Drought selection on *Arabidopsis* populations and their microbiomes

Talia L. Karasov[1,2], Manuela Neumann[2,3], Gautam Shirsekar[2], Grey Monroe[2,4], PATHODOPSIS Team, Detlef Weigel[2]*, Rebecca Schwab[2]

[1]School of Biological Sciences, University of Utah, Salt Lake City, Utah, USA
[2]Department of Molecular Biology, Max Planck Institute for Biology Tübingen, Tübingen, Germany
[3]Current address: Robert Bosch GmbH, 71272 Renningen, Germany
[4]Current address: Department of Plant Sciences, University of California Davis, Davis, CA, USA
*Corresponding author: weigel@tue. mpg. de

Formatting guide: https://www. nature. com/nature/for-authors/formatting-guide

PATHODOPSIS Team (in alphabetical order): Cristina Barragan, Ilja Bezrukov, Alba González Hernando, Julia Hildebrandt, Sonja Kersten, Patricia Lang, Sergio Latorre, Miriam Lucke, Anette Habring, Claudia Friedemann, Fiona Paul, Derek Lundberg, Ulrich Lutz, Fernando Rabanal, Julian Regalado, Thanvi Srikant, Bridgit Waithaka, Anjar Wibowo, Wei Yuan

Summary

Microbes affect plant health, stress tolerance[1] and life history[2]. In different regions of the globe, plants are colonized by distinct pathogenic and commensal microbiomes, but the factors driving their geographic variation are largely unknown[3]. We identified and measured the core leaf microbiome of *Arabidopsis thaliana* in its native range, from almost 300 populations across Europe. Comparing the distribution of the approximately 500 major bacterial phylotypes, we discovered marked, geography-dependent differences in microbiome composition within *A. thaliana* and between *A. thaliana* and other Brassicaceae, with two distinct microbiome types segregating along a latitudinal gradient. The differences in microbiome composition mirror the spatial genetics of *A. thaliana*, with 52-68% of variance in the first two principal coordinates of microbiome type explained by host genotype. Microbiome composition is best predicted by drought-associated metrics that are well known to be a major selective agent on *A. thaliana* populations. The reproducible and predictable associations between specific microbes and water availability raise the possibility that drought not only directly shapes genetic variation in *A. thaliana*, but does so also indirectly through its effects on the leaf microbiome.

Results

The crucifer *Arabidopsis thaliana* is an annual species that today can be found on at least six continents[4], residing in environments with disparate temperatures, water availability, salinity and surrounding ecosystems. These distinct environments have imposed strong selection on *A. thaliana*, leaving environment-associated signatures of selection throughout its genome[5]. While spatial differences in abiotic factors are well-appreciated, differences in the resident microbiota are also likely to influence plant fitness and the course of local adaptation, and concrete evidence is beginning to emerge that there is regional variation in the microbiota that is associated with *A. thaliana*[6]. A recent survey of root microbiomes[3] found regional differentiation between populations, with microbes in the roots reflecting the composition of microbes in the soil. Host location was similarly significantly correlated with both root- and leaf-associated microbial composition of another Brassicaceae species, *Boechera stricta*[7].
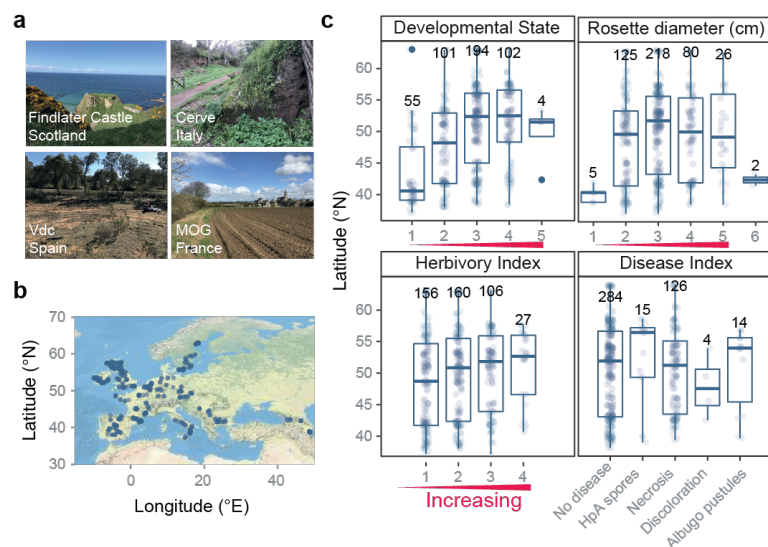
Figure 1 | Representative sampling of *A. thaliana* phyllosphere microbiomes across Europe. *Arabidopsis thaliana* plants were collected from distinct ecosystems across Europe. **a, b**. **c,** Images taken at each site enabled the assessment of plant health and development. The *x*-axis represents qualitative values, as described in Methods, except for rosette diameter, which is classified in intervals of 0-1 cm (1), 1-2 cm (2), 2-3 cm (3) etc. Developmental state and herbivory index are ordered in increasing progression. Disease index corresponds to different macroscopic disease symptoms. The central horizontal line in each box indicates the median, the bounds of the box the upper and lower quartiles. The number of individuals in each group are indicated above the boxes.

These and other studies have reproducibly demonstrated that plant populations vary in the presence and relative abundance of microbes[8], but little is known about what drives these differences. There are many possible factors and it is often infeasible to tease apart proximal from distal causes. Host genetics can influence microbiome composition[7,9], and spatial structuring of host genetic variation may in turn result in spatial structuring of the resident microbiome, but the two might also be independently affected by physical distance, including abiotic factors that vary geographically[3,7]. Numerous abiotic variables including temperature, rainfall, humidity, sun exposure and soil composition could impact microbial abundance in plants. For example, Thiergart and colleagues[3] found pH to be a significant predictor of *A. thaliana* rhizosphere bacterial composition, consistent with pH as a major explanatory variable of soil bacterial composition[10].

Because previous studies have typically been limited in the number of populations[3] or the geographic range surveyed[6], it has been difficult to disentangle these multifarious influences. In general, we still have a poor understanding of which environmental factors are the best predictors (and likely the causative agents) of differences in plant-associated microbial communities. We also do not know how much of their composition is explained by the relative contributions of host genetics versus environmental variables.

In this study, through continental-scale assessment of the bacteria that colonize the leaves of *A. thaliana* in Europe, we identify environmental and host genetic factors that are strongly associated with distinct bacterial microbiomes. We further determine the environmental variables that best predict microbiome composition, we test the relative contributions of host genetics and abiotic factors to these predictable patterns, and we characterize the association between soil microbiomes, host genetics and companion species and the phyllosphere microbiota.

From February to May 2018, we visited 267 *A. thaliana* populations across Europe at the end of their vegetative growth period and onset of flowering[2] (Figure 1a). In each population we collected whole rosettes from two *A. thaliana* plants, a neighboring Brassicaceae (primarily *Capsella bursa pastoris*), if present, and two soil samples. We evaluated *A. thaliana* life history traits including developmental state and rosette diameter (Figure 1c, Figure S1), and extracted information on climate variables for the collection sites[11]. We assessed the microbial composition of the leaf and soil samples by sequencing the v3-v4 region of the 16S rRNA locus (hereafter rDNA) and classifying distinct 16S sequences as alternative sequence variants (ASV) using DADA2 (ref. [12]). Each ASV was considered a distinct bacterial lineage which we term here a phylotype. Host genetics and absolute microbial load were assessed by shotgun sequencing plant tissue which generate metagenomic sequences of host and microbial genomes[13].

Phyllosphere composition is distinct from the soil and is host species-specific

There is considerable debate as to the origin of the microbes that colonize plant tissues, although soil often has a measurable influence[3,14,15]. A study across 17 European *A. thaliana* populations[3] found significant differentiation between root and non-root-associated microbes, but no significant differences between *A. thaliana* and neighboring grasses[3]. Intra-species comparisons in a common-garden experiment had suggested that host genetics can explain about 10% of the variance in composition of the *A. thaliana* leaf bacterial microbiome[9]. At the basis of these comparisons is the question of how much host influence there is in microbiome assembly, either because of active recruitment of specific microbes from the environment, or because of differential ability of microbes to grow in and on plant hosts.
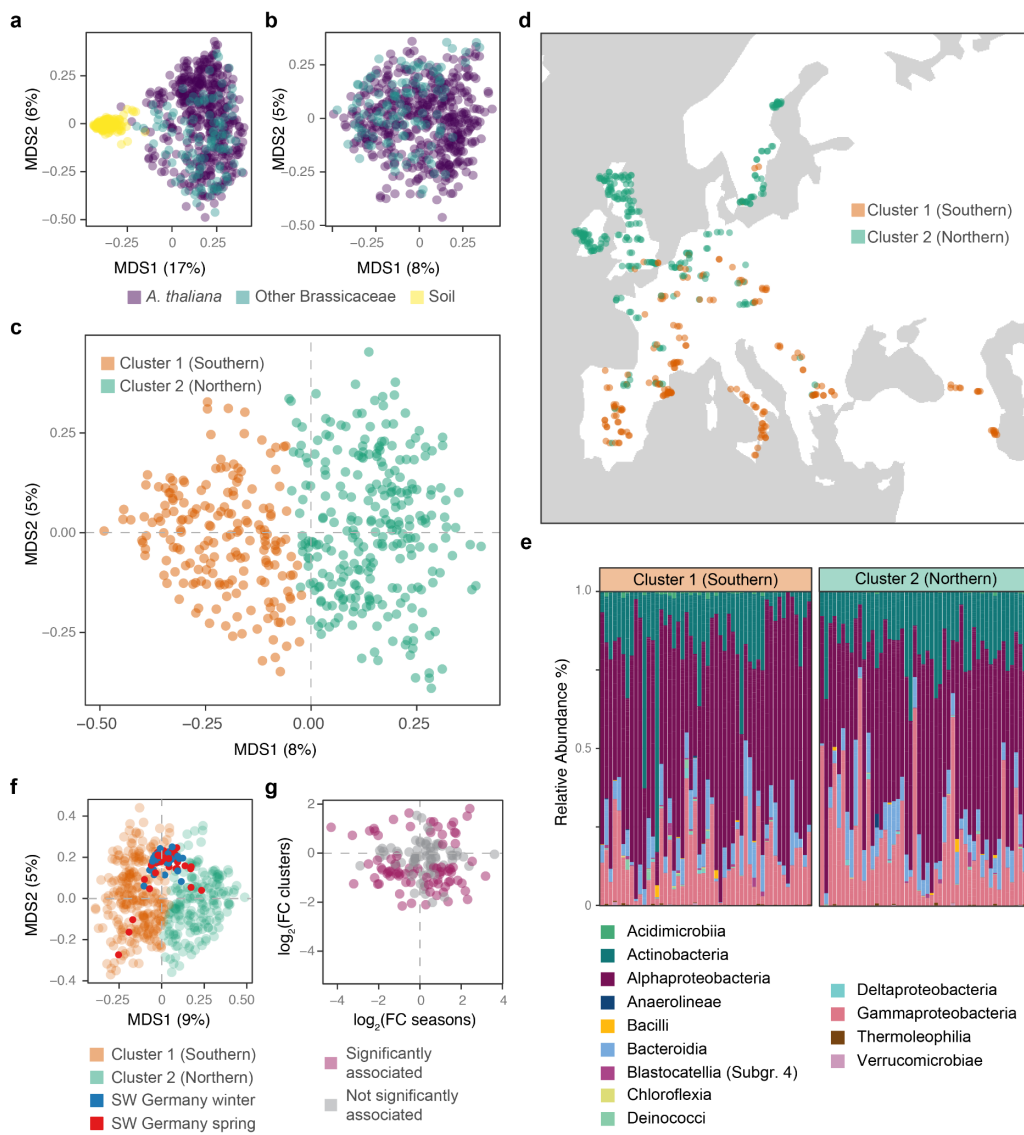
Figure 2 | Two distinct microbiome types in *A. thaliana* along a latitudinal cline. a, b, The *A. thaliana* leaf microbiome is significantly differentiated from that of surrounding soil (a) and less so (though still significant) from surrounding Brassicaceae (b). **c, d.** k-means clustering (k=2) (c) identified two microbiome types that turned out to have a North-South latitudinal cline (c). **e,** Distribution of higher taxonomic levels across the Southern and Northern clusters. **f,** Comparison of seasonal variation in microbiome composition in southwest Germany (winter and spring samples) with the European variation (Cluster 1 and 2). **g,** Absence of correlation in fold-changes (FC) in phylotype abundance between the Southern and Northern clusters (y-axis) and between the winter and spring samples from southwestern Germany (x-axis). Phylotypes that are significantly associated with the geographic clusters are indicated.

To explicitly test for enrichment of microbes in the phyllosphere, we compared soil, leaves of *A. thaliana* plants as well as leaves from neighboring Brassicaceae across all 267 sites via multidimensional scaling (Hellinger transformation). As expected, there was broad-scale separation between the microbiomes of the phyllosphere and the soil (Figure 2a). Modeling[16] the effect of host versus soil on the abundance of core microbial phylotypes revealed that 91% (524/575) of phylotypes were differentially

abundant between the *A. thaliana* phyllosphere and soil (False Discovery Rate FDR<0.01), indicating extensive filtering of microbes that colonize the plant, whether from soil or from another environmental source. Because we collected *A. thaliana* in parallel with other Brassicaceae species, we could assess the role of host identity in microbial composition at the plant species level. Differential abundance testing[16] identified 205 out of 575 phylotypes (36%) that distinguished the phyllospheres of *A. thaliana* and those of neighboring Brassicaceae (Figure S2, S3). These results indicate that inter-species genetic or phenological differences have a strong influence on microbiome composition. On a phylotype-by-phylotype basis, the abundance of a phylotype in *A. thaliana* was poorly predicted by the phylotype's abundance in soil or in the surrounding Brassicaceae companion species (Figure S3).

Phyllosphere microbial composition exhibits a latitudinal gradient

We tested the geographic differentiation of *A. thaliana* microbiomes and their constituent members using dimensionality reduction techniques for the entire community and through independent testing of the spatial distribution for each bacterial phylotype (ASV). The former can reveal overarching trends in composition, while the latter provides information on which microbes contribute to observed trends. Dimensionality reduction suggested spatial variation in the composition of the phyllosphere microbiome, with loading on the first and second principal coordinate axes (Fig 2c) correlated with latitude (Pearson's R = 0.75, p = 2.2x10^{-16}, and R = -0.24, p = 1.35x10^{-7}, respectively). Silhouette scoring[17] indicated that the microbiomes were best characterized as two distinct groups or types, and we used k-means clustering to categorize our samples into two distinct groups (Figure 2c, Figure S4). Clustering with k=2 revealed two microbiome types that were strongly differentiated by geography, with one dominating on plants in Northern Europe and the other on plants in Southern Europe (Figure 2d). Among individual phylotypes, the relative abundance of a third (33%) of phylotypes was significantly associated with latitude (linear regression, FDR < 0. 01), but only 2% were correlated with longitude, confirming that Northern European *A. thaliana* populations reproducibly harbor a different microbiota than those in the South.

Phyllosphere composition is not static throughout the lifetime of a plant, but changes with the season and developmental stage[18]. To test whether the latitudinal gradient in microbiota we observed could be explained by seasonal and host developmental differences, we compared our spatial phyllosphere dataset with a multi-year dataset collected from a single location in Southern Germany[19]. Projecting the seasonal phylotype compositional data into the MDS biplots of the spatial data did not reveal any preferential association of the season of collection with microbiome cluster type (Figure 2f). Comparing changes in the abundance of single phylotypes between seasons and between the two cluster types (Figure 2g) similarly did not point to the observed large-scale geographic variation reflecting environmental variation that aligns with different seasons in a single geographic region (Wald-test of multinomial frequency estimates, p > 0. 01).

The association between latitude and phylotype abundance was phylotype-specific, differing within and between bacterial families (Figure 3a). Previous studies demonstrated that *Pseudomonas* and *Sphingomonas* were reproducibly abundant genera across *A. thaliana* populations[19–21], and strains of both genera can affect *A. thaliana* health[19,22,23]. Focusing on these two abundant genera, linear regression assessing the relationship between each core phylotype and latitude revealed that four of the five most abundant Sphingomonads exhibited latitudinal clines (Figure 3a, b, FDR<0. 01) while the most abundant Pseudomonad phylotypes were not distributed along latitudinal gradients (Figure 3a, b, c). A result of the phylotype specificity of association with latitude was that the differentiation between microbiome clusters was significant at the phylotype level, but not at higher levels of taxonomic classification (Figure 2e). Thus,

even though *A. thaliana* plants are colonized by different phylotypes in Northern and Southern Europe, the composition of microbiomes at the level of bacterial classes is broadly the same (Figure 2e).

Common *A. thaliana* pathogens are patchily distributed and do not correlate with latitude

Arabidopsis thaliana exhibits extensive genetic variation at loci involved in immunity, and these loci may participate in shaping microbial composition, especially those of pathogens. Immune loci, however, vary often within each population, but little across large geographic distances[24,25]. If there is pervasive selection for immune variation within populations, why do many bacterial microbes then show latitudinal gradients in their association with *A. thaliana*? To test whether a similar cline exists specifically for bacterial pathogens, we focused on a bacterial lineage that is known to be a prevalent pathogenic threat for *A. thaliana*. A previous study[19] had identified a single *Pseudomonas* phylotype, ATUE5 (previously OTU5), that was very common in local populations in Southwest Germany, and that was highly pathogenic in the lab. ATUE5 turned out to be also the most abundant Pseudomonadaceae phylotype across all of Europe. Because ATUE5 is an important driver of total microbial load[19], we wanted to learn whether distribution of this phylotype was spatially structured and perhaps a contributor to the observed microbiome types (Figure 3c). The phylotype that matched exactly to ATUE5 was the seventh most common phylotype overall, ranging in relative abundance from 0-64% of the microbial community (mean = 1.8%). ATUE5 was present across the continent, without significant latitudinal differentiation (Pearson's R = 0. 01, p = 0. 92).

While ATUE5 was ubiquitously present, its distribution was not uniform. Interpolation by ordinary Kriging of its abundance across the collection range indicated instead a very patchy presence (Figure 3c). In contrast, the most frequent *Sphingomonas* phylotype (and most frequent phylotype overall) showed a significant latitudinal cline (Figure 3c). High levels of ATUE5 colonization were largely limited to single populations or populations that were very close to each other, with a spatial autocorrelation restricted to distances under 50 km (Figure S5). In total, the spatial autocorrelation analyses revealed that the *Pseudomonas* pathogen ATUE5 is widely distributed, but in contrast to the overall leaf-associated microbiota does not show obvious latitudinal patterns. We conclude that ATUE5 likely acts as a common, though not uniform, selective pressure across the *A. thaliana* range.

Host genetics is associated with differences in microbiome composition

Arabidopsis thaliana, a globally distributed species, exhibits strong population structure across Eurasia, with a pattern of isolation by distance[26] and greater differentiation along latitudinal than along longitudinal gradients[4]. Since the geographic genetic differentiation is almost certainly the result of both demographic and adaptive processes, a central challenge is to decipher how much of the spatial genetic variation is adaptive. Recent work has suggested that appreciable fractions of geographically structured genetic diversity in *A. thaliana* are associated with climate-driven selective pressures, particularly water availability and drought[27], but also with different groups of insect predators[28]. Whether this extends to microbial threats is unknown, but our observation of distinct microbiome types in different parts of the *A. thaliana* range is compatible with different microbiota affecting host genetic differentiation, spatially structured genetics selecting different microbiomes, or both.

Figure 3 | Latitudinal clines in microbial abundances and association of a host immune gene with microbiome type. a, Linear relationships between relative abundance (RA %) of the most common phylotypes. The y-axis represents $-\log_{10}$-transformed FDR-corrected p-value from regressing the abundance of the phylotype on latitude. **b,** There is a significant latitudinal cline for the relative abundance (RA %) of the most abundant Sphingomonads (top), but not for the most abundant Pseudomonads (bottom). **c,** Interpolation of the abundance of the top Sphingomonadaceae phylotype and the top Pseudomonadaceae phylotype, the known opportunistic pathogen ATUE5, revealed a continuous spatial gradient for the top Sphingomonad (left), but a patchy distribution with regional hotspots for the Pseudomonad (right). **d,** The relationship between microbiome type and polymorphism in plant immune genes was assessed with the F_{st} population differentiation index. The most extreme F_{st} values were found in the immune regulator *ACD6*.

Based on the type of genes that are associated with some of the most obvious signatures of selection, interactions with pathogens seem to be a major force shaping variation in the genomes of *A. thaliana* and its relatives[29–31], even though it remains largely a matter of speculation which pathogens are responsible for these observed signatures of selection. It is also still unknown whether selection on host genes involved in pathogen responses has led to knock-on effects on microbiome composition at large. Moreover, while it is clear that differences in microbial exposure between plants have led to the maintenance of immune diversity, much of this diversity is maintained within populations[24], and there is little if any geographic structuring of polymorphisms in these genes, suggesting the widespread distribution of the selective pathogens.

If variation in most immune genes is not geographically structured, what then explains the geographic differentiation in microbiome types of *A. thaliana*? Is it variation in other genes than members of the immune system, or is it purely driven by the environment? One possible explanation is that the different microbiome types belie similarity in community function, with microbial communities selecting for similar traits across populations, even if the constituent members differ. Such a scenario is also consistent with the same host proteins often being targeted by disparate microbes[32]. Another possibility is that the majority of microbes does not impose strong selective pressures on their hosts but instead colonizes them at low levels without extractive or beneficial interactions. In our study we were able to determine microbiome composition, microbiome load and host genotype from the same wild individuals, and we could therefore assess the relationship between host genotype and microbiome composition in an unbiased manner.

The observed host genetic diversity was consistent with previous surveys (Figure S6), indicating that we had sampled a broad diversity of *A. thaliana* genotypes across Europe. To determine the relationship between host genotype and microbiome composition, we fitted a mixed-effect model that included relatedness as a random effect and the loading on the first axis of the decomposition of the microbiome composition as the phenotypic response variable, finding that plant genotype alone explains 68% of the variance in the loading on the first principle coordinate axis, MDS1, and 52% of the variance in the loading on MDS2 (pseudo-h^2 = 0.68, s.e=0.10 for MDS1 and pseudo-h^2 = 0.52, s.e.=0.12 for MDS2).

Because immune genes are prime targets for interactions with pathogenic microbes, we tested for differentiation, as measured with the fixation index F_{st}, in immune gene alleles present in plants colonized by the different microbiome types. Among a generous, though not exhaustive, list of 1,103 genes with connection to pathogen response and defense[33], the top three SNPs were in the coding region or introns of the immune regulator *ACD6* (empirical p = 0. 0001) (Figure 3d, S7). Alleles at the *ACD6* locus can differentially confer resistance to *Pseudomonas syringae* and *Hyaloperonospora* pathogens in a greenhouse environment through constitutive effects on immunity[34]. The full *ACD6* haplotypes associated with each microbiome type have not yet been reconstructed, as the short reads used for genotypic comparisons did not allow for resolution of the full-length alleles. Nonetheless, our results demonstrate a striking association between microbiome type and polymorphisms in a known central regulator of immune activation. Whether resident microbiota select for *ACD6* allele type or instead *ACD6* allele type influences microbiome type remains to be determined.

Drought Indices are the strongest predictors of microbiome composition

While it was satisfying to have discovered that *ACD6*, an immune regulator with known major allelic variation, is associated with microbiome composition, common-garden experiments had suggested that host genetics plays only a minor role in shaping bacterial microbiome composition[9]. Assuming that the core members of the *A. thaliana* phyllosphere have dispersed over large geographic regions, the most obvious candidates for geographic structuring of microbiome types are abiotic, especially climate-related factors.

We took advantage of curated public data on climate variables to investigate effects of climate on microbiome composition. Since the microbiome might at least in part reflect overall plant growth and health[35], we included plant growth and health traits as potential confounders. Altogether, we considered 39 covariates that could influence microbiome composition (Figure S8, Table S1). To identify those covariates that most significantly predict microbiome composition, we first removed covariates that were highly correlated with others and then performed random forest classification using the two microbiome types as response variables (Figure 4). The resulting model indicated the covariate with greatest explanatory power to be the mean annual Palmer Drought Severity Index (PDSI) for the six months predating collection, a metric of the dryness of the local environment based on recent precipitation and temperature[36]. PDSI was similarly the best predictor for the loading of a sample on MDS1. In general, environmental covariates were stronger predictors than were plant health and life history traits. In contrast, environmental covariates (including PDS1) had poor explanatory power for soil microbiome composition, explaining less than 1% of the variance in the loading on the first principal coordinate axis.

Because PDSI is correlated with latitude, one possible explanation for the predictive capacity of PDSI is that PDSI itself is not causatively associated with microbiome differences, but instead is a proxy for other variables correlated with latitude. While we cannot exclude this possibility with the current data, we tested whether inclusion of information about latitude and PDSI improves prediction outcomes

beyond models that include latitude alone. We found significant improvement in predictive capacity (p =4. 2*10⁻⁷ for logistic regression with cluster identity and p=2. 7*10⁻⁷ for linear regression on MDS1) when PDSI was included in the model, indicating that the association between microbiome type and PDSI extends beyond latitudinal correlation. The predictive relationship between microbiome composition and PDSI was further found within regions and sampling tours (p =2. 3*10⁻⁷ for logistic regression with cluster identity, and p =0. 047 for linear regression on MDS1).
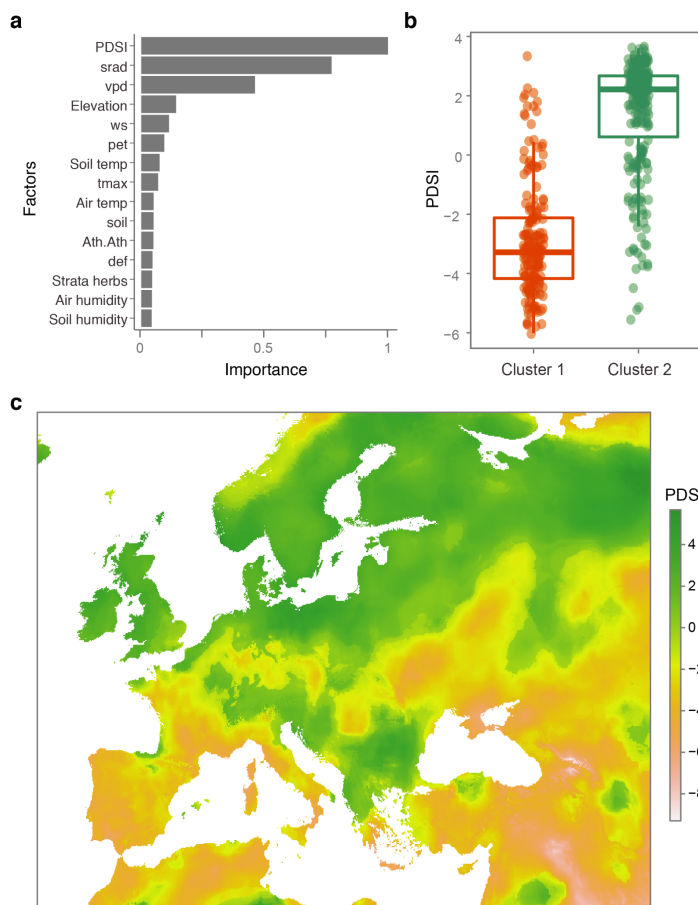
Figure 4 | PDSI is the best predictor of microbiome type. a, Random forest modeling was used to determine environmental variables associated with microbiome type. Abbreviations explained in Methods. **b,** PDSI of the location was the best predictor of microbiome type, explaining more than 50% of the variance. **c,** Mean PDSI throughout Europe for January to April 2018.

Previous work on non-host associated soils supports the importance of water availability in determining microbiome composition[10]. Together with our findings, this implicates water availability as one of the most, if not the most important environmental factor to determine which microbes colonize different sites. To probe for possible relative contributions of genotype and drought, we performed mixed effect modeling and estimated the marginal R^2 for PDSI to be 50%. In comparison, all environmental covariates explained only 1% of the variance in total microbiome composition in soil, as measured by the loading on the first principal coordinate axis. We conclude that drought is very likely to affect which microbes can access the host plant or proliferate in and on the host. Drought might do so directly, by affecting the physiological state of the plant, indirectly by shaping host genetics, or both.

Discussion

Our results reveal several robust trends. Firstly, colonization of *A. thaliana* leaves serves as a strong filter from the surrounding environment, with the majority of microbes differing in abundance between the soil and *A. thaliana* leaves, and more than a quarter differing between *A. thaliana* and its neighboring plant species. Plant genetics clearly matters for determining which microbes manage to establish in and on the plant. While these trends have been observed before, the strength of our study is the demonstration of how reproducible and ubiquitous these effects are on a continental scale. Secondly, geography and its corresponding abiotic factors significantly influence the microbes that colonize *A. thaliana* populations: a plant in Spain will very likely be colonized by a different suite of microbes than a plant in Sweden. These

parallel differences in geography and microbiomes also correlate with plant genetics, though the geographic distribution of microbiome composition is likely simpler than is the population structure of *A. thaliana*[4]. A major task for future work will be to disentangle the direct contribution of geography-dependent climate differences on the microbiome and those that are mediated by adaptive differences in plant genetics. We identify polymorphisms in an immunity gene, *ACD6,* that are associated with microbiome type and with PDSI. Specific alleles of *ACD6* confer drought tolerance[37], adding further complexity to our understanding of the relationship between drought, microbes and plant genetics. Lastly, our analyses suggest that which microbial community colonizes a plant is primarily dictated by water availability or rain and its associated microbiota. This again raises the question of how the different microbial communities influence plant phenotype. Drought not only plays a major selective role in *A. thaliana* populations[27], but it is also known to affect the ability of plants to withstand pathogen attack. An important question will be whether different background microbiomes in plants that are more likely to experience drought in the wild will help or hamper defense against pathogens[38]. Conversely, we need to learn whether the resident microbiome aids in protection against drought stress, or whether it exacerbates the effects of limited water availability. Future manipulative work that examines reciprocal transplants of plants and their microbial communities can shed light on the adaptive consequences of the reproducible shifts in microbiome composition that we have observed across Europe in natural populations of *A. thaliana*.

Methods

Sample collection

Arabidopsis thaliana and other Brassicaceae were sampled during local Springtime in 2018. Most Brassicaceae companion samples were *Capsella bursa-pastoris*, and the rest were *Cardamine hirsuta*. A full list of sampling locations and dates is provided in Table S1. Rosettes were separated from the roots using alcohol-wipe-sterilized scissors and forceps, then washed with water and ground with a sharp disposable spatula (Roth) in RNAlater (Sigma, now ThermoFisher). For each *A. thaliana* plant for which soil was accessible, 1-3 tablespoons of soil were collected from the location where the plant had been removed, and placed in a clean airtight bag. Samples were then maintained in electrical coolers (Severin Kühlbox KB2922) until the end of the sampling trip (between 1-7 days). In the lab, samples were stored at 4°C. Within 0-3 days RNAlater was removed from plant samples. Samples were centrifuged for 1 min at 1000 g, the supernatant was removed and samples were washed with 1 ml autoclaved water. For storage at -80°C, plant tissue was transferred with ethanol sterilized forceps to screw cap freezer tubes containing 1.0 mm Garnet Sharp Particles (BioSpec Products, Cat. No. 11079110GAR). A ~200 mg aliquot from each soil sample was transferred to a screw cap freezer tube using an ethanol sterilized spatula, with great effort to exclude plant and insect pieces. Prior to aliquoting, soil bags were kept at -80°C and defrosted at 4°C overnight, unless aliquoting was done immediately upon arrival in the lab at the end of the sampling trip.

Plant phenotyping

Scores presented in Figure 1 and Figure S1 are:

Developmental state: vegetative (1), just bolting (2), flowering (3), mature (4), drying (5)

Herbivory index: no (1), weak (2), strong (3), very strong (4) herbivory.

Disease index: no disease (1), visible *H. arabidopsidis* spores (2), necrosis (3), leaf coloration suggesting infection (4), *Albugo* ssp. pustules (5).

For rosette diameter, a 1cm rosette diameter listing corresponds to any rosette diameter less than or equal to 1cm.

DNA extraction

Plant DNA was extracted from plant samples according to the protocol from ref.[19]. Soil DNA was extracted using Qiagen Mag Attract PowerSoil DNA EP Kit (384) (Cat. 27100-4-EP). On dry ice, soil samples were transferred from tubes to PowerBead DNA plates using sterile individual funnels. Plates were stored up to two weeks at -80°C until processing. The Qiagen protocol was adapted to a 96-well-pipette (Integra Viaflo96). PowerBead solution and SL Solution were pre-warmed at 55-60°C to avoid precipitation. RNase A was added to the PowerBead solution just prior to use. From step 17 of the protocol, instead of starting epMotion protocol, the following steps were performed: To each well of the 2 ml deep-well-plate containing maximum 850 µl of supernatant, 750 µl of Bead Solution were added and mixed with Eppendorf MixMate at 650 rpm for 10-20 minutes. Plates were placed on a magnet for 5 minutes, the supernatant solution discarded, and the beads washed three times with 500 µl Wash solution. Beads were eluted with 100 µl Elution Buffer. The eluate was transferred to PCR plates and stored at -20°C until library preparation.

16S rDNA ASV identification

Primers targeting the consensus v3-v4 rDNA region from 341 bp (5'-CCTACGGGAGGCAGCAG-3') to 806 bp (5'-GGACTACNVGGGTWTCTAAT-3') were used to amplify 16S rDNA sequences with the protocol described in ref.[19]. Briefly, amplification was achieved with a two-step PCR protocol in which 100uM PNA was used in the initial PCR to block amplification of chloroplast. Amplicons were sequenced on the MiSeq (Illumina) platform using the MiSeq Reagent Kit v3 (600 cycle). Samples with lower coverage were preferentially sequenced to greater depth in subsequent runs in a total of four runs of the Miseq. Output from all runs was pooled for downstream analysis. Primer sequences were removed prior to analysis with a combination of usearch (version 11[39]) and custom bash scripting. 16S rDNA sequences were quality-trimmed using DADA2[12] (version 1.10.1). The forward read was truncated at position 260, the reverse read at position 210 due to decreased quality of the second read. Reads were truncated when the quality score dropped to less than or equal to 2 (trunQ=2). Chimeras were removed with the removeBimeraDenovo function (method='consensus') and ASVs called denovo using DADA2. The resulting reads were then aligned using AlignSeqs from the DECIPHER package[40] (version 2.8.1). A phylogenetic tree of the de novo called ASVs was constructed using fasttreeMP [41]. Taxonomic assignment of reads was performed with comparisons of 16S rDNA sequences to the Silva database[42] (nr v132 training set).

Only samples with 1,000 or more reads after filtering for mitochondria and chloroplast were included. We began with 939 samples, in which we found 195,545 ASVs. 918 samples had a sufficient number of reads, and removing ASVs that were not found in any single sample with more than 50 reads, we were left with 10,566 ASVs. We identified a core set of 575 ASVs by filtering for those ASVs that were present in at least 5% of *A. thaliana* samples. ASVs classified as belonging to the taxonomic class Cyanobacteria were removed from the dataset to eliminate possible misassignment of plant chloroplast DNA that can vary between plant genotypes and skew subsequent analyses.

Climate variable data acquisition

The majority of climate variables were obtained from Terraclimate[11] using the data for 2018 (http://www. climatologylab. org/terraclimate. html), a dataset with approximately 4 km spatial resolution. For random forest modeling and climate associations, we calculated the average value of each climate metric over the six months preceding the date of collection. The following variables were included in the random forest modeling from the Terraclimate dataset: tmax, maximum temperature; tmin, minimum temperature; vp, vapor pressure: ppt, precipitation accumulation; srad, downward surface shortwave radiation; ws, wind-speed; pet, reference evapotranspiration (ASCE Penman-Montieth); q, runoff; aet, actual evapotranspiration; def, climate water deficit soil, soil moisture; swe, snow water equivalent; PDSI, Palmer Drought Severity Index; vpd, vapor pressure deficit.

We further analyzed associations with Koeppen-Geiger climatic zones[43,44] which were inferred in R using the package kgc and the regional classifications from ref. [45]. Initial assessments of the density of microbes throughout Europe were calculated via ordinary Kriging using the R package automap[46] (version 1.0-14). Four models were tested during variogram fitting, "Sph", "Exp", "Gau" and "Ste". Interpolation was performed either on the abundance data untransformed or on log10 transformed values with 0. 0001 added to allow for zero counts to be included. Global information on the major vegetation types was obtained using the Globcover 2009 map (released December 2010) from the European Space Agency (http://due. esrin. esa. int/page_globcover. php). Measures of soil properties were obtained using the ISRIC (global gridded soil information) Soil Grids (https://soilgrids. org/#!/?layer=geonode:taxnwrb_250m).

At the time of collection we took several measurements of the soil and air temperature and humidity (Soil temp; Air temp; Soil hum; Air hum), the surrounding plant community and the location type: distance between the focal and the closest neighboring *A. thaliana* plant (Ath.Ath); distance between the focal and the closest other plant (Ath.other); immediate plant density (Ground cover); visible *H. arabidopsidis* infection on focal plant (HpA plant) or at site (HpA site); visible *Albugo* spp. infection on focal plant (Albugo tour); fraction of herbal plants in the surrounding (Strata herb); estimated sun exposure (Sun), slope (Slope) and ground humidity (Humidity ground). Measurements are listed and detailed in Supplementary Table 1.

Feature selection and random forest modeling

Features of interest were first identified by feature selection in the R package caret[47] (version 6.0-86) using repeated cross-validation (3 repeats). Prediction variables were preprocessed by centering, scaling and nearest-neighbor imputation for samples that lacked data for a variable. A training set was generated with 75% of the data. Random forest regression was performed to minimize the root mean squared error with repeated cross validation. Variable importance was assessed via generalized cross-validation in the package caret[47].

Differential abundance

Differential abundance of ASVs in soil vs. *A. thaliana*, and *A. thaliana* vs. other Brassicaceae was assessed using the edgeR[16] package in R (version 3.28.1). We estimated a common negative binomial dispersion parameter, and abundance-dispersion trends by Cox-Reid approximate profile likelihoods[48]. We then fit a quasi-likelihood negative binomial generalized log-linear model to the count data. We tested for differential abundance by a likelihood ratio test.

Classification and regression

Phylotypic clusters were identified by k-means clustering of Hellinger-transformed ASV count matrices. The optimal number of clusters was determined through both partitioning around medioids[49] using the pamk function in the R package fpc[50] (version 2.2.9) and through silhouette analysis[17] in the cluster (version 2.1.2) package in R[51].

To determine the relative effect sizes of drought, latitude and plant identity on MDS loadings, phenotypes were modeled with the gaston[52] package in R. The model was $Y_i \sim PDSI_i + Lat_i + k_{i,j} + \varepsilon_i$ where Yi is the phenotype for each ith accession, $k_{i,j}$ is the genetic relatedness between the ith and jth accessions[52], Lat_i is the Latitude of the collection location of the ith accession, and ε_i is the unaccounted error. The kinship matrix was constructed using several methods including the R package gaston[52] as well as the centered kinship matrix in gemma (version 0.98.3)[53]. The different methods yielded unstable estimates of kinship, likely due to the low coverage of the plant genomes. To account for the low coverage, we employed a method designed for kinship estimation in low coverage data, SEEKIN [54] using the homogeneous parameter. Mixed effect modeling with a kinship matrix was computed both with lmekin[55] and with gemma. The proportion of phenotypic variance explained by the environmental covariates was estimated with the function "r.squaredLR" from the package MuMIn (version 1.43.1) and the pseudo-heritability was estimated using the kinship matrix and lmekin as well as in gemma (-gk=1, maf=0.1). In the manuscript we report the lower estimate for pseudo-heritability as estimated in gemma with the centered kinship matrix also estimated in gemma.

Plant polymorphism calling and filtering

Raw reads were mapped to the TAIR10 reference genome of *A. thaliana* with bwa-mem (bwa 0. 7. 15)[56]. Single nucleotide polymorphism (SNP) calling was performed using GATK (version 3. 5) HaplotypeCaller using recommended best practices[57] with some modifications. Filtering for individuals with greater than 25% missing data (across all the SNPs) and bi-allelic SNPs with greater than 25% missing data (across all the individuals) resulted in a final set of 527 individuals with 409,850 bi-allelic SNPs for further analysis.

Assessing population structure of *A. thaliana* plants

Wright's fixation index (F_{st}) was calculated using the method of Cockeram and Weir[58]. The 1001 Genomes[4] dataset (without individuals from North America) was merged with the dataset from this study to perform principal component analysis. Genotypes from this study were projected into the principal component space of the 1001 Genomes genotypes using the SmartPCA tool of EIGENSOFT (version 6)[59].

References

1. Zolla, G., Badri, D. V., Bakker, M. G., Manter, D. K. & Vivanco, J. M. Soil microbiomes vary in their ability to confer drought tolerance to Arabidopsis. *Appl. Soil Ecol.* **68**, 1–9 (2013).
2. Wagner, M. R. *et al.* Natural soil microbes alter flowering phenology and the intensity of selection on flowering time in a wild Arabidopsis relative. *Ecol. Lett.* **17**, 717–726 (2014).
3. Thiergart, T. *et al.* Root microbiota assembly and adaptive differentiation among European Arabidopsis populations. *Nat Ecol Evol* **4**, 122–131 (2020).
4. 1001 Genomes Consortium. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell* **166**, 481–491 (2016).
5. Hancock, A. M. *et al.* Adaptation to Climate Across the Arabidopsis thaliana Genome. *Science* vol. 334 83–86 (2011).
6. Bartoli, C. *et al.* In situ relationships between microbiota and potential pathobiota in Arabidopsis

thaliana. *ISME J.* **12**, 2024–2038 (2018).

7. Wagner, M. R. *et al.* Host genotype and age shape the leaf and root microbiomes of a wild perennial plant. *Nat. Commun.* **7**, 12151 (2016).

8. Rojas, J. A. *et al.* Oomycete Species Associated with Soybean Seedlings in North America—Part II: Diversity and Ecology in Relation to Environmental and Edaphic Factors. *Phytopathology®* **107**, 293–304 (2017).

9. Horton, M. W. *et al.* Genome-wide association study of Arabidopsis thaliana leaf microbial community. *Nat. Commun.* **5**, 5320 (2014).

10. Delgado-Baquerizo, M. *et al.* A global atlas of the dominant bacteria found in soil. *Science* **359**, 320–325 (2018).

11. Abatzoglou, J. T., Dobrowski, S. Z., Parks, S. A. & Hegewisch, K. C. TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958-2015. *Sci Data* **5**, 170191 (2018).

12. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).

13. Karasov, T. L., Neumann, M. & Duque-Jaramillo, A. The relationship between microbial biomass and disease in the Arabidopsis thaliana phyllosphere. *bioRxiv* (2019).

14. Lundberg, D. S. *et al.* Defining the core Arabidopsis thaliana root microbiome. *Nature* **488**, 86–90 (2012).

15. Bonito, G. *et al.* Plant host and soil origin influence fungal and bacterial assemblages in the roots of woody plants. *Mol. Ecol.* **23**, 3356–3370 (2014).

16. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

17. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).

18. Beilsmith, K., Perisin, M. & Bergelson, J. Natural Bacterial Assemblages in Arabidopsis thaliana Tissues Become More Distinguishable and Diverse during Host Development. *MBio* **12**, (2021).

19. Karasov, T. L. *et al.* Arabidopsis thaliana and Pseudomonas Pathogens Exhibit Stable Associations over Evolutionary Timescales. *Cell Host Microbe* **24**, 168–179.e4 (2018).

20. Regalado, J. *et al.* Combining whole-genome shotgun sequencing and rRNA gene amplicon analyses to improve detection of microbe--microbe interaction networks in plant leaves. *ISME J.* 1–15 (2020).

21. Lundberg, D. S. *et al.* Contrasting patterns of microbial dominance in the Arabidopsis thaliana phyllosphere. *bioRxiv* 2021.04.06.438366 (2021) doi:10.1101/2021.04.06.438366.

22. Innerebner, G., Knief, C. & Vorholt, J. A. Sphingomonas strains protect Arabidopsis thaliana against leaf pathogenic Pseudomonas syringae in a controlled model system. *Appl. Environ. Microbiol.* (2011).

23. Shalev, O., Karasov, T. L., Lundberg, D. S., Ashkenazy, H. & Weigel, D. Protective host-dependent antagonism among Pseudomonas in the Arabidopsis phyllosphere. *bioRxiv* 2021.04.08.438928 (2021) doi:10.1101/2021.04.08.438928.

24. Karasov, T. L. *et al.* The long-term maintenance of a resistance polymorphism through diffuse interactions. *Nature* **512**, 436–440 (2014).

25. Vetter, M., Karasov, T. L. & Bergelson, J. Differentiation between MAMP Triggered Defenses in Arabidopsis thaliana. *PLoS Genet.* **12**, e1006068 (2016).

26. Platt, A. *et al.* The scale of population structure in Arabidopsis thaliana. *PLoS Genet.* **6**, e1000843 (2010).

27. Exposito-Alonso, M. *et al.* Natural selection on the Arabidopsis thaliana genome in present and future climates. *Nature* **573**, 126–129 (2019).

28. Züst, T. *et al.* Natural enemies drive geographic variation in plant defenses. *Science* **338**, 116–119 (2012).

29. Bakker, E. G., Toomajian, C., Kreitman, M. & Bergelson, J. A genome-wide survey of R gene polymorphisms in Arabidopsis. *Plant Cell* **18**, 1803–1818 (2006).

30. Clark, R. M. *et al.* Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. *Science* **317**, 338–342 (2007).

31. Koenig, D. *et al.* Long-term balancing selection drives evolution of immunity genes in Capsella. *Elife* **8**, (2019).

32. Mukhtar, M. S. *et al.* Consortium EUE, Vandenhaute J, Roth FP, Hill DE, Ecker JR, Vidal M, Beynon J, Braun P, Dangl JL (2011) Independently evolved virulence effectors converge onto hubs in a pant immune system network. *Science* **333**, 596–601 (2011).

33. Glander, S. *et al.* Assortment of Flowering Time and Immunity Alleles in Natural Arabidopsis thaliana Populations Suggests Immunity and Vegetative Lifespan Strategies Coevolve. *Genome Biol. Evol.* **10**, 2278–2291 (2018).

34. Todesco, M. *et al.* Natural allelic variation underlying a major fitness trade-off in Arabidopsis thaliana. *Nature* **465**, 632–636 (2010).

35. McMullan, M. *et al.* Evidence for suppression of immunity as a driver for genomic introgressions and host range expansion in races of Albugo candida, a generalist parasite. *Elife* **4**, (2015).

36. Palmer, W. C. *Meteorological Drought.* (U.S. Department of Commerce, Weather Bureau, 1965).

37. Okuma, E., Nozawa, R., Murata, Y. & Miura, K. Accumulation of endogenous salicylic acid confers drought tolerance to Arabidopsis. *Plant Signal. Behav.* **9**, e28085 (2014).

38. Colaianni, N. R. *et al.* A complex immune response to flagellin epitope variation in commensal communities. *Cell Host Microbe* **29**, 635–649.e9 (2021).

39. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).

40. Wright, E. S. Using DECIPHER v2. 0 to analyze big biological sequence data in R. *R J.* **8**, (2016).

41. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).

42. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–6 (2013).

43. Koppen, W. Das Geographische System der Klimate, Handb. (1936).

44. Köppen, W. Versuch einer Klassifikation der Klimate, vorzugsweise nach ihren Beziehungen zur Pflanzenwelt. *Geogr. Z.* **6**, 593–611 (1900).

45. Rubel, F. & Kottek, M. Observed and projected climate shifts 1901--2100 depicted by world maps of the Köppen-Geiger climate classification. *Meteorol. Z.* **19**, 135–141 (2010).

46. Hiemstra, P. Package 'automap': Automatic interpolation package, R package version 1.0-14. (2013).

47. Kuhn, M. *et al.* Package 'caret'. *R J.* 223 (2020).

48. Cox, D. R. & Reid, N. Parameter Orthogonality and Approximate Conditional Inference. *J. R. Stat. Soc. Series B Stat. Methodol.* **49**, 1–39 (1987).

49. Kaufman, L. & Rousseeuw, P. J. Partitioning around medoids (program pam). *Finding groups in data: an introduction to cluster analysis* **344**, 68–125 (1990).

50. Hennig, C. & Imports, M. Package 'fpc'. https://cran.r-project.org/web/packages/fpc/fpc.pdf (2015).

51. Maechler, M. *et al.* Package 'cluster'. *Dosegljivo na* (2013).

52. Perdry, H. Claire Dandine-Roulland gaston: Genetic Data Handling (QC, GRM, LD, PCA). *Linear Mixed Models version* **1**,.

53. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821–824 (2012).

54. Dou, J. *et al.* Estimation of kinship coefficient in structured and admixed populations using sparse sequencing data. *PLoS Genet.* **13**, e1007021 (2017).

55. Therneau, T. M. & Therneau, M. T. M. Package 'coxme'. *Mixed effects cox models. R package version* **2**, (2015).

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

57. Van der Auwera, G. A. *et al.* From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Curr. Protoc. Bioinformatics* **43**, 11–10 (2013).

58. Weir, B. S. & Cockerham, C. C. Estimating F-Statistics for the Analysis of Population Structure. *Evolution* **38**, 1358–1370 (1984).

59. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).

Data and Software Availability

v3-v4 16S rDNA sequence data were deposited in the European Nucleotide Archive (ENA) under the Primary Accession ENA: PRJEB44379. Metadata and processed read data sets including phyloseq objects are available at Zenodo under DOI 10.5281/zenodo.5140512. Scripts for data processing, analyses and figure generation can be accessed at https://github. com/tkarasov/pathodopsis.

Acknowledgements

Author Contributions

TLK, RS, GS and DW devised the study. TLK, RS, GS, MN and the PATHODOPSIS collection team collected and prepared the samples. TLK, GS and MN processed the samples. TLK, RS and GS analyzed the data. GM provided climate data. TLK, RS and DW wrote the manuscript.

Competing interest declaration

We declare no competing interests.

Nagoya Protocol Compliance

Respective national authorities of all sampled countries Party to the Nagoya Protocol were contacted ahead of collections. Where needed, advised measures were taken and resulted in sampling and export permit KC3M-160/11. 04. 2018 (Bulgaria), ABSCH-IRCC-FR-253846-1 (France) and ABSCH-IRCC-ES-259169-1 (Spain).

Figure S1: Spatial distribution of plants with various developmental and health states. Arbitrary scales (see Methods) except for rosette diameter.
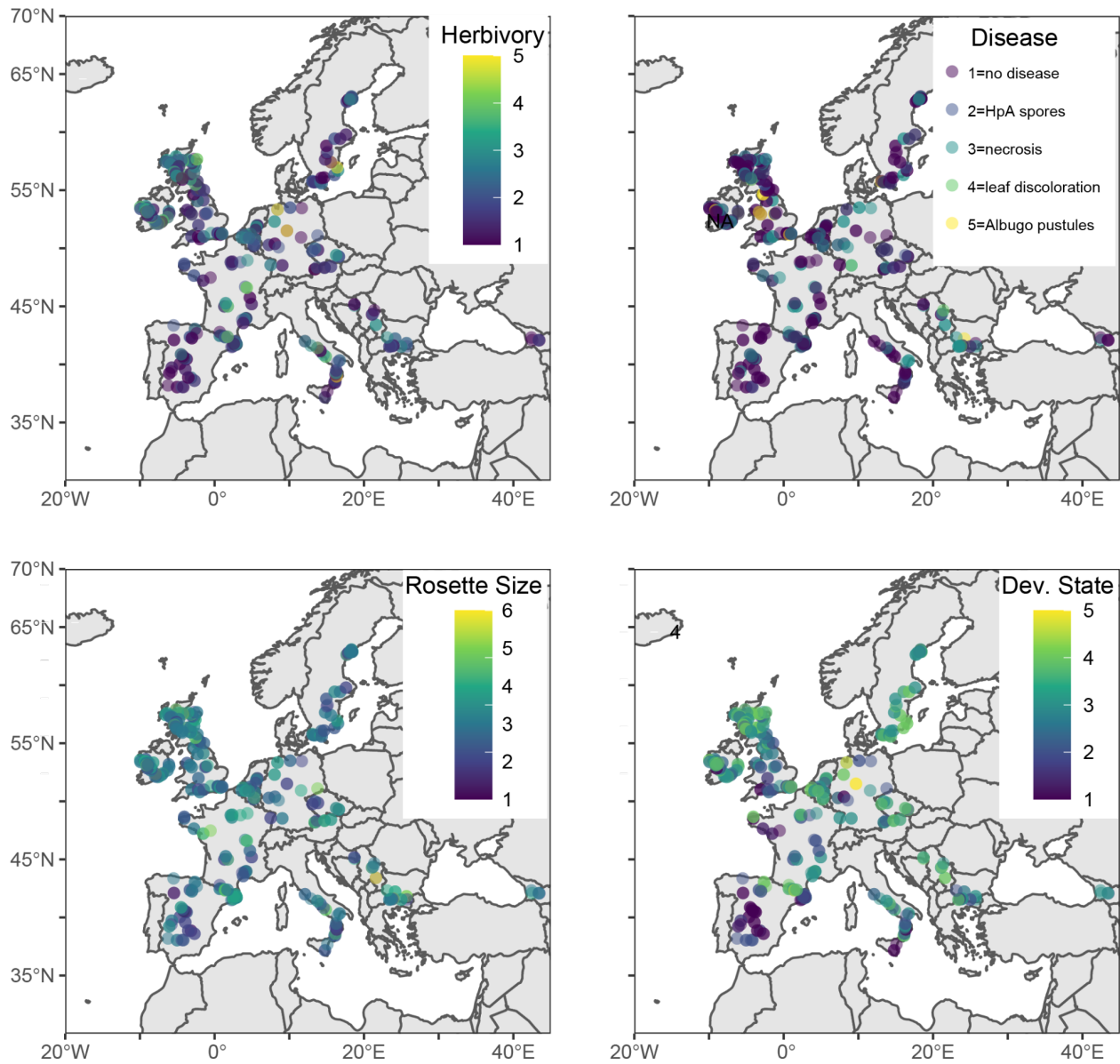
Figure S2: Differential abundance of phylotypes in soil, *A. thaliana* phyllospheres and phyllospheres of other Brassicaceae. Differential abundance analysis of phylotypes identified 91% of phylotypes as differentially abundant between *A. thaliana* and soil and 36% of phylotypes differentially abundant between *A. thaliana* and other Brassicaceae. y-axis shows the -log$_{10}$(FDR) for the association between the phylotype and latitude in linear regression.
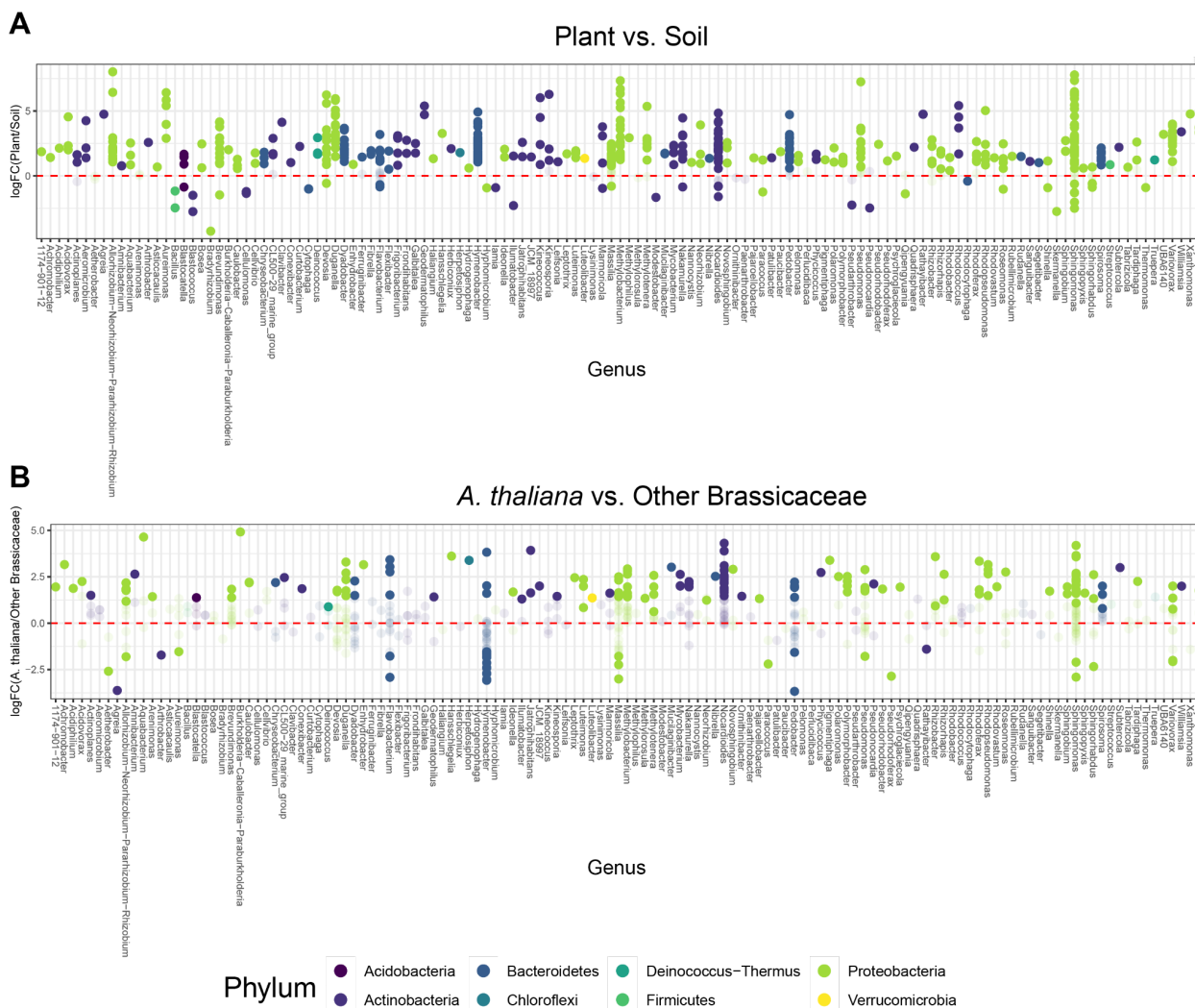
Figure S3: Within-site correlation of phylotype abundance.

Correlation coefficients were calculated for the co-occurrence of a phylotype within a site between the two *A. thaliana* collected at the site, *A. thaliana* x other Brassicaceae and *A. thaliana* x soil sample.
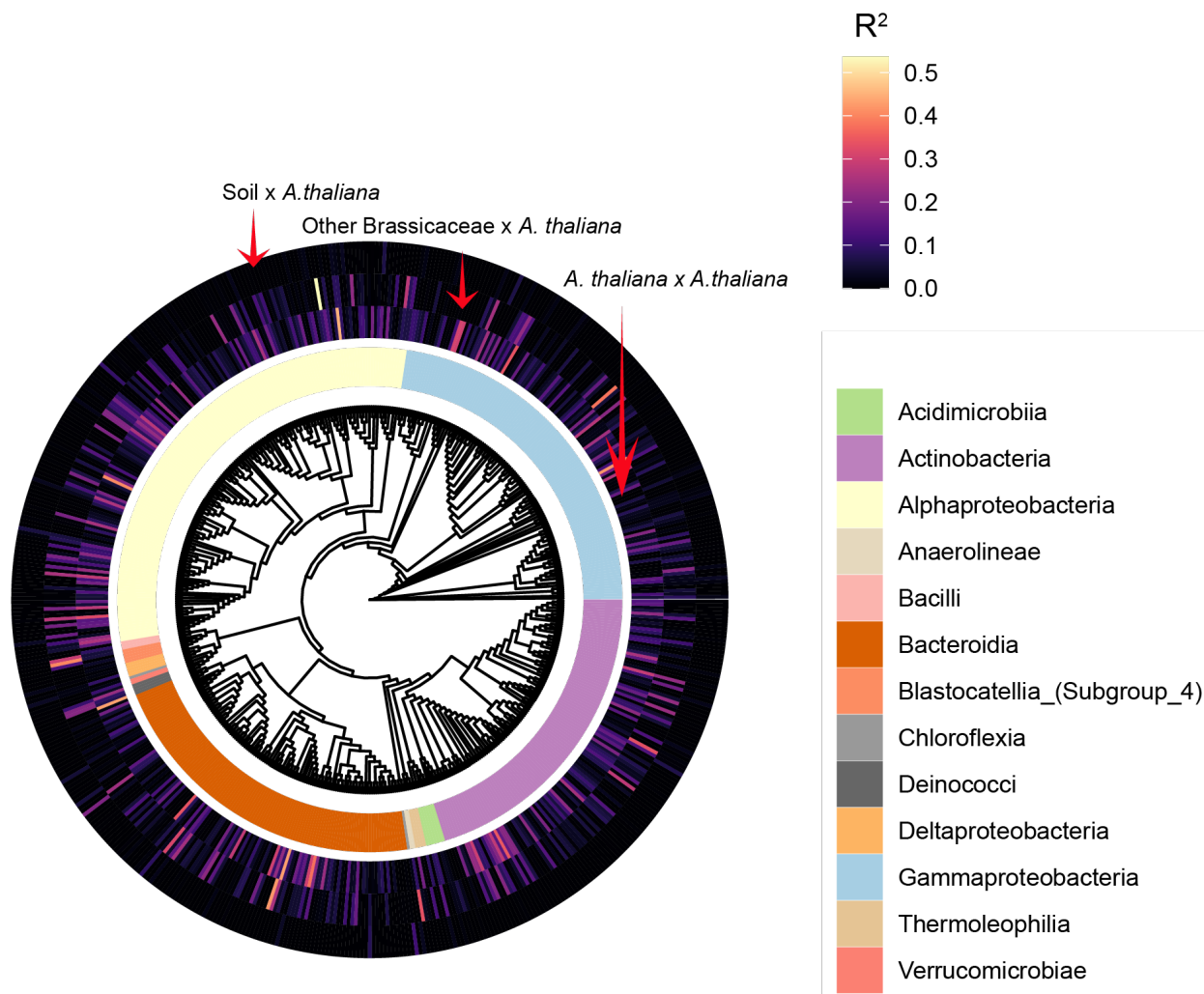
**Figure S4: Silhouette width estimated for members of two clusters.**

Silhouette scores for membership assignment to each of the microbiome types. For each cluster, "number of individuals in the cluster | average distance between a sample and members of the other cluster" is indicated.
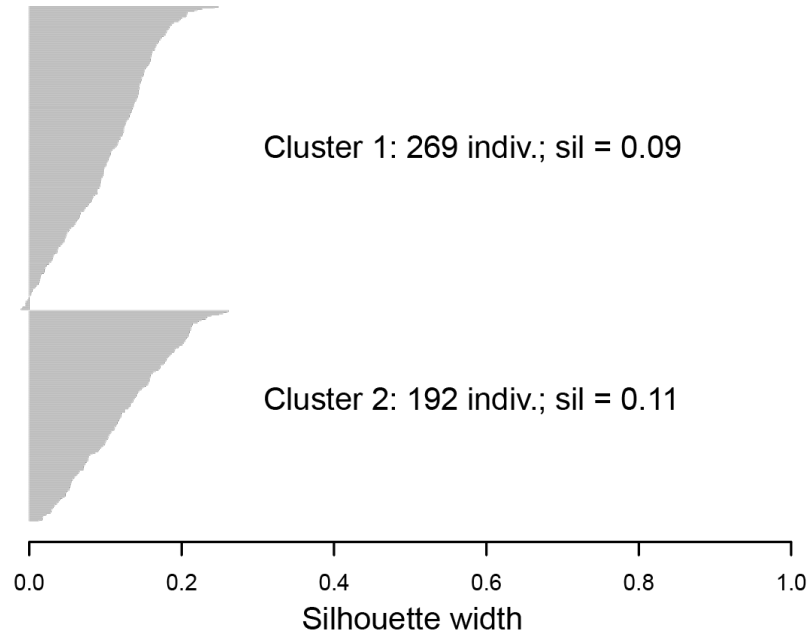
Figure S5: Distance-Semivariance plot for ATUE5. The relationship between the spatial distance between two plants, and the correlation of the abundance of ATUE5 between the two plants.
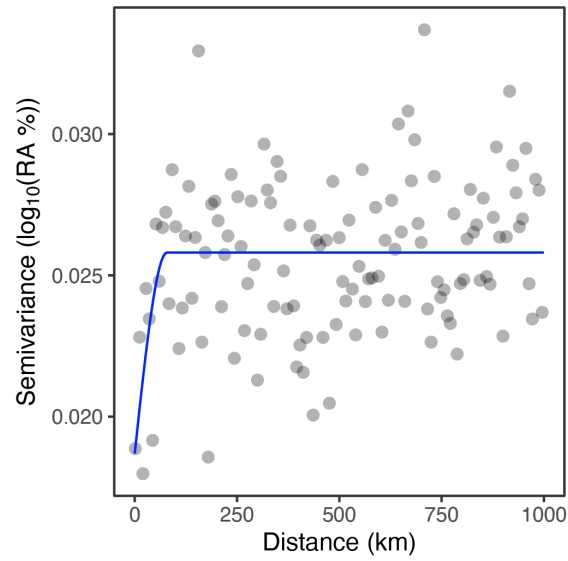
Figure S6: Projection of *A. thaliana* genotypes from this study projected into genotypic PC space from 1001 Genomes project. Individuals from this study ("Pathodopsis") align well with the broader 1001 Genomes ("1001g") population.
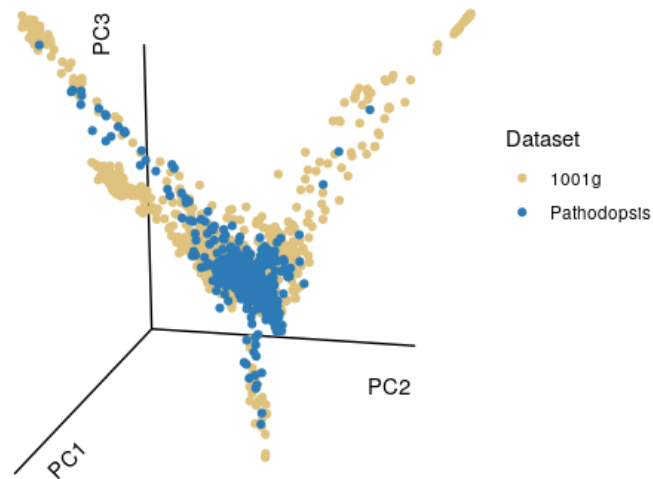
Figure S7: F$_{st}$ around *ACD6*. The fixation index F$_{st}$ was estimated for SNPs in a list of known immune-associated genes. The most extreme values of F$_{st}$ lie in the immune regulator *ACD6* which has previously been associated with autoimmunity in *A. thaliana* populations.
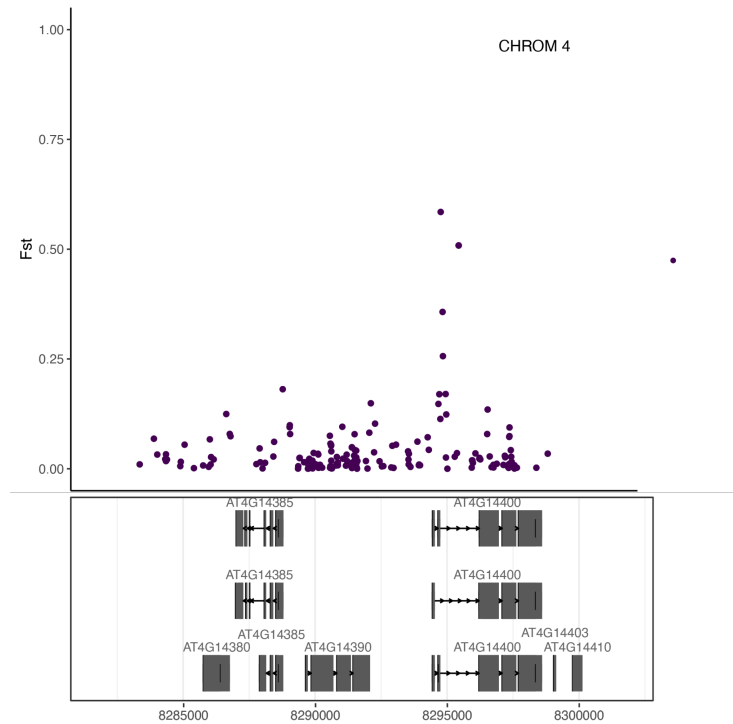
Figure S8: Correlogram of relationship between environmental and developmental covariates used in random forest modeling. Covariates are detailed in Methods.