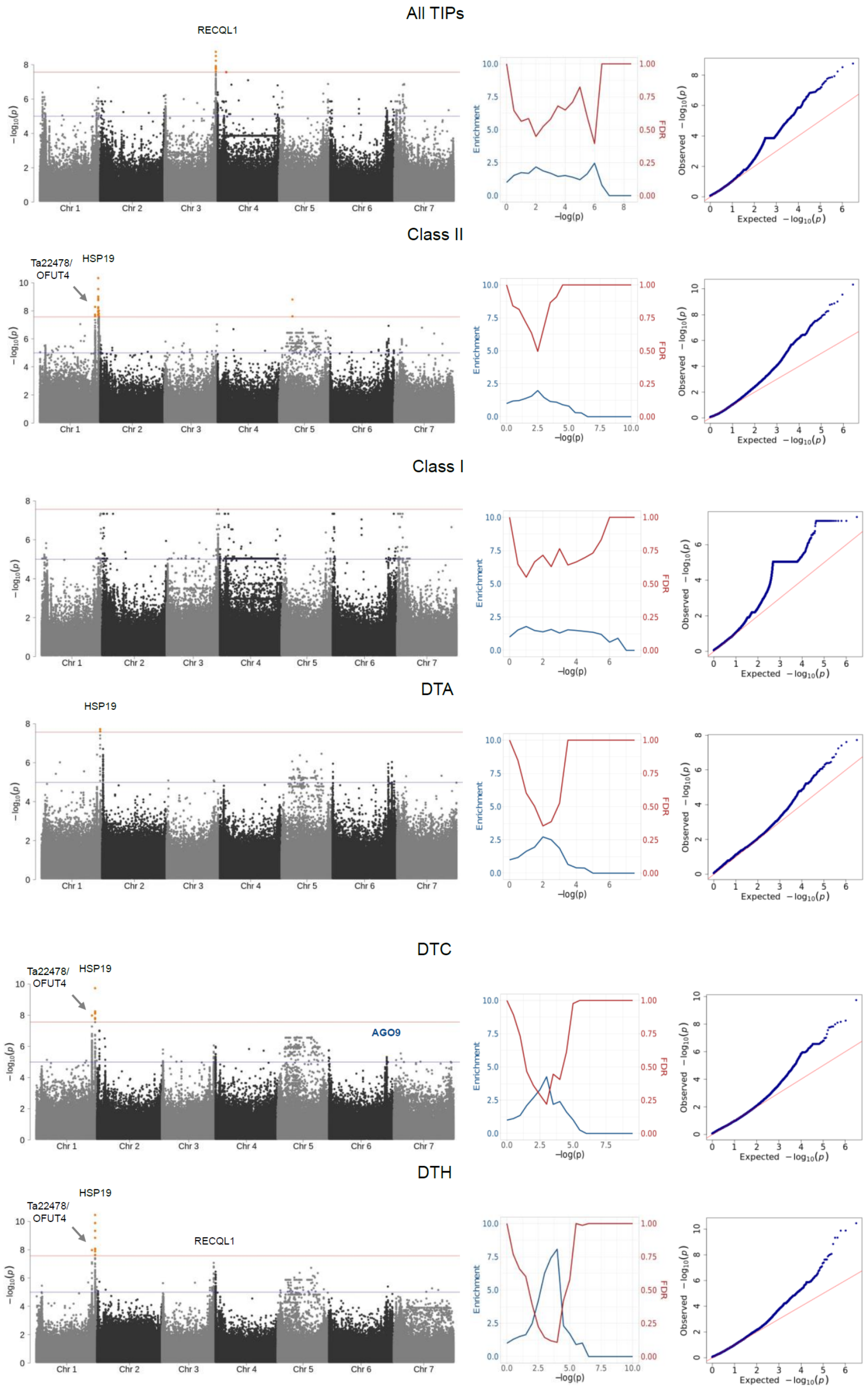


Figure S6. Genomic distribution of TIPs and TAPs along all seven chromosomes of *T. arvense*. Color columns indicate to which biogeographical population each accession belongs to.



All TIPs

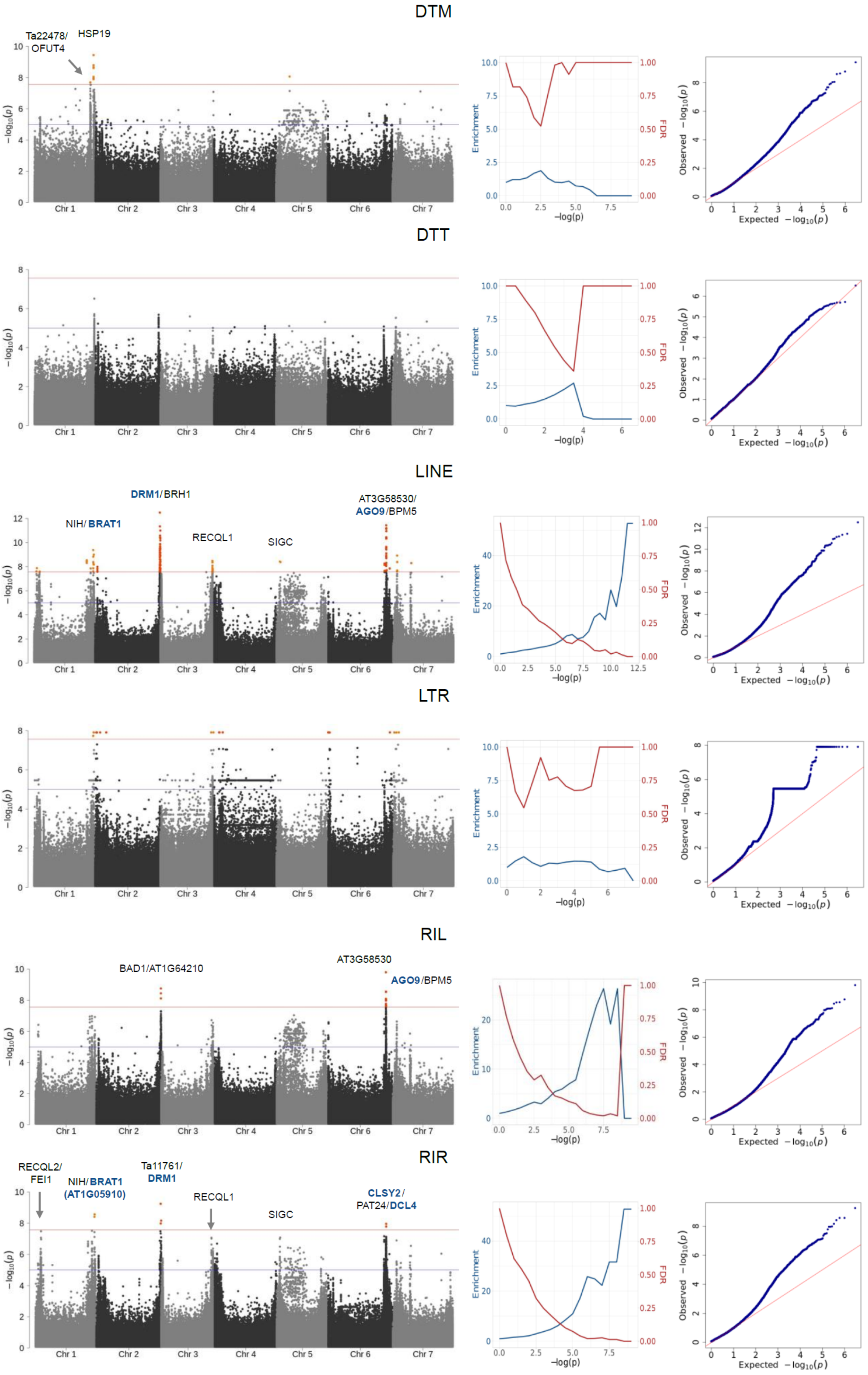
Class II

Class I

DTA

DTC

DTH


DTM
Ta22478/
OFUT4
HSP19
DTT
LINE
DRM1/BRH1
NIH/BRAT1
RECQL1
SIGC
AT3G58530/
AGO9/BPM5
LTR
RIL
BAD1/AT1G64210
AT3G58530
AGO9/BPM5
RIR
RECQL2/
FEI1
NIH/BRAT1
(AT1G05910)
Ta11761/
DRM1
RECQL1
SIGC
CLSY2/
PAT24/DCL4
Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
$-\log_{10}(p)$
Enrichment
FDR
$-\log(p)$
Observed $-\log_{10}(p)$
Expected $-\log_{10}(p)$

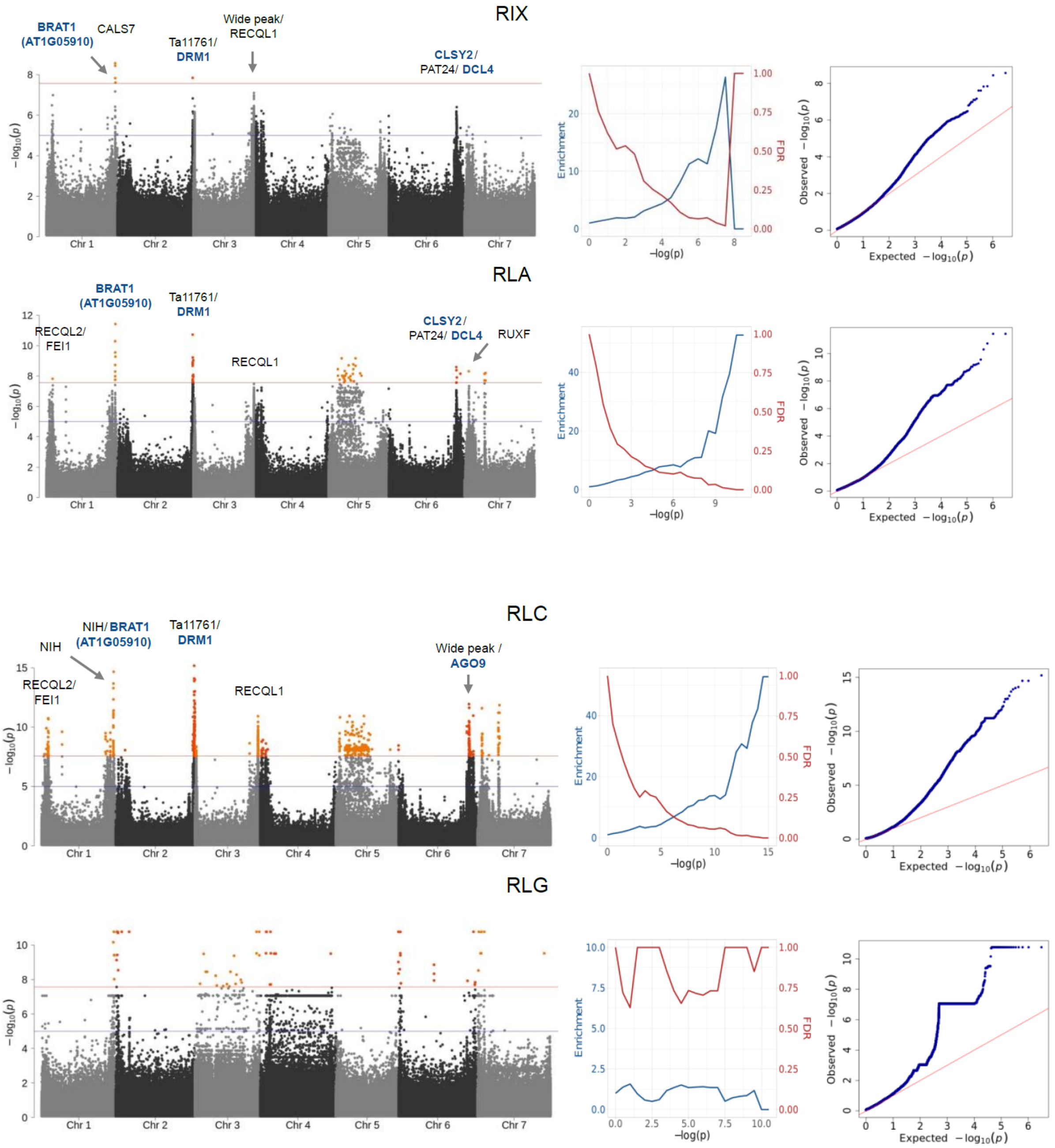


Figure S7. Complete GWA results for TIP load. Left: Manhattan plots for each TIP superfamily load. The genome-wide significance (red line) corresponds to a full Bonferroni correction, the suggestive line (blue) to a more generous hard threshold of –log(p)=5. Genes next to top variants are labeled with 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.

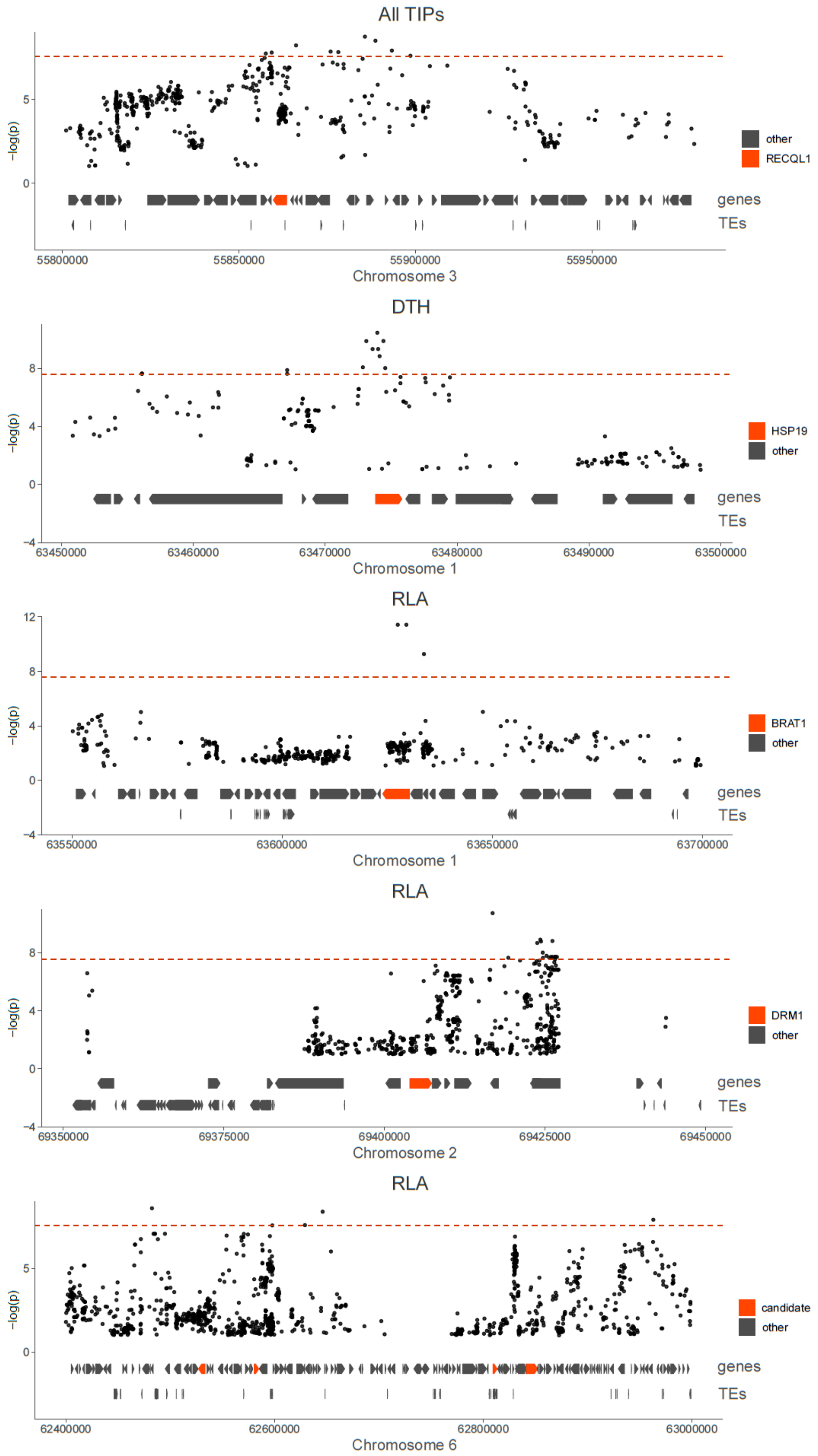

All TIPs
−log(p)
5
0
other
RECQL1
genes
TEs
55800000
55850000
55900000
55950000
Chromosome 3
DTH
−log(p)
8
4
0
−4
HSP19
other
genes
TEs
63450000
63460000
63470000
63480000
63490000
63500000
Chromosome 1
RLA
−log(p)
12
8
4
0
−4
BRAT1
other
genes
TEs
63550000
63600000
63650000
63700000
Chromosome 1
RLA
−log(p)
8
4
0
−4
DRM1
other
genes
TEs
69350000
69375000
69400000
69425000
69450000
Chromosome 2
RLA
−log(p)
5
0
candidate
other
genes
TEs
62400000
62600000
62800000
63000000
Chromosome 6

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.

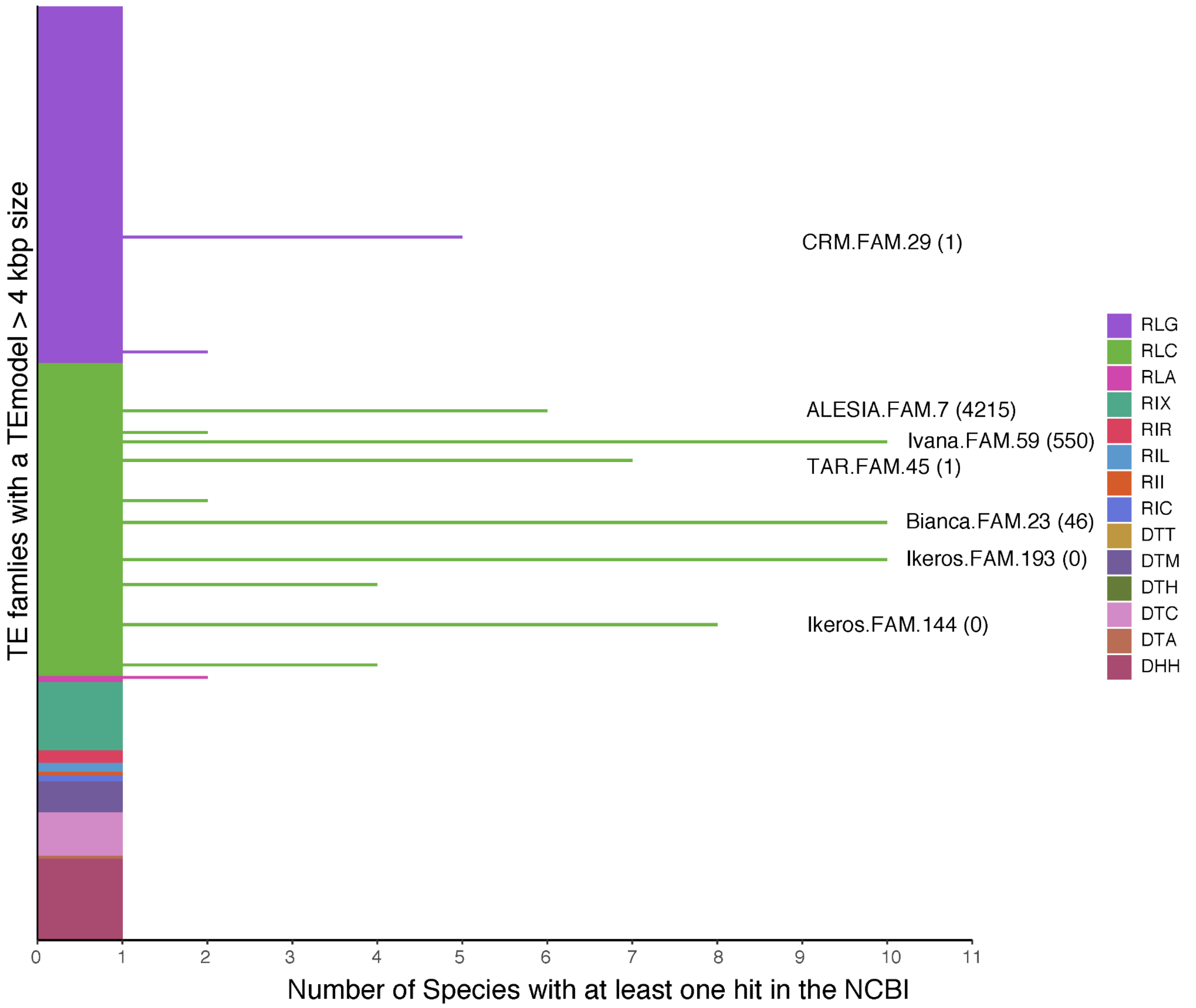


Figure S9. BLASTN hits of *T. arvense* TE families with model sizes > 4 kb against the NCBI NT database (36), **June 2022 release.** We filtered the matches using the 80/80/80 rule, and further constrained matches to fulfill > 2kb length criteria. The x-axis denotes the number of species with at least 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.