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

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

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

[2] 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 glucosinolate contents. 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 it 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) essays, and such essays 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 it 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, more generally, 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, where more and larger datasets are being generated for many species (37–43), the use of non-target reads will likely have 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 consistently occurred in 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 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 incl. *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*, and 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).

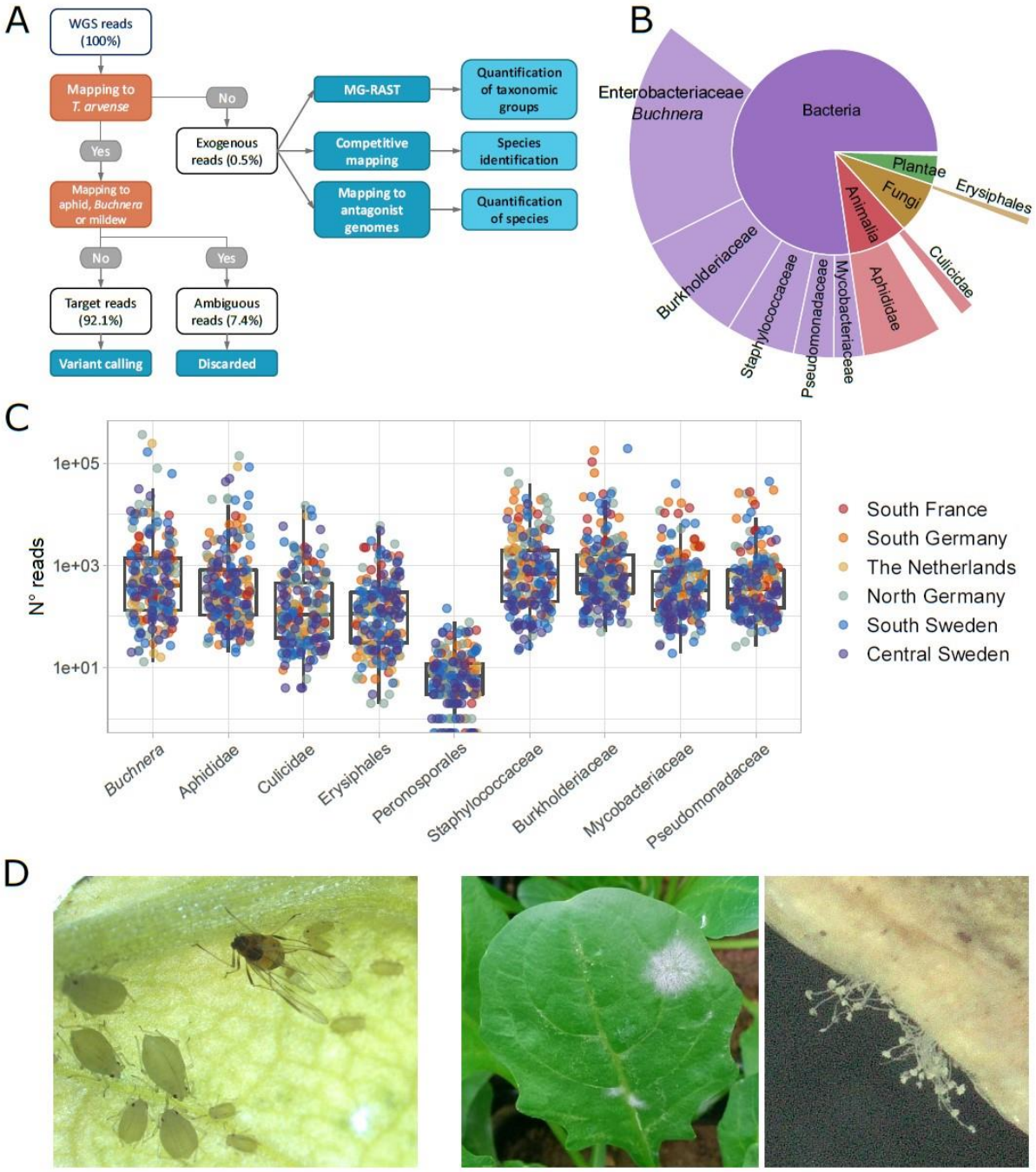


Fig 1. Classification of sequencing reads in *Thlaspi arvense* WGS data. (A) Workflow of the analyses, including reads classification (orange nodes) into target, ambiguous and exogenous reads, and downstream analysis (dark blue nodes) (see Methods). (B) Fractions of exogenous reads assigned to different 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 our visual observation had been that aphid and mildew infections in the glasshouse were not random, but more prevalent on 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 we found that there were no significant greenhouse differences between the offspring of plants that had been attacked by aphids or mildew 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 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 indeed 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, albeit somewhat 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. The glucosinolates levels were measured on the same lines in a separate experiment not affected by pests (S7 Table). We found positive correlations of aphid read counts with an aliphatic glucosinolate, allyl glucosinolate (sinigrin), which is by far the most abundant glucosinolate in the leaves of *T. arvense*, and a stronger negative correlation with benzyl glucosinolate (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.

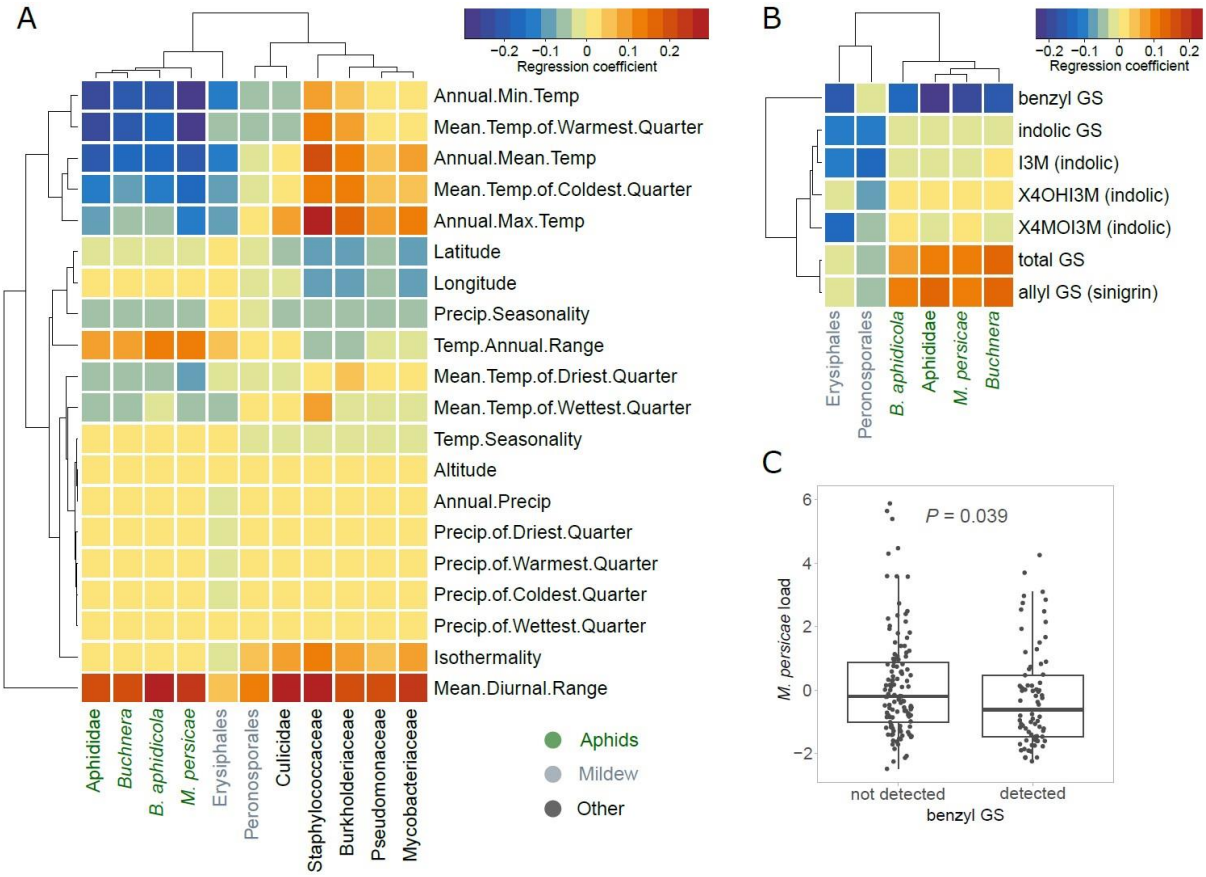


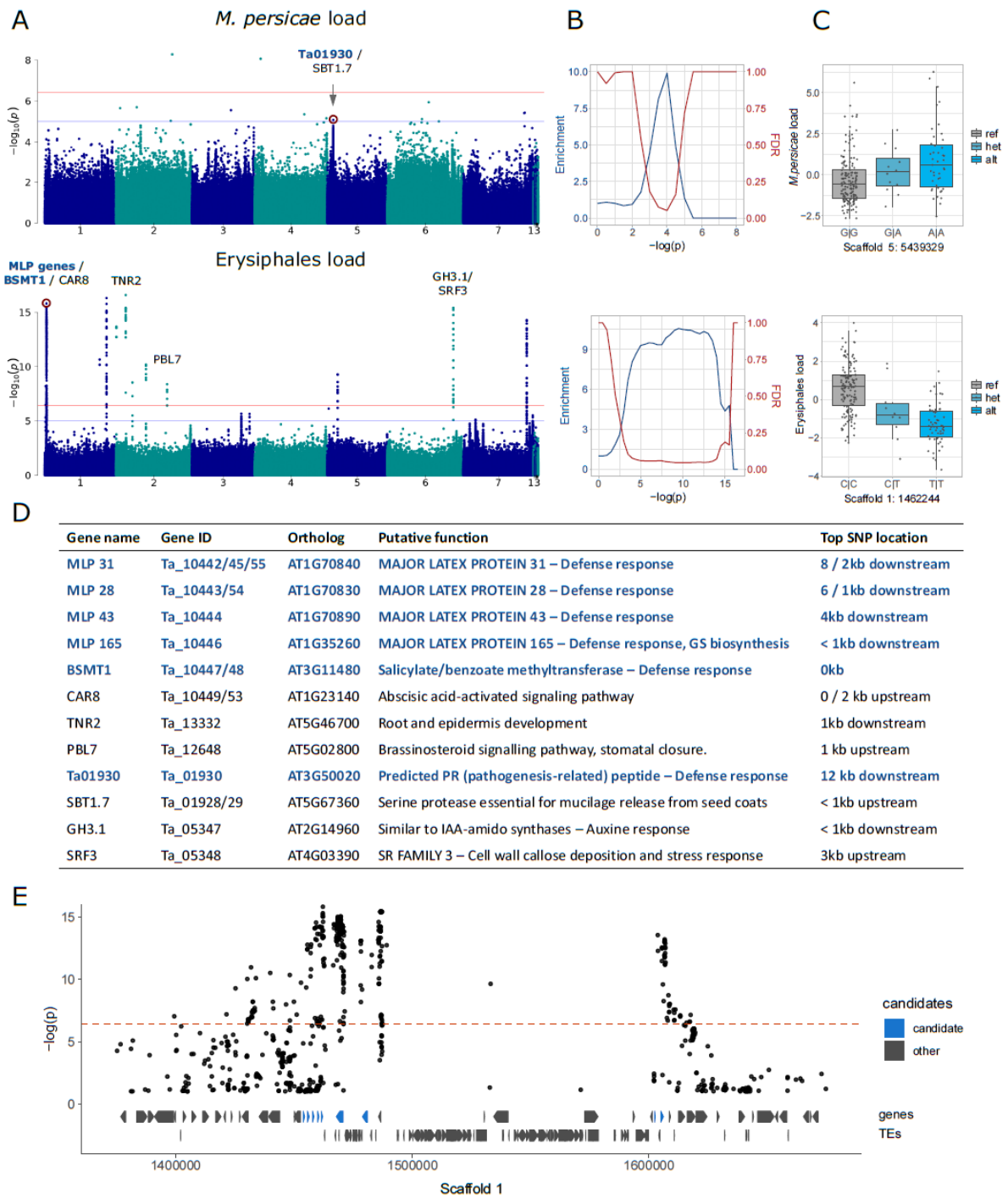
Fig 2. Relationships between climates of origin or glucosinolate levels of plants and the exogenous reads loads. (A) Correlations with bioclimatic variables. (B) Correlations with baseline glucosinolate (GS) levels measured in the same pennycress lines in another experiment. All correlations in (A) and (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 or 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 only 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 mild 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 neighbouring candidate genes involved in defense (Fig 3E). There were other significant peaks for Erysiphales load which were close to other genes possibly contributing to resistance such as *PBL7*, involved in signaling and stomatal closure or *SRF3*, reinforcing cell walls by callose deposition.



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.

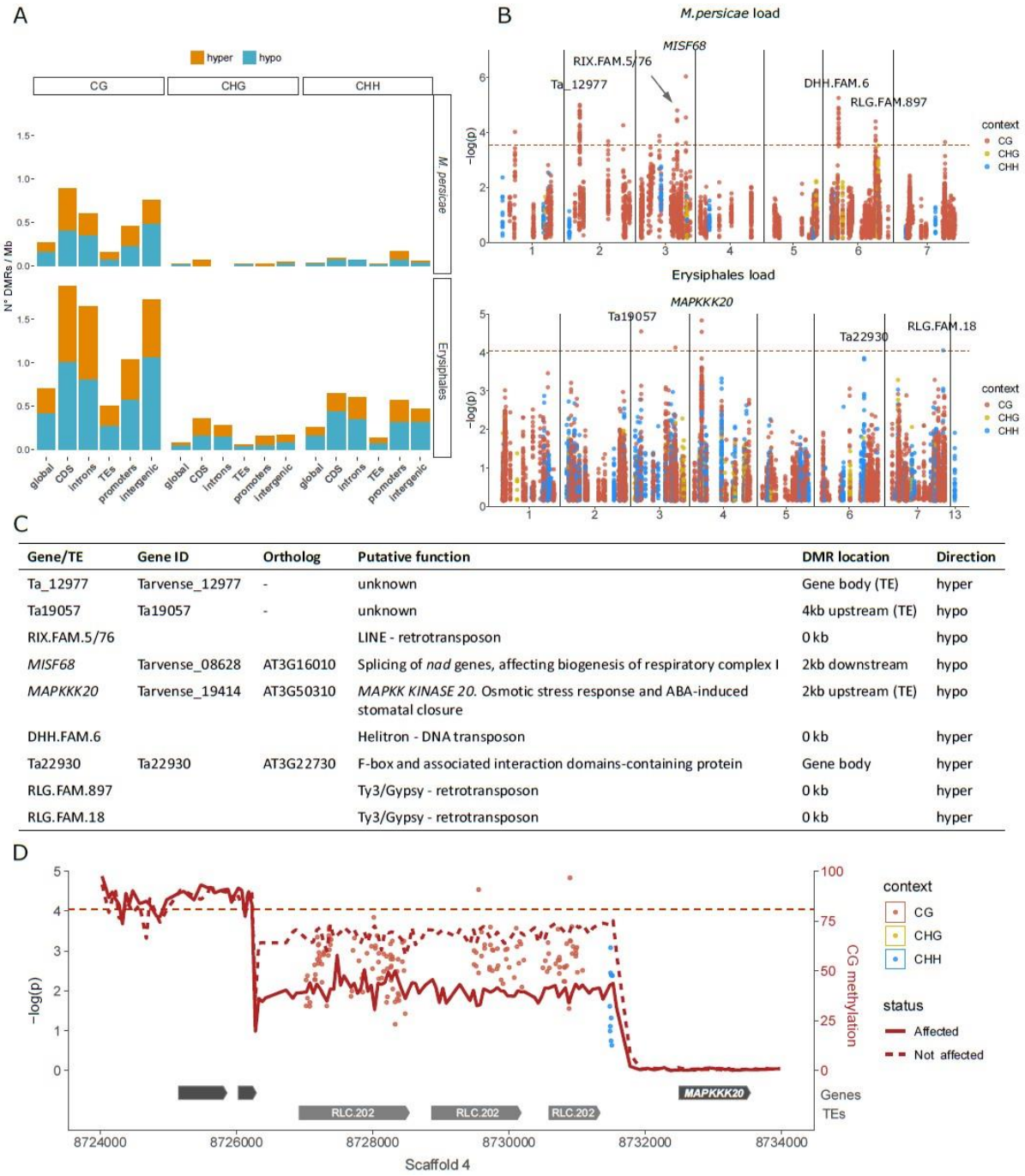
Aphid and mildew loads correlate with differential methylation at genes and transposons

Variation in phenotypes - such as our indirect estimates of pest resistance - may not only be associated with DNA sequence but also with epigenetic changes such as 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 in this study and to test for associations between DNA methylation variation and pest attack. For the sake of 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 the 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).



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) Zoom-in of the 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, could 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 it can in principle be applied to many other situations such as field experiments. It is not error-free, but we highlight some possible 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 started off by classifying all reads into target (only mapping to *T. arvense*), ambiguous (mapping to *T. arvense* and at the same time to at least one of the pest genomes) and exogenous reads (not mapping to *T. arvense*) (Fig 1A). We then demonstrated how important it was to remove the ambiguous reads prior to variant calling, because 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 is low but 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 originated from climates less favorable to aphids to perform worse (higher pathogen loads) in our glasshouse. 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 thermally more fluctuating climates, which are less favorable to aphids, were less well defended and performed worse in our glasshouse. We found similar but weaker patterns also for Erysiphales load.

Since 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 GWA. For aphids (*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 they can also reflect plastic responses to environmental stresses such as herbivores or pathogens. Our data do not allow to confidently distinguish between these two sources of DNA methylation variation and should thus 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 the 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 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 instead 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 suspect 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* (eg. LINE, Helitron and Ty3/Gypsi TEs isolated from genes). Although we cannot confidently tell apart inherited versus induced DNA methylation variants, as far as we know 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 might 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 can be used to quantify pathogen distributions in the field. With decreasing cost of sequencing and more and more large-scale and single-species sequencing projects eg. (37–43), the number of data sets suitable for such analyses will likely rapidly increase in the 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, Staphylocaccaceae, 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 23 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.

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 neighbouring genes potentially affecting *exogenous reads* load.

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 before and above the gene *MPKKK20*.

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 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. 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. 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