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

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

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

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

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

Key words

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

Abstract

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

35 **Introduction**

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

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

58 Whether inherited or induced, some strategies of plants for defense against pathogens and
59 herbivores include: i) physical barriers such as reinforced cell walls, leaf protective layers or closing
60 stomata, ii) production of specialized (secondary) metabolites that reduce palatability or are toxic to
61 pests, iii) oxidative bursts, iv) the activation of signaling cascades to induce systemic responses and
62 v) RNA interference mechanisms to silence pathogen genes (18–22). In Brassicaceae, a particularly
63 important and diverse class of defense metabolites are glucosinolates, which often show local
64 adaptation driven by variation in pests and can also be induced by herbivore and pathogen attacks
65 (1,2,23).

66 Studying natural variation in plant resistance, along with associated genetic and epigenetic variation,
67 can identify genes involved in defense and their regulators, including vital genes whose function

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

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

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

97

98

99

100 **Results**

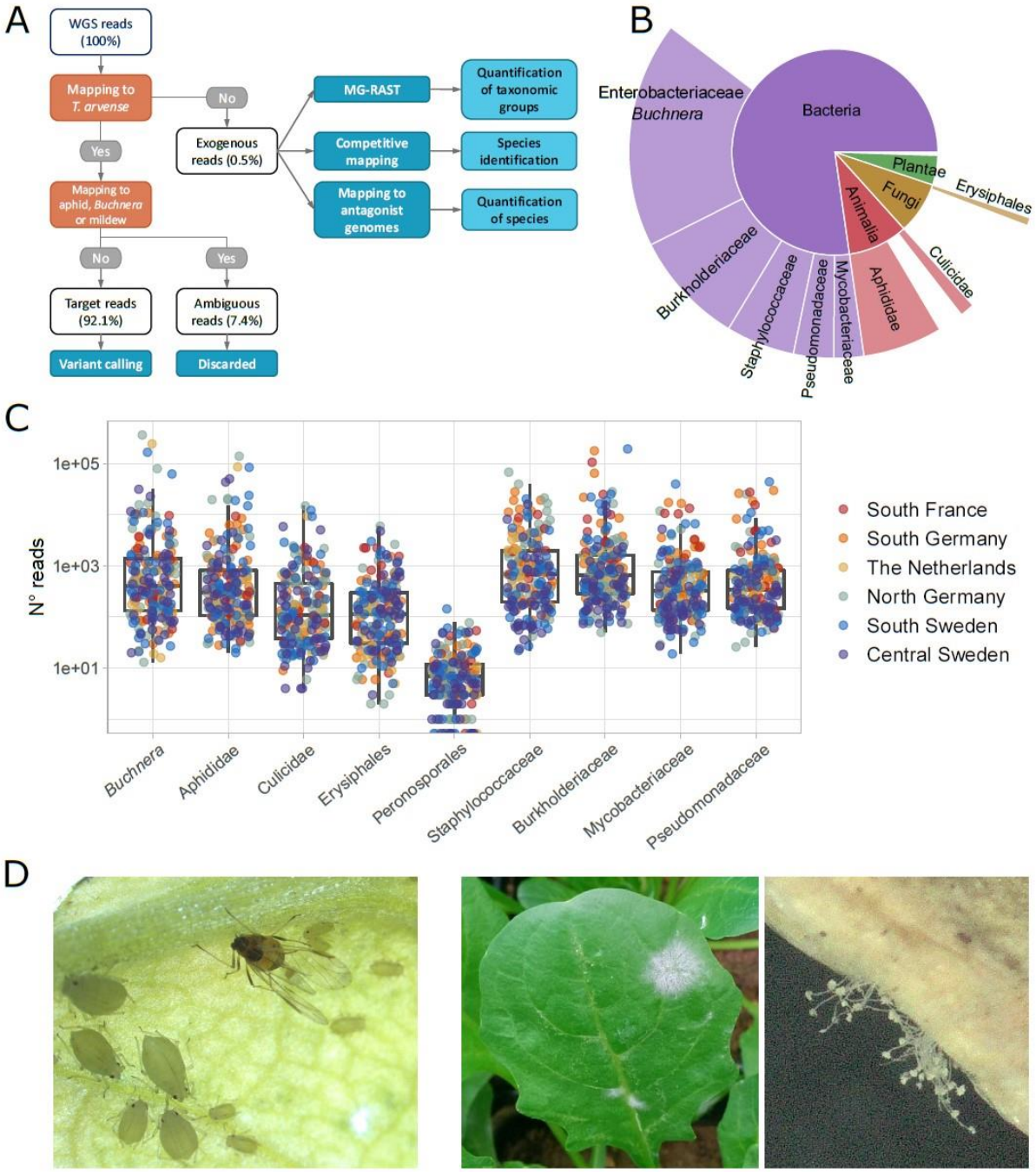
101 *Reads classification and species identification*

102 Starting from our previously published sequencing data (32), the first step of our analysis was to
103 separate the Whole Genome Sequencing (WGS) reads of each sample into the ~99.5% mapping to
104 the *Thlaspi arvense* reference genome (29) and the ~0.5% that did not, hereafter called "exogenous
105 reads" (Fig 1A). Initially, we used all mapped reads for calling variants in *Thlaspi*, but after some
106 difficulties with Genome Wide Associations (see below) we suspected that some plant reads were
107 false and mapped to the *T. arvense* genome only because of the high cross-taxa similarity of some
108 genomic regions. We therefore remapped all reads to the genomes of the aphid *Acyrthosiphon*
109 *pisum*, its endosymbiont *Buchnera aphidicola* and the powdery mildew *Blumeria graminis*, and
110 found that, on average, 7.4% of the reads mapped to both *T. arvense* and at least one of the pests.
111 We removed these ambiguous reads from our analyses and used only the *T. arvense* target reads,
112 92.1% on average, for variant calling (Fig 1A, S1 Table).

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

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

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



138

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

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

144

145 *Exogenous read counts are a heritable Thlaspi phenotype*

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

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

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

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

Taxonomic group	Data type	Population differences (R^2 and *P*-value)	SNP-based heritability
Myzus persicae	Mapping to reference genome	0.245 (*P* = 0.029)	0.190 (0.055-0.488)
Buchnera aphidicola	Mapping to reference genome	0.256 (*P* = 0.016)	0.169 (0.042-0.490)
Buchnera	MG-RAST - genus	0.223 (*P* = 0.090)	0.115 (0.016-0.505)
Aphididae	MG-RAST - family	0.226 (*P* = 0.082)	0.189 (0.052-0.496)
Culicidae	MG-RAST - family	0.166 (*P* = 0.519)	0.183 (0.055-0.465)

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

167

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

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

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

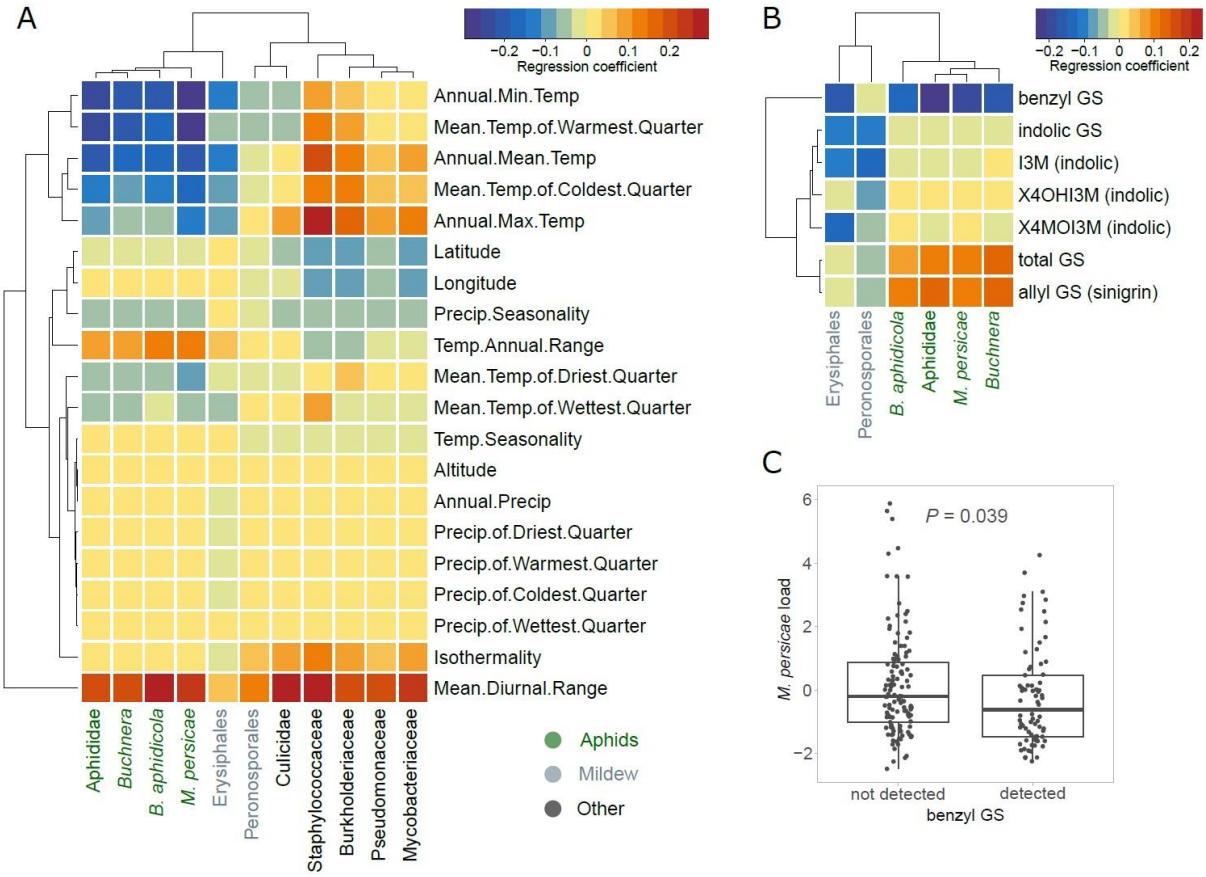


Fig 2. Relationships between climates of origin or glucosinolate levels of plants and the exogenous reads loads. (A) Correlations with bioclimatic variables. (B) Correlations with baseline glucosinolate (GS) levels measured in the same pennycress lines in another experiment. All correlations in (A) and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

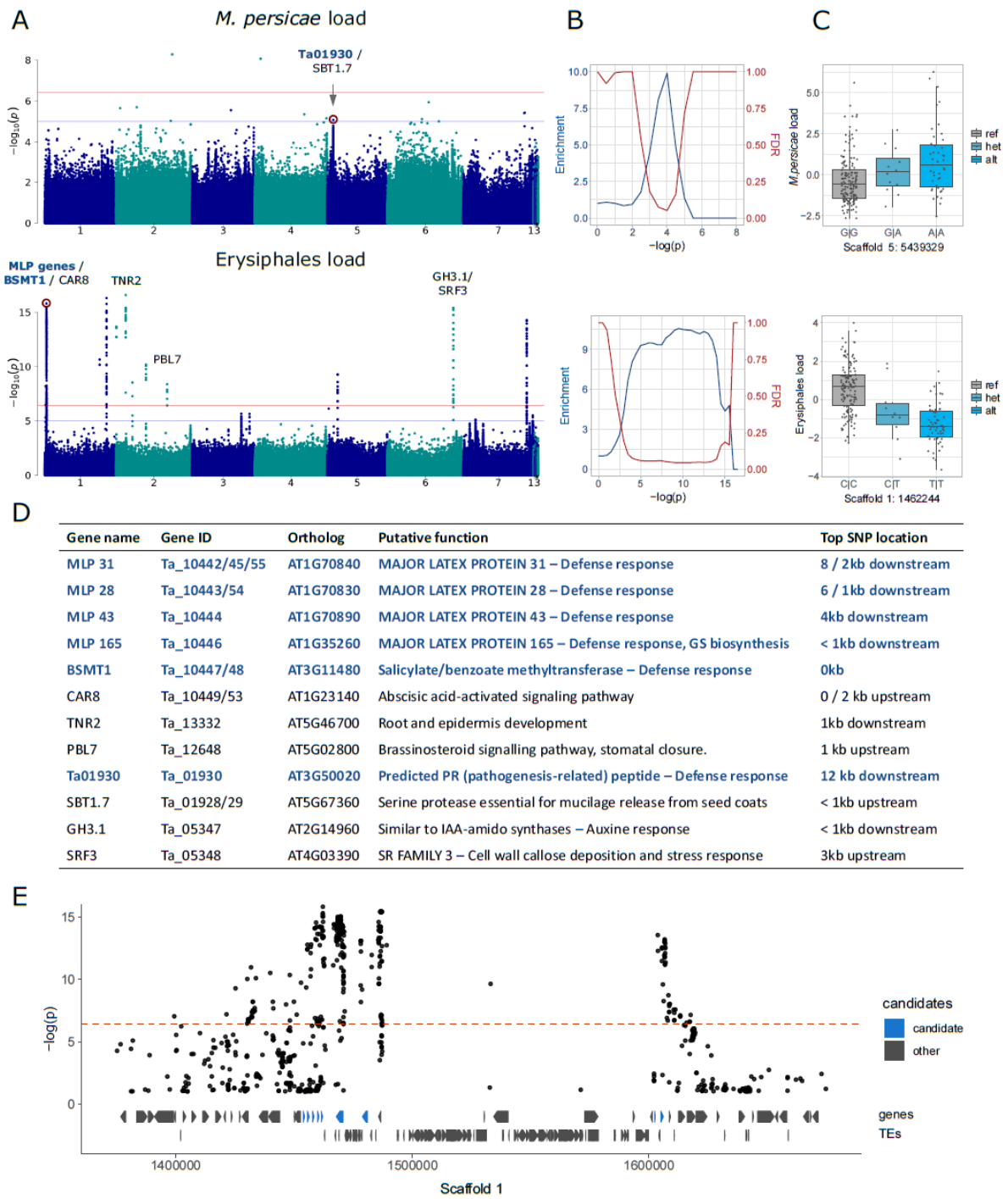
GWA identifies peaks near defense genes

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

210 which eliminated the observed massive GWA peaks, indicating that they had indeed reflected false
211 associations (S2 Fig).

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

227



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

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

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

239

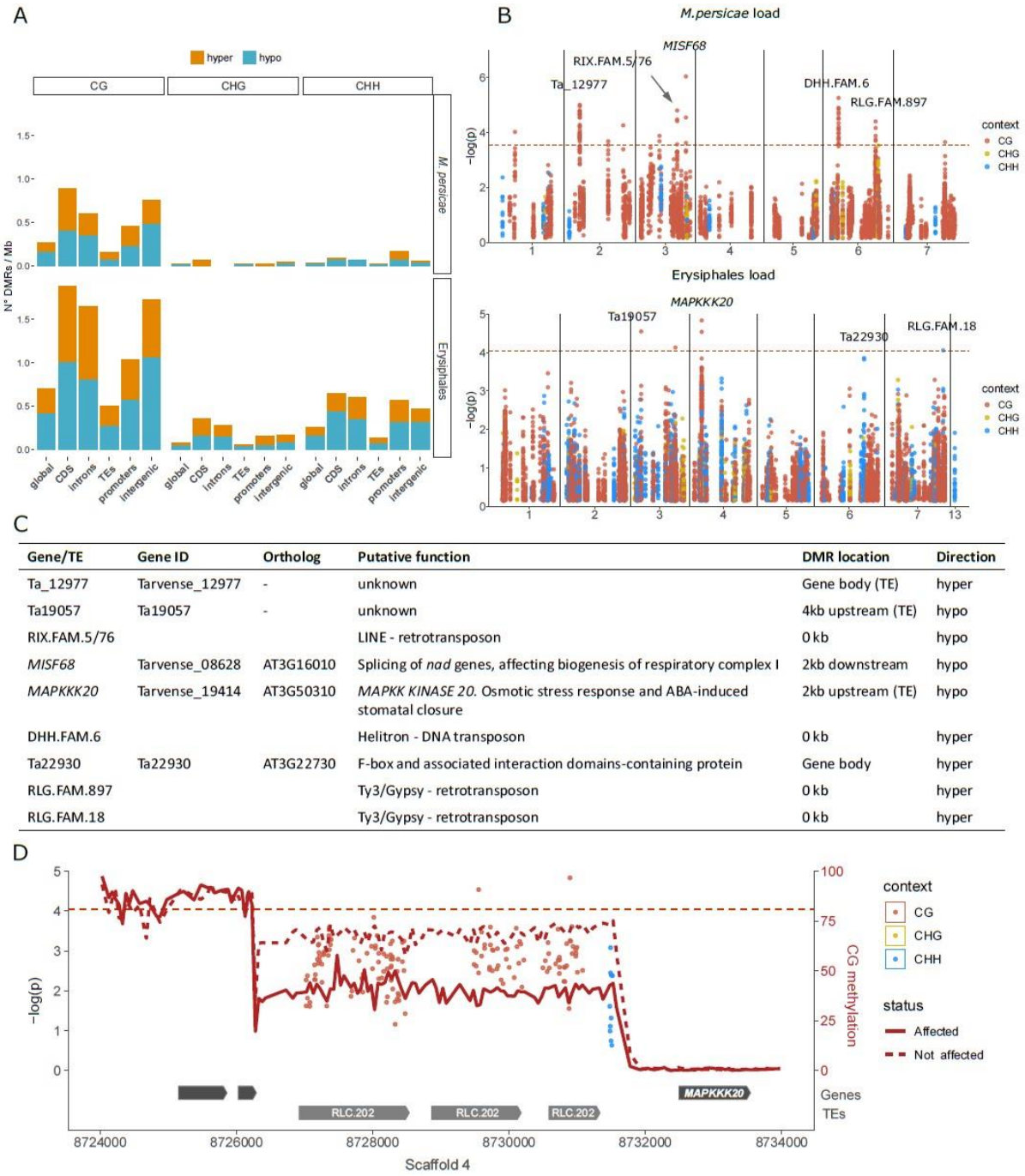
240 *Aphid and mildew loads correlate with differential methylation at genes and transposons*

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

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

262 For a more detailed investigation, we turned to EWA, leveraging the power of the entire *Thlaspi*
263 collection. We tested for associations between *M. persicae* or Erysiphales loads and the methylation
264 at individual cytosines located within the DMRs. As in GWA, we corrected for population structure
265 using an IBS matrix. For both types of pest loads, we found associations in the proximity of genes
266 and especially within TEs, but no genomic feature was particularly enriched for low *P*-value
267 associations (S5A Fig). *M. persicae* load was associated with methylation at several genomic
268 locations, especially TEs (Fig 4B), but these associations had strongly inflated *P*-values (S5B Fig). For
269 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*.



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)

280 Candidate genes and TEs marked in panel B, their putative functions, genomic locations of associated DMRs, and whether
281 affected samples were hyper- or hypomethylated. (D) Zoomed-in Manhattan plot for Erysiphales load around the peak in
282 Scaffold 4, with gene and TE models given below. The CG methylation in the 20 most and least affected samples was
283 calculated over 50 bp bins (see S5C Fig for methylation in other contexts).

284

285 **Discussion**

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

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

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

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

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

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

An important general insight from our GWA analyses was the frequent ambiguity of reads that mapped to both pest and host plant genome. Such ambiguous reads generated false variants only present in samples with pest DNA, which resulted in highly significant false associations, and it was

346 therefore important to remove these reads before variant calling. Another potential reason for
347 sequence similarity between host and pathogens could be defense mechanisms such as RNA
348 interference. If *T. arvense* produces small or micro RNAs to silence pathogen genes, this would
349 originate from genomic regions of high similarity between host and pathogen, and thus reflect a true
350 association. However, a BLAST (57) of the region in which the suspicious associations occurred did
351 not reveal any similarity to aphid or mildew genes, but instead to the highly conserved ribosomal
352 RNA coding regions. While genetic variants are generally inherited from parents and thus reflect
353 evolutionary processes, DNA methylation variants can be heritable but can also reflect plastic
354 responses to environmental stresses like herbivores or pathogens. Our data do not allow to
355 confidently distinguish between these two sources of DNA methylation variation, and thus should be
356 interpreted with caution, especially with regard to the directionality of associations. A beneficial
357 DNA methylation variant is expected to be associated with lower pathogen load when already
358 present before pathogen arrival, but with higher pathogen load when plastically induced by
359 pathogens during the experiment. For both *M. persicae* and Erysiphales, the majority of DMRs were
360 hypomethylated in affected samples, which is in accordance with the loss of methylation observed in
361 *A. thaliana* and *T. arvense* upon aphid feeding, and in diploid wheat upon powdery mildew infection
362 (8,9,31), but we also detected hypermethylation at several loci. *M. persicae* load was associated with
363 differential methylation at only few genes but several TEs, which is in accordance with the aphid or
364 stress-induced TE reactivation observed in *A. thaliana* (9,14). Erysiphales load was associated with
365 hypomethylated Copia TEs upstream of *MAPKKK20*, a gene involved in ABA-mediated signaling and
366 stomatal closure. Since stomatal closure is a known defense mechanism to block pathogen access
367 (21), it is tempting to conclude that hypomethylation of the *MAPKKK20* promoter might induce its
368 overexpression and consequent stomatal closure, thereby preventing mildew access to the leaf
369 blade. Overall, we found associations between pathogen load and TE methylation that could act
370 both in *cis* (eg. Copia TE methylation in *MAPKKK20* promoter) and in *trans*, possibly through
371 transposon reactivation (eg. LINE, Helitron and Ty3/Gypsi TEs isolated from genes). Although we
372 cannot confidently distinguish inherited versus induced DNA methylation variants, to our knowledge
373 this is the first coupled GWA - EWA analysis conducted on a large natural plant collection.

374 In summary, our study offers first insights into the defense mechanisms of *Thlaspi arvense*, including
375 candidate genes and alleles which may be of interest for breeding efforts in this novel biofuel and
376 cover crop. It also provides a proof of principle that exogenous reads from large sequencing efforts,
377 usually discarded if not mapping to the target genome, can be leveraged to extract additional
378 information about important biotic interactions of the target species, including its antagonists and
379 microbiome components. We combined this approach with data from a common environment

380 experiment to show that pest and microbiome load were geographically structured, as expected
381 from locally adapted traits, and associated with both genetic and DNA methylation variants. In
382 principle, our approach can be applied to many other designs. For example, field-collected samples
383 could be used to quantify geographic pathogen distributions. With the decreasing cost of sequencing
384 and increasing large-scale and single-species sequencing projects e.g. (37–43), the number of data-
385 sets suitable for such analyses is set to rapidly increase in the near future.

386

387 **Materials and Methods**

388 *Plant growth and sequencing*

389 The WGS data used in this study were already published in Galanti et al. (2022) (32). Please refer to
390 this publication for details on data generation and methods. Briefly, we collected 207 *Thlaspi*
391 *arvense* accessions from 36 European populations in July 2018, and we grew their offspring in a
392 glasshouse at University of Tübingen (48°32'21.3"N, 9°02'04.2"E) between August and October 2019.
393 The glasshouse was located in biodiverse surroundings, and insects and pests could enter when the
394 windows opened for temperature regulation. A few weeks after germination, we noticed aphid and
395 mildew infestations. After 46 d we sampled the 3rd or 4th true leaf of each plant and snap-froze it in
396 liquid nitrogen. We extracted DNA using the DNeasy Plant Mini Kit (Qiagen, Hilden, DE), sonicated
397 (Covaris) 300 ng of genomic DNA and used the NEBNext Ultra II DNA Library Prep Kit for Illumina
398 (New England Biolabs) to prepare the libraries. Half way through the protocol we split the DNA into
399 1/3 for genomic libraries and 2/3 for bisulfite libraries. For the bisulfite conversion we used the EZ-
400 96 DNA Methylation-Gold MagPrep (ZYMO) kit. We sequenced paired-end for 150 cycles using
401 Illumina NovaSeq 6000 (Illumina, San Diego, CA) for genomic libraries and HiSeq X Ten (Illumina, San
402 Diego, CA) for bisulfite libraries.

403

404 *Reads mapping and classification*

405 Upon demultiplexing the raw reads, we used cutadapt (58) for quality (minimum quality of 20) and
406 adaptor trimming, excluding reads shorter than 35 bp. We used FastQC and MultiQC (59,60) to
407 estimate the duplication rate, and calculated the total deduplicated reads, which we later used for
408 correcting the number of exogenous reads. We then classified the reads based on their mapping
409 behavior. First we aligned reads to the *T. arvense* reference genome (29) with BWA-MEM v0.7.17
410 (61), excluding multimapping reads (-c 1) and marking duplicates with MarkDuplicatesSpark (62,63).
411 We then mapped all samples again (61) to the three putative exogenous genomes of pea aphid

412 (*Acrophyson pisum*), the aphid symbiont *Buchnera aphidicola* and powdery mildew (*Blumeria*
413 *graminis*), using available resources (www.ncbi.nlm.nih.gov/assembly/GCF_005508785.2,
414 www.ncbi.nlm.nih.gov/assembly/GCA_001939165.1) (64). After this, we used a custom script to
415 collect all read IDs within a sample mapping to any of the three exogenous genomes, and removed
416 any of these reads from the *T. arvense* alignment bam files. We thus removed all ambiguous reads
417 before proceeding with variant calling. To compare coverage of specific regions with and without
418 ambiguous reads, we used samtools bedcov (65). The numbers of reads classified by their mapping
419 behaviour are reported in S1 Table.

420

421 *Variant calling*

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

431

432 *Identification and classification of exogenous reads*

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

439 We ran MG-RAST mostly with default parameters, without assembled reads, excluding dereplicated
440 sequences, and dynamically trimming reads with a minimum Phred score of 15 in more than 5
441 consecutive bases. We set the "sequence screening" to *A. thaliana*, the closest relative of *T. arvense*
442 available. We used two different approaches to extract read counts. First we classified all reads up to
443 family level using the web-based Analysis tool from MG-RAST. We used RefSeq as query annotation

444 database and filtered reads classified with low confidence using default settings: e-value 5, 60 %-
445 identity, length 15 and min.abundance of 1 (S2 Table). Out of the hundreds of taxonomic groups
446 identified by MG-RAST, we selected only a small subset for follow-up analyses, based on their
447 biological relevance, our visual observations and/or abundance: Aphididae, Culicidae,
448 Peronosporales, Staphylocaccaceae, Burkholderiaceae, Mycobacteriaceae and Pseudomonadaceae
449 (Table 1). Additionally, we used a custom Python script to download individual "taxonomy" or
450 "sequence_breakdown" results from MG-RAST API (68) and extracted the counts of the genus
451 *Buchnera,* including bacterial symbionts of many aphid species, and of the order Erysiphales, to
452 quantify the observed mildew infection (Table 1). All the code for extracting counts for all families or
453 specific taxonomic groups are available on GitHub (https://github.com/junhee-jung/MG-RAST-read-
454 counter).

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

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

465

466 *Exogenous reads heritability and species identification*

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

473 Even though MG-RAST classifies reads based on all taxonomic ranks, the accuracy of species
474 identification of course strongly depends on the sequences available in the query databases. MG-
475 RAST assigned our aphid reads to *A. pisum*, but this did not fit with our visual observations and with

476 the poor performance of this species on Brassicaceae (70). We therefore selected three plausible
477 aphid species and test which of these had mostly likely attacked our experiment. In addition to *A.*
478 *pisum*, we tested two other aphid species commonly attacking Brassicaceae: *Brevicoryne brassicae*
479 and *Myzus persicae*. While not all three species have reference genomes available, all mitochondrial
480 genomes are available on NCBI (71) under accession numbers MN232006, NC_011594 and
481 NC_056270. We downloaded these sequences, aligned them to each other (72), removed INDELs to
482 retain only SNPs and combined them into a single pseudo-reference fasta file (S3 Table). We then
483 mapped the exogenous reads from 40 randomly selected samples to this pseudo-reference, allowing
484 for unique mappings only and counted the reads mapping to either of the three aphid species. We
485 used the same approach for mildew except that we included only two possible species: *Blumeria*
486 *graminis*, as suggested by MG-RAST, and *Erysiphe cruciferarum* which is known to attack
487 Brassicaceae but was not in the MG-RAST query database and seemed plausible from visual
488 inspection (Fig 2B). For the mildew pseudo-reference (S4 Table) we used the Internal Transcribed
489 Spacer (ITS), which is publicly available for both species on NCBI (71) under accession numbers
490 MT644878 and AF031283.

491

492 *Quantification of glucosinolates*

493 Using seed material collected from the sequenced plants, we conducted a follow-up experiment to
494 estimate the glucosinolates (GS) content of all 207 lines in the absence of pathogens. Briefly, we
495 sowed the seeds in petri dishes, stratified them at 4°C in the dark for two weeks and transplanted
496 the germinated seedlings to individual 9 x 9 cm pots. We grew the plants in a growth chamber with a
497 14/10 h light/dark cycle at 21/17 °C and a relative humidity of ~45%. Two weeks after germination
498 the plants were vernalized at 4°C for two more weeks in order to minimize phenological and
499 developmental differences between winter and summer annuals. Ten days after vernalization, we
500 collected the 3rd or 4th true leaf and snap-froze it in liquid nitrogen. After freeze drying, we weighed
501 all samples and extracted the material threefold in 80% methanol, adding *p*-hydroxybenzl
502 glucosinolate (Phytoplan, Heidelberg, Germany) as internal standard. After centrifugation, we
503 applied the supernatants onto ion-exchange columns with diethylaminoethyl (DEAE) Sephadex A25
504 (Sigma Aldrich, St.Louis, MO, USA) in 0.5 M acetic acid buffer, pH 5. We added purified sulfatase,
505 converting glucosinolates to desulfo glucosinolates. After one day, we eluted desulfo glucosinolates
506 in water and analyzed them on a HPLC coupled to a DAD detector (HPLC-1200 Series, Agilent
507 Trechnologies, Inc., Santa Clara, CA, USA) equipped with a Supelcosil LC 18 column (3 µm, 150×3
508 mm, Supelco, Bellefonte, PA, USA). We analysed the samples with a gradient from water to

509 methanol starting at 5% methanol for 6 min and then increased from 5 to 95% within 13 min with a
510 hold at 95% for 2 min, followed by a column equilibration cycle. We identified different
511 glucosinolates based on their retention times and UV spectra in comparison to respective standards
512 and an in-house database. We integrated peaks at 229 nm and calculated respective glucosinolate
513 concentrations in relation to the internal standard and sample dry mass, using response factors as
514 reported by Agerbirk et al. (2015) (73).

515

516 *Drivers of exogenous reads variation*

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

524

525 *GWA analysis*

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

533 To validate our results and test for overlap with existing gene functional annotations, we performed
534 an enrichment analysis of variants neighboring *a priori* candidate genes as described in Atwell et al.
535 (2010) (52). Briefly, we attributed *a priori* candidate status to all variants located within 20 kb from
536 orthologs (78) of *A. thaliana* genes annotated with the GO term "defense response" (GO:0006952),
537 including nine genes similar to *AtBSMT1* (S8 Table). We then calculated the enrichment of these
538 variants compared to the background frequency and an upper bound for the FDR for increasing -
539 log(p) thresholds (32,52). The code for these analyses is available on https://github.com/Dario-
540 Galanti/multipheno_GWAS/tree/main/gemmaGWAS.

541

542 *Methylation and DMR calling*

543 For the methylome analyses we used the EpiDiverse toolkit (79), specifically designed for large
544 WGBS datasets. We used the WGBS pipeline (https://github.com/EpiDiverse/wgbs) for read
545 mapping and methylation calling, retained only uniquely-mapping reads longer than 30 bp, and
546 obtained individual-sample bedGraph files for each sequence context. We then called DMRs using
547 the DMR pipeline (79), with a minimum coverage of 4x. We compared the 20 samples with the most
548 and the least *M. periscae* and Eriysiphales loads, resulting in two sets of DMRs for each sequence
549 context. Since this was only the first step of our methylation analysis, meant to identify potential
550 regions of interest, we retained all DMRs with a FDR < 20%. To understand the genomic preferences
551 of DMRs, we intersected them with genomic features and calculated their densities in each by
552 dividing their number by the total Mb covered by each genomic feature in the genome.

553

554 *EWA analysis*

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

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

573 in genomic features is available on https://github.com/Dario-
574 Galanti/EWAS/tree/main/gemmaEWAS.

575

576 **Supporting Information**

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

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

586 **S3 Fig. GWA results for all exogenous reads.** Manhattan plots, enrichment of *a priori* candidate
587 variants and QQplots for all exogenous reads. The genome-wide significance (horizontal red lines),
588 was calculated based on unlinked variants (53), the blue line corresponds to -log(p) =5. The top
589 variants are labeled with the neighboring genes potentially affecting *exogenous reads* load.

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

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

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

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

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

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

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

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

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

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

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

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

622

623 **Code and data availability statement**

624 The seed material from the sequenced lines is available at the Nottingham Arabidopsis Stock Centre
625 (NASC) under stock numbers N950001 to 950204. Genomic and bisulfite sequencing raw data are
626 available on the ENA Sequence Read Archive (www.ebi.ac.uk/ena/) under study accession number
627 PRJEB50950. The reference genome and annotations were previously published by Nunn et al.
628 (2022) (29). GWA and EWA results in a format compatible with the Integrative Genomics Viewer
629 (https://www.igv.org/) are available on Zenodo (https://zenodo.org/records/10011535).

630 All code used in this study is available and documented on GitHub. The scripts for variant calling,
631 filtering and imputation are on https://github.com/Dario-Galanti/BinAC_varcalling, and the scripts
632 for the classification of sequencing reads and MG-RAST analysis are in https://github.com/Dario-
633 Galanti/Exoreads_treasure and https://github.com/junhee-jung/MG-RAST-read-counter

respectively. The pipelines for methylation and DMR calling from WGBS data can be found on the EpiDiverse GitHub (https://github.com/EpiDiverse). The workflow for downstream analysis of methylation data is on https://github.com/Dario-Galanti/WGBS_downstream/tree/main/WGBS_simpleworkflow. Finally, the scripts for running GWA and EWA analysis are on https://github.com/Dario-Galanti/multipheno_GWAS/tree/main/gemmaGWAS and https://github.com/Dario-Galanti/EWAS/tree/main/gemmaEWAS respectively.

Funding

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

Acknowledgments

We thank the EpiDiverse network for its amazing support and discussions, in particular Adrián Contreras-Garrido, Bárbara Díez Rodríguez, Iris Sammarco, Adam Nunn, Daniela Ramos and Anupoma Troyee for their close collaboration. We also thank Frank Reis for his feedback and expert tips for the project, and Cecilia Heyworth for proofreading the manuscript. We thank Peter Stadler at the University of Leipzig and David Langenberger from ecSeq, which helped with computing and hosted the EpiDiverse servers. The BinAC cluster is managed by the High Performance and Cloud Computing Group at the Zentrum für Datenverarbeitung of the University of Tübingen, and supported by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant no INST 37/935-1 FUGG.

Author Contributions

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

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

664 Caroline Müller: Investigation, Writing – Review & Editing.

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

667

668 **References**

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

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

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

678 4. Corwin JA, Copeland D, Feusier J, Subedy A, Eshbaugh R, Palmer C, et al. The Quantitative Basis of the
679 Arabidopsis Innate Immune System to Endemic Pathogens Depends on Pathogen Genetics. PLOS Genet.
680 2016 Feb 11; 12(2):e1005789.
681 https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1005789

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

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

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

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

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

699 10. Kinoshita T, Seki M. Epigenetic Memory for Stress Response and Adaptation in Plants. Plant Cell Physiol.
700 2014 Nov 1; 55(11):1859–63. https://academic.oup.com/pcp/article/55/11/1859/2756009

701 11. Espinas NA, Saze H, Saijo Y. Epigenetic Control of Defense Signaling and Priming in Plants. Front Plant Sci.
702 2016 Aug 11; 7:1201. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4980392/

703 12. Lämke J, Bäurle I. Epigenetic and chromatin-based mechanisms in environmental stress adaptation and
704 stress memory in plants. Genome Biol. 2017 Jun 27; 18(1):124. https://doi.org/10.1186/s13059-017-1263-
705 6

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

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

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

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

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

718 18. Wojtaszek P. Oxidative burst: an early plant response to pathogen infection. Biochem J. 1997 Mar 15;
719 322(Pt 3):681–92. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1218243/

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

723 20. Kant MR, Jonckheere W, Knegt B, Lemos F, Liu J, Schimmel BCJ, et al. Mechanisms and ecological
724 consequences of plant defence induction and suppression in herbivore communities. Ann Bot. 2015 Jun 1;
725 115(7):1015–51. https://doi.org/10.1093/aob/mcv054

726 21. Melotto M, Zhang L, Oblessuc PR, He SY. Stomatal Defense a Decade Later. Plant Physiol . 2017 Jun 1;
727 174(2):561–71. https://doi.org/10.1104/pp.16.01853

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

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

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

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

739 26. Roman-Reyna V, Pinili D, Borja FN, Quibod IL, Groen SC, Alexandrov N, et al. Characterization of the Leaf
740 Microbiome from Whole-Genome Sequencing Data of the 3000 Rice Genomes Project. Rice. 2020 Oct 9;
741 13(1):72. https://doi.org/10.1186/s12284-020-00432-1

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

745 28. Geng Y, Guan Y, Qiong L, Lu S, An M, Crabbe MJC, et al. Genomic analysis of field pennycress (Thlaspi
746 arvense) provides insights into mechanisms of adaptation to high elevation. BMC Biol. 2021 Jul 22;
747 19(1):143. https://doi.org/10.1186/s12915-021-01079-0

748 29. Nunn A, Rodríguez-Arévalo I, Tandukar Z, Frels K, Contreras-Garrido A, Carbonell-Bejerano P, et al.
749 Chromosome-level Thlaspi arvense genome provides new tools for translational research and for a newly
750 domesticated cash cover crop of the cooler climates. Plant Biotechnol J. 2022.
751 https://onlinelibrary.wiley.com/doi/abs/10.1111/pbi.13775

752 30. Hu Y, Wu X, Jin G, Peng J, Leng R, Li L, et al. Rapid Genome Evolution and Adaptation of Thlaspi arvense
753 Mediated by Recurrent RNA-Based and Tandem Gene Duplications. Front Plant Sci. 2022; 12.
754 https://www.frontiersin.org/articles/10.3389/fpls.2021.772655

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

758 32. Galanti D, Ramos-Cruz D, Nunn A, Rodríguez-Arévalo I, Scheepens JF, Becker C, et al. Genetic and
759 environmental drivers of large-scale epigenetic variation in Thlaspi arvense. PLOS Genet. 2022 ott;
760 18(10):e1010452. https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1010452

761 33. Dorn KM, Fankhauser JD, Wyse DL, Marks MD. A draft genome of field pennycress (Thlaspi arvense)
762 provides tools for the domestication of a new winter biofuel crop. DNA Res. 2015 Apr 1; 22(2):121–31.
763 https://academic.oup.com/dnaresearch/article/22/2/121/332636

764 34. Frels K, Chopra R, Dorn KM, Wyse DL, Marks MD, Anderson JA. Genetic Diversity of Field Pennycress
765 (Thlaspi arvense) Reveals Untapped Variability and Paths Toward Selection for Domestication. Agronomy.
766 2019 Jun; 9(6):302. https://www.mdpi.com/2073-4395/9/6/302

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

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

773 37. Kajiya-Kanegae H, Nagasaki H, Kaga A, Hirano K, Ogiso-Tanaka E, Matsuoka M, et al. Whole-genome
774 sequence diversity and association analysis of 198 soybean accessions in mini-core collections. DNA Res.
775 2021 Feb 1; 28(1):dsaa032. https://doi.org/10.1093/dnares/dsaa032

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

62. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Angel G del, Levy-Moonshine A, et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Curr Protoc

845 Bioinforma. 2013 [cited 2021 Jun 11]; 43(1):11.10.1-11.10.33.
846 https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/0471250953.bi1110s43

847 63. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Auwera GAV der, et al. Scaling accurate
848 genetic variant discovery to tens of thousands of samples. bioRxiv. 2018 Jul 24 [cited 2021 Jun
849 11];201178. https://www.biorxiv.org/content/10.1101/201178v3

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

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

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

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

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

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

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

871 71. National Center for Biotechnology Information (NCBI). Bethesda (MD): National Library of Medicine (US).
872 1988. NCBI. https://www.ncbi.nlm.nih.gov/

873 72. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality
874 protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011 Jan [cited 2023 Oct 12];
875 7(1):539. https://www.embopress.org/doi/full/10.1038/msb.2011.75

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

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

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

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

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

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

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

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