eLife

Reviewed Preprint
v2 • October 29, 2024
Revised by authors

Reviewed Preprint
v1 • April 3, 2024

Evolutionary Biology

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

Dario Galanti ✉, **Jun Hee Jung, Caroline Müller, Oliver Bossdorf**

Plant Evolutionary Ecology, Institute of Evolution and Ecology, University of Tübingen, 72076 Tübingen, Germany • Royal Botanic Gardens, Kew, Richmond upon Thames, UK • Chemical Ecology, Bielefeld University, Universitätststr. 25, 33615 Bielefeld, Germany

eLife Assessment

This **important** study presents a significant methodological advance by leveraging previously discarded, unmapped DNA sequence reads to estimate pest infestation loads across plant accessions, and map variation in these apparent pest loads to defense genes. The bioinformatics approach is **compelling**, and the results should bear broad implications for phenotype-genotype prediction, especially regarding the use of unmapped reads for GWAS.

https://doi.org/10.7554/eLife.95510.2.sa3

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

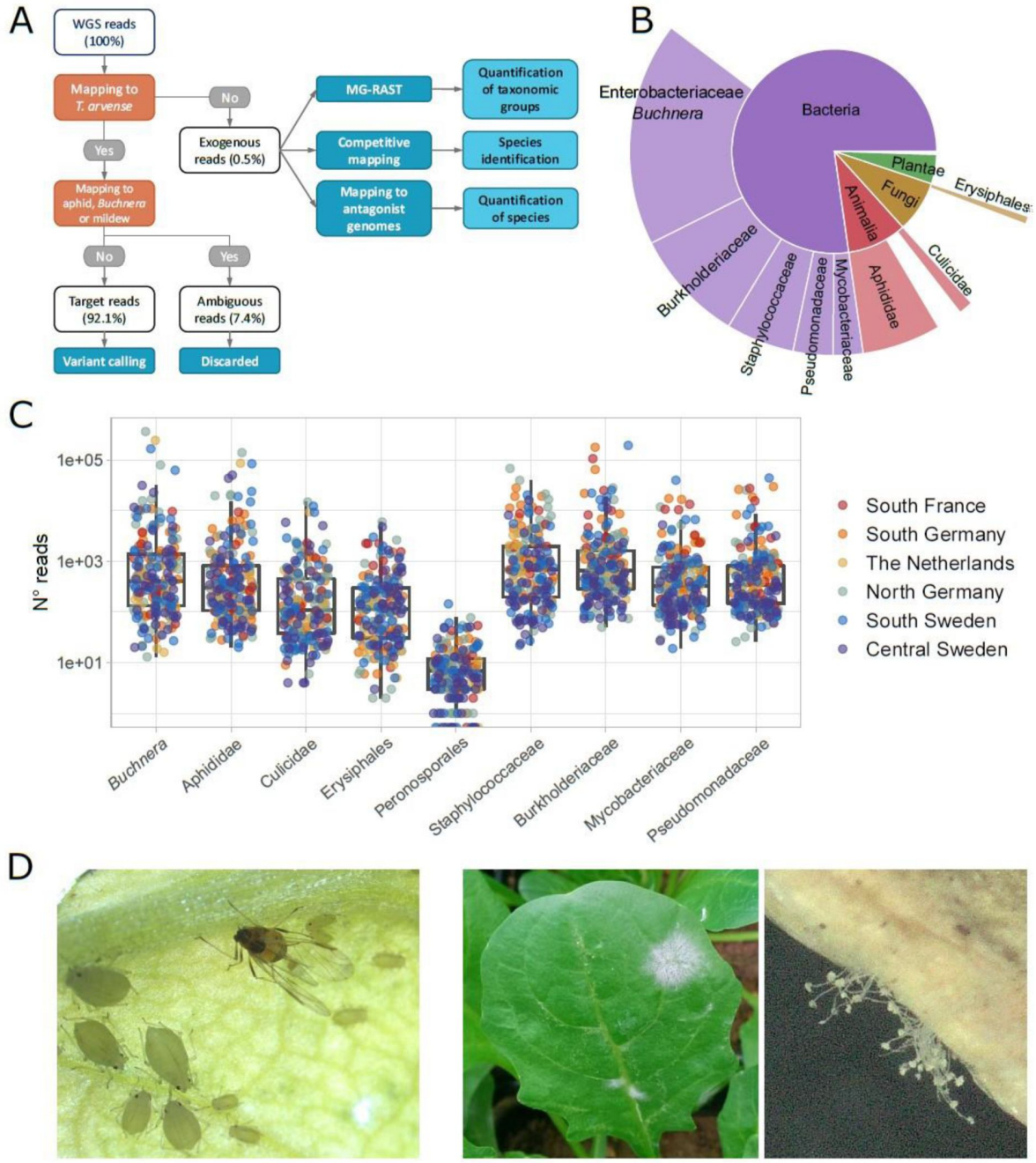


Fig 1.

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.

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).

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 read-counts. 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 (see Methods).

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 variation in enemy densities 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).

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 (1,52), 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.

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* (51).

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)

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 in the glasshouse (**Fig 2C**). We also detected three indole glucosinolates, but these did not show any significant correlations with aphid loads.

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 (53) 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.

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

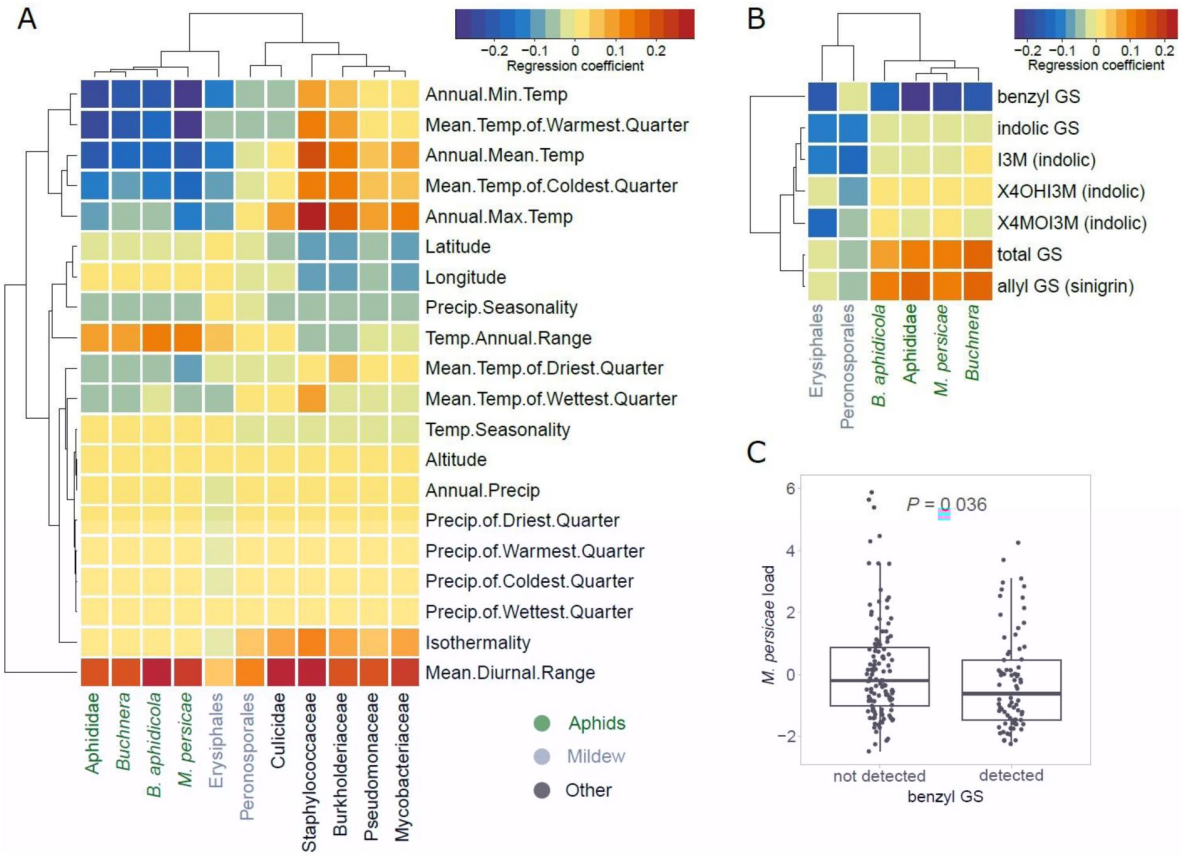
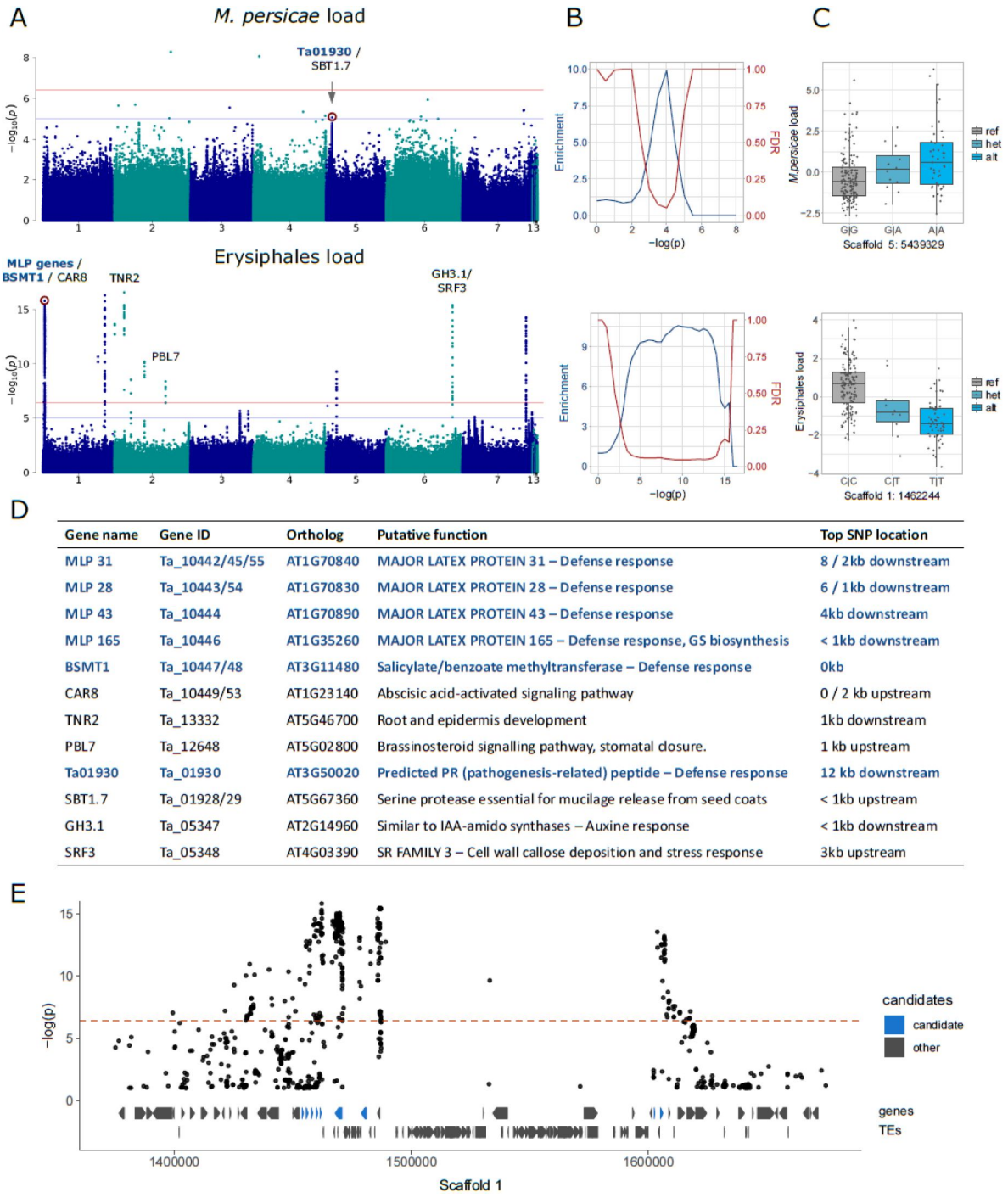


Fig 2.

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. The *P* value is based on the Welch's *t*-test for unequal variances.



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 (54), 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 (53)) 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.

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.

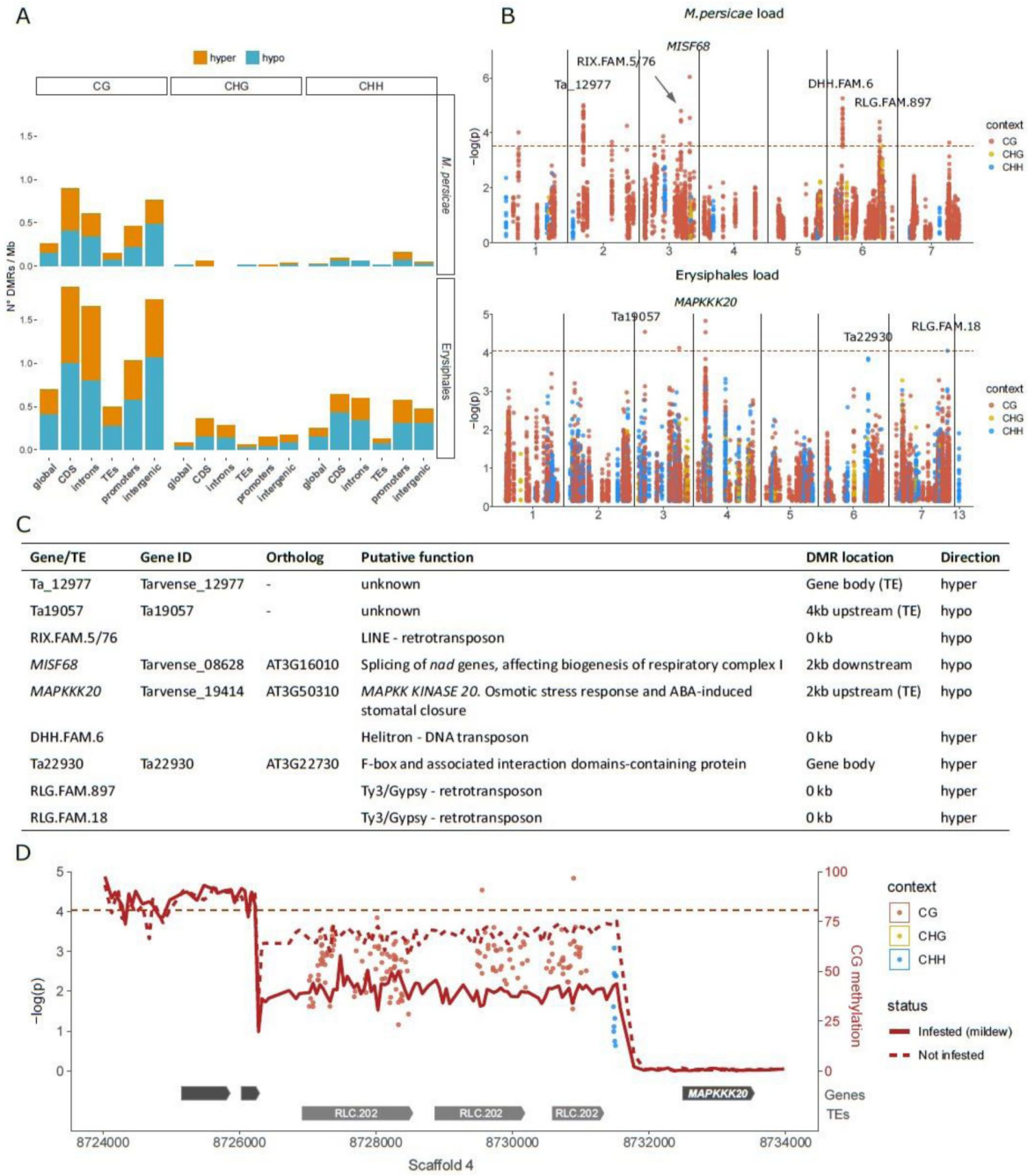
Our analysis had two steps: First we called Differentially Methylated Regions (DMRs) between the 20 samples with the most and 20 samples with the least *M. persicae* or Erysiphales loads, and then we conducted Epigenome Wide Association (EWA) analyses on individual positions located within these DMRs, using the complete dataset (188 lines - 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 infested samples (higher pathogen load), 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*.

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



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 infested 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 infested samples was calculated over 50 bp bins (see S5C Fig for methylation in other contexts).

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* (55), 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 (56).

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 3A** 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 3E**). *MLP165*, the closest gene to the most significant variant in the peak, is indirectly involved in GS biosynthesis in *A. thaliana* (57), 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 3D**). 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 (58) 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 potentially act both in *cis* (eg. Copia TE methylation in *MAPKKK20* promoter) and in *trans*, for example 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 (59) for quality (minimum quality of 20) and adaptor trimming, excluding reads shorter than 35 bp. We used FastQC and MultiQC (60,61) 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 (62), excluding multimapping reads (-c 1) and marking duplicates with MarkDuplicatesSpark (63,64). We then mapped all samples again (62) 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*) (65). 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 (67). 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 (63,64), 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 (67) 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 (68) 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 (69) 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 (70) and removing reads with MAPQ < 20 and duplicates with MarkDuplicatesSpark (63,64). 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, correcting 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.

To test for variation between populations we used a general linear model with population as a predictor. To measure SNP-based heritability, i.e. the proportion of variance explained by kinship, we used the marker_h2() function from the R package *heritability* (51), which uses a genetic distance matrix as predictor to compute REML-estimates of the genetic and residual variance. We used the same IBS matrix as for GWAS and for the correlations with climatic variables.

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 (71). 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 (72) under accession numbers *MN232006*, *NC_011594* and *NC_056270*. We downloaded these sequences, aligned them to each other (73), 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) (74).

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 (75), as described in Galanti et al. 2022 (32). We then used the R package *lme4qtl* (76) 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 (77). To obtain the IBS matrix we used PLINK v 1.90b6.12 (78). 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) (53). Briefly, we attributed *a priori* candidate status to all variants located within 20 kb from orthologs (79) of *A. thaliana* genes annotated with the GO term “defense response” (GO:0006952), including nine genes similar to *AtBSMT1* (S8 Table). We then calculated enrichment for incremental −log(p) thresholds as the ratio between observed frequency (significant *a priori* candidate / significant variants) and background frequency (total *a priori* candidate / total variants), and an upper bound for the FDR (32,53). We further assessed the significance of the enrichment through a previously established genome rotation scheme (80). Briefly, we calculated a null distribution of enrichments by randomly rotating the

P-values and *a priori* candidate status of the genetic variants within each chromosome for 1M permutations. We then assessed significance by comparing the observed enrichment at the Bonferroni threshold to the null distribution. 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 (81), 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), which left 188 samples for analysis. 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 (82) 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 (53), 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*.

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.

Acknowledgements

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 (2012) **Natural Enemies Drive Geographic Variation in Plant Defenses** *Science* **338**:116–9

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

3. Chan EKF, Rowe HC, Kliebenstein DJ (2010) **Understanding the Evolution of Defense Metabolites in Arabidopsis thaliana Using Genome-wide Association Mapping** *Genetics* **185**: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. (2016) **The Quantitative Basis of the Arabidopsis Innate Immune System to Endemic Pathogens Depends on Pathogen Genetics** *PLOS Genet* **12**

5. Thoen MPM, Olivas NHD, Kloth KJ, Coolen S, Huang PP, Aarts MGM, et al. (2017) **Genetic architecture of plant stress resistance: multi-trait genome-wide association mapping** *New Phytol* **213**:1346–62

6. Hanson AA, Lorenz AJ, Hesler LS, Bhusal SJ, Bansal R, Michel AP, et al. (2018) **Genome-Wide Association Mapping of Host-Plant Resistance to Soybean Aphid** *Plant Genome* **11**

7. Jaouannet M, Morris JA, Hedley PE, Bos JIB (2015) **Characterization of Arabidopsis Transcriptional Responses to Different Aphid Species Reveals Genes that Contribute to Host Susceptibility and Non-host Resistance** *PLOS Pathog* **11**

8. Geng S, Kong X, Song G, Jia M, Guan J, Wang F, et al. (2019) **DNA methylation dynamics during the interaction of wheat progenitor Aegilops tauschii with the obligate biotrophic fungus Blumeria graminis f. sp. tritici** *New Phytol* **221**:1023–35

9. Annacondia ML, Markovic D, Reig-Valiente JL, Scaltsoyiannes V, Pieterse CMJ, Ninkovic V, et al. (2021) **Aphid feeding induces the relaxation of epigenetic control and the associated regulation of the defense response in Arabidopsis** *New Phytol* **230**:1185–200

10. Kinoshita T, Seki M (2014) **Epigenetic Memory for Stress Response and Adaptation in Plants** *Plant Cell Physiol* **55**:1859–63

11. Espinas NA, Saze H, Saijo Y (2016) **Epigenetic Control of Defense Signaling and Priming in Plants** *Front Plant Sci* **7**

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

13. He Y, Li Z (2018) **Epigenetic Environmental Memories in Plants: Establishment, Maintenance, and Reprogramming** *Trends Genet*

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

15. Law JA, Jacobsen SE (2010) **Establishing, maintaining and modifying DNA methylation patterns in plants and animals** *Nat Rev Genet* **11**:204–20

16. Zhang H, Lang Z, Zhu JK (2018) **Dynamics and function of DNA methylation in plants** *Nat Rev Mol Cell Biol* **1**

17. Liu J, He Z (2020) **Small DNA Methylation, Big Player in Plant Abiotic Stress Responses and Memory** *Front Plant Sci* **11**

18. Wojtaszek P (1997) **Oxidative burst: an early plant response to pathogen infection** *Biochem J* **322**:681–92

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

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

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

22. Muhammad T, Zhang F, Zhang Y, Liang Y (2019) **RNA Interference: A Natural Immune System of Plants to Counteract Biotic Stressors** *Cells* **8**

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

24. Nalam V, Louis J, Patel M, Shah J (2018) **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* **8**

25. Sangiovanni M, Granata I, Thind AS, Guarracino MR (2019) **From trash to treasure: detecting unexpected contamination in unmapped NGS data** *BMC Bioinformatics* **20** 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. (2020) **Characterization of the Leaf Microbiome from Whole-Genome Sequencing Data of the 3000 Rice Genomes Project** *Rice* **13** https://doi.org/10.1186/s12284-020-00432-1

27. Gathercole LAP, Nocchi G, Brown N, Coker TLR, Plumb WJ, Stocks JJ, et al. (2021) **Evidence for the Widespread Occurrence of Bacteria Implicated in Acute Oak Decline from Incidental Genetic Sampling** *Forests* **12**

28. Geng Y, Guan Y, Qiong L, Lu S, An M, Crabbe MJC, et al. (2021) **Genomic analysis of field pennycress (Thlaspi arvense) provides insights into mechanisms of adaptation to high elevation** *BMC Biol* **19** 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. (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 Biotechnol J*

30. Hu Y, Wu X, Jin G, Peng J, Leng R, Li L, et al. (2022) **Rapid Genome Evolution and Adaptation of Thlaspi arvense Mediated by Recurrent RNA-Based and Tandem Gene Duplications** *Front Plant Sci* **12**

31. Troyee AN, Medrano M, Müller C, Alonso C (2022) **Variation in DNA methylation and response to short-term herbivory in Thlaspi arvense** *Flora* **293**

32. 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**

33. Dorn KM, Fankhauser JD, Wyse DL, Marks MD (2015) **A draft genome of field pennycress (Thlaspi arvense) provides tools for the domestication of a new winter biofuel crop** *DNA Res* **22**:121–31

34. Frels K, Chopra R, Dorn KM, Wyse DL, Marks MD, Anderson JA (2019) **Genetic Diversity of Field Pennycress (Thlaspi arvense) Reveals Untapped Variability and Paths Toward Selection for Domestication** *Agronomy* **9**

35. Chopra R, Johnson EB, Emenecker R, Cahoon EB, Lyons J, Kliebenstein DJ, et al. (2019) **Progress toward the identification and stacking of crucial domestication traits in pennycress** *Plant Biology*

36. Zhao R, Yang X, Li M, Peng X, Wei M, Zhang X, et al. (2021) **Biodiesel preparation from Thlaspi arvense L. seed oil utilizing a novel ionic liquid core-shell magnetic catalyst** *Ind Crops Prod* **162**

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

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

39. Habyarimana E, Gorthy S, Baloch FS, Ercisli S, Chung G (2022) **Whole-genome resequencing of Sorghum bicolor and S. bicolor × S. halepense lines provides new insights for improving plant agroecological characteristics** *Sci Rep* **12**

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

41. Metheringham CL, Plumb WJ, Stocks JJ, Kelly LJ, Gorriz MN, Moat J, et al. (2022) **Rapid polygenic adaptation in a wild population of ash trees under a novel fungal epidemic** *bioRxiv*

42. Nocchi G, Brown N, Coker TLR, Plumb WJ, Stocks JJ, Denman S, et al. (2022) **Genomic structure and diversity of oak populations in British parklands** *PLANTS PEOPLE PLANET* **4**:167–81

43. Friis G, Smith EG, Lovelock CE, Ortega A, Marshall A, Duarte CM, et al. (2024) **Rapid diversification of grey mangroves (Avicennia marina) driven by geographic isolation and extreme environmental conditions in the Arabian Peninsula** *Mol Ecol* **33**

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

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

46. (2021) **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 (1999) **Acceptability of different species of Brassicaceae as hosts for the cabbage aphid** *Entomol Exp Appl* **91**:105–9

48. Warwick SI, Francis A, Susko DJ (2002) **The biology of Canadian weeds. 9. Thlaspi arvense L. (updated)** *Can J Plant Sci* **82**

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

50. Singh KS, Cordeiro EMG, Troczka BJ, Pym A, Mackisack J, Mathers TC, et al. (2021) **Global patterns in genomic diversity underpinning the evolution of insecticide resistance in the aphid crop pest Myzus persicae** *Commun Biol* **4**:1–16

51. (2019) **Kruijer, Willem, and with a contribution from Ian White, contains data collected by Padraic Flood and Rik Kooke. 2019. "Heritability: Marker-Based Estimation of Heritability Using Individual Plant or Plot Data." https://CRAN.R-project.org/package=heritability.**

52. Gao J, Fang C, Zhao B (2019) **The latitudinal herbivory hypothesis revisited: To be part is to be whole** *Ecol Evol* **9**:3681–8

53. Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, et al. (2010) **Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines** *Nature* **465**:627–31

54. Sobota RS, Shriner D, Kodaman N, Goodloe R, Zheng W, Gao YT, et al. (2015) **Addressing population-specific multiple testing burdens in genetic association studies** *Ann Hum Genet* **79**:136–47

55. Kim JH, Jander G (2007) **Myzus persicae (green peach aphid) feeding on Arabidopsis induces the formation of a deterrent indole glucosinolate** *Plant J* **49**:1008–19

56. Klingauf F, Şengonca Ç, Bennewitz H (1972) **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* **9**:53–7

57. (2000) **The Arabidopsis Information Resource (TAIR). 2000. www.arabidopsis.org/servlets/TairObject?id=137911&type=locus**

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

59. Martin M (2011) **Cutadapt removes adapter sequences from high-throughput sequencing reads** *EMBnet.journal* **17**:10–2

60. Andrews S (2010) **FASTQC. A quality control tool for high throughput sequence data** *BibSonomy*

61. Ewels P, Magnusson M, Lundin S, Käller M (2016) **MultiQC: summarize analysis results for multiple tools and samples in a single report** *Bioinformatics* **32**:3047–8 https://doi.org/10.1093/bioinformatics/btw354

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

63. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, et al. (2013) **From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline** *Curr Protoc Bioinforma* **43**:11–11

64. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Auwera GAV der, et al. (2018) **Scaling accurate genetic variant discovery to tens of thousands of samples** *bioRxiv*

65. Frantzeskakis L, Kracher B, Kusch S, Yoshikawa-Maekawa M, Bauer S, Pedersen C, et al. (2018) **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* **19** https://doi.org/10.1186/s12864-018-4750-6

66. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. (2021) **Twelve years of SAMtools and BCFtools** *GigaScience* **10** https://doi.org/10.1093/gigascience/giab008

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

68. Browning BL, Zhou Y, Browning SR (2018) **A One-Penny Imputed Genome from Next-Generation Reference Panels** *Am J Hum Genet* **103**:338–48

69. Paczian T, Trimble WL, Gerlach W, Harrison T, Wilke A, Meyer F (2019) **The MG-RAST API explorer: an on-ramp for RESTful query composition** *BMC Bioinformatics* **20** https://doi.org/10.1186/s12859-019-2993-0

70. Robinson KM, Hawkins AS, Santana-Cruz I, Adkins RS, Shetty AC, Nagaraj S, et al. (2017) **Aligner optimization increases accuracy and decreases compute times in multi-species sequence data** *Microb Genomics* **3**

71. Prince D, Mugford S, Vincent T, Hogenhout S (2014) **Pea Aphid Survival Assays on Arabidopsis thaliana** *BIO-Protoc* **4**

72. (1988) **National Center for Biotechnology Information (NCBI). Bethesda (MD): National Library of Medicine (US). 1988. NCBI. https://www.ncbi.nlm.nih.gov/**

73. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. (2011) **Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega** *Mol Syst Biol* **7**

74. Agerbirk N, Olsen CE, Heimes C, Christensen S, Bak S, Hauser TP (2015) **Multiple hydroxyphenethyl glucosinolate isomers and their tandem mass spectrometric distinction in a geographically structured polymorphism in the crucifer Barbarea vulgaris** *Phytochemistry* **115**:130–42

75. (2020) **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**

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

77. Zhou X, Stephens M (2012) **Genome-wide efficient mixed-model analysis for association studies** *Nat Genet* **44**:821–4

78. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. (2007) **PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses** *Am J Hum Genet* **81**:559–75

79. Emms DM, Kelly S (2019) **OrthoFinder: phylogenetic orthology inference for comparative genomics** *Genome Biol* **20** https://doi.org/10.1186/s13059-019-1832-y

80. Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, Zheng H, et al. (2005) **The Pattern of Polymorphism in Arabidopsis thaliana** *PLOS Biology* **3** https://doi.org/10.1371/journal.pbio.0030196

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

82. Quinlan AR, Hall IM (2010) **BEDTools: a flexible suite of utilities for comparing genomic features** *Bioinformatics* **26**:841–2 https://doi.org/10.1093/bioinformatics/btq033

Editors

Reviewing Editor
Sergio Rasmann
University of Neuchâtel, Neuchâtel, Switzerland

Senior Editor
Meredith Schuman
University of Zurich, Zürich, Switzerland

Reviewer #1 (Public Review):

Galanti et al. present an innovative new method to determine the susceptibility of large collections of plant accessions towards infestations by herbivores and pathogens. This work resulted from an unplanned infestation of plants in a greenhouse that was later harvested for sequencing. When these plants were extracted for DNA, associated pest DNA was extracted and sequenced as well. In a standard analysis, all sequencing reads would be mapped to the plant reference genome and unmapped reads, most likely originating from 'exogenous' pest

DNA, would be discarded. Here, the authors argue that these unmapped reads contain valuable information and can be used to quantify plant infestation loads.

For the present manuscript, the authors re-analysed a published dataset of 207 sequenced accessions of Thlaspi arvense. In this data, 0.5% of all reads had been classified as exogenous reads, while 99.5% mapped to the T. arvense reference genome. In a first step, however, the authors repeated read mapping against other reference genomes of potential pest species and found that a substantial fraction of 'ambiguous' reads mapped to at least one such species. Removing these reads improved the results of downstream GWAs, and is in itself an interesting tool that should be adopted more widely.

The exogenous reads were primarily mapped to the genomes of the aphid Myzus persicae and the powdery mildew Erysiphe cruciferarum, from which the authors concluded that these were the likely pests present in their greenhouse. The authors then used these mapped pest read counts as an approximate measure of infestation load and performed GWA studies to identify plant gene regions across the T. arvense accessions that were associated with higher or lower pest read counts. In principle, this is an exciting approach that extracts useful information from 'junk' reads that are usually discarded. The results seem to support the authors' arguments, with relatively high heritabilities of pest read counts among T. arvense accessions, and GWA peaks close to known defence genes. Nonetheless, I do feel that more validation would be needed to support these conclusions, and given the radical novelty of this approach, additional experiments should be performed.

A weakness of this study is that no actual aphid or mildew infestations of plants were recorded by the authors. They only mention that they anecdotally observed differences in infestations among accessions. As systematic quantification is no longer possible in retrospect, a smaller experiment could be performed in which a few accessions are infested with different quantities of aphids and/or mildew, followed by sequencing and pest read mapping. Such an approach would have the added benefit of allowing causally linking pest read count and pest load, thereby going beyond correlational associations.

On a technical note, it seems feasible that mildew-infested leaves would have been selected for extraction, but it is harder to explain how aphid DNA would have been extracted alongside plant DNA. Presumably, all leaves would have been cleaned of live aphids before they were placed in extraction tubes. What then is the origin of aphid DNA in these samples? Are these trace amounts from aphid saliva and faeces/honeydew that were left on the leaves? If this is the case, I would expect there to be substantially more mildew DNA than aphid DNA, yet the absolute read counts for aphids are actually higher. Presumably read counts should only be used as a relative metric within a pest organism, but this unexpected result nonetheless raises questions about what these read counts reflect. Again, having experimental data from different aphid densities would make these results more convincing.

Comments on revised version:

The authors have addressed many technical details in their revision, but they did not address my more fundamental concerns about validation of their results. I still believe that validation would be needed, but I also acknowledge that an additional experiment that reliably tests a causal relationship between read counts and pest abundance would go beyond the scope of a revision. Nonetheless, the authors currently only show variation in pest read counts among plant accessions, not in pest abundance. While the two measures are likely correlated, I hope that future studies will address more directly how pest abundance and read counts are causally linked, and whether pest read counts truly are a robust measure of pest abundance across a range of conditions and systems

https://doi.org/10.7554/eLife.95510.2.sa2

Reviewer #2 (Public Review):

Summary:

Galanti et al investigate genetic variation in plant pest resistance using non-target reads from whole-genome sequencing of 207 field lines spontaneously colonized by aphids and mildew. They calculate significant differences in pest DNA load between populations and lines, with heritability and correlation with climate and glucosinolate content. By genome-wide association analyses they identify known defence genes and novel regions potentially associated with pest load variation. Additionally, they suggest that differential methylation at transposons and some genes are involved in responses to pathogen pressure. The authors present in this study the potential of leveraging non-target sequencing reads to estimate plant biotic interactions, in general for GWAS, and provide insights into the defence mechanisms of Thlaspi arvense.

Strengths:

The authors ask an interesting and important question. Overall, I found the manuscript very well-written, with a very concrete and clear question, a well-structured experimental design, and clear differences from previous work. Their important results could potentially have implications and utility for many systems in phenotype-genotype prediction. In particular, I think the use of unmapped reads for GWAS is intriguing.

Comments on revised version:

The revisions to the manuscript have significantly enhanced its clarity and scientific rigor. Methodological clarifications, especially regarding the normalization of read counts, now provide a stronger foundation for the presented results. Statistical enhancements, including more robust methods for controlling population structure and refined GWAS approaches, have solidified the reliability of the findings, effectively linking genetic variants and epigenetic modifications to pest loads. The discussion section has been improved to offer a more cautious interpretation of the correlations between transposable element (TE) methylation and pathogen load, emphasizing the associative nature of these findings. Additionally, increased transparency in data handling, particularly the treatment of ambiguous reads, has significantly reduced potential biases. These improvements have made the manuscript better suited to the readership, providing clearer insights into the genomic and epigenetic underpinnings of plant pest resistance.

https://doi.org/10.7554/eLife.95510.2.sa1

Author response:

The following is the authors' response to the original reviews.

> ***Reviewer #1 (Public Review):***
>
> *Galanti et al. present an innovative new method to determine the susceptibility of large collections of plant accessions towards infestations by herbivores and pathogens. This work resulted from an unplanned infestation of plants in a greenhouse that was later harvested for sequencing. When these plants were extracted for DNA, associated pest DNA was extracted and sequenced as well. In a standard analysis, all sequencing reads would be mapped to the plant reference genome and unmapped reads, most likely originating from 'exogenous' pest DNA, would be discarded. Here, the authors argue that these unmapped reads contain valuable information and can be used to quantify plant infestation loads.*

For the present manuscript, the authors re-analysed a published dataset of 207 sequenced accessions of Thlaspi arvense. In this data, 0.5% of all reads had been classified as exogenous reads, while 99.5% mapped to the T. arvense reference genome. In a first step, however, the authors repeated read mapping against other reference genomes of potential pest species and found that a substantial fraction of 'ambiguous' reads mapped to at least one such species. Removing these reads improved the results of downstream GWAs, and is in itself an interesting tool that should be adopted more widely.

The exogenous reads were primarily mapped to the genomes of the aphid Myzus persicae and the powdery mildew Erysiphe cruciferarum, from which the authors concluded that these were the likely pests present in their greenhouse. The authors then used these mapped pest read counts as an approximate measure of infestation load and performed GWA studies to identify plant gene regions across the T. arvense accessions that were associated with higher or lower pest read counts. In principle, this is an exciting approach that extracts useful information from 'junk' reads that are usually discarded. The results seem to support the authors' arguments, with relatively high heritabilities of pest read counts among T. arvense accessions, and GWA peaks close to known defence genes. Nonetheless, I do feel that more validation would be needed to support these conclusions, and given the radical novelty of this approach, additional experiments should be performed.

A weakness of this study is that no actual aphid or mildew infestations of plants were recorded by the authors. They only mention that they anecdotally observed differences in infestations among accessions. As systematic quantification is no longer possible in retrospect, a smaller experiment could be performed in which a few accessions are infested with different quantities of aphids and/or mildew, followed by sequencing and pest read mapping. Such an approach would have the added benefit of allowing causally linking pest read count and pest load, thereby going beyond correlational associations.

On a technical note, it seems feasible that mildew-infested leaves would have been selected for extraction, but it is harder to explain how aphid DNA would have been extracted alongside plant DNA. Presumably, all leaves would have been cleaned of live aphids before they were placed in extraction tubes. What then is the origin of aphid DNA in these samples? Are these trace amounts from aphid saliva and faeces/honeydew that were left on the leaves? If this is the case, I would expect there to be substantially more mildew DNA than aphid DNA, yet the absolute read counts for aphids are actually higher. Presumably read counts should only be used as a relative metric within a pest organism, but this unexpected result nonetheless raises questions about what these read counts reflect. Again, having experimental data from different aphid densities would make these results more convincing.

We agree with the reviewer that additional aphid counts at the time of (or prior to) sequencing would have been ideal, but unfortunately we do not have these data. However, compared to such counts one strength of our sequencing-based approach is that it (presumably) integrates over longer periods than a single observation (e.g. if aphid abundances fluctuated, or winged aphids visited leaves only temporarily), and that it can detect pathogens even when invisible to our eyes, e.g. before a mildew colony becomes visible. Moreover, the key point of our study is that we can detect variation in pest abundance even in the absence of count data, which are really time consuming to collect.

Conducting a new experiment, with controlled aphid infestations and continuous monitoring of their abundances, to test for correlation between pest abundance and the number of detected reads would require resequencing at least 30-50% of the collection for the results to be reliable. It would be a major experimental study in itself.

Regarding the origin of aphid reads and the differences in read-counts between e.g. aphids and mildew, we believe this should not be of concern. DNA contamination is very common in all kinds of samples, but these reads are simply discarded in other studies. For example, although we collected and handled samples using gloves, MG-RAST detected human reads (Hominidae, S2 Table), possibly from handling the plants during transplanting or phenotyping 1-2 weeks before sequencing. Therefore, although we did remove aphids from the leaves at collection, aphid saliva or temporary presence on leaves must have been enough to leave detectable DNA traces. Additionally, the fact that the *M. persicae* load strongly correlates with the *Buchnera aphidicola* load (R2=0.86, S6 Table), is reassuring. This obligate aphid symbiont is expected to be found in high amounts when sequencing aphids (see e.g. The International Aphid Genomics Consortium (2010))

The higher amount of aphid compared to mildew reads, can probably be explained by aphids having expanded more than mildew at the time of plant collection, but most importantly, as already mentioned by the reviewer, the read-counts were meant to compare plant accessions rather then pests to one another. We are interested in relative not absolute values. Comparisons between pest species are a challenge because they can be influenced by several factors such as the availability of sequences in the MG-RAST database and the DNA extraction kit used, which is plant-specific and might bias towards certain groups. All these potential biases are not a concern when comparing different plants as they are equally subject to these biases.

> ***Reviewer #2 (Public Review):***
>
> *Summary:*
>
> *Galanti et al investigate genetic variation in plant pest resistance using non-target reads from whole-genome sequencing of 207 field lines spontaneously colonized by aphids and mildew. They calculate significant differences in pest DNA load between populations and lines, with heritability and correlation with climate and glucosinolate content. By genome-wide association analyses they identify known defence genes and novel regions potentially associated with pest load variation. Additionally, they suggest that differential methylation at transposons and some genes are involved in responses to pathogen pressure. The authors present in this study the potential of leveraging non-target sequencing reads to estimate plant biotic interactions, in general for GWAS, and provide insights into the defence mechanisms of Thlaspi arvense.*
>
> *Strengths:*
>
> *The authors ask an interesting and important question. Overall, I found the manuscript very well-written, with a very concrete and clear question, a well-structured experimental design, and clear differences from previous work. Their important results could potentially have implications and utility for many systems in phenotype-genotype prediction. In particular, I think the use of unmapped reads for GWAS is intriguing.*

Thank you for appreciating the originality and potential of our work.

> *Weaknesses:*
>
> *I found that several of the conclusions are incomplete, not well supposed by the data and/or some methods/results require additional details to be able to be judged. I believe these analyses and/or additional clarifications should be considered.*

Thank you very much for the supportive and constructive comments. They helped us to improve the manuscript.

Recommendations for the authors:

Reviewing Editor (Recommendations For The Authors):

The authors address an interesting and significant question, with a well-written manuscript that outlines a clear experimental design and distinguishes itself from previous work. However, some conclusions seem incomplete, lacking sufficient support from the data, or requiring additional methodological details for proper evaluation. Addressing these limitations through additional analyses or clarifications is recommended.

Reviewer #2 (Recommendations For The Authors):

Major comments:

- So far it is not clear to me how read numbers were normalised and quantified. For instance, Figure 1C only reports raw read numbers. In L149: "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". Was library size considered? Is the load the ratio between exogenous vs no exogenous reads? It is described in L461, but according to this, read counts were normalised and duplicated reads were removed. Now, why read counts were used? As opposite to total coverage / or count of bases per base? I cannot follow how variation in sequencing quality was considered. I can imagine that samples with higher sequencing depth will tend to have higher exogenous reads (just higher resolution and power to detect something in a lower proportion).

Correcting for sequencing depth/library size is indeed very important. As the reviewer noted, we had explained how we did this in the methods section (L464), and we now also point to it in the results (L151):

“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, correcting for the sequencing depth of each sample.”

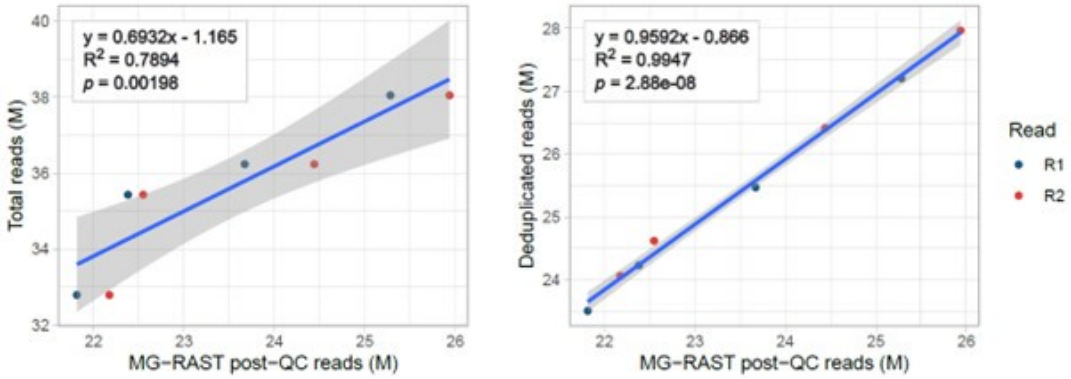
We showed the uncorrected read-counts only in Fig 1 to illustrate the orders of magnitude but used the corrected read-counts (also referred to as “loads”) for all subsequent analyses.

In our view, theoretically, the best metric to correct the number of reads of a specific contaminant organism, is the total number of DNA fragments captured. Importantly, this is not well reflected by the total number of raw reads because of PCR and optical duplicates occurring during library prep and sequencing. For this reason we estimated the total number of reads captured multiplying total raw reads (after trimming) by the deduplication rate obtained from FastQC (methods L409-411). This metric reflects the amount of DNA fragments sampled better than the raw reads. Also it better reflects MG-RAST metrics as this software also deduplicates reads (Author response image 1 below). We also removed duplicates in our strict mappings to the *M. persicae* and *B. aphidicola* genomes.

Coverage is not a good option for correction, because it is defined for a specific reference genome and many of the read-counts output by MG-RAST do not have a corresponding full assembly. Moreover, coverage and base counts are influenced by read size, which depends on library prep and is not included in the read-counts produced by MG-RAST.

Author response image 1.

Linear correlations between the number of MG-RAST reads post-QC and either total (left) or deduplicated (right) reads from fastq files of four full samples (not only unmapped reads).



> *- The general assumption is that plants with different origins will have genetic variants or epigenetic variations associated with pathogen resistance, which can be tracked in a GWAS. However, plants from different regions will also have all variants associated with their origin (isolation by state as presented in the manuscript). In line 169: "Having established that our method most likely captured variation in plant resistance, we were interested in the ecological drivers of this variation". It is not clear to me how variation in plant resistance is differentiated from geographical variation (population structure). in L203: "We corrected for population structure using an IBS matrix and only tested variants with Minor Allele Frequency (MAF) > 0.04 (see Methods).". However, if resistant variants are correlated with population structure as shown in Table 1, how are they differentiated? In my opinion, the analyses are strongly limited by the correlation between phenotype and population structure.*

The association of any given trait with population structure is surely a very important aspect in GWAS studies and when looking at correlations of traits with environmental variables. If a trait is strongly associated with population structure, then disentangling variants associated with population structure vs. the ones associated with the trait can indeed be challenging, a good example being flowering time in *A. thaliana* (e.g. Brachi et al. 2013).

In our case, although the pest and microbiome loads are associated with population structure to some extent, this association is not very strong. This can be observed for example in Fig. 1C, where there is no clear separation of samples from different regions. This means that we can correct for population structure (in both GWAS and correlations with climatic variables) without removing the signals of association. It is possible that other associations were missed if specific variants were indeed strongly associated with structure, but these would be unreliable within our dataset, so it is prudent to exclude them.

> *- Similarly, in L212: "we still found significant GWA peaks for Erysiphales but not for other types of exogenous reads (excluding isolated, unreliable variants) (Figure 3A and S3 Figure)." In a GWA analysis, multiple variants will constitute an association pick (as shown for instance in main Figure 3A) only when the pick is accentuated by lockage disequilibrium around the region under selection (or around the variant explaining phenotypic variation in this case). However, in this case, I suspect there is a strong component of population structure (which still needs to be corroborated as suggested in the previous comment). But if variants are filtered by population structure, the only*

> *variants considered are those polymorphic within populations. In this case, I do not think clear picks are expected since most of the signal, correlated with population has been removed. Under this scenario, I wonder how informative the analyses are.*

As mentioned above, the traits we analyse (aphid and mildew loads) are only partially associated with population structure. This is evident from Fig. 1C (see answer above) but also from the SNP-based heritability (Table 1, last column) which measures indeed the proportion of variance explained by genetic population structure. Although some variance is explained (i.e. the reviewer is correct that there is some association) there is still plenty of leftover variance to be used for GWAS and correlations with environmental variables. The fact that we still find GWAS peaks confirms this, as otherwise they would be lost by the population structure correction included in our mixed model.

> *- How were heritability values calculated? Were related individuals filtered out? I suggest adding more detail in both the inference of heritability and the kinship matrix (IBS matrix). Currently missing in methods (for heritability I only found the mention of an R package in the caption of Table 1).*

We somehow missed this in the methods and thank the reviewer for noticing. We now added this paragraph to the chapter “Exogenous reads heritability and species identification”:
“To test for variation between populations we used a general linear model with population as a predictor. To measure SNP-based heritability, i.e. the proportion of variance explained by kinship, we used the marker_h2() function from the R package heritability (Kruijer and Kooke 2019), which uses a genetic distance matrix as predictor to compute REML-estimates of the genetic and residual variance. We used the same IBS matrix as for GWAS and for the correlations with climatic variables.”

We also added the reference to the R package *heritability* to the Table 1 caption.

> *- Figure 2C. in line 188: "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 (Figure 3C)". It is not clear to me how to see these values in Figure 2C. From the boxplot, the difference in aphid loads between detected and not detected benzyl seems significantly lower. From the boxplot distribution is not clear how this difference is statistically significant. It rather seems like a sampling bias (a lot of non-detected vs low detected values). Is the difference still significant when random subsampling of groups is considered?*

Here the “70% less reads” refers to the uncorrected read-counts directly (difference in means between samples where benzyl-GS were detected vs. not). We agree with the reviewer that this is confusing when referred to figure 2C which depicts the corrected *M. persicae* load (residuals). We therefore removed that information.

Regarding the significance of the difference, we re-calculated the p value with the Welch's t-test, which accounts for unequal variances, and with a bootstrap t-test. Both tests still found a significant difference. We now report the p value of the Welch’s t-test.

> *- I think additional information regarding the read statistics needs to be improved. At the moment some sections are difficult to follow. I found this information mainly in Supplementary Table 1. I could not follow the difference in the manuscript and supplementary materials between read (read count), fragment, ambiguous fragments, target fragments, etc. I didn't find information regarding mean coverage per sample and relative plant vs parasite coverage. This lack of clarity led me to some confusion. For instance, in L207: "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". I find it difficult to follow how non-Thlaspi reads will interfere with genotyping. I think the fact that the large pick is lost after filtering reads is already quite insightful. However, in principle I would expect the relative coverage between non-Thlaspi:Thlaspi reads to be rather low in all cases. I would say below 1%. Thus, genotyping should be relatively accurate for the plant variants for the most part. In particular, considering genotyping was done with GATK, where low-frequency variants (relative coverage) should normally be called reference allele for the most part.*

We agree with the reviewer that some clarification over these points is necessary! We modified Supplementary Table 1 to include coverage information for all samples before and after removal of ambiguous reads and explained thoroughly how each value in the table was obtained. Regarding reads and fragments, we define each fragment as having two reads (R1 and R2). The classification into Target, Ambiguous and Unmapped reads was based on fragments, so we used that term in the table, but referring to reads has the same meaning in this context as for example an unmapped read is a read whose fragment was classified as unmapped.

We did not include the pest coverage specifically, because this cannot be calculated for any of the read counts obtained with MG-RAST as this tool is mapping to online databases where genome size is not necessarily known. What is more meaningful instead are the read counts, which are in Supplementary tables 2 and 6. Importantly as mentioned in other answers, if different taxa are differently represented in the databases this does not affect the comparison of read counts across different samples, but only the comparison of different taxa which was not used for any further analyses.

Regarding the ambiguous reads causing unreliable variants, these occur only in very few regions of the Thlaspi genome that are highly conserved in evolution or of very low complexity. In these regions reads generated from both plant or for instance aphid DNA, can map, but the ones from aphid might contain variants when mapping to the Thlaspi reference genome (L207 and L300). The reviewer is right that there is only a very small difference in average coverage when removing those ambiguous reads (~1X, S1 Table), but that is not true for those few regions where coverage changes massively when removing ambiguous reads as shown on the right side Y axes of S2 Figure. Therefore these unreliable variants are not low-frequency and therefore not removed by GATK.

> *- L215. I am not very convinced with the enrichment analyses, justified with a reference (52). For instance, how many of the predicted picks are not close to resistance genes? How was the randomisation done? At the moment, the manuscript reads rather anecdotally by describing only those picks that effectively are "close" to resistance genes. For instance, if random windows (let's say 20kb windows) are sampled along the genome, how often there are resistant genes in those random windows, and how is the random sampling compared with observed picks (windows).*

Enrichment is by definition an increase in the proportion of true positives (observed frequency: proportion of significant SNPs located close to *a priori* candidate genes) compared to the background frequency (number of all SNPs located close to *a priori* candidate genes). So the background likelihood of SNPs to fall into *a priori* candidate SNPs (i.e. the occurrence of *a priori* candidate genes in randomly sampled windows, as suggested by the reviewer) is already taken into account as the background frequency. We now explained more extensively how enrichment is calculated in the relevant methods section (L545-549), but it is an extensively used method, established in a large body of literature, so it can be found in many papers (e.g. Atwell et al. 2010, Brachi et al. 2010, Kawakatsu et al. 2016, Kerdaffrec et al. 2017, Sasaki et al. 2015-2019-2022, Galanti et al. 2022, Contreras-Garrido et al. 2024).

Although we had already calculated an upper bound for the FDR based on the *a priori* candidates, as in previous literature, we now further calculated the significance of the enrichment for the Bonferroni-corrected -log(p) threshold for Erysiphales. Calculating significance requires adopting a genome rotation scheme that preserves the LD structure of the data, as described in the previously mentioned literature (eg. Kawakatsu et al. 2016, Sasaki et al. 2022). Briefly, we calculated a null distribution of enrichments by randomly rotating the p values and *a priori* candidate status of the genetic variants within each chromosome, for 10 million permutations. We then assessed significance by comparing the observed enrichment to the null distribution. We found that the enrichment at the Bonferroni corrected -log(p) threshold is indeed significant for Erysiphales (p = 0.016). We added this to the relevant methods section and the code to the github page.

In addition, many other genes very close (few kb max) to significant SNPs were not annotated with the “defense response” GO term but still had functions relatable to it. Some examples are *CAR8*, involved in ABA signalling, *PBL7* in stomata closure and *SRF3* in cell wall building and stress response (Fig 3D). This means that our enrichment is actually most likely underestimated compared to if we had a more complete functional annotation.

> *- L247. Additional information is needed regarding sampling. It is not clear to me why methylation analyses are restricted to 20 samples, contrary to whole genome analyses.*

The sampling is best described in the original paper (on natural DNA methylation variation; Galanti et al. 2022), although the most important parts are repeated in the first chapter of the methods.
Regarding methylation analysis, they are not restricted to 20 samples. Only the DMR calling was restricted to the 20 vs. 20 samples with the most divergent values (of pest loads) to identify regions of variation. This analysis was used to subset the genome to potential regions associated with pest presence rather than thoroughly testing actual methylation variants associated with pest presence. The latter was done in the second step, EWAS, which was based on the whole dataset with the exclusions of samples with high non-conversion rate. This left 188 samples for EWAS. We added this number in the new manuscript (L251 and L571).

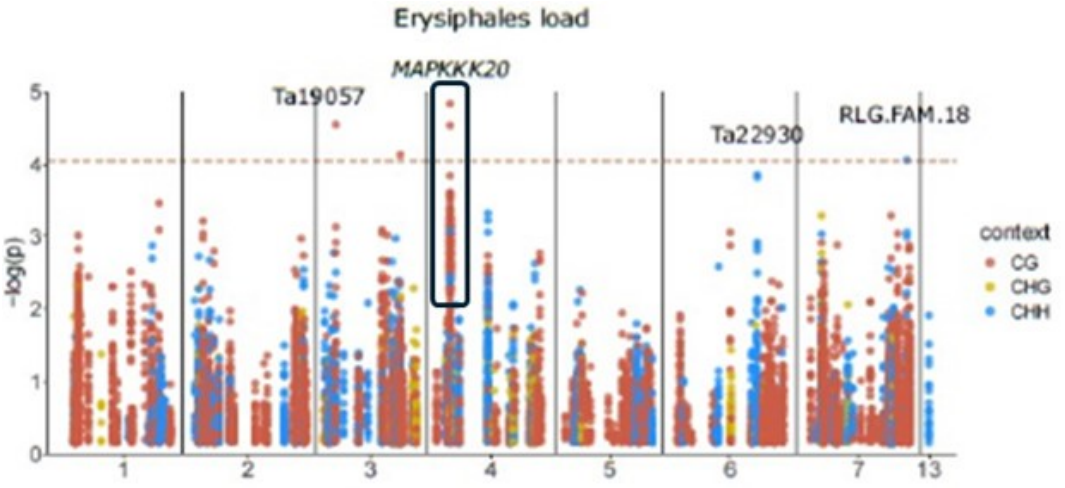
To clarify, we made a few additions to the results (L250) and methods (last two subchapters) sections, where we explain the above.

> *- No clear association with TEs: in L364: "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)." I find the whole discussion related to transposable elements, first, rather anecdotical, and second very speculative. To claim: "Overall, we found associations between pathogen load and TE methylation", I believe a more detailed analysis is needed. For instance, how often there is an association? In general, there are some rather anecdotical examples, several of which are presented as association with pathogen load on the basis of being "in proximity" to a particular region/pick. The same regions contain multiple other genes and annotations, but the authors limit the discussion to the particular gene or TE concordant with the hypothesis. This is for both the discussion and results sections.*

Here we are referring to associations in a purely statistical sense. The fact that "Overall, we found associations between pathogen load and TE methylation" is simply a conclusion drawn from Fig. 4b, without implying any causality. Some methylation variants are statistically associated with the traits (aphid or mildew loads), and whether they are true positives or causal is of course more difficult to assess.

Regarding the methylation variants associated with mildew load in proximity of *MAPKKK20*, those are the only two significant ones, located close to each other and close to many other variants that, although not significant, have low *P*-values (Author response image 2 below), so it is the most obvious association warranting further exploration. The reviewer is correct that there are other genes flanking the large DMR that covers the TEs (Fig. 4D), but the DMR is downstream of these genes, so less likely to affect their transcription.

Author response image 2.



Regarding all other associations found with *M. persicae* load, we stated that these are not really reliable due to a skewed *P*-value distribution (L269, S5B Fig), but we think that for future reference it is still worth reporting the closeby genes and TEs.

We slightly changed the wording of the passage the reviewer is citing above to make it clearer that we are only offering potential explanations for the associations we observe with TE methylation, but by no means we state that TE reactivation is surely what is happening.

> *- One conclusion in the manuscript is that DMRs have been mostly the result of hypomethylation. This is shown for instance in supplementary Figure 4. However, no general statistic is shown of methylation distribution (not only restricted to DMRs). Was the ratio methylation over de-methylation proportional along the genome? Thus the finding in DMRs is out of the genome-wide distribution? Or on the contrary, the DMRs are just a random sampling of the global distribution. The same for different annotated regions. For instance, I would expect that in general coding regions would be less methylated (not restricted to DMRs).*

Complete and exhaustive analyses of the methylomes were already published in the original manuscript (Galanti et al 2022). However, the variation among these methylomes is complex and influenced by multiple factors including genetic background and environment of origin, and talking about these things would have been beyond the scope of our paper. In this paper, we just took advantage of the existing methylome information to identify the few genomic regions that are consistently differentially methylated between samples with extreme values

of pest loads. As for the GWAS, the phenotypes are only partially associated with population structure, so the 20 samples with the lowest and the 20 with the highest pathogen loads are not e.g. all Swedish vs. all German but they are a mixture, which allowed us to correct for population structure running EWAS with a mixed model that includes a genetic distance matrix.

In this study we called DMRs between two defined groups: samples with the lowest amounts of pathogen DNA (not-infected; the “control” group) vs. samples with the highest amounts of pathogens (infected or the “treatment” group), so we could define a directionality (“hyper vs. “hypo” methylation). However, this is not the case for population DMRs called between many different combinations of populations. This is why the hyper- and hypomethylated regions found here cannot be compared to the genome-wide averages, which are influenced by other factors than the pathogens. Even with relaxed thresholds we indeed found very few DMRs associated to pathogen presence here.

Specifically about coding regions, the reviewer is correct that they are less methylated, especially because *T. arvense* has largely lost gene body methylation (Nunn et al. 2021, Galanti et al. 2022), but this is unrelated and was discussed in the original publication (Galanti et al. 2022).

> *Minor comments:- Figure 1B: it would be good to add also percentage values.*

As the figure is already tightly packed, we rather keep it simple. As the chart gives a good impression of frequencies of different kingdoms, and the frequences of several relevant groups. Also, as explained in a previous answer, comparing different taxonomic groups could be imprecise (as opposed to comparing the same group between different samples), so exact percentages seem unnecessary. If needed, the exact percentages can still be calculated from S2 Table.

> *- L159: It is not clear to me what "enemy variation" is referring to here.*

We are referring to variation in enemy densities (attack rates) in the field, that could potentially be carried over to the greenhouse to cause the patterns of infection we observed. We changed it to “variation in enemy densities” to make it more clear.

> *- L259: "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". Not clear to me what "affected samples" is referring to. Samples with lower load?*

Affected samples have a higher load of pathogen reads. We changed it to “infested” to make it more clear.

> *- L336. Figure should be Fig 3E.*

We fixed it, thanks for noticing.

ADDITIONAL CHANGES

We updated reference 43 to point to the published paper rather than the preprint.

We corrected the phenotype names in S3 Fig, to make them consistent with the rest of the manuscript and increased font size on the axes to make it more readable.

https://doi.org/10.7554/eLife.95510.2.sa0