Genome - epigenome dynamics and natural variation in *Thlaspi arvense*

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Dario Galanti

aus Milan, Italien

Tübingen

2024

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	04.06.2024
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Oliver Bossdorf
2. Berichterstatter:	Prof. Dr. Detlef Weigel
3. Berichterstatter:	Prof. Dr. Frank Johannes

This work is licensed under a CC BY 4.0 licence
(https://creativecommons.org/licenses/by/4.0/legalcode).

Table of Contents

Declaration of Author Contributions

This thesis entitled "Genome - epigenome dynamics and natural variation in *Thlaspi arvense*" is based on my PhD dissertation work at the University of Tübingen, supervised by Prof. Dr. Oliver Bossdorf. The three main chapters include independent scientific manuscripts, each with co-authors, that have been (or will be) published. The author contributions for each chapter are as follows:

Chapter I: Genetic and environmental drivers of large-scale epigenetic variation in *Thlaspi arvense*

Dario Galanti, Daniela Ramos-Cruz, Adam Nunn, Isaac Rodríguez-Arévalo, J. F. Scheepens, Claude Becker, Oliver Bossdorf.

Published in PLOS Genetics, Volume 18, Issue 10 (October 2022)

https://doi.org/10.1371/journal.pgen.1010452

OB, CB, DG and JFS designed the study. DG collected and analysed the data with contributions from all coauthors. DG and OB wrote the manuscript with revisions from all coauthors.

Chapter II: Transposon dynamics in the oilseed crop *Thlaspi arvense*

*Adrián Contreras-Garrido, ***Dario Galanti**, Andrea Movilli, Claude Becker, Oliver Bossdorf, Hajk-Georg Drost, Detlef Weigel.

* These authors contributed equally

Published in PLOS Genetics, Volume 20, Issue 1 (January 2024)

https://doi.org/10.1371/journal.pgen.1011141

All the authors designed the study. DG and CB collected the data. ACG and DG analysed the data with contributions from all coauthors. ACG and DG wrote the manuscript with revisions from all coauthors.

Chapter III: Discarded sequencing reads uncover natural variation in pest resistance in *Thlaspi arvense*

Dario Galanti, Jun Hee Jung, Caroline Müller, Oliver Bossdorf.

Reviewed Preprint in eLife, selected for publication on 07/02/2024.

https://doi.org/10.7554/eLife.95510.3

JHJ, DG and OB designed the study. DG and CM collected data. DG and JHJ analysed the data. DG, JHJ and OB wrote the manuscript with revisions from all coauthors.

Abstract

Plants play a vital role in sustaining all life on Earth. However, anthropogenic environmental changes are now threatening plant populations. Understanding how plants adapted to environmental changes in the past is key to understanding how they will respond in the future and one way to do this is the study of intraspecific natural variation – the heritable differences occurring in different populations of a species. This natural variation is the result of adaptive processes, and it allows to study phenotypes and their genetic bases, and to discover beneficial alleles with potential applications in breeding. Besides DNA sequence variation, plant populations also harbour extensive epigenetic variation, which can affect phenotypes and be induced by the environment or accumulated in the form of stochastic epimutations. In sexually reproducing species, most but not all heritable epigenetic variation is controlled by genetic variation. In turn, epigenetic changes can affect not only gene expression, but also the silencing of transposable elements, regulating novel genetic variation.

To understand these processes and interactions, I worked with *Thlaspi arvense*, which compared to the model species *Arabidopsis thaliana* has a more complex genome, rich in transposable elements. This species is also a new biofuel and winter cover crop, and breeding efforts to domesticate it are underway.

I surveyed genome-wide genetic and epigenetic natural variation in a large collection of European *T. arvense* accessions, grown in a common environment. I found extensive genetic and epigenetic variation, and that genetic variants at genes involved in the DNA methylation machinery were associated with variable levels of methylation across the whole genome. However, DNA methylation patterns were also associated with climate of origin. Although the genetic variants explained the majority of the observed Differentially Methylated Regions (DMRs), a fraction of DMRs was more strongly associated with environment than DNA sequence, and this fraction was sequence-context dependent, increasing from CG to CHG to CHH.

To understand how short genetic variants and DNA methylation interact with transposable element dynamics, I analysed transposon insertion polymorphisms detected against the reference accession and found several genes associated with the rate of transposition. As DNA methylation is the main mechanism silencing transposons, many of these genes were indeed part of the genomic machinery depositing and maintaining DNA methylation. Nevertheless this was only the case for retrotransposons, whose new insertions became indeed methylated, while DNA transposons, whose new insertions were not methylated, were not associated with DNA methylation machinery genes but with a single gene coding for Heat Shock Protein 19 (*HSP19*). Since this gene is absent in *A. thaliana*, this shows how moving away from classical model species can bring new insights into genome

dynamics. In this work we also investigated the insertion behaviour of different TE families and identified an *Alesia* family preferentially inserting into genes and coding sequences. Since there is growing interest in domesticating *T. arvense*, this TE family could potentially be used to generate new phenotypes of interest.

In the final chapter, I recycled sequencing data not belonging to *T. arvense* to indirectly estimate the abundance of pests and pathogens in my common-environment experiment, and thus natural variation in plant resistance. I found that resistance variation was related to the environment of plant origins, suggesting local adaptation. Moreover, pathogen read-counts allowed me to map genes associated with this resistance variation. Many of these genes were already known to be involved in plant defense and are of potential interest for breeding. Using DNA methylation information I also detected epialleles associated with pathogen presence, located close to genes and transposons.

Altogether my thesis provides evidence of strong and complex interactions between the genome, the epigenome and the environment, which are important to understand adaptive processes and predict the effects of climate change on plant populations. My work also brought insights with potential applications for breeding *T. arvense* as a future crop.

Zusammenfassung

Pflanzen spielen eine wichtige Rolle bei der Erhaltung allen Lebens auf der Erde, doch die vom Menschen verursachten Umweltveränderungen bedrohen die Pflanzenpopulationen. Zu verstehen, wie sich Pflanzen in der Vergangenheit an Umweltveränderungen angepasst haben, ist der Schlüsselfaktor, um vorherzusagen, wie sie in Zukunft reagieren werden. Eine Möglichkeit, dies zu erreichen, ist die Untersuchung der intraspezifischen natürlichen Variation - der vererbbaren Unterschiede, die in verschiedenen Populationen einer Art auftreten. Diese natürliche Variation ist das Ergebnis von Anpassungsprozessen und ermöglicht es, Phänotypen und ihre genetischen Grundlagen zu untersuchen und vorteilhafte Allele zu entdecken, die in der Züchtung eingesetzt werden können. Neben der DNA-Sequenzvariation weisen Pflanzenpopulationen auch eine umfangreiche epigenetische Variation auf, die sich auf den Phänotyp auswirken kann und durch die Umwelt induziert oder in Form von stochastischen Epimutationen akkumuliert wird. Bei sich sexuell fortpflanzenden Arten wird die meiste, aber nicht die gesamte vererbbare epigenetische Variation durch genetische Variation kontrolliert. Epigenetische Veränderungen können sich nicht nur auf die Genexpression auswirken, sondern auch Transposons unterdrücken, wodurch neue genetische Variationen reguliert werden.

Um diese Prozesse und Wechselwirkungen zu verstehen, habe ich mit *Thlaspi arvense* gearbeitet, die im Vergleich zur Modellart *Arabidopsis thaliana* ein komplexeres Genom hat, das reich an Transposons ist. Diese Art ist auch eine neue Biokraftstoff- und Winterzwischenfrucht, und es werden derzeit Züchtungsbemühungen unternommen, um sie zu domestizieren.

Ich untersuchte die genomweite genetische und epigenetische natürliche Variation in einer großen Sammlung von europäischen *T. arvense*-Akzessionen, die in konstanten Umweltbedingungen angebaut wurden. Ich fand eine umfangreiche genetische und epigenetische Variation und stellte fest, dass genetische Varianten in Genen, die an der DNA-Methylierungsmaschinerie beteiligt sind, mit unterschiedlichen Methylierungsgraden im gesamten Genom verbunden sind. Die DNA-Methylierungsmuster waren jedoch auch vom Herkunftsklima abhängig. Obwohl die genetischen Varianten den Großteil der beobachteten Differentially Methylated Regions (DMRs) erklärten, war ein Teil der DMRs stärker mit der Umwelt als mit der DNA-Sequenz assoziiert, und dieser Teil war abhängig vom Sequenzkontext und nahm von CG über CHG bis CHH zu.

Um zu verstehen, wie kurze genetische Varianten und DNA-Methylierung mit der Dynamik von Transposons interagieren, analysierte ich die Transposon-Insertions-Polymorphismen, die im Vergleich zur Referenzakzession festgestellt wurden, und fand mehrere Gene, die mit der Transpositionsrate in Verbindung stehen. Da die DNA-Methylierung der Hauptmechanismus ist, der Transposons

unterdrückt, waren viele dieser Gene tatsächlich Teil der genomischen Maschinerie, die der DNA-Methylierung zugrunde liegt und aufrechterhält. Dies war jedoch nur bei Retrotransposons der Fall, deren neue Insertionen tatsächlich methyliert wurden, während DNA-Transposons, deren neue Insertionen nicht methyliert wurden, nicht mit Genen der DNA-Methylierungsmaschinerie in Verbindung gebracht wurden, sondern mit einem einzigen Gen, das für das Hitzeschockprotein 19 kodiert (*HSP19*). Da dieses Gen in *A. thaliana* nicht vorkommt, zeigt dies, wie die Abkehr von klassischen Modellarten neue Erkenntnisse über die Genomdynamik bringen kann. In dieser Arbeit haben wir auch das Insertionsverhalten verschiedener TE-Familien untersucht und eine Alesia-Familie identifiziert, die bevorzugt in Gene und kodierende Sequenzen insertiert. Da ein wachsendes Interesse an der Domestizierung von *T. arvense* besteht, könnte diese TE-Familie möglicherweise zur Erzeugung neuer, interessanter Phänotypen genutzt werden.

Im letzten Kapitel habe ich Sequenzierungsdaten, die nicht zu *T. arvense* gehören, aus meinem Experiment mit konstanten Umweltbedingungen wiederverwendet, um indirekt die Häufigkeit von Schädlingen und Krankheitserregern und damit die natürliche Variation der Pflanzenresistenz zu schätzen. Ich fand heraus, dass die Variation der Resistenz mit den Umweltbedingungen des Pflanzenursprungs zusammenhing, was auf eine lokale Anpassung hindeutet. Darüber hinaus konnte ich durch Zählen von Pathogenen, Gene kartieren, die mit dieser Resistenzvariation in Verbindung stehen. Von vielen dieser Gene war bereits bekannt, dass sie an der Pflanzenabwehr beteiligt sind, und sie sind von potenziellem Interesse für die Züchtung. Mithilfe von DNA-Methylierungsinformationen konnte ich auch Epiallele nachweisen, die mit der Anwesenheit von Pathogenen in Verbindung stehen und sich in der Nähe von Genen und Transposons befinden.

Insgesamt liefert meine Arbeit Beweise für starke und komplexe Wechselwirkungen zwischen dem Genom, dem Epigenom und der Umwelt, die wichtig sind, um Anpassungsprozesse zu verstehen und die Auswirkungen des Klimawandels auf Pflanzenpopulationen vorherzusagen. Meine Arbeit brachte auch Erkenntnisse für potenziellen Anwendungen zur Zucht von *T. arvense* als zukünftige Nutzpflanze.

Introduction

Plant adaptation and natural variation in a changing world

Plants living in the wild are constantly exposed to biotic stresses such as herbivores and pathogens, and abiotic stresses posed by the climate and the environment. These stresses are usually heterogeneous in space and time, and plants were forced to develop strategies to withstand them. Although this varies greatly depending on species' life histories, plants usually deal with spatial variation through local adaptation and with temporal variation through plastic responses. Local adaptation is the result of the natural selection of alleles advantageous for living under specific conditions, while plasticity is a reversible physiological or developmental response to deal with an environmental change or stress (Williams 2018, Fusco and Minelli 2010). The two aspects are intertwined because plasticity can also be adaptive, as plants with greater plasticity will be selected for in highly variable environments (Laitinen and Nikoloski 2019). Additionally, not all species deal with spatial and temporal variation in the same way. For example, while sexually reproducing species with fast generation time can rely on fast adaptation through recombination and selection of alleles, clonally reproducing species with long generation time rely more on plasticity, which can allow broad environmental distributions without adaptation based on genetic variation. For example, while *Arabidopsis thaliana* has strong population structure resulting from local adaptation to different environments, the *Populus nigra* cv. 'Italica' clone has very low genetic variation but was planted and lives all across Europe (Rodríguez et al. 2022).

Abiotic stressors such as climatic conditions and extreme weather events are spatially structured based on latitude, altitude and other factors. In turn, these abiotic factors affect insects and pathogen communities. For example, aphids are very sensitive to temperature, and fungal pathogens to humidity (Dampc et al 2021, Velasquez et al. 2018, Talley et al. 2002). This implies that climate is a main driver of both biotic and abiotic stressors for plant communities, and considering the currently rapid climate change and the increase of extreme weather events, these stressors are likely to increase in the near

future. Understanding plasticity and adaptation in response to these changes is key to predict their effects on plant populations. One way to achieve this is to study interspecific natural variation, as it is the outcome that evolutionary processes had on a plant species exposed to different conditions, over thousands of years. This makes natural variation a great tool to study local adaptation but also to discover genes and alleles that are beneficial in specific environments, including essential genes whose function cannot be uncovered through knock-out experiments. Overall, understanding natural variation and adaptation can help (i) to predict and mitigate the effects of climate change through informed species conservation practices, (ii) to unravel the genetic bases of traits and discover natural advantageous alleles and consequently (iii) to leverage this information for breeding more robust and resilient crops.

Epigenetic regulation of the genome

According to the "modern synthesis", the evolutionary framework developed by population geneticists between the 1920s and 1940s, and further refined in the subsequent decades (Fisher 1919; Huxley 1943; Provine 1978), evolution is the result of natural selection acting on randomly generated genetic variation. Although this basic principle still holds true, the discovery of epigenetic mechanisms and recent studies on their heritability and phenotypic effects suggest a more complex picture where genome evolution is more interactive with the environment. Most simplistically, in addition to the information coded in its sequence, DNA harbours epigenetic marks that can affect chromatin compaction and gene expression, can be acquired in response to environmental cues, and can be heritable (Kinoshita and Seki 2014, Lämke and Bäurle 2017). Additionally, epigenetic mechanisms are very important for silencing transposable elements (TEs), major components of eukaryotic genomes with the ability to self-replicate and translocate to new genomic locations. Due to their environmental responsiveness and heritability, epigenetic mechanisms have been proposed as a means of rapid evolution (Bossdorf et al. 2008; Richards et al. 2010).

The currently most studied epigenetic mark is DNA methylation. It refers to the addition of a methyl group to the 5th, or more rarely 4th, atom of the cytosine ring, and is involved in silencing TEs and regulating gene expression. Other types of base modifications exist, including for example adenine methylation, but their functions are less clear, so we will not discuss them here (Liang et al. 2018). Cytosine methylation is usually a repressive mark on regulatory regions and transposons, which are silenced by heavy methylation in all contexts (TE-like), but GC methylation alone is often also present in gene bodies of active genes (gbM) (Niederhuth et al. 2016, Schmitz et al. 2019). Histone modifications include several chemical modifications of the amino acids present in the histone tails, such as methylation or acetylation, and depending on the modification can be repressive or permissive for gene expression. Both DNA methylation and histone modifications are involved in plastic responses to stresses and can contribute to somatic stress memory, i.e. the ability of an individual to better stand successive stresses of the same kind (Lämke and Bäurle 2017, He and Li 2018). For example DNA methylation changes associated with somatic stress memory were found for hyperosmotic stress, phosphate starvation, bacterial infections and herbivory (Secco et al. 2015, He and Li 2018, Lämke and Bäurle 2017). Even more stresses were associated with histone mark modifications (He and Li 2018, Lämke and Bäurle 2017). Although there is evidence that DNA methylation can be inherited through sexual reproduction, this is not as clear for histone modifications (He and Li 2018), i.e. so far the former seems to have greater evolutionary potential.

In plants, DNA methylation can occur in the three sequence contexts: CG, CHG and CHH (where H is A, T or C). Distinguishing between these contexts is important because they differ in the molecular machineries for depositing and maintaining methylation (Law and Jacobsen 2010, Zhang et al. 2018), which has consequences for their dynamics and stability (Fig 1). In *Arabidopsis thaliana*, CG methylation is maintained in a copy-paste manner during replication by *DMT1*, which recognizes hemi-methylated sites in the daughter strand and re-establishes full methylation. CHG methylation is largely maintained by *CMT3-SUVH4* self-reinforcing loops between DNA and histone methylation, in which *CMT3* binds histone 3 lysine 9 dimethylation (H3K9me2), depositing CHG methylation, and in turn

SUVH4 (and its paralogs *SUVH5* and *6*) binds CHG methylation depositing H3K9me2. Finally, CHH methylation is deposited *de-novo* by *DRM2,* through the RNA-directed DNA methylation pathway (RdDM), and by *CMT2* at some loci, mostly in histone H1 - containing heterochromatin (Law and Jacobsen 2010, Zhang et al. 2018). CHG and CHH methylation partially share maintenance pathways, as *CMT2* can methylate in CHG (Stroud et al. 2014) and *CMT3* also affects CHH methylation at few loci (Cao et al. 2003). Altogether, there is a gradient of similarity and decreasing stability from CG to CHG to CHH. Although least stable, CHH is the most abundant context and often the most responsive to stresses (Liu and He 2020).

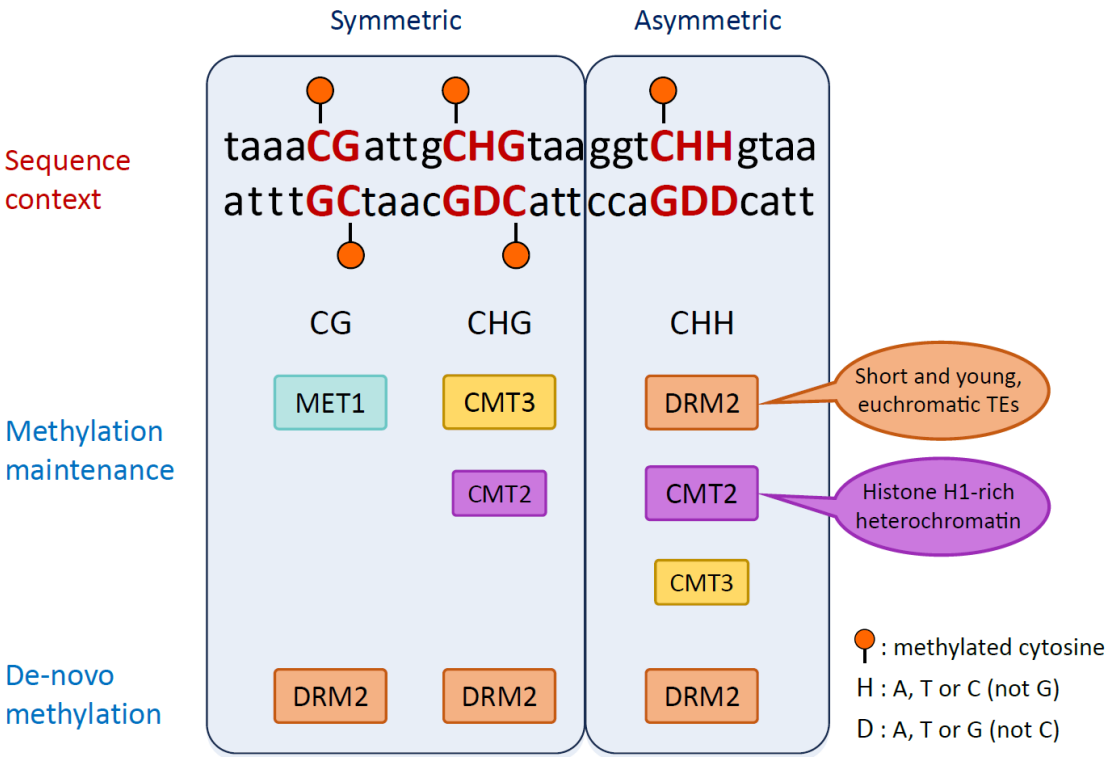


Fig 1. DNA methylation deposition and maintenance in different sequence contexts. Sketch of the molecular pathways and genes depositing and maintaining DNA methylation in different sequence contexts in *A. thaliana*, inspired by Ramos-Cruz 2023. *De-novo* methylation is deposited in all contexts by DRM2 in the RNA directed DNA methylation pathway (RdDM). CG methylation is copy pasted on the daughter strand during replication by *MET1*. CHG methylation is maintained by *CMT3* and only partially by *CMT2*, which work in a feedback-loop manner recognising H3K9me2 deposited by *SUVH4*, *5* and *6*. CHH methylation is maintained by recurrent de-novo deposition via RdDM and *DRM2* on short euchromatic TEs and by *CMT2* on histone H1 - rich heterochromatin and the bodies of long TEs. *CMT3* is also responsible for CHH methylation at few loci (Cao et al. 2003).

Given all this, one should expect that DNA methylation in less stable and more stress-responsive contexts (CHH) might contribute most to plasticity, while the more stable CG methylation might be inherited for longer and more involved in adaptation, with CHG methylation laying somewhere in between. But most importantly, methylation can only be adaptive if it is inherited and not controlled by genetic variation.

Genome-epigenome interactions and their evolutionary consequences

Given the complex molecular machinery for maintaining and regulating DNA methylation, it is not surprising that previous studies have found various kinds of genetic control over DNA methylation. Genetic polymorphisms can control DNA methylation in *cis*, for example new TE insertions are usually methylated also on their flanks and can silence genes by methylating their promoters (Martin et al. 2009), or in *trans*, when genetic mutations affect genes involved in the DNA methylation machinery, resulting in methylation changes across the entire genome (Dubin et al. 2015, Kawakatsu et al. 2016, Sasaki et al. 2019). In such situations it might seem as if a methylation variant is inherited across generations, but in fact it is simply re-established in each generation by the underlying genetic variant. When either variant has a phenotypic effect, the ultimate causal one, that could be selected for, would still be the genetic variant.

There is not only genetic control of epigenetic variation, but also reverse causal relationships where epigenetic marks affect genetic variation, for example by controlling transposition. Transposons are major drivers of genome evolution, but if not properly controlled they can easily disrupt the function of genes and cause high mortality. Because of this, they are usually heavily methylated and targeted by repressive histone marks. TEs are widely distributed among eukaryotes and can make up large fractions of plant genomes, up to more than 80% in extreme cases such as maize and barley (Wells and Feschotte 2020). By replicating and decaying at a neutral fashion over evolutionary time, TE sequences

became highly diverse, and they have been classified into a complex taxonomic system, which in plants includes thousands of families separated into two classes based on their replication intermediate (Wicker et al. 2007). Retrotransposons (Class I) replicate through an RNA intermediate that is then reverse transcribed into complementary DNA and inserted into the target site, in a copy-and-paste fashion. DNA transposons (Class II) replicate through a DNA intermediate which is usually (Subclass 1) cleaved from the donor site and inserted into the target site in a cut-and-paste fashion, but can also be copy-and-pasted when a single strand is cleaved and repaired at the donor site (Subclass 2) (Fig 2) (Wicker et al. 2007, Wells and Feschotte 2020). Classes and subclasses are further subdivided into orders, based on the exact replication mechanism, and superfamilies based on the protein domains coded in the TE structure and the presence and length of the Target Site Duplication (TSD), a repeat generated on both sides of the transposon upon insertion. Finally, TEs in the same superfamily are split into different families, defined as groups of TEs sharing at least 80% sequence identify in at least 80% of their sequence (Wicker et al. 2007). In eukaryotes and particularly plants, the most abundant retrotransposons contain Long Terminal Repeats (LTRs) on their extremities (Wang et al. 2021) and share high sequence similarity with retroviruses, suggesting a common evolutionary history (Vázquez et al. 2000). The most widespread eukaryotic DNA TEs contain Terminal Inverted Repeats (TIRs) at their extremities and are transposed via the DDE DNA transposase coded in their own sequence (Feschotte and Pritham 2007). For more details on TE replication mechanisms and classification refer to Wicker et al. (2007) and Wells and Feschotte (2020).

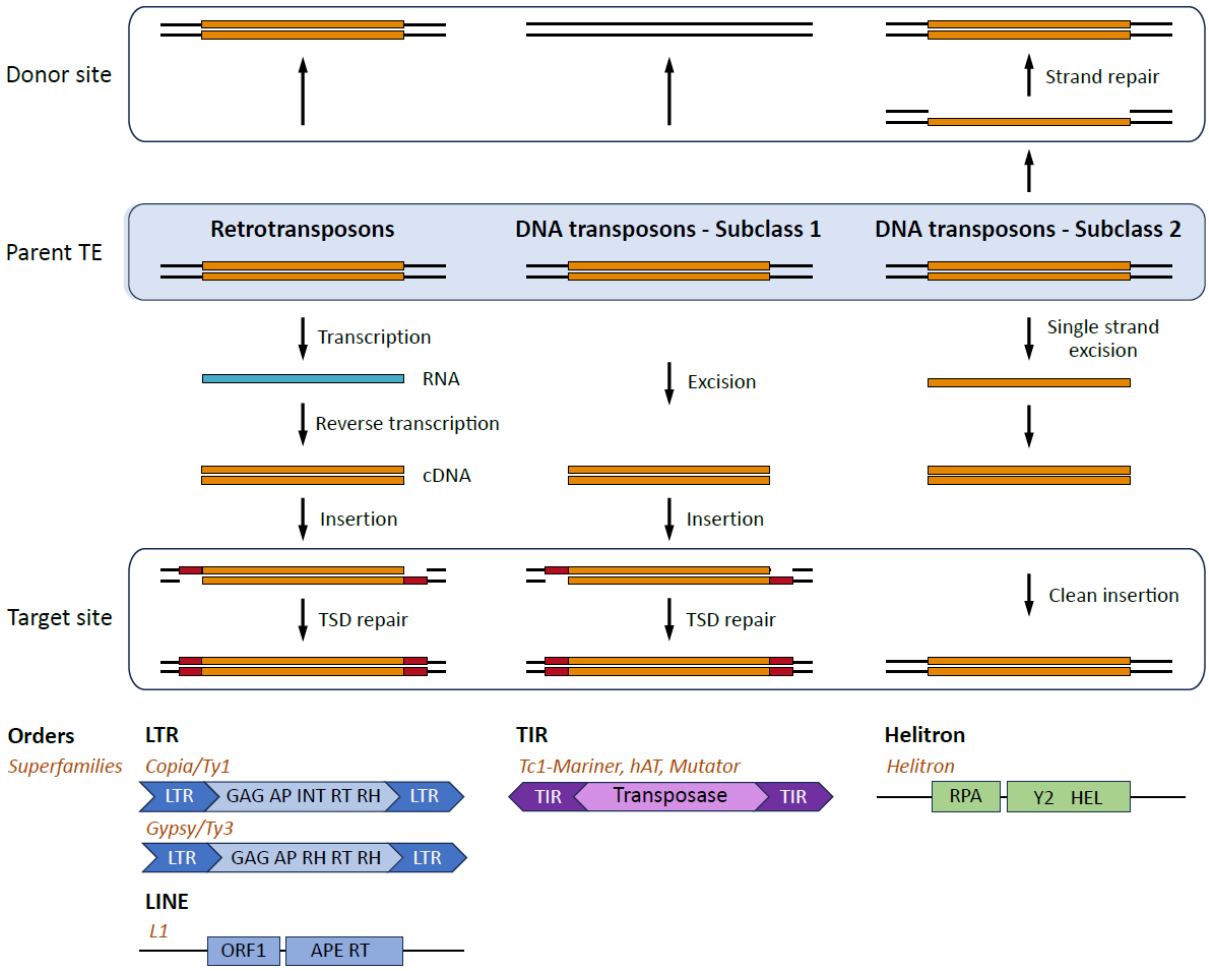


Fig 2. Sketch of transposition mechanisms and structure of different transposon classes and subclasses. (Top) Transposition mechanisms: Retrotransposons (Class I) are transposed in a copy-and-paste fashion, while DNA transposons (Class II) usually adopt a cut-and-paste mechanism, but in some cases (*Helitrons*) can be copy-and-pasted (Wicker et al. 2007, Wells and Feschotte 2020). In most TE superfamilies the DNA cleavage before insertion generates sticky ends, which are then repaired upon insertion, generating the Target Site Duplication (TSD) (in red). (Bottom) Structure of the most common TE superfamilies found in plants, for each class/subclass above. Structural repeats: Long Terminal Repeats (LTRs) and Terminal Inverted Repeats (LTRs). Long Interspersed Nuclear Element (LINE). Protein coding domains: Capsid protein (GAG), Aspartic proteinase (AP), Integrase (INT), Reverse transcriptase (RT), RNA Helicase (RH), Apurinic endonuclease (APE), Replication protein A (RPA, plant specific), Tyrosine recombinase with YY motif (Y2) and Helicase (HEL).

As mentioned before, controlling TE mobility is extremely important for organisms, because TEs can quickly generate new variation, but they can also disrupt vital genes. Transposon insertions were indeed shown to be generally more deleterious than point mutations as they are on average less likely to spread in a population (Baduel et al. 2021). Nevertheless, the discovery of TE families that are stress-

responsive and show target site preferences (Gao et al. 2012, Quadrana et al. 2016, Pietzenuk et al. 2016, Roquis et al. 2021) indicates that they are important for plasticity and adaptation (Li et al. 2018). Stress-induced TE reactivation is usually coupled with DNA demethylation and transitioning to active chromatin states (Roquis et al. 2021, Annacondia et al. 2021), confirming the importance of epigenetic mechanisms in controlling TE activity.

In addition to controlling transposition, recent studies further highlighted the role of epigenetic marks in regulating genetic mutation rates. Methylated DNA usually has a higher mutation rate, due to 5mC being more prone to C to T transitions by deamination (Pfeifer 2006), but for example 5-hydroxymethylcytosine (5hmC), although rare in plant genomes (Erdmann 2015), is less efficiently deaminated and shows reduced mutation rate (Zilberman 2016, Tomkova and Schuster-Bockler 2018). Additionally, mutation rate is affected by several other factors such as CG content, which stabilizes the DNA double helix, and several other epigenetic marks (Monroe et al. 2022). In particular the histone mark H3K4me1 was the strongest predictor of the mutation bias observed in gene bodies of essential genes in *A. thaliana* (Monroe et al. 2022, Monroe et al. 2023). Monroe and colleagues (Monroe et al. 2022) estimated that epigenetic marks could predict up to 90% of the mutation bias observed across gene bodies. Although this result might partially reflect biases induced by sequencing errors (Wang et al 2023, Monroe et al. 2023), it still indicates that epigenetic modifications might be directly linked to mutation rates.

Altogether, there is clearly reciprocal causation and complex interactions between the genome and the epigenome, resulting in a complex system for genome evolution that is more than mere selection on randomly generated DNA sequence variation, and involves a genetics-epigenetics crosstalk interacting with the environment.

Natural genetic and epigenetic variation

Understanding mechanisms of genome-epigenome interactions is key to predicting evolutionary dynamics, but it is equally important to observe the effects of these mechanisms in real life and study natural populations in the wild. In the last 20 years, several studies unveiled extensive DNA methylation variation both within and between plant populations of several species. This variation was often associated with climate of origin (Vaughn et al. 2007, Lira-Medeiros et al. 2010, Paun et al 2010, Gugger et al. 2016, Gaspar et al. 2018) and also with phenotypes (Cubas et al. 1999, Cortijo et al. 2014, Kooke et al. 2015), sometimes independently of genetic variation (Vaughn et al. 2007, Cubas et al. 1999, Cortijo et al. 2014). Additionally, natural epigenetic variation was shown to arise (i) stochastically through spontaneous epimutations, at higher rates than genetic mutations (Becker et al. 2011; Johannes and Schmitz 2018), or (ii) through environmental induction (Chinnusamy and Zhu 2009, Lämke and Bäurle 2017). The combination of these findings raised the hypothesis that transgenerational epigenetic variation may provide plants with an additional means of rapid adaptive evolution (Bossdorf et al. 2008; Richards et al. 2010). However, the ecological and evolutionary significance of this epigenetic variation remains poorly understood as very large and high-resolution datasets are required to show that natural epigenetic variation can indeed create heritable phenotypes independently of DNA sequence variation (Richards et al. 2017). This is mainly due to the necessity of combining large-scale surveys with high-accuracy at the genomic level. Large-scale surveys are fundamental to capture a large enough variation to be representative of the adaptive ability of a species and to have enough statistical power. High-accuracy is required at the DNA methylation level in order to capture at least most differential methylation in the genome and at the DNA sequence level, due to the tight link between genetics and epigenetics. Such datasets led to major breakthroughs, but so far they could only be generated for the model plant *A. thaliana* (Dubin et al. 2015; Kawakatsu et al. 2016), which has a small genome with unusually low TE content and DNA methylation compared to most plants (Alonso et al. 2015). Further data from species with different life histories, ploidy, reproduction modes and genomic complexity are fundamental to generalise these

findings, and it can now finally be generated, thanks to technological advances and decreasing sequencing costs (Richards et al. 2017).

Thlaspi arvense as a new model species and crop

The need to study epi-genomic natural variation in complex and diverse plant species prompted the EpiDiverse European Training Network (https://epidiverse.eu/), the larger project within which this work has been embedded. EpiDiverse included several sub-projects across three different plant species, the sexually and clonally reproducing *Fragaria vesca*, the clone *Populus nigra* cv. *'italica'* and the annual crucifer *Thlaspi arvense* - the study organism of my thesis.

Thlaspi arvense (field pennycress) is an annual herbaceous plant, native to Eurasia and widely distributed in the temperate regions of Europe, Asia and the Americas (Warwick et al. 2002). It is increasingly studied both as a model species (Geng et al. 2021, Nunn et al 2021, Hu et al. 2022, Troyee et al. 2022, Galanti et al. 2022), and as a potential biofuel and cover crop (Dorn et al. 2015, Frels et al. 2019, Chopra et al. 2019, Zhao et al. 2021). As a model, it has the advantage of being closely related to *A. thaliana*, which allows to leverage many available genomic resources such as functional gene ontology. Additionally, it reproduces sexually by selfing, with low outcrossing rates of 3-10% (Best and McIntyre 1976), which leads to reduced heterozygosity. Despite this similarity to *A. thaliana* in terms of phylogeny and reproduction mode, *T. arvense* harbours a much larger portion of TEs and a higher global DNA methylation (Nunn et al. 2022), making it an interesting model for studying a more complex and TE-rich genome, properties that are very common in the plant kingdom, using highly accurate genomic tools.

In addition to its potential as a model species, *T. arvense* is a rising crop, as it is being domesticated to be used for biodiesel production. The seeds contain high amounts (up to 40%) of oil, and they also have high nutritional value for livestock. Moreover, winter-annual accessions overwinter as rosettes

and are highly cold tolerant, providing soil cover during winter that can reduce erosion and nutrient run-off (Moser 2012, Dorn et al 2015). However, several traits including reducing seed dormancy and erucic acid seed content still need improvement before pennycress can be effectively used as a crop. Therefore finding natural alleles influencing any of these crop-traits, but of course also climate tolerance and pest resistance, will be highly useful for breeding.

Aims of my thesis

To study genome-epigenome-environment relationships at large-scale and high-resolution in *Thlaspi arvense*, I collected seeds of 207 lines from 36 populations on a latitudinal gradient in Europe, grew all lines in a common environment and generated Whole Genome Sequencing (WGS) and Bisulfite Sequencing (WGBS) libraries for all. With this dataset, I could then investigate several of the knowledge gaps highlighted above. My work is divided into three main research chapters:

- In **Chapter I**, I analysed natural genetic and epigenetic variation across Europe. I used the 207 wild accessions to understand how genetic background and environment of origin influence DNA methylation patterns across the *Thlaspi arvense* genome.

- In **Chapter II**, I investigated, together with my colleague Adrian Contreras, transposon polymorphisms in a total of 280 *T. arvense* lines, including 73 lines from America, China and Armenia published elsewhere. I described the diversity and abundance of transposons in *T. arvense*, investigated their expansion history and insertion behaviour, and the genetic bases of different transposition mechanisms by means of Genome Wide Association.

- In **Chapter III** I followed up on an idea from my colleague Jun Hee Jung to quantify aphid, mildew and microbe colonisation of *T. arvense* during the common-environment experiment described above, based on WGS reads that did not map to *T. arvense*. We used these data to

estimate variation in pest resistance among the *T. arvense* lines, and to discover associations with climate of origin and with genetic and DNA methylation variants.

References

Alonso, Conchita, Ricardo Pérez, Pilar Bazaga, and Carlos M. Herrera. 2015. "Global DNA Cytosine Methylation as an Evolving Trait: Phylogenetic Signal and Correlated Evolution with Genome Size in Angiosperms." Frontiers in Genetics 6. https://doi.org/10.3389/fgene.2015.00004.

Annacondia, Maria Luz, Dimitrije Markovic, Juan Luis Reig-Valiente, Vassilis Scaltsoyiannes, Corné M. J. Pieterse, Velemir Ninkovic, R. Keith Slotkin, and German Martinez. 2021. "Aphid Feeding Induces the Relaxation of Epigenetic Control and the Associated Regulation of the Defense Response in Arabidopsis." New Phytologist 230 (3): 1185–1200. https://doi.org/10.1111/nph.17226.

Baduel, Pierre, Basile Leduque, Amandine Ignace, Isabelle Gy, José Gil, Olivier Loudet, Vincent Colot, and Leandro Quadrana. 2021. "Genetic and Environmental Modulation of Transposition Shapes the Evolutionary Potential of Arabidopsis thaliana." Genome Biology 22 (1): 138. https://doi.org/10.1186/s13059-021-02348-5.

Becker, Claude, Jörg Hagmann, Jonas Müller, Daniel Koenig, Oliver Stegle, Karsten Borgwardt, and Detlef Weigel. 2011. "Spontaneous Epigenetic Variation in the Arabidopsis thaliana Methylome." Nature 480 (7376): 245–49. https://doi.org/10.1038/nature10555.

Best, K. F., and G. I. Mcintyre. 1975. "THE BIOLOGY OF CANADIAN WEEDS: 9. Thlaspi Arvense L." Canadian Journal of Plant Science 55 (1): 279–92. https://doi.org/10.4141/cjps75-039.

Bossdorf, Oliver, Christina Richards, and Massimo Pigliucci. 2008. "Epigenetics for Ecologists." Ecology Letters 11 (2): 106–15. https://doi.org/10.1111/j.1461-0248.2007.01130.x.

Cao, Jun, Korbinian Schneeberger, Stephan Ossowski, Torsten Günther, Sebastian Bender, Joffrey Fitz, Daniel Koenig, et al. 2011. "Whole-Genome Sequencing of Multiple Arabidopsis thaliana Populations." Nature Genetics 43 (10): 956–63. https://doi.org/10.1038/ng.911.

Chinnusamy, Viswanathan, and Jian-Kang Zhu. 2009. "Epigenetic Regulation of Stress Responses in Plants." Current Opinion in Plant Biology, Genome Studies and Molecular Genetics, 12 (2): 133–39. https://doi.org/10.1016/j.pbi.2008.12.006.

Chopra, Ratan, Nicole Folstad, Joseph Lyons, Tim Ulmasov, Cynthia Gallaher, Liam Sullivan, Abby McGovern, et al. 2019. "The Adaptable Use of Brassica NIRS Calibration Equations to Identify Pennycress Variants to Facilitate the Rapid Domestication of a New Winter Oilseed Crop." Industrial Crops and Products 128 (February): 55–61. https://doi.org/10.1016/j.indcrop.2018.10.079.

Cortijo, Sandra, René Wardenaar, Maria Colomé-Tatché, Arthur Gilly, Mathilde Etcheverry, Karine Labadie, Erwann Caillieux, et al. 2014. "Mapping the Epigenetic Basis of Complex Traits." Science 343 (6175): 1145–48. https://doi.org/10.1126/science.1248127.

Cubas, Pilar, Coral Vincent, and Enrico Coen. 1999. "An Epigenetic Mutation Responsible for Natural Variation in Floral Symmetry." Nature 401 (6749): 157–61. https://doi.org/10.1038/43657.

Dampc, Jan, Mateusz Mołoń, Tomasz Durak, and Roma Durak. 2021. "Changes in Aphid—Plant Interactions under Increased Temperature." Biology 10 (6): 480. https://doi.org/10.3390/biology10060480.

Dorn, Kevin M., Johnathon D. Fankhauser, Donald L. Wyse, and M. David Marks. 2015. "A Draft Genome of Field Pennycress (Thlaspi Arvense) Provides Tools for the Domestication of a New Winter Biofuel Crop." DNA Research 22 (2): 121–31. https://doi.org/10.1093/dnares/dsu045.

Dubin, Manu J, Pei Zhang, Dazhe Meng, Marie-Stanislas Remigereau, Edward J Osborne, Francesco Paolo Casale, Philipp Drewe, et al. 2015. "DNA Methylation in Arabidopsis Has a Genetic Basis and Shows Evidence of Local Adaptation." eLife 4. https://doi.org/10.7554/eLife.05255.

Erdmann, Robert M, Amanda L Souza, Clary B Clish, and Mary Gehring. 2015. "5-Hydroxymethylcytosine Is Not Present in Appreciable Quantities in Arabidopsis DNA." G3 Genes|Genomes|Genetics 5 (1): 1–8. https://doi.org/10.1534/g3.114.014670.

Feschotte, Cédric, and Ellen J. Pritham. 2007. "DNA Transposons and the Evolution of Eukaryotic Genomes." Annual Review of Genetics 41 (Volume 41, 2007): 331–68. *https://doi.org/10.1146/annurev.genet.40.110405.090448*.

Fisher, R. A. 1919. The Correlation between Relatives on the Supposition of Mendelian Inheritance. Earth and Environmental Science Transactions of the Royal Society of Edinburgh. 1919; 52 (02): 399–433.

Frels, Katherine, Ratan Chopra, Kevin M. Dorn, Donald L. Wyse, M. David Marks, and James A. Anderson. 2019. "Genetic Diversity of Field Pennycress (Thlaspi Arvense) Reveals Untapped Variability and Paths Toward Selection for Domestication." Agronomy 9 (6): 302. https://doi.org/10.3390/agronomy9060302.

Fusco, Giuseppe, and Alessandro Minelli. 2010. "Phenotypic Plasticity in Development and Evolution: Facts and Concepts." Philosophical Transactions of the Royal Society B: Biological Sciences 365 (1540): 547–56. https://doi.org/10.1098/rstb.2009.0267.

Gao, Dongying, Jinfeng Chen, Mingsheng Chen, Blake C. Meyers, and Scott Jackson. 2012. "A Highly Conserved, Small LTR Retrotransposon That Preferentially Targets Genes in Grass Genomes." PLOS ONE 7 (2): e32010. https://doi.org/10.1371/journal.pone.0032010.

Gáspár, Bence, Oliver Bossdorf, and Walter Durka. 2018. "Structure, Stability and Ecological Significance of Natural Epigenetic Variation: A Large-Scale Survey in Plantago Lanceolata." New Phytologist 0 (0). https://doi.org/10.1111/nph.15487.

Geng, Yupeng, Yabin Guan, La Qiong, Shugang Lu, Miao An, M. James C. Crabbe, Ji Qi, Fangqing Zhao, Qin Qiao, and Ticao Zhang. 2021. "Genomic Analysis of Field Pennycress (Thlaspi Arvense) Provides Insights into Mechanisms of Adaptation to High Elevation." BMC Biology 19 (1): 143. https://doi.org/10.1186/s12915-021-01079-0.

Gugger, Paul F., Sorel Fitz-Gibbon, Matteo Pellegrini, and Victoria L. Sork. 2016. "Species-Wide Patterns of DNA Methylation Variation in Quercus Lobata and Their Association with Climate Gradients." Molecular Ecology 25 (8): 1665–80. https://doi.org/10.1111/mec.13563.

He, Yuehui, and Zicong Li. 2018. "Epigenetic Environmental Memories in Plants: Establishment, Maintenance, and Reprogramming." Trends in Genetics, August. https://doi.org/10.1016/j.tig.2018.07.006.

Hu, Yanting, Xiaopei Wu, Guihua Jin, Junchu Peng, Rong Leng, Ling Li, Daping Gui, Chuanzhu Fan, and Chengjun Zhang. 2022. "Rapid Genome Evolution and Adaptation of Thlaspi Arvense Mediated by Recurrent RNA-Based and Tandem Gene Duplications." Frontiers in Plant Science 12. https://www.frontiersin.org/articles/10.3389/fpls.2021.772655.

Huxley, Julian Sorell. 1943. "Evolution, the Modern Synthesis." London Georg Allen Unwin.

Johannes, Frank, and Robert J. Schmitz. 2018. "Spontaneous Epimutations in Plants." New Phytologist, September. https://doi.org/10.1111/nph.15434.

Kawakatsu, Taiji, Shao-shan Carol Huang, Florian Jupe, Eriko Sasaki, Robert J. Schmitz, Mark A. Urich, Rosa Castanon, et al. 2016. "Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions." Cell 166 (2): 492–505. https://doi.org/10.1016/j.cell.2016.06.044.

Kinoshita, Tetsu, and Motoaki Seki. 2014. "Epigenetic Memory for Stress Response and Adaptation in Plants." Plant and Cell Physiology 55 (11): 1859–63. https://doi.org/10.1093/pcp/pcu125.

Kooke, Rik, Frank Johannes, René Wardenaar, Frank Becker, Mathilde Etcheverry, Vincent Colot, Dick Vreugdenhil, and Joost J. B. Keurentjes. 2015. "Epigenetic Basis of Morphological Variation and Phenotypic Plasticity in Arabidopsis thaliana." The Plant Cell, February, tpc.114.133025. https://doi.org/10.1105/tpc.114.133025.

Laitinen, Roosa A E, and Zoran Nikoloski. 2019. "Genetic Basis of Plasticity in Plants." Journal of Experimental Botany 70 (3): 739–45. https://doi.org/10.1093/jxb/ery404.

Lämke, Jörn, and Isabel Bäurle. 2017. "Epigenetic and Chromatin-Based Mechanisms in Environmental Stress Adaptation and Stress Memory in Plants." Genome Biology 18 (1): 124. https://doi.org/10.1186/s13059-017-1263-6.

Law, Julie A., and Steven E. Jacobsen. 2010. "Establishing, Maintaining and Modifying DNA Methylation Patterns in Plants and Animals." Nature Reviews Genetics 11 (3): 204–20. https://doi.org/10.1038/nrg2719.

Li, Zi-Wen, Xing-Hui Hou, Jia-Fu Chen, Yong-Chao Xu, Qiong Wu, Josefa González, and Ya-Long Guo. 2018. "Transposable Elements Contribute to the Adaptation of Arabidopsis thaliana." Genome Biology and Evolution 10 (8): 2140–50. https://doi.org/10.1093/gbe/evy171.

Liang, Zhe, Lisha Shen, Xuean Cui, Shengjie Bao, Yuke Geng, Guoliang Yu, Fan Liang, et al. 2018. "DNA N6-Adenine Methylation in Arabidopsis thaliana." Developmental Cell 45 (3): 406-416.e3. https://doi.org/10.1016/j.devcel.2018.03.012.

Lira-Medeiros, Catarina Fonseca, Christian Parisod, Ricardo Avancini Fernandes, Camila Souza Mata, Monica Aires Cardoso, and Paulo Cavalcanti Gomes Ferreira. 2010. "Epigenetic Variation in Mangrove Plants Occurring in Contrasting Natural Environment." PLOS ONE 5 (4): e10326. https://doi.org/10.1371/journal.pone.0010326.

Liu, Junzhong, and Zuhua He. 2020. "Small DNA Methylation, Big Player in Plant Abiotic Stress Responses and Memory." Frontiers in Plant Science 11. https://www.frontiersin.org/article/10.3389/fpls.2020.595603.

Martin, Antoine, Christelle Troadec, Adnane Boualem, Mazen Rajab, Ronan Fernandez, Halima Morin, Michel Pitrat, Catherine Dogimont, and Abdelhafid Bendahmane. 2009. "A Transposon-Induced Epigenetic Change Leads to Sex Determination in Melon." Nature 461 (7267): 1135–38. https://doi.org/10.1038/nature08498.

Monroe, J. Grey, Kevin D. Murray, Wenfei Xian, Thanvi Srikant, Pablo Carbonell-Bejerano, Claude Becker, Mariele Lensink, et al. 2023. "Reply to: Re-Evaluating Evidence for Adaptive Mutation Rate Variation." Nature 619 (7971): E57–60. https://doi.org/10.1038/s41586-023-06315-x.

Monroe, J. Grey, Thanvi Srikant, Pablo Carbonell-Bejerano, Claude Becker, Mariele Lensink, Moises Exposito-Alonso, Marie Klein, et al. 2022. "Mutation Bias Reflects Natural Selection in Arabidopsis thaliana." Nature, January, 1–5. https://doi.org/10.1038/s41586-021-04269-6.

Moser, Bryan R. 2012. "Biodiesel from Alternative Oilseed Feedstocks: Camelina and Field Pennycress." Biofuels 3 (2): 193–209. https://doi.org/10.4155/bfs.12.6.

Niederhuth, Chad E., Adam J. Bewick, Lexiang Ji, Magdy S. Alabady, Kyung Do Kim, Qing Li, Nicholas A. Rohr, et al. 2016. "Widespread Natural Variation of DNA Methylation within Angiosperms." Genome Biology 17 (September): 194. https://doi.org/10.1186/s13059-016-1059-0.

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 n/a (n/a). https://doi.org/10.1111/pbi.13775.

Paun, Ovidiu, Richard M. Bateman, Michael F. Fay, Mikael Hedrén, Laure Civeyrel, and Mark W. Chase. 2010. "Stable Epigenetic Effects Impact Adaptation in Allopolyploid Orchids (Dactylorhiza: Orchidaceae)." Molecular Biology and Evolution 27 (11): 2465–73. https://doi.org/10.1093/molbev/msq150.

Pfeifer, G. P. 2006. "Mutagenesis at Methylated CpG Sequences." In DNA Methylation: Basic Mechanisms, edited by Walter Doerfler and Petra Böhm, 259–81. Current Topics in Microbiology and Immunology. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-31390-7_10.

Pietzenuk, Björn, Catarine Markus, Hervé Gaubert, Navratan Bagwan, Aldo Merotto, Etienne Bucher, and Ales Pecinka. 2016. "Recurrent Evolution of Heat-Responsiveness in Brassicaceae COPIA Elements." Genome Biology 17 (1): 209. https://doi.org/10.1186/s13059-016-1072-3.

Provine, W. B. 1978. "The Role of Mathematical Population Geneticists in the Evolutionary Synthesis of the 1930s and 1940s." Studies in History of Biology 2: 167–92. http://europepmc.org/abstract/med/11610409.

Ramos Cruz, D. (2023). Physiological role and molecular effects of benzoxazinoids (Bx). Wien. https://usearch.univie.ac.at/primo-explore/search?query=any,contains,Benzoxazinoids&tab=default_tab&search_scope=UWI_UBBestand&sortby=date&vid=UWI&facet=frbrgroupid,include,1086382807&lang=en_US&mode=basic&offset=0.

Richards, Christina L., Conchita Alonso, Claude Becker, Oliver Bossdorf, Etienne Bucher, Maria Colomé-Tatché, Walter Durka, et al. 2017. "Ecological Plant Epigenetics: Evidence from Model and Non-Model Species, and the Way Forward." Ecology Letters 20 (12): 1576–90. https://doi.org/10.1111/ele.12858.

Richards Christina L., Bossdorf Oliver, and Verhoeven Koen J. F. 2010. "Understanding Natural Epigenetic Variation." New Phytologist 187 (3): 562–64. https://doi.org/10.1111/j.1469-8137.2010.03369.x.

Rodríguez, Bárbara Díez, Cristian Peña, Paloma Pérez Bello, Julius Bette, Lena Lerbs, Tabea Mackenbach, Sven Wulle, et al. 2022. "An Uncommon Garden Experiment: Microenvironment Has Stronger Influence on Phenotypic Variation than Epigenetic Memory in the Clonal Lombardy Poplar." bioRxiv. https://doi.org/10.1101/2022.03.22.485169.

Roquis, David, Marta Robertson, Liang Yu, Michael Thieme, Magdalena Julkowska, and Etienne Bucher. 2021. "Genomic Impact of Stress-Induced Transposable Element Mobility in Arabidopsis." Nucleic Acids Research 49 (18): 10431–47. https://doi.org/10.1093/nar/gkab828.

Sasaki, Eriko, Taiji Kawakatsu, Joseph R. Ecker, and Magnus Nordborg. 2019. "Common Alleles of CMT2 and NRPE1 Are Major Determinants of CHH Methylation Variation in Arabidopsis thaliana." PLOS Genetics 15 (12): e1008492. https://doi.org/10.1371/journal.pgen.1008492.

Schmitz, Robert J., Zachary A. Lewis, and Mary G. Goll. 2019. "DNA Methylation: Shared and Divergent Features across Eukaryotes." Trends in Genetics 35 (11): 818–27. https://doi.org/10.1016/j.tig.2019.07.007.

Stroud, Hume, Truman Do, Jiamu Du, Xuehua Zhong, Suhua Feng, Lianna Johnson, Dinshaw J. Patel, and Steven E. Jacobsen. 2014. "Non-CG Methylation Patterns Shape the Epigenetic Landscape in Arabidopsis." Nature Structural & Molecular Biology 21 (1): 64–72. https://doi.org/10.1038/nsmb.2735.

Talley, Sharon M, Phyllis D Coley, and Thomas A Kursar. 2002. "The Effects of Weather on Fungal Abundance and Richness among 25 Communities in the Intermountain West." BMC Ecology 2 (June): 7. https://doi.org/10.1186/1472-6785-2-7.

Tomkova, Marketa, and Benjamin Schuster-Böckler. 2018. "DNA Modifications: Naturally More Error Prone?" Trends in Genetics: TIG 34 (8): 627–38. https://doi.org/10.1016/j.tig.2018.04.005.

Troyee, A. Niloya, Mónica Medrano, Caroline Müller, and Conchita Alonso. 2022. "Variation in DNA Methylation and Response to Short-Term Herbivory in Thlaspi Arvense." Flora 293 (August): 152106. https://doi.org/10.1016/j.flora.2022.152106.

Vaughn, Matthew W., Miloš Tanurdžić, Zachary Lippman, Hongmei Jiang, Robert Carrasquillo, Pablo D. Rabinowicz, Neilay Dedhia, et al. 2007. "Epigenetic Natural Variation in Arabidopsis thaliana." PLOS Biology 5 (7): e174. https://doi.org/10.1371/journal.pbio.0050174.

Vázquez, Rafael P., Mariano Hern, M. José Mart, and Rosa de Frutos. 2000. "Evolution of Gypsy Endogenous Retrovirus in the Drosophila Obscura Species Group." Molecular Biology and Evolution 17 (8): 1185–93. https://doi.org/10.1093/oxfordjournals.molbev.a026401.

Velásquez, André C., Christian Danve M. Castroverde, and Sheng Yang He. 2018. "Plant and Pathogen Warfare under Changing Climate Conditions." Current Biology: CB 28 (10): R619–34. https://doi.org/10.1016/j.cub.2018.03.054.

Wang, Dandan, Zeyu Zheng, Ying Li, Hongyin Hu, Zhenyue Wang, Xin Du, Shangzhe Zhang, et al. 2021. "Which Factors Contribute Most to Genome Size Variation within Angiosperms?" Ecology and Evolution 11 (6): 2660–68. https://doi.org/10.1002/ece3.7222.

Wang, Long, Alexander T. Ho, Laurence D. Hurst, and Sihai Yang. 2023. "Re-Evaluating Evidence for Adaptive Mutation Rate Variation." Nature 619 (7971): E52–56. https://doi.org/10.1038/s41586-023-06314-y.

Warwick, S. I., A. Francis, and D. J. Susko. 2002. "The Biology of Canadian Weeds. 9. Thlaspi Arvense L. (Updated)." Canadian Journal of Plant Science 82 (4): 803–23. https://doi.org/10.4141/P01-159.

Wells, Jonathan N., and Cédric Feschotte. 2020. "A Field Guide to Eukaryotic Transposable Elements." Annual Review of Genetics 54 (1): 539–61. https://doi.org/10.1146/annurev-genet-040620-022145.

Wicker, Thomas, François Sabot, Aurélie Hua-Van, Jeffrey L. Bennetzen, Pierre Capy, Boulos Chalhoub, Andrew Flavell, et al. 2007. "A Unified Classification System for Eukaryotic Transposable Elements." Nature Reviews Genetics 8 (12): 973–82. https://doi.org/10.1038/nrg2165.

Williams, George Christopher. 2018. Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought. Princeton University Press.

Zhang, Huiming, Zhaobo Lang, and Jian-Kang Zhu. 2018. "Dynamics and Function of DNA Methylation in Plants." Nature Reviews Molecular Cell Biology, May, 1. https://doi.org/10.1038/s41580-018-0016-z.

Zhao, Ru, Xinyu Yang, Muzhi Li, Xiaojin Peng, Mengxia Wei, Xiucheng Zhang, Lei Yang, and Jialei Li. 2021. "Biodiesel Preparation from Thlaspi Arvense L. Seed Oil Utilizing a Novel Ionic Liquid Core-Shell Magnetic Catalyst." Industrial Crops and Products 162 (April): 113316. https://doi.org/10.1016/j.indcrop.2021.113316.

Zilberman, Daniel. 2017. "An Evolutionary Case for Functional Gene Body Methylation in Plants and Animals." Genome Biology 18 (1): 87. https://doi.org/10.1186/s13059-017-1230-2.

Chapter I

Genetic and environmental drivers of large-scale epigenetic variation in *Thlaspi arvense*

Dario Galanti, Daniela Ramos-Cruz, Adam Nunn, Isaac Rodríguez-Arévalo, J. F. Scheepens, Claude Becker, Oliver Bossdorf.

https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1010452

PLOS GENETICS

RESEARCH ARTICLE

Genetic and environmental drivers of large-scale epigenetic variation in *Thlaspi arvense*

Dario Galanti[1], **Daniela Ramos-Cruz**[2,3], **Adam Nunn**[4,5], **Isaac Rodríguez-Arévalo**[2,3], **J. F. Scheepens**[6], **Claude Becker**[2,3], **Oliver Bossdorf**[1]*

1 Plant Evolutionary Ecology, Institute of Evolution and Ecology, University of Tübingen, Tübingen, Germany, **2** Gregor Mendel Institute of Molecular Plant Biology, Austrian Academy of Sciences, Vienna BioCenter (VBC), Vienna, Austria, **3** LMU Biocenter, Faculty of Biology, Ludwig Maximilians University Munich, 82152 Martinsried, Germany, **4** ecSeq Bioinformatics GmbH, Leipzig, Germany, **5** Institute for Computer Science, University of Leipzig, Leipzig, Germany, **6** Plant Evolutionary Ecology, Institute for Ecology, Evolution and Diversity, Faculty of Biological Sciences, Goethe University Frankfurt, Frankfurt am Main, Germany

* oliver.bossdorf@uni-tuebingen.de

Check for updates

OPEN ACCESS

Citation: Galanti D, Ramos-Cruz D, Nunn A, Rodríguez-Arévalo I, Scheepens JF, Becker C, et al. (2022) Genetic and environmental drivers of large-scale epigenetic variation in *Thlaspi arvense*. PLoS Genet 18(10): e1010452. https://doi.org/10.1371/journal.pgen.1010452

Editor: Nathan M. Springer, University of Minnesota, UNITED STATES

Received: March 24, 2022

Accepted: September 28, 2022

Published: October 12, 2022

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pgen.1010452

Copyright: © 2022 Galanti et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Seed material from the sequenced lines was collected complying with the Nagoya protocol (permit number: TREL1734890A/19) and is available at the

Abstract

Natural plant populations often harbour substantial heritable variation in DNA methylation. However, a thorough understanding of the genetic and environmental drivers of this epigenetic variation requires large-scale and high-resolution data, which currently exist only for a few model species. Here, we studied 207 lines of the annual weed *Thlaspi arvense* (field pennycress), collected across a large latitudinal gradient in Europe and propagated in a common environment. By screening for variation in DNA sequence and DNA methylation using whole-genome (bisulfite) sequencing, we found significant epigenetic population structure across Europe. Average levels of DNA methylation were strongly context-dependent, with highest DNA methylation in CG context, particularly in transposable elements and in intergenic regions. Residual DNA methylation variation within all contexts was associated with genetic variants, which often co-localized with annotated methylation machinery genes but also with new candidates. Variation in DNA methylation was also significantly associated with climate of origin, with methylation levels being lower in colder regions and in more variable climates. Finally, we used variance decomposition to assess genetic versus environmental associations with differentially methylated regions (DMRs). We found that while genetic variation was generally the strongest predictor of DMRs, the strength of environmental associations increased from CG to CHG and CHH, with climate-of-origin as the strongest predictor in about one third of the CHH DMRs. In summary, our data show that natural epigenetic variation in *Thlaspi arvense* is significantly associated with both DNA sequence and environment of origin, and that the relative importance of the two factors strongly depends on the sequence context of DNA methylation. *T. arvense* is an emerging biofuel and winter cover crop; our results may hence be relevant for breeding efforts and agricultural practices in the context of rapidly changing environmental conditions.

Nottingham Arabidopsis Stock Centre (NASC) under stock numbers N950001 to 950204. Genomic and bisulfite sequencing raw data are available on the ENA Sequence Read Archive (www.ebi.ac.uk/ena/) under study accession number PRJEB50950. Reference genome and annotations were previously published by Nunn et al. (2022). From the annotation, we subset confident de novo gene annotations with an annotation edit distance score < 1.0 (source: T_arvense_v2). Mean and weighted methylation values extracted from sequencing data, DMRs, Gene Body Methylation classification and GWA results (-log(p)>1) in a format compatible with the Integrative Genomics Viewer (software.broadinstitute.org/software/igv/) are available on Zenodo (doi.org/10.5281/zenodo.6361977). All the code used in this study is available and documented on Github. Pipelines for methylation and DMR calling from WGBS data are on the EpiDiverse Github (https://github.com/EpiDiverse). The workflow for downstream analysis of methylation data is on https://github.com/Dario-Galanti/WGBS_downstream. Scripts for downstream processing of DMRs and DMRs variance decomposition are on https://github.com/Dario-Galanti/popDMRs_refine_VCA. Scripts for variant calling, filtering and imputation, performed on the Baden-Württemberg BinAC cluster, are on https://github.com/Dario-Galanti/BinAC_varcalling. Finally, scripts for running GWA analysis and the enrichment of a priori candidates are on https://github.com/Dario-Galanti/multipheno_GWAS.

Funding: This work is part of the European Training Network EpiDiverse (https://epidiverse.eu), which received funding from the EU Horizon 2020 program under Marie Skłodowska-Curie grant agreement No 764965. This work was also supported by the European Union's Horizon 2020 research and innovation program via the European Research Council (ERC) Grant agreement No. 716823 "FEAR-SAP" to C.B., by the Austrian Academy of Sciences and by the German Research Foundation (DFG) through grant no INST 37/935-1 FUGG. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

Variation within species is an important level of biodiversity, and it is key for future adaptation. Besides variation in DNA sequence, plants also harbour heritable variation in DNA methylation, and we want to understand the evolutionary significance of this epigenetic variation, in particular how much of it is under genetic control, and how much is associated with the environment. We addressed these questions in a high-resolution molecular analysis of 207 lines of the common plant field pennycress (*Thlaspi arvense*), which we collected across Europe, propagated under standardized conditions, and sequenced for their genetic and epigenetic variation. We found large geographic variation in DNA methylation, associated with both DNA sequence and climate of origin. Genetic variation was generally the stronger predictor of DNA methylation variation, but the strength of environmental association varied between different sequence contexts. Climate-of-origin was the strongest predictor in about one third of the differentially methylated regions in the CHH context, which suggests that epigenetic variation may play a role in the short-term climate adaptation of pennycress. As pennycress is currently being domesticated as a new biofuel and winter cover crop, our results may be relevant also for agriculture, particularly in changing environments.

Introduction

Besides variation in DNA sequence, natural plant populations usually also harbour variation in epigenetic modifications of the DNA. This is particularly well documented for DNA methylation, usually referring to the addition of a methyl group to the 5th atom of the cytosine ring, a modification associated with silencing of transposable elements (TEs) and the regulation of gene expression. Variation in DNA methylation can arise if methylation marks are altered by chance during mitosis or meiosis (epimutations) [1,2], or if they are induced in response to environmental changes [3,4]. Some DNA methylation differences are stably inherited through meiosis, which has led some to hypothesize that DNA methylation variation could be under natural selection and contribute to adaptation [5–7]. These ideas are fuelled by the observation that DNA methylation variation in natural plant populations is often non-random and geographically structured [8–12]. However, the DNA methylation variation observed in the field is always a combination of stable (= heritable) and plastic (= non-heritable) components. In order to tease these apart and describe the heritable component of DNA methylation variation, one must analyse the offspring of different populations grown in a common environment. To date, common-environment analyses of natural DNA methylation variation that cover many populations and broad environmental gradients are still rare.

In plants, DNA methylation can occur in the three sequence contexts: CG, CHG and CHH (where H is A, T or C). Distinguishing between these contexts is sensible because they differ in the molecular machineries for depositing, maintaining and removing methylation [13,14], which has consequences for their dynamics and stability. In *Arabidopsis thaliana*, CG methylation (mCG) is mostly maintained in a copy-paste manner during replication, CHG methylation (mCHG) by DNA-histone methylation self-reinforcing loops and CHH methylation (mCHH) by recursive *de-novo* methylation deposited by the RNA-directed DNA methylation pathway (RdDM) and partially by CMT2 [13,14]. In addition, CHG and CHH methylation partially share maintenance pathways [15,16]. Overall, there is a gradient of similarity and decreasing stability from CG to CHG to CHH. Although less stable, CHH is the most abundant context and often the most responsive to stresses [17]. Besides the sequence contexts, the

dynamics of DNA methylation also strongly depend on the genomic features in which it occurs. While heterochromatic regions and TEs are usually heavily methylated to repress transcription, methylation is often lower and more variable in genes and regulatory regions [18–20]. In addition, while DNA methylation is almost exclusively a repressive mark on TEs and in regulatory regions, its function is less clear in gene bodies, as several constitutively expressed housekeeping genes often harbour CG but not CHG and CHH methylation in their coding sequences (CDS) [20,21]. If methylation in different genomic features has different functions, then also different selective pressures are to be expected [22]. Finally, for both influences of sequence context and genomic features on methylation variation, there appears to be high species-specificity in plants [20].

To study such complex dynamics, DNA methylation can be quantified at different levels, from global (or genome-wide) methylation, to average methylation limited to sequence contexts or genomic features, to the methylation of genomic regions or individual positions. While genetic single nucleotide polymorphisms (SNPs) can have large effects, this does not seem to be the case for DNA methylation polymorphisms, which affect transcription only when accumulating over a broader genomic region [23–25]. For this reason, the study of differentially methylated regions (DMRs) became very popular in high-resolution studies [11,12,18,26].

Given the complex molecular machinery for regulation and maintenance of DNA methylation, it is not surprising that previous studies have demonstrated various kinds of genetic control over DNA methylation variation. Genetic polymorphisms can control DNA methylation in *cis*, for example, when a TE insertion next to a gene promoter induces the methylation of the latter [25], or in *trans*, when genetic mutations affect genes involved in the DNA methylation machinery [11,12,27]. In the latter case, variation in individual DNA loci often affects methylation levels across the entire genome. In addition, a number of genes have been found to affect methylation levels indirectly, acting upstream or in aid of the methylation machinery. In particular, ubiquitination, a post-translational modification affecting histone tails and protein turnover, affects DNA methylation in plants and animals in several ways [28–33]. For example, in plants ORTH/VIM E3 ubiquitin ligases recruit DNA METHYLTRANSFERASE 1 (MET1) for methylation maintenance through ubiquitination of histone tails [30,31]. However, in spite of this functional understanding of several mechanisms of genetic control, we still lack a good understanding of the degree of genetic determination of DNA methylation variation in wild plant populations.

If DNA methylation variation is under natural selection–whether independently from DNA sequence or linked to it–we expect this to result in patterns of association between methylation variation and the environment. Several previous studies indeed found correlations between methylation patterns and habitat or climate in different plant species [8–12,34]. However, most of these studies were either conducted in the field, based on only few natural populations, or used low-resolution molecular methods, which limited their generalizability and/or their power to detect environment-methylation associations and to separate genetically controlled from independent components of DNA methylation variation [5]. The only available data that does not suffer from any of these limitations comes from *Arabidopsis thaliana* [11,12,18], a plant with an exceptionally small and simple genome, with low numbers of TEs, and low global DNA methylation [35]. Given these unrepresentative genomic properties, it is currently unclear to which extent findings from population epigenomic studies with *A. thaliana* can be generalised across the plant kingdom. As the abundance and genomic distribution of TEs is a major driver of variation in DNA methylation, species with higher TE contents could differ not only in the extent of DNA methylation, but also in the dynamics of epimutation accumulation, and the DNA methylation-based machinery controlling TE mobility. To

understand the extent of these differences, and the genetic and environmental drivers of natural DNA methylation variation, it is critical to expand our scope and collect large-scale, high-resolution data also for other plant species.

Here, we present a detailed genomic analysis of 207 lines of the plant *Thlaspi arvense* (field pennycress) that we collected across a latitudinal gradient in Europe, cultivated in a common environment, and profiled for genomic and epigenomic variation. Like *A*. *thaliana*, *T*. *arvense* is an annual and mostly selfing member of the Brassicaceae family, but it has a significantly larger genome of approx. 500Mb, which is richer in TEs and DNA methylation [36]. The species is an interesting study object also because it is currently being domesticated into a new biofuel and cover crop [37–41]. The genomic work with *T*. *arvense* is facilitated by recently published high-quality reference genomes [36,42]. In our study, we demonstrate that European populations of *T*. *arvense* harbour substantial natural epigenetic variation, which is associated with DNA sequence variation as well as with climate of origin, but in a highly context-dependent manner. In our data, genetic variation was generally the stronger predictor of DNA methylation variation. Genome-wide association analyses identified several candidate loci, but there was a fraction of the DNA methylation variation that was most strongly associated with climate of origin, suggesting a link with climate adaptation.

Results

The 207 *Thlaspi arvense* lines we worked with came from 36 natural populations which we sampled across Europe in 2018, on a latitudinal gradient from Southern France to Central Sweden, with three populations each in Southern France and The Netherlands, seven in Southern Germany, eight in Central Germany and South Sweden, respectively, and another seven populations in Central Sweden (Fig 1A and S1 Table). In each population, we collected seeds of 4–6 different lines (S1 Table). We grew all lines under common environmental conditions, extracted their DNA and generated Whole Genome Sequencing (WGS) and Whole Genome Bisulfite Sequencing (WGBS) libraries, which upon deduplication, were sequenced with an average coverage of 19.7x and 30.3x, respectively (S2 Table). Bisulfite non-conversion rates were calculated from chloroplast DNA and ranged between 0.14 and 1.9% (S2 Table). Variant calling retrieved around nine million SNPs and short INDELs with genotypes called in >90% of the lines. Methylation calling retrieved about 16 million, 18.4 million and 95.3 million positions in CG, CHG and CHH contexts respectively, with up to 25% missing calls per position. The global weighted DNA methylation, calculated as the ratio between all methylated and all total read counts at every analysed cytosine [43], was estimated at 16.9% (average of all lines).

We found significant genetic and epigenetic population structure across Europe. A principal component analysis (PCA) based on genetic variants showed two main clades: a larger one including almost all lines from France, Germany and the Netherlands, and a smaller one that consisted almost exclusively of Swedish lines (Fig 1B). The larger clade also showed a clear latitudinal gradient. PCAs based on DNA methylation variation generally also found two major clades, with the CG methylation-based patterns most closely resembling the genetically-based ones, and a decreasing similarity between genetic and epigenetic population structure from CG to CHG to CHH methylation (Fig 1B and S1). Restricting methylation to specific genomic features also revealed that mCG of genes and promoters has stronger geographic patterns that methylation of TEs (S1 Fig).

Average methylation

To understand the structure of DNA methylation variation in *T*. *arvense*, we first examined patterns of weighted average methylation [43] across all lines. We not only distinguished

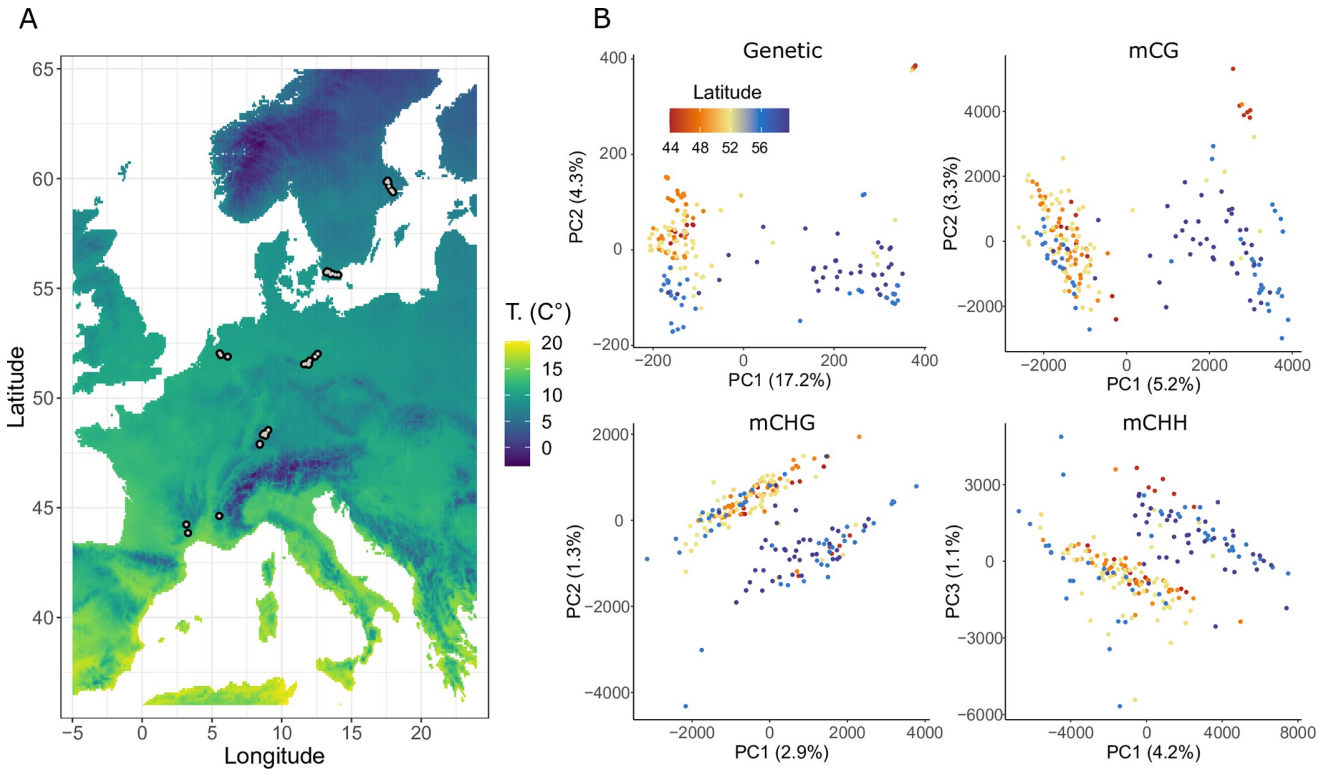


Fig 1. Geographic distribution and population structure of the 207 sampled *Thlaspi arvense* lines. (A) Geographic locations of the 36 populations. The background colours are gridded satellite data of average daily temperature (T.) from the Copernicus programme [44]. (B) PCA plots of all 207 lines based on DNA sequence ("Genetic") and DNA methylation in different sequence contexts ("mCG", "mCHG" and "mCHH").

https://doi.org/10.1371/journal.pgen.1010452.g001

between the three sequence contexts CG, CHG and CHH, but we also assigned cytosines to different genomic features: CDS, introns, promoters, TEs and intergenic regions. For genes and TEs, we used available annotations [36], while for promoters we considered the 2 kb upstream sequences of genes (or until the boundary of the previous gene if closer). We considered intergenic space, anything not belonging to these categories. Across all genomic features, the average methylation was much higher in CG context than in CHG and CHH; for the latter two it was generally similar (Fig 2A). TEs were the most highly methylated genomic features, followed by intergenic regions, whereas promoters and especially gene bodies (CDS and introns) showed very low average methylation (Fig 2A). For instance, while for CG sites in TEs the weighted average methylation was around 80%, it was below 2% for CHH sites in genes. Although these patterns are conserved in the whole collection, there is large residual variation between lines, which is particularly high in TEs (up to 12%) and decreases gradually moving to intergenic regions, promoters and particularly genes (Fig 2A). Finally, partially due to TEs covering about 60% of the *T. arvense* genome [36], its global weighted methylation of 16.9% (average of all lines) is much higher than that of *A. thaliana* (5.8%)[12] and many other Brassicaceae [20].

To better understand the observed values of average methylation, and in particular the low gene body methylation, we further examined the distributions of methylation values of individual CDS, TEs and promoters, averaging across all lines. Interestingly, while context-specific methylation levels were very consistent for TEs, almost exclusively methylated, we found

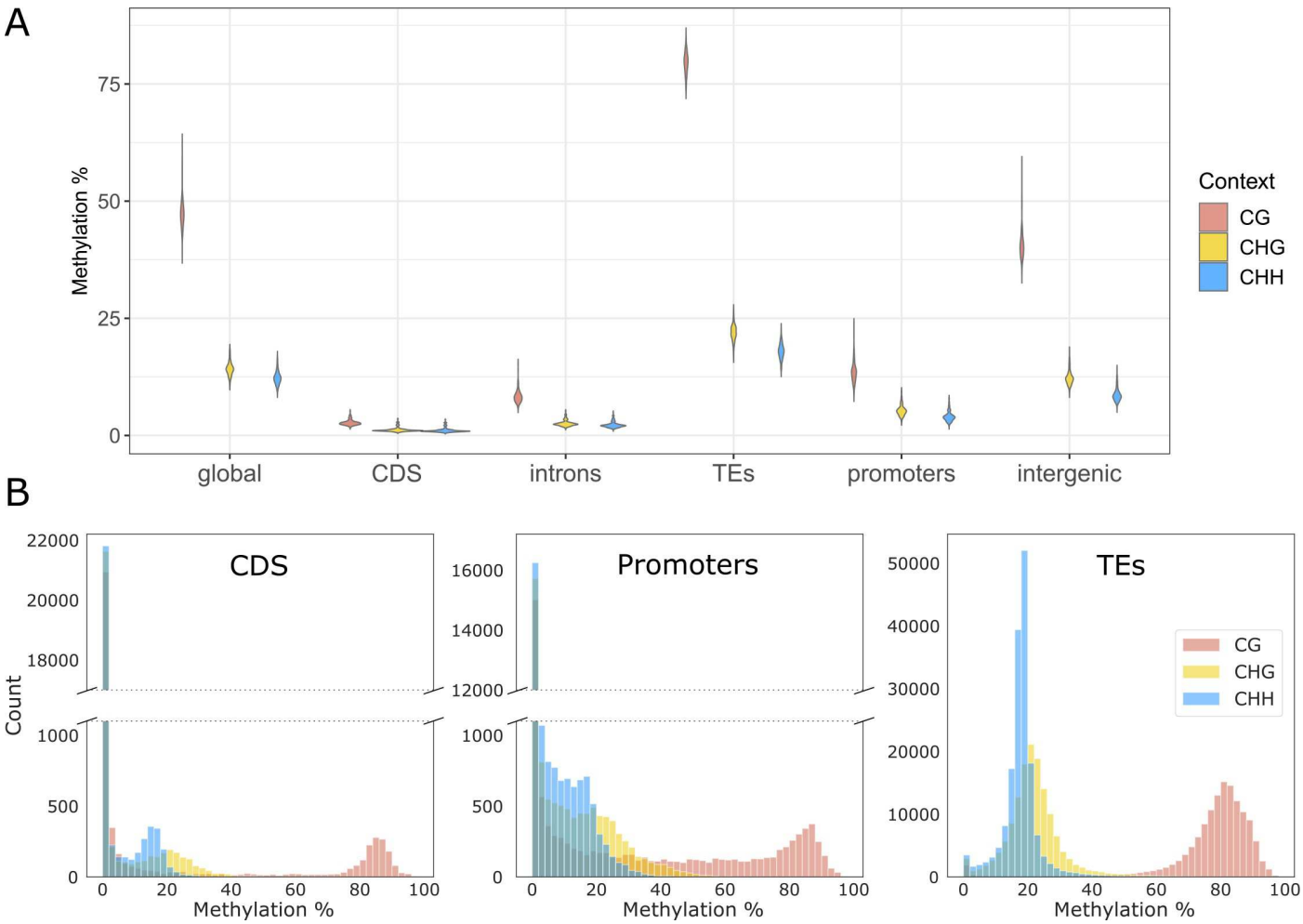


Fig 2. Average methylation and distributions of methylation values for different sequence contexts and genomic features in *T. arvense*. (A) Weighted average methylation levels of genomic features; violin plots represent variation between lines. (B) Distributions of individual methylation values for coding sequences (CDS), promoters and transposable elements (TEs) obtained averaging across all 207 lines.

https://doi.org/10.1371/journal.pgen.1010452.g002

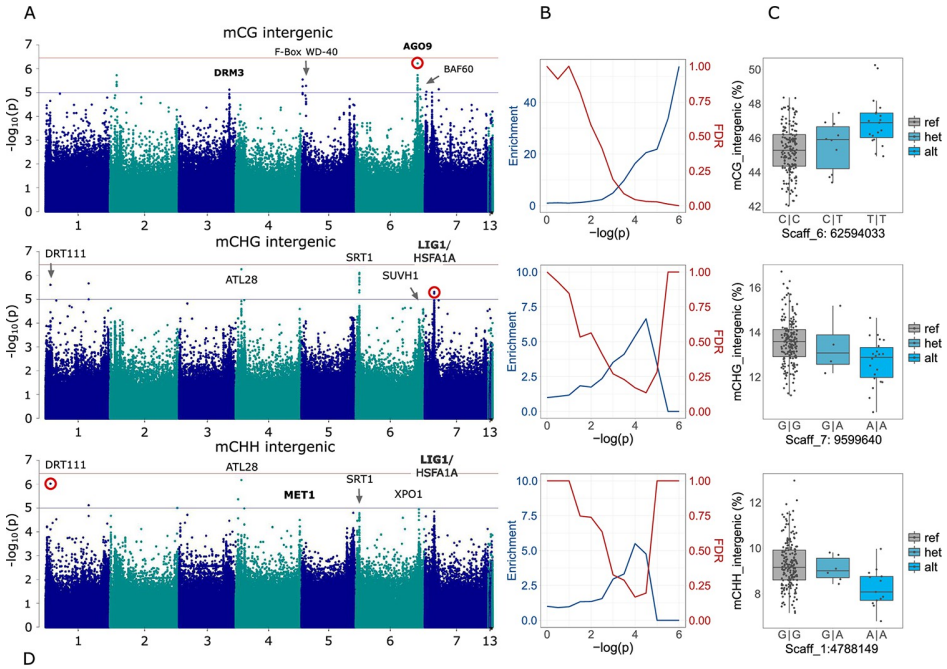
bimodal distributions for CDS and promoters, with a large majority of unmethylated and a smaller fraction of methylated features (Fig 2B). Using a binomial test [20,45], we found than only a small portion of genes is significantly methylated in each sequence context (7.5, 6.5 and 7.3% on average for CG, CHG and CHH respectively), with rather small variation between lines (S3 Table). Intersecting genes consistently methylated (methylated in at least 70% of the lines) in each of the three sequence contexts, we confirmed that a large fraction of these was methylated in all context, showing a TE-like methylation signature (TEm), and a much smaller fraction was methylated only in CG, showing a gene body methylation signature (gbM) (S2A Fig)[36]. Moreover, the fraction of methylated genes, tended to cooccur with TEs, since TEm genes were about eight times more likely than the average gene to overlap with TEs, and gbM genes were twice as likely. Even though many TEm genes might be pseudogenes, a gene ontology (GO) enrichment analysis found enrichment for some housekeeping-like GO terms such as nucleotide biosynthesis and protein catalysis (S2B Fig). In contrast, the few genes methylated only in CG, were only enriched for few molecular functions (S2B Fig).

Genetic basis of methylation variation

To understand the genetic basis of the observed methylation variation, we employed genome-wide association (GWA) analyses that tested for statistical associations between every biallelic genetic variant and the average methylation of every sequence context and genomic feature (S4 Table). For this analysis we used the (unweighted) mean methylation, as weighted methylation is strongly influenced by structural and copy number variants, which could distort GWA and produce misleading results when looking for individual genes affecting methylation levels. We restricted our analyses to genetic variants with a minor allele frequency (MAF) ≥ 0.04; however, repeating all analyses with a MAF>0.01 did not influence the results relevantly. Since large numbers of unmethylated genes (Fig 2B) could potentially obscure association patterns in methylated genes, we re-ran these analyses for average methylation levels based only on genes with methylation $> 5\%$ (across all lines). In all GWA analyses, we corrected for population structure using an Isolation-By-State (IBS) distance matrix. Although our experimental design and number of lines hardly provided sufficient power to meet a full Bonferroni threshold, we found that many of the genetic variants that were most strongly associated with methylation levels were close to genes with predicted functions related to DNA methylation (Figs 3A, 3D and S3 Fig). For instance, one strong candidate was an orthologue of *ARGONAUTE 9 (AGO9)*, coding a DICER-like protein involved in RNA silencing; *AGO9* natural variation is associated with mCHH in TEs in *A*. *thaliana* [12]. Another candidate was an orthologue of *DOMAINS REARRANGED METHYLTRANSFERASE 3 (DRM3)*, which despite being catalytically mutated, is necessary for RdDM and non-CG methylation maintenance in Arabidopsis [46–48]. Reflecting the multigenic basis of methylation, even the higher -log(p) variants had relatively small size effects of about 1.5% methylation (Fig 3C).

To confirm the suspected enrichment of methylation-related genes among stronger associations, we conducted an enrichment analysis based on all genetic variants within 20kb from *a priori* candidate genes–orthologues of *A*. *thaliana* genes known to affect methylation (S5 Table). For many genomic features and sequence contexts, we indeed found an enrichment of these *a priori* candidates among the genetic variants most strongly associated with average methylation levels (e.g. mCG in Fig 3B), but in most cases the top variants were not neighbouring any *a priori* candidates (drop of the enrichment for high -log(p) thresholds in mCHG and mCHH in Fig 3B; see S3 Fig for more results). Nevertheless, a search of the neighbouring regions of these variants identified several new candidates that may not affect methylation directly, but have predicted functions with a potential for indirect effects on DNA methylation. These include e.g. the histone deacetylase *SIRTUIN 1 (SRT1)*, the DNA-damage-repair/toleration *(DRT111)*, *the DNA-repair gene STRUCTURAL MAINTENANCE OF CHROMOSOMES 5 (SMC5)* and several E3 ubiquitin ligases such as F-box transcription factors and RING-H2 finger proteins (Fig 3; see S6 Table for all genes located within 15kb from variants significant at -log(p) > 5). Overall, our results showed that natural DNA methylation variation in *T*. *arvense* was significantly associated with underlying DNA sequence variation, but only some of the top genetic variants were known methylation machinery genes, whereas there were many additional, less well-characterized genes that appeared to play a role, possibly through less direct effects on methylation.

The GWA results strongly differed between sequence contexts, with a unique profile of genetic variants associated with average mCG, while the results were very similar for mCHG and mCHH (Figs 3A, 3D and S3 Fig). In mCG, some of the top candidates were *AGO9*, the methyltransferase *DRM3*, the F-box/WD-40 repeat-containing gene *Tarvense_02099*, involved in histone methylation, and two orthologues of the SWI/SNF chromatin remodelling component *BAF60* (S6 Table). In mCHG and mCHH, the strongest associations included *SRT1*,



Candidate gene	Gene ID	Ortholog	Putative function	Top SNP location
DRT111	Ta_11093	AT1G30480	DNA-DAMAGE-REPAIR/TOLERATION 111	9kb downstream
DRM3	**Ta_08517**	**AT3G17310**	**Catalytically mutated methyltransferase (RdDM)**	**13kb upstream**
ATL28	Ta_19900	AT2G35420	RING-H2 finger protein (protein modification)	3kb upstream
MET1	**Ta_24743**	**AT5G49160**	**DNA methyltransferase 1**	**3kb upstream**
F-Box WD-40	Ta_02099	AT3G52030	Histone H3-K4 methylation	<1kb upstream
SRT1	Ta_00697	AT5G55760	Sirtuin family HDAC (histone deacetylase)	1kb downstream
AGO9	**Ta_05665**	**AT5G21150**	**ARGONAUTE 9 (siRNA silencing)**	**14kb downstream**
XPO1	Ta_05994	AT5G17020	EXPORTIN 1 (embryo and pollen development)	0kb
BAF60	Ta_06233/ Ta_06235	AT5G14170	SWI/SNF complex component (chromatin remodeling)	2kb downstream / 1kb upstream
SUVH1	Ta_06979	AT5G04940	SU(VAR)3-9 homolog (histone methyltransferase)	3kb upstream
LIG1 / HSFA1A	**Ta_17381** / Ta_17386	**AT1G08130** / AT4G17750	**DNA Ligase 1 (DNA demethylation)** / HEAT SHOCK FACTOR 1A (heat stress, transposon activation)	**15kb downstream of LIG1** / 3kb upstream of HSFA1A

Fig 3. Genome-wide association analyses for genetic control of average DNA methylation. We show only the results for intergenic methylation; for full results see S3 Fig. (A) Manhattan plots, with the top variants labelled with the neighbouring genes potentially affecting methylation. The genome-wide significance (horizontal red lines), was calculated based on unlinked variants as in Sobota et al. (2015) [49], the suggestive-line (blue) corresponds to–log(p) = 5. (B) Corresponding to each Manhattan plot on the left, enrichment of *a priori* candidates and expected false discovery rates (both as in Atwell et al. 2010 [50]) for stepwise significance thresholds. (C) The allelic effects of the red-marked variants in the corresponding Manhattan plots on the left, with genotypes on the x-axes and the average methylation on the y-axes. (D) The candidate genes marked in panel A, their putative functions and distances to the top variant of the neighbouring peaks. Bold font indicates *a priori* candidates that were included in the enrichment analyses.

https://doi.org/10.1371/journal.pgen.1010452.g003

SMC5, the *DNA LIGASE 1 (LIG1)*, involved in DNA demethylation, and *DRT111*. Lastly, we tested whether variation in number of gbM genes between lines was associated to genetic variants and detected a clear peak in Scaffold_3, including, among a few additional genes, *LOG2--LIKE UBIQUITIN LIGASE3 (LUL3)*, which codes a ubiquitin ligase (S2C Fig).

Methylation relationships with climate of origin

To test for environmental associations of methylation variation, we compiled bioclimatic data (see Methods section for details) for our 36 study populations and analysed the relationships

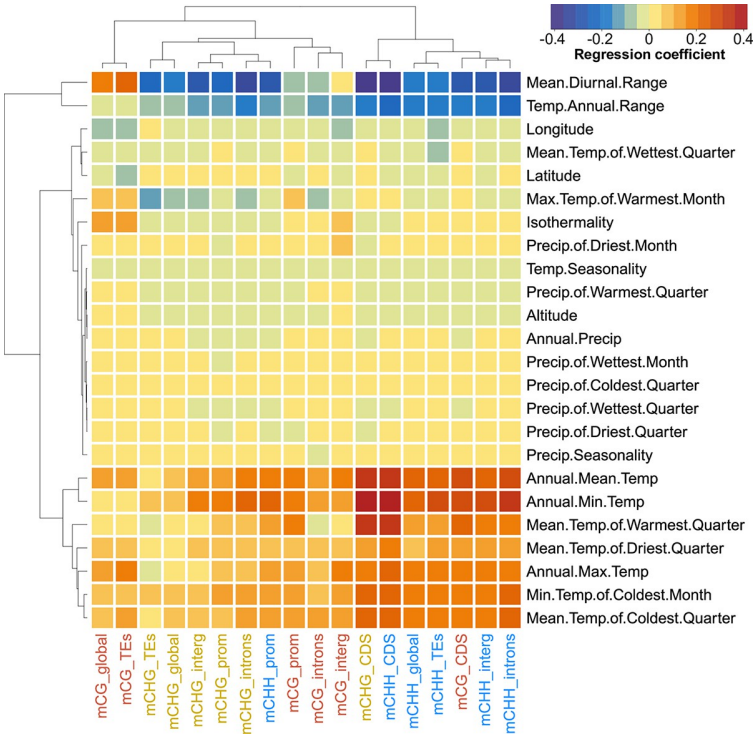


Fig 4. Climate-methylation associations. A Heatmap of the correlations between mean methylation and different climatic variables (Precip: precipitation; Temp: temperature), separately for different sequence contexts and genomic features (prom: promoter; interg: intergenic; TEs: Transposable Elements; CDS: coding sequences). Both rows and columns are clustered by their multivariate similarity in association patterns.

https://doi.org/10.1371/journal.pgen.1010452.g004

between climatic variables and the mean methylation in different sequence contexts and genomic features, correcting for population structure with the same IBS matrix used in the GWA analyses. We found that average methylation was positively correlated with several climate variables reflecting variation in mean temperatures, but negatively with variables related to temperature variability, such as the mean diurnal range and annual temperature range (Fig 4). Moreover, associations with temperature were more pronounced for minimum temperature variables than for maximum temperature variables. In other words, plants originating from colder origins or such with more fluctuating temperature environments had lower overall methylation. In contrast to the temperature variables, methylation was not associated with the precipitation variation of the population of origin, and there was also little association with latitude (Fig 4). The latter at first appears counterintuitive, because latitude is usually correlated with temperature, but in our case latitude is confounded with altitude–more southern samples were collected at higher elevations (S1 Table)–and therefore poorly correlated with temperature.

The described climate-methylation associations were generally stronger in CHG and CHH contexts, particularly for methylation that occurred in CDS (Fig 4). With the exception of mCG in CDS, which had climate associations similar to mCHH, other methylation variables clustered mostly by sequence context, with some similarity between CG and CHG. Finally, global and TEs mCG were the only types of methylation positively associated with temperature variability (Fig 4).

DMR variance decomposition

Having established associations of methylation variation with genetic background and environment of origin, we sought to investigate the relative importance of these two drivers in our study system, and how this might vary between sequence contexts and genomic features. To address these questions, we analysed methylation variation at the level of DMRs. We identified around 44k DMRs in CG, 12k DMRs in CHG and 77k DMRs in CHH (see Methods for details on the DMR calling), and quantified their overlap with different genomic features. Most DMRs were located in TEs, and decreasing numbers in intergenic regions, promoters and genes (Fig 5B).

To quantify the degrees of genetic versus environmental determination, we then analysed three mixed models for each DMR that included either a distance matrix based on genetic variants in *cis*, on genetic variants in *trans*, or on multivariate climatic distances. Across all DMRs, genetic similarity based on *trans*-variants explained the largest proportions of methylation variance in all contexts (Fig 5A). Most variance was explained in CHG-DMRs, followed closely by CG-DMRs, but in CHH-DMRs the amounts of variance explained were generally much lower. Interestingly, the explanatory power of environmental variation relative to that of genetic variation gradually increased from the more stable mCG towards the less stable mCHG and mCHH (Fig 5A and 5C).

Although genetic variation in *trans* was on average the strongest predictor of methylation variance, there were large differences between individual DMRs, and we observed that

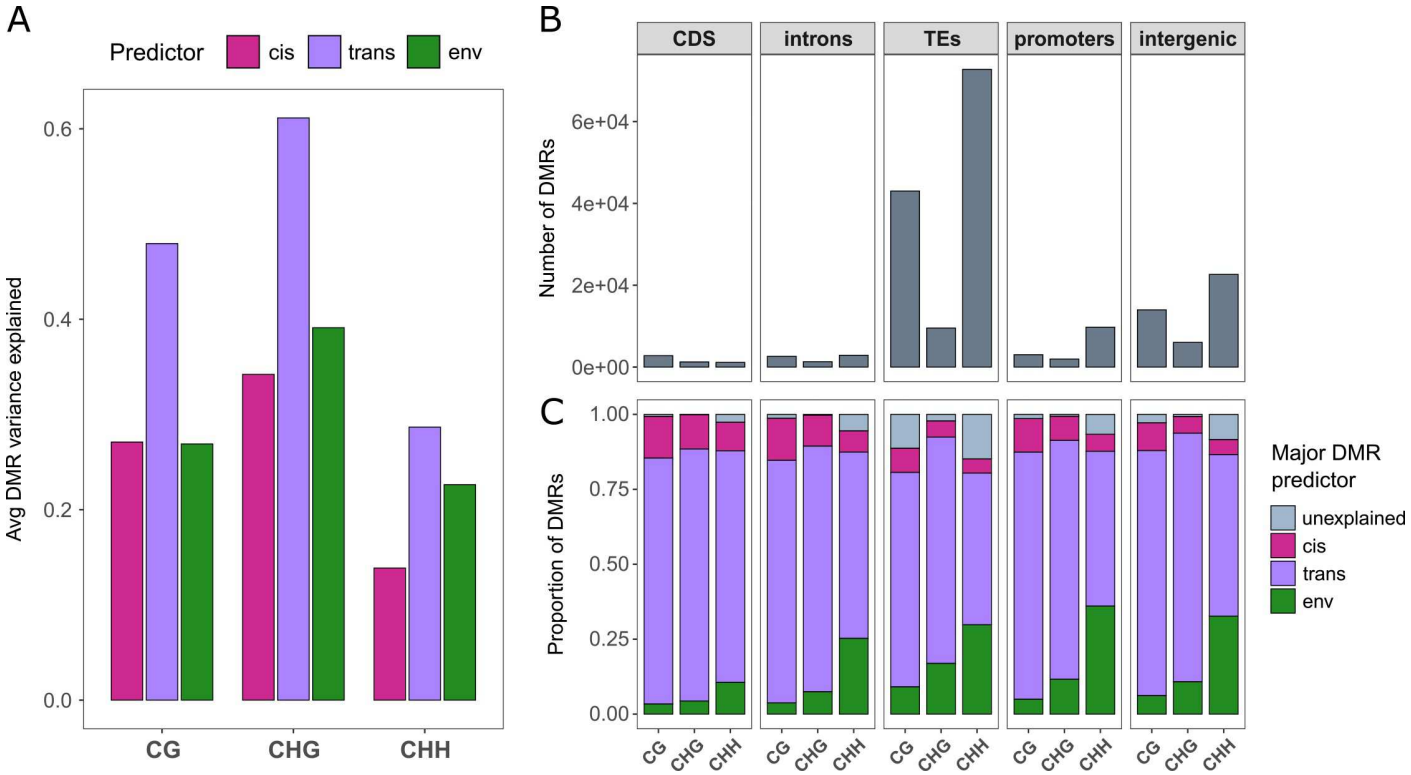


Fig 5. Genetic versus environmental predictors of DMR variance. (A) The variance in DMR weighted methylation explained by genetic similarity in *cis*, genetic similarity in *trans* and climatic similarity, averaged across all DMRs. (B) The number of DMRs identified in different genomic features and sequence contexts, and (C) the fractions of these individual DMRs where *cis*-variation, *trans*-variation or climatic variation are the major predictors. DMRs where none of the three predictors explained >10% of the variance are classified as "unexplained".

https://doi.org/10.1371/journal.pgen.1010452.g005

sometimes genetic variation in *cis* or climatic distance, too, could be the strongest predictor. To study this more systematically, we classified all DMRs based on their strongest predictor, and we found that the fraction of DMRs in which climate was a stronger predictor of methylation variance than any of the genetic distances increased from CG to CHG to CHH (Fig 5C). In CHH, 25–30% of all DMRs had climatic distance as their strongest predictor. To find out if *cis*-, *trans*- and climate-predicted DMRs were enriched close to genes responsible for different functions, we ran separate GO enrichment analyses for the genes neighbouring these three classes of DMRs. However, only for the *trans*-predicted DMRs we found significant enrichment of a few GO terms (S4 Fig), while there were none for the other two DMR classes.

Discussion

Understanding natural epigenetic variation requires combining large-scale surveys of natural populations with high-resolution genomics and environmental data. Here, we studied European populations of *T*. *arvense* to assess how climate of origin and genetic background shaped their heritable DNA methylation variation. We found epigenetic population structure and confirmed the genomic patterns of methylation of the *T*. *arvense* genome [36] in a large natural collection. Most importantly, both genetic background and climate of origin were significantly associated with methylation variation, but their relative predictive power varied depending on DNA sequence context.

Our analysis of population structure detected two main clades, one composed of lines from all surveyed countries and a smaller one with almost exclusively lines from Sweden. A latitudinal gradient was also clear within the larger clade. The epigenetic population structure generally resembled the genetic one, with decreasing degrees of similarity from CG to CHG and to CHH sequence contexts (Fig 1B). These differences between contexts might reflect their different stability, caused by differences in the maintenance machineries [13,14] and possibly different proportions of genetic versus environmental control. Moreover, mCG shows stronger geographic patterns in genes and promoters than in TEs, possibly indicating a higher stability or selection for this kind of methylation (S1 Fig).

Across all lines, we calculated a global weighted methylation of 16.9%, which is high in the Brassicaceae family [20], particularly in comparison to *A*. *thaliana* (5.8%) [12]. The high global methylation is related to the high TE content of the *T*. *arvense* genome (~60%)[36], but also to a higher CHH methylation (12.3% across all lines) than it is known for most other angiosperms [20]. The levels of CG and CHG methylation (47.4% and 14.2% across all lines), in contrast, are more similar to other Brassicaceae [20]. As expected, we found that methylation was very unevenly distributed not only between sequence contexts, but also between genomic features, with high levels of methylation particularly in CG context, and in TEs and intergenic regions (Fig 2A). Gene body methylation was generally very low, with lines carrying on average ~93% of the CDS unmethylated in all contexts (S3 Table), and the results were similar, albeit much less extreme, for promoters (Fig 2B)[36]. When methylated, CDS were usually methylated in all contexts, showing TE-like patterns, while CG-only methylation, typical of many housekeeping genes in other species [20], was almost completely absent (S2 Fig). This uncommonly low gbM is present in other Brassicaceae [20], in particular in the close relative *Eutrema salsugineum* and might have evolved before speciation between Thlaspi and Eutrema [20,36]. Although the loss of *CHROMOMETHYLTRANSFERASE 3 (CMT3)* was previously associated to the loss of gene body methylation [51], this gene is expressed, although possibly mutated, in Thlaspi. If *CMT3* is indeed mutated in Thlaspi, the mutation is likely to affect all lines equally, since we found no variants neighbouring *CMT3* associated with variation in the

number of gbM genes. Instead a significant peak in Scaffold_7, pointing towards other genes, might explain this variation (S2C Fig). TE-like methylated genes, which are usually pseudogenes in many species, were enriched for some constitutive functions and were about eight times more likely than average to overlap with TEs. This might indicate that the extensive TE expansion that occurred in the Thlaspi genome also affected some housekeeping genes, without compromising viability (S2 Fig). Overall these findings suggest that gene body methylation in *T*. *arvense* differs from most previously studied plant species [20].

To understand the genetic basis of methylation variation in *T*. *arvense*, we used GWA analyses, testing for associations between DNA sequence variation and average methylation levels in different sequence contexts and genomic features. With a strict Bonferroni correction, we did not detect any significant genetic variants, which probably resulted from a combination of our moderate number of only 207 sequenced lines, the nested sampling design, and the high number of tests (compared to *A*. *thaliana*) in a ~500 Gb genome. However, for some methylation phenotypes, we found strong enrichment of *a priori* candidates neighbouring genes known to play a role in DNA methylation from *A*. *thaliana* studies, and this indicates that many of our top peaks are likely to be true positives (Figs 3 and S3). Examples include the peaks detected next to the genes *AGO9*, *DRM3* and *LIG1*, which are all part of the DNA methylation machinery of *A*. *thaliana* [13,14], and which were also among our *a priori* candidates (S5 Table). In addition to these 'expected' candidates, we found several additional peaks next to genes that were indirectly linked to DNA methylation, with predicted functions such as histone acetylation, DNA repair and ubiquitination (S6 Table). The latter in particular is a post-translational modification which was previously shown to affect methylation in several ways [28–33]. These new candidate genes were not in our *a priori* list, which explains the drop of enrichment at high -log(p) in several GWA analyses (Figs 3B and S3). Our results show that while there appears to be partial overlap in the genetic control of DNA methylation between *T*. *arvense* and *A*. *thaliana*, there are also important differences. Some of our strongest candidates have not been associated with DNA methylation before, particularly not in natural populations. Functional characterization of these "new" candidates will be necessary to confirm our findings and understand the mechanisms of action of these genes.

Finally, some interesting associations warrant further exploration and could uncover functional differences with *A*. *thaliana* in the methylation machineries of different sequence contexts. For example we find a peak for mCG, next to a *DRM3* orthologue, involved in RdDM and non-CG methylation maintenance in Arabidopsis [46–48], and vice versa a peak for mCHH of promoters and TEs right next (3kb upstream) to an orthologue of the mCG maintenance methyltransferase *MET1*. On the contrary, the high similarity between mCHG and mCHH in regard to their genetic basis, as shown by the strong overlap of GWA results, seems to be a common feature in the plant kingdom [13,14].

Natural epigenetic variation was not only associated with genetic background in our study, but also with climate of origin. These correlations were generally much stronger than those with latitude or longitude, which supports the idea that the observed correlations reflect adaptive processes and not just the combination of epigenetic drift and isolation-by-distance. Specifically, we found average methylation to be positively correlated with mean temperature but negatively with temperature variability (Fig 4). Our field survey particularly captured the cold end of the distribution range of *T*. *arvense* (Mean Annual Temp. 6.5–11.1˚C). Previous studies showed that cold can induce DNA demethylation in plants [52–54] and that demethylation in turn can be associated with expression of cold-resistance genes and increased freezing tolerance [55,56]. The observed negative correlations between methylation and temperature might therefore reflect adaptation to cold and the fact that we captured the cold end of the distribution. This interpretation is further supported by the fact that correlations with minimum

temperatures were generally stronger than with bioclimatic variables capturing maximum temperature (Fig 4) and explains why a similar study found negative correlations between temperature and methylation in Arabidopsis accessions sampled on a range including many warmer locations [12]. The negative relationship between DNA methylation and temperature variability (Mean Diurnal Range and Temperature Annual Range) is more challenging to interpret, as there have so far been no experimental tests manipulating environmental variability in temperature. However, lower DNA methylation is often associated with lower genome stability [57,58], and it is conceivable that in fluctuating and thus less predictable environmental conditions, lower genome stability and higher transposon activity could be adaptive. Supporting this hypothesis, Arabidopsis *cmt2* mutants with slightly lower and more variable CHH methylation in TEs, were shown to be more common in regions with high temperature seasonality [59]. Finally, we did not find any association between DNA methylation and the precipitation of the population origins. However, this may largely be a result of our latitudinal sampling design, which maximized temperature but not precipitation variation. None of our sampling sites were particularly dry or particularly wet/oceanic (Annual Prec. 475–869 mm).

To better understand the predictive power of climate of origin versus genetic background, we finally analysed the variance in methylation levels of individual DMRs. We found that, across all DMRs, genetic variation in *trans* generally explained more DMR variation than climatic variation or genetic variation in *cis*. However, there was a trend from CG to CHG to CHH that the explanatory power of climate increased relative to that of genetic background (Fig 5A). In CHH, climate was the strongest predictor of methylation variation in over one quarter of the individual DMRs; in promoters this was true for even 35% of the DMRs (Fig 5B). These results further support the idea that methylation variation, particularly in CHG and CHH, is not only involved in plant responses to short-term stress [17] but also in longer-term environmental adaptation. Moreover, the observation that sometimes climate was the strongest predictor, indicates that at least part of the climate-methylation associations could be independent of DNA sequence variation [5]. Clearly, further work is needed to support these speculations, in particular high-resolution analyses that disentangle the genomic versus epigenomic basis of relevant phenotypes related to climatic tolerances. We attempted to get some hints of the functional basis of the observed genomic-methylation and climate-methylation relationships by analysing GO enrichment in the neighbouring genes of *trans*-, *cis*- and environmentally-associated DMRs, and we found some enrichment, mostly related to housekeeping functions, for *trans*-DMRs, but none for *cis*- and environmentally-associated DMRs (S4 Fig). However, the functional annotation had GO terms for only less than half of our candidate genes, so our GO enrichment analysis had rather limited power.

In summary, our study is the first large-scale investigation of DNA methylation variation in natural plant populations beyond the Arabidopsis model. We found that *T*. *arvense* natural DNA methylation variation is shaped by genetic and environmental factors, and that the relative contributions of the two drivers vary strongly between sequence contexts. Methylation variation in CG is generally the most similar to, and best predicted by, genetic variation. Moving to CHG and particularly CHH, the genetic determination decreases making environmental determination relatively higher. Our results thus indicate that DNA methylation could play a role in the large-scale environmental adaptation of *T*. *arvense*. Further experimental research, in particular dissecting adaptive phenotypes, is necessary to corroborate this hypothesis. There are currently efforts underway to develop *T*. *arvense* into a new biofuel and winter cover crop [37–41], and any insights into the genomic basis of climate and other environmental adaptation will be highly relevant to these efforts, particularly to deal with future climates.

Materials and methods

Sampling and plant growth

In July 2018, we collected *T*. *arvense* seeds from 36 natural populations in six European regions, spanning from southern France to central Sweden, and used them to conduct a common environment experiment in Tübingen, Germany. The experiment started at the end of August 2018 and lasted about two months. Upon sowing 207 lines in 9x9 cm pots filled with soil, we stratified them for 10 d at 4˚C in the dark. We then transferred the seeds to a glasshouse and transplanted seedlings to individual pots upon germination. The glasshouse had a 15/9 h light/dark cycle (6 a.m. to 9 p.m.) with temperature and humidity conditions averaging 18˚C and 30% at night and 22˚C and 25% during the day. External conditions influenced these parameters, resembling natural growing conditions. 46 d after the end of the stratification period, we collected the 3rd or 4th true leaf and snap-froze it in liquid nitrogen.

Library preparation and sequencing

Using the DNeasy Plant Mini Kit (Qiagen, Hilden, DE), we extracted DNA from disrupted leaf tissue obtained from the 3rd or 4th true leaf. For each sample, we sonicated (Covaris) 300 ng of genomic DNA to a mean fragment size of ~350 bp and used the resulting DNA for both genomic and bisulfite libraries. The NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs) was used for library preparation and was combined with EZ-96 DNA Methylation-Gold MagPrep (ZYMO) for bisulfite libraries. Briefly, the procedure involved: i) end repair and 3' adenylation of sonicated DNA fragments, ii) NEBNext adaptor ligation and U excision, iii) size selection with AMPure XP Beads (Beckman Coulter, Brea, CA), iv) splitting DNA for bisulfite (2/3) and genomic (1/3) libraries, v) bisulfite treatment and cleanup of bisulfite libraries, vi) PCR enrichment and index ligation using Kapa HiFi Hot Start Uracil + Ready Mix (Agilent) for bisulfite libraries (14 cycles) and NEBNext Ultra II Q5 Master Mix for genomic libraries (4 cycles), vii) final size selection and cleanup. Finally, we sequenced paired-end for 150 cycles. Genomic libraries were sequenced on Illumina NovaSeq 6000 (Illumina, San Diego, CA), while bisulfite libraries were sequenced on HiSeq X Ten (Illumina, San Diego, CA).

Variant calling, filtering and imputation

Base calling and demultiplexing of raw sequencing data were performed by Novogene using the standard Illumina pipeline. After quality and adaptor trimming using cutadapt v2.6 (M. Martin 2011), we aligned reads to the reference genome [36] with BWA-MEM v0.7.17 [60]. We then performed variant calling with GATK4 v4.1.8.1 [61,62] following the best practices for Germline short variant discovery (https://gatk.broadinstitute.org/hc/en-us/articles/360035535932-Germline-short-variant-discovery-SNPs-Indels-). Briefly, we marked duplicates with MarkDuplicatesSpark and ran HaplotypeCaller to obtain individual sample GVCF files. We combined individual GVCF files running GenomicsDBImport and GenotypeGVCFs successively and parallelizing by scaffold, obtaining single-scaffold multisample vcf files. We then re-joined these files with GatherVcfs. Upon assessment of quality parameters distributions, we removed low quality variants using VariantFiltration with different filtering 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). Using vcftools v0.1.16 [63], we further filtered scaffolds with less than three variants and variants with multiple alleles or more than 10% missing values. Prior to imputation, we only applied a mild Minor Allele Frequency (MAF) > 0.01 filtering not to

reduce imputation accuracy [64]. Imputation with BEAGLE 5.1 [65] recovered the few missing genotype calls left, outputting a complete multisample vcf file.

Methylation analysis

The EpiDiverse WGBS pipeline is specifically designed for bisulfite reads mapping and methylation calling in non-model species (https://github.com/EpiDiverse/wgbs) [66]. We used it to perform quality control (FastQC), base quality and adaptor trimming (cutadapt), bisulfite aware mapping (erne-bs5), duplicates detection (Picard MarkDuplicates), alignment statistics and methylation calling (Methyldackel). In the mapping step, we only retained uniquely-mapping reads longer than 30bp. The pipeline outputs context-specific (CG, CHG and CHH) individual-sample bedGraph files, which we filtered for coverage > 3 and combined in multisample unionbed files with methylated/total read counts for every position and sample (we used custom scripts and bedtools [67]). We retained all cytosines with coverage > 3 in at least 75% of the lines and used this dataset for all subsequent analyses.

For describing general patterns of methylation, we calculated weighted methylation as the fraction between all methylated and all total read counts at every cytosine included in the calculation [43]. In this way we also calculated the bisulfite non-conversion rates, including all cytosines with coverage > 10 [2] in two regions of Scaffold_364 (51–60.5 KB and 95–110 KB), selected for high similarity to chloroplast DNA and confidently unmethylated. For analyses of variation between lines (GWA and correlation with climate) we used mean methylation, which is obtained by calculating the methylation of each position first (methylated/total read count) and then averaging all positions included in the calculation [43]. Weighted methylation corrects for coverage, but is highly influenced by structural and copy number variants, which are likely abundant in a species with such a high TE content [36]. As we were interested in true variation of methylation levels, mean methylation was more suited for comparing methylation of whole genomic features between lines.

To extract the mean and weighted methylation of genomic features, we intersected (bedtools) [67] unionbed files with genomic features (genes, CDS, introns, TEs, promoters and intergenic regions) and averaged methylation of all intersected cytosines. For introns, we only included regions annotated as intronic on both strands. We also extracted weighted methylation of individual CDS, promoters and TEs across all samples and plotted their distributions. We then used this information to calculate the mean methylation of genes, excluding lowly methylated ones (average mC $< 5\%$ across all lines) and used it for GWA. For PCA, we used the R [68] function prcomp(). Genome wide PCAs were only based on positions without missing values as these were already a large amount (always > 1 million). Instead when restricting to genomic features we allowed for 2% NAs and imputed these with the "missMDA" R package [69] to include a larger amount of positions (always > 0.8 million). Nevertheless comparison of PCA plots with and without imputation gave very similar results.

Gene Body Methylation classification

To test whether genes were methylated in their CDS, in any of the sequence contexts, we adopted a method from previous authors [20,45]. First we used a binomial test to determine, for each cytosine in CDS, whether it had significantly higher methylation than expected from bisulfite non-conversion rates ($P < 0.01$). We then computed the fraction of methylated cytosines in all CDS and lines, separately for each sequence context. Finally we tested if the fraction of methylated cytosines of each individual CDS, was higher than the average of all CDS, with a one-sided binomial test. In other words, we tested whether a specific CDS had a higher density of methylated positions than all CDS on average. Upon correcting for multiple testing with the p.adjust() R function [68], we considered "methylated" CDS with FDR<0.05. We restricted the analysis to genes with at least

10 covered cytosines (coverage > 3) in each sequence context, for at least 90% of the lines. If a CDS had less than 6 cytosines covered in a line, we coded it as a missing value. Such analysis revealed the methylation status of 22703 genes in each line and sequence context. We defined as "gbM", genes with mCG FDR < 0.05, and mCHG and mCHH FDR > 0.05. We defined as "teM", genes with mCHG or mCHH FDR < 0.05 [12]. For GO enrichment analysis we used genes consistently methylated, i.e. methylated in at least 70% of the lines.

Population genetic and GWA analysis

For basic genetic population structure analysis, including PCA plots and generation of the IBS matrix, we applied a mild MAF filtering (MAF>0.01) and performed variants pruning with PLINK v1.90b6.12 [70], using a window of 50 variants, sliding by five and a maximum LD of 0.8. Upon this filtering, we also used PLINK to generate the IBS matrix used in several analyses to correct for population structure or for DMRs variance decomposition. For PCA, we used the R [68] function prcomp().

We ran GWA analyses for multiple phenotypes using a custom script based on the R package "rrBLUP" [71], which allows to run mixed models correcting for population structure with the above-mentioned IBS matrix. We used biallelic variants and applied a MAF > 0.04 cutoff. For Manhattan and QQplots we used the "qqman" package [72], calculating the genome-wide significance threshold according to Sobota et al. (2015) [49]. We ran GWA analyses using each average methylation context (CG, CHG and CHH) feature (global, CDS, introns, TEs, promoters and intergenic regions) combination as phenotype. For genes we also calculated mean methylation of methylated genes, excluding lowly methylated ones (average methylation > 5% across all lines), ending up with a total of 24 methylation phenotypes (S4 Table). Since a few samples had higher than usual non-conversion rates (S2 Table), leading to an overestimation of their average methylation, we calculated, for each individual sample, the surplus non-conversion rate from a baseline of 0.6%, and subtracted it from the mean methylation values. The baseline of 0.6% allowed us to correct only the ~20% of samples with highest non-conversion rates. Occasionally, we observed a positive correlation between mean methylation and coverage across lines, probably due to library preparation bias. In these cases we fit a linear model to the data using the logarithm of coverage (from bam files), as this gave the best fit in all cases, and used the residuals for GWA analysis. Finally, we applied Inverse Normal Transformation to mean methylation phenotypes that deviated strongly from normality. A list of all methylation phenotypes used and corrections and transformations applied, can be found in S4 Table.

With the double aim of validating GWA results and comparing with previous *A*. *thaliana* studies, we performed enrichment of variants neighbouring *a priori* candidate genes, according to the method established by Atwell et al. (2010) [50]. We made a few additions to the methylation candidate gene list used by Kawakatsu et al. (2016) [12], kindly provided by the authors, extracted all *T*. *arvense* orthologues that we could retrieve from orthofinder [73] analysis and used them for our *a priori* candidate genes list (S5 Table). Briefly, we attributed "*a priori* candidate" status to all variants within 20kb from genes in the list and calculated enrichment for increasing -log(p) thresholds as the ratio between Observed frequency (sign. *a priori* candidates/sign. variants) and Background frequency (total *a priori* candidates/total variants). Using the same formula adopted by Atwell et al. (2010) [50], we additionally calculated an upper bound for the FDR among candidates.

Climate-methylation correlations

To obtain bioclimatic variables for the 25 years predating the experiment, we downloaded temperature and precipitation variables from the "E-OBS daily gridded meteorological data for

Europe" database (v21.0), freely available on the Copernicus website [44]. All downstream analyses were conducted in R [68]. We extracted data for our population locations with the "ncdf4" package [74], calculated monthly averages and extracted bioclimatic variables with "dismo" [75]. Finally, we averaged bioclimatic variables from 1994 to 2018, the year of collection (S7 Table). To test for climatic patterns in methylation, we ran mixed models for all mean methylation variables (the same as we used for GWA) and bioclimatic variables combinations, using the relmatLmer() function from the R package "lme4qtl" [76] and correcting for population structure using the same IBS matrix used for GWA analysis.

DMR calling

The EpiDiverse toolkit [66] includes a DMR pipeline based on metilene [77], which calls DMRs between all possible pairwise comparisons between user-defined groups. We used this tool to call DMRs using the 36 populations as groups, a minimum coverage of five ($cov > 4$) and default values for all other parameters. We complemented the pipeline with a custom downstream workflow to obtain DMRs for the whole collection from comparison-specific DMRs. Briefly, since the pipeline output had an enrichment of short and close DMRs (particularly in CHH), we joined all comparison-specific DMRs that were closer than 146bp and had the same directionality (higher methylation in the same group). 146bp was chosen for consistency with the pipeline fragmentation parameter. We then merged DMRs from all pairwise comparisons (bedtools) [67] in a unique file and re-extracted weighted methylation of the resulting regions from all samples. Finally, we filtered DMRs with a minimum methylation difference of 20% (CG) or 15% (CHG and CHH) in at least 5% of the samples. This ensured to select DMRs with variability at the level of the whole collection.

DMR variance decomposition

To quantify the variance in methylation explained by *cis*-variants, *trans*-variants, and by the environment, we ran three mixed models for each individual DMR using the marker_h2() function from the R package "heritability" [78]. Each model had one random factor matrix, capturing one of the three predictors. For *cis* we used an IBS matrix generated with PLINK v1.90b6.12 [70] from variants within 50kb from the DMR middle point. For *trans* we used the same IBS matrix used for all other analyses, described in the previous chapter. For the environment we calculated the Euclidean distance between locations, based on all Bioclimatic Variables averaged over 25 years before the sampling (1994–2018), and further reversed and normalized the matrix to obtain a similarity matrix in a 0 to 1 range. To summarize the results we: i) averaged *cis*, *trans* and environment explained variance across all DMRs and ii) classified each DMR based on the mayor predictor.

Supporting information

S1 Fig. PCA plots of all 207 lines. (A) Complement to Fig 1B with latitude-coloured PCA plots for the missing PC. (B) latitude-coloured PCA plots based on methylation of specific genomic features (genes, TEs and promoters).
(PDF)

S2 Fig. Genes methylated in each context, GO enrichment analysis and GWA. (A) Venn diagram of the number of genes methylated in each context in at least 70% of the lines, which were also used for the GO enrichment. Genes methylated only in CG are labelled as "gbM", genes methylated in either CHG or CHH as "TE-like" [12]. (B) GO enrichment analysis of methylated genes corresponding to (A). Only significant results for GO terms with minimum

gene count of four are reported. GO categories are: Biological Process (BP), Cellular Component (CC) and Molecular Function (MF). (C) GWA for number of gbM genes, including Manhattan plot, enrichment of *a-priori* candidates and qqplot.
(PDF)

S3 Fig. Complete methylation GWA results. Manhattan plots, enrichment of *a priori* candidate variants and QQplots for all mean methylation phenotypes. more5met: mean methylation of genes with methylation > 5% across all lines. The genome-wide significance (horizontal red lines), was calculated based on unlinked variants as in Sobota et al. (2015) [49], the suggestive-line (blue) corresponds to–log(p) = 5. Top variants are labelled with the neighbouring genes potentially affecting methylation.
(PDF)

S4 Fig. GO enrichment analysis of genes neighbouring trans-DMRs. Genes neighbouring (2kb max) *cis*, *trans* and *env*-DMRs were used for individual GO term enrichment analysis, but only the *trans*-DMRs gene set was enriched for any significant term.
(PDF)

S1 Table. Geographic locations of all *T. arvense* populations. Geographic coordinates, elevation and size of all populations.
(PDF)

S2 Table. Mapping statistics. Number of deduplicated mapped reads, average coverage and non-conversion rates calculated from chloroplast DNA. WGS: Whole Genome Sequencing; WGBS: Whole Genome Bisulfite Sequencing.
(CSV)

S3 Table. Number of genes methylated in each line. Numbers and fractions of genes per line methylated in each sequence context, in CG only (gbM) and in either CHG or CHH (TEm) [12].
(XLSX)

S4 Table. List of all mean methylation variables used for GWA and climate correlations. Coverage correction indicates that, prior to GWA, residuals were extracted from a linear model with log(coverage) as predictor. INT indicates Inverse Normal Transformation. more5-met: Mean methylation of genes with methylation > 5% across all lines.
(PDF)

S5 Table. List of *Thlaspi arvense a priori* candidate genes. *T. arvense* genes and the respective *A. thaliana* orthologues with known roles in methylation. We used this list for the enrichment of a priori candidate variants performed upon GWA.
(CSV)

S6 Table. GWA candidate genes. List of all genes located within 15kb from variants significant to -log(p) > 5, including methylation phenotypes where the association was found, *a priori* candidate status and relevant functional putative roles. Genes with predicted function possibly affecting methylation are highlighted in bold.
(XLSX)

S7 Table. Bioclimatic variables. Bioclimatic variables used in this study, obtained from monthly averages extracted from the Copernicus programme website [44] and averaged for 1993–2018.
(CSV)

Acknowledgments

We thank the entire EpiDiverse network for its amazing support and discussions, in particular Adrián Contreras-Garrido for providing orthofinder results and discussing analysis, and Bárbara Díez Rodríguez and Iris Sammarco for really useful suggestions. We thank Detlef Weigel for his input on data analysis, and Magnus Nordborg and Eriko Sasaki for their feedback on GWA analysis and for sharing their list of candidate genes. Finally, we thank Anupoma Troyee and Valentina Vaglia for helping with sampling, Sabine Silberhorn, Christiane Karasch-Wittmann, Eva Schloter, Julia Rafalski and Elodie Kugler for the greenhouse experiment, and Katharina Jandrasits for help with library preparation. For computing, we acknowledge Prof. Peter Stadler at the University of Leipzig and David Langenberger from ecSeq, for hosting the EpiDiverse servers. We also acknowledge the High Performance and Cloud Computing Group at the Zentrum für Datenverarbeitung of the University of Tübingen for managing the BinAC server.

Author Contributions

Conceptualization: J. F. Scheepens, Claude Becker, Oliver Bossdorf.

Data curation: Dario Galanti, Adam Nunn, Isaac Rodríguez-Arévalo.

Formal analysis: Dario Galanti.

Funding acquisition: Claude Becker, Oliver Bossdorf.

Investigation: Dario Galanti, Daniela Ramos-Cruz.

Methodology: Dario Galanti, Oliver Bossdorf.

Project administration: Oliver Bossdorf.

Resources: Daniela Ramos-Cruz, Claude Becker.

Software: Dario Galanti, Adam Nunn.

Supervision: J. F. Scheepens, Claude Becker, Oliver Bossdorf.

Validation: Dario Galanti.

Visualization: Dario Galanti.

Writing – original draft: Dario Galanti, Oliver Bossdorf.

Writing – review & editing: Dario Galanti, Daniela Ramos-Cruz, Adam Nunn, Isaac Rodríguez-Arévalo, J. F. Scheepens, Claude Becker, Oliver Bossdorf.

References

1. Schmitz RJ, Schultz MD, Lewsey MG, O'Malley RC, Urich MA, Libiger O, et al. Transgenerational epigenetic instability is a source of novel methylation variants. Science. 2011 Oct 21; 334(6054):369–73. https://doi.org/10.1126/science.1212959 PMID: 21921155
2. Becker C, Hagmann J, Müller J, Koenig D, Stegle O, Borgwardt K, et al. Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. Nature. 2011 Dec; 480(7376):245–9.
3. Lämke J, Bäurle I. Epigenetic and chromatin-based mechanisms in environmental stress adaptation and stress memory in plants. Genome Biology. 2017 Jun 27; 18(1):124. https://doi.org/10.1186/s13059-017-1263-6 PMID: 28655328
4. He Y, Li Z. Epigenetic Environmental Memories in Plants: Establishment, Maintenance, and Reprogramming. Trends in Genetics. 2018 Aug 22; Available from: http://www.sciencedirect.com/science/article/pii/S0168952518301276

5. 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 Dec 1; 20 (12):1576–90. https://doi.org/10.1111/ele.12858 PMID: 29027325
6. Schmid MW, Heichinger C, Schmid DC, Guthörl D, Gagliardini V, Bruggmann R, et al. Contribution of epigenetic variation to adaptation in Arabidopsis. Nature Communications. 2018 Oct 25; 9(1):4446. https://doi.org/10.1038/s41467-018-06932-5 PMID: 30361538
7. Münzbergová Z, Latzel V, Šurinová M, Hadincová V. DNA methylation as a possible mechanism affecting ability of natural populations to adapt to changing climate. Oikos. 2019; 128(1):124–34.
8. Paun O, Bateman RM, Fay MF, Hedrén M, Civeyrel L, Chase MW. Stable Epigenetic Effects Impact Adaptation in Allopolyploid Orchids (Dactylorhiza: Orchidaceae). Mol Biol Evol. 2010 Nov 1; 27 (11):2465–73. https://doi.org/10.1093/molbev/msq150 PMID: 20551043
9. Lira-Medeiros CF, Parisod C, Fernandes RA, Mata CS, Cardoso MA, Ferreira PCG. Epigenetic Variation in Mangrove Plants Occurring in Contrasting Natural Environment. PLOS ONE. 2010 Apr 26; 5(4): e10326. https://doi.org/10.1371/journal.pone.0010326 PMID: 20436669
10. Gugger PF, Fitz-Gibbon S, Pellegrini M, Sork VL. Species-wide patterns of DNA methylation variation in Quercus lobata and their association with climate gradients. Molecular Ecology. 2016 Apr 1; 25 (8):1665–80. https://doi.org/10.1111/mec.13563 PMID: 26833902
11. Dubin MJ, Zhang P, Meng D, Remigereau M-S, Osborne EJ, Paolo Casale F, et al. DNA methylation in Arabidopsis has a genetic basis and shows evidence of local adaptation. eLife. 2015;4. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4413256/
12. Kawakatsu T, Huang SC, Jupe F, Sasaki E, Schmitz RJ, Urich MA, et al. Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions. Cell. 2016 Jul 14; 166(2):492–505. https://doi.org/10.1016/j.cell.2016.06.044 PMID: 27419873
13. Law JA, Jacobsen SE. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nature Reviews Genetics. 2010 Mar; 11(3):204–20. https://doi.org/10.1038/nrg2719 PMID: 20142834
14. Zhang H, Lang Z, Zhu J-K. Dynamics and function of DNA methylation in plants. Nature Reviews Molecular Cell Biology. 2018 May 21;1. https://doi.org/10.1038/s41580-018-0016-z PMID: 29784956
15. Cao X, Aufsatz W, Zilberman D, Mette MF, Huang MS, Matzke M, et al. Role of the DRM and CMT3 Methyltransferases in RNA-Directed DNA Methylation. Current Biology. 2003 Dec 16; 13(24):2212–7. https://doi.org/10.1016/j.cub.2003.11.052 PMID: 14680640
16. Stroud H, Do T, Du J, Zhong X, Feng S, Johnson L, et al. Non-CG methylation patterns shape the epigenetic landscape in Arabidopsis. Nat Struct Mol Biol. 2014 Jan; 21(1):64–72. https://doi.org/10.1038/nsmb.2735 PMID: 24336224
17. Liu J, He Z. Small DNA Methylation, Big Player in Plant Abiotic Stress Responses and Memory. Frontiers in Plant Science. 2020; 11. Available from: https://www.frontiersin.org/article/10.3389/fpls.2020.595603 https://doi.org/10.3389/fpls.2020.595603 PMID: 33362826
18. Schmitz RJ, Schultz MD, Urich MA, Nery JR, Pelizzola M, Libiger O, et al. Patterns of population epigenomic diversity. Nature. 2013 Mar; 495(7440):193–8. https://doi.org/10.1038/nature11968 PMID: 23467092
19. Seymour DK, Koenig D, Hagmann J, Becker C, Weigel D. Evolution of DNA methylation patterns in the Brassicaceae is driven by differences in genome organization. PLoS Genet. 2014 Nov; 10(11): e1004785. https://doi.org/10.1371/journal.pgen.1004785 PMID: 25393550
20. Niederhuth CE, Bewick AJ, Ji L, Alabady MS, Kim KD, Li Q, et al. Widespread natural variation of DNA methylation within angiosperms. Genome Biology. 2016 Sep 27; 17:194. https://doi.org/10.1186/s13059-016-1059-0 PMID: 27671052
21. Schmitz RJ, Lewis ZA, Goll MG. DNA Methylation: Shared and Divergent Features across Eukaryotes. Trends in Genetics. 2019 Nov 1; 35(11):818–27. https://doi.org/10.1016/j.tig.2019.07.007 PMID: 31399242
22. Hazarika RR, Serra M, Zhang Z, Zhang Y, Schmitz RJ, Johannes F. Molecular properties of epimutation hotspots. Nat Plants. 2022 Jan 27;1–11.
23. Cubas P, Vincent C, Coen E. An epigenetic mutation responsible for natural variation in floral symmetry. Nature. 1999 Sep; 401(6749):157–61. https://doi.org/10.1038/43657 PMID: 10490023
24. Manning K, Tör M, Poole M, Hong Y, Thompson AJ, King GJ, et al. A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening. Nature Genetics. 2006 Aug; 38(8):948–52. https://doi.org/10.1038/ng1841 PMID: 16832354
25. Martin A, Troadec C, Boualem A, Rajab M, Fernandez R, Morin H, et al. A transposon-induced epigenetic change leads to sex determination in melon. Nature. 2009 Oct; 461(7267):1135–8. https://doi.org/10.1038/nature08498 PMID: 19847267

26. Cortijo S, Wardenaar R, Colomé-Tatché M, Gilly A, Etcheverry M, Labadie K, et al. Mapping the Epigenetic Basis of Complex Traits. Science. 2014 Mar 7; 343(6175):1145–8. https://doi.org/10.1126/science.1248127 PMID: 24505129

27. 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 Genetics. 2019 dic; 15(12): e1008492. https://doi.org/10.1371/journal.pgen.1008492 PMID: 31887137

28. Bostick M, Kim JK, Estève P-O, Clark A, Pradhan S, Jacobsen SE. UHRF1 Plays a Role in Maintaining DNA Methylation in Mammalian Cells. Science. 2007 Sep 21; 317(5845):1760–4. https://doi.org/10.1126/science.1147939 PMID: 17673620

29. Sharif J, Muto M, Takebayashi S, Suetake I, Iwamatsu A, Endo TA, et al. The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. Nature. 2007 Dec; 450(7171):908–12.

30. Kraft E, Bostick M, Jacobsen SE, Callis J. ORTH/VIM proteins that regulate DNA methylation are functional ubiquitin E3 ligases. The Plant Journal. 2008; 56(5):704–15. https://doi.org/10.1111/j.1365-313X.2008.03631.x PMID: 18643997

31. Kim J, Kim JH, Richards EJ, Chung KM, Woo HR. Arabidopsis VIM Proteins Regulate Epigenetic Silencing by Modulating DNA Methylation and Histone Modification in Cooperation with MET1. Molecular Plant. 2014 Sep 1; 7(9):1470–85. https://doi.org/10.1093/mp/ssu079 PMID: 25009302

32. Chen J, Liu J, Jiang J, Qian S, Song J, Kabara R, et al. F-box protein CFK1 interacts with and degrades de novo DNA methyltransferase in Arabidopsis. New Phytologist. 2021; 229(6):3303–17. https://doi.org/10.1111/nph.17103 PMID: 33216996

33. Wang J, Qiu Z, Wu Y. Ubiquitin Regulation: The Histone Modifying Enzyme's Story. Cells. 2018 Aug 27; 7(9):118. https://doi.org/10.3390/cells7090118 PMID: 30150556

34. Gáspár B, Bossdorf O, Durka W. Structure, stability and ecological significance of natural epigenetic variation: a large-scale survey in Plantago lanceolata. New Phytologist; 0(0). Available from: https://nph.onlinelibrary.wiley.com/doi/abs/10.1111/nph.15487 PMID: 30222201

35. 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. Available from: https://www.frontiersin.org/articles/10.3389/fgene.2015.00004/full https://doi.org/10.3389/fgene.2015.00004 PMID: 25688257

36. Nunn A, Rodríguez-Arévalo I, Tandukar Z, Frels K, Contreras-Garrido A, Carbonell-Bejerano P, 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 Biotechnology Journal. 2022 Aug; 1–20. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/pbi.13775 PMID: 34990041

37. Tsogtbaatar E, Cocuron J-C, Sonera MC, Alonso AP. Metabolite fingerprinting of pennycress (Thlaspi arvense L.) embryos to assess active pathways during oil synthesis. J Exp Bot. 2015 Jul; 66(14):4267–77. https://doi.org/10.1093/jxb/erv020 PMID: 25711705

38. 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 Apr 1; 22(2):121–31. https://doi.org/10.1093/dnares/dsu045 PMID: 25632110

39. Claver A, Rey R, López MV, Picorel R, Alfonso M. Identification of target genes and processes involved in erucic acid accumulation during seed development in the biodiesel feedstock Pennycress (Thlaspi arvense L.). Journal of Plant Physiology. 2017 Jan 1; 208:7–16. https://doi.org/10.1016/j.jplph.2016.10.011 PMID: 27889523

40. Chopra R, Johnson EB, Daniels E, McGinn M, Dorn KM, Esfahanian M, et al. Translational genomics using Arabidopsis as a model enables the characterization of pennycress genes through forward and reverse genetics. The Plant Journal. 2018; 96(6):1093–105. https://doi.org/10.1111/tpj.14147 PMID: 30394623

41. Chopra R, Johnson EB, Emenecker R, Cahoon EB, Lyons J, Kliebenstein DJ, et al. Progress toward the identification and stacking of crucial domestication traits in pennycress. Plant Biology; 2019 Apr. Available from: http://biorxiv.org/lookup/doi/10.1101/609990

42. Geng Y, Guan Y, Qiong L, Lu S, An M, Crabbe MJC, et al. Genomic analysis of field pennycress (Thlaspi arvense) provides insights into mechanisms of adaptation to high elevation. BMC Biology. 2021 Jul 22; 19(1):143. https://doi.org/10.1186/s12915-021-01079-0 PMID: 34294107

43. Schultz MD, Schmitz RJ, Ecker JR. 'Leveling' the playing field for analyses of single-base resolution DNA methylomes. Trends Genet. 2012 Dec 1; 28(12):583–5. https://doi.org/10.1016/j.tig.2012.10.012 PMID: 23131467

44. Copernicus Climate Change Service. E-OBS daily gridded meteorological data for Europe from 1950 to present derived from in-situ observations. ECMWF; 2020 [cited 2022 Feb 9]. Available from: https://cds.climate.copernicus.eu/doi/10.24381/cds.151d3ec6

45. Takuno S, Gaut BS. Gene body methylation is conserved between plant orthologs and is of evolutionary consequence. Proc Natl Acad Sci. 2013 Jan 29; 110(5):1797–802. https://doi.org/10.1073/pnas.1215380110 PMID: 23319627

46. Henderson IR, Deleris A, Wong W, Zhong X, Chin HG, Horwitz GA, et al. The De Novo Cytosine Methyltransferase DRM2 Requires Intact UBA Domains and a Catalytically Mutated Paralog DRM3 during RNA–Directed DNA Methylation in Arabidopsis thaliana. PLOS Genetics. 2010 ott; 6(10):e1001182. https://doi.org/10.1371/journal.pgen.1001182 PMID: 21060858

47. Zhong X, Hale CJ, Nguyen M, Ausin I, Groth M, Hetzel J, et al. DOMAINS REARRANGED METHYLTRANSFERASE3 controls DNA methylation and regulates RNA polymerase V transcript abundance in Arabidopsis. PNAS. 2015 Jan 20; 112(3):911–6. https://doi.org/10.1073/pnas.1423603112 PMID: 25561521

48. Costa-Nunes P, Kim JY, Hong E, Pontes O. The cytological and molecular role of DOMAINS REARRANGED METHYLTRANSFERASE3 in RNA-dependent DNA methylation of Arabidopsis thaliana. BMC Research Notes. 2014 Oct 14; 7(1):721. https://doi.org/10.1186/1756-0500-7-721 PMID: 25316414

49. Sobota RS, Shriner D, Kodaman N, Goodloe R, Zheng W, Gao Y-T, et al. Addressing population-specific multiple testing burdens in genetic association studies. Ann Hum Genet. 2015 Mar; 79(2):136–47. https://doi.org/10.1111/ahg.12095 PMID: 25644736

50. Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, et al. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature. 2010 Jun; 465(7298):627–31. https://doi.org/10.1038/nature08800 PMID: 20336072

51. Bewick AJ, Ji L, Niederhuth CE, Willing EM, Hofmeister BT, Shi X, et al. On the origin and evolutionary consequences of gene body DNA methylation. Proc Natl Acad Sci. 2016 Aug 9; 113(32):9111–6. https://doi.org/10.1073/pnas.1604666113 PMID: 27457936

52. Steward N, Kusano T, Sano H. Expression of ZmMET1, a gene encoding a DNA methyltransferase from maize, is associated not only with DNA replication in actively proliferating cells, but also with altered DNA methylation status in cold-stressed quiescent cells. Nucleic Acids Research. 2000 Sep 1; 28(17):3250–9. https://doi.org/10.1093/nar/28.17.3250 PMID: 10954592

53. Conde D, Le Gac A-L, Perales M, Dervinis C, Kirst M, Maury S, et al. Chilling-responsive DEMETER-LIKE DNA demethylase mediates in poplar bud break. Plant Cell Environ. 2017 Oct; 40(10):2236–49. https://doi.org/10.1111/pce.13019 PMID: 28707409

54. Lai Y-S, Zhang X, Zhang W, Shen D, Wang H, Xia Y, et al. The association of changes in DNA methylation with temperature-dependent sex determination in cucumber. Journal of Experimental Botany. 2017 May 17; 68(11):2899–912. https://doi.org/10.1093/jxb/erx144 PMID: 28498935

55. Xie HJ, Li H, Liu D, Dai WM, He JY, Lin S, et al. ICE1 demethylation drives the range expansion of a plant invader through cold tolerance divergence. Molecular Ecology. 2015 Feb 1; 24(4):835–50. https://doi.org/10.1111/mec.13067 PMID: 25581031

56. Xie H, Sun Y, Cheng B, Xue S, Cheng D, Liu L, et al. Variation in ICE1 Methylation Primarily Determines Phenotypic Variation in Freezing Tolerance in Arabidopsis thaliana. Plant and Cell Physiology. 2019 Jan 1; 60(1):152–65. https://doi.org/10.1093/pcp/pcy197 PMID: 30295898

57. Slotkin RK, Martienssen R. Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet. 2007 Apr; 8(4):272–85. https://doi.org/10.1038/nrg2072 PMID: 17363976

58. Zhou D, Robertson KD. Role of DNA Methylation in Genome Stability. Elsevier Inc.; 2016. p. 409–24. Available from: http://www.scopus.com/inward/record.url?scp=85022019460&partnerID=8YFLogxK

59. Shen X, Jonge JD, Forsberg SKG, Pettersson ME, Sheng Z, Hennig L, et al. Natural CMT2 Variation Is Associated With Genome-Wide Methylation Changes and Temperature Seasonality. PLOS Genet. 2014 dic; 10(12):e1004842. https://doi.org/10.1371/journal.pgen.1004842 PMID: 25503602

60. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009 Jul 15; 25(14):1754–60. https://doi.org/10.1093/bioinformatics/btp324 PMID: 19451168

61. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Angel G del, Levy-Moonshine A, et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Current Protocols in Bioinformatics. 2013; 43(1):11.10.1–11.10.33. https://doi.org/10.1002/0471250953.bi1110s43 PMID: 25431634

62. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Auwera GAV der, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. 2018 Jul 24;201178.

63. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011 Aug 1; 27(15):2156–8. https://doi.org/10.1093/bioinformatics/btr330 PMID: 21653522

64. Roshyara NR, Kirsten H, Horn K, Ahnert P, Scholz M. Impact of pre-imputation SNP-filtering on genotype imputation results. BMC Genet. 2014 Aug 12; 15:88. https://doi.org/10.1186/s12863-014-0088-5 PMID: 25112433

65. Browning BL, Zhou Y, Browning SR. A One-Penny Imputed Genome from Next-Generation Reference Panels. The American Journal of Human Genetics. 2018 Sep 6; 103(3):338–48. https://doi.org/10.1016/j.ajhg.2018.07.015 PMID: 30100085

66. Nunn A, Can SN, Otto C, Fasold M, Díez Rodríguez B, Fernández-Pozo N, et al. EpiDiverse Toolkit: a pipeline suite for the analysis of bisulfite sequencing data in ecological plant epigenetics. NAR Genomics and Bioinformatics. 2021 Dec 1; 3(4):lqab106. https://doi.org/10.1093/nargab/lqab106 PMID: 34805989

67. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar 15; 26(6):841–2. https://doi.org/10.1093/bioinformatics/btq033 PMID: 20110278

68. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; 2020. Available from: https://www.R-project.org/

69. Husson F, Josse J. missMDA: Handling Missing Values with Multivariate Data Analysis. 2020 [cited 2022 Feb 1]. Available from: https://CRAN.R-project.org/package=missMDA

70. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics. 2007 Sep 1; 81(3):559–75. https://doi.org/10.1086/519795 PMID: 17701901

71. Endelman J. rrBLUP: Ridge Regression and Other Kernels for Genomic Selection. 2019 [cited 2022 Feb 1]. Available from: https://CRAN.R-project.org/package=rrBLUP

72. Turner S. qqman: Q-Q and Manhattan Plots for GWAS Data. 2021 [cited 2022 Feb 1]. Available from: https://CRAN.R-project.org/package=qqman

73. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biology. 2019 Nov 14; 20(1):238. https://doi.org/10.1186/s13059-019-1832-y PMID: 31727128

74. Pierce D. ncdf4: Interface to Unidata netCDF (Version 4 or Earlier) Format Data Files. 2021 [cited 2022 Feb 1]. Available from: https://CRAN.R-project.org/package=ncdf4

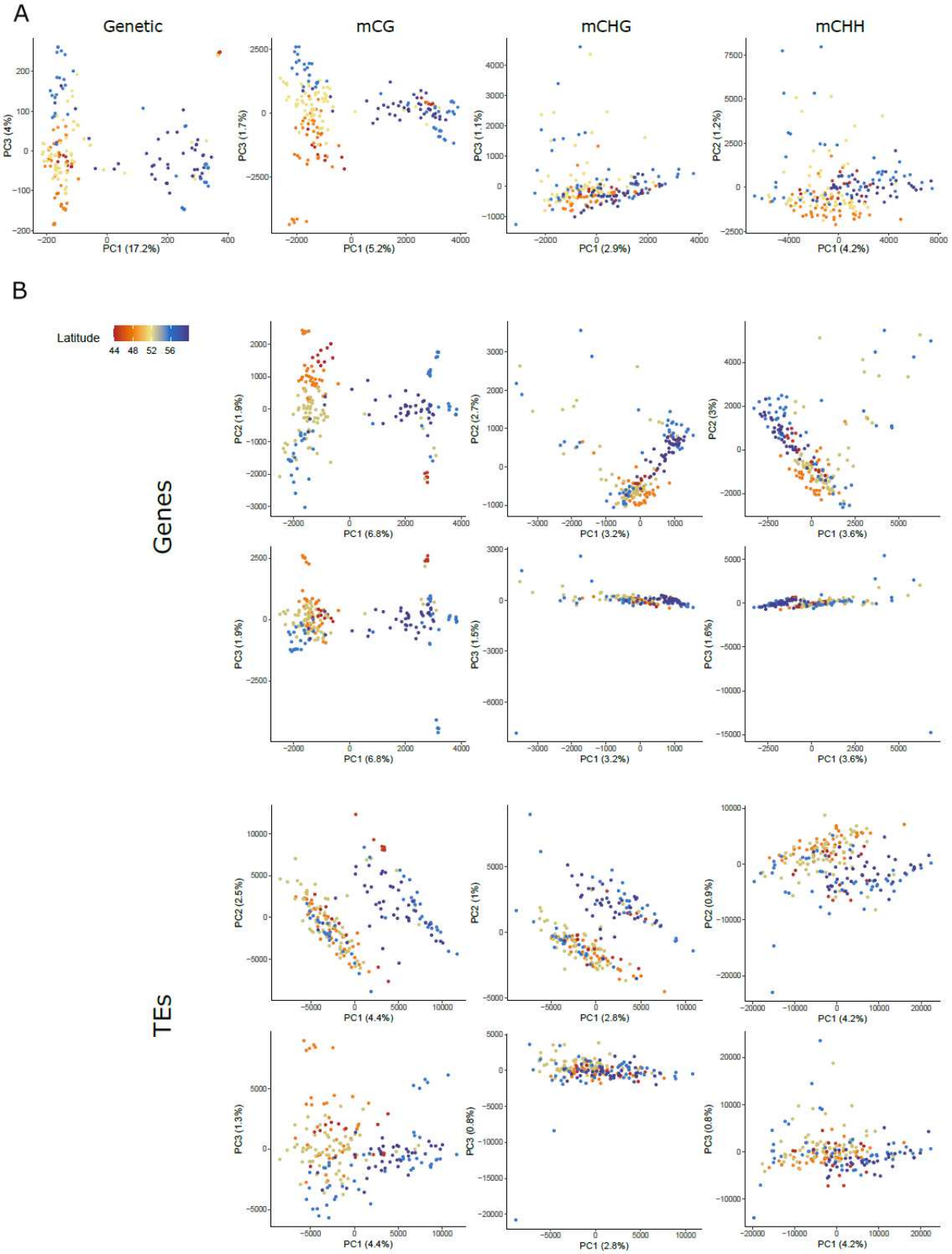
75. Hijmans RJ, Phillips S, Elith JL and J. dismo: Species Distribution Modeling. 2021 [cited 2022 Feb 1]. Available from: https://CRAN.R-project.org/package=dismo

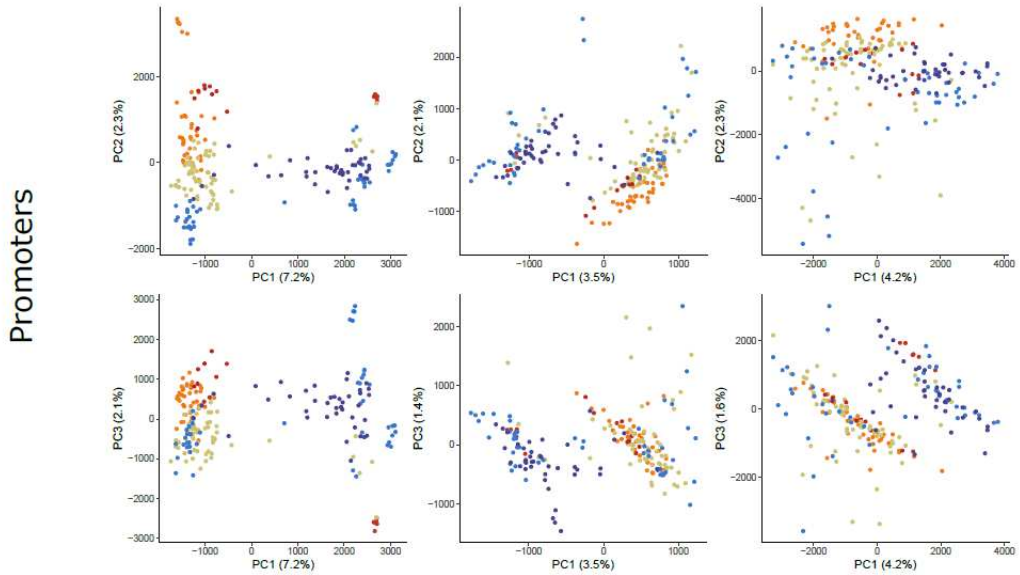
76. Ziyatdinov A, Vázquez-Santiago M, Brunel H, Martinez-Perez A, Aschard H, Soria JM. lme4qtl: linear mixed models with flexible covariance structure for genetic studies of related individuals. BMC Bioinformatics. 2018 Feb 27; 19(1):68. https://doi.org/10.1186/s12859-018-2057-x PMID: 29486711

77. Jühling F, Kretzmer H, Bernhart SH, Otto C, Stadler PF, Hoffmann S. metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. Genome Res. 2016 Feb 1; 26(2):256–62. https://doi.org/10.1101/gr.196394.115 PMID: 26631489

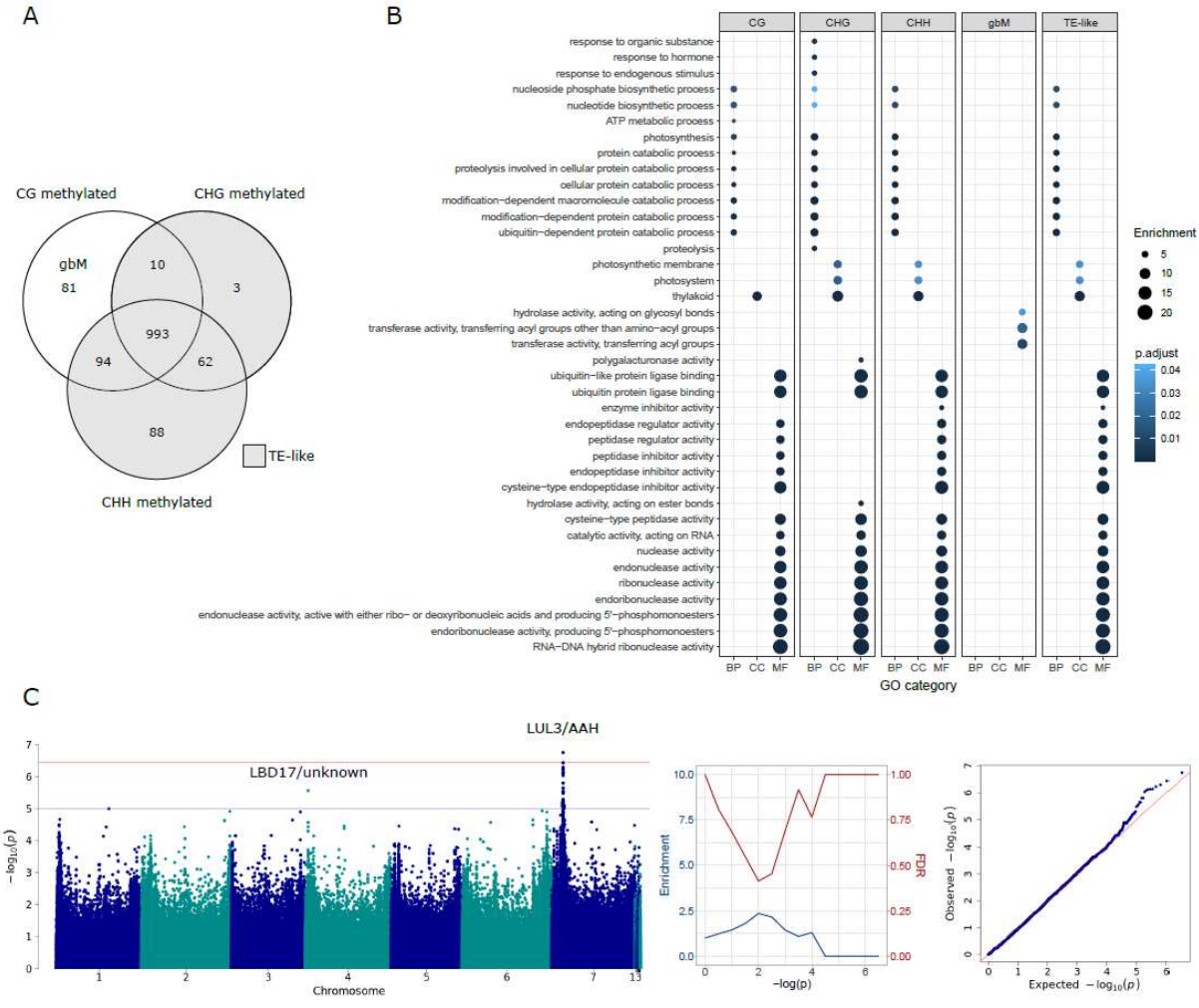
78. Kruijer W, Kooke with a contribution from IW (the internal function pin) C data collected by PF and R. heritability: Marker-Based Estimation of Heritability Using Individual Plant or Plot Data. 2019 [cited 2022 Feb 1]. Available from: https://CRAN.R-project.org/package=heritability

Supporting Information

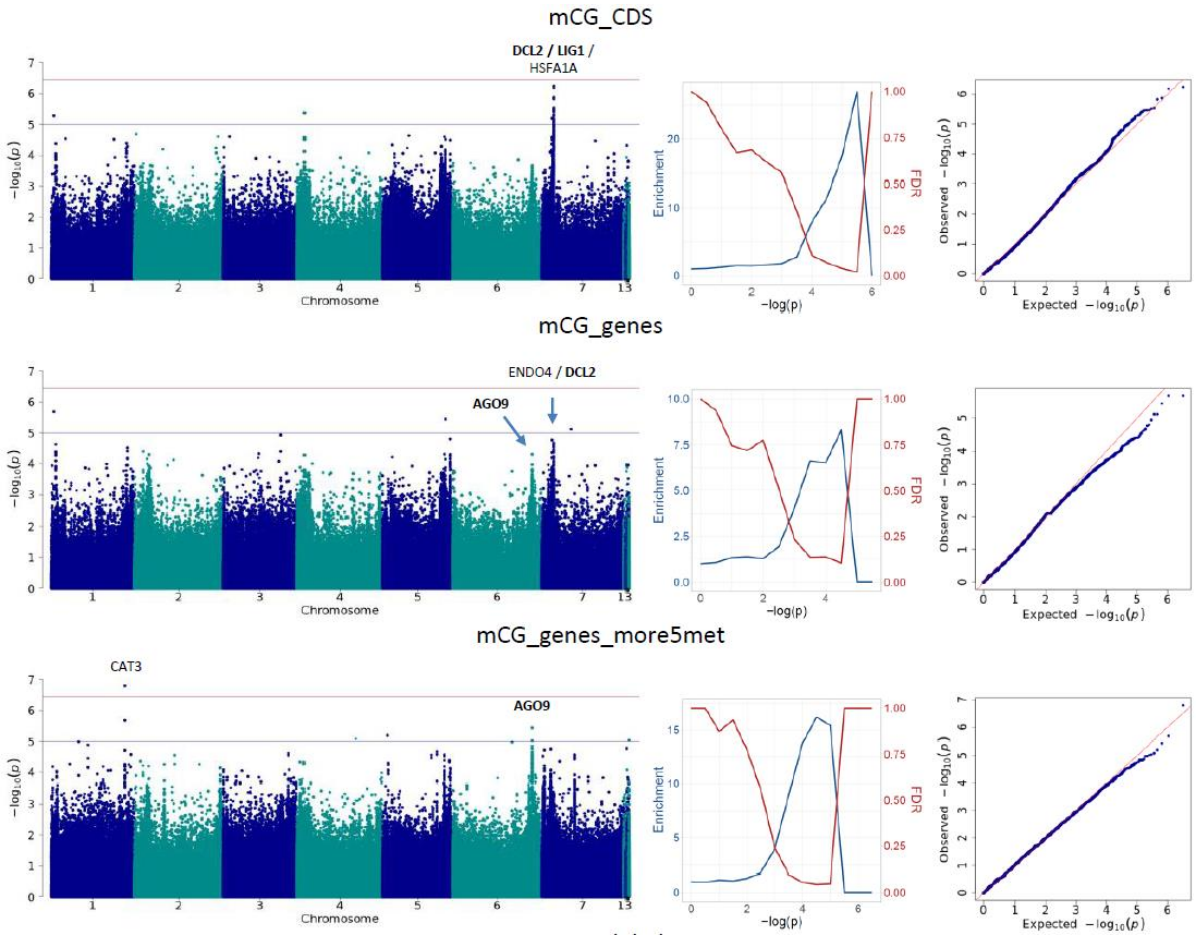




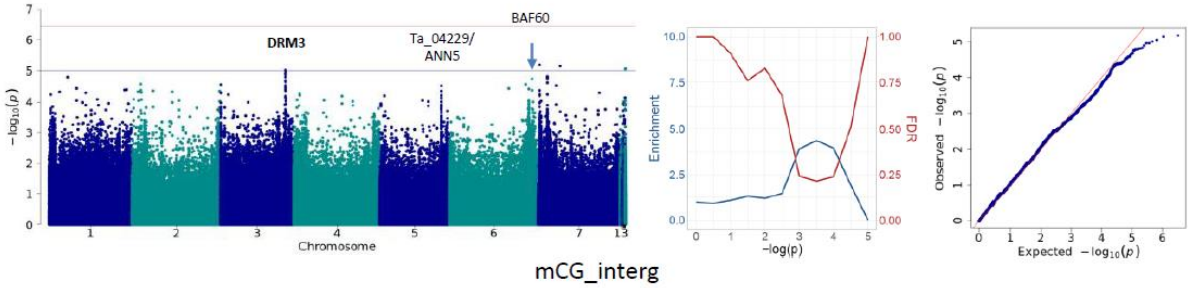
S1 Fig. PCA plots of all 207 lines. (A) Complement to Fig 1B with latitude-coloured PCA plots for the missing PC. (B) latitude-coloured PCA plots based on methylation of specific genomic features (genes, TEs and promoters).



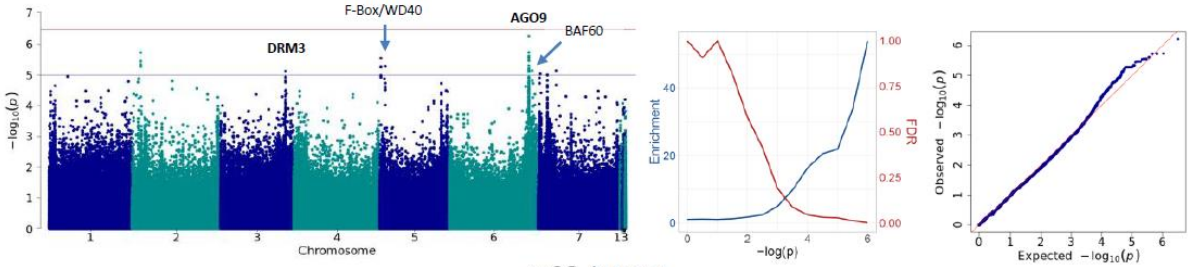
S2 Fig. Genes methylated in each context, GO enrichment analysis and GWA. (A) Venn diagram of the number of genes methylated in each context in at least 70% of the lines, which were also used for the GO enrichment. Genes methylated only in CG are labelled as "gbM", genes methylated in either CHG or CHH as "TE-like" [12]. (B) GO enrichment analysis of methylated genes corresponding to (A). Only significant results for GO terms with minimum gene count of four are reported. GO categories are: Biological Process (BP), Cellular Component (CC) and Molecular Function (MF). (C) GWA for number of gbM genes, including Manhattan plot, enrichment of *a-priori* candidates and qqplot.



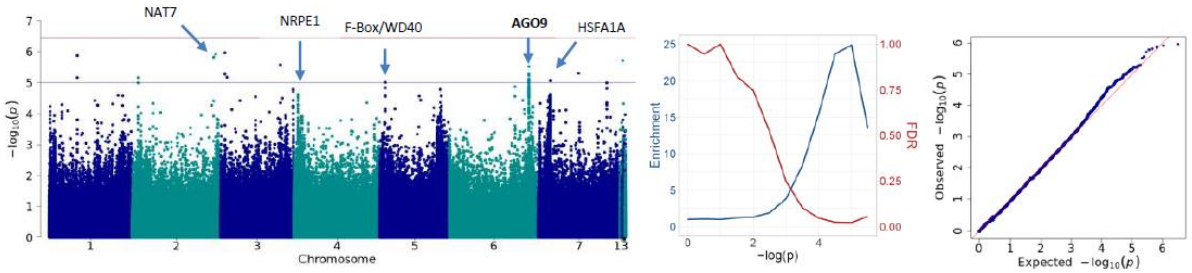
mCG_global



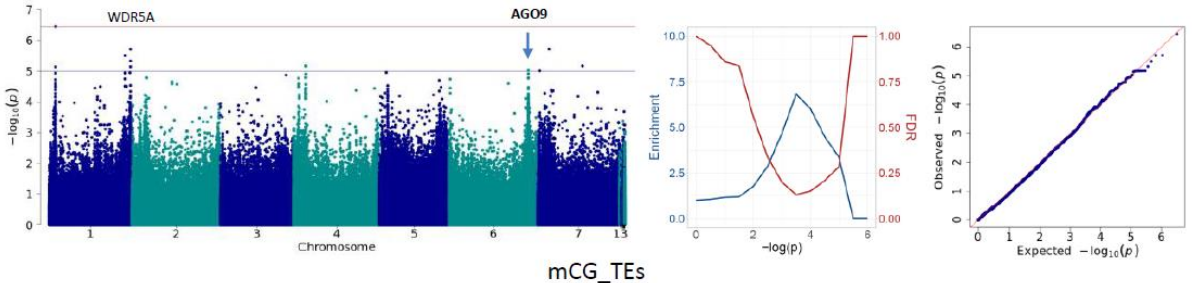
mCG_interg



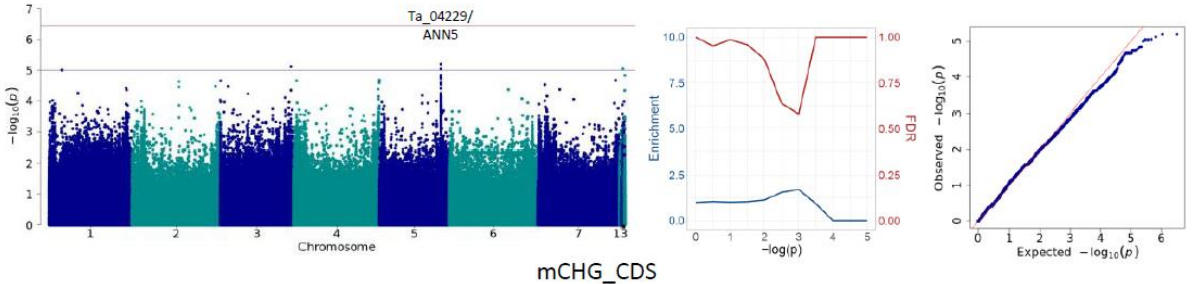
mCG_introns



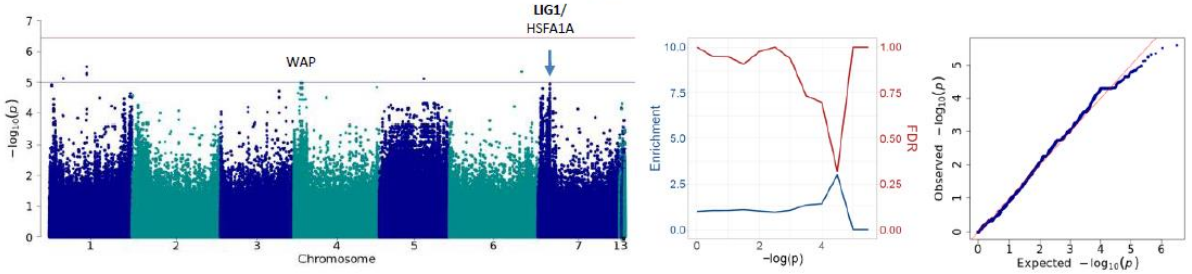
mCG_prom



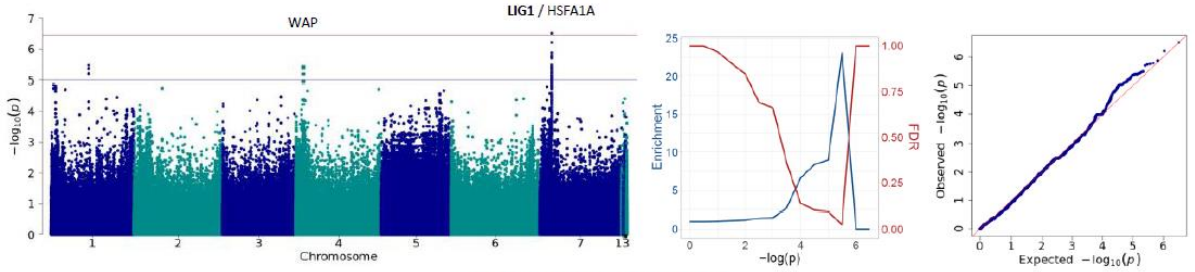
mCG_TEs



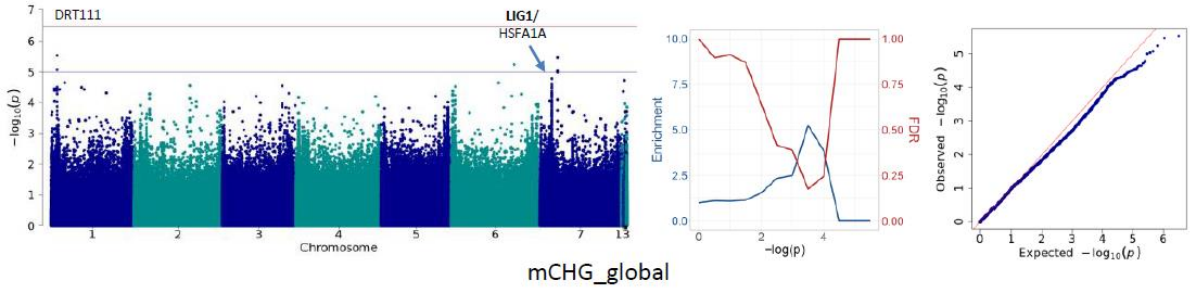
mCHG_CDS



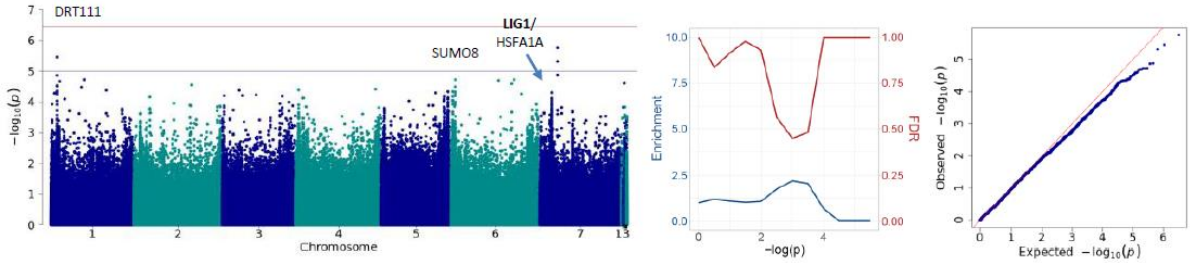
mCHG_genes



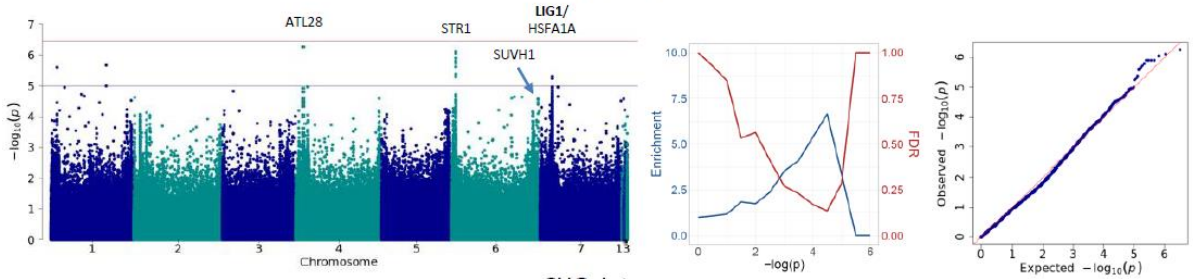
mCHG_genes_more5met



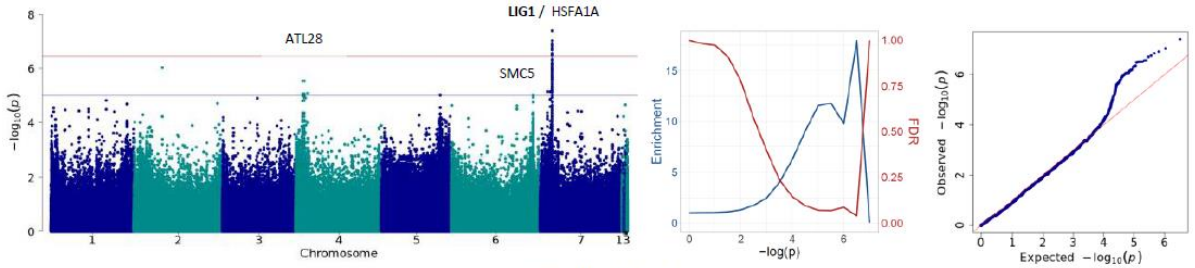
mCHG_global



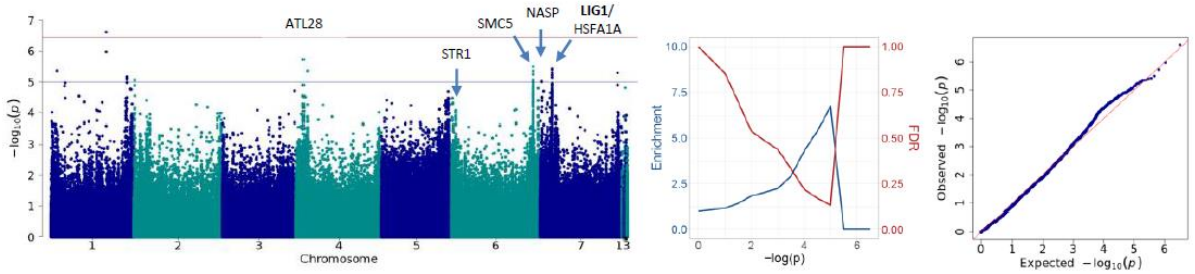
mCHG_interg



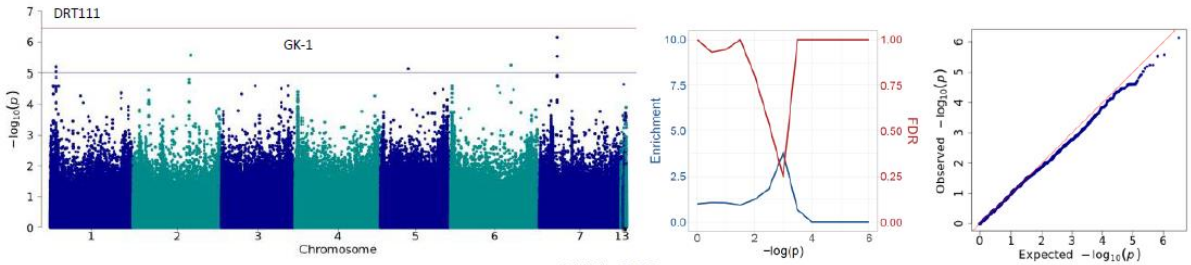
mCHG_introns



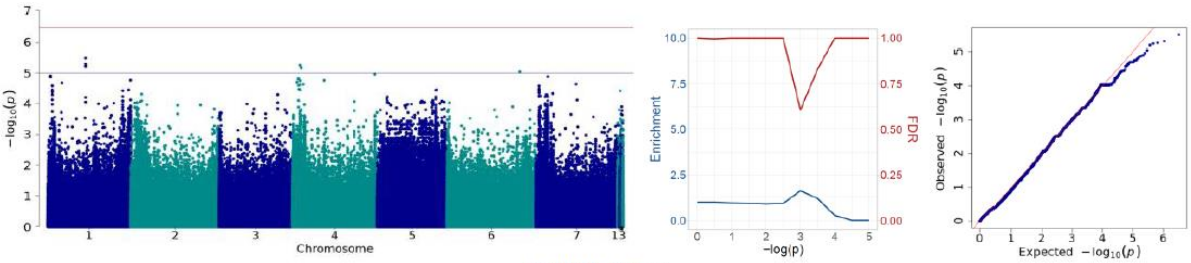
mCHG_prom



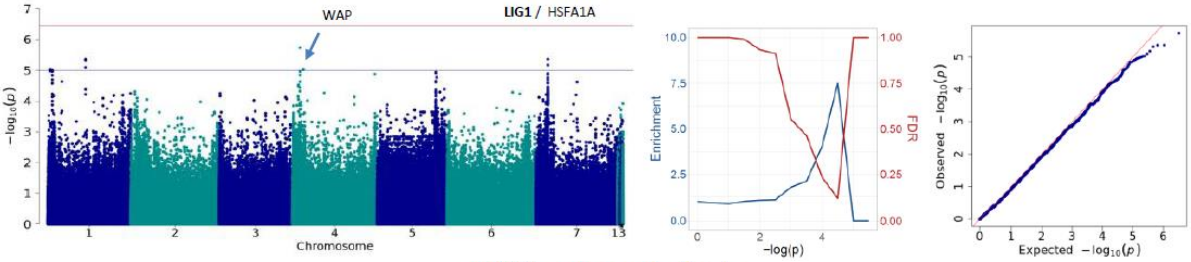
mCHG_TEs



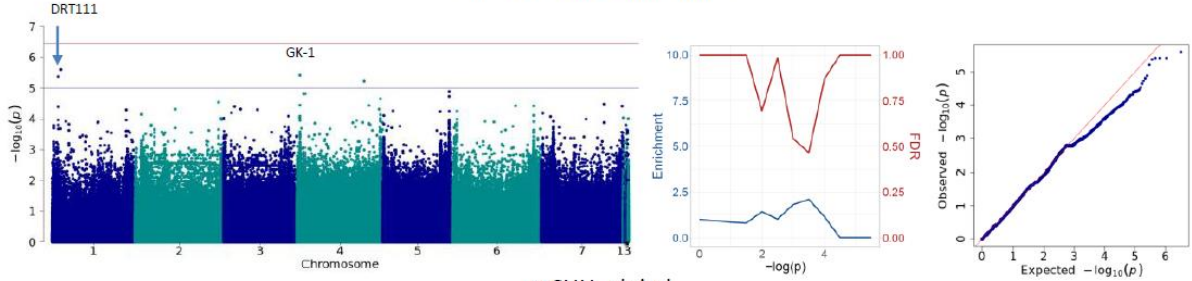
mCHH_CDS



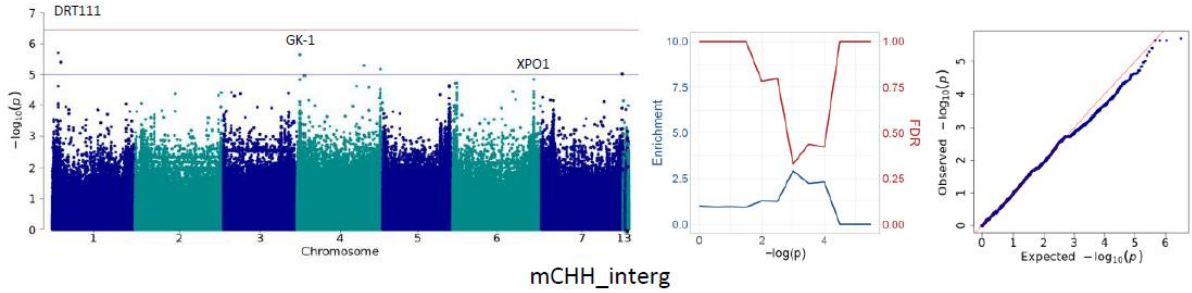
mCHH_genes



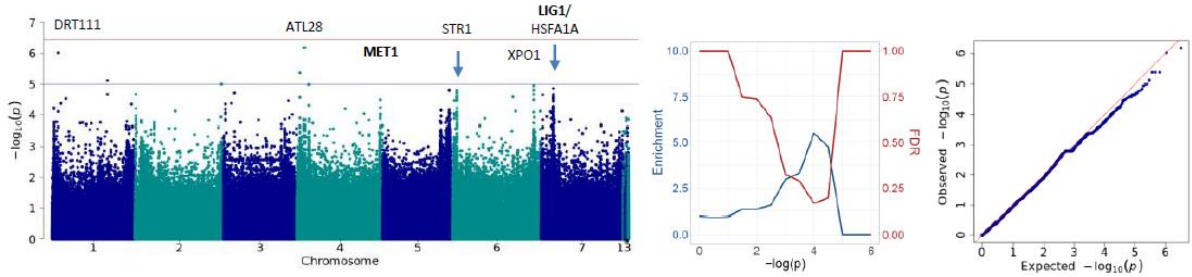
mCHH_genes_more5met

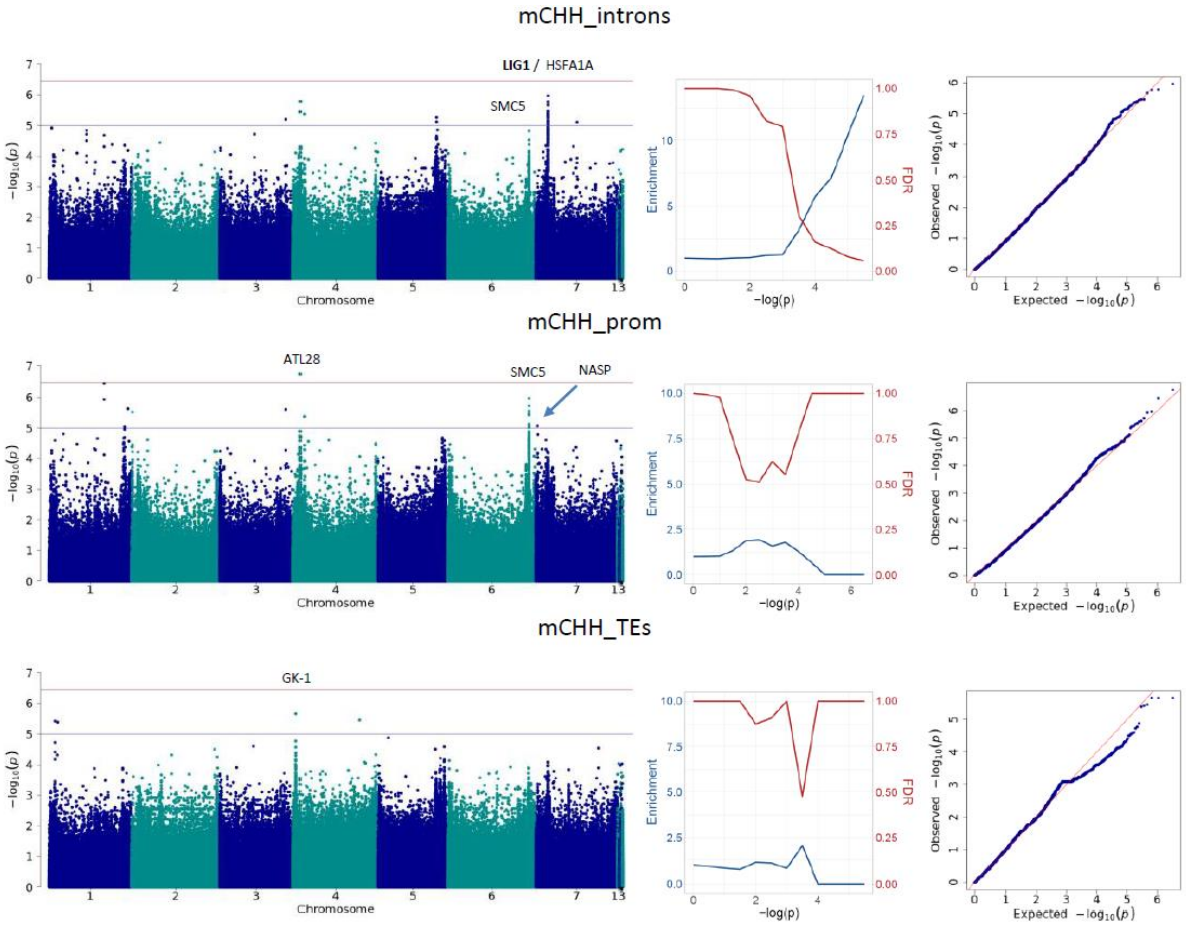


mCHH_global

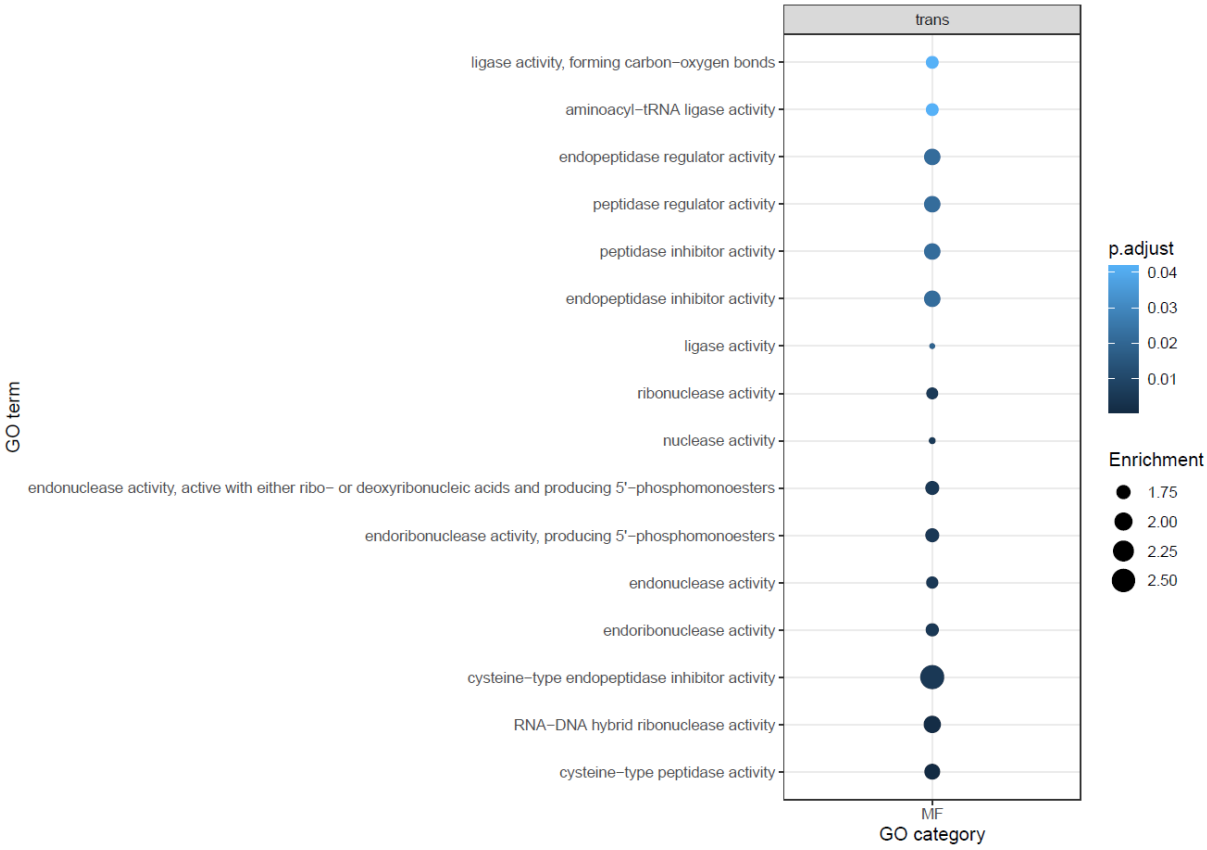


mCHH_interg





S3 Fig. Complete methylation GWA results. Manhattan plots, enrichment of *a priori* candidate variants and QQplots for all mean methylation phenotypes. more5met: mean methylation of genes with methylation > 5% across all lines. The genome-wide significance (horizontal red lines) was calculated based on unlinked variants as in Sobota et al. (2015) [49], the suggestive-line (blue) corresponds to –log(p)=5. Top variants are labelled with the neighbouring genes potentially affecting methylation.



S4 Fig. GO enrichment analysis of genes neighbouring trans-DMRs. Genes neighbouring (2kb max) *cis*, *trans* and *env*-DMRs were used for individual GO term enrichment analysis, but only the *trans*-DMRs gene set was enriched for any significant term.

Region	Population	N° lines	Closest town	Latitude	Longitude	Altitude (m)
France	FR_01	5	Les Rives	43.8508071	3.282875	743
France	FR_02	4	Mostuéjouls	44.2385639	3.1596174	848
France	FR_03	6	Miscon	44.6281220	5.523118	821
South Germany	DE_01	6	Tübingen	48.5402860	9.034686	458
South Germany	DE_02	6	Löffingen	47.8766732	8.427379	708
South Germany	DE_03	6	Braunlingen	47.9202121	8.434509	769
South Germany	DE_04	6	Balingen	48.2802594	8.837293	539
South Germany	DE_06	6	Hirrlingen	48.4121117	8.8835621	431
South Germany	DE_07	6	Empfingen	48.3843302	8.7295661	515
South Germany	DE_08	4	Wittershausen	48.3252288	8.6434947	536
The Netherlands	NL_01	6	Wageningen	51.9550616	5.6372513	8
The Netherlands	NL_02	4	Veenendaal	52.0403204	5.5523601	8
The Netherlands	NL_03	6	Herwen	51.8861228	6.1316716	10
North Germany	DE_09	6	Halle	51.5138186	11.9201814	83
North Germany	DE_10	6	Schwittersdorf	51.5626351	11.7071825	187
North Germany	DE_11	6	Eisleben	51.5409550	11.5963349	221
North Germany	DE_12	6	Halle	51.5457637	11.9575064	90
North Germany	DE_13	6	Plötz	51.6362233	11.9366542	82
North Germany	DE_14	6	Rothen	51.7079562	12.0079172	88
North Germany	DE_15	6	Bossdorf	52.0151260	12.583979	151
North Germany	DE_16	6	Coswig	51.8852458	12.393296	55
South Sweden	SE_01	6	Lund	55.7224293	13.184996	48
South Sweden	SE_02	6	Lund	55.7316952	13.2524585	74
South Sweden	SE_03	6	Lund	55.7729103	13.2571215	21
South Sweden	SE_04	6	Eslöv	55.7483891	13.3845241	24
South Sweden	SE_05	6	Veberöd	55.6372295	13.5000471	34
South Sweden	SE_06	6	Vressel	55.6676921	13.6234294	20
South Sweden	SE_07	6	Vanstad	55.622051	13.835959	81
South Sweden	SE_08	5	Onslunda	55.6085961	14.0608974	108
Central Sweden	SE_09	5	Stockholm	59.3695897	17.9951086	9
Central Sweden	SE_10	6	Sollentuna	59.4340638	17.9218504	23
Central Sweden	SE_11	6	Rotebro	59.4773200	17.8533700	31
Central Sweden	SE_12	5	Uppsala	59.8191923	17.6535739	26
Central Sweden	SE_13	6	Uppsala	59.8447498	17.5101196	30
Central Sweden	SE_14	6	Uppsala	59.908915	17.599931	26
Central Sweden	SE_15	6	Sigtuna	59.652574	17.6884701	15

S1 Table. Geographic locations of all *T. arvense* populations. Geographic coordinates, elevation and size of all populations.

The remaining supplementary tables are too large to be reported here, so please refer to our publication: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1010452#sec018

S2 Table. Mapping statistics. Number of deduplicated mapped reads, average coverage and non-conversion rates calculated from chloroplast DNA. WGS: Whole genome Sequencing; WGBS: Whole Genome Bisulfite Sequencing.

S3 Table. Number of genes methylated in each line. Numbers and fractions of genes per line methylated in each sequence context, in CG only (gbM) and in either CHG or CHH (TEm) [12].

S4 Table. List of all mean methylation variables used for GWA and climate correlations. Coverage correction indicates that, prior to GWA, residuals were extracted from a linear model with log(coverage) as predictor. INT indicates Inverse Normal Transformation. more5met: Mean methylation of genes with methylation > 5% across all lines.

S5 Table. List of *Thlaspi arvense a priori* candidate genes. *T. arvense* genes and the respective *A. thaliana* orthologues with known roles in methylation. We used this list for the enrichment of a priori candidate variants performed upon GWA.

S6 Table. GWA candidate genes. List of all genes located within 15kb from variants significant to -log(p) > 5, including methylation phenotypes where the association was found, *a priori* candidate status and relevant functional putative roles. Genes with predicted function possibly affecting methylation are highlighted in bold.

S7 Table. Bioclimatic variables. Bioclimatic variables used in this study, obtained from monthly averages extracted from the Copernicus programme website [44] and averaged for 1993-2018.

Chapter II

Transposon dynamics in the emerging oilseed crop *Thlaspi arvense*

---

*Adrián Contreras-Garrido, ***Dario Galanti**, Andrea Movilli, Claude Becker, Oliver Bossdorf, Hajk-Georg Drost, Detlef Weigel.

* These authors contributed equally

https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1011141

PLOS GENETICS

RESEARCH ARTICLE

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, Tübingen, Germany, **2** Plant Evolutionary Ecology, University of Tübingen, Tübingen, Germany, **3** LMU Biocenter, Faculty of Biology, Ludwig Maximilians University Munich, Martinsried, Germany, **4** Computational Biology Group, Max Planck Institute for Biology Tübingen,Tübingen, Germany

☯ These authors contributed equally to this work.
* drost@tue.mpg.de(H-GD), weigel@tue.mpg.de (DW)

Check for updates

OPEN ACCESS

Citation: Contreras-Garrido A, Galanti D, Movilli A, Becker C, Bossdorf O, Drost H-G, et al. (2024) Transposon dynamics in the emerging oilseed crop *Thlaspi arvense*. PLoS Genet 20(1): e1011141. https://doi.org/10.1371/journal.pgen.1011141

Editor: Yalong Guo, Institute of Botany, Chinese Academy of Sciences, CHINA

Received: August 9, 2023

Accepted: January 17, 2024

Published: January 31, 2024

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pgen.1011141

Copyright: © 2024 Contreras-Garrido et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Code used for analysis and figures can be found at: https://github.com/acontrerasg/Tarvense_transposon_dynamics. Sequencing reads can be found at the European Nucleotide Archive (ENA) under accession number

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 might contribute to the broad geographic success and adaptive capacity of the emerging oil crop *Thlaspi arvense* (field pennycress). We classified the TE inventory based on a high-quality genome assembly, estimated the age of retrotransposon TE families and comprehensively assessed their mobilization potential. A survey of 280 accessions from 12 regions across the Northern hemisphere allowed us to quantify over 90,000 TE insertion polymorphisms (TIPs). Their distribution mirrored the 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. 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 transcriptome divergence. The number of retrotransposon TIPs is associated with variation at genes related to epigenetic regulation, including an apparent knockout mutation in *BROMODOMAIN AND ATPase DOMAIN-CONTAINING PROTEIN 1* (*BRAT1*), while DNA transposons are associated with variation at the *HSP19* heat shock protein gene. 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 summary

Transposable elements (TEs) are often considered genomic parasites, but they can also generate phenotypic novelty that helps organisms to adapt to new environments. To understand how TEs might contribute to phenotypic diversity and adaptive potential in the emerging oilseed crop *Thlaspi arvense* (field pennycress), we examined the dynamics

PRJEB62093. See S3 Table for details of the datasets. Datasets were uploaded to Zenodo under the DOI: 10.5281/zenodo.6372331. The workflow was based on custom bash and python scripts available at https://github.com/acontrerasg/Tarvense_transposon_dynamics. All the code for short variants calling, filtering and imputation can be found on GitHub (https://github.com/Dario-Galanti/BinAC_varcalling).

Funding: The study was supported by Marie Skłodowska Curie ETN EpiDiverse (EU Horizon 2020 Grant Agreement No. 764965; C.B., O.B., D. W.), the European Research Council (Grant Agreement No. 716823 "FEAR-SAP"; C.B.), the Novo Nordisk Foundation Novozymes 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.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: D.W. holds equity in Computomics, which advises breeders. D.W. advises KWS SE, a plant breeder and seed producer. All the other authors have declared that no competing interests exist.

of TE variation in a geographically diverse sample of this species. By surveying almost 300 wild accessions from North America and Eurasia we discovered over 90,000 polymorphic TE insertions. We identified not only genetic factors that vary between populations and that are associated with TE mobilization, but also TE families that are most likely to generate genetic diversity of interest to breeders.

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 [14–17], 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 (S1 Table), 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 (S2 Table) 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 (S2 Table).

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 Ty1 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 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 (S1 Fig). Next, we focused 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).

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 TE families. To learn how TE mobility has shaped genomes 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 (S3 Table), 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

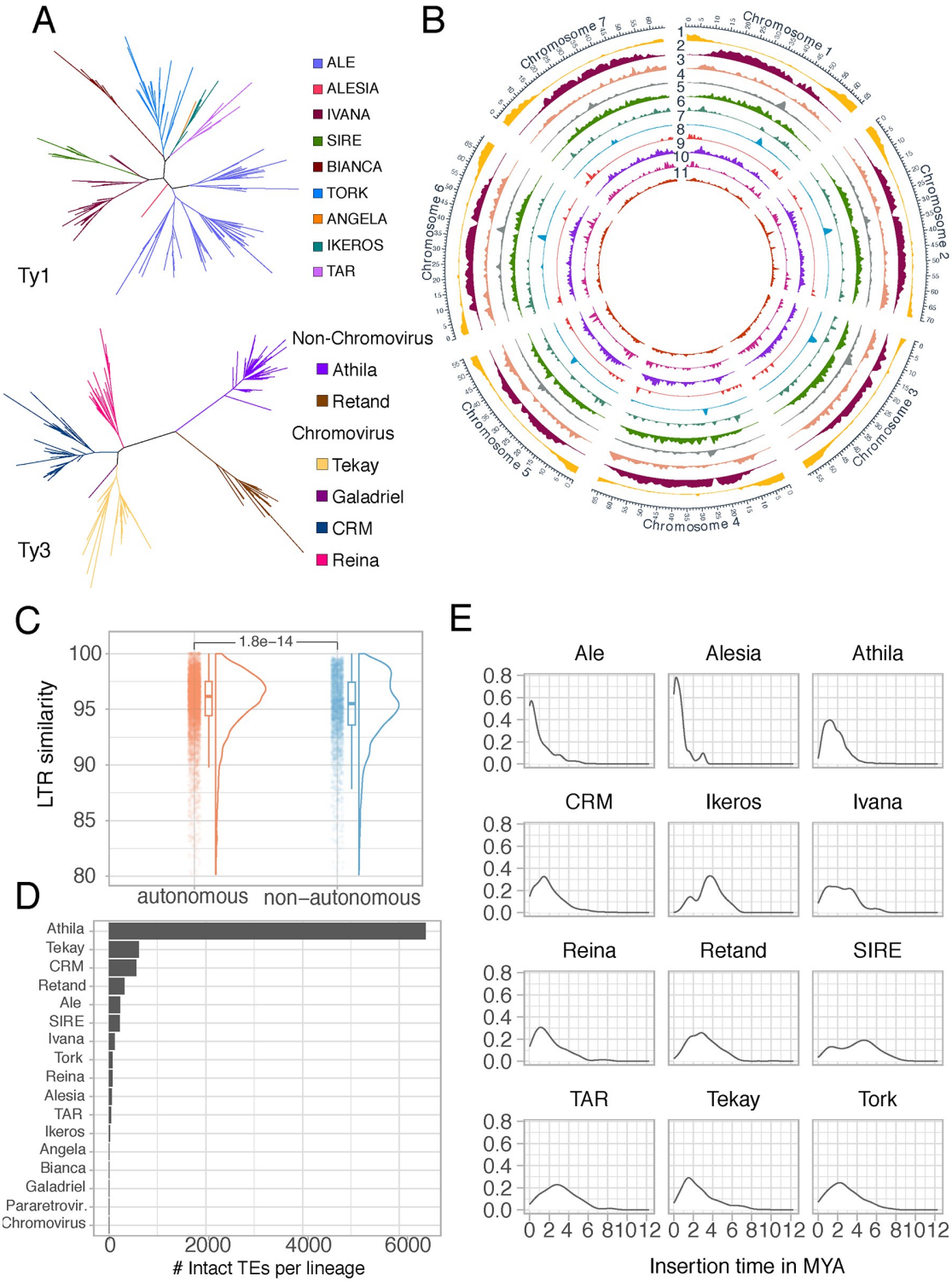


Fig 1. Genome-wide distribution and classification of TE families and superfamilies in the *T. arvense* reference genome MN106-Ref. (A) Phylogenetic tree of LTR retrotransposons based on the reverse transcriptase domain. **(B)** Genome-wide distribution of TE family and superfamily abundances. The tracks 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).

https://doi.org/10.1371/journal.pgen.1011141.g001

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 (S2 Fig). 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 the latter view is that TIPs in the gene-rich fraction of the genome, near the telomeres, have higher allele frequencies (Fig 2E), while TIPs in the pericentromeric regions are more abundant, but have lower allele frequencies (see S3 Fig for a statistical assessment).

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 (Figs 2B and 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.

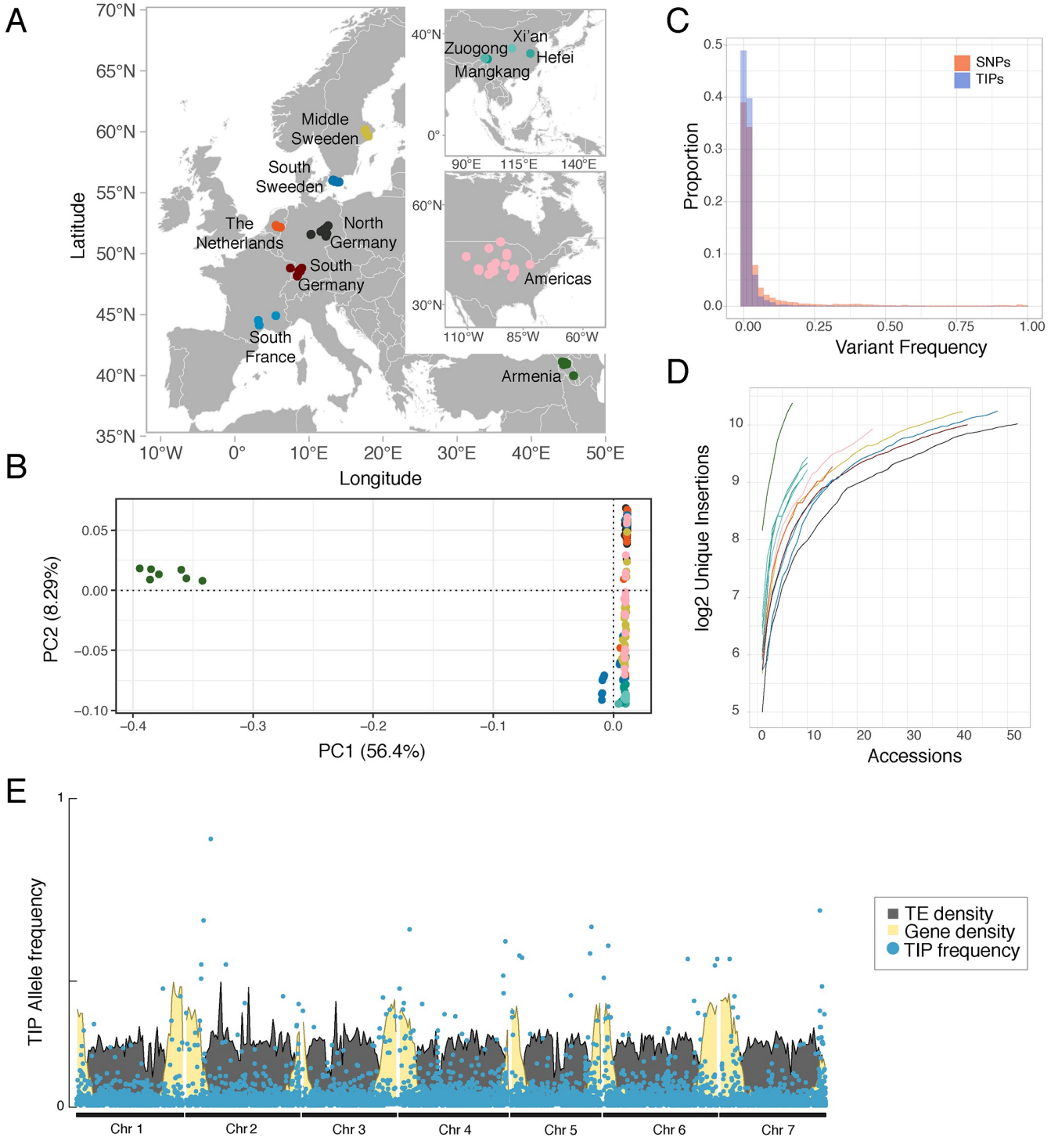


Fig 2. The genome-wide landscape of TE insertion polymorphisms in *T. arvense*. **(A)** Distribution of accessions across their native Eurasian and naturalized North American range in the Northern hemisphere (omitting a sample from Chile, included in the Americas group). **(B)** A SNP-based principal 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 S4 Fig. **(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)** Average TIP frequencies over 100 kb windows along the genome, compared to gene and TE densities, also displayed in 100 Kb windows. Map source: naturalearthdata.com.

https://doi.org/10.1371/journal.pgen.1011141.g002

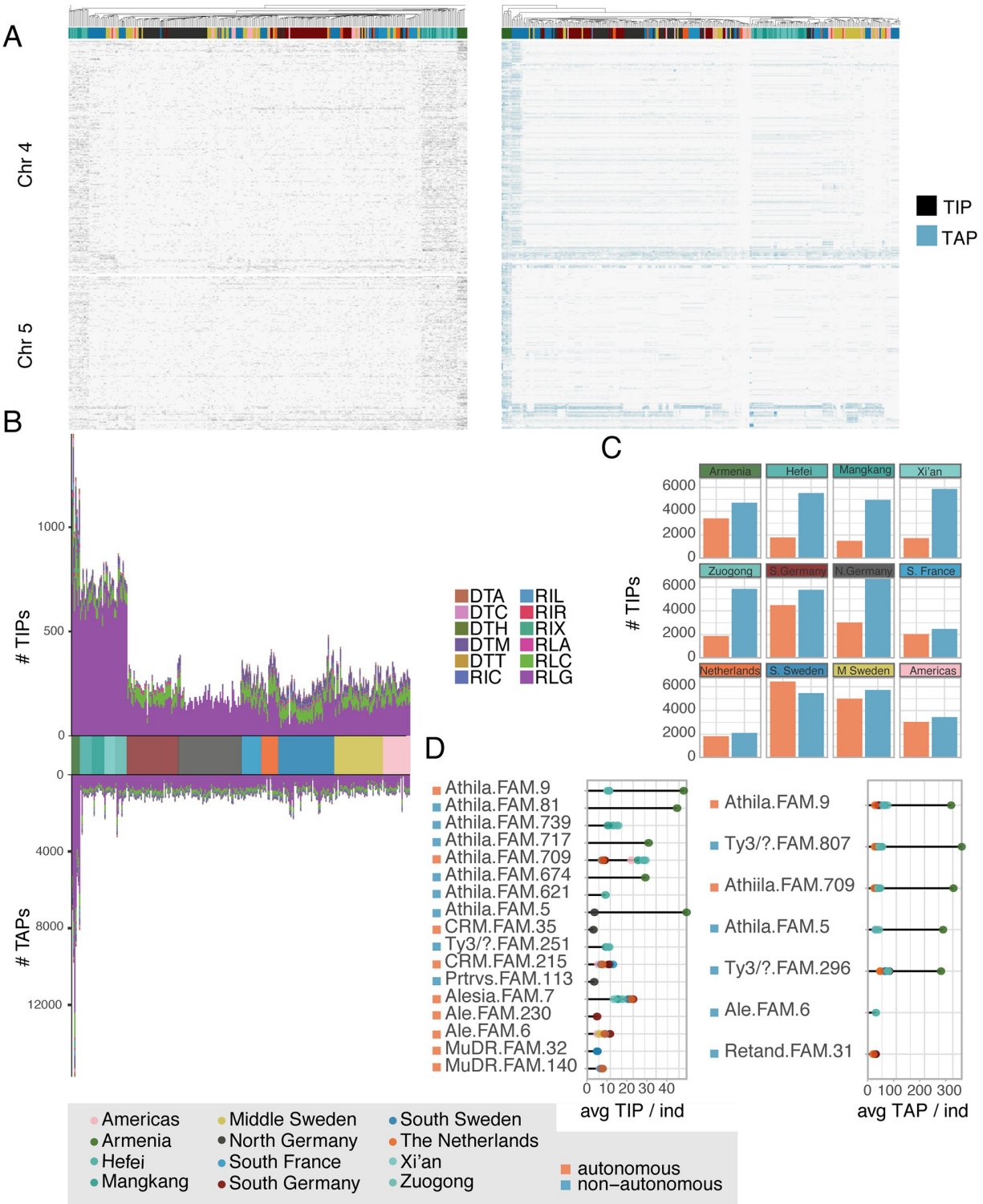


Fig 3. The *T. arvense* mobilome. (A) Genomic distribution of TIPs and TAPs in chromosomes 4 and 5, where we observe major TIP/TAP hotspots. TIPs and TAPs along the other chromosomes are shown in S6 Fig. **(B)** Contribution of different superfamilies to transposon insertion polymorphisms (TIPs) and 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.

https://doi.org/10.1371/journal.pgen.1011141.g003

Host control of TE mobility

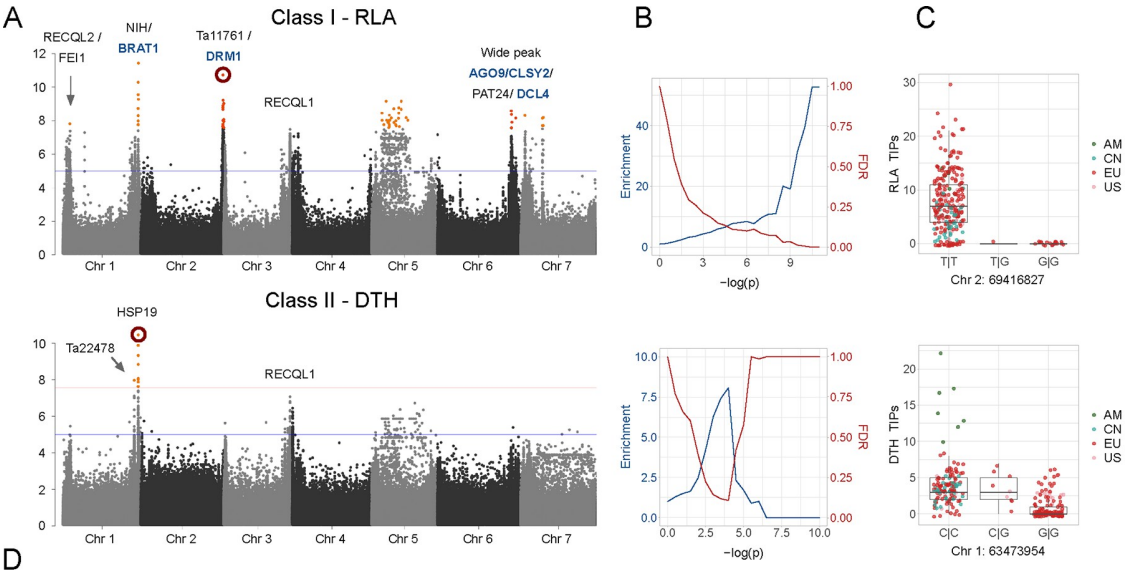
In *A*. *thaliana*, natural genetic variation affects TE mobility and genome-wide patterns of TE distribution, driven by functional changes in key epigenetic regulators [14,30–32]. The rich inventory of TE polymorphisms in *T*. *arvense* offered an opportunity to investigate the genetic basis of TE mobility 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–4D). 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–4D), class II TEs were mostly associated with allelic variation at an ortholog of *O*. *sativa HEAT SHOCK PROTEIN 19* (*HSP19)*. Only class I TE superfamilies were enriched for significant associations close to DNA methylation machinery genes (Fig 4B), and this difference was consistent for most superfamilies that belonged to either class I or class II (S7 Fig). The most prominent hits for class I TIPs were near orthologs 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] (Figs 4A–4D, S7 and S8). Another category of genes that emerged in our GWA are genes encoding DNA and RNA helicases such as *RECQL1* and *2* (Figs 4 and S8). Some of our GWA peaks extend over several genes and might reflect associations with less well characterized genes, but others have the strongest associations in individual genes such as *HSP19* and *BRAT1* (S8 Fig). For *HSP19*, the top SNPs are located in introns and it is difficult to predict their effect. *BRAT1* has two highly significant, fully linked SNPs in exons 1 and 4. The SNP in exon 4 (Chr1:63627484) introduces a stop codon that removes part of the ATPase domain and the entire chromatin binding bromodomain, and this mutation almost certainly completely eliminates BRAT1's anti-silencing activity [7].

Since accessions that diverged earlier from the reference had potentially more time to accumulate TIPs, we also estimated the age of all insertions [14] and repeated the GWA using only TIPs younger than 500,000 years. The results were similar to using all TIPs, suggesting that this potential reference bias is unlikely to drive any of the identified associations (S9 Fig).

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; where H stands for all three nucleotides but G), 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.

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 (< 1 kb), of those, 8 copies are within genes (S3 Table). Across all 4,215 Alesia.FAM.7 TIPs, that is insertions not present in



D

Candidate gene	Gene ID	Ortholog	Putative function	Top SNP location
RECQL2	Tarvense_11198	AT1G31360	RECQ HELICASE L2 (ATP-dependent DNA helicase)	5kb (downstream)
FEI1	Tarvense_11200	AT1G31420	Cell wall elongation	0 kb (intron)
Ta22478	Tarvense_22478	AT1G14990	Uncharacterized – transmembrane protein	0 kb (intron)
HSP19	Tarvense_21603	OsHSP19	class II heat shock protein	0 kb (intron)
NIH	Tarvense_21621	AT1G06670	NUCLEAR DEIH-BOXHELICASE (RNA/DNA helicase)	0.1 kb (upstream)
BRAT1	**Tarvense_21555**	**AT1G05910**	**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)**

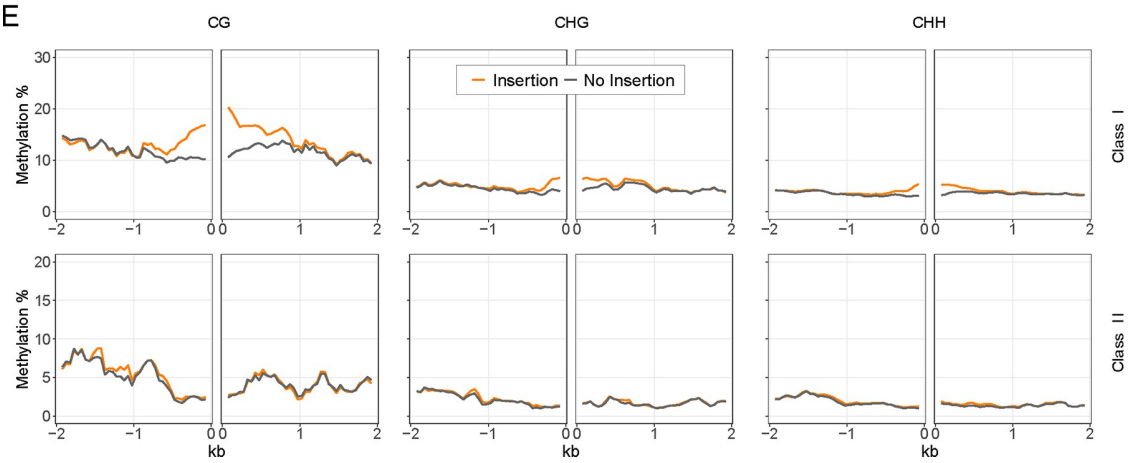


Fig 4. GWA analysis for TIP load of a class I and a class II TE superfamily. Results including all superfamilies are shown in S7 Fig. (**A**) Manhattan plots with candidate genes indicated next to neighboring variants. The red line corresponds to a genome-wide significance with full Bonferroni correction, the blue line to a more generous threshold of –log(p) = 5. (**B**) Enrichment and expected FDR of *a priori* candidate 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.

https://doi.org/10.1371/journal.pgen.1011141.g004

the reference genome, we found a strong enrichment nearby and within genes, which was the case for ~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*. *oleracea*, *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 (S10 Fig) and Alesia.FAM.7 was indeed the only deeply conserved TE family with evidence for recent activity.

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, Ale elements are preferentially centrophilic whereas Alesia elements 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.

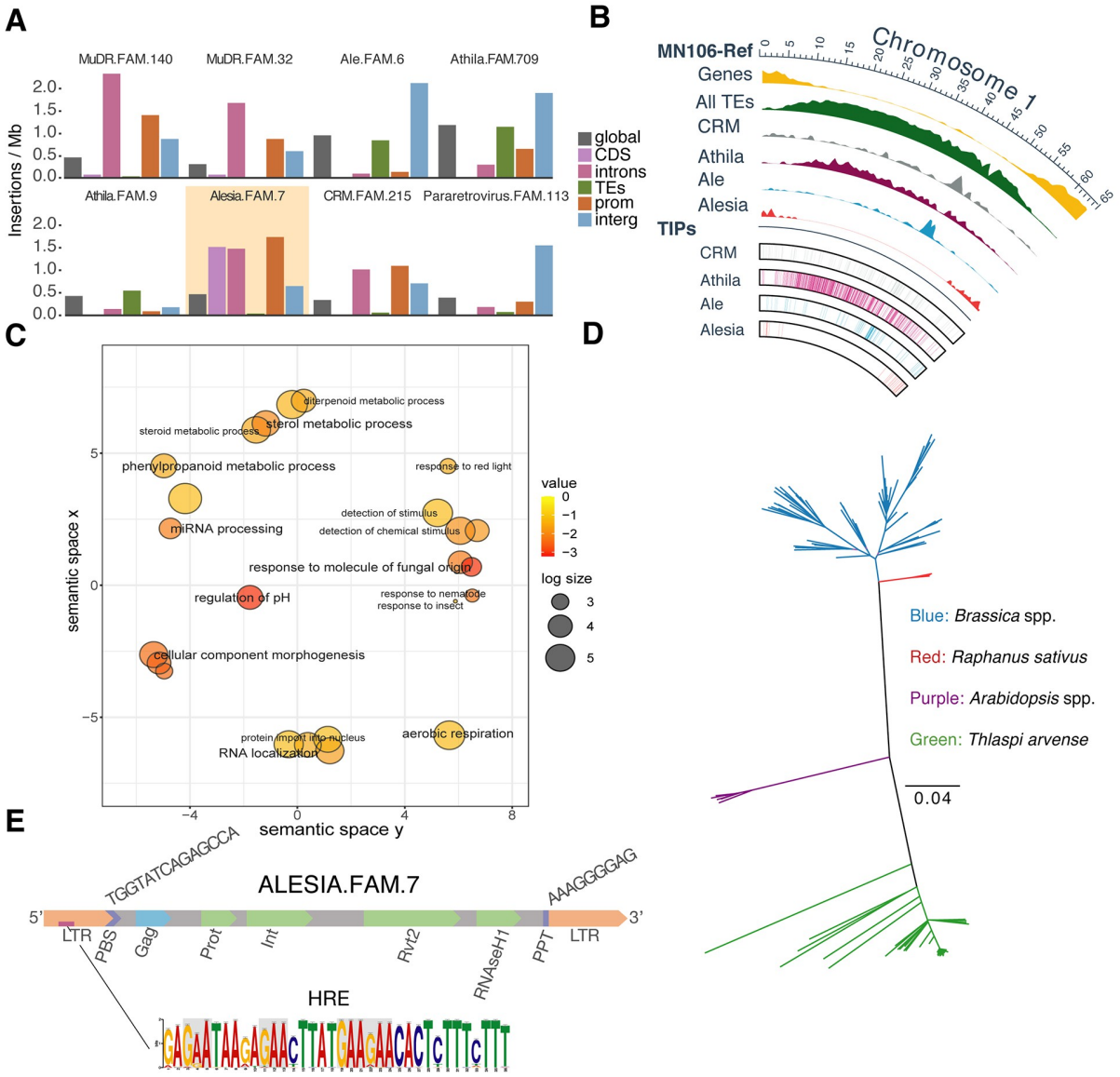


Fig 5. Summary statistics and characterization of the Alesia.FAM.7 family in *T. arvense* and other Brassicaceae. **(A)** Distribution of several TE families across different genomic contexts in *T. arvense* accessions. While several other families, such as MuDR.FAM.140 or CRM.FAM.215, are also often 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 by a purple segment.

https://doi.org/10.1371/journal.pgen.1011141.g005

We learned that one third of the *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, 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]. That class I and class II TEs in *T*. *arvense* apparently differ in their genetic requirements for silencing can be potentially linked to our observation that DNA methylation spreads more rapidly from class I than class II TE insertions in this species.

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. TE insertions in exons might disrupt genes, while intronic insertions might modulate alternative splicing or reduce the accumulation of correctly spliced transcripts. An example is provided by *A*. *thaliana* accessions in which an intronic COPIA insertion in the disease resistance gene *RPP7* shifts the balance between full-length and truncated transcripts [54]. 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] (S3 Table). We used as reference the assembly generated in [26], together with the gene and TE annotation also generated in that study. To visualize the accession locations in the world map, we used free vector and raster map data from naturalearthdata.com. We reinforced this dataset by sequencing 12 different accessions, 7 Armenian and 5 European, using Illumina paired-end 2x150 bp WGS (S3 Table). 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 two assemblies for *Eutrema salsugineum* and *Schrenkiella parvula* (NCBI ID: PRJNA73205; Phytozome genome ID: 574 respectively) as outgroup species.

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) and its Viridiplantae v3.0 database. 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 likely have frameshifts and stop codons in these sequences. Using these identified protein domains, we evaluated whether a given TE family is autonomous, i.e., whether it codes for the entire machinery needed for transposition. An LTR retrotransposon family was considered autonomous with the following domains identified by DANTE: retrotranscriptase, RT; capsid related domain, GAG; RNase H, RH; protease, PROT; and integrase, INT. Autonomous non-LTR retrotransposons, LINEs, had to contain: retrotranscriptase, RT. DNA TE families had to contain: transposase, TPase. DNA TEs of the Helitron superfamily had to contain in addition: DNA helicase, HEL.

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 [55]. Using RAxML [56], 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 [57] 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 [58] intersecting with" -f 0.8 -r" parameters.

To analyze the extent of conservation of TE families larger than 2 kb across Brassicaceae, we ran BLASTN [59] 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 [55] with default settings and constructed a tree with RaxML [56] 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 [60] 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).

Expression analysis of reference TEs was performed using TEspeX [61]. We obtained paired-end RNA seq data from 27 samples comprising nine different tissues from the MN106-Ref reference accession [26]. We obtained raw counts for each library by mapping the reads to both transcripts of protein coding genes and to the TE consensus library. Raw counts were normalized as suggested [61] (RPM: raw counts/total mapped reads x 1 million). We used a non-parametric Wilcoxon rank-sum test to compare expression between autonomous and non-autonomous TE families.

Short variant calling

We called variants with GATK4 [62], 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 ($\leq$25 bp) using cutadapt v2.6 [63]. We aligned trimmed reads to the reference genome [26] with BWA-MEM v0.7.17 [64], 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 [65], 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 [66], 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* [67] 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 (S4 Fig).

Lastly, we analyzed the genetic relatedness among accessions from different geographic regions constructing a maximum likelihood tree using *TREEMIX* [68] 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 [69]. 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, using sample SAMEA9464759 from ENA project PRJEB46635 [26]. We calculated read coverage (RC) in 100 bp windows, adjusted for GC content [70], and excluded windows with abnormal coverage, arbitrarily defined as threefold

lower or higher than the genome wide, GC content adjusted mean. Any TIPs in these regions, which corresponded to ~16% of the reference genome, 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 estimate the age of insertions, we used this same classification and calculated the maximum pairwise divergence (number of SNPs) between each combination of two carriers, in the 70 kb region around the insertion [14], using simply the number of private SNPs for singletons. We then extrapolated the most likely age based on *A*. *thaliana* mutation rate [71], assuming 1 generation per year.

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 a visual browser. Over 70% of TIPs were deemed correct, which is in line with reports from other studies in *A*. *thaliana* [32] and tomato [72].

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 [73] using 600 mg 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 6–10 kb 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* [74]. 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* [75], correcting for population structure with an Isolation-By-State (IBS) matrix. Starting from the complete VCF file obtained from variant calling, we used *PLINK* [67] 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 TiPs 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* [58] 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 *OrthoFinder2* [76]. We assessed enrichment with *clusterProfiler* [77] and piped all terms with p value < 0.05 to *REVIGO* [78], using default parameters.

Code availability

Code used for analysis and figures can be found at: https://github.com/acontrerasg/Tarvense_transposon_dynamics.

Supporting information

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

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

S3 Table. List of datasets that were uploaded to the Zenodo repository: https://doi.org/10.5281/zenodo.10161730 (10.5281/zenodo.6372331).
(XLSX)

S1 Fig. Comparison of autonomous and non-autonomous TE families in *T. arvense* MN106-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. **(E**) TE expression in autonomous vs. non-autonomous TEs.
(TIF)

S2 Fig. 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.
(TIF)

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

S4 Fig. 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.
(TIF)

S5 Fig. 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 result recapitulates the clustering pattern observed with the SNP-PCA.
(TIF)

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

S7 Fig. 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.
(TIF)

S8 Fig. Zoom-in of GWA peaks with candidate genes highlighted. The genome-wide significance (dotted red line) corresponds to a full Bonferroni correction. DNA methylation machinery genes used for the enrichment of *a priori* candidates are depicted in blue, other genes that might affect transposition in red. The putative knock-out SNP disrupting the function of

BRAT1 is depicted in green.
(TIF)

S9 Fig. GWA results for genome-wide load of TIPs younger than 500,000 years. Left: Manhattan plots for load of TIPs for each TE superfamily. 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 links to DNA methylation, for significance threshold increments. Right: QQplots of p-values.
(TIF)

S10 Fig. BLASTN hits of *T. arvense* TE families with model sizes > 4 kb against the NCBI NT database, 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.
(TIF)

Acknowledgments

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. We also thank Kevin M. Dorn for sharing *T. arvense* seeds. For computing, we acknowledge Prof. Peter Stadler at the University of Leipzig and David Langenberger from ecSeq, for hosting the EpiDiverse servers, and the High Performance and Cloud Computing Group at the Zentrum für Datenverarbeitung of the University of Tübingen for managing the BinAC server.

Author Contributions

Conceptualization: Adrián Contreras-Garrido, Dario Galanti, Andrea Movilli, Oliver Bossdorf, Hajk-Georg Drost, Detlef Weigel.

Data curation: Adrián Contreras-Garrido, Dario Galanti, Andrea Movilli.

Formal analysis: Adrián Contreras-Garrido, Dario Galanti.

Funding acquisition: Claude Becker, Oliver Bossdorf, Detlef Weigel.

Investigation: Adrián Contreras-Garrido, Dario Galanti, Claude Becker, Oliver Bossdorf, Hajk-Georg Drost, Detlef Weigel.

Methodology: Adrián Contreras-Garrido, Dario Galanti.

Project administration: Oliver Bossdorf, Hajk-Georg Drost, Detlef Weigel.

Resources: Adrián Contreras-Garrido, Dario Galanti, Andrea Movilli, Claude Becker, Oliver Bossdorf, Hajk-Georg Drost, Detlef Weigel.

Software: Adrián Contreras-Garrido.

Supervision: Oliver Bossdorf, Hajk-Georg Drost, Detlef Weigel.

Validation: Adrián Contreras-Garrido, Dario Galanti, Andrea Movilli, Hajk-Georg Drost, Detlef Weigel.

Visualization: Adrián Contreras-Garrido, Dario Galanti.

Writing – original draft: Adrián Contreras-Garrido, Dario Galanti, Claude Becker, Oliver Bossdorf, Hajk-Georg Drost, Detlef Weigel.

Writing – review & editing: Adrián Contreras-Garrido, Dario Galanti, Andrea Movilli, Claude Becker, Oliver Bossdorf, Hajk-Georg Drost, Detlef Weigel.

References

1. Wells JN, Feschotte C. A Field Guide to Eukaryotic Transposable Elements. Annu Rev Genet. 2020. https://doi.org/10.1146/annurev-genet-040620-022145 PMID: 32955944
2. Tenaillon MI, Hollister JD, Gaut BS. A triptych of the evolution of plant transposable elements. Trends Plant Sci. 2010; 15: 471–478. https://doi.org/10.1016/j.tplants.2010.05.003 PMID: 20541961
3. Wicker T, Gundlach H, Spannagl M, Uauy C, Borrill P, Ramírez-González RH, et al. Impact of transposable elements on genome structure and evolution in bread wheat. Genome Biol. 2018; 19: 103. https://doi.org/10.1186/s13059-018-1479-0 PMID: 30115100
4. 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. https://doi.org/10.1038/nrg2165 PMID: 17984973
5. 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. https://doi.org/10.1186/s13100-018-0144-1 PMID: 30622655
6. Arkhipova IR. Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories. Mob DNA. 2017; 8: 19. https://doi.org/10.1186/s13100-017-0103-2 PMID: 29225705
7. Zhang H, Lang Z, Zhu J-K. Dynamics and function of DNA methylation in plants. Nat Rev Mol Cell Biol. 2018; 19: 489–506. https://doi.org/10.1038/s41580-018-0016-z PMID: 29784956
8. Bouyer D, Kramdi A, Kassam M, Heese M, Schnittger A, Roudier F, et al. DNA methylation dynamics during early plant life. Genome Biol. 2017; 18: 179. https://doi.org/10.1186/s13059-017-1313-0 PMID: 28942733
9. Sigman MJ, Slotkin RK. The First Rule of Plant Transposable Element Silencing: Location, Location, Location. Plant Cell. 2016; 28: 304–313. https://doi.org/10.1105/tpc.15.00869 PMID: 26869697
10. Srikant T, Drost H-G. How stress facilitates phenotypic innovation through epigenetic diversity. Front Plant Sci. 2020; 11: 606800. https://doi.org/10.3389/fpls.2020.606800 PMID: 33519857
11. Pecinka A, Dinh HQ, Baubec T, Rosa M, Lettner N, Mittelsten Scheid O. Epigenetic regulation of repetitive elements is attenuated by prolonged heat stress in Arabidopsis. Plant Cell. 2010; 22: 3118–3129. https://doi.org/10.1105/tpc.110.078493 PMID: 20876829
12. Cavrak VV, Lettner N, Jamge S, Kosarewicz A, Bayer LM, Mittelsten Scheid O. How a retrotransposon exploits the plant's heat stress response for its activation. PLoS Genet. 2014; 10: e1004115. https://doi.org/10.1371/journal.pgen.1004115 PMID: 24497839
13. 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. https://doi.org/10.1038/nature09861 PMID: 21399627
14. Baduel P, Leduque B, Ignace A, Gy I, Gil J Jr, Loudet O, et al. Genetic and environmental modulation of transposition shapes the evolutionary potential of Arabidopsis thaliana. Genome Biol. 2021; 22: 138. https://doi.org/10.1186/s13059-021-02348-5 PMID: 33957946
15. Ou S, Collins T, Qiu Y, Seetharam AS, Menard CC, Manchanda N, et al. Differences in activity and stability drive transposable element variation in tropical and temperate maize. bioRxiv. 2022. p. 2022.10.09.511471. https://doi.org/10.1101/2022.10.09.511471
16. Benoit M, Drost H-G, Catoni M, Gouil Q, Lopez-Gomollon S, Baulcombe DC, et al. Environmental and epigenetic regulation of Rider retrotransposons in tomato. bioRxiv. 2019. p. 517508. https://doi.org/10.1371/journal.pgen.1008370 PMID: 31525177
17. Esposito S, Barteri F, Casacuberta J, Mirouze M, Carputo D, Aversano R. LTR-TEs abundance, timing and mobility in Solanum commersonii and S. tuberosum genomes following cold-stress conditions. Planta. 2019; 250: 1781–1787. https://doi.org/10.1007/s00425-019-03283-3 PMID: 31562541

18. Paszkowski J. Controlled activation of retrotransposition for plant breeding. Curr Opin Biotechnol. 2015; 32: 200–206. https://doi.org/10.1016/j.copbio.2015.01.003 PMID: 25615932
19. McGinn M, Phippen WB, Chopra R, Bansal S, Jarvis BA, Phippen ME, 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 PMID: 30230695
20. García Navarrete T, Arias C, Mukundi E, Alonso AP, Grotewold E. Natural variation and improved genome annotation of the emerging biofuel crop field pennycress (Thlaspi arvense). G3. 2022. https://doi.org/10.1093/g3journal/jkac084 PMID: 35416986
21. 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. https://doi.org/10.1093/dnares/dsu045 PMID: 25632110
22. Hill J, Nelson E, Tilman D, Polasky S, Tiffany D. Environmental, economic, and energetic costs and benefits of biodiesel and ethanol biofuels. Proc Natl Acad Sci U S A. 2006; 103: 11206–11210. https://doi.org/10.1073/pnas.0604600103 PMID: 16837571
23. 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.
24. Frels K, Chopra R, Dorn KM, Wyse DL, Marks MD, Anderson JA. Genetic Diversity of Field Pennycress (Thlaspi arvense) Reveals Untapped Variability and Paths Toward Selection for Domestication. Agronomy. 2019; 9: 302.
25. Warwick SI, Francis A, Susko DJ. The biology of Canadian weeds. 9. Thlaspi arvense L. (updated). Can J Plant Sci. 2002; 82: 803–823.
26. Nunn A, Rodríguez-Arévalo I, Tandukar Z, Frels K, Contreras-Garrido A, Carbonell-Bejerano P, 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 PMID: 34990041
27. 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. https://doi.org/10.3389/fpls.2021.772655 PMID: 35058947
28. 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. https://doi.org/10.1371/journal.pgen.1010452 PMID: 36223399
29. Bourgeois Y, Boissinot S. On the Population Dynamics of Junk: A Review on the Population Genomics of Transposable Elements. Genes. 2019;10. https://doi.org/10.3390/genes10060419 PMID: 31151307
30. Dubin MJ, Zhang P, Meng D, Remigereau M-S, Osborne EJ, Paolo Casale F, et al. DNA methylation in Arabidopsis has a genetic basis and shows evidence of local adaptation. Elife. 2015; 4: e05255. https://doi.org/10.7554/eLife.05255 PMID: 25939354
31. Sasaki E, Gunis J, Reichardt-Gomez I, Nizhynska V, Nordborg M. Conditional GWAS of non-CG transposon methylation in Arabidopsis thaliana reveals major polymorphisms in five genes. PLoS Genet. 2022; 18: e1010345. https://doi.org/10.1371/journal.pgen.1010345 PMID: 36084135
32. 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. https://doi.org/10.7554/eLife.15716 PMID: 27258693
33. Erdmann RM, Picard CL. RNA-directed DNA Methylation. PLoS Genet. 2020; 16: e1009034. https://doi.org/10.1371/journal.pgen.1009034 PMID: 33031395
34. Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, et al. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature. 2010; 465: 627–631. https://doi.org/10.1038/nature08800 PMID: 20336072
35. 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. https://doi.org/10.1186/s13059-016-1072-3 PMID: 27729060
36. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, et al. Database resources of the national center for biotechnology information. Nucleic Acids Res. 2022; 50: D20–D26. https://doi.org/10.1093/nar/gkab1112 PMID: 34850941
37. Krämer U. Planting molecular functions in an ecological context with Arabidopsis thaliana. Elife. 2015;4. https://doi.org/10.7554/eLife.06100 PMID: 25807084
38. 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. https://doi.org/10.1186/s13059-018-1577-z PMID: 30454069

39. Sultana T, Zamborlini A, Cristofari G, Lesage P. Integration site selection by retroviruses and transposable elements in eukaryotes. Nat Rev Genet. 2017; 18: 292–308. https://doi.org/10.1038/nrg.2017.7 PMID: 28286338

40. 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. https://doi.org/10.1038/s41467-019-11385-5 PMID: 31366887

41. 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. https://doi.org/10.1002/ajb2.1709 PMID: 34415576

42. Zhang S-J, Liu L, Yang R, Wang X. Genome Size Evolution Mediated by Gypsy Retrotransposons in Brassicaceae. Genomics Proteomics Bioinformatics. 2020; 18: 321–332. https://doi.org/10.1016/j.gpb.2018.07.009 PMID: 33137519

43. Geng Y, Guan Y, Qiong L, Lu S, An M, Crabbe MJC, et al. Genomic analysis of field pennycress (Thlaspi arvense) provides insights into mechanisms of adaptation to high elevation. BMC Biol. 2021; 19: 143. https://doi.org/10.1186/s12915-021-01079-0 PMID: 34294107

44. Kim S, Park M, Yeom S-I, Kim Y-M, Lee JM, Lee H-A, et al. Genome sequence of the hot pepper provides insights into the evolution of pungency in Capsicum species. Nat Genet. 2014; 46: 270–278. https://doi.org/10.1038/ng.2877 PMID: 24441736

45. 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. https://doi.org/10.1371/journal.pgen.1008492 PMID: 31887137

46. Ichino L, Boone BA, Strauskulage L, Jake Harris C, Kaur G, Gladstone MA, et al. MBD5 and MBD6 couple DNA methylation to gene silencing through the J-domain protein SILENZIO. Science. 2021 [cited 4 Jun 2021]. https://doi.org/10.1126/science.abg6130 PMID: 34083448

47. Specchia V, Bozzetti MP. The Role of HSP90 in Preserving the Integrity of Genomes Against Transposons Is Evolutionarily Conserved. Cells. 2021;10. https://doi.org/10.3390/cells10051096 PMID: 34064379

48. Cappucci U, Noro F, Casale AM, Fanti L, Berloco M, Alagia AA, et al. The Hsp70 chaperone is a major player in stress-induced transposable element activation. Proc Natl Acad Sci U S A. 2019; 116: 17943–17950. https://doi.org/10.1073/pnas.1903936116 PMID: 31399546

49. Specchia V, Piacentini L, Tritto P, Fanti L, D'Alessandro R, Palumbo G, et al. Hsp90 prevents phenotypic variation by suppressing the mutagenic activity of transposons. Nature. 2010; 463: 662–665. https://doi.org/10.1038/nature08739 PMID: 20062045

50. Volff J-N. Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. Bioessays. 2006; 28: 913–922. https://doi.org/10.1002/bies.20452 PMID: 16937363

51. Jangam D, Feschotte C, Betrán E. Transposable Element Domestication As an Adaptation to Evolutionary Conflicts. Trends Genet. 2017; 33: 817–831. https://doi.org/10.1016/j.tig.2017.07.011 PMID: 28844698

52. Almeida MV, Vernaz G, Putman ALK, Miska EA. Taming transposable elements in vertebrates: from epigenetic silencing to domestication. Trends Genet. 2022; 38: 529–553. https://doi.org/10.1016/j.tig.2022.02.009 PMID: 35307201

53. Drost H-G, Sanchez DH. Becoming a Selfish Clan: Recombination Associated to Reverse-Transcription in LTR Retrotransposons. Genome Biol Evol. 2019; 11: 3382–3392. https://doi.org/10.1093/gbe/evz255 PMID: 31755923

54. 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. https://doi.org/10.1073/pnas.1312545110 PMID: 23940361

55. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013; 30: 772–780. https://doi.org/10.1093/molbev/mst010 PMID: 23329690

56. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014; 30: 1312–1313. https://doi.org/10.1093/bioinformatics/btu033 PMID: 24451623

57. Drost H-G. LTRpred: de novo annotation of intact retrotransposons. JOSS. 2020; 5: 2170.

58. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26: 841–842. https://doi.org/10.1093/bioinformatics/btq033 PMID: 20110278

59. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009; 10: 421. https://doi.org/10.1186/1471-2105-10-421 PMID: 20003500

60. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME Suite. Nucleic Acids Res. 2015; 43: W39–49. https://doi.org/10.1093/nar/gkv416 PMID: 25953851

61. Ansaloni F, Gualandi N, Esposito M, Gustincich S, Sanges R. TEspeX: consensus-specific quantification of transposable element expression preventing biases from exonized fragments. Bioinformatics. 2022; 38: 4430–4433. https://doi.org/10.1093/bioinformatics/btac526 PMID: 35876845

62. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20: 1297–1303. https://doi.org/10.1101/gr.107524.110 PMID: 20644199

63. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011; 17: 10–12.

64. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25: 1754–1760. https://doi.org/10.1093/bioinformatics/btp324 PMID: 19451168

65. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011; 27: 2156–2158. https://doi.org/10.1093/bioinformatics/btr330 PMID: 21653522

66. Browning BL, Zhou Y, Browning SR. A One-Penny Imputed Genome from Next-Generation Reference Panels. Am J Hum Genet. 2018; 103: 338–348. https://doi.org/10.1016/j.ajhg.2018.07.015 PMID: 30100085

67. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81: 559–575. https://doi.org/10.1086/519795 PMID: 17701901

68. Pickrell JK, Pritchard JK. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. 2012; 8: e1002967. https://doi.org/10.1371/journal.pgen.1002967 PMID: 23166502

69. Baduel P, Quadrana L, Colot V. Efficient Detection of Transposable Element Insertion Polymorphisms Between Genomes Using Short-Read Sequencing Data. In: Cho J, editor. Plant Transposable Elements: Methods and Protocols. New York, NY: Springer US; 2021. pp. 157–169.

70. Yoon S, Xuan Z, Makarov V, Ye K, Sebat J. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res. 2009; 19: 1586–1592. https://doi.org/10.1101/gr.092981.109 PMID: 19657104

71. 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. https://doi.org/10.1126/science.1180677 PMID: 20044577

72. Domínguez M, Dugas E, Benchouaia M, Leduque B, Jiménez-Gómez JM, Colot V, et al. The impact of transposable elements on tomato diversity. Nat Commun. 2020; 11: 4058. https://doi.org/10.1038/s41467-020-17874-2 PMID: 32792480

73. 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. https://doi.org/10.1093/nar/gkac1115 PMID: 36453992

74. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, et al. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods. 2018; 15: 461–468. https://doi.org/10.1038/s41592-018-0001-7 PMID: 29713083

75. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 2012; 44: 821–824. https://doi.org/10.1038/ng.2310 PMID: 22706312

76. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019; 20: 238. https://doi.org/10.1186/s13059-019-1832-y PMID: 31727128

77. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012; 16: 284–287. https://doi.org/10.1089/omi.2011.0118 PMID: 22455463

78. Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One. 2011; 6: e21800. https://doi.org/10.1371/journal.pone.0021800 PMID: 21789182

Supporting information

Order	Superfamily	Key	Number of families	Number of copies	% Genomic 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

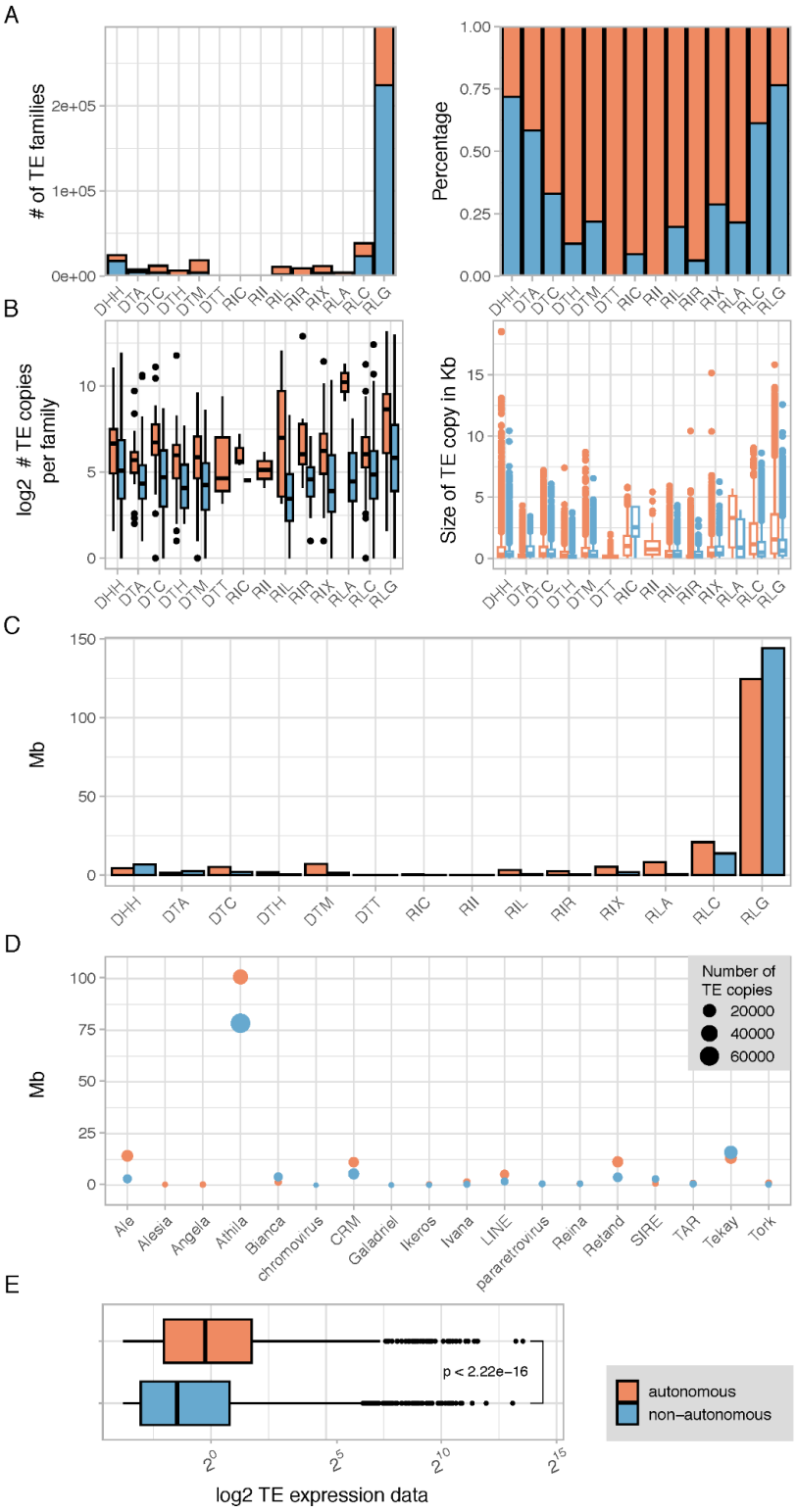
S1 Table. Summary statistics of previously annotated TEs for the *T. arvense* reference genome MN106-Ref [26].

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

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

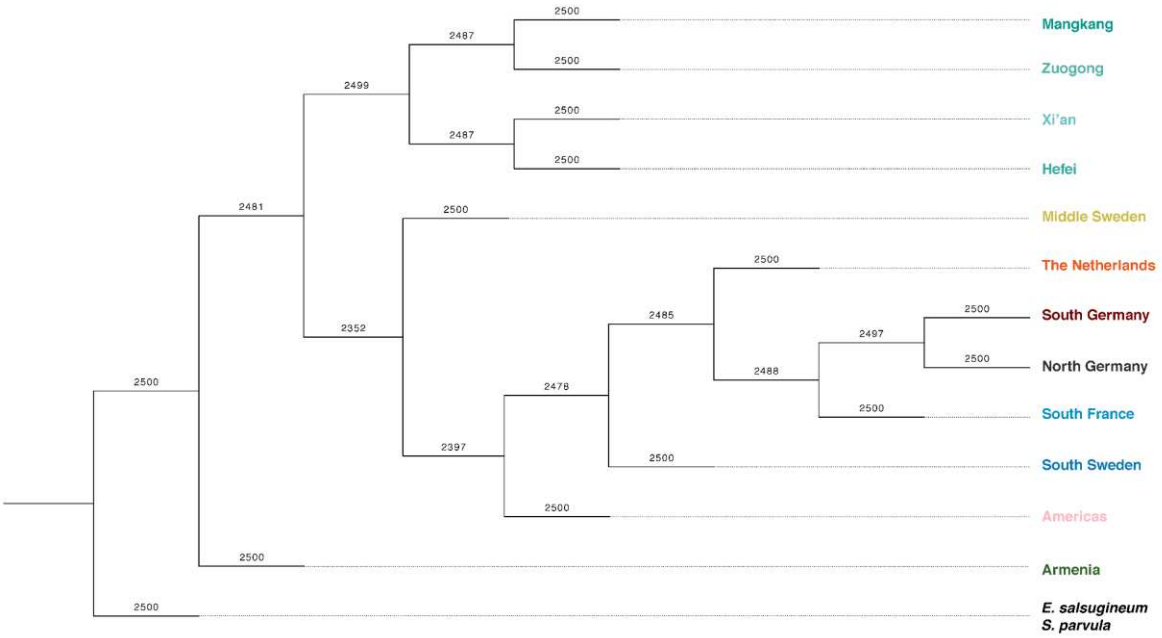
S3A:	Accession numbers of samples sequenced in this study.
S3B:	Metadata of all accessions used in this study.
S3C:	Association of TE family name and the inferred lineage.
S3D:	Complete list of TIPs discovered in this study.
S3E:	Complete list of TAPs discovered in this study.
S3F:	Distribution of Alesia.FAM.7 in the reference genome.
S3G:	Detailed GO enrichment results of genes located within 2 kb 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.
S3I:	GWA results in a format compatible with IGV.
S3J:	GWA results of TIPs younger than 500,000 years in a format compatible with IGV.
S3K:	Bed file with the age estimation of all TIPs.

S3 Table. Additional information. These datasets were uploaded to the Zenodo repository: https://doi.org/10.5281/zenodo.10161731.

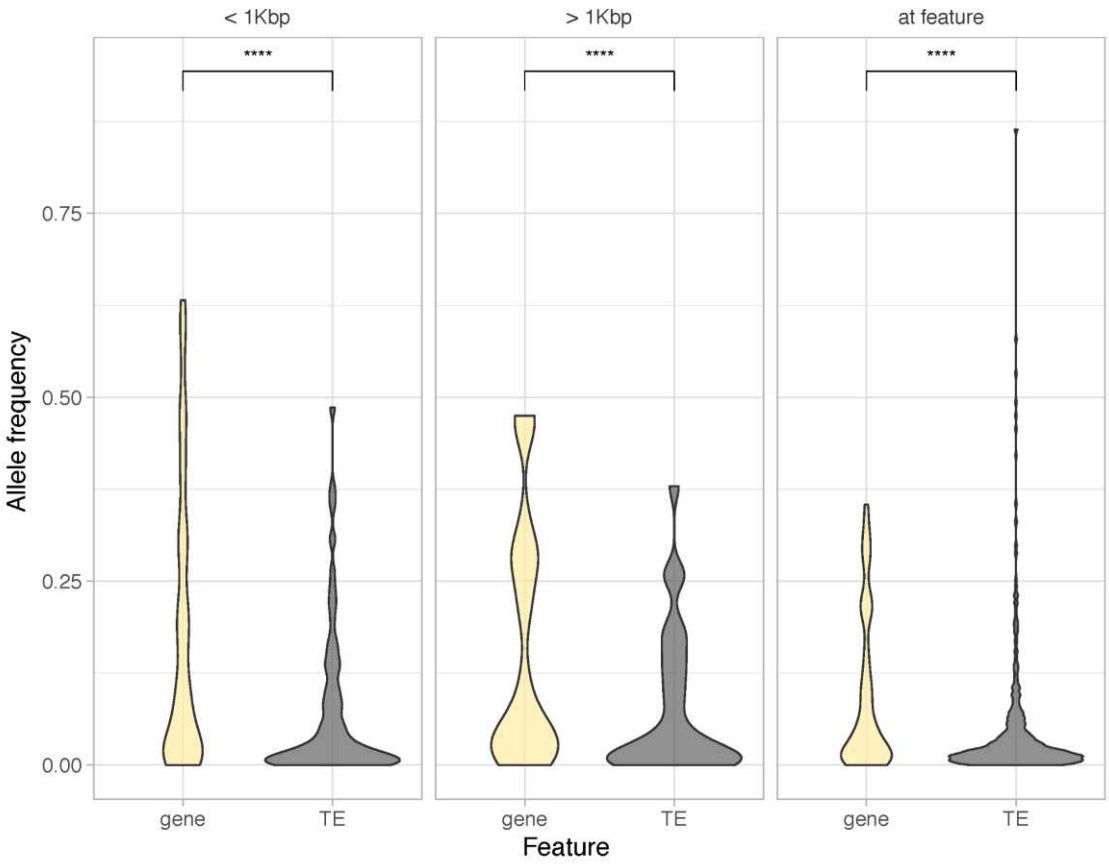


S1 Fig. Comparison of autonomous and non-autonomous TE families in *T. arvense* MN106-Ref. (A) Absolute (left) and relative (right) fraction of autonomous and non-autonomous elements in each TE

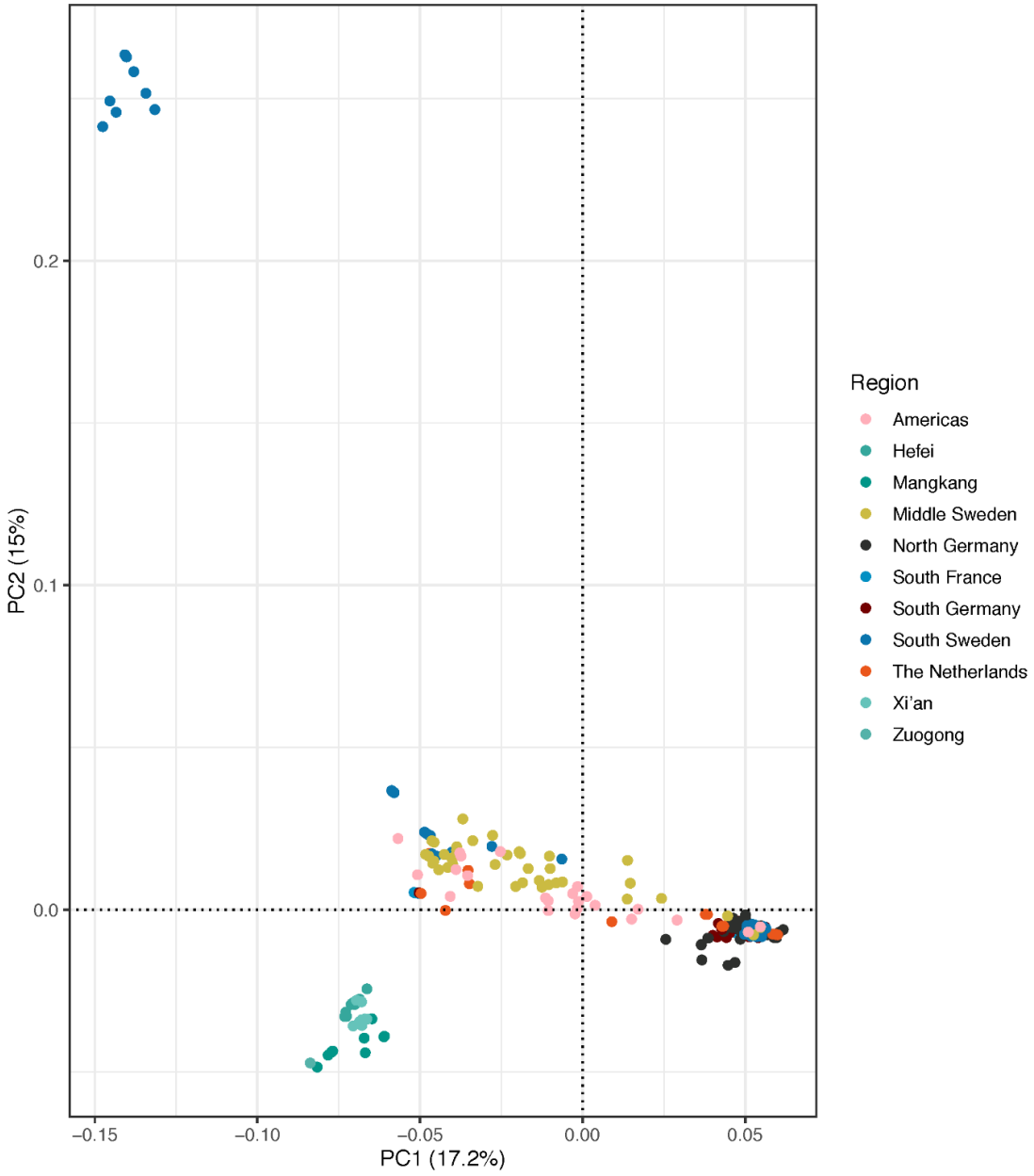
superfamily. (B) Comparison of the fraction of autonomous and non-autonomous elements in each TE superfamily (left). Size comparison of the TE copies according to their autonomy per superfamily (right). (C) Contribution of each superfamily and their autonomous/non-autonomous fraction to total genome size in Mb. (D) Distribution of size and copy number per LTR retrotransposon lineage. (E) TE expression in autonomous vs. non-autonomous TEs.



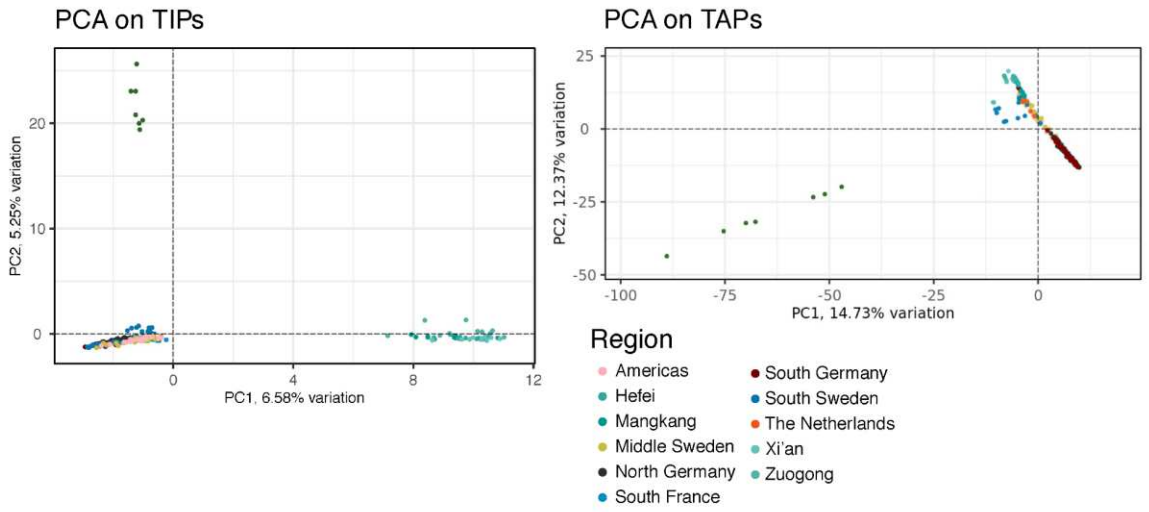
S2 Fig. SNP-based maximum likelihood tree of *T. arvense* populations. Based on a model without migration, 2,500 bootstraps. Node weights represent bootstrap values. Outgroup species at the bottom.



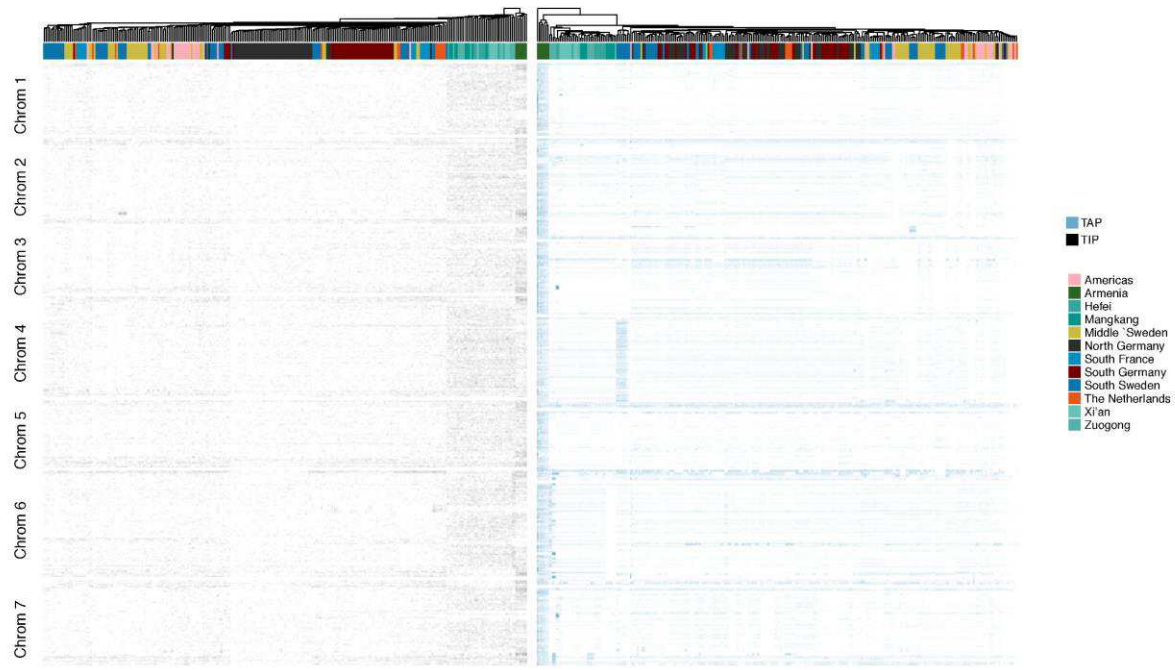
S3 Fig. Frequency distribution of TIPs overlapping with annotated genes and TEs. TIP allele frequencies near other TEs are significantly lower than near genes (Wilcoxon Rank Sum test, $p < 2.22E-16$).



S4 Fig. 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.

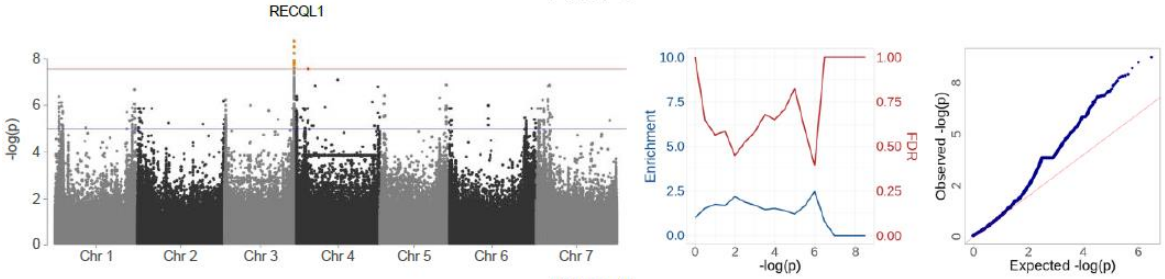


S5 Fig. PCA analysis of 279 individuals of *T. arvense*. A presence/absence matrix of either TIPs (left) or TAPs, (right) was used as input to calculate PCA. This result recapitulates the clustering pattern observed with the SNP-PCA.

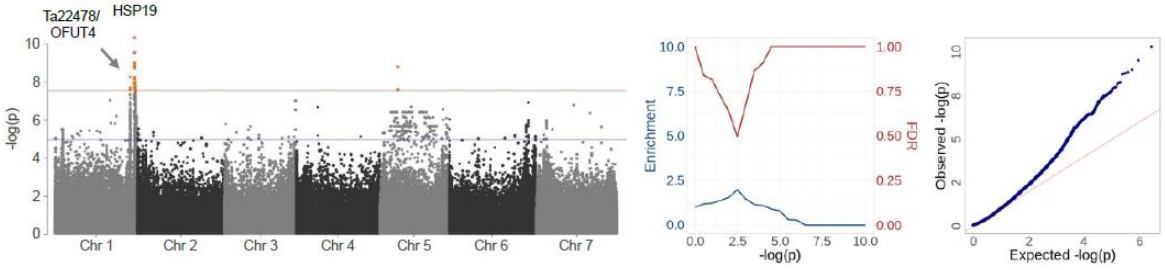


S6 Fig. Genomic distribution of TIPs and TAPs along all seven chromosomes of *T. arvense*. Color columns indicate to which biogeographical population each accession belongs to.

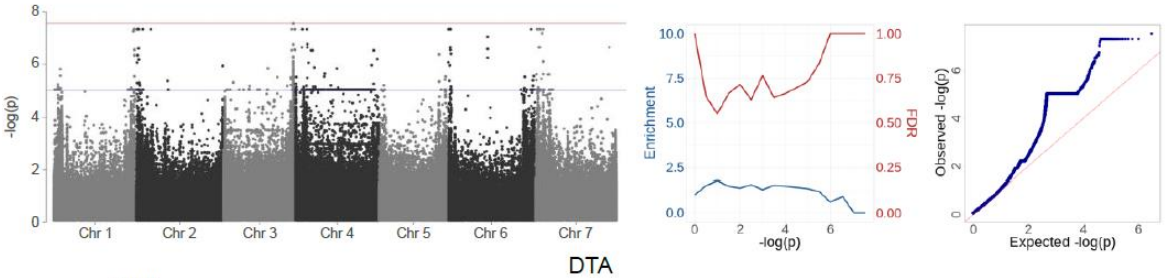
All TIPs



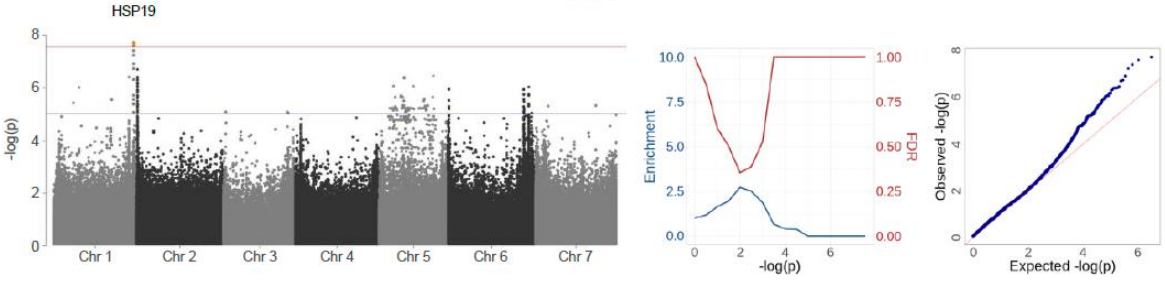
Class II



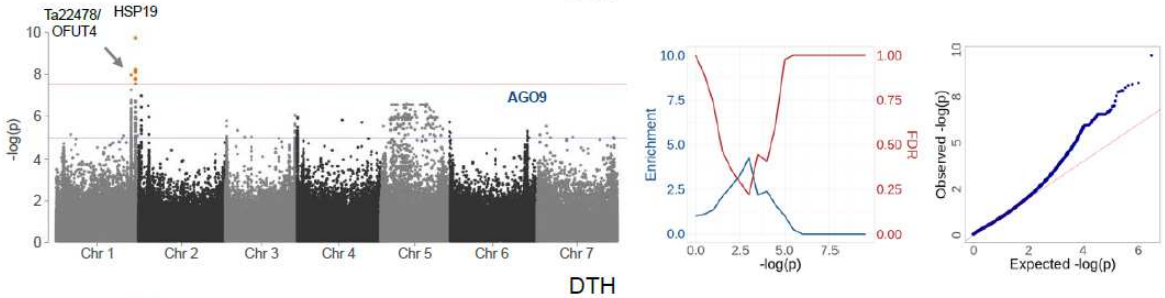
Class I



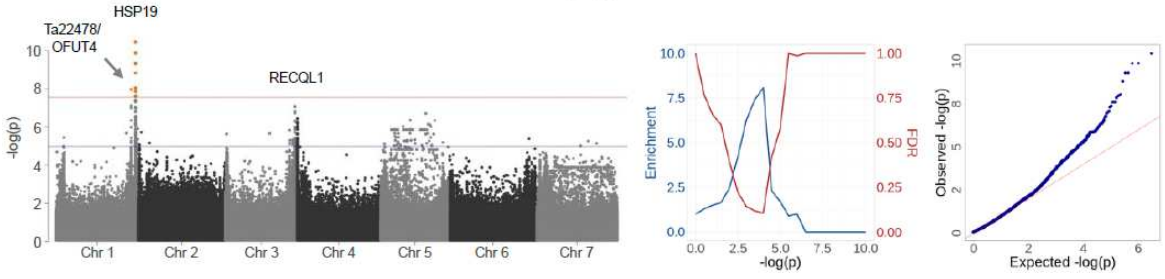
DTA



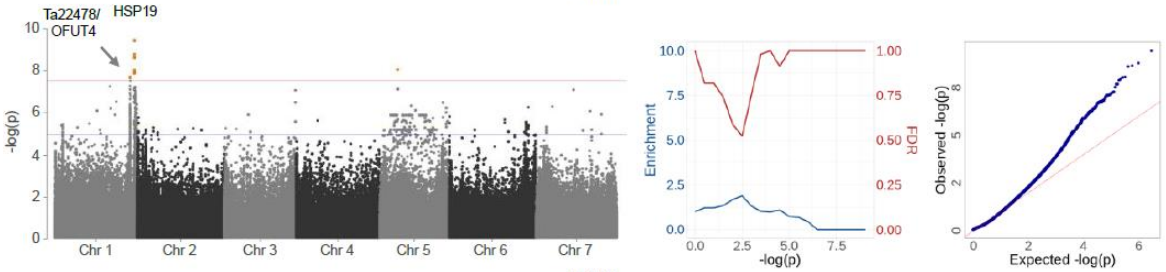
DTC



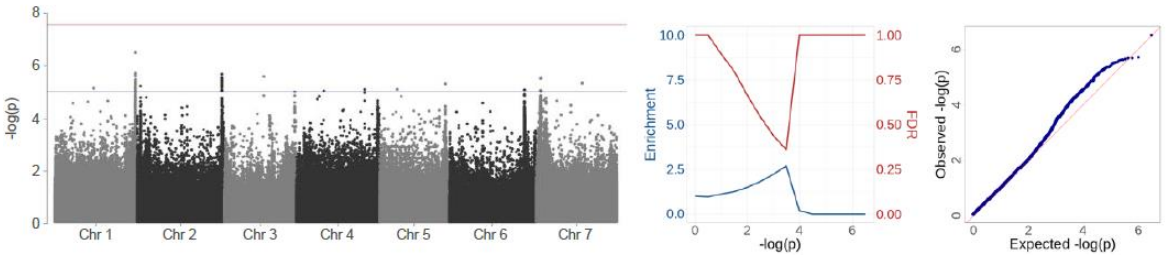
DTH



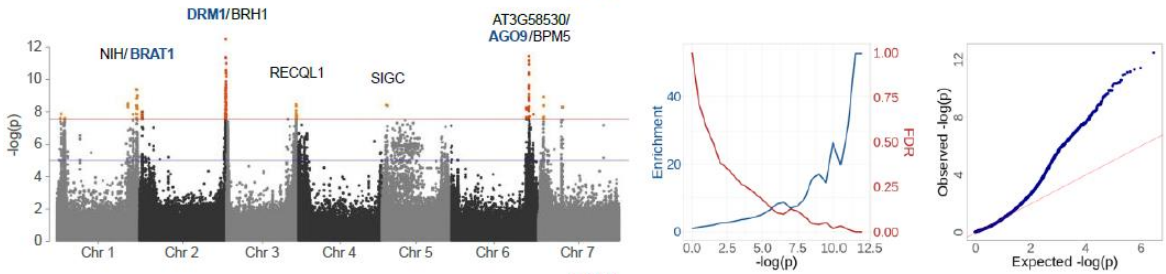
DTM



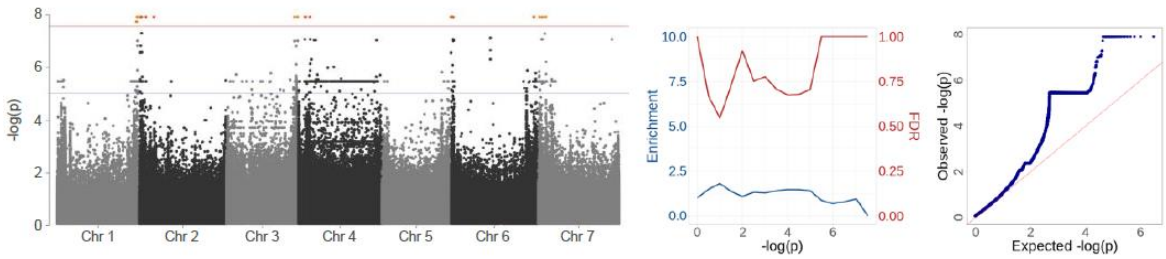
DTT



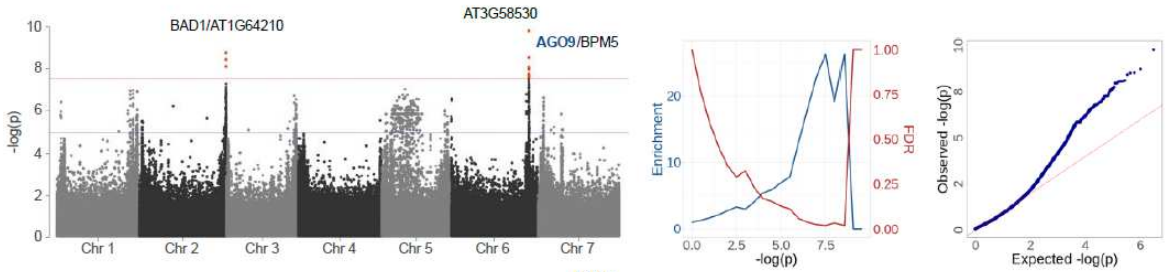
LINE



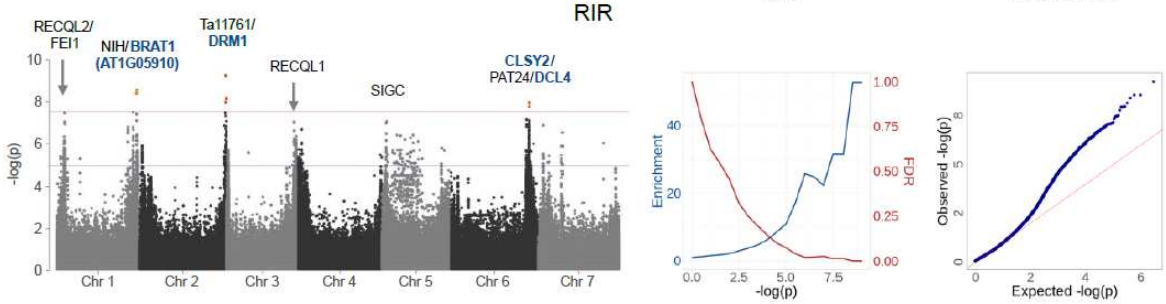
LTR

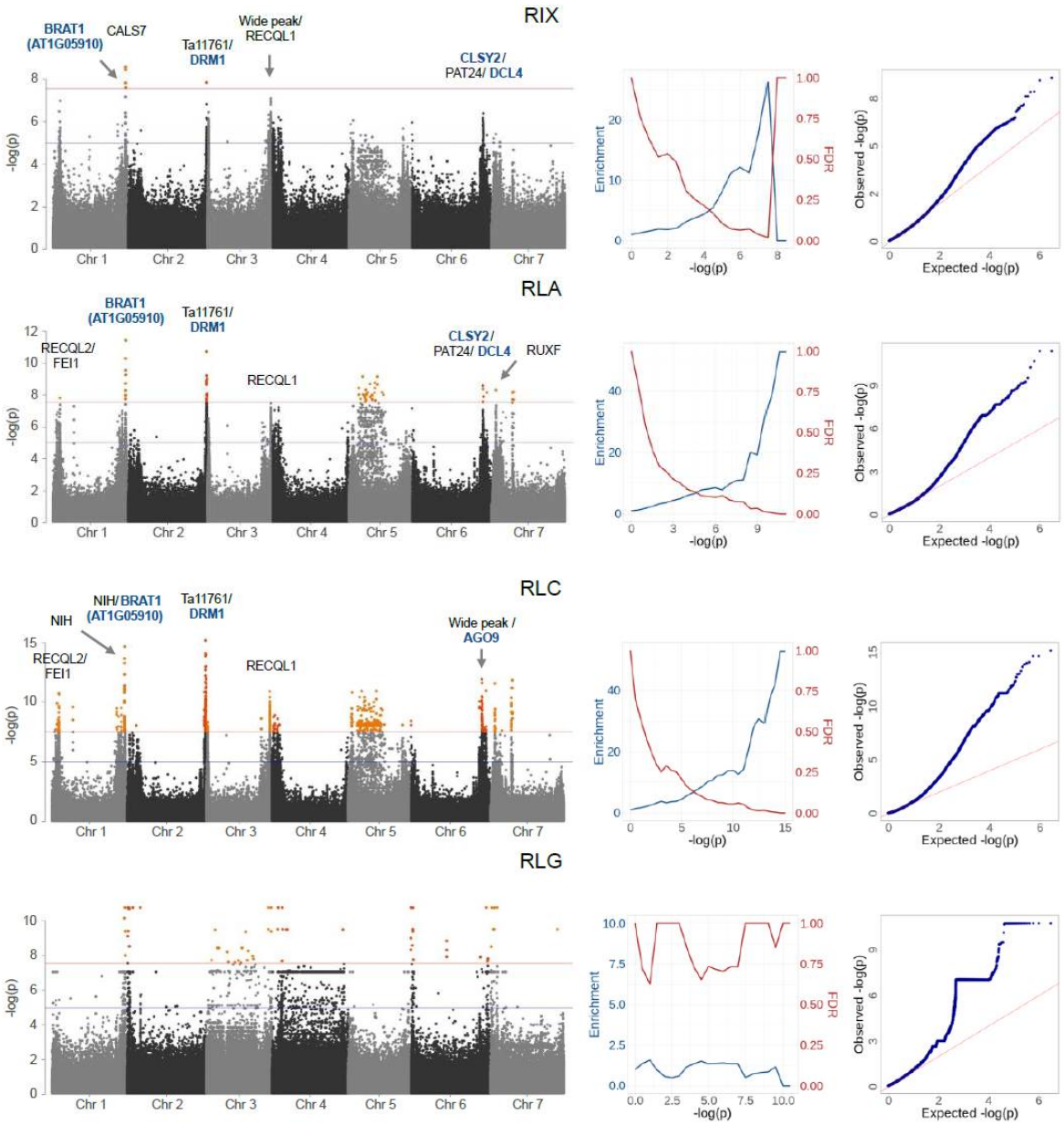


RIL

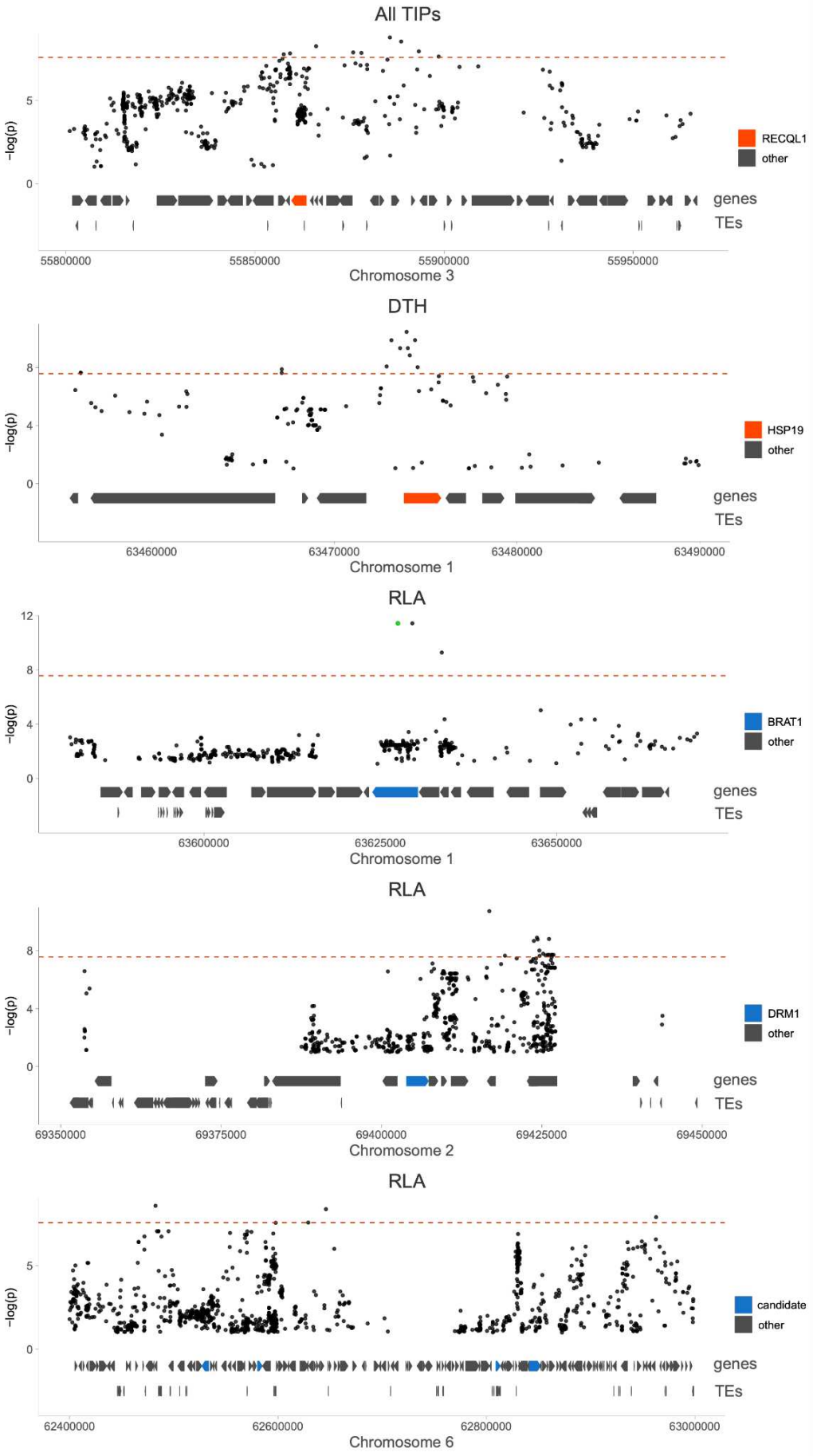


RIR

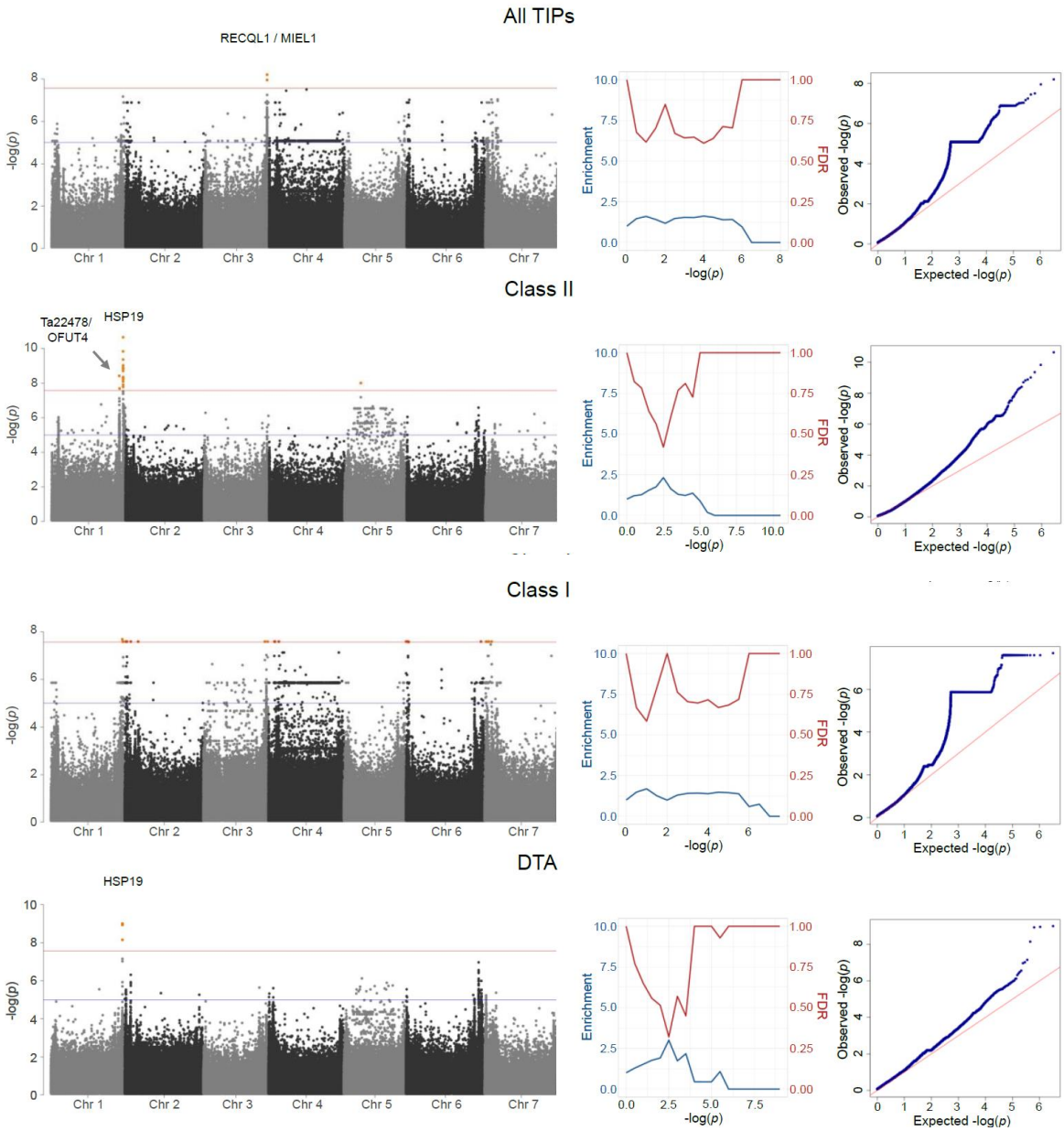




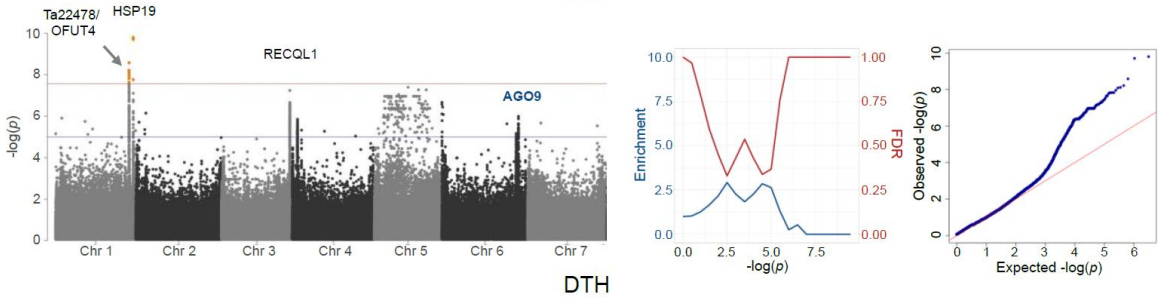
S7 Fig. Complete GWA results for TIP load. Left: Manhattan plots for each TIP superfamily load. The genome-wide significance (red line) corresponds to a full Bonferroni correction, the suggestive line (blue) to a more generous hard threshold of –log(p) =5. Genes next to top variants are labelled with 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)
RECQL1
other
genes
TEs
55800000
55850000
55900000
55950000
Chromosome 3
DTH
HSP19
63460000
63470000
63480000
63490000
Chromosome 1
RLA
BRAT1
63600000
63625000
63650000
Chromosome 1
RLA
DRM1
69350000
69375000
69400000
69425000
69450000
Chromosome 2
RLA
candidate
62400000
62600000
62800000
63000000
Chromosome 6

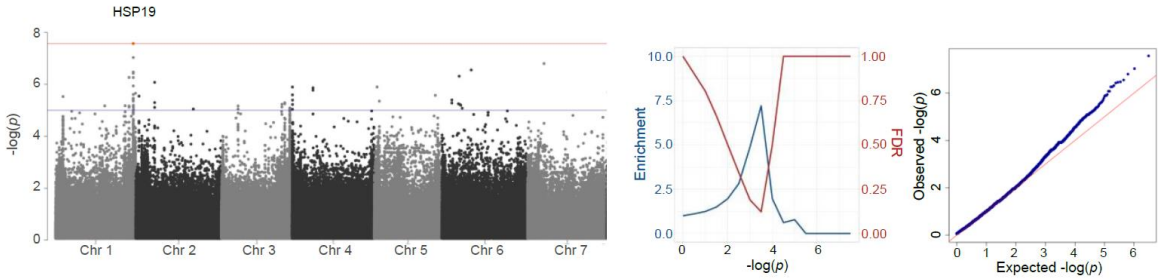
S8 Fig. Zoom-in of GWA peaks with candidate genes highlighted. The genome-wide significance (dotted red line) corresponds to a full Bonferroni correction. DNA methylation machinery genes used for the enrichment of *a-priori* candidates are depicted in blue, other genes that might affect transposition in red. The putative knock-out SNP disrupting the function of *BRAT1* is depicted in green.



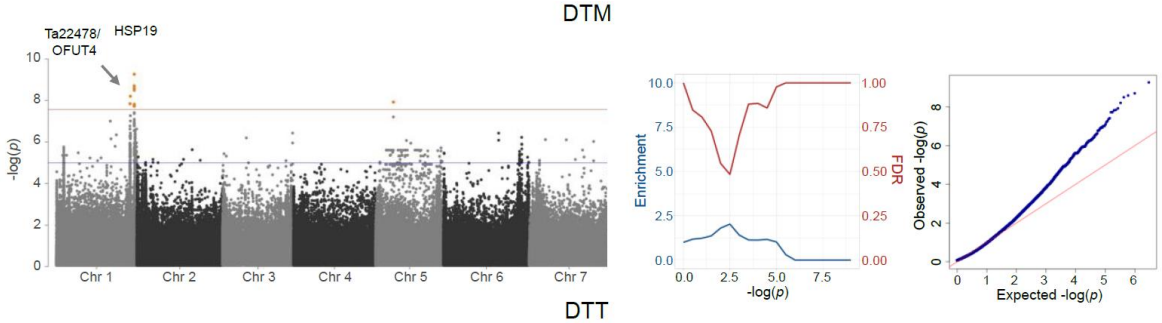
DTC



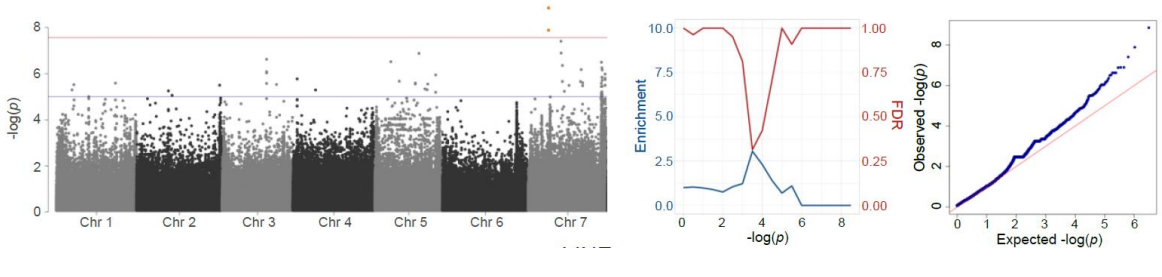
DTH



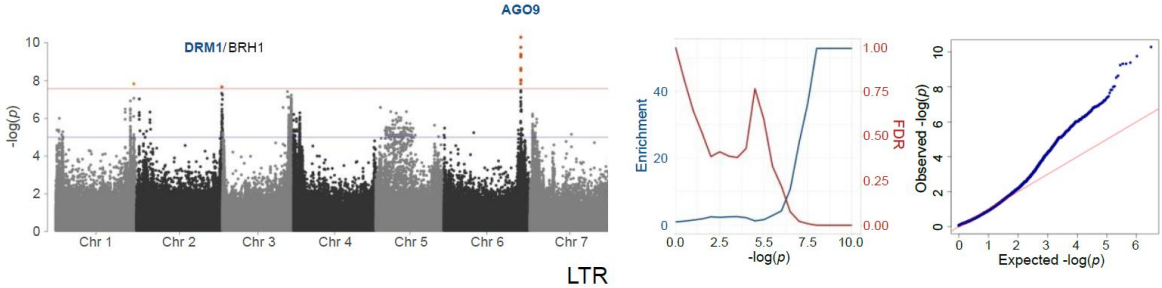
DTM



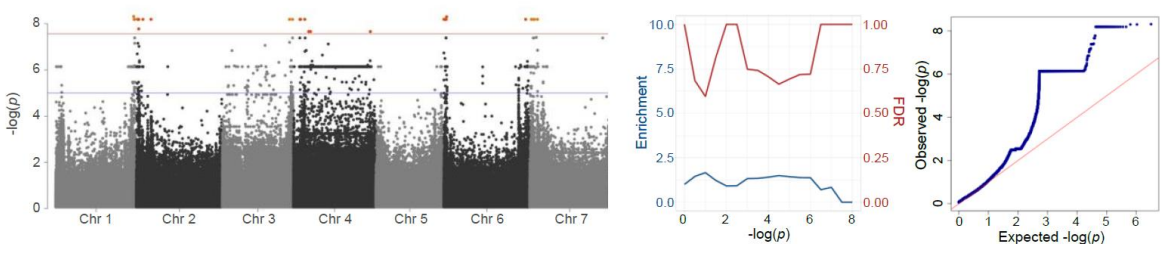
DTT



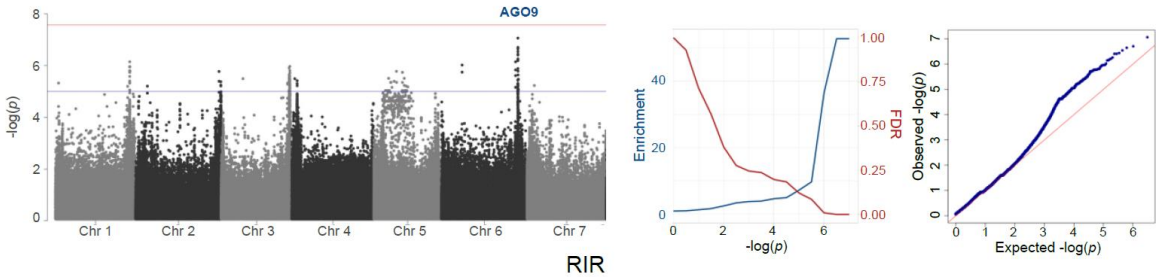
LINE



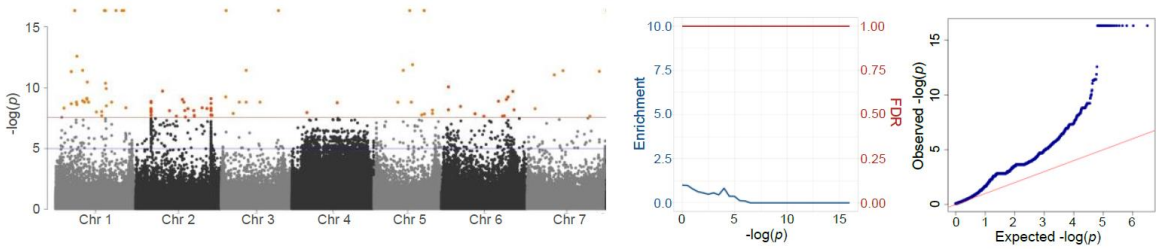
LTR



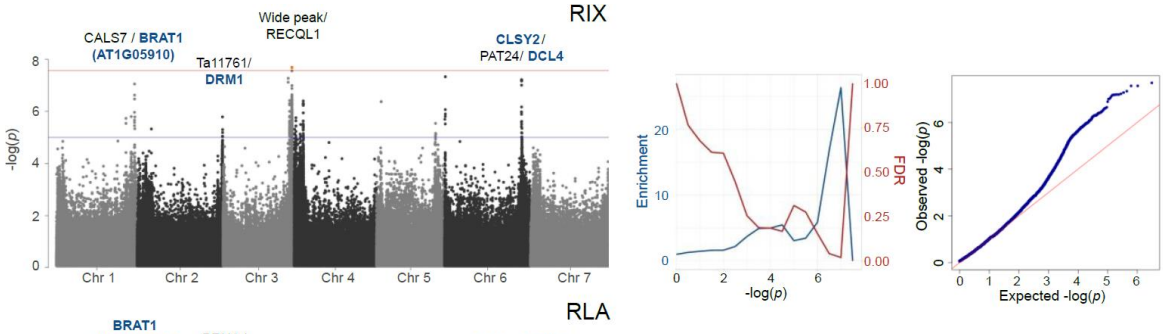
RIL



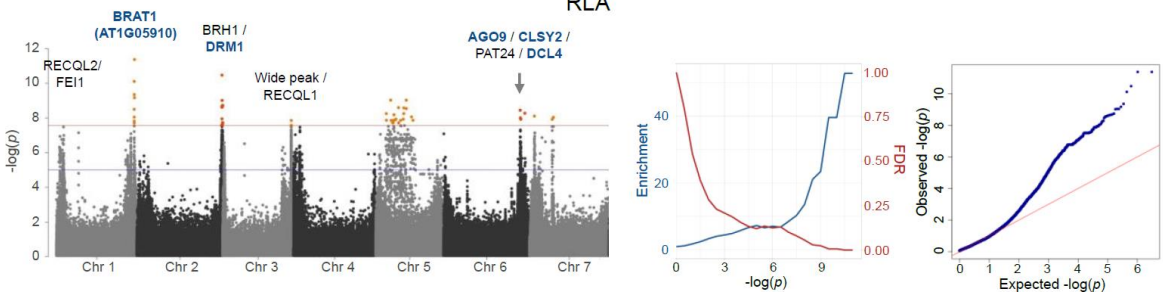
RIR



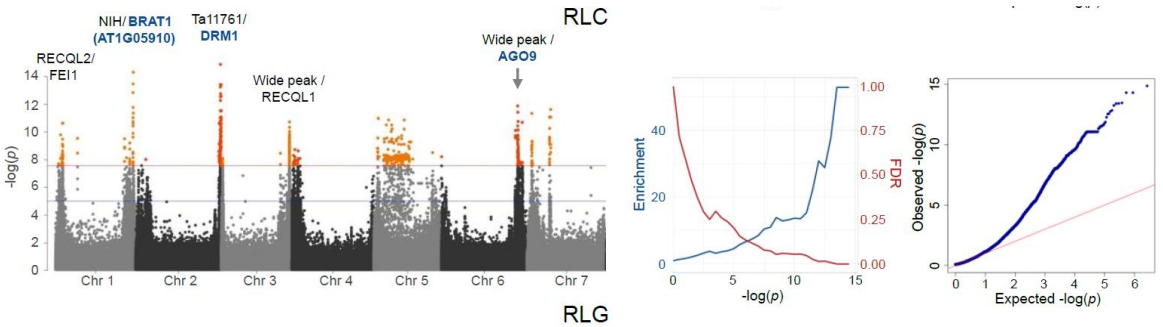
RIX



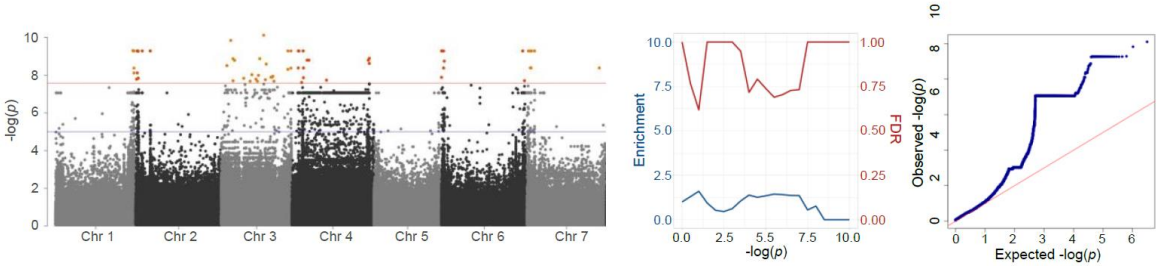
RLA



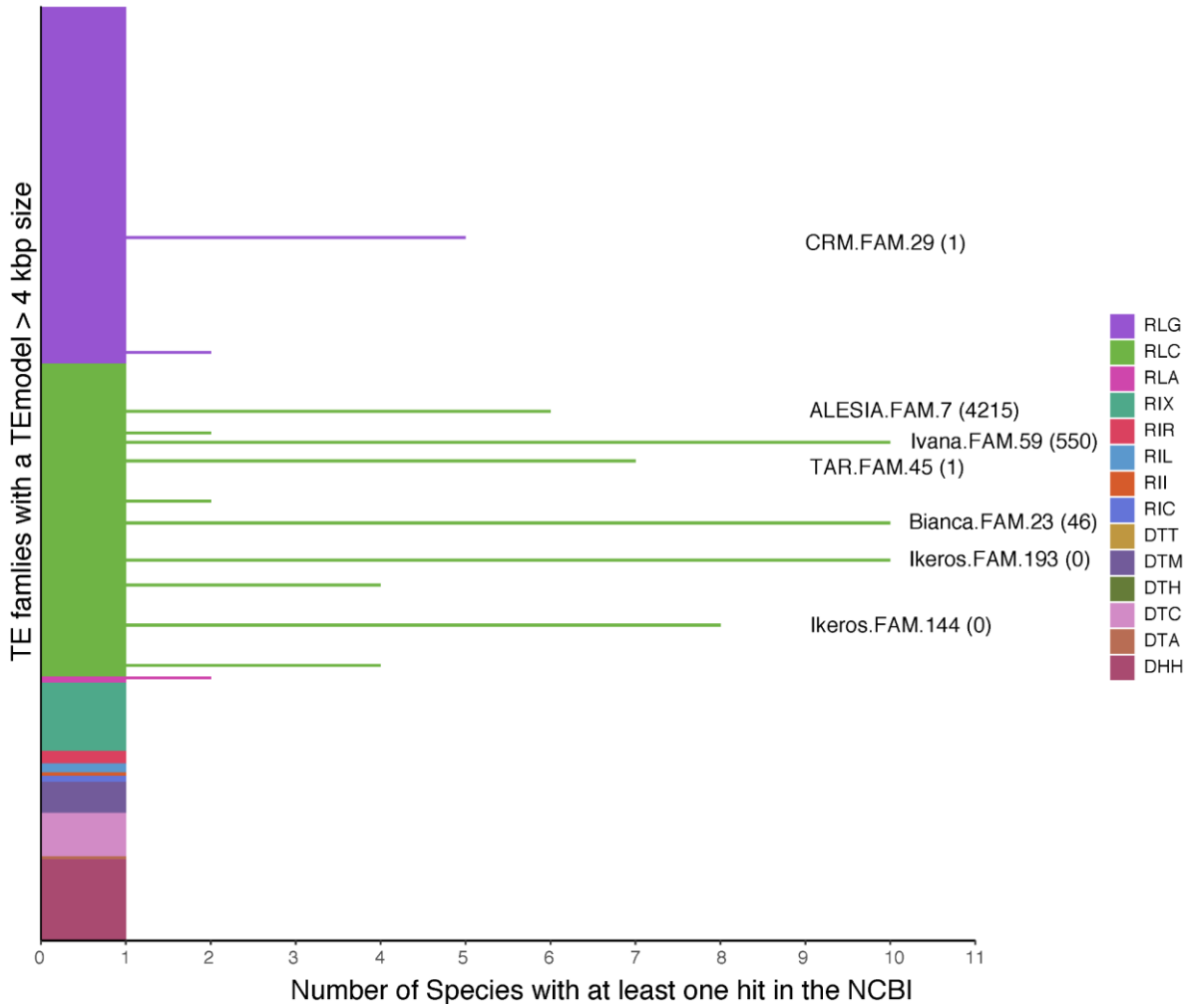
RLC



RLG



S9 Fig. GWA results for genome-wide load of TIPs younger than 500,000 years. Left: Manhattan plots for load of TIPs for each TE superfamily. 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 labelled with names, blue font indicates genes with link to DNA methylation included in the enrichment analyses. Middle: Enrichment and expected FDR of genes with links to DNA methylation, for significance threshold increments. Right: QQplots of p-values.



S10 Fig. BLASTN hits of *T. arvense* TE families with model sizes > 4kb against the NCBI NT database, 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 1hit. Each family has at least one hit, namely *T. arvense* itself. TE families with more than 5hits 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.

Chapter III

Discarded sequencing reads uncover natural variation in pest resistance in *Thlaspi arvense*

Dario Galanti, Jun Hee Jung, Caroline Müller, Oliver Bossdorf.

https://elifesciences.org/articles/95510

Discarded sequencing reads uncover natural variation in pest resistance in *Thlaspi arvense*

Dario Galanti[1,2], Jun Hee Jung[1], Caroline Müller[3] and Oliver Bossdorf[1]

[1] Plant Evolutionary Ecology, Institute of Evolution and Ecology, University of Tübingen, 72076 Tübingen, Germany.

[2] Royal Botanic Gardens, Kew, Richmond upon Thames, UK.

[3] Chemical Ecology, Bielefeld University, Universitätststr. 25, 33615 Bielefeld, Germany.

Key words

Aphids, DMRs, DNA methylation, EWAS, glucosinolates, GWAS, mildew, natural variation, pennycress, plant defense

Abstract

Understanding the genomic basis of natural variation in plant pest resistance is an important goal in plant science, but it usually requires large and labour-intensive phenotyping experiments. Here, we explored the possibility that non-target reads from plant DNA sequencing can serve as phenotyping proxies for addressing such questions. We used data from a whole-genome and -epigenome sequencing study of 207 natural lines of field pennycress (*Thlaspi arvense*) that were grown in a common environment and spontaneously colonized by aphids, mildew and other microbes. We found that the numbers of non-target reads assigned to the pest species differed between populations, had significant SNP-based heritability, and were associated with climate of origin and baseline glucosinolates content. Specifically, pennycress lines from cold and thermally fluctuating habitats, presumably less favorable to aphids, showed higher aphid DNA load, i.e. decreased aphid resistance. Genome-wide association analyses identified genetic variants at known defense genes but also novel genomic regions associated with variation in aphid and mildew DNA load. Moreover, we found several differentially methylated regions associated with pathogen loads, in particular differential methylation at transposons and hypomethylation in the promoter of a gene involved in stomatal closure, likely induced by pathogens. Our study provides first insights into the defense mechanisms of *Thlaspi arvense*, a rising crop and model species, and demonstrates that non-target whole genome sequencing reads, usually discarded, can be leveraged to estimate intensities of plant biotic interactions. With rapidly increasing numbers of large sequencing datasets worldwide, this approach should have broad application in fundamental and applied research.

Introduction

Plant pests, such as pathogens and herbivores, can cause major yield losses in crops and often require the massive use of pesticides to control their damage. Natural plant populations, on the other hand, are constantly exposed to such biotic stressors and their higher genetic diversity often allows these populations to become locally adapted. Since many pest species are sensitive to climatic conditions, their pressure on plant communities is spatially heterogeneous, maintaining geographically structured genetic variation in plant defenses (1,2). For these reasons, natural plant populations are highly suitable to study defense mechanisms and evolution of defenses, and also a very useful source of beneficial and resistance alleles for specific pathogens and environmental conditions. This genetic variation in defense-related genes can for example be screened through Genome Wide Association (GWA) (3–6) or approaches based on known candidate genes (2).

Many pest species are also highly sensitive to temporal variation in weather conditions. This temporal heterogeneity in pathogen pressure can induce plastic responses in plants, involving gene expression and epigenetic changes (7–9), which may also be studied through stress experiments (7–9). Some plastic epigenetic responses can have a transient stability and be transmitted to the next generations through inheritance of epigenetic marks (10–13). In particular, DNA methylation has been shown to respond to biotic and abiotic stresses through gene expression regulation and transposable elements (TEs) reactivation, and can be inherited across generations (9,12,14). In plants, DNA methylation can occur in the three sequence contexts CG, CHG and CHH (H being A, T or C), which differ in their molecular machineries depositing, maintaining and removing methylation and consequently also in their transgenerational stability (15,16). While CG methylation is usually more stable across generations, CHH methylation is less stable and more responsive to stress and the sensitivity of CHG methylation lies somewhere in between (15–17).

Whether inherited or induced, some strategies of plants for defense against pathogens and herbivores include: i) physical barriers such as reinforced cell walls, leaf protective layers or closing stomata, ii) production of specialized (secondary) metabolites that reduce palatability or are toxic to pests, iii) oxidative bursts, iv) the activation of signaling cascades to induce systemic responses and v) RNA interference mechanisms to silence pathogen genes (18–22). In Brassicaceae, a particularly important and diverse class of defense metabolites are glucosinolates, which often show local adaptation driven by variation in pests and can also be induced by herbivore and pathogen attacks (1,2,23).

Studying natural variation in plant resistance, along with associated genetic and epigenetic variation, can identify genes involved in defense and their regulators, including vital genes whose function cannot be determined through knockout experiments. Such knowledge, and especially the discovery

of natural resistance alleles, are crucial sources for the breeding of more pest-resistant crop varieties. Nevertheless, because of the diversity of resistance mechanisms and their often multigenic nature, plant defense mechanisms remain difficult to study. In particular, antixenosis (the prevention of pathogen settlement) and antibiosis (the repression of pathogen growth and reproduction) require extensive and time-consuming phenotyping, based for example on choice (24) or settling (9) assays, and such assays are extremely challenging to perform on large collections. On the other hand, there are increasing numbers of large sequencing datasets, which may also be used to quantify contaminants or microbiome composition (25–27) and thus as proxies for resistance phenotyping. In this study we investigated such usage of exogenous reads, i.e. reads not mapping to the target reference genome, as a source of information for quantifying herbivore and pathogen abundance in large collections.

We worked with field pennycress (*Thlaspi arvense*), an annual plant in the Brassicaceae family that is increasingly studied as a model species (28–32) and new biofuel and winter cover crop (33–36). In a previous study, we investigated natural epigenetic variation in a collection of 207 *Thlaspi* lines from across Europe (32). Prior to their whole-genome and -epigenome sequencing these lines had been grown in a common environment, an open glasshouse where the plants were spontaneously colonized by aphids and powdery mildew, as well as by other microbes. At the time of sequencing, pathogen contamination was still very limited but appeared highly variable, and preliminary analyses showed that it resulted in sizeable amounts of non-target reads assigned to the pest species, i.e. contamination of the DNA samples. Inspired by other recent studies on non-target reads (25–27), we asked if there was systematic variation in the numbers of aphid and pathogen reads among different *T. arvense* lines, and whether these data, together with our whole-genome plant sequencing data, could provide insights into the genomic basis of plant resistance variation.

The goals of our study were thus two-fold: i) to contribute to a mechanistic understanding of pest resistance in *Thlaspi arvense*, and ii) to explore whether non-target reads from plant sequencing can be used as proxies for studying plant biotic interactions. Considering that we are moving towards an increasingly sequencing-prone world, with more and larger datasets being generated for many species (37–43), the use of non-target reads has very broad potential.

Results

Reads classification and species identification

Starting from our previously published sequencing data (32), the first step of our analysis was to separate the Whole Genome Sequencing (WGS) reads of each sample into the ~99.5% mapping to the *Thlaspi arvense* reference genome (29) and the ~0.5% that did not, hereafter called "exogenous reads" (Fig 1A). Initially, we used all mapped reads for calling variants in *Thlaspi*, but after some difficulties with Genome Wide Associations (see below) we suspected that some plant reads were false and mapped to the *T. arvense* genome only because of the high cross-taxa similarity of some genomic regions. We therefore remapped all reads to the genomes of the aphid *Acyrthosiphon pisum*, its endosymbiont *Buchnera aphidicola* and the powdery mildew *Blumeria graminis*, and found that, on average, 7.4% of the reads mapped to both *T. arvense* and at least one of the pests. We removed these ambiguous reads from our analyses and used only the *T. arvense* target reads, 92.1% on average, for variant calling (Fig 1A, S1 Table).

We next attempted a taxonomic classification of the exogenous reads, in multiple steps. First, we used MG-RAST (44,45) to assign reads to taxonomic groups based on public sequencing databases. Out of the 78% of the exogenous reads that passed the MG-RAST quality control (S1 Table) the majority belonged to bacteria and smaller fractions to fungi, plants and animals (Fig 1B and S2 Table). For subsequent group-level analyses, we then focused on nine taxonomic groups that occurred consistently within our samples (Fig 1C), and that were particularly abundant or relevant: Erysiphales (fungi), Peronosporales (oomycetes), Aphididae and Culicidae (both insects), and five bacterial families.

Visual inspection (Fig 1D) and other sources of information narrowed down the observed aphid and mildew species to a few candidates. For aphids we considered *Acyrthosiphon pisum* (indicated by MG-RAST), *Myzus persicae* (visual match, and a generalist attacking Brassicaceae (46)) and *Brevicoryne brassicae* (attacks Brassicaceae including *Thlaspi* (47)). For powdery mildew we considered *Blumeria graminis* (indicated by MG-RAST), and *Erysiphe cruciferarum* (attacks Brassicaceae (48)). To decide among these species, we then used a competitive mapping approach (49), where the exogenous reads were aligned to a pseudo-reference composed of the same DNA sequences from the different candidate species (see Methods for details, S3 and S4 Tables). The majority (77%) of the aphid reads mapped to *M. persicae*, 18% to *B. brassicae*, and 5% to *A. pisum*, while 98% of the mildew reads mapped to *E. cruciferarum*, with only 2% to the other mildew species (S5 Table). Based on these results, we concluded that the plants in our experiment had been attacked by *Myzus persicae* and *Erysiphe cruciferarum*.

Finally, to compare the power of a large database approach (MG-RAST) versus using specific reference genomes, we also remapped all exogenous reads to the *M. persicae* and *B. aphidicola* genome assemblies (50) (www.ncbi.nlm.nih.gov/assembly/GCA_001939165.1) and used the counts from these two mappings as additional phenotypes, besides the nine taxonomic groups selected through MG-RAST (Table 1, S6 Table).

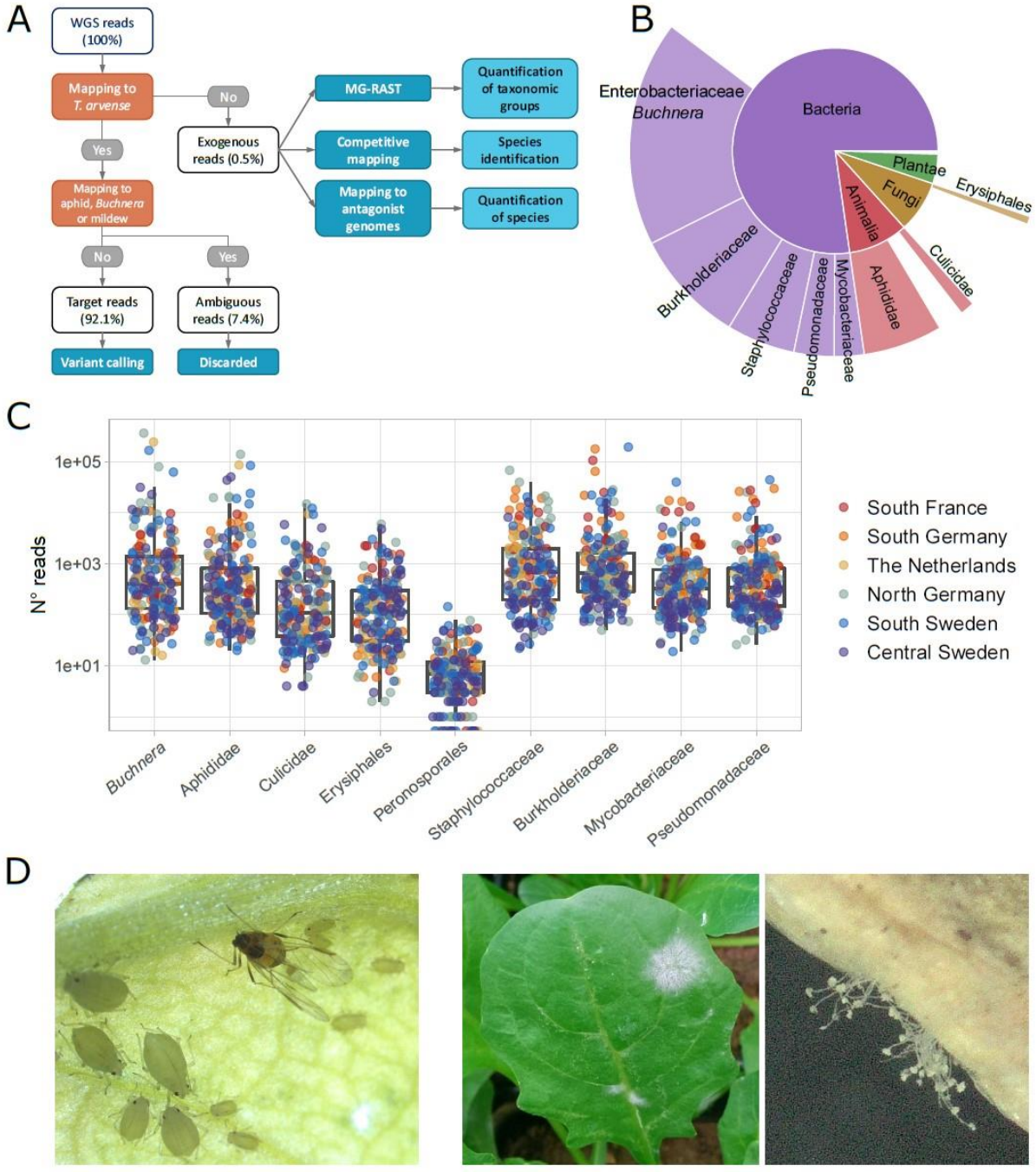


Fig 1. Classification of sequencing reads in *Thlaspi arvense* WGS data. (A) Workflow of the analyses, including reads classification (orange nodes) into target, ambiguous and exogenous reads, and downstream analysis (dark blue nodes) (see Methods). (B) Fractions of exogenous reads assigned to different taxonomic groups by MG-RAST (44,45). (C) Read counts

assigned to nine selected groups in our 207 *T. arvense* samples from different European regions. (D) Aphids and mildew occurring on *T. arvense* leaves during our experiment.

Exogenous read counts are a heritable Thlaspi phenotype

As we had observed that aphid and mildew infections in the glasshouse were not random, but prevalent on plants from some origins than others, i.e. possibly reflecting heritable variation in plant resistance, we next tested for population differences and SNP-based heritability in pest and microbiome loads (see Methods). Prior to these analyses, to avoid biases caused by different sequencing depths, we corrected the read counts for the total numbers of deduplicated reads in each library and used the residuals as unbiased estimates of aphid, mildew and microbe loads.

For most of the nine taxonomic groups, there were significant population effects, with 20-40% of the variance in read counts explained, as well as significant SNP-based heritability, typically in the range of 0.18 - 0.30 (Table 1). The highest heritability of 0.47 was for read counts of Erysiphales, indicating particularly strong variation for resistance to mildew. Both SNP-based heritability and population differences tended to be stronger for aphid and *Buchnera* data based on read mapping to the reference genomes than for those based on MG-RAST, demonstrating that the former method is stronger and thus preferable if high-quality genome assemblies are available.

An alternative explanation for different aphid and mildew loads in the greenhouse could be that enemy variation in the field was transmitted to the greenhouse, through maternal carry-over effects, or even as seed contamination. However, we had recorded aphid and mildew occurrence during seed sampling in the field and found no significant differences in the glasshouse between the offspring of plants that had been attacked in the field versus those that had not (S1 Fig).

Table 1: Population differences and SNP-based heritability for different types of exogenous read counts. Population differences were tested with a linear model, SNP-based heritabilities (and their confidence intervals) estimated with the R package *heritability*.

Taxonomic group	Data type	Population differences (R^2 and *P*-value)	SNP-based heritability
Myzus persicae	Mapping to reference genome	0.245 (*P* = 0.029)	0.190 (0.055-0.488)
Buchnera aphidicola	Mapping to reference genome	0.256 (*P* = 0.016)	0.169 (0.042-0.490)
Buchnera	MG-RAST - genus	0.223 (*P* = 0.090)	0.115 (0.016-0.505)
Aphididae	MG-RAST - family	0.226 (*P* = 0.082)	0.189 (0.052-0.496)

Culicidae	MG-RAST - family	0.166 (P = 0.519)	0.183 (0.055-0.465)
Erysiphales	MG-RAST - order	0.326 (P < 0.001)	0.468 (0.238-0.712)
Peronosporales	MG-RAST - family	0.253 (P = 0.020)	0.266 (0.096-0.553)
Staphylococcaceae	MG-RAST - family	0.390 (P < 0.001)	0.301 (0.124-0.567)
Burkholderiaceae	MG-RAST - family	0.275 (P = 0.005)	0.256 (0.092-0.538)
Mycobacteriaceae	MG-RAST - family	0.362 (P < 0.001)	0.294 (0.120-0.560)
Pseudomonadaceae	MG-RAST - family	0.273 (P = 0.006)	0.192 (0.052-0.505)

Aphid and mildew loads correlate with climate of origin and glucosinolates content of plants

Having established that our method most likely captured variation in plant resistance, we were interested in the ecological drivers of this variation. As climate is known to be a major influence on many biotic interactions as well as plant defenses (1,51), we correlated the observed read counts with the climates of origin of the plants. We found negative correlations between aphid read counts and several temperature variables, in particular annual minimum temperature (Fig 2A). Aphid read counts were also positively correlated with temperature variability, i.e. the diurnal and annual ranges of temperature (Fig 2A). In other words, plants from warmer and more stable climates had consistently lower levels of aphid infestation in our glasshouse, possibly because these plants had evolved greater resistance under such benign climatic conditions where aphids thrive. We found similar, although weaker patterns, for the number of Erysiphales reads. The other analysed taxonomic groups showed different and often weaker patterns of correlation with climate, except that the read counts of several bacterial groups were positively correlated with annual maximum temperature and in particular diurnal temperature range.

Since glucosinolates are major defense metabolites of Brassicaceae, and their variation could thus be an explanation for variance in plant resistance, we also tested for correlations between the baseline amounts of these metabolites and the frequencies of aphid and mildew reads. Glucosinolate levels were measured on the same *T. arvense* lines in a separate experiment not affected by pests (S7 Table). We found positive correlations of aphid read counts with sinigrin, an aliphatic glucosinolate which is by far the most abundant in the leaves of *T. arvense*, and a stronger negative correlation with benzyl glucosinolates (glucotropaeolin) (Fig 2B). Although the baseline levels of benzyl glucosinolates were very low and probably sometimes below the detection level, plant lines where benzyl glucosinolate was detected had significantly lower aphid loads (over 70% less reads) in the glasshouse (Fig 3C). We

also detected three indole glucosinolates, but these did not show any significant correlations with aphid loads.

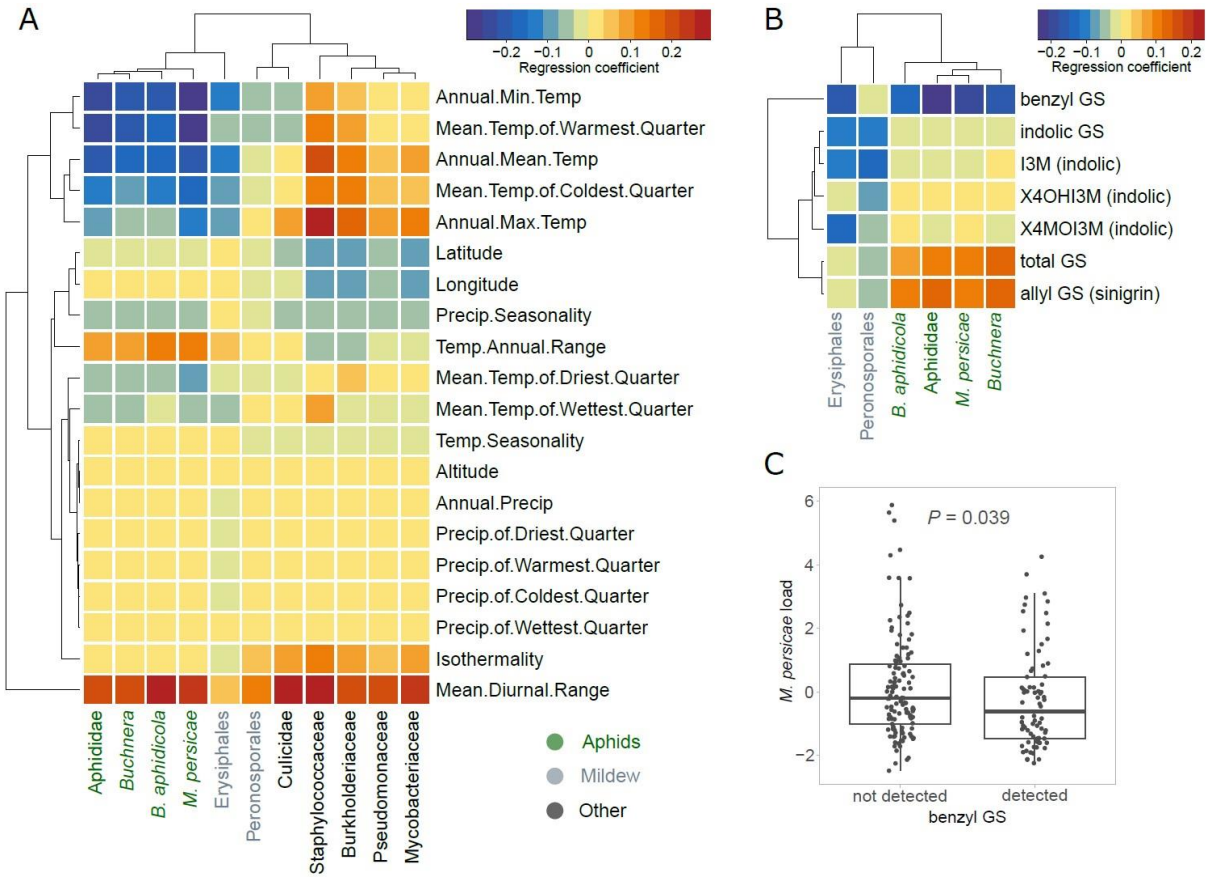


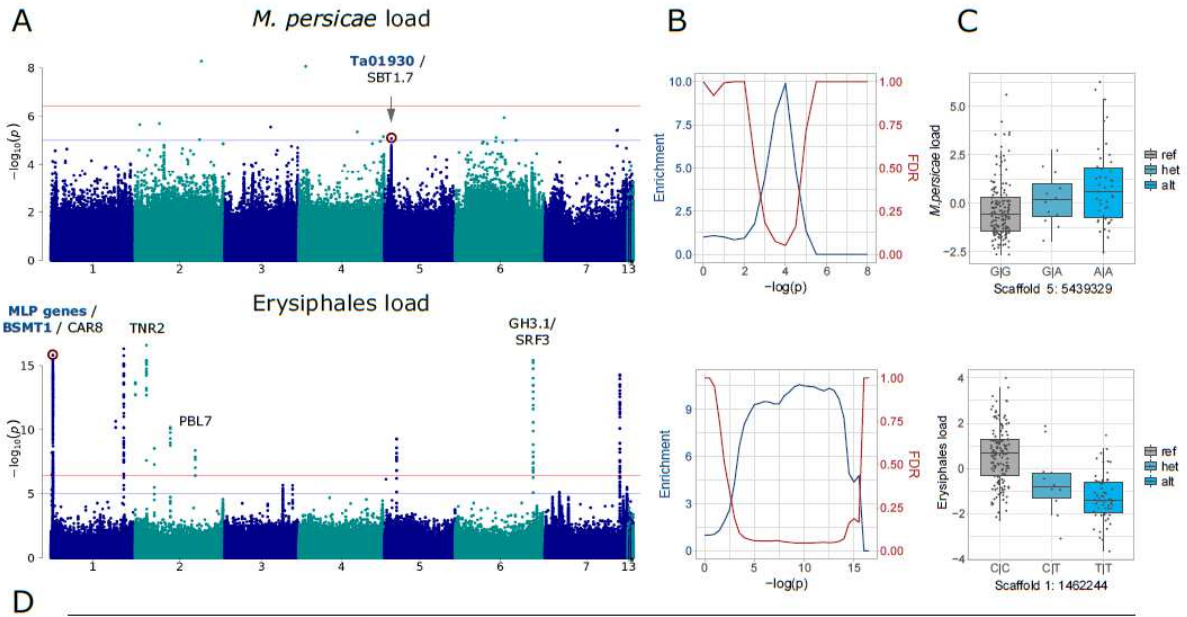
Fig 2. Relationships between climates of origin or glucosinolate levels of plants and the exogenous reads loads. (A) Correlations with bioclimatic variables. (B) Correlations with baseline glucosinolate (GS) levels measured in the same pennycress lines in another experiment. All correlations in (A) and (B) were done after correction for population structure. Aphid-related read counts are in green, mildew-related in gray, others in black. (C) Boxplot of the aphid reads residuals in samples where benzyl GS was detected vs. not.

GWA identifies peaks near defense genes

To further investigate the genetic basis of variation in aphid, mildew and microbe loads, we next employed GWA and tested for associations between exogenous read counts and biallelic genetic variants (SNPs and short INDELs). We corrected for population structure using an IBS matrix and only tested variants with Minor Allele Frequency (MAF) > 0.04 (see Methods). Initially, we called genetic variants using all reads that mapped to the *T. arvense* genome and found massive peaks in some highly conserved regions of the genome, which had very high mapping coverage (S2 Fig). We suspected that this might be because some non-*Thlaspi* reads were very similar to these highly conserved regions and, by mapping there, generated false variants only in samples containing many non-*Thlaspi* reads. We

therefore identified and removed ambiguous reads prior to variant calling, which eliminated the observed massive GWA peaks, indicating that they had indeed reflected false associations (S2 Fig).

After excluding the ambiguous reads, we still found significant GWA peaks for Erysiphales but not for other types of exogenous reads (excluding isolated, unreliable variants) (Fig 3A and S3 Fig). Nevertheless, when clear peaks were visible, regardless of their significance, they were usually located close to genes involved in plant defense response. An enrichment analysis (52) confirmed that stronger variants were indeed enriched close to these defense genes (S8 Table) for some exogenous read counts (Fig 3B and S3 Fig). For *M. persicae* load there was a peak in the proximity of *Tarvense_01930*, encoding a predicted pathogenesis-related peptide. The top variant in this peak had a slight but clear allelic effect on *M. persicae* load (Fig 3C). For Erysiphales load we detected a more persistent enrichment, with a highly significant peak in Scaffold 1, located in a region with several defense genes, including *MAJOR LATEX PROTEINS (MLP)* and two genes similar to *Arabidopsis thaliana SALICYLATE METHYLTRANSFERASE 1 (BSMT1)* (Fig 3D and E). This region is wide due to ancient TE colonization, but the top variants are clearly neighboring candidate genes involved in defense (Fig 3E). Other significant peaks for Erysiphales load were close to other genes that possibly contribute to resistance such as *PBL7*, involved in signaling and stomatal closure or *SRF3*, reinforcing cell walls by callose deposition.



D

Gene name	Gene ID	Ortholog	Putative function	Top SNP location
MLP 31	Ta_10442/45/55	AT1G70840	MAJOR LATEX PROTEIN 31 – Defense response	8 / 2kb downstream
MLP 28	Ta_10443/54	AT1G70830	MAJOR LATEX PROTEIN 28 – Defense response	6 / 1kb downstream
MLP 43	Ta_10444	AT1G70890	MAJOR LATEX PROTEIN 43 – Defense response	4kb downstream
MLP 165	Ta_10446	AT1G35260	MAJOR LATEX PROTEIN 165 – Defense response, GS biosynthesis	< 1kb downstream
BSMT1	Ta_10447/48	AT3G11480	Salicylate/benzoate methyltransferase – Defense response	0kb
CAR8	Ta_10449/53	AT1G23140	Abscisic acid-activated signaling pathway	0 / 2 kb upstream
TNR2	Ta_13332	AT5G46700	Root and epidermis development	1kb downstream
PBL7	Ta_12648	AT5G02800	Brassinosteroid signalling pathway, stomatal closure.	1 kb upstream
Ta01930	Ta_01930	AT3G50020	Predicted PR (pathogenesis-related) peptide – Defense response	12 kb downstream
SBT1.7	Ta_01928/29	AT5G67360	Serine protease essential for mucilage release from seed coats	< 1kb upstream
GH3.1	Ta_05347	AT2G14960	Similar to IAA-amido synthases – Auxine response	< 1kb downstream
SRF3	Ta_05348	AT4G03390	SR FAMILY 3 – Cell wall callose deposition and stress response	3kb upstream

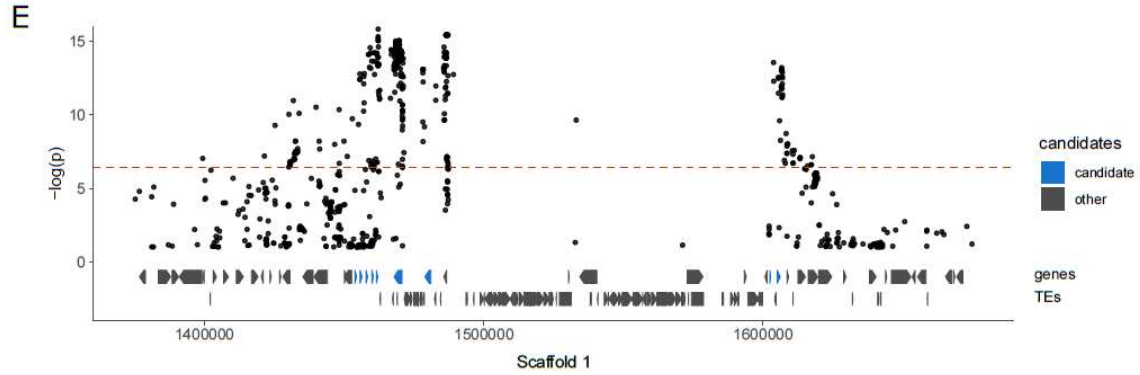


Fig 3. Genome-wide association analyses for aphid and mildew loads. We show only the results for *M. persicae* and MG-RAST Erysiphales read-counts; for full results see S3 Fig. (A) Manhattan plots, annotated with genes potentially affecting aphid/mildew colonization. The genome-wide significance (horizontal red line) was calculated based on unlinked variants (53), the blue line corresponds to –log(p) = 5. (B) Corresponding to the Manhattan plots on the left, enrichment of *a priori* candidates and expected false discovery rates (as in (52)) for increasing significance thresholds. (C) Allelic effects of the red-marked variants in the corresponding Manhattan plots, with genotypes on the x-axes and the read-count residuals on the y-axes. (D) The candidate genes marked in panel A, their putative functions and distances to the top variant of the neighboring peak. Candidates in dark blue are the *a priori* candidates included in the enrichment analyses and involved in defense

response (GO:0006952). GS: glucosinolates. (E) Zoom-in for the Manhattan plot of Erysiphales load, around the first peak in Scaffold 1, with gene and TE models below, and *a priori* candidates in blue.

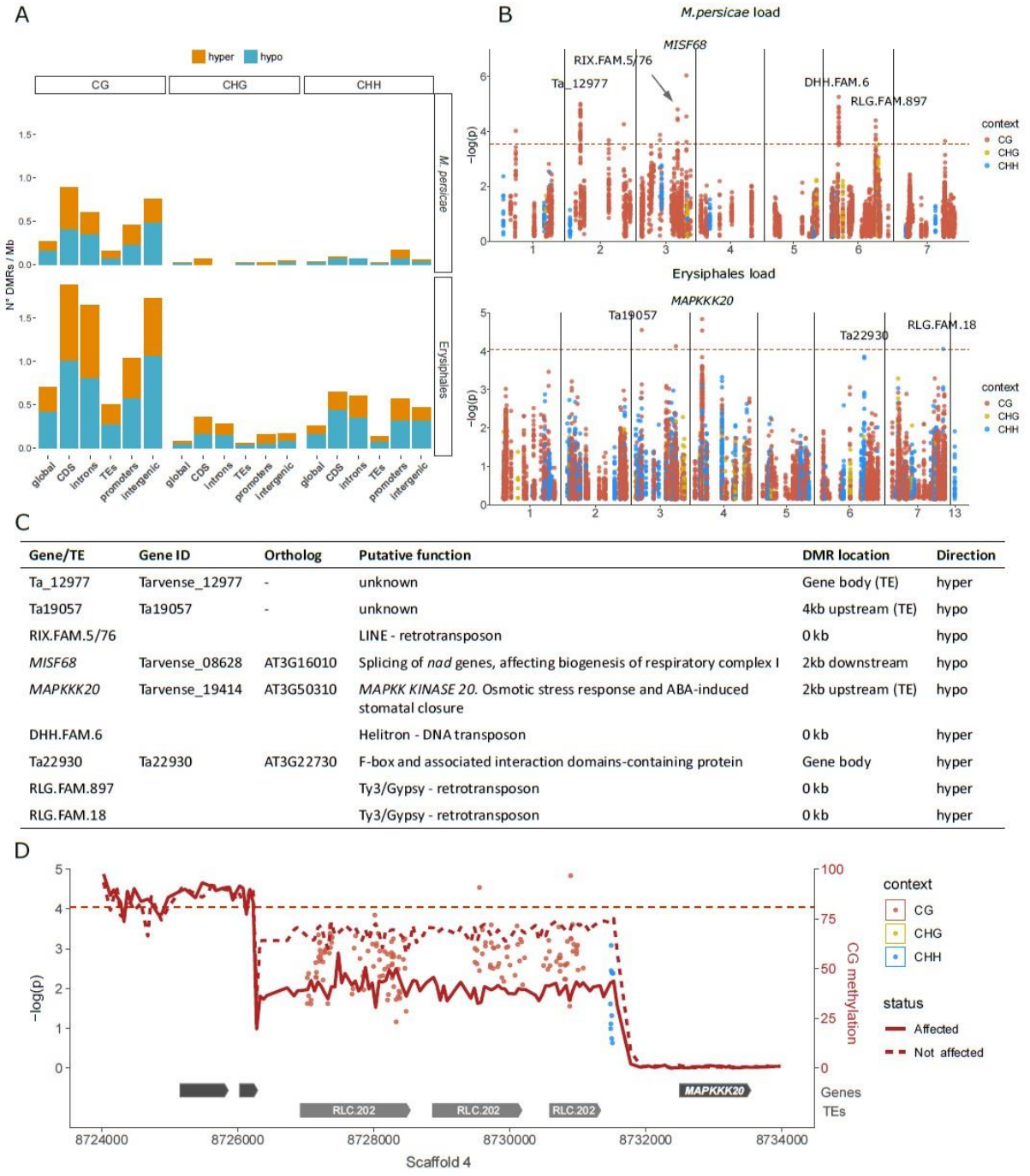
Aphid and mildew loads correlate with differential methylation at genes and transposons

Variation in phenotypes, such as our indirect estimates of pest resistance, may not only be associated with DNA sequence but also with epigenetic changes like DNA methylation. This phenotype-associated epigenetic variation can include both heritable and plastic components. The Whole Genome Bisulfite Sequencing (WGBS) data from our previous study (32) allowed us to also explore these questions and to test for associations between DNA methylation variation and pest attack. For simplicity, we limited this analysis to *M. persicae* and Erysiphales loads.

Our analysis had two steps: First we called Differentially Methylated Regions (DMRs) between the 20 samples with the most and least *M. persicae* or Erysiphales loads, and then we conducted Epigenome Wide Association (EWA) analyses on individual positions located within these DMRs (see Methods). This approach allowed us to target genomic regions of interest, while strongly reducing the multiple-testing problem of millions of cytosines in the whole genome and correcting for population structure. Using a relaxed False Discovery Rate (FDR) of 20%, we identified 162 DMRs for *M. persicae* load and 548 DMRs for Erysiphales load (S4 Fig, S9 and S10 Table). The majority of these were in the CG context, especially for *M. persicae*-related DMRs (Fig 4A and S4 Fig). As observed previously (32), DMRs in CHH were generally shorter than in the other sequence contexts (S4A Fig). Since the genome of *T. arvense* is rich in TEs and intergenic regions, the majority of DMRs were located in those features (S4B Fig). However, the DMR density was higher in proximity of genes and particularly in coding sequences (Fig 4A), and even DMRs assigned as intergenic (Fig 4A) were often located close to genes or promoters. In accordance with previous studies (8,9), most DMRs were hypomethylated in the affected samples, indicating that genes needed for defense might be activated through demethylation.

For a more detailed investigation, we turned to EWA, leveraging the power of the entire *Thlaspi* collection. We tested for associations between *M. persicae* or Erysiphales loads and the methylation at individual cytosines located within the DMRs. As in GWA, we corrected for population structure using an IBS matrix. For both types of pest loads, we found associations in the proximity of genes and especially within TEs, but no genomic feature was particularly enriched for low *P*-value associations (S5A Fig). *M. persicae* load was associated with methylation at several genomic locations, especially TEs (Fig 4B), but these associations had strongly inflated *P*-values (S5B Fig). For Erysiphales load the *P*-value distribution was more well-behaved (S5B Fig), and we found a clear association with hypomethylation of *Copia* family 202 TEs upstream of *MAPKK KINASE 20* (*MAPKKK20*), a gene involved

in abscisic acid (ABA) stress response and stomatal closure (Fig 4B, C and D). A coverage analysis confirmed that none of the *T. arvense* lines carries insertions or deletions of the TEs upstream of *MAPKKK20*.



C

Gene/TE	Gene ID	Ortholog	Putative function	DMR location	Direction
Ta_12977	Tarvense_12977	-	unknown	Gene body (TE)	hyper
Ta19057	Ta19057	-	unknown	4kb upstream (TE)	hypo
RIX.FAM.5/76			LINE - retrotransposon	0 kb	hypo
MISF68	Tarvense_08628	AT3G16010	Splicing of *nad* genes, affecting biogenesis of respiratory complex I	2kb downstream	hypo
MAPKKK20	Tarvense_19414	AT3G50310	*MAPKK KINASE 20.* Osmotic stress response and ABA-induced stomatal closure	2kb upstream (TE)	hypo
DHH.FAM.6			Helitron - DNA transposon	0 kb	hyper
Ta22930	Ta22930	AT3G22730	F-box and associated interaction domains-containing protein	Gene body	hyper
RLG.FAM.897			Ty3/Gypsy - retrotransposon	0 kb	hyper
RLG.FAM.18			Ty3/Gypsy - retrotransposon	0 kb	hyper

Fig 4. Differential methylation associated with aphid and mildew loads. (A) Differentially Methylated Region (DMR) densities in different genomic features when comparing the 20 samples with the most vs. the least *M. persicae* (top) or Erysiphales (bottom) load. CDS: coding sequences. (B) Manhattan plots from EWA analyses based on individual cytosines within DMRs, with sequence contexts in different colours and annotation of genes close to low *P*-value cytosines. The genome-wide Bonferroni significance thresholds (dashed red lines) were calculated based on the number of DMRs. (C) Candidate genes and TEs marked in panel B, their putative functions, genomic locations of associated DMRs, and whether affected samples were

hyper- or hypomethylated. (D) Zoomed-in Manhattan plot for Erysiphales load around the peak in Scaffold 4, with gene and TE models given below. The CG methylation in the 20 most and least affected samples was calculated over 50 bp bins (see S5C Fig for methylation in other contexts).

Discussion

Plant pests are a major threat to food safety, causing large yield losses, and new crops such as the potential biofuel plant *Thlaspi arvense* must be able to resist pathogen and herbivore attacks. A powerful source for obtaining resistant varieties is natural variation in plant defenses, but phenotyping large collections can be very time-consuming and error-prone. Here we describe how an unplanned pest infestation in a glasshouse experiment, together with available WGS data, can be used to estimate aphid, mildew and microbial loads, and thus variation in plant resistance. The approach is straightforward, makes use of WGS data without microbiome-specific DNA extraction, and can in principle be applied to many other situations such as field experiments. It is not error-free, but we highlight some potential pitfalls, show how to reduce noise, and illustrate its potential to detect associations with climatic, genetic and epigenetic variation.

An important first step in our analyses was the identification and classification of pest-related reads in the plant WGS data. We began by classifying all reads as target (only mapping to *T. arvense*), ambiguous (mapping to *T. arvense* and at the same time to at least one of the pest genomes) or exogenous (not mapping to *T. arvense*) (Fig 1A). We demonstrated the importance of removing ambiguous reads prior to variant calling, as this prevented calling false positive variants caused by exogenous DNA that also mapped to highly conserved or repetitive sequences in the *T. arvense* genome. We then classified the exogenous reads using MG-RAST (44,45) or by confident mapping to specific pest genomes, and selected the eleven most relevant and/or abundant taxonomic groups to focus our analyses on. To obtain unbiased pest/microbe loads we also corrected the read-counts for the total number of deduplicated reads of each sample. A competitive mapping approach allowed us to identify the aphid and mildew species that had occurred in our experiment as the generalist aphid *Myzus persicae* and the powdery mildew *Erysiphe cruciferaum* (Fig 2).

Since we suspected a non-random colonization of pests and microbes in our *T. arvense* collection, we tested for population differences as well as SNP-based heritability. We found significant population differences for most pest and microbe loads, and often heritabilities above 15%, which although low, is still indicative of genetic determination (5). Moreover, Erysiphales load had a particularly high heritability of 47% (Table 1). We therefore next asked what could explain the observed variation in pest loads in our experiment. As pathogen abundances in the field are often determined by climatic

conditions, we expected plants originating from climates less favorable to aphids to perform worse in our glasshouse, i.e. to have higher pathogen loads. As expected, aphid counts were negatively correlated with temperature of origin (particularly minimum temperature), and positively with temperature variability (Mean Diurnal Range and Temperature Annual Range) (Fig 3A), suggesting that plants from colder and more thermally fluctuating climates, which are less favorable to aphids, were less well defended and performed worse in our glasshouse. We found similar but weaker patterns for Erysiphales load.

As we expected the observed climate-associated variation in pest loads to be at least partially explained by variation in chemical defenses, and since in *Arabidopsis thaliana* glucosinolates, the main defensive compounds, are known to be geographically structured in response to aphid distributions (1), we also tested for association of aphid and mildew loads with glucosinolates in our collection. In accordance with literature on *A. thaliana* (54), we observed a positive correlation of aphid loads with total glucosinolates as well as with the most abundant glucosinolate sinigrin (aliphatic glucosinolate), but a negative correlation with benzyl glucosinolates (Fig 3B). These findings suggest that glucosinolate composition, rather than total amount, is important for aphid defense, and that while benzyl glucosinolates might have a deterrent effect, sinigrin might on the contrary attract *M. persicae* or act as a stimulant, which would be in accordance with previous observations (55).

To detect genetic variants associated with pest and microbe loads, we then conducted a GWA study. For aphid (*M. persicae*) load, we detected only one non-significant peak in Scaffold 5, close to a pathogenesis-related coding gene (Fig 4A and D). For Erysiphales load, however, there were several significant associations neighboring genes directly involved in defense, mostly members of the *MLP* family, clustered in a large peak on the first arm of Scaffold 1 (Fig 4E). *MLP165*, the closest gene to the most significant variant in the peak, is indirectly involved in GS biosynthesis in *A. thaliana* (56), which might explain why baseline GS levels were associated with Erysiphales load (Fig 3B). Further GWA peaks for Erysiphales pointed towards other genes indirectly involved in the defense response through phytohormone signaling (eg. *CAR8, PBL7, GH3.1*) or preventing pathogen access through cell wall reinforcement or stomatal closure (*SRF3, PBL7*) (Fig 4D). Further experiments would be necessary to confirm the functionality of these genes.

An important general insight from our GWA analyses was the frequent ambiguity of reads that mapped to both pest and host plant genome. Such ambiguous reads generated false variants only present in samples with pest DNA, which resulted in highly significant false associations, and it was therefore important to remove these reads before variant calling. Another potential reason for sequence similarity between host and pathogens could be defense mechanisms such as RNA interference. If *T.*

arvense produces small or micro RNAs to silence pathogen genes, this would originate from genomic regions of high similarity between host and pathogen, and thus reflect a true association. However, a BLAST (57) of the region in which the suspicious associations occurred did not reveal any similarity to aphid or mildew genes, but instead to the highly conserved ribosomal RNA coding regions. While genetic variants are generally inherited from parents and thus reflect evolutionary processes, DNA methylation variants can be heritable but can also reflect plastic responses to environmental stresses like herbivores or pathogens. Our data do not allow to confidently distinguish between these two sources of DNA methylation variation, and thus should be interpreted with caution, especially with regard to the directionality of associations. A beneficial DNA methylation variant is expected to be associated with lower pathogen load when already present before pathogen arrival, but with higher pathogen load when plastically induced by pathogens during the experiment. For both *M. persicae* and Erysiphales, the majority of DMRs were hypomethylated in affected samples, which is in accordance with the loss of methylation observed in *A. thaliana* and *T. arvense* upon aphid feeding, and in diploid wheat upon powdery mildew infection (8,9,31), but we also detected hypermethylation at several loci. *M. persicae* load was associated with differential methylation at only few genes but several TEs, which is in accordance with the aphid or stress-induced TE reactivation observed in *A. thaliana* (9,14). Erysiphales load was associated with hypomethylated Copia TEs upstream of *MAPKKK20*, a gene involved in ABA-mediated signaling and stomatal closure. Since stomatal closure is a known defense mechanism to block pathogen access (21), it is tempting to conclude that hypomethylation of the *MAPKKK20* promoter might induce its overexpression and consequent stomatal closure, thereby preventing mildew access to the leaf blade. Overall, we found associations between pathogen load and TE methylation that could act both in *cis* (eg. Copia TE methylation in *MAPKKK20* promoter) and in *trans*, possibly through transposon reactivation (eg. LINE, Helitron and Ty3/Gypsi TEs isolated from genes). Although we cannot confidently distinguish inherited versus induced DNA methylation variants, to our knowledge this is the first coupled GWA - EWA analysis conducted on a large natural plant collection.

In summary, our study offers first insights into the defense mechanisms of *Thlaspi arvense*, including candidate genes and alleles which may be of interest for breeding efforts in this novel biofuel and cover crop. It also provides a proof of principle that exogenous reads from large sequencing efforts, usually discarded if not mapping to the target genome, can be leveraged to extract additional information about important biotic interactions of the target species, including its antagonists and microbiome components. We combined this approach with data from a common environment experiment to show that pest and microbiome load were geographically structured, as expected from locally adapted traits, and associated with both genetic and DNA methylation variants. In principle, our

approach can be applied to many other designs. For example, field-collected samples could be used to quantify geographic pathogen distributions. With the decreasing cost of sequencing and increasing large-scale and single-species sequencing projects e.g. (37–43), the number of data-sets suitable for such analyses is set to rapidly increase in the near future.

Materials and Methods

Plant growth and sequencing

The WGS data used in this study were already published in Galanti et al. (2022) (32). Please refer to this publication for details on data generation and methods. Briefly, we collected 207 *Thlaspi arvense* accessions from 36 European populations in July 2018, and we grew their offspring in a glasshouse at University of Tübingen (48°32'21.3"N, 9°02'04.2"E) between August and October 2019. The glasshouse was located in biodiverse surroundings, and insects and pests could enter when the windows opened for temperature regulation. A few weeks after germination, we noticed aphid and mildew infestations. After 46 d we sampled the 3rd or 4th true leaf of each plant and snap-froze it in liquid nitrogen. We extracted DNA using the DNeasy Plant Mini Kit (Qiagen, Hilden, DE), sonicated (Covaris) 300 ng of genomic DNA and used the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs) to prepare the libraries. Half way through the protocol we split the DNA into 1/3 for genomic libraries and 2/3 for bisulfite libraries. For the bisulfite conversion we used the EZ-96 DNA Methylation-Gold MagPrep (ZYMO) kit. We sequenced paired-end for 150 cycles using Illumina NovaSeq 6000 (Illumina, San Diego, CA) for genomic libraries and HiSeq X Ten (Illumina, San Diego, CA) for bisulfite libraries.

Reads mapping and classification

Upon demultiplexing the raw reads, we used cutadapt (58) for quality (minimum quality of 20) and adaptor trimming, excluding reads shorter than 35 bp. We used FastQC and MultiQC (59,60) to estimate the duplication rate, and calculated the total deduplicated reads, which we later used for correcting the number of exogenous reads. We then classified the reads based on their mapping behavior. First we aligned reads to the *T. arvense* reference genome (29) with BWA-MEM v0.7.17 (61), excluding multimapping reads (-c 1) and marking duplicates with MarkDuplicatesSpark (62,63). We then mapped all samples again (61) to the three putative exogenous genomes of pea aphid (*Acrophyson pisum*), the aphid symbiont *Buchnera aphidicola* and powdery mildew (*Blumeria graminis*), using available resources (www.ncbi.nlm.nih.gov/assembly/GCF_005508785.2, www.ncbi.nlm.nih.gov/assembly/GCA_001939165.1) (64). After this, we used a custom script to

collect all read IDs within a sample mapping to any of the three exogenous genomes, and removed any of these reads from the *T. arvense* alignment bam files. We thus removed all ambiguous reads before proceeding with variant calling. To compare coverage of specific regions with and without ambiguous reads, we used samtools bedcov (65). The numbers of reads classified by their mapping behaviour are reported in S1 Table.

Variant calling

For variant calling we used GATK4 v4.1.8.1 (62,63), following the best practices for germline short variant discovery (https://gatk.broadinstitute.org/hc/en-us/articles/360035535932-Germline-short-variant-discovery-SNPs-Indels-) with few adjustments for large datasets (32). Briefly, starting from the bam files generated after the removal of ambiguous reads, we i) ran HaplotypeCaller, ii) combined the resulting GVCF files with GenomicsDBImport and GenotypeGVCFs and iii) filtered out low quality variants with VariantFiltration (see Galanti et al. (2022) (32) for more details). Finally, we used vcftools v0.1.16 (66) to retain biallelic variants with MAF > 0.01 and a maximum of 10% missing genotype calls. We imputed these missing calls with BEAGLE 5.1 (67) to obtain a complete multisample vcf file.

Identification and classification of exogenous reads

To identify exogenous reads, we extracted all unmapped reads from the bam files created aligning WGS reads to the *T. arvense* genome (29). We selected reads with both mates unmapped (SAM flag 12) and excluded supplementary alignments (SAM flag 256 after running MarkDuplicatesSpark) with samtools (65). We then recovered these reads from the trimmed fastq files with seqtk subset (https://github.com/lh3/seqtk) to obtain fastq files of unmapped reads only. We used these as input for MG-RAST (44,45), a web-based tool for phylogenetic analysis of metagenomes.

We ran MG-RAST mostly with default parameters, without assembled reads, excluding dereplicated sequences, and dynamically trimming reads with a minimum Phred score of 15 in more than 5 consecutive bases. We set the "sequence screening" to *A. thaliana*, the closest relative of *T. arvense* available. We used two different approaches to extract read counts. First we classified all reads up to family level using the web-based Analysis tool from MG-RAST. We used RefSeq as query annotation database and filtered reads classified with low confidence using default settings: e-value 5, 60 %-identity, length 15 and min.abundance of 1 (S2 Table). Out of the hundreds of taxonomic groups identified by MG-RAST, we selected only a small subset for follow-up analyses, based on their biological relevance, our visual observations and/or abundance: Aphididae, Culicidae, Peronosporales,

Staphylococcaceae, Burkholderiaceae, Mycobacteriaceae and Pseudomonadaceae (Table 1). Additionally, we used a custom Python script to download individual "taxonomy" or "sequence_breakdown" results from MG-RAST API (68) and extracted the counts of the genus *Buchnera,* including bacterial symbionts of many aphid species, and of the order Erysiphales, to quantify the observed mildew infection (Table 1). All the code for extracting counts for all families or specific taxonomic groups are available on GitHub (https://github.com/junhee-jung/MG-RAST-read-counter).

In addition to the nine read groups selected from MG-RAST results, we also performed a highly confident mapping of exogenous reads to the *M. persicae* and *B. aphidicola* genome assemblies (50) (www.ncbi.nlm.nih.gov/assembly/GCA_001939165.1), to test whether mapping to a high quality assembly of the exact pathogen has a higher sensitivity than MG-RAST. We mapped with BWA-MEM v0.7.17 (61), using a seed length of 25 bp (69) and removing reads with MAPQ < 20 and duplicates with MarkDuplicatesSpark (62,63). We then counted all reads in the bam files.

Finally, we log transformed all read counts to approximate normality, and corrected for the total number of deduplicated reads by extracting residuals from the following linear model, log(Read_count + 1) ~ log(Deduplicated_reads), which allowed us to quantify non-*Thlaspi* loads, accounting for the sequencing depth of each sample.

Exogenous reads heritability and species identification

To exclude the possibility that aphid and mildew infestation patterns were carried-over from the field, through seed contamination or maternal effects, we used aphid and mildew presence/absence data collected in the field. We found no difference in aphid or mildew loads between samples with and without aphids or mildew on the original parental plant in the field (S1 Fig). Nevertheless, to exclude a possible bias, we excluded one outlier sample with particularly high aphid load and aphids observed in the field (S1 Fig) from the analyses.

Even though MG-RAST classifies reads based on all taxonomic ranks, the accuracy of species identification of course strongly depends on the sequences available in the query databases. MG-RAST assigned our aphid reads to *A. pisum*, but this did not fit with our visual observations and with the poor performance of this species on Brassicaceae (70). We therefore selected three plausible aphid species and test which of these had mostly likely attacked our experiment. In addition to *A. pisum*, we tested two other aphid species commonly attacking Brassicaceae: *Brevicoryne brassicae* and *Myzus persicae*. While not all three species have reference genomes available, all mitochondrial genomes are available

on NCBI (71) under accession numbers MN232006, NC_011594 and NC_056270. We downloaded these sequences, aligned them to each other (72), removed INDELs to retain only SNPs and combined them into a single pseudo-reference fasta file (S3 Table). We then mapped the exogenous reads from 40 randomly selected samples to this pseudo-reference, allowing for unique mappings only and counted the reads mapping to either of the three aphid species. We used the same approach for mildew except that we included only two possible species: *Blumeria graminis*, as suggested by MG-RAST, and *Erysiphe cruciferarum* which is known to attack Brassicaceae but was not in the MG-RAST query database and seemed plausible from visual inspection (Fig 2B). For the mildew pseudo-reference (S4 Table) we used the Internal Transcribed Spacer (ITS), which is publicly available for both species on NCBI (71) under accession numbers MT644878 and AF031283.

Quantification of glucosinolates

Using seed material collected from the sequenced plants, we conducted a follow-up experiment to estimate the glucosinolates (GS) content of all 207 lines in the absence of pathogens. Briefly, we sowed the seeds in petri dishes, stratified them at 4°C in the dark for two weeks and transplanted the germinated seedlings to individual 9 x 9 cm pots. We grew the plants in a growth chamber with a 14/10 h light/dark cycle at 21/17 °C and a relative humidity of ~45%. Two weeks after germination the plants were vernalized at 4°C for two more weeks in order to minimize phenological and developmental differences between winter and summer annuals. Ten days after vernalization, we collected the 3rd or 4th true leaf and snap-froze it in liquid nitrogen. After freeze drying, we weighed all samples and extracted the material threefold in 80% methanol, adding *p*-hydroxybenzl glucosinolate (Phytoplan, Heidelberg, Germany) as internal standard. After centrifugation, we applied the supernatants onto ion-exchange columns with diethylaminoethyl (DEAE) Sephadex A25 (Sigma Aldrich, St.Louis, MO, USA) in 0.5 M acetic acid buffer, pH 5. We added purified sulfatase, converting glucosinolates to desulfo glucosinolates. After one day, we eluted desulfo glucosinolates in water and analyzed them on a HPLC coupled to a DAD detector (HPLC-1200 Series, Agilent Trechnologies, Inc., Santa Clara, CA, USA) equipped with a Supelcosil LC 18 column (3 µm, 150×3 mm, Supelco, Bellefonte, PA, USA). We analysed the samples with a gradient from water to methanol starting at 5% methanol for 6 min and then increased from 5 to 95% within 13 min with a hold at 95% for 2 min, followed by a column equilibration cycle. We identified different glucosinolates based on their retention times and UV spectra in comparison to respective standards and an in-house database. We integrated peaks at 229 nm and calculated respective glucosinolate concentrations in relation to the internal standard and sample dry mass, using response factors as reported by Agerbirk et al. (2015) (73).

Drivers of exogenous reads variation

To test for associations between glucosinolate variation, as well as climate of origin, and the observed pest loads, we extracted average bioclimatic variables for the 25 years predating our experiment for our 36 study populations from the Copernicus website (74), as described in Galanti et al. 2022 (32). We then used the R package "lme4qtl" (75) to run mixed models that included either bioclimatic variables or glucosinolate contents as explanatory variables, and the exogenous read counts as dependent variables, while correcting for population structure with the same IBS matrix as in GWA and EWA analyses (see below).

GWA analysis

We conducted GWA with mixed models that corrected for population structure with a genetic Isolation By State (IBS) matrix as a random factor, as implemented in GEMMA (76). To obtain the IBS matrix we used PLINK v 1.90b6.12 (77). Starting from the imputed multisample vcf file obtained from variant calling, we pruned variants with LD > 0.8 in 50 variants windows, sliding by five. To produce the genetic variants used for GWAS, we also started from the imputed multisample vcf file from variant calling and filtered out variants with MAF < 0.04. As phenotypes we used the number of exogenous reads corrected for the total number of deduplicated reads, as described above.

To validate our results and test for overlap with existing gene functional annotations, we performed an enrichment analysis of variants neighboring *a priori* candidate genes as described in Atwell et al. (2010) (52). Briefly, we attributed *a priori* candidate status to all variants located within 20 kb from orthologs (78) of *A. thaliana* genes annotated with the GO term "defense response" (GO:0006952), including nine genes similar to *AtBSMT1* (S8 Table). We then calculated the enrichment of these variants compared to the background frequency and an upper bound for the FDR for increasing -log(p) thresholds (32,52). The code for these analyses is available on https://github.com/Dario-Galanti/multipheno_GWAS/tree/main/gemmaGWAS.

Methylation and DMR calling

For the methylome analyses we used the EpiDiverse toolkit (79), specifically designed for large WGBS datasets. We used the WGBS pipeline (https://github.com/EpiDiverse/wgbs) for read mapping and methylation calling, retained only uniquely-mapping reads longer than 30 bp, and obtained individual-sample bedGraph files for each sequence context. We then called DMRs using the DMR pipeline (79),

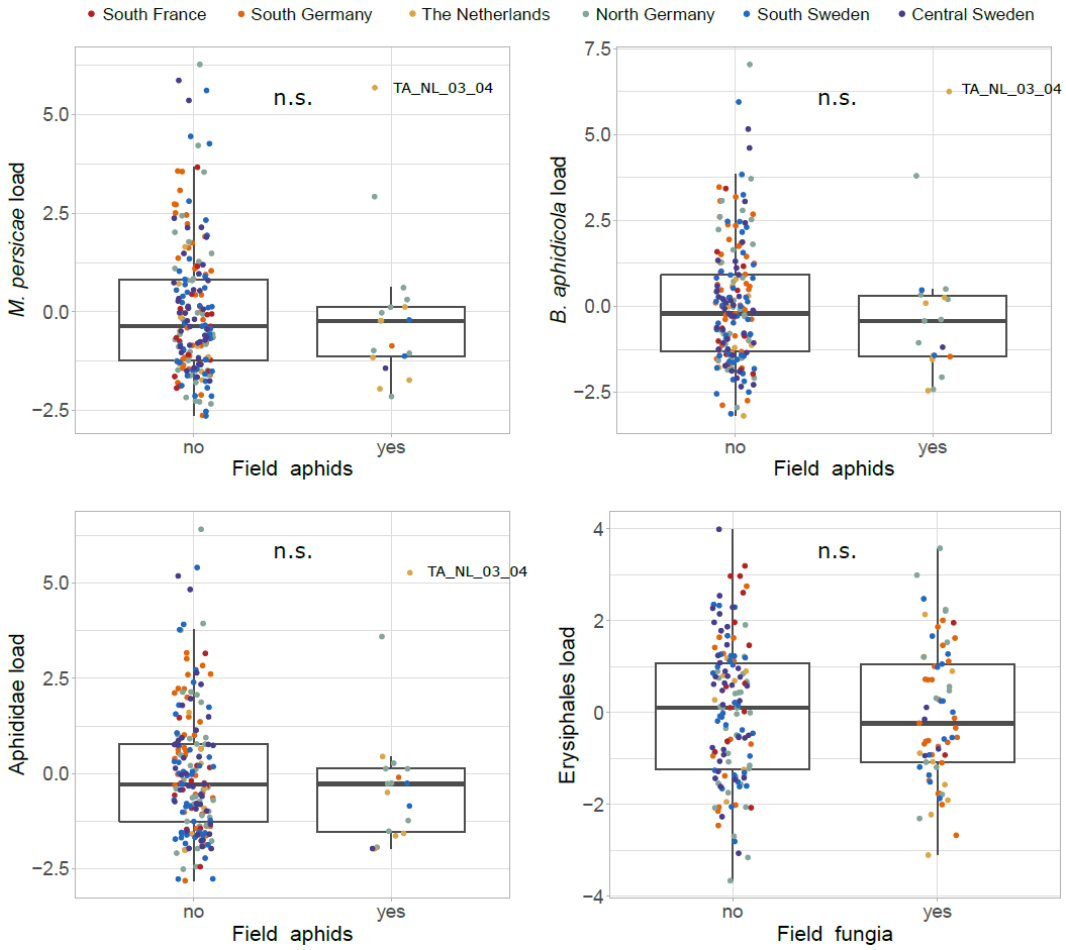
with a minimum coverage of 4x. We compared the 20 samples with the most and the least *M. periscae* and Eriysiphales loads, resulting in two sets of DMRs for each sequence context. Since this was only the first step of our methylation analysis, meant to identify potential regions of interest, we retained all DMRs with a FDR < 20%. To understand the genomic preferences of DMRs, we intersected them with genomic features and calculated their densities in each by dividing their number by the total Mb covered by each genomic feature in the genome.

EWA analysis

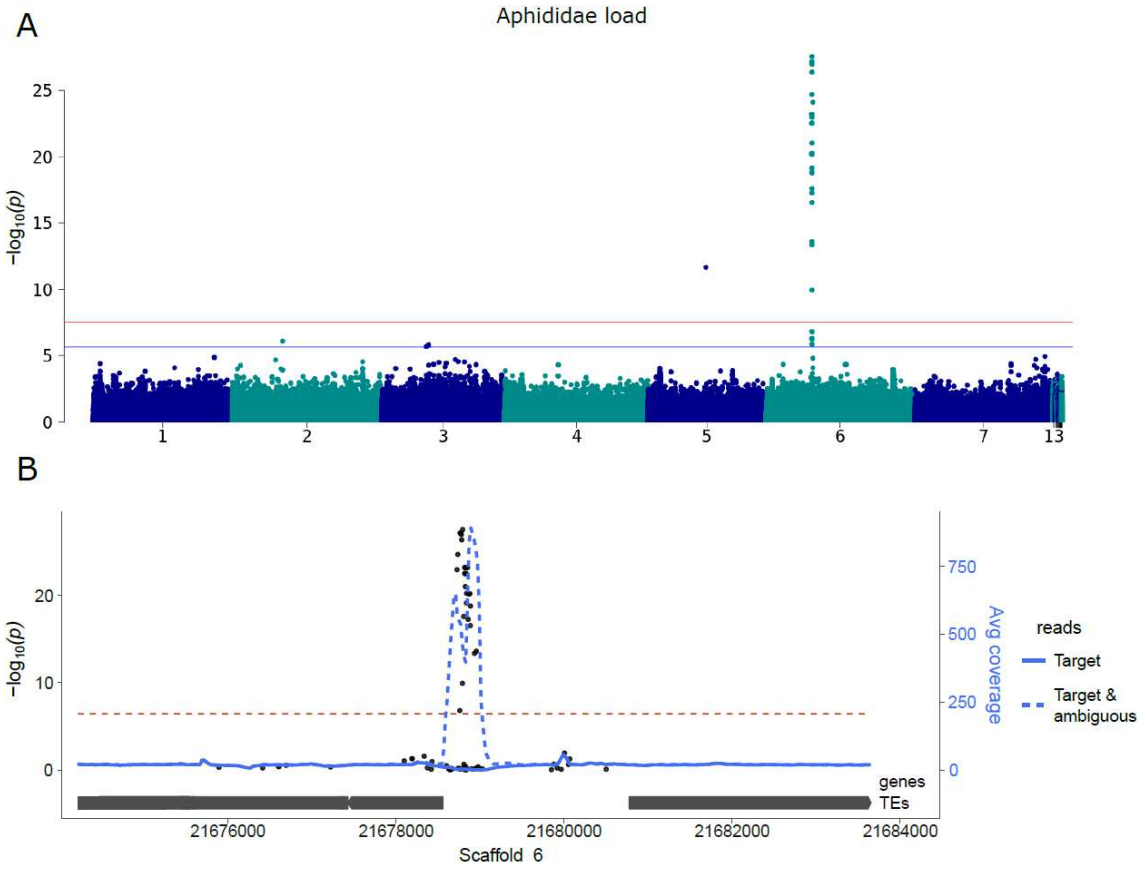
Following the DMR calling, we investigated methylation-phenotype relationships in more detail, using EWA. We ran EWA similarly to GWA, enabling the "-notsnp" option available in GEMMA (76), and correcting for population structure with the same IBS matrix. To exclude possible biases, we excluded all samples with a bisulfite non-conversion rate >1 (32). To generate the methylation input files we first used custom scripts (https://github.com/Dario-Galanti/WGBS_downstream/tree/main/WGBS_simpleworkflow) (32) to unite individual-sample bedGraph files into unionbed files and retain positions with coverage > 3 in at least 95% of the samples and a methylation difference of at least 5% in at least two samples. We then intersected the unionbed files with the DMRs of the corresponding sequence context using bedtools (80) and converted unionbed to BIMBAM format as input for GEMMA.

We ran EWA for individual positions within the DMRs and calculated Bonferroni thresholds based on the number of DMRs, assuming that cytosines within the same DMR are mostly autocorrelated. To observe in which genomic features associations with lower *P*-values were located, we performed enrichment analyses similar to the ones performed for defense *a priori* candidate genes in GWA (52), but based on whole genomic features. Starting from all cytosines used for EWAS, we calculated the background frequency as the fraction of all cytosines located in each genomic feature and then calculated the observed frequency in the same way for -log(p) 0.5 increments, with enrichment as the ratio of observed and expected frequencies. All code used for EWA and the enrichment analysis in genomic features is available on https://github.com/Dario-Galanti/EWAS/tree/main/gemmaEWAS.

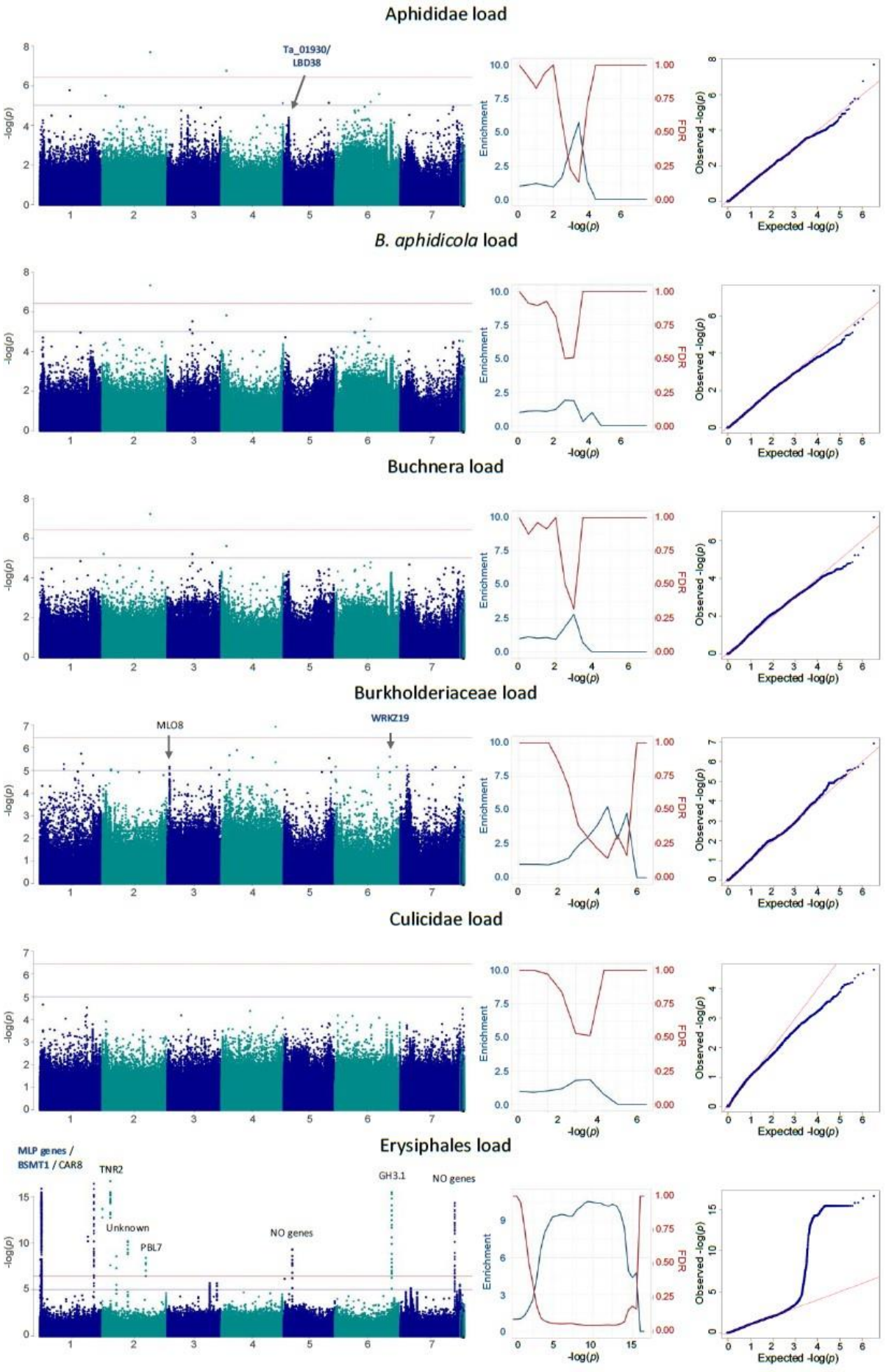
Supporting Information

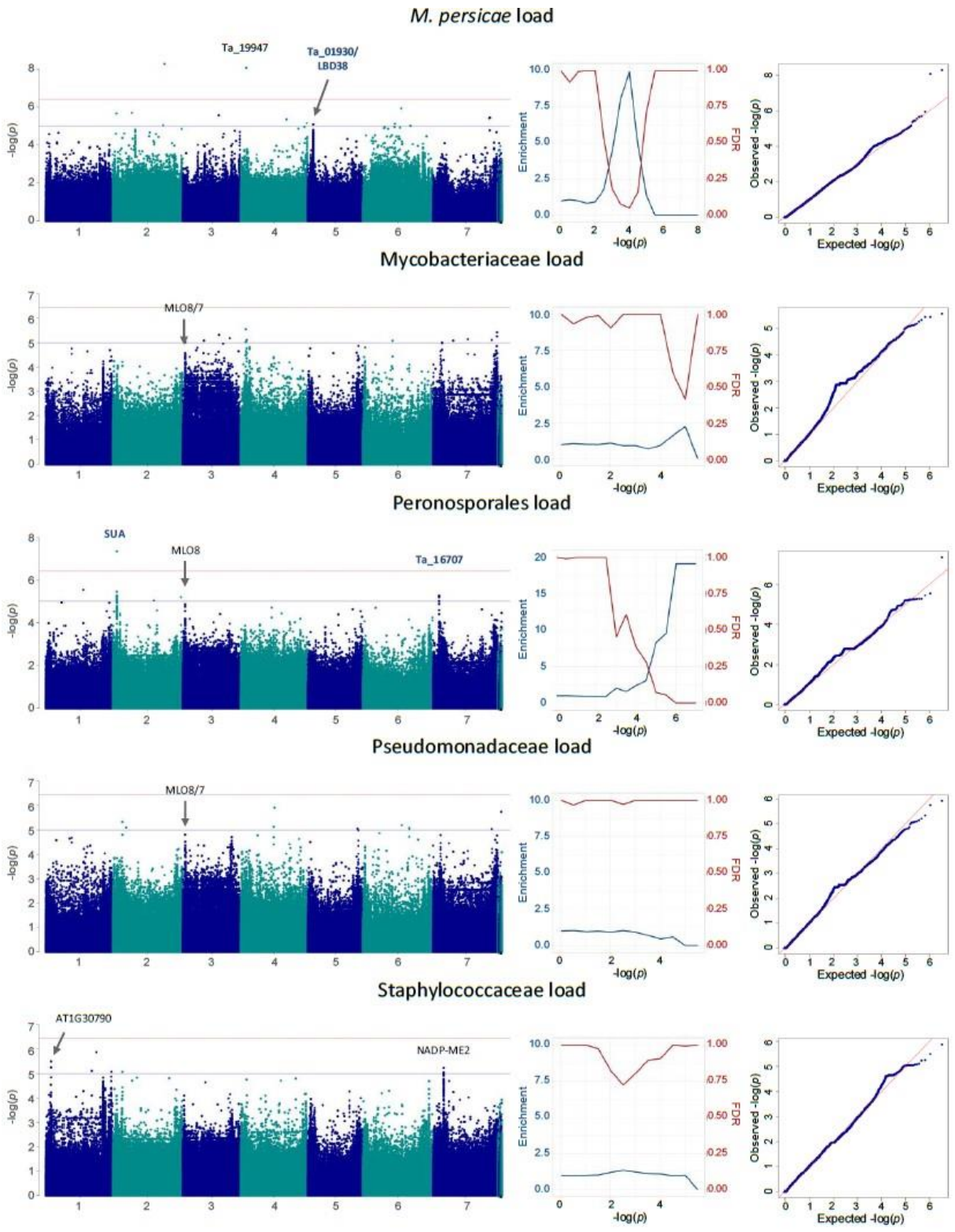


S1 Fig. Pest loads in samples with or without pests in the field. Offspring from plants with or without pathogens in the field did not differ (n.s.) in pathogen loads in the glasshouse, suggesting the pathogens were not carried with the seeds. A single outlier (TA_NL_03_04) with aphids in the field also showed high levels of aphid load in the glasshouse and was excluded from the analysis.

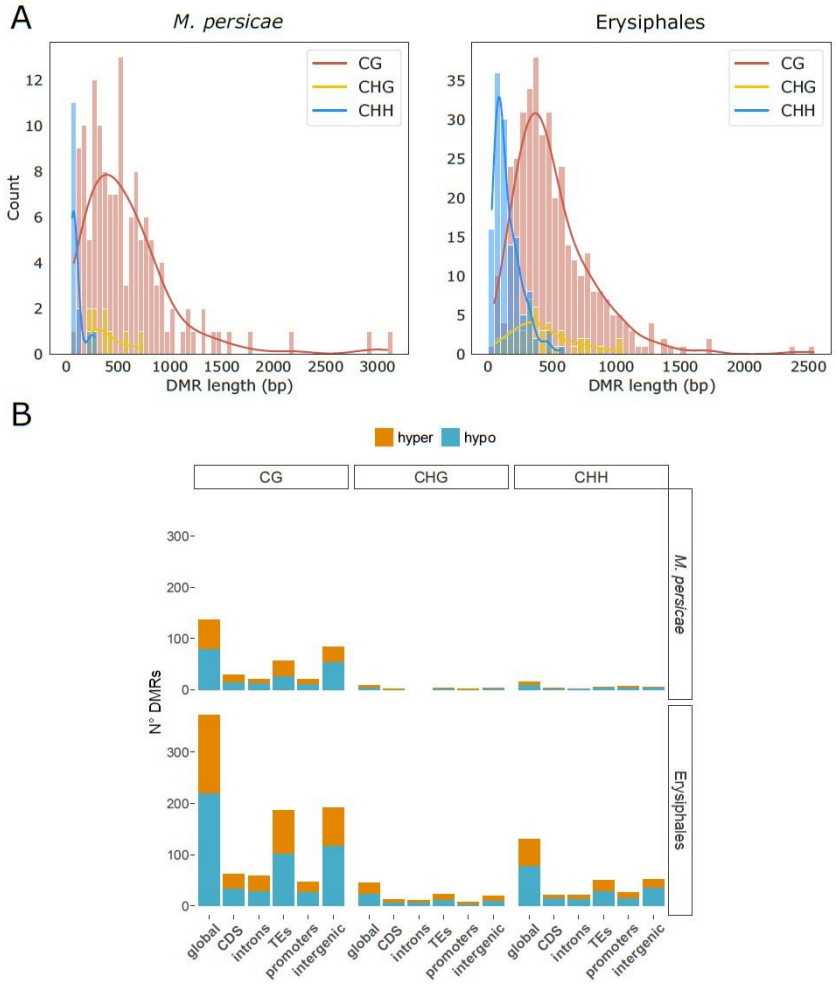


S2 Fig. Example of a GWA peak caused by ambiguous reads. (A) Manhattan plot of GWA with aphid load, without removing ambiguous reads prior to variant calling. The genome-wide significance (horizontal red line) was calculated based on unlinked variants (53), the suggestive line (blue) corresponds to -log(p) =5. (B) Zoomed-in peak on Scaffold 6 with coverage before and after removal of ambiguous reads.

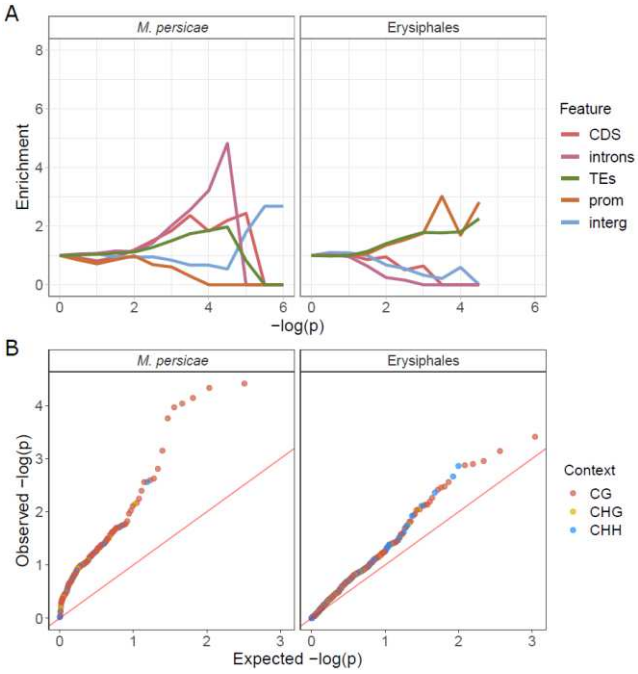

Aphididae load
Ta_01930/
LBD38
B. aphidicola load
Buchnera load
Burkholderiaceae load
MLO8
WRKZ19
Culicidae load
Erysiphales load
MLP genes /
BSMT1 / CAR8
TNR2
Unknown
PBL7
NO genes
GH3.1
NO genes
-log(p)
Enrichment
FDR
Observed -log(p)
Expected -log(p)



S3 Fig. GWA results for all exogenous reads. Manhattan plots, enrichment of *a priori* candidate variants and QQplots for all exogenous reads. The genome-wide significance (horizontal red lines) was calculated based on unlinked variants (53), the blue line corresponds to -log(p) =5. The top variants are labeled with the neighboring genes potentially affecting exogenous reads load, the *a priori* candidates included in the enrichment analyses and involved in defense response (GO:0006952) are in bold blue.



S4 Fig. Lengths and genomic locations of Differentially Methylated Regions (DMRs). (A) DMRs length distributions in different sequence contexts. (B) Numbers of hypermethylated and hypomethylated DMRs in each genomic feature.



S5 Fig. EWA enrichment in different genomic features and *P*-value distributions. (A) Enrichment of associations in different genomic features for increasing -log(p) thresholds, compared to expectations from all cytosines. (B) *P*-value distributions of EWA. (C) Complement to Fig 4D showing CHG and CHH methylation upstream and above the gene *MPKKK20*.

Supplementary tables are too large to be reported here, so please refer to the preprint of the manuscript: https://www.biorxiv.org/content/10.1101/2023.10.17.562203v2.supplementary-material

S1 Table. Classification of reads. Total reads, duplication rates, deduplicated raw reads, target, ambiguous and exogenous reads and MG-RAST reads passing QC.

S2 Table. MG-RAST classification of exogenous reads. Results of MG-RAST classification of exogenous reads at the family level, using RefSeq as query database.

S3 Table. Competitive mapping pseudo-reference for aphid identification. Fasta file combining mitochondrial sequences of *B. Brassicaceae, A. pisum* and *M. persicae*, with structural variants removed.

S4 Table. Competitive mapping pseudo-reference for mildew identification. Fasta file combining ITS sequences of *Blumeria graminis* and *Erysiphe cruciferarum*, with structural variants removed.

S5 Table. Results of competitive mapping for pest identification. Number of reads mapping uniquely to the pseudo-reference genomes of different aphid (mitochondrial DNA for either *B. Brassicaceae, A. pisum* or *M. persicae*) or mildew species (ITS sequence for either *B. graminis* or *E. cruciferaum*).

S6 Table. Exogenous reads used for downstream analyses. Classes of exogenous reads used in the analyses, including the nine groups from MG-RAST and two from mapping to the *M. persicae* and *B. aphidicola* reference genomes.

S7 Table. Quantification of glucosinolates. Glucosinolate concentrations in leaves of *T. arvense*, obtained from offspring of the sequenced plants, not affected by any herbivores nor pathogens.

S8 Table. *A priori* "defense response" candidate genes used for the GWA enrichment analysis. List of *T. arvense* candidate genes used for the GWA enrichment analysis: orthologs of *A. thaliana* genes annotated with the GO term "defense response".

S9 Table. Differentially Methylated Regions based on *M. persicae* load. List of DMRs called between the 20 samples with the highest and the lowest *M. persicae* load.

S10 Table. Differentially Methylated Regions based on Erysiphales load. List of DMRs called between the 20 samples with the highest and the lowest Erysiphales load.

Code and data availability statement

The seed material from the sequenced lines is available at the Nottingham Arabidopsis Stock Centre (NASC) under stock numbers N950001 to 950204. Genomic and bisulfite sequencing raw data are available on the ENA Sequence Read Archive (www.ebi.ac.uk/ena/) under study accession number PRJEB50950. The reference genome and annotations were previously published by Nunn et al. (2022) (29). GWA and EWA results in a format compatible with the Integrative Genomics Viewer (https://www.igv.org/) are available on Zenodo (https://zenodo.org/records/10011535).

All code used in this study is available and documented on GitHub. The scripts for variant calling, filtering and imputation are on https://github.com/Dario-Galanti/BinAC_varcalling, and the scripts for the classification of sequencing reads and MG-RAST analysis are in https://github.com/Dario-Galanti/Exoreads_treasure and https://github.com/junhee-jung/MG-RAST-read-counter respectively. The pipelines for methylation and DMR calling from WGBS data can be found on the EpiDiverse GitHub (https://github.com/EpiDiverse). The workflow for downstream analysis of methylation data is on https://github.com/Dario-Galanti/WGBS_downstream/tree/main/WGBS_simpleworkflow. Finally, the scripts for running GWA and EWA analysis are on https://github.com/Dario-Galanti/multipheno_GWAS/tree/main/gemmaGWAS and https://github.com/Dario-Galanti/EWAS/tree/main/gemmaEWAS respectively.

Funding

This work was supported by the EU Horizon 2020 program under Marie Skłodowska-Curie grant agreement 764965 (Innovative Training Network EpiDiverse; https://epidiverse.eu; PhD fellowship of DG), as well as by the Deutsche Forschungsgemeinschaft (DFG) as part of the priority programme 2125 Deconstruction and Reconstruction of the Plant Microbiota, "DECRyPT" (grant 401829393 to OB; PhD fellowship of JHJ). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments

We thank the EpiDiverse network for its amazing support and discussions, in particular Adrián Contreras-Garrido, Bárbara Díez Rodríguez, Iris Sammarco, Adam Nunn, Daniela Ramos and Anupoma Troyee for their close collaboration. We also thank Frank Reis for his feedback and expert tips for the project, and Cecilia Heyworth for proofreading the manuscript. We thank Peter Stadler at the University of Leipzig and David Langenberger from ecSeq, which helped with computing and hosted

the EpiDiverse servers. The BinAC cluster is managed by the High Performance and Cloud Computing Group at the Zentrum für Datenverarbeitung of the University of Tübingen, and supported by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant no INST 37/935-1 FUGG.

Author Contributions

Dario Galanti: Conceptualization, Methodology, Formal analysis, Investigation, Data Curation, Writing – Original Draft Preparation, Visualization.

Jun Hee Jung: Conceptualization, Methodology, Data Curation, Writing – Review & Editing.

Caroline Müller: Investigation, Writing – Review & Editing.

Oliver Bossdorf: Conceptualization, Methodology, Writing – Original Draft Preparation, Supervision, Funding Acquisition.

References

1. Züst T, Heichinger C, Grossniklaus U, Harrington R, Kliebenstein DJ, Turnbull LA. Natural Enemies Drive Geographic Variation in Plant Defenses. Science. 2012 Oct 5; 338(6103):116–9. https://www.science.org/doi/full/10.1126/science.1226397

2. Kerwin R, Feusier J, Corwin J, Rubin M, Lin C, Muok A, et al. Natural genetic variation in Arabidopsis thaliana defense metabolism genes modulates field fitness. Kant MR, editor. eLife. 2015 Apr 13; 4:e05604. https://doi.org/10.7554/eLife.05604

3. Chan EKF, Rowe HC, Kliebenstein DJ. Understanding the Evolution of Defense Metabolites in Arabidopsis thaliana Using Genome-wide Association Mapping. Genetics. 2010 Jul 1; 185(3):991–1007. https://doi.org/10.1534/genetics.109.108522

4. Corwin JA, Copeland D, Feusier J, Subedy A, Eshbaugh R, Palmer C, et al. The Quantitative Basis of the Arabidopsis Innate Immune System to Endemic Pathogens Depends on Pathogen Genetics. PLOS Genet. 2016 Feb 11; 12(2):e1005789. https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1005789

5. Thoen MPM, Olivas NHD, Kloth KJ, Coolen S, Huang PP, Aarts MGM, et al. Genetic architecture of plant stress resistance: multi-trait genome-wide association mapping. New Phytol. 2017; 213(3):1346–62. https://nph.onlinelibrary.wiley.com/doi/abs/10.1111/nph.14220

6. Hanson AA, Lorenz AJ, Hesler LS, Bhusal SJ, Bansal R, Michel AP, et al. Genome-Wide Association Mapping of Host-Plant Resistance to Soybean Aphid. Plant Genome. 2018; 11(3):180011. https://acsess.onlinelibrary.wiley.com/doi/abs/10.3835/plantgenome2018.02.0011

7. Jaouannet M, Morris JA, Hedley PE, Bos JIB. Characterization of Arabidopsis Transcriptional Responses to Different Aphid Species Reveals Genes that Contribute to Host Susceptibility and Non-host Resistance. PLOS Pathog. 2015 mag; 11(5):e1004918. https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1004918

8. Geng S, Kong X, Song G, Jia M, Guan J, Wang F, et al. DNA methylation dynamics during the interaction of wheat progenitor Aegilops tauschii with the obligate biotrophic fungus Blumeria graminis f. sp. tritici. New Phytol. 2019; 221(2):1023–35. https://onlinelibrary.wiley.com/doi/abs/10.1111/nph.15432

9. Annacondia ML, Markovic D, Reig-Valiente JL, Scaltsoyiannes V, Pieterse CMJ, Ninkovic V, et al. Aphid feeding induces the relaxation of epigenetic control and the associated regulation of the defense response in Arabidopsis. New Phytol. 2021; 230(3):1185–200. https://nph.onlinelibrary.wiley.com/doi/abs/10.1111/nph.17226

10. Kinoshita T, Seki M. Epigenetic Memory for Stress Response and Adaptation in Plants. Plant Cell Physiol. 2014 Nov 1; 55(11):1859–63. https://academic.oup.com/pcp/article/55/11/1859/2756009

11. Espinas NA, Saze H, Saijo Y. Epigenetic Control of Defense Signaling and Priming in Plants. Front Plant Sci. 2016 Aug 11; 7:1201. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4980392/

12. Lämke J, Bäurle I. Epigenetic and chromatin-based mechanisms in environmental stress adaptation and stress memory in plants. Genome Biol. 2017 Jun 27; 18(1):124.https://doi.org/10.1186/s13059-017-1263-6

13. He Y, Li Z. Epigenetic Environmental Memories in Plants: Establishment, Maintenance, and Reprogramming. Trends Genet. 2018 Aug 22; http://www.sciencedirect.com/science/article/pii/S0168952518301276

14. Roquis D, Robertson M, Yu L, Thieme M, Julkowska M, Bucher E. Genomic impact of stress-induced transposable element mobility in Arabidopsis. Nucleic Acids Res. 2021 Oct 11; 49(18):10431–47. https://doi.org/10.1093/nar/gkab828

15. Law JA, Jacobsen SE. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet. 2010 Mar; 11(3):204–20. https://www.nature.com/articles/nrg2719

16. Zhang H, Lang Z, Zhu JK. Dynamics and function of DNA methylation in plants. Nat Rev Mol Cell Biol. 2018 May 21; 1. https://www.nature.com/articles/s41580-018-0016-z

17. Liu J, He Z. Small DNA Methylation, Big Player in Plant Abiotic Stress Responses and Memory. Front Plant Sci. 2020; 11. https://www.frontiersin.org/article/10.3389/fpls.2020.595603

18. Wojtaszek P. Oxidative burst: an early plant response to pathogen infection. Biochem J. 1997 Mar 15; 322(Pt 3):681–92. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1218243/

19. War AR, Paulraj MG, Ahmad T, Buhroo AA, Hussain B, Ignacimuthu S, et al. Mechanisms of plant defense against insect herbivores. Plant Signal Behav. 2012 Oct 1; 7(10):1306–20. https://doi.org/10.4161/psb.21663

20. Kant MR, Jonckheere W, Knegt B, Lemos F, Liu J, Schimmel BCJ, et al. Mechanisms and ecological consequences of plant defence induction and suppression in herbivore communities. Ann Bot. 2015 Jun 1; 115(7):1015–51. https://doi.org/10.1093/aob/mcv054

21. Melotto M, Zhang L, Oblessuc PR, He SY. Stomatal Defense a Decade Later. Plant Physiol . 2017 Jun 1; 174(2):561–71. https://doi.org/10.1104/pp.16.01853

22. Muhammad T, Zhang F, Zhang Y, Liang Y. RNA Interference: A Natural Immune System of Plants to Counteract Biotic Stressors. Cells. 2019 Jan; 8(1):38. https://www.mdpi.com/2073-4409/8/1/38

23. Kutyniok M, Müller C. Crosstalk between above- and belowground herbivores is mediated by minute metabolic responses of the host Arabidopsis thaliana. J Exp Bot. 2012 Oct 1; 63(17):6199–210. https://doi.org/10.1093/jxb/ers274

24. Nalam V, Louis J, Patel M, Shah J. Arabidopsis-Green Peach Aphid Interaction: Rearing the Insect, No-choice and Fecundity Assays, and Electrical Penetration Graph Technique to Study Insect Feeding Behavior. Bio-Protoc. 2018 Aug 5; 8(15):e2950.

25. Sangiovanni M, Granata I, Thind AS, Guarracino MR. From trash to treasure: detecting unexpected contamination in unmapped NGS data. BMC Bioinformatics. 2019 Apr 18; 20(4):168. https://doi.org/10.1186/s12859-019-2684-x

26. Roman-Reyna V, Pinili D, Borja FN, Quibod IL, Groen SC, Alexandrov N, et al. Characterization of the Leaf Microbiome from Whole-Genome Sequencing Data of the 3000 Rice Genomes Project. Rice. 2020 Oct 9; 13(1):72. https://doi.org/10.1186/s12284-020-00432-1

27. Gathercole LAP, Nocchi G, Brown N, Coker TLR, Plumb WJ, Stocks JJ, et al. Evidence for the Widespread Occurrence of Bacteria Implicated in Acute Oak Decline from Incidental Genetic Sampling. Forests. 2021 Dec; 12(12):1683. https://www.mdpi.com/1999-4907/12/12/1683

28. Geng Y, Guan Y, Qiong L, Lu S, An M, Crabbe MJC, et al. Genomic analysis of field pennycress (Thlaspi arvense) provides insights into mechanisms of adaptation to high elevation. BMC Biol. 2021 Jul 22; 19(1):143. https://doi.org/10.1186/s12915-021-01079-0

29. Nunn A, Rodríguez-Arévalo I, Tandukar Z, Frels K, Contreras-Garrido A, Carbonell-Bejerano P, 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://onlinelibrary.wiley.com/doi/abs/10.1111/pbi.13775

30. 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. 2022; 12. https://www.frontiersin.org/articles/10.3389/fpls.2021.772655

31. Troyee AN, Medrano M, Müller C, Alonso C. Variation in DNA methylation and response to short-term herbivory in Thlaspi arvense. Flora. 2022 Aug 1; 293:152106. https://www.sciencedirect.com/science/article/pii/S0367253022001037

32. 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 ott; 18(10):e1010452. https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1010452

33. 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 Apr 1; 22(2):121–31. https://academic.oup.com/dnaresearch/article/22/2/121/332636

34. Frels K, Chopra R, Dorn KM, Wyse DL, Marks MD, Anderson JA. Genetic Diversity of Field Pennycress (Thlaspi arvense) Reveals Untapped Variability and Paths Toward Selection for Domestication. Agronomy. 2019 Jun; 9(6):302. https://www.mdpi.com/2073-4395/9/6/302

35. Chopra R, Johnson EB, Emenecker R, Cahoon EB, Lyons J, Kliebenstein DJ, et al. Progress toward the identification and stacking of crucial domestication traits in pennycress. Plant Biology; 2019 Apr. http://biorxiv.org/lookup/doi/10.1101/609990

36. Zhao R, Yang X, Li M, Peng X, Wei M, Zhang X, et al. Biodiesel preparation from Thlaspi arvense L. seed oil utilizing a novel ionic liquid core-shell magnetic catalyst. Ind Crops Prod. 2021 Apr 1; 162:113316. https://www.sciencedirect.com/science/article/pii/S0926669021000807

37. Kajiya-Kanegae H, Nagasaki H, Kaga A, Hirano K, Ogiso-Tanaka E, Matsuoka M, et al. Whole-genome sequence diversity and association analysis of 198 soybean accessions in mini-core collections. DNA Res. 2021 Feb 1; 28(1):dsaa032. https://doi.org/10.1093/dnares/dsaa032

38. Colgan TJ, Arce AN, Gill RJ, Ramos Rodrigues A, Kanteh A, Duncan EJ, et al. Genomic Signatures of Recent Adaptation in a Wild Bumblebee. Mol Biol Evol. 2022 Feb 1; 39(2):msab366. https://doi.org/10.1093/molbev/msab366

39. Habyarimana E, Gorthy S, Baloch FS, Ercisli S, Chung G. Whole-genome resequencing of Sorghum bicolor and S. bicolor × S. halepense lines provides new insights for improving plant agroecological characteristics. Sci Rep. 2022 Apr 1; 12(1):5556. https://www.nature.com/articles/s41598-022-09433-0

40. Mekbib Y, Tesfaye K, Dong X, Saina JK, Hu GW, Wang QF. Whole-genome resequencing of Coffea arabica L. (Rubiaceae) genotypes identify SNP and unravels distinct groups showing a strong geographical pattern. BMC Plant Biol. 2022 Feb 14; 22(1):69. https://doi.org/10.1186/s12870-022-03449-4

41. Metheringham CL, Plumb WJ, Stocks JJ, Kelly LJ, Gorriz MN, Moat J, et al. Rapid polygenic adaptation in a wild population of ash trees under a novel fungal epidemic. bioRxiv; 2022. p. 2022.08.01.502033. https://www.biorxiv.org/content/10.1101/2022.08.01.502033v3

42. Nocchi G, Brown N, Coker TLR, Plumb WJ, Stocks JJ, Denman S, et al. Genomic structure and diversity of oak populations in British parklands. PLANTS PEOPLE PLANET. 2022; 4(2):167–81. https://onlinelibrary.wiley.com/doi/abs/10.1002/ppp3.10229

43. Friis G, Smith EG, Lovelock CE, Ortega A, Marshell A, Duarte CM, et al. Rapid lineage diversification of gray mangroves (Avicennia marina) driven by isolation in cryptic glacial refugia and extreme environmental conditions in the Arabian Peninsula. Preprints; 2022 Nov. https://www.authorea.com/users/439811/articles/571762-rapid-lineage-diversification-of-gray-mangroves-avicennia-marina-driven-by-isolation-in-cryptic-glacial-refugia-and-extreme-environmental-conditions-in-the-arabian-peninsula?commit=bde361987d90a63b509ba7cd490054a9ee20bb70

44. Meyer F, Paarmann D, D'Souza M, Olson R, Glass E, Kubal M, et al. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics. 2008 Sep 19; 9(1):386. https://doi.org/10.1186/1471-2105-9-386

45. Keegan KP, Glass EM, Meyer F. MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function. Methods Mol Biol Clifton NJ. 2016;1399:207–33.

46. CABI. Myzus persicae (green peach aphid). CABI Compend . 2021 Dec 18 [cited 2023 Apr 7];CABI Compendium:35642. https://www.cabidigitallibrary.org/doi/10.1079/cabicompendium.35642

47. Gabryś B, Pawluk M. Acceptability of different species of Brassicaceae as hosts for the cabbage aphid. Entomol Exp Appl. 1999; 91(1):105–9. https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1570-7458.1999.00471.x

48. Warwick SI, Francis A, Susko DJ. The biology of Canadian weeds. 9. Thlaspi arvense L. (updated). Can J Plant Sci. 2002 Oct 1; 82(4):803–23. http://www.nrcresearchpress.com/doi/abs/10.4141/P01-159

49. Feuerborn TR, Palkopoulou E, van der Valk T, von Seth J, Munters AR, Pečnerová P, et al. Competitive mapping allows for the identification and exclusion of human DNA contamination in ancient faunal genomic datasets. BMC Genomics. 2020 Nov 30; 21(1):844. https://doi.org/10.1186/s12864-020-07229-y

50. Singh KS, Cordeiro EMG, Troczka BJ, Pym A, Mackisack J, Mathers TC, et al. Global patterns in genomic diversity underpinning the evolution of insecticide resistance in the aphid crop pest Myzus persicae. Commun Biol. 2021 Jul 7; 4(1):1–16. https://www.nature.com/articles/s42003-021-02373-x

51. Gao J, Fang C, Zhao B. The latitudinal herbivory hypothesis revisited: To be part is to be whole. Ecol Evol. 2019; 9(7):3681–8. https://onlinelibrary.wiley.com/doi/abs/10.1002/ece3.2759

52. Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, et al. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature. 2010 Jun; 465(7298):627–31. https://www.nature.com/articles/nature08800

53. Sobota RS, Shriner D, Kodaman N, Goodloe R, Zheng W, Gao YT, et al. Addressing population-specific multiple testing burdens in genetic association studies. Ann Hum Genet. 2015 Mar; 79(2):136–47. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4334751/

54. Kim JH, Jander G. Myzus persicae (green peach aphid) feeding on Arabidopsis induces the formation of a deterrent indole glucosinolate. Plant J. 2007; 49(6):1008–19. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-313X.2006.03019.x

55. Klingauf F, Şengonca Ç, Bennewitz H. Einfluß von Sinigrin auf die Nahrungsaufnahme polyphager und oligophager Blattlausarten (Aphididae) (Effect of Sinigrin on Sucrose Uptake by Some Polyphagous and Oligophagous Aphids (Aphididae)). Oecologia. 1972; 9(1):53–7. https://www.jstor.org/stable/4214736

56. The Arabidopsis Information Resource (TAIR). 2000. www.arabidopsis.org/servlets/TairObject?id=137911&type=locus

57. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009 Dec 15; 10(1):421. https://doi.org/10.1186/1471-2105-10-421

58. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal . 2011 May 2 [cited 2021 Dec 2];17(1):10–2. http://journal.embnet.org/index.php/embnetjournal/article/view/200

59. Andrews S. FASTQC. A quality control tool for high throughput sequence data | BibSonomy. 2010 [cited 2023 Feb 13]. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

60. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016 Oct 1 [cited 2023 Feb 13];32(19):3047–8. https://doi.org/10.1093/bioinformatics/btw354

61. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009 Jul 15 [cited 2021 Jun 11]; 25(14):1754–60. https://doi.org/10.1093/bioinformatics/btp324

62. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Angel G del, Levy-Moonshine A, et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Curr Protoc Bioinforma. 2013 [cited 2021 Jun 11]; 43(1):11.10.1-11.10.33. https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/0471250953.bi1110s43

63. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Auwera GAV der, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. 2018 Jul 24 [cited 2021 Jun 11];201178. https://www.biorxiv.org/content/10.1101/201178v3

64. Frantzeskakis L, Kracher B, Kusch S, Yoshikawa-Maekawa M, Bauer S, Pedersen C, et al. Signatures of host specialization and a recent transposable element burst in the dynamic one-speed genome of the fungal barley powdery mildew pathogen. BMC Genomics. 2018 May 22; 19(1):381. https://doi.org/10.1186/s12864-018-4750-6

65. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021 Feb 1 [cited 2022 Aug 15]; 10(2):giab008. https://doi.org/10.1093/gigascience/giab008

66. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011 Aug 1 [cited 2022 Jun 11]; 27(15):2156–8. https://doi.org/10.1093/bioinformatics/btr330

67. Browning BL, Zhou Y, Browning SR. A One-Penny Imputed Genome from Next-Generation Reference Panels. Am J Hum Genet. 2018 Sep 6 [cited 2021 Jun 11]; 103(3):338–48. https://www.sciencedirect.com/science/article/pii/S0002929718302428

68. Paczian T, Trimble WL, Gerlach W, Harrison T, Wilke A, Meyer F. The MG-RAST API explorer: an on-ramp for RESTful query composition. BMC Bioinformatics. 2019 Nov 8 [cited 2023 Apr 19]; 20(1):561. https://doi.org/10.1186/s12859-019-2993-0

69. Robinson KM, Hawkins AS, Santana-Cruz I, Adkins RS, Shetty AC, Nagaraj S, et al. Aligner optimization increases accuracy and decreases compute times in multi-species sequence data. Microb Genomics. 2017 Jul 8 [cited 2023 Feb 13]; 3(9):e000122. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5643015/

70. Prince D, Mugford S, Vincent T, Hogenhout S. Pea Aphid Survival Assays on Arabidopsis thaliana. BIO-Protoc. 2014 [cited 2023 Oct 11]; 4(19). https://bio-protocol.org/e1251

71. National Center for Biotechnology Information (NCBI). Bethesda (MD): National Library of Medicine (US). 1988. NCBI. https://www.ncbi.nlm.nih.gov/

72. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011 Jan [cited 2023 Oct 12]; 7(1):539. https://www.embopress.org/doi/full/10.1038/msb.2011.75

73. Agerbirk N, Olsen CE, Heimes C, Christensen S, Bak S, Hauser TP. Multiple hydroxyphenethyl glucosinolate isomers and their tandem mass spectrometric distinction in a geographically structured polymorphism in the crucifer Barbarea vulgaris. Phytochemistry. 2015 Jul 1; 115:130–42. https://www.sciencedirect.com/science/article/pii/S0031942214003665

74. Copernicus Climate Change Service. E-OBS daily gridded meteorological data for Europe from 1950 to present derived from in-situ observations. ECMWF; 2020 [cited 2022 Feb 9]. https://cds.climate.copernicus.eu/doi/10.24381/cds.151d3ec6

75. Ziyatdinov A, Vázquez-Santiago M, Brunel H, Martinez-Perez A, Aschard H, Soria JM. lme4qtl: linear mixed models with flexible covariance structure for genetic studies of related individuals. BMC Bioinformatics. 2018 Feb 27 [cited 2022 Feb 1]; 19(1):68. https://doi.org/10.1186/s12859-018-2057-x

76. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 2012 Jul [cited 2020 Feb 10]; 44(7):821–4. https://www.nature.com/articles/ng.2310

77. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet. 2007 Sep 1 [cited 2021 Jun 17]; 81(3):559–75. https://www.sciencedirect.com/science/article/pii/S0002929707613524

78. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019 Nov 14 [cited 2021 Jul 31]; 20(1):238. https://doi.org/10.1186/s13059-019-1832-y

79. Nunn A, Can SN, Otto C, Fasold M, Díez Rodríguez B, Fernández-Pozo N, et al. EpiDiverse Toolkit: a pipeline suite for the analysis of bisulfite sequencing data in ecological plant epigenetics. NAR Genomics Bioinforma. 2021 Dec 1 [cited 2021 Dec 2]; 3(4):lqab106. https://doi.org/10.1093/nargab/lqab106

80. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar 15 [cited 2021 Jun 11]; 26(6):841–2. https://doi.org/10.1093/bioinformatics/btq033

Additional work

Besides the three main studies included as chapters in my PhD thesis, I carried out various additional work during my PhD that resulted in further publications and coauthorships: a book chapter in the gitbook "Ecological Epigenetics", two published papers and another manuscript.

- **Galanti D**. 2022. Epigenetics in Evolution in *Ecological Epigenetics.* Gitbook. https://epidiverse.gitbook.io/project/-MfxkdBDZggX_vc_sG5l/ecology/epigenetics-in-evolution. - 90% contribution.

- Can SN, Nunn A, **Galanti D** , Langenberger D, Becker C, Volmer K, et al. The EpiDiverse Plant Epigenome Wide Association Studies (EWAS) Pipeline. Epigenomes. 2021 Jun;5(2):12. https://www.mdpi.com/2075-4655/5/2/12. - 5% contribution.

- Díez Rodríguez B, **Galanti D**, Nunn A, Peña-Ponton C, Pérez-Bello P, et al. "Epigenetic variation in the Lombardy poplar along climatic gradients is independent of genetic structure and persists across clonal reproduction". https://www.biorxiv.org/content/10.1101/2022.11.17.516862v1. - 7% contribution.

- Sammarco I, Díez Rodríguez B, **Galanti D**, Nunn A, Becker C, Bossdorf O, Münzbergová Z, Latzel V. 2023. "DNA Methylation in the Wild: Epigenetic Transgenerational Inheritance Can Mediate Adaptation in Clones of Wild Strawberry (*Fragaria Vesca*)." *New Phytologist* 241 (4): 1621–35. https://doi.org/10.1111/nph.19464. - 7% contribution.

Discussion

Since their first appearance about 420 million years ago, vascular plants evolved to colonize the most diverse corners of our planet, becoming the major source of organic carbon and therefore vitally important for all life on earth. Understanding how plants evolved in the past and how they adapt to different environmental conditions in the short-term, is key to predict how they might respond to the threats posed by the increasingly rapid climate change. One way to understand plant adaptation is to study intraspecific natural variation, i.e. heritable variation occurring in different populations of a single species, because such natural variation represents a snapshot of the adaptive processes that took place under different environmental conditions in the species' distribution. Analysing natural genetic variation allows us to study these processes, unravel the genetic bases of traits and discover natural alleles that can be used for breeding. Additionally, plant populations harbor extensive epigenetic variation, which often correlates with climate of origin and can potentially affect the phenotype, but its potential role in adaptive evolution is much less understood than that of genetic variation. It is therefore key to understand how natural epigenetic variation arises, whether it is heritable, and how it interacts with DNA sequence variation to possibly contribute to local adaptation. Given the diversity of vascular plants, encompassing different reproduction modes, ploidies, genome sizes and genome evolutions, it is crucial to expand our understanding of local adaptation from a few models to many diverse species.

In my PhD project I studied natural genetic and epigenetic variation in *Thlaspi arvense*, a plant species with a quite complex genome, characterized by extensive TE colonization and DNA methylation. *T. arvense* is increasingly studied as a new model species (Geng et al. 2021, Nunn et al 2022, Hu et al. 2022, Troyee et al. 2022, Galanti et al. 2022), but it is also being developed into a new crop to produce biodiesel from the oil contained in its seeds (Dorn et al. 2015, Frels et al. 2019, Chopra et al. 2019, Zhao et al. 2021). With these features and objectives in mind, I produced and analysed whole genome sequencing and bisulfite sequencing data from a large collection of over 200 *T. arvense* accessions. In

Chapter I I described the DNA methylation variation of the collection and the genetic and environmental drivers of this variation. In **Chapter II** I focused on TE insertion/deletion polymorphisms and identified the genetic basis of variation in transposon mobility. In **Chapter III**, I used sequencing reads to estimate aphid and mildew infestation of each accession grown in a common environment, and identified genetic, epigenetic and environmental factors associated with these differences in infestation.

Below, I discuss in more detail the contribution of this work towards understanding how short genetic variants and TE insertions/deletions interact with one another and with epigenetic variation. I further elaborate on how this complex genome-epigenome system interacts with the environment to provide plasticity and adaptation. In my discussion I also refer to two additional studies I contributed to that focused on natural DNA methylation variation in *Populus nigra* (Rodríguez et al. 2022) and *Fragaria vesca* (Sammarco et al. 2023). Finally, I discuss broader applications of using non-target reads to study antagonist and microbe colonisation of plant collections and identify their genetic and epigenetic basis.

Evolutionary consequences of genome-epigenome interactions

One key issue that hinders our understanding of the drivers and consequences of epigenetic variation, is its complex interaction with DNA sequence variation. On one side, when a genetic polymorphism is controlling an epigenetic one, the latter will seem to be heritable, while it is simply reestablished in every generation because of the genetic polymorphism. This can happen with genetic polymorphisms acting in *cis*, such as TE insertions (Matrin et al. 2009) triggering the RdDM to redeposit methylation on the insertions, or in *trans*, e.g. with variants at genes involved in depositing, maintaining, or removing epigenetic marks (Dubin et al. 2015, Kawakatsu et al. 2016, Sasaki et al. 2019). On the other side, newly arisen epigenetic variation can release transposons from silencing (Roquis et al. 2021) and influence mutation rates (Monroe et al. 2022 and 2023), potentially generating novel genetic variation.

This reciprocal causation can be difficult to disentangle and, to use a common saying, answering whether “the chicken or the egg came first”, can be arduous. It requires at least to investigate full genomes and methylomes of many individuals. In **Chapter I** and in two additional studies (Sammarco et al. 2023, Rodriguez et al 2022), we attempted to address these challenges. We observed extensive methylation variation in natural populations of all three species and investigated its potential drivers. While in the poplar clone “Italica” there is very little sequence variation and DNA methylation variation correlates mostly with climate of origin, in the two sexually reproducing species most DNA methylation variation observed could be explained by underlying *trans*-acting sequence variants. In **Chapter I** I investigated these genetic bases of DNA methylation variation and identified many of these variants, mostly located into genes already known to control DNA methylation or histone modifications. Compared to related species, *T. arvense* has low CHG methylation, which is more similar to CHH than to the much higher CG methylation. This was reflected by our GWA results as the sets of genes controlling CHG and CHH methylation were largely overlapping, while this was very different for CG. Despite this strong genetic component of DNA methylation, I also observed correlations with climate of origin and decided to investigate the relative contributions of genetic versus environmental determination. What stood out clearly for both species, was that the proportion of genetic versus environmental determination highly depended on the sequence context, with an increasing proportion of DMRs best explained by the environment of origin when moving from the more stable CG context to the less stable and more environmentally responsive CHG and especially CHH contexts. Although we generated the two common gardens clonally for *F. vesca* and via seed for *T. arvense*, CHH methylation seemed less stable in the former which did not undertake any potential meiotic resetting. In *F. vesca* we indeed found that many CHH DMRs best explained by the environment in the field became “unexplained” in the common garden, indicating that CHH methylation was already converging to a homogenous common-garden state.

Taken together, these findings clearly indicate that, although DNA methylation is strongly influenced by genetic variation at methylation machinery genes, it is also responsive to environmental conditions,

and the extents of the two processes strongly depend on the sequence context. In a scenario where CHH methylation responds quickly to the environment but is only stable for a few generations, while CHG and CG methylation are gradually more stable but might take longer to respond to environmental changes, one tempting conclusion is that plasticity and adaptation are not completely distinct, but intermediate processes exist based on gradually fast and stable acquisition of new epimutations.

Another very important aspect to consider is the interplay between DNA methylation and transposition. In **Chapter II** we thoroughly investigated this relationship and found that in *T. arvense* not all transposons are equally controlled by DNA methylation. In *A. thaliana*, several pathways exist to silence different transposon groups (Kawakatsu et al. 2016, Sasaki et al. 2019, Baduel et al. 2021), but are all associated with DNA methylation. In *T. arvense*, we found that new retrotransposon insertions became quickly methylated and their activity was indeed highly dependent on the effectiveness of the DNA methylation machinery, i.e. mutations at genes involved in DNA methylation dynamics affect retrotransposons activity. For example we found a knock-out mutation in an exon of the DNA demethylating gene *BRAT1*, which was associated with lower retrotransposition. However, new DNA transposon insertions remained unmethylated and their activity was not linked to the same genes determining retrotransposition, but instead to a single gene coding a heat shock protein orthologue of *Oryza sativa HSP19*. Although *T. arvense HSP19* does not have a clear orthologue in *A. thaliana*, it contains the same α-crystalline domain as At*ACD15* and *21*, members of the *MBD5/6* complex that silences genes and TEs downstream of DNA methylation (Boone et al. 2023). If *HSP19* were to interact with this complex, downstream of DNA methylation, this would explain how this mechanism would bypass the need for DNA methylation to drive DNA transposon silencing.

Another important interaction between genome and epigenome that has been investigated by others is that DNA methylation and other epigenetic marks seem to affect genetic mutation rates, resulting in mutation biases across genomes (see the introduction for an extended discussion) (Monroe et al. 2022, 2023).

Taken together, all these interactions draw a picture in which genomes and epigenomes are not distinct entities with separate roles, but they closely interact with one another and in response to the environment, to provide a complex system capable of both plasticity, adaptation and a plethora of intermediate effects. Moreover, we showed the importance of expanding the study of these mechanisms to several species as these might reveal different systems. We should particularly exploit recent technological advances to expand our studies to more complex genomes.

Thlaspi arvense as a new model and crop

As explained above, *Thlaspi arvense* is an excellent model for studying quite complex genomes with extensive TE-colonisation and DNA methylation. Moreover, this species is emerging as a valuable cover crop that can protect the soil in winter and generate biodiesel without interfering with summer crops, using the oil extracted from its seeds.

The study of transposition in *T. arvense* is particularly important because of the distinct TE colonisation history of this species compared to the well-studied close relative and model species *A. thaliana*, which contains very low amounts of TEs compared to most other plants (Tenaillon et al. 2010). The genome of *T. arvense* has been vastly colonised by retrotransposons of the *Ty3* superfamily, which inserted in all chromosomes in the wide pericentromeric region and pushed genes to the edges of the chromosome arms. Investigating TE insertions in *T. arvense* natural populations, we identified genetic transposition regulators that appear to differ for Class I versus Class II transposons. As mentioned above, the number of retrotransposon (Class I) insertions was linked to variation at genes with known roles in epigenetic regulation, while DNA transposon (Class II) mobility was associated with variation at *HSP19*. This evidence was further supported by the observation that DNA methylation is deposited over new retrotransposon but not DNA transposon insertions, suggesting that while the former are indeed controlled by the DNA methylation machinery through *de-novo* deposition, the latter are likely

controlled by another mechanism bypassing DNA methylation. This difference was not detected as clearly in studies with *A. thaliana*, which lacks *HSP19*. It is therefore likely that this gene provides alternative functionality in *T. arvense*, demonstrating how using a species with a different and more complex TE architecture can provide new insights.

Additionally, we discovered an *Alesia* family of recently active TEs that targets genes. Combining information from this family with its environmental (possibly heat) and genetic regulators could potentially be used to induce genetic variation to be leveraged for breeding purposes. Ongoing research aims to determine its responsiveness to heat, as suggested by sequence similarity to other heat-responsive TE families.

In the context of domesticating *T. arvense* into a new crop, **Chapter III** of my thesis focused on its genetic variation in pest resistance and led to the identification of genomic regions and genes of interest that could potentially be employed in breeding programs for resistant varieties.

Non-target sequencing reads to study pest resistance in plants

With decreasing costs of sequencing, large sequencing datasets are currently increasing rapidly, and in addition to sequences of the target species, they usually contain minor portions of reads from other organisms such as symbionts or antagonists. A few studies already investigated the possibility of 'recycling' these reads, that are usually discarded, to draw information on organisms co-occurring with the target species (Sangiovanni et al. 2019, Roman-Reyna et al. 2020, Gathercole et al. 2021). In **Chapter III** we went several steps further, demonstrating how powerful this information can indeed be. We used the non-target reads to estimate the abundances of aphids and mildew that attacked the *T. arvense* accessions in our glasshouse experiment and found associations with environmental, genetic and even epigenetic factors. The numbers of aphid and mildew reads were associated with environmental variables related to the proliferation of these pests in their locations of origin.

Moreover, the pest abundance data allowed us to map defense genes involved in resistance to these pests and identify beneficial alleles, proving the effectiveness of this method. *T. arvense* has several natural enemies such as aphids, mildew, caterpillars, beetles and more (Warwick et al. 2002), therefore identifying resistance alleles has high potential value for breeding it into a crop that can be used successfully for large-scale biodiesel production.

Using the full methylomes generated in **Chapter I**, we additionally found DNA methylation variants associated with pest abundance. However, a challenge with our data was that we could not be sure whether these methylation variants were present before the pest attack (beneficial methylation states are expected to co-occur with low pest load), or whether they were plastically induced by the pathogens (beneficial methylation states are expected to co-occur with high pest load). We discussed this extensively in **Chapter III** and were possibly able to unravel it in at least one case. We found a large DMR, hypomethylated in affected samples, located on *Copia* TEs in the promoter of the *MPKKK20* gene, involved in stomata closure. Most likely the *MPKKK20* promoter loses methylation upon mildew infection to activate the gene and close the stomata, preventing further mildew access to the leaf blade. Additional experiments such as DNA methylation and expression analysis of the *MPKKK20* promoter and gene respectively, before and after mildew infection, would be necessary to corroborate this hypothesis.

Overall, we showed how powerful non-target reads can be to quantify organisms co-occurring with the target species, and we are confident this will be vastly applicable, given the rising number of large sequencing projects (e.g. Kajiya-Kanegae 2021, Colgan et al. 2022, Habyarimana et al. 2022, Mekbib et al. 2022, Metheringham et al. 2022, Nocchi et al. 2022, Friis et al. 2024). Other potential applications of non-target reads include for example co-GWAS approaches to identify host-pathogen variants that coevolved (Dexter et al. 2023), but this requires high coverage also for the pathogen, for variant calling, and therefore a specifically designed sequencing strategy.

Future perspectives

The work I presented in my thesis opens future perspectives for both basic research in plant biology, including the study of epigenetic and transposable element regulation, and for applied breeding, as we identified potential beneficial alleles for pest resistance and the *Alesia* TE family which has the potential to generate novel variation at genes, likely to impact phenotypes. Many of these aspects would benefit from further exploration. The most important and obvious follow-up would be to investigate whether natural epigenetic variants affect phenotypic traits. Although I did not present this data in any of the chapters, during my PhD I scored several phenotypic traits on the whole collection, and in **Chapter III** I developed a method to carry out Epigenome Wide Association analysis similarly to its genetic counterpart GWA. Comparing results from GWA and EWA would certainly be a powerful strategy for discovering DNA methylation variants associated with phenotypes. Another option for disentangling phenotypic variation explained by genetic versus DNA methylation variants could be genomic prediction (Van Raden et al. 2017). Quantifying whether the addition of DNA methylation information to genomic prediction models improves the predictions compared to using only genetic variants would be a potential way to answer whether natural DNA methylation carries additional phenotypic information missing in the nucleotide sequence.

References

Baduel, Pierre, Basile Leduque, Amandine Ignace, Isabelle Gy, José Gil, Olivier Loudet, Vincent Colot, and Leandro Quadrana. 2021. "Genetic and Environmental Modulation of Transposition Shapes the Evolutionary Potential of *Arabidopsis thaliana*." *Genome Biology* 22 (1): 138. https://doi.org/10.1186/s13059-021-02348-5.

Boone, Brandon A., Lucia Ichino, Shuya Wang, Jason Gardiner, Jaewon Yun, Yasaman Jami-Alahmadi, Jihui Sha, et al. 2023. "ACD15, ACD21, and SLN Regulate the Accumulation and Mobility of MBD6 to Silence Genes and Transposable Elements." Science Advances 9 (46): eadi9036. https://doi.org/10.1126/sciadv.adi9036.

Chopra, Ratan, Nicole Folstad, Joseph Lyons, Tim Ulmasov, Cynthia Gallaher, Liam Sullivan, Abby McGovern, et al. 2019. "The Adaptable Use of Brassica NIRS Calibration Equations to Identify Pennycress Variants to Facilitate the Rapid Domestication of a New Winter Oilseed Crop." *Industrial Crops and Products* 128 (February): 55–61. https://doi.org/10.1016/j.indcrop.2018.10.079.

Colgan, Thomas J, Andres N Arce, Richard J Gill, Ana Ramos Rodrigues, Abdoulie Kanteh, Elizabeth J Duncan, Li Li, Lars Chittka, and Yannick Wurm. 2022. “Genomic Signatures of Recent Adaptation in a Wild Bumblebee.” *Molecular Biology and Evolution* 39 (2): msab366. https://doi.org/10.1093/molbev/msab366.

Dexter, Eric, Peter D Fields, and Dieter Ebert. 2023. “Uncovering the Genomic Basis of Infection Through Co-Genomic Sequencing of Hosts and Parasites.” *Molecular Biology and Evolution* 40 (7): msad145. https://doi.org/10.1093/molbev/msad145.

Dorn, Kevin M., Johnathon D. Fankhauser, Donald L. Wyse, and M. David Marks. 2015. “A Draft Genome of Field Pennycress (*Thlaspi Arvense*) Provides Tools for the Domestication of a New Winter Biofuel Crop.” *DNA Research* 22 (2): 121–31. https://doi.org/10.1093/dnares/dsu045.

Dubin, Manu J, Pei Zhang, Dazhe Meng, Marie-Stanislas Remigereau, Edward J Osborne, Francesco Paolo Casale, Philipp Drewe, et al. 2015. “DNA Methylation in Arabidopsis Has a Genetic Basis and Shows Evidence of Local Adaptation.” *eLife* 4. https://doi.org/10.7554/eLife.05255.

Frels, Katherine, Ratan Chopra, Kevin M. Dorn, Donald L. Wyse, M. David Marks, and James A. Anderson. 2019. “Genetic Diversity of Field Pennycress (*Thlaspi Arvense*) Reveals Untapped Variability and Paths Toward Selection for Domestication.” *Agronomy* 9 (6): 302. https://doi.org/10.3390/agronomy9060302.

Friis, Guillermo, Edward G. Smith, Catherine E. Lovelock, Alejandra Ortega, Alyssa Marshell, Carlos M. Duarte, and John A. Burt. 2024. “Rapid Diversification of Grey Mangroves (Avicennia Marina) Driven by Geographic Isolation and Extreme Environmental Conditions in the Arabian Peninsula.” *Molecular Ecology* n/a (n/a): e17260. https://doi.org/10.1111/mec.17260.

Galanti, Dario, Daniela Ramos-Cruz, Adam Nunn, Isaac Rodríguez-Arévalo, J. F. Scheepens, Claude Becker, and Oliver Bossdorf. 2022. “Genetic and Environmental Drivers of Large-Scale Epigenetic Variation in *Thlaspi Arvense*.” *PLOS Genetics* 18 (10): e1010452. https://doi.org/10.1371/journal.pgen.1010452.

Gathercole, Louise A. P., Gabriele Nocchi, Nathan Brown, Timothy L. R. Coker, William J. Plumb, Jonathan J. Stocks, Richard A. Nichols, Sandra Denman, and Richard J. A. Buggs. 2021. “Evidence for the Widespread Occurrence of Bacteria Implicated in Acute Oak Decline from Incidental Genetic Sampling.” *Forests* 12 (12): 1683. https://doi.org/10.3390/f12121683.

Geng, Yupeng, Yabin Guan, La Qiong, Shugang Lu, Miao An, M. James C. Crabbe, Ji Qi, Fangqing Zhao, Qin Qiao, and Ticao Zhang. 2021. “Genomic Analysis of Field Pennycress (*Thlaspi Arvense*) Provides Insights into Mechanisms of Adaptation to High Elevation.” *BMC Biology* 19 (1): 143. https://doi.org/10.1186/s12915-021-01079-0.

Habyarimana, Ephrem, Sunita Gorthy, Faheem S. Baloch, Sezai Ercisli, and Gyuhwa Chung. 2022. “Whole-Genome Resequencing of Sorghum Bicolor and S. Bicolor × S. Halepense Lines Provides New Insights for Improving Plant Agroecological Characteristics.” *Scientific Reports* 12 (1): 5556. https://doi.org/10.1038/s41598-022-09433-0.

Hu, Yanting, Xiaopei Wu, Guihua Jin, Junchu Peng, Rong Leng, Ling Li, Daping Gui, Chuanzhu Fan, and Chengjun Zhang. 2022. “Rapid Genome Evolution and Adaptation of *Thlaspi Arvense* Mediated by Recurrent RNA-Based and Tandem Gene Duplications.” *Frontiers in Plant Science* 12. https://www.frontiersin.org/articles/10.3389/fpls.2021.772655.

Kajiya-Kanegae, Hiromi, Hideki Nagasaki, Akito Kaga, Ko Hirano, Eri Ogiso-Tanaka, Makoto Matsuoka, Motoyuki Ishimori, et al. 2021. "Whole-Genome Sequence Diversity and Association Analysis of 198 Soybean Accessions in Mini-Core Collections." *DNA Research* 28 (1): dsaa032. https://doi.org/10.1093/dnares/dsaa032.

Kawakatsu, Taiji, Shao-shan Carol Huang, Florian Jupe, Eriko Sasaki, Robert J. Schmitz, Mark A. Urich, Rosa Castanon, et al. 2016. "Epigenomic Diversity in a Global Collection of *Arabidopsis thaliana* Accessions." *Cell* 166 (2): 492–505. https://doi.org/10.1016/j.cell.2016.06.044.

Martin, Antoine, Christelle Troadec, Adnane Boualem, Mazen Rajab, Ronan Fernandez, Halima Morin, Michel Pitrat, Catherine Dogimont, and Abdelhafid Bendahmane. 2009. "A Transposon-Induced Epigenetic Change Leads to Sex Determination in Melon." *Nature* 461 (7267): 1135–38. https://doi.org/10.1038/nature08498.

Mekbib, Yeshitila, Kassahun Tesfaye, Xiang Dong, Josphat K. Saina, Guang-Wan Hu, and Qing-Feng Wang. 2022. "Whole-Genome Resequencing of Coffea Arabica L. (Rubiaceae) Genotypes Identify SNP and Unravels Distinct Groups Showing a Strong Geographical Pattern." *BMC Plant Biology* 22 (1): 69. https://doi.org/10.1186/s12870-022-03449-4.

Metheringham, Carey L., William J. Plumb, Jonathan J. Stocks, Laura J. Kelly, Miguel Nemesio Gorriz, Justin Moat, Richard J. A. Buggs, and Richard A. Nichols. 2022. "Rapid Polygenic Adaptation in a Wild Population of Ash Trees under a Novel Fungal Epidemic." bioRxiv. https://doi.org/10.1101/2022.08.01.502033.

Monroe, J. Grey, Kevin D. Murray, Wenfei Xian, Thanvi Srikant, Pablo Carbonell-Bejerano, Claude Becker, Mariele Lensink, et al. 2023. "Reply to: Re-Evaluating Evidence for Adaptive Mutation Rate Variation." *Nature* 619 (7971): E57–60. https://doi.org/10.1038/s41586-023-06315-x.

Monroe, J. Grey, Thanvi Srikant, Pablo Carbonell-Bejerano, Claude Becker, Mariele Lensink, Moises Exposito-Alonso, Marie Klein, et al. 2022. "Mutation Bias Reflects Natural Selection in *Arabidopsis thaliana*." *Nature*, January, 1–5. https://doi.org/10.1038/s41586-021-04269-6.

Nocchi, Gabriele, Nathan Brown, Timothy L. R. Coker, William J. Plumb, Jonathan J. Stocks, Sandra Denman, and Richard J. A. Buggs. 2022. "Genomic Structure and Diversity of Oak Populations in British Parklands." *PLANTS, PEOPLE, PLANET* 4 (2): 167–81. https://doi.org/10.1002/ppp3.10229.

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* n/a (n/a). https://doi.org/10.1111/pbi.13775.

Rodríguez, Bárbara Díez, Dario Galanti, Adam Nunn, Cristian Peña-Ponton, Paloma Pérez-Bello, Iris Sammarco, Katharina Jandrasits, et al. 2022. "Epigenetic Variation in the Lombardy Poplar along Climatic Gradients Is Independent of Genetic Structure and Persists across Clonal Reproduction." bioRxiv. https://doi.org/10.1101/2022.11.17.516862.

Roman-Reyna, Veronica, Dale Pinili, Frances N. Borja, Ian L. Quibod, Simon C. Groen, Nickolai Alexandrov, Ramil Mauleon, and Ricardo Oliva. 2020. "Characterization of the Leaf Microbiome from Whole-Genome Sequencing Data of the 3000 Rice Genomes Project." *Rice* 13 (1): 72. https://doi.org/10.1186/s12284-020-00432-1.

Roquis, David, Marta Robertson, Liang Yu, Michael Thieme, Magdalena Julkowska, and Etienne Bucher. 2021. "Genomic Impact of Stress-Induced Transposable Element Mobility in Arabidopsis." *Nucleic Acids Research* 49 (18): 10431–47. https://doi.org/10.1093/nar/gkab828.

Sammarco, Iris, Bárbara Díez Rodríguez, Dario Galanti, Adam Nunn, Claude Becker, Oliver Bossdorf, Zuzana Münzbergová, and Vít Latzel. 2023. "DNA Methylation in the Wild: Epigenetic Transgenerational Inheritance Can Mediate Adaptation in Clones of Wild Strawberry (Fragaria Vesca)." *New Phytologist* 241 (4): 1621–35. https://doi.org/10.1111/nph.19464.

Sangiovanni, Mara, Ilaria Granata, Amarinder Singh Thind, and Mario Rosario Guarracino. 2019. "From Trash to Treasure: Detecting Unexpected Contamination in Unmapped NGS Data." *BMC Bioinformatics* 20 (4): 168. https://doi.org/10.1186/s12859-019-2684-x.

Sasaki, Eriko, Taiji Kawakatsu, Joseph R. Ecker, and Magnus Nordborg. 2019. "Common Alleles of CMT2 and NRPE1 Are Major Determinants of CHH Methylation Variation in *Arabidopsis thaliana*." *PLOS Genetics* 15 (12): e1008492. https://doi.org/10.1371/journal.pgen.1008492.

Tenaillon, Maud I., Jesse D. Hollister, and Brandon S. Gaut. 2010. "A Triptych of the Evolution of Plant Transposable Elements." Trends in Plant Science 15 (8): 471–78. https://doi.org/10.1016/j.tplants.2010.05.003.

Troyee, A. Niloya, Mónica Medrano, Caroline Müller, and Conchita Alonso. 2022. "Variation in DNA Methylation and Response to Short-Term Herbivory in *Thlaspi arvense*." *Flora* 293 (August): 152106. https://doi.org/10.1016/j.flora.2022.152106.

VanRaden, Paul M., Melvin E. Tooker, Jeffrey R. O'Connell, John B. Cole, and Derek M. Bickhart. 2017. "Selecting Sequence Variants to Improve Genomic Predictions for Dairy Cattle." Genetics Selection Evolution 49 (1): 32. https://doi.org/10.1186/s12711-017-0307-4.

Warwick, S. I., A. Francis, and D. J. Susko. 2002. "The Biology of Canadian Weeds. 9. *Thlaspi arvense* L. (Updated)." *Canadian Journal of Plant Science* 82 (4): 803–23. https://doi.org/10.4141/P01-159.

Zhao, Ru, Xinyu Yang, Muzhi Li, Xiaojin Peng, Mengxia Wei, Xiucheng Zhang, Lei Yang, and Jialei Li. 2021. "Biodiesel Preparation from *Thlaspi arvense* L. Seed Oil Utilizing a Novel Ionic Liquid Core-Shell Magnetic Catalyst." *Industrial Crops and Products* 162 (April): 113316. https://doi.org/10.1016/j.indcrop.2021.113316.

Acknowledgments

First of all, I want to thank my supervisor Oliver! None of this would have been possible without you Oliver! Thanks for your constant support, in both science and life, and for believing in me even in the toughest moments. You taught me how a supportive, motivating, fun and friendly environment is key for doing great science! Thanks for sharing your knowledge, for your tips on navigating the science world, handling situations and making life choices. I will always value your advice!

I also really want to thank Detlef Weigel and Claude Becker for being my second supervisors, for always giving me great advice to improve, and for the papers we published together! You are great scientists, working with you has been an honour.

So many people to thank in the PlantEvoEco group! Niek you were a great second supervisor, always available, supportive and helpful. Thanks to all the students and technicians that helped during data collection. Corinna for a great thesis, Christiane, Eva and Sabine for all the extensive practical help. Martina, thanks for having my back when I was struggling with German. A big thank you to all my colleagues that made my experience in the group unforgettable! Bence, Robert, Shirley, Anna and Franziska for their warm welcoming and first 3 years. Frank, Junhee and Madalin thanks for the chats, bike rides and for sharing fried grasshoppers! Junhee thanks to your brain and our aphids. Thanks to everyone else in the group for sharing nice chats, science and gourmet mensa meals.

Another amazing group of people that I really want to thank are the members of EpiDiverse! I feel so lucky I could be part of this! The science, meetings, conferences and holidays all around Europe were the best! A special thanks to Adam for showing me the wonderland of bioinformatics and to Adrian for always teasing my brain, introducing me to the TE world and for our great collaboration. To Iris and Barbara for many fruitful and fun discussions, for sharing code, ideas and papers. And to everybody else for the amazing science and really fun times.

A special thanks also to Dani, Troyee and Valentina that bore with my workaholic self and helped me cross the whole of Europe to collect Thlaspi seeds!

Finally I want to thank my family that has always been more than supportive and gave me the opportunity to pursue my love for science. And all my friends, whether they were back home or scattered around the world. Thanks to all the amazing people I met in Tübingen, that made these 5 years super fun and full of adventures, thanks for being there when I needed you, for the bbqs, chats at Jäger, parties, walks, rides, drinks and endless laughs!