Plant genetic effects on microbial hubs impact fitness across field trials

Benjamin Brachi[1, 2], Daniele Filiault[3*], Hannah Whitehurst[1*], Paul Darme[1], Pierre Le Gars[1], Marine Le Mentec[1], Timothy C. Morton[1], Envel Kerdaffrec[3], Fernando Rabanal[3], Alison Anastasio[1], Mathew S. Box[4], Susan Duncan[4], Feng Huang[1,5], Riley Leff[1], Polina Novikova[3], Matthew Perisin[1], Takashi Tsuchimatsu[3], Roderick Woolley[1], Caroline Dean[4], Magnus Nordborg[3], Svante Holm[6], Joy Bergelson[1]

Affiliations:

[1] Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637

[2] Univ. Bordeaux, INRAE, BIOGECO, F-33610 Cestas, France

[3] Gregor Mendel Institute (GMI), Austrian Academy of Sciences, Vienna Biocenter (VBC), Dr. Bohr-Gasse 3, 1030 Vienna, Austria

[4] John Innes Center, Norwich, UK

[5] South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, China

[6] Mid-Sweden University, Sundsvall, Sweden

*contributed equally to the work

Abstract:

Although complex interactions between hosts and microbial associates are increasingly well documented, we still know little about how and why hosts shape microbial communities in nature. In addition, host genetic effects on microbial communities vary widely depending on the environment, obscuring conclusions about which microbes are impacted and which plant functions are important. We characterized the leaf microbiota of 200 *A. thaliana* genotypes in eight field experiments and detected consistent host effects on specific, broadly distributed microbial OTU's. Host genetics disproportionately influenced hubs within the microbial communities, with their impact then percolating through the community, as evidenced by a decline in the heritability of particular OTUs with their distance to the nearest hub. By simultaneously measuring host performance, we found that host genetics associated with microbial hubs explained over 10% of the variation in lifetime seed production among host genotypes across sites and years. We successfully cultured one of these microbial hubs and demonstrated its growth-promoting effects on plants grown in sterile conditions. Finally, genome-wide association mapping identified many putatively causal genes with small effects on the relative abundance of microbial hubs across sites and years, and these genes were enriched for those involved in the synthesis of specialized metabolites, auxins and the immune system. Using untargeted metabolomics, we corroborate the consistent association of variation in specialized metabolites and microbial hubs across field sites. Together, our results reveal that host natural variation impacts the microbial communities in consistent ways across environments and that these effects contribute to fitness variation among host genotypes.

Main

Hosts harbor complex microbial communities that are thought to impact health and development [1]. Human microbiota has been implicated in a variety of diseases, including obesity and cancer [2]. Efforts are thus underway to determine the host factors shaping these communities [3,4], and to use next-generation probiotics to inhibit colonization by pathogens [5]. Similarly, in agriculture, there is great hope of shaping the composition of the microbiota in order to mitigate disease and increase crop yield in a sustainable fashion. Indeed, the Food and Agriculture Organization of the United Nations has made the use of biological control and growth-promoting microbial associations a clear priority for improving food production [6].

Plant-associated microbes can be beneficial in many ways, including improving access to nutrients, activating or priming the immune system, and competing with pathogens. For example, seeds inoculated with a combination of naturally occurring microbes were found to be protected from a sudden-wilt disease that emerged after continuous cropping [7]. Thus, it would be advantageous to breed crops that promote the growth of beneficial microbes under a variety of field conditions, a prospect that is made more likely by the demonstration of host genotypic effects on their microbiota [8–10]. However, microbial communities are complex entities that are influenced by the combined impact of host factors, environment and microbe-microbe interactions [11]. Indeed, several studies have found a strong influence of the environment on estimates of host genotype effects [8,12,13]. Although most, if not all, studies exploring the influence that host genotype exerts on microbial communities suggest that such plant control could be beneficial to plant performance, almost nothing is known about the relationship between host genotype effects on microbial communities and on plant performance or fitness. As a consequence, the extent to which host plants can control

3

66 microbial communities to their advantage, especially in a consistent manner across multiple
67 environments, remains unclear.

68 Here, we combine large-scale field experiments in natural environments, extensive
69 microbial community analysis, and genome-wide association mapping to: (i) determine how
70 host genotype affects different microbial community members, and thus shapes the overall
71 microbiome; (ii) estimate host genotype effects on microbial communities across eight
72 environments and investigate the contribution of those effects to the performance of plant
73 genotypes; and (iii) use genome-wide association mapping to identify key pathways that
74 shape the leaf microbial communities across multiple environmental conditions.
75

76 **Snapshot of microbial community variation**

77 We performed a set of field experiments that included natural inbred lines of *Arabidopsis*
78 *thaliana* (hereafter "accessions") originally collected throughout Sweden, mainly in two
79 climatically contrasted regions of the country (Supplementary Table 1); *A. thaliana* in the
80 north of Sweden experiences long, snowy winters, and as a consequence plants are typically
81 found on south-facing slopes of rocky cliffs. *Arabidopsis* populations in the south of Sweden,
82 on the other hand, tend to be associated with agricultural or disturbed fields that experience
83 highly variable snow cover over the winter months. We used replicate experiments in four
84 representative *Arabidopsis* sites, two each in the north (sites NM and NA) and south (sites SU
85 and SR) of Sweden. Experiments were repeated across two years, for a total of eight
86 experiments.

87 Each experiment was organized in a complete randomized block design including 24
88 replicates of 200 sequenced accessions [14], established as seedlings in a mixture of native
89 and potting soil and timed to coincide with local germination flushes in late summer.
90 Immediately upon snowmelt in early spring, we sampled and freeze-dried 5 to 6 whole

rosettes per accession. DNA was extracted from the freeze-dried rosettes and both the ITS1 portion of the Internal Transcribed Spacer (ITS) and the V5 to V7 regions of the 16S RNA gene were sequenced to characterize the fungal and bacterial communities, respectively [9,11,15]. The sequences obtained were clustered into Operational Taxonomic Units (OTUs) using Swarm to generate community matrices [14] (see "Count table filtering" section in the methods). The frequency distributions of OTUs were highly skewed, with the top ten most common OTUs contributing on average 59% of the reads in each experiment (ranging from 45 to 78%). Taxonomic assignments indicate that the fungal communities were dominated by Leotimycetes and Dothideomycetes while the bacterial communities included high proportions of Alphaproteobacteria and Actinobacteria (Extended Data Fig. 1).

In a principal coordinate analysis, differences between northern and southern sites explained 10 and 5% of the overall diversity in the fungal and bacterial communities, respectively, while differences between the two consecutive years explained 5 and 3%. This level of differentiation among experiments likely underestimates that present in the native soil, as it has been shown that hosts filter the microbial community to reduce site-to-site differences [17,18] (Fig. 1). In addition, there may have been a homogenizing effect of using a combination of local and potting soil. Irrespective of how well our treatments mimicked natural microbial communities, our analysis of eight common garden experiments permits assessment of the consistency across time and space of plant genetic effects on their associated microbial communities.

5

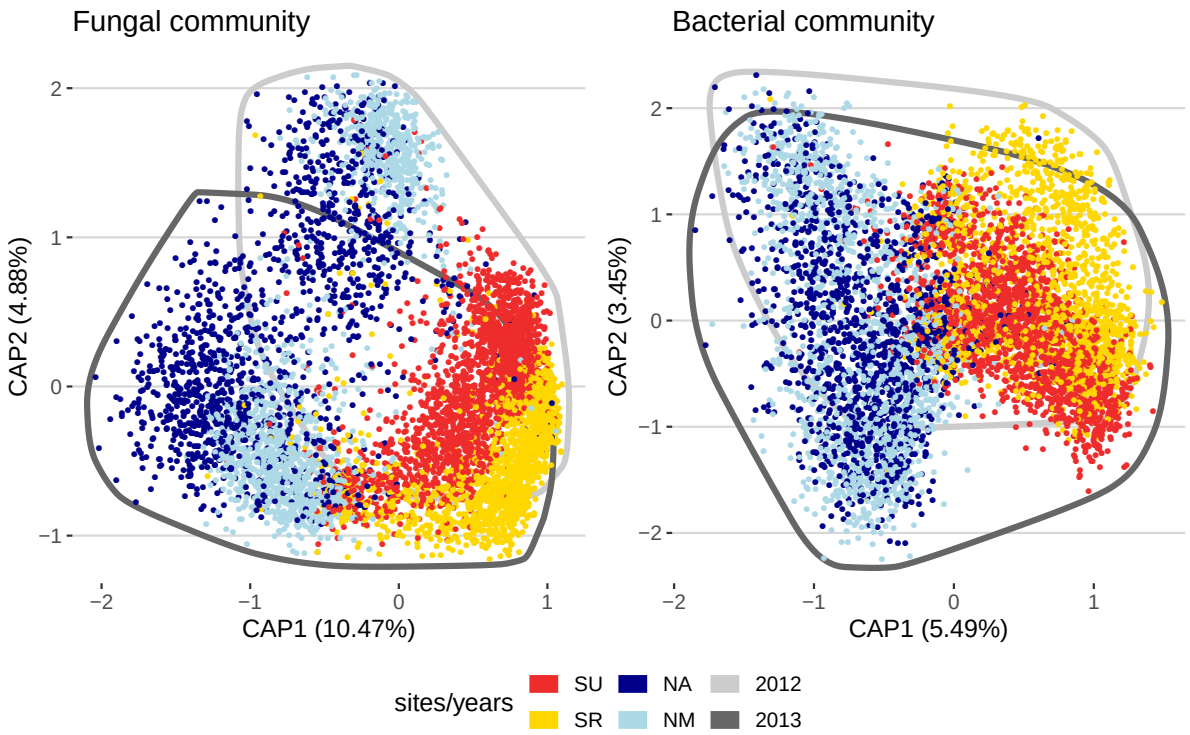
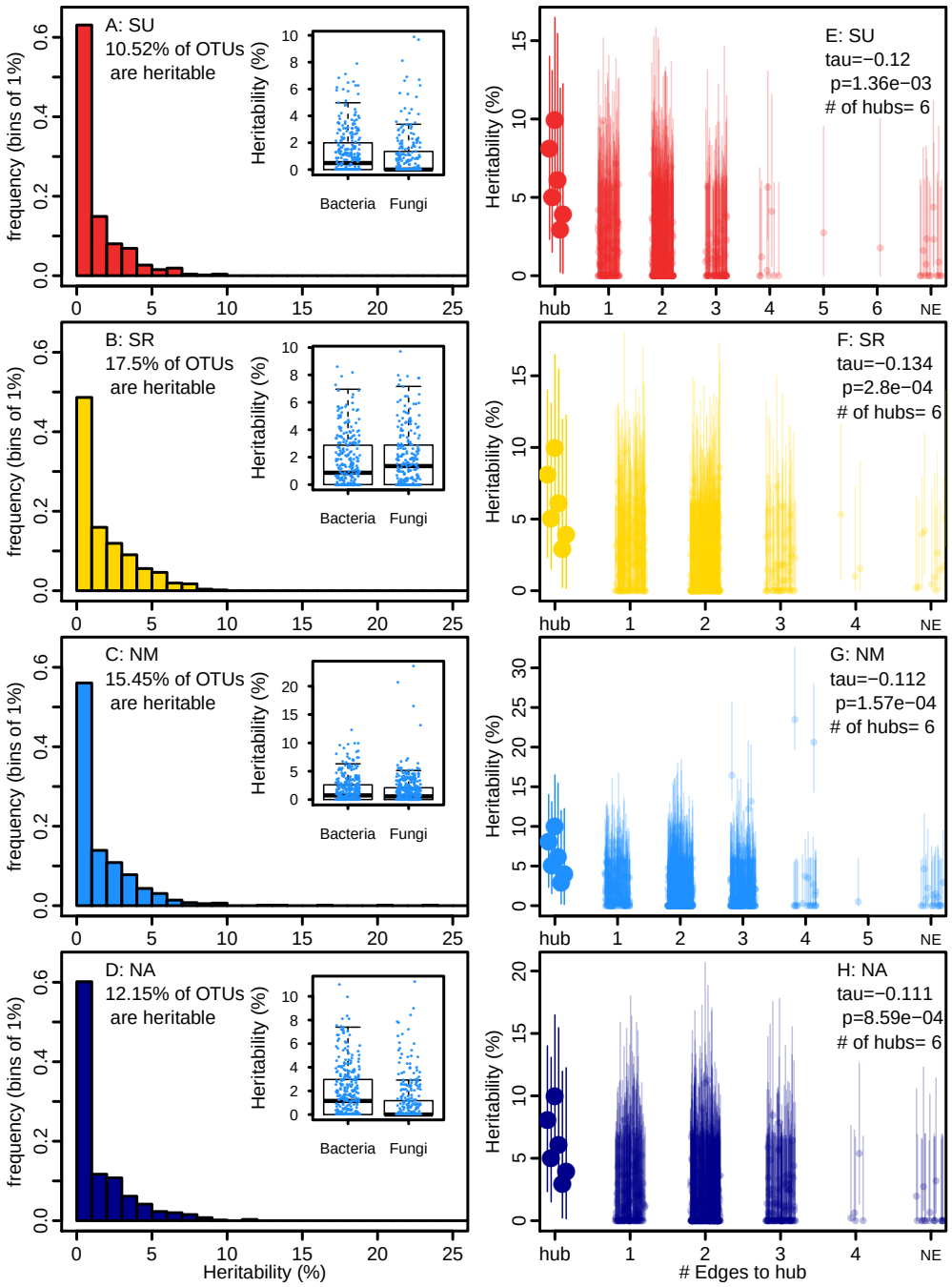


Fig. 1 | Plants grown in different environments have different microbial communities. The plots represent the projection of each sample on the plane defined by the first two constrained components of the fungal and bacterial communities, describing variation among sites and years. The percentages in parentheses are the proportion of the total inertia (square root of the Bray-Curtis dissimilarity) explained by each component. The colors of the points indicate the site from which samples were collected. Experiments from the South are represented in red (SU) and yellow (SR), and experiments from the North in light blue (NR) and dark blue (NA). All points from 2012 and 2013 are encircled by a dark and lighter grey line respectively.

Host effects on the microbiota

Our experiments provided a unique opportunity to investigate associations between host genetic variation and their resident microbiomes, within the context of environmental variation across time and space. We computed the proportion of variance explained by the host genotype (hereafter heritability or H^2) based on simple unconstrained principal coordinates (PCoA) within each experiment. Within each experiment, we found significant heritability of components of the microbial communities (Extended Data Table 1), suggesting that genetic variation in the host significantly impacts at least a fraction of the microbiota, in line with results of previous studies [8–11,19,20].

6



131

132 **Fig. 2 | The effect of host genetic variation on the microbial community targets**
133 **relatively few OTUs and percolates through hubs**. This figure corresponds to observations
134 in the set of four experiments sampled in 2013, see Extended Data Figure 3 for experiments
135 performed in 2012. **A-D:** Each frame presents the distribution of heritability estimates for
136 individual OTUs in one site. In each frame, the inset graph is a box and whiskers plot
137 contrasting the heritability (y-axis) of bacterial (B) and fungal (F) OTUs. **E-F:** The heritable
138 hubs are represented by large dots, at a distance of 0 (hub). The other OTUs are represented
139 by smaller dots and the x-axis represents their distance to the nearest heritable hub(s) within
140 the sparse covariance networks. The number of heritable hubs detected in each experiment is
141 indicated in the legend. The correlation coefficients presented are Kendall rank correlations
142 calculated for OTUs with a distance to the heritable hub(s) above 0. NE stands for “no edge”.

7

Significant heritability of principal components could arise from host genotypes exerting weak control over a large number of community members, or by targeting a few microbes that then influence the relative abundance of others through microbe-microbe interactions. Random-effects linear modeling of log-transformed OTU counts revealed significant genotypic effects (with the 95% confidence interval of heritability not overlapping 0) for between 10.52 and 22.65% of all OTUs, depending on the site and year (Fig. 2A-D and Extended Data Fig. 2A-D). Thus, the influence of the host appears focused on relatively few OTUs. We found no evidence that either fungal or bacterial communities are systematically more impacted by host effects than the other (Fig. 2A-D and Extended Data Fig. 2A-D), nor that mean relative abundance was strongly correlated with OTU heritability (Extended Data Fig. 3).

Having found that host effects are concentrated on a small proportion of OTUs, we investigated the possibility that these heritable OTUs trigger a broader community level change in the microbiota. First, we computed networks of microbe co-occurrence for each experiment. We explored the ecological importance of heritable OTUs by computing networks of microbe co-occurrence for each experiment using the SPIEC-EASI pipeline [21]. Although our networks included both fungal and bacterial OTUs, most microbe-microbe interactions occurred within each domain, with an average of only 8.71 [min=6.94, max=10.38]% of edges connecting fungal and bacterial OTUs. We quantified the ecological importance of OTUs using two common characteristics of nodes in a network (“Degree” and “Between-ness centrality”) [11], defining ecologically important “hubs” in each network as OTUs in the 95% tail of both of these statistics (Extended data Fig. 4). We identified on average 16.5 microbial hubs per experiment (ranging from 11 to 24), representing 77 unique OTUs across all eight experiments. These hubs were connected to an average of 20.09 [min=14.50, max=25.23]% of the edges in the networks, indicating that they are likely

8

important in structuring the microbial community. In addition, hubs were involved in proportionally more interactions between fungi and bacteria than the rest of the community (Extended Data Table 3).

Next, we asked whether heritable OTUs are more likely to be ecological hubs, because this could open the door to community-level impacts. Across all eight experiments, we detected 23 OTUs that were both heritable and hubs at least once (Extended Data Table 2, Supplementary Table 2). This represents a significant enrichment of hub OTUs amongst heritable OTUs (Wilcoxon rank sum test: N=8, W=57, *p*-value= 0.00699), suggesting that host effects on the microbiota preferentially influence the relative abundance of ecologically important microbes. In fact, hub OTUs were often among the OTUs with the highest heritability within each experiment.

Finally, we sought evidence of community level impacts of heritable hubs by mapping heritability onto the ecological network. In six out of eight experiments, we observed a significant negative relationship between heritability and the distance (number of network edges) to the nearest heritable hub (combined *p*-value=$4.104e^{-15}$, using Fisher's method for combining *p*-values)[22](Fig. 2E-H and Extended Data Fig. 2E-H). This suggests that host genetic variation most strongly affects a few microbial hubs that then influence other microbes, most likely through microbe-microbe interactions.

Not only do heritable hubs have an impact that appears to percolate through the microbial community, they tend to be widely distributed among accessions, sites and years. We were able to identify 278 fungal and bacterial OTUs that were found in at least 50% of samples in all experiments. Interestingly, OTUs that were heritable hubs at least once were over-represented in this core microbiota (χ^2=34.68, df=1, *p*-value=3.891e-9). This was not an artifact of their being widespread; significant heritability estimates were detected across the entire range of prevalence, with prevalence of an OTU explaining less than 1.4% of OTU

9

193 heritability across all experiments (F-statistic=29.48, df=4176, *p*-value=5.964e-08, Extended
194 Data Figure 5). Thus, ecologically important OTUs with greatest associations to host
195 genotypes were unusual in being widespread among plants in multiple experiments. Host
196 effects on the fungal OTU #8 (hereafter F8) are especially important; this OTU was heritable
197 in five out of the seven experiments in which it was a hub (Extended Data Table 2),
198 suggesting that natural variation in *A. thaliana* influences its microbiota with some
199 consistency across environments. The widespread prevalence of these heritable hubs suggests
200 that variation at particular host genes associate with particular hubs across time and space,
201 potentially providing a means to impact the microbiota in a robust fashion.

202 **Variation in performance of host genotypes explained by their influence on microbial**
203 **hubs**

204 The extent to which natural variation among host genotypes in their associated
205 microbes translates into fitness differences has yet to be determined. Our experiments
206 included additional replicates of all genotypes that were left to flower and mature in the field.
207 We harvested mature stems in early summer and used high-throughput image analysis to
208 estimate lifetime seed production (LSP) from mature stem size, using an independently
209 validated method (Extended Data Fig. 6) [23]. We observed that plant LSP estimates were
210 positively correlated across experiments (Extended Data Fig. 7), suggesting fitness variation
211 among accessions was relatively consistent across sites. We therefore asked whether host
212 effects on microbial hubs contributed to some genotypes producing more seeds across all
213 environments investigated. Specifically, we used random intercept models to estimate
214 genotype effects on both heritable microbial hubs and LSP in a series of analyses that jointly
215 considered all eight experiments and investigated the relationship between these two effects
216 (see methods "Heritable hubs and LSP across environments").

We found that the host genotype explained, on average, 6.88% (with a 95% confidence interval [5.52, 8.34]) of our estimate of plant LSP. Host genotype effects (blups) on the relative abundances of 19 of our 23 heritable microbial hubs were similarly modest, explaining up to 4% of the variation (Fig. 3A, four heritable hubs were not detected in more than 2 experiments and were removed for this analysis). In order to estimate genetic correlations between host genotype effects on LSP and on microbial hubs, we performed a multiple regression. After using model selection to identify significant relationships, we detected positive correlations between accession effects on LSP and accession effects on three heritable hubs, F8, B38 and B13, as well as a negative correlation between accession effects on LSP and accession effects on F5 (Fig. 3B). The variation explained by host genotype on the relative abundances of microbial hubs explained 12.4% of the host genotype effects on LSP.

These results reveal that a sizable percentage of genetic variance in LSP is shared with genetic variation associated with the relative abundance of a few broadly distributed microbial hubs, consistent with a causal relationship between genotype and LSP mediated by heritable microbial hubs. Of course, the proportion of shared genetic variation between LSP and heritable microbial hubs is unlikely to be equally important across time and space. In fact, in analyses performed on an experiment-by-experiment basis, we found that relationships between host effects on hubs and LSP were stronger in southern Sweden, where we detected significant relationships in both sites and both years (Extended data Table 4).

Overall, our results highlight the importance for plants to control their leaf microbial community and suggest that breeding plants for their effects on specific members of microbial communities has the potential to significantly increase plant productivity.

11

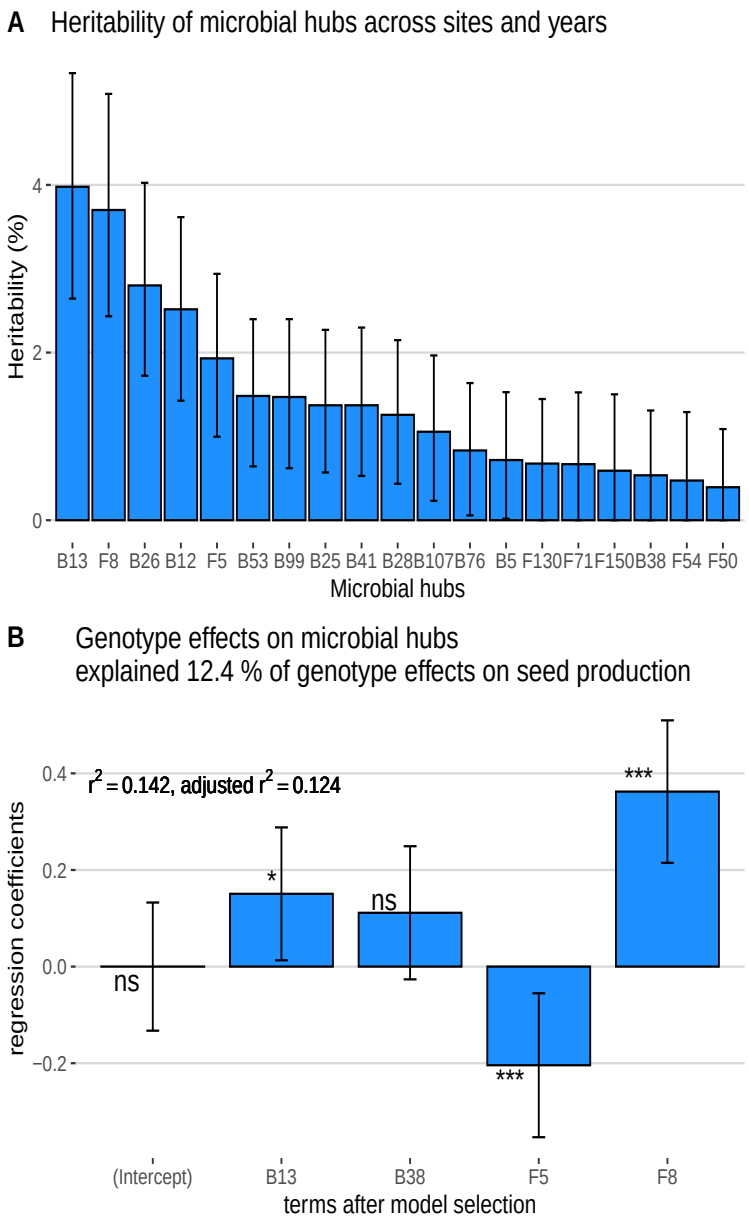


Fig. 3 | Relationship between host genotype seed production and influence on microbial hubs across sites and years. A. Proportion of heritable hub relative counts explained by host effects across all sites and years. **B**. Coefficients for the linear regression explaining lifetime see production variation among accession with accession effects on microbial hubs across experiments (after model selection).

Effect of hubs on growth in controlled condition

In an effort to verify the correlations between host performance and the relative abundance of microbial hubs, we returned to the field to collect wild *A. thaliana* leaves [24], cultured approximately 2400 microbial isolates from within these leaves, and sequenced both the 16S

12

RNA gene and gyrase-B. Among heritable hubs, only B38 was successfully cultured; this isolate derived from Vårhallarna, in southern Sweden, and was identified by a 100% match in 16S sequencing (Extended data Table 5). We subsequently performed shotgun whole genome sequencing of B38 which we identified as *Brevundimonas sp*. The assembled and annotated genome did not identify putative pathogenic or virulence genetic factors present in the genome.

To test the effects of B38 on host growth, we grew Arabidopsis plants of an accession (#6136) from the South of Sweden chosen to have intermediate relative abundance of B38 in the field. Plants were grown under sterile conditions in ½ MS media under long day conditions in the growth chamber, with and without B38 inoculation. Approximately two weeks after germination, over 600 plants were randomly selected for either drip inoculation with the control or B38 inoculum, and measured for surface area growth over the following two weeks. Accounting for variation in plant growth among trials and plates within trials, we found that plants treated with B38 grew 5.375 (standard error=1.973) mm^2 larger than control plants (F=7.3981, df=1, p-value=$6.7e^{-3}$) between day 7 and 14, corresponding to a 10.22% growth increase.

The microbial hubs could in principle influence host fitness directly, for example by contributing to growth, or indirectly through their influence on other beneficial members of the microbial community [25]. Here we show that B38 directly improves host growth over early life stages in isolation from the rest of the microbial community. This result is consistent with our field observations, where we found a positive correlation between genetic variation associated with B38 and with LSP, suggesting that in this instance the correlation is causative. The possibility of additional indirect interactions in the field cannot, of course, be excluded.

13

Mapping the genetic bases of consistent variation in the relative abundances of microbial hubs across experiments

Our observation that host control of the relative abundances of four microbial hubs explains ~12% of variation in LSP among Arabidopsis genotypes grown in 8 field trials suggests the potential to reveal host genes that can enhance plant performance in the presence of microbes, particularly across environments. Towards this end, we performed genome-wide association mapping for host genotype effects on microbial hubs (N=19) and LSP across all experiments. Despite significant differences among accessions, GWAs yielded few peaks with *p*-values below accepted significance thresholds after correction for multiple testing. Specifically, we found only two significant associations, both for microbial hub B41. The first is located on chromosome 1 at position 29909876 in AT1G79510 annotated as a pseudo-gene. The second is on chromosome 4 on positions 15704377, 15704472 and 15704478. These consecutive SNPs are located between *YUC-1* (AT4G32540), involved in auxin biosynthesis, and *LEUNIG* (AT4G32551), involved in the development of the leaf blade and floral organs.

A potentially more powerful strategy to detect minor QTL involves computing local association scores along the genome. The assumption underlying this method is that neighboring markers in linkage disequilibrium with causal mutations will also carry association signals; thus, aggregating *p*-values increases power [26]. This method identified 340 non-overlapping loci (hereafter QTLs), with sizes ranging from 93 to 150,926 bp including a total of 25,529 SNPs. Out of the 340 QTLs, only 27 included SNPs associated with multiple traits (Supplementary Table 3), suggesting a modest level of pleiotropy.

To investigate functions underlying these associations, we tested pathway and GO term enrichment (Biological processes only)[27,28]. Using a combination of methods accounting for multiple testing, overlapping gene lists, and the potential aggregation of

14

functions and associations along the genome [29–32], we identified 29 enriched GO terms related to biological processes across 16 traits (Supplementary Table 4 and 5), including four genes involved in the response to virus (GO:0009615) and nematodes (GO:0009624), hypersensitive response (GO:0009626) and response to chitin (GO:0010200), all of which are related to interactions with other organisms. Three enriched GO terms directly concern auxins and their transport (GO:0009926, GO:0010540, GO:0009734); auxins have previously been documented to contribute to shaping plant interactions with beneficial bacteria [33,34]. Specialized metabolites also appear involved in shaping the relative abundance of microbial hubs. Indeed hub B107 is associated with genes in the geranylgeranyl diphosphate metabolism (GO:0033385), the universal precursor of monoterpenes, which are volatile compounds with anti-microbial properties, [see 35 for a review] that potentially shape within rosettes microbial communities. In addition, loci associated with B76 are enriched in genes related to specialized metabolite biosynthesis (GO:0044550) and genes involved in the synthesis of sinapoyl glucose and sinapoyl malate (PWY-3301), an intermediate in the synthesis of phenylpropanoids. Genes involved in the synthesis of glucosinolates from phenylalanine (#11 Bz [36] aka glucotropaeolin, PWY-2821) and hexahomomethionine (specifically #69 mSOo [36] aka 8-(methylsulfinyl)octyl-glucosinolate PWYQT-4475) are also enriched in loci associated with B5 and F71, respectively.

The functions highlighted by our analysis are in line with other studies suggesting the involvement of specialized metabolites, auxins and the immune system in influencing the leaf microbial communities [37,38]. Our analysis also highlights less obvious players, like growth lipid metabolism and brassinosteroids (Supplementary Table 5). This is especially true with regard to beneficial members of the community. For example, loci associated with the relative abundance of the beneficial microbial hub B38 are enriched for transition metal

15

ion transport (GO:0000041), response to carbohydrates (GO:0009743), and fatty acid biosynthesis (PWY-4381).

Plant specialized metabolites correlated with microbial hub abundance

Our biological processes and pathway enrichment analysis suggest that specialized metabolites are involved in shaping microbial hubs. To support this result, we quantified 20 abundant compounds using untargeted metabolomics in a subset of the field samples in which we characterized the rosette microbiome. We found that the relative abundance of 14 out of 19 hubs were significantly correlated with at least one of 11 specialized metabolites (after correction for multiple testing), six of which displayed significant heritability in the field across sites ranging from 1 to 38% (Extended Data Fig. 8A & B).

The molecule known as #69 mSOo (here 260_GSL_8MSO) displayed the strongest relationship with multiple microbial hubs in the field (Extended Data Fig. 8A, Extended Data Table 6), as well as significant heritability under field conditions (Extended Data Fig. 8B). However, the variation among accessions of this abundant glucosinolate was less evident in the greenhouse and in sterile conditions (Extended Data Fig. 8B), leaving open the possibility that the correlation is induced by one or more of the microbial hubs. In contrast, other molecules significantly related with the abundance of microbial hubs in the field across experiments (354_C_Cy-GRGF_785 and 358_F_R-K-R_577, Extended Data Table 6) are heritable in all conditions, and variation among accessions in the field is positively correlated with the variation among accessions in the greenhouse. This suggests that these flavonoids are constitutively and consistently produced by accessions and influence microbial hubs in a manner that is robust to heterogeneity among field experiments.

Conclusion

In this study, we show that not only does host genetic variation influence the microbiome, but it does so in consistent ways. Host genotype effects are centered on

16

ecologically important hub species, and percolate through the microbial community, most likely as a result of microbe-microbe interactions. Our replicate field experiments were likely instrumental in allowing us to reveal consistent host effects on the leaf microbiome via common and widespread hub species.

Furthermore, we found that the influence of host genetics on a handful of prevalent microbial hubs has a far-reaching impact on the community, associated with a substantial fraction of the variation in our fitness estimates among accessions. Although these relationships are correlational, we were able to culture one of the identified hubs and confirm a direct positive effect on host fitness experimentally.

Understanding how host performance or fitness components are influenced by their ability to shape microbial communities could provide a basis for breeding crops favoring microbes that are beneficial both to growth and resistance to pathogens. We successfully mapped variation in host microbe interactions using genome-wide association, and our results suggest that natural and artificial selection can act on plant traits such as leaf specialized metabolites, auxins and the immune system to improve plant performances through effects on microbial communities [39,40]. In addition, we found that at least some plant metabolites are expressed in a consistent manner that is robust to variation among our experiments and correlates with the relative abundance of microbial hubs. Our results therefore suggest that ongoing efforts to harness the microbiome for agricultural purposes can be successful and highlight the value of explicitly considering abiotic variation in those efforts.

17

Methods:

Field experiments

This study uses a set of 200 diverse accessions (inbred lines, Supplementary Table 1) that were previously re-sequenced [14]. The seeds were produced simultaneously in the greenhouse of the University of Chicago under long day conditions, except for a 12-week vernalization period at 4˚C, required to induce flowering. The seeds for the common garden experiments were cold stratified in water at 4˚C for 3 days before being planted in trays of 66 open-bottom wells, each measuring 4 cm in diameter. The soil used was a 90:10 mix of standard greenhouse soil and soil from each of the four sites in which the experiments were installed:

- SU: Ullstorp (Agricultural field, lat: 56.067, long: 13.945)
- SR: Ratchkegården (Agricultural field, lat: 55.906, long: 14.260)
- NM: Ramsta (Agricultural field, lat: 62.85, long: 18.193)
- NA: Ådal (South facing slope, lat: 62.862, long 18.331)

Each experiment included 3 complete randomized blocks including 8 replicates per accession. Experiments were sown in pairs (2 in the North and 2 in the South) over 6 days, corresponding to the sowing of one block a day, alternating between the 2 experiments (between August 7th and 12th in the North, and between August 31st and September 5th in the South). The trays were placed in a common garden the morning after sowing under row tunnels to avoid disturbance by precipitation and to favor germination (on the campus of Mid Sweden University and Lund University, in the North and in the South, respectively). Trays were watered as needed and missing seedlings were transplanted between cells within blocks and then thinned to one per cell after 9 days. Seventeen days after sowing, trays were laid in the field in their final location over tilled soil. For each experiment, the blocks were laid

18

399 across the most obvious environmental gradient (exposition, shading, slope, soil humidity…).
400 The pierced bottom of the cells allowed the roots to grow through and reach the soil, as was
401 verified upon harvest. The same protocol was followed in 2011 and 2012.

402 **Sample collection and processing**

403 The rosettes used to characterize the microbial community were harvested in the
404 spring of 2012 and 2013 only a few days after the plants were exposed, following snow melt.
405 We harvested 2 randomly selected replicates per accession in each experimental block. Upon
406 harvest, the roots were removed and the rosettes were washed twice in successive baths of TE
407 and 70% ethanol to remove loosely attached microbes from the leaf surface. The rosettes
408 were then placed in sealed paper envelopes and placed on dry ice. The rosettes were kept at -
409 80˚C until lyophilized. Freeze-dried rosettes were then transferred to 2 ml tubes along with 3
410 2mm silica beads. For 2 successive years, the tubes were randomized and separated in 34 and
411 46 sets of 96 tubes, respectively. Our randomization strategy maintained approximately the
412 same number of tubes from each of the 12 experimental units (3 blocks in 4 experiments) in
413 order to avoid confounding biologically meaningful effects. We powdered the samples using
414 a Geno/Grinder® (from Spex SamplePrep, USA, NJ) for 1min at 1750rpm, before
415 transferring 10 - 20 mg to 2ml 96-well plates, along with two zirconia/silica beads (diameter
416 = 2.3mm), for DNA extraction.

417 **DNA extraction**

418 DNA extraction started with 2 enzymatic digestions to maximize yield from Gram-
419 negative bacteria [41]. First, we added 250µl of TES with 50 units.μl^{-1} of Lysozyme (Ready-
420 Lys Lysozyme, Epicenter) to each well. The plates were then shaken using the Geno-Grinder
421 for 2 min at 1750 rpm, briefly spun and incubated 30 min at room temperature. Second, we
422 added 250µl of TES with 2% SDS and 1 $mg.mL^{-1}$ of proteinase K. The plates were then

briefly vortexed and incubated at 55˚C for 4 hours. The protocol then followed [42], adapted to the 96-well plate format and automated pipetting on a Tecan Freedom Evo Liquid Handler. We added 500 µl of Chloroform:Isoamyl Alcohol (24:1), pipette mixed, and centrifuged the plates at 6600 g for 15 min. We transferred 450 µl of the aqueous supernatant to a new plate containing 500µl of 100% isopropanol. The plates were then sealed, inverted 50 times, incubated at -20˚C for 1 hour, and centrifuged at 6600 g for 15 min. The Isopropanol was then removed and the pellets were washed twice with 500 µl of 70% Ethanol, dried and re-suspended in 100 µl of TE. After 5 min incubation on ice, the plates were centrifuged 12 min at 6600 g and the supernatant was pipetted into a new plate.

PCR and Sequencing

To describe the microbial communities, we amplified and sequenced fragments of the taxonomically informative genes *16S* and *ITS* for bacteria and fungi, respectively. For bacteria we amplified the hypervariable regions V5, V6 and V7 of the *16S* gene using the primers 799F (5'-AACMGGATTAGATACCCKG-3') and 1193R (5'-ACGTCATCCCCACCTTCC-3') [9,43]. For fungi, we amplified the ITS-1 region using the primers ITS1F (5'-CTTGGTCATTTAGAGGAAGTAA-3') [15] and ITS2 (5'-GCTGCGTTCTTCATCGATGC-3') [44]. To the 5' end of these primers we added a 2bp linker, a 10bp pad region, a 6bp barcode and the adapter to the Illumina flowcell, following [45]. The appropriate linkers were chosen using the PrimerProspector program [46]. The PCR reactions were realized in 25 µl including: 10 µl of Hot Start Master Mix 2.5x (5prime), 1µl of a 1/10 dilution of the DNA template, 4µl of SBT-PAR buffer, and 5 µl of the forward and reverse primers (1 µM). The SBT-PAR buffer is a modified version of the TBT-PAR PCR buffer described in [47] with the trehalose replaced by sucrose (Sucrose, BSA, Tween20). The PCR program consisted of an initial denaturing step at 94˚C for 2'30", followed by 35 cycles of a denaturing step (94˚C for 30"), an annealing step (54.3˚C for 40"), and an

20

448 extension step (68˚C for 40"). A final extension step at 68˚C was performed for 7' before
449 storing the samples at 4˚C. For each plate, the PCRs were performed in triplicates, pooled,
450 and purified using 90 µl of a magnetic bead solution prepared and used following [48]. The
451 purified PCR products were quantified with Picogreen following the manufacturer's
452 instruction [49] and pooled into an equimolar mix. Between 5 and 7 plates (480 to 672
453 samples) were pooled in each MiSeq run. If the bioanalyzer traces for pooled libraries
454 showed only one dominant peak, they were sequenced directly following the standard MiSeq
455 library preparation protocols for amplicons. In cases where the bioanalyzer trace presented
456 peaks for smaller fragments (remaining primers, primer dimers, small PCR products), the
457 libraries were first concentrated 20X on a speedvac (55˚C for 2 to 3 hours), purified with 0.9
458 volume of magnetic bead solution, and/or size selected using a Blue Pippin (range mode
459 between 300 and 800 bp).

460 The sequencing was performed using MiSeq 500 cycle V2 kits (251 cycles per read
461 and 6 cycles of index reads twice), using a loading concentration of 12.5pM for *ITS*
462 fragments and 8pM for *16S* fragments following the standard Illumina protocol. Sequencing
463 primers were designed and spiked in following [45]. The sequencing primer for the first read
464 of *16S* fragments was prolonged into the conserved beginning of the fragment amplified to
465 reach a sufficient melting temperature. This primer modification produced no change in the
466 Blast results of the primers against the GreenGene database. A total of 11 sequencing runs
467 were performed for each of the fungal and bacterial communities.

468

469 **Sequence processing and clustering**

470 The demultiplexed fastq files generated by MiSeq reporter for the first read of each
471 run were quality filtered and truncated to remove potential primer sequences and low quality
472 basecalls using the program cutadapt [50]. The reads were then further filtered and converted

473 to fasta files using the FASTX-Toolkit (-q 30 -p 90 -Q33). The fasta files for each run were
474 then de-replicated using AWK code provided in the swarm git repository
475 (https://github.com/torognes/swarm)[16]. The resulting de-replicated fasta files were filtered
476 for PCR chimeras using the vsearch uchime_denovo command
477 (https://github.com/torognes/vsearch). The de-replicated fasta files for each run were then
478 combined and further de-replicated at the study level. The fasta files were then used as input
479 for OTU clustering using swarm (-t 4 -c 20000). The clustering identified 150,412 and
480 251,065 OTUs for the fungal and bacterial communities, respectively. The output files were
481 combined into two separate community matrices using a custom python script (available at
482 https://bitbucket.org/bbrachi/microbiota.git). The taxonomy of each OTU was determined
483 using the quiime2 2019.1 v8 feature classifier trained on the UNITE V6 and SILVA 119
484 database for Bacteria and Fungi, respectively [51,52].
485
486 **Count table filtering**
487 The count tables obtained for both the bacterial and fungal communities were filtered in
488 successive steps by removing:

489 1) samples corresponding to empty wells and additional plant genotypes present in the
490 experiments sampled by mistake (leaving 7476 and 7240 samples for the fungal and
491 bacterial count tables, respectively).
492 2) samples with less than 1000 reads (leaving 6678 and 6819 samples for the fungal and
493 bacterial count tables, respectively)
494 3) OTUs represented by less than 10 read in 5 samples (leaving 1381 and 993 OTUs for
495 the fungal and bacterial count tables, respectively)
496 4) for the bacterial community, OTUs assigned to plant mitochondria (leaving 993 OTUs
497 in the bacterial count table, no OTUs assigned to plant mitochondria)

5) for a second time, samples with less than 1000 reads (leaving 6656 and 6783 samples for the fungal and bacterial count tables, respectively).

The final count tables used in the study included 993 OTUs and 6783 samples for the bacterial communities and 1381 OTUs and 6656 samples for the fungal community.

Differentiation of the microbial communities among sites and years

This analysis was performed for the fungal and bacterial communities independently, including all samples and only OTUs with read counts above 0.01% of total read counts (after the filtering described above) across sites and years. To investigate how the microbial communities differed among sites and years, we performed a constrained ordination on log transformed read counts using the capscale function in the R-package Vegan [53] and following [54]. The log transformation offers the advantage of removing large differences in scale among variables. The capscale function performs canonical analysis of principal coordinates, an analysis similar to redundancy analysis (rda), but based on the decomposition of a Bray-Curtis dissimilarity matrix among samples (instead of euclidean distance in the case of rda). This allows identification of the dimension that maximized the variance explained by components, while discriminating groups of samples, here sites and years [54].

Core microbiota

In order to define a core microbiota, we counted, for each OTU, the number of site/year combinations in which it was prevalent. We defined "prevalent" as being present in at least 50% of the samples in a given site/year. We performed this analysis using count tables for each experiment with the filtering described in the previous paragraph. Therefore, for an OTU to be designated as a member of the core microbiota, it needed to have non-zero counts in more than 50% of the samples within each site/year combinations and, due to

23

previously described filtering, needed to be represented by at least 10 reads in 5 of those samples across all site/year combinations (see "Count table filtering").

Heritability of the microbiota

In this analysis, count tables were split per site and year before filtering for OTUs represented by more than 0.01% of the reads (after the filtering described in the section "Count table filtering") for each of the bacterial and fungal communities. The resulting 16 count tables were normalized to 1000 reads per sample and used to calculate 16 Bray-Curtis pairwise dissimilarity matrices among samples. These matrices were then decomposed into 10 principal coordinates. For each component we estimated broad sense heritability (hereafter H^2), *i.e.* the proportion of variance explained by a random intercept effect capturing the identity of the accessions present in the experiment (plate effects had limited impact on H^2 estimates but were included in the models). Mixed models were fitted using the function lmer in the lme4 R package [55]. We computed 95% confidence intervals using 1000 bootstraps, and components were considered to have significant H^2 when their confidence intervals did not overlap 0 (lower bound of the confidence interval ≥ 0.01).

Heritability of individual OTUs

This analysis was also performed per site, year and community, as in the microbiota H^2 estimation analysis. In this analysis, counts were transformed to centered log-ratios using a dedicated function in the R package mixOmics [56,57]. H^2 estimates and confidence intervals were computed for individual OTUs using the method described in the previous paragraph (without the plate effect). H^2 estimates for our estimate of LSP (see below) were estimated the same way using a box-cox transformation.

Microbe–microbe interaction networks

24

Microbe-microbe interaction networks were computed for the fungal and bacterial communities together, using the count tables per site/ year and filtering OTUs represented by less than 0.01% of the reads within each community. The count tables were then combined into the same table and analyzed using the SPIEC-EASI (v1.1) pipeline [21]. This method computes sparse microbial ecological networks in a fashion robust to compositional bias and uses conditional independence to identify true ecological interactions, meaning that a connection between 2 OTUs will be significant when one provides information about the other, given the state of all other OTUs in the network. This means that covariance among OTUs induced by micro-environmental and host genetic variation is controlled. SPIEC-EASI was run using the neighborhood selection framework and model selection was regularized with parameters set to a minimum lambda ratio of $1e^{-2}$ and a sequence of 50 lambda values (see documentation for SPIEC-EASI and the huge R package, which provides regularization functions)[58] .

Network statistics

The inferences of microbe-microbe ecological interactions inferred using SPIEC-EASI were passed to the igraph package [59], which was used for enforcing simplicity of graphs (no loops or duplicated edges), computing degree and betweenness centrality of vertices, computing distances between vertices, and plotting. Within each of the 8 networks thus computed, hubs were defined as OTUs with degree and betweenness centrality both in the 5% tail of their respective distributions. We then checked the overlap between heritable OTUs and hubs, and the over-representation of heritable OTUs among hubs was tested using a simple χ^2 test across all site/year combinations. The relationship between distances to heritable hubs (OTUs that are both hubs and have significant H^2) and heritability was investigated using Spearman's rank correlation coefficient. Distances were calculated as the number of edges between OTUs and the closest heritable hub in the network. OTUs not

25

571 connected to heritable hubs were assigned a distance equal to one more than the maximum
572 distance observed for OTUs connected to heritable hubs.

573 **Estimation of seed production**

574 The experiments each included 8 replicates per block per accession (24 replicates per
575 experiment). While we harvested 2 replicates per block (6 replicates per experiment) for
576 microbiota analysis, the remaining plants were left to grow, flower and produce seeds in the
577 field. We harvested the mature stems of all remaining plants at the end of the spring, when all
578 plants had finished flowering and siliques were mature, and stored them flat in individual
579 paper envelopes. We estimated lifetime see production (LSP) by the size of the mature stems.
580 After removing remaining traces of roots and rosettes, each mature plant was photographed
581 on a black background, using a DSLR camera (Nikon 60D) mounted on a copy-stand and
582 equipped with a 60mm macro lens (Nikon 60mm). The photographs were segmented (using
583 custom scripts in R based on the EBimage package [60] to isolate plants from the image
584 background and estimate the total surface of the image they occupied.

585 We validated this method with mature plants harvested from a previous experiment
586 that was planted in NM in fall 2010, and that included the 200 accessions used in this study.
587 We counted siliques and estimated the average silique size for 1607 mature stems that were
588 also photographed. The total silique length produced per plant (number * average size) was
589 highly correlated with our size estimates based on image analysis (Spearman's rho=0.84) and
590 displayed a clear linear relationship.

591 **Relationship between host effects on microbial hubs and fecundity**

592 To investigate the relationship between host genotype effects on heritable hubs and
593 LSP in each experiment, we computed estimates of accession effects (Best unbiased linear
594 predictors or BLUPs) for both log-ratio transformed heritable hubs and box-cox transformed
595 LSP estimates. We then fitted multiple regressions for each site/year combination aiming to

explain LSP variation among accessions with their influence over microbial hubs and following **eq. 1**.

$$f_i \sim \sum_{j=1}^{n}[(\beta_j.h_{ij}) + (\gamma_j.h_{ij}^2)] + \varepsilon_i \quad \textbf{(eq.1)}$$

where f_i is the LSP estimate of the ith accession (blup), h_{ij} is the effect of the ith accession on the jth hub. β_j is the regression coefficient for the jth hub (h_j) and γ_j is the regression coefficient for the jth hub squared. $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ captures residual variance per accession. We then performed forward/backward model selection to obtain the final models presented in (Extended Data Table 4).

Heritable hubs and LSP across environments

We next investigated host effects on heritable hubs and LSP across all 8 experiments. Similarly to previous analyses, count tables were split per site and year before filtering for OTUs represented by more than 0.01% of the reads (after the filtering described in the section "Count table filtering") for each of the bacterial and fungal communities. The resulting 16 count tables were then combined into one before fitting a mixed-model following eq. 2:

$$Y_{ijk} \sim \beta.exp_j + a_k + \varepsilon_i \quad \textbf{(eq. 2)}$$

where Y_{ijk} are the transformed counts for a heritable hub measured the ith time in experiment j (exp_j) (N=8, four sites and two years) and for accession k (N=200), β is the vector of fixed experiment effects (N=8) and $a_k \sim \mathcal{N}(0, \sigma_a^2)$ is a random intercept estimated by restricted maximum likelihood for each accession. $\varepsilon_i \sim \mathcal{N}(0, \sigma_e^2)$ captures the residual variance.

Microbial hub heritability (H^2) across experiments was estimated as the percentage of variance explained by the random accession intercept:

27

$$H^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}.$$

LSP data was analyzed the same way, except we performed Box-Cox transformation of the data. The lambda parameter for the Box-Cox transformation was estimated using the same model, but without the random accession term.

For both heritable microbial hubs and LSP, we retrieved random intercept accession effects (BLUPS) and fitted a multiple linear regression following:

$$F_i \sim \sum_{j=1}^{n} [(\beta_j . H_{ij}) + (\gamma_j . H_{ij}^2)] + \varepsilon_i \qquad \textbf{(eq. 3)}$$

where F_i is the effect of the ith accession (N=200) on LSP (across all experiments) , H_{ij} is the effect of accession i on hub j across all experiments, H_{ij}^2 is the squared effect of accession i on hub j. β_j and γ_j are the corresponding regression coefficient for hub j and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ captures the residual variance per accession. The final model was obtained after backward/forward model selection based on AIC.

Isolation, culture and identification of microbial hubs

Bacteria sampling from wild A. thaliana plants: We collected 2 leaves from 10 plants at 5 locations in Sweden (Extended data Table 5) . The leaves were first cleaned by rinsing individually in ddH2O, and subsequently surface-sterilized by dipping 70% EtOH for 3-5 seconds. The leaves were ground in individual 1.5 mL tubes. The leaf material was stored in 20% glycerol at -20ºC. Wild *A. thaliana* microbial isolates were collected using modified methods that were previously described (Bai et al 2015). Briefly, the leaf and glycerol mixture was plated on nine distinct media, including; R2A, Minimal media containing Methanol, Tryptic Soy Agar, Tryptone Yeast extract Glucose Agar, Yeast Extract Manitol Agar [24]; 0.1 Tryptic Soy Agar [61]; Potato Dextrose Agar, 0.2 Potato Dextrose Agar, and

28

Malt Extract Agar [62,63]. Colonies were picked over the next 14 days, restreaked, and grown in liquid media in an orbital shaker for 1-4 days. A portion of the inoculum was saved in 15-20% glycerol, and the rest of the liquid culture was pelleted by centrifugation and decanted for DNA extraction. We performed a double enzymatic digest for all isolates, which was performed using the Tecan: 30 minute incubation with 350 U Ready-Lyse Lysozyme and 245 U RNase A (QIAGEN, Germantown, MD) in 250ul TES (10 mM Tris-HCl pH ~8, 1 mM EDTA, 100 mM NaCl), followed by the addition of 2 mg/mL Proteinase K in 250ul TES + 2% SDS and a 4-6 hour incubation at 55C. The SDS-protein complexes were precipitated with .3 volume 5M NaCl and pelleted by a brief centrifugation. The clear supernatant was pipetted into a clean plate, and a standard .5 volume SPRI bead DNA extraction was performed with 2x 70% EtOH washes. Clean DNA was resuspended into MilliQ water. The samples were then amplified for 16S sequencing using the same primers binding regions as previously, 799F and 1193R, and sequenced by either Sanger or Illumina MiSeq (PE 300). Illumina adapters were designed and generated as described by Illumina with internal barcodes to increase sample count capacity per lane [64]. Isolate B38 was identified by 100% match to the B38 representative sequence from the previous analysis.

B38 whole genome assembly

We used a low-input method for Illumina library prep [Baym]. Briefly, ~2 ng extracted DNA was used in a reduced volume (5ul) tagmentation reaction with TDE1 (incubate 55C for 10 mins, room temperature for 5 mins). The tagmentation reaction was added to a 15 ul PCR reaction, adding the Illumina adapters (Kapa HiFi Hotstart PCR kit KK502, standard Illumina adapters and cycling). The library was cleaned with .8x volume SPRI beads, quantified on the Bioanalyzer, and run on the MIseq2500 using paired end 300 chemistry. Reads were trimmed for adapters (BBDuk, ktrim=r k=23 mink=11 hdist=1 tbo)

29

and quality across a sliding window (k=4, trimq=20). Reads were assembled using SPAdes (-isolate) and annotated with PROKKA. **Plant growths assays with B38**

Plant growth: *Arabidopsis thaliana* accession 6136 from Southern Sweden was used in the growth assays. In our field experiments it displayed average relative counts for B38 (rank 102 of 199). The plant assay used slightly modified methods as previously described [65]. The seeds were exposed to chlorine gas for sterilization: in a bell jar with dessicant, an open 1.5 mL tube with seeds was placed next to a 50 mL beaker with 40 mL Chlorox bleach and 1 mL hydrochloric acid, sealed with parafilm, and incubated for 4 hours. Sterilized seeds were subsequently sown on 24-well tissue plates containing 1.5mL of ½ MS media (Murashige & Skoog medium incl. Nitsch vitamins, bioWORLD, Dublin, Ohio) containing 500mg/L MES, pH 5.7 - 5.8. Plates were wrapped in parafilm, and vernalized in the dark at 4ºC for 4 days. The plates were individually wrapped with micropore tape to prevent environmental contamination and transferred to a growth chamber with 16 hours of light at 16ºC. The plants were treated with either B38 or control inoculum between days 13-15 post-vernalization. The plates were returned to the chamber to grow for another 14 days.

B38 inoculation: The B38 isolate grew in R2A liquid media in an orbital shaker until OD_{600}=0.2, approximately 3 days. To ensure no environmental contamination, a portion of the inoculum was saved for DNA extraction and subsequent 16S Sanger sequencing verification. The liquid cultures were pelleted by centrifuging 1800 RCF 18C for 7 minutes, decanted and resuspended in 0.1 M $MgSO_4$. The plants in each 24-well plate were randomly selected to receive the infection (B38 + 0.1 M $MgSO_4$) or control (0.1 M $MgSO_4$) treatment. Each plant was drip inoculated using pipettes with 180ul of the selected treatment. The plates were re-wrapped in micropore tape and returned to the growth chamber.

Measuring plant growth: We performed 3 trials of 11, 28, and 23 plates, totalling 62 24-well plates. Plants were not treated and removed from the experiment if they had less than

30

3 true leaves, cracked agar, or failed to germinate, resulting in a total of 1094 plants. The plants were individually photographed immediately before inoculation, then again at 7 and 14 days post-inoculation. The images were processed using a custom script employing cv2 in Python [66], which quantified plant surface area in each well by scaling based on the wells' size, converting images into binary images, and measuring non-white pixels within each well (i.e. plant surface area). The output images were manually inspected, and any images which failed to be accurately processed were manually measured using the same pipeline described above, but using Image J.

Due to the high humidity of the plates and the drip inoculation, 422 plants showed signs of water log stress. Plants were scored for symptoms of stress induced by water logging (blindly with regard to B38 inoculation) as categorized by translucent/white leaves or stunted growth, and were removed from the experiment.

We used a linear mixed model (eq. 4) accounting for variation in plant growth among trials and plates within trials to estimate the effect of B38 inoculation.

$$G_{ij} \sim \beta.T_{ij} + p_j + \varepsilon_{ij} \qquad \textbf{(eq. 4)}$$

In equation 4, G_{ij} is the growth of ith plant in the jth plate/assay combination. β is the estimate of the treatment effect compared to the controls (intercept) and T_{ij}, is the treatment (inoculation with a B38 or control solution). $p_j \sim \mathcal{N}(0, \sigma_p^2)$ the random intercept effect capturing variation among plates in assays (N=62 plates across three trials).

$\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_e^2)$ captures the residual variances.

Genome-wide association mapping

Single polymorphism calling and filtering

31

Single nucleotide polymorphisms (SNP) used in this study were generated from the sequences generated in the context of the 1001genome project [67] and published in Long, Q. *et al.* [14]. As pipelines evolve, we re-ran SNP calling to ensure optimal quality.

For each sequenced individual, we performed 3' adapter removal (either TruSeq or Nextera), quality trimming (quality 15 and 10 for 5' and 3'-ends, respectively) and N-end trimming with cutadapt (v1.9) [50]. After processing, we only kept reads of approximately half the length of the original read-length. We mapped all paired-end (PE) reads to the *A. thaliana* TAIR10 reference genome with BWA-MEM (v0.7.8) [68,69]. We used Samtools (v0.1.18) to convert file formats [70] and Sambamba (v0.6.3) to sort and index bam files [71]. We removed duplicated reads with Markduplicates from Picard (v1.101) (http://broadinstitute.github.io/picard/) and performed local realignment around indels with GATK/RealignerTargetCreator and GATK/IndelRealigner functions from GATK (v3.5) [72,73] by providing known indels from The 1001 Genomes Consortium (1001 Genomes Consortium 2016). Similarly, we conducted base quality recalibration with the functions GATK/BaseRecalibrator and GATK/PrintReads by providing known indels and SNPs from The 1001 Genomes Consortium.

For variant calling, we employed GATK/HaplotypeCaller on each sample in 'GVCF mode', followed by joint genotyping of a single cohort of 220 individuals with GATK/GenotypeGVCFs. To filter SNP variants, we followed the protocol of variant quality score recalibration (VQSR) from GATK. First, we created a set of 191,968 training variants from the intersection between the 250k SNP array [74] used to genotype the RegMap panel [75] and the SNPs from The 1001 Genomes Consortium. Second, this training set was further filtered by the behavior in the population of several annotation profiles (DP < 10686, InbreedingCoeff > -0.1, SOR < 2, FS < 10, MQ > 45, QD > 20) to leave 175,224 training high-quality variants. Third, we executed GATK/VariantRecalibrator with the latter as the

training set, an *a priori* probability of 15, the maximum number of Gaussian distributions set at 4, and annotations MQ, MQRankSum, ReadPosRankSum, FS, SOR, DP, QD and InbreedingCoeff enabled. Finally, we applied a sensitivity threshold of 99.5 with GATK/ApplyRecalibration and restricted our set to bi-allelic SNPs with GATK/SelectVariants for a total of 2,303,415 SNPs in the population.

Preparation for use in genome-wide association analysis involved further filtering of individuals and SNPs using Plink1.9 [76,77]. Individuals not included in this study were removed and SNPs with over 5% missing data and with minor allele frequencies below 5% in our collection of accessions were removed.

Phenotype preparation and association analysis

Association mapping analyses were performed for the 11 heritable microbial hubs for which we estimated host genotype effects across experiment and accession LSP estimates. Association analyses were performed using a classical one trait mixed model accounting for genetic relatedness among accessions (kinship) [78] .

In order to take advantage of linkage disequilibrium and gain power by grouping association statistics in contiguous markers, we computed local association scores [26]. We followed the instructions provided by the authors and defined the parameter X_i as the 0.999 quantile of the distribution of $-log(p-value)-1$ rounded to the closest integer for each trait investigated (19 microbial hubs and LSP). The approach highlights regions, which we call QTLs.

The null association model (without fixed SNP effect) from Gemma allows us to estimate SNP-based heritability or pseudo-heritability [79], which is the proportion of variance explained by the random accession effect, accounting for the genetic similarity among accessions. To investigate if the regions highlighted by the local score approach included true positives, we computed SNP based heritability for each trait, each time using

three sets of SNPs to compute the kinship matrix: 1) All the SNPs in the genome over 10% frequency, 2) all the SNPs within QTLs identified by the local score approach, and 3) all SNPs not included in the QTLs identified by the local score approach.

Pathway enrichment analysis

To investigate biological functions associated with LSP of accessions or their influence over microbial hubs, we searched for enrichment in annotated pathways (BIOCYC) and GO categories (Biological processes only) in *Arabidopsis thaliana*. Gene-set enrichment methods are designed for assays that directly assign *p*-values or effects to individual genes (i.e. RNAseq experiments). Here, for each trait, each gene was attributed the largest absolute SNP effect within a distance of 5kb on each side and followed the setRank procedure that accounts for overlapping categories and multiple testing. We set the parameter "setPCutoff" to 0.01 and the "fdrCutoff" to 0.05 [29]. To account for specificities of gene set enrichment in the context of association mapping, we also tested the enrichment of the gene groups identified by setRank using a weighted Kolmogorov-Smirnov score [30] and a permutation scheme accounting for the non-independence of marker effects due to linkage disequilibrium along the genome, as well as the potential clustering of genes with similar function [31,32]. Briefly, enrichment was calculated using a weighted Kolmogorov sum using gene effect rank (and not a gene effect significance threshold)[30]. Enrichments were then tested against an empirical distribution generated from 1e5 permutations. For each permutation, chromosomes are randomly re-ordered and re-oriented and the whole genome is shifted (or "rotated") by a random number, before re-assigning SNP effects to genes and calculating enrichment for the groups of genes of interest. We considered only categories with an empirical *p*-values below 0.05.

Untargeted metabolomics

Plant material and sample preparation.

34

This analysis uses three sets of samples. The first are samples collected from the experiments in Sweden and correspond to a subset of those used for the microbial community. In particular we chose samples from the four experiments established in 2012 and focused on a subset of 50 accessions selected to span the genetic variation among hosts in our mapping population. The second set of samples correspond to 6 replicates of the same 50 genotypes grown in the University of Chicago greenhouse during the summer 2014 under long day conditions (16-hour light period), in standard culture soil. After 28 days, plants were vernalized for three weeks at 4°C and leaf samples were collected after vernalization, immediately flash frozen in liquid nitrogen, freeze-dried and stored at room temperature. The third set corresponds to 3 replicates of the same 50 genotypes, grown on sterile agar medium (Murashige and Skoog with Nitsch vitamins) in individual well plates in a growth chamber with a 16-hour light period (long day condition). Seeds were sterilized by a 70% ethanol bath for 10 minutes, and manipulated under a sterile hood. Samples were collected after 28 days of growth, flash frozen, freeze-dried, and stored at room temperature.

Dried samples from the 3 sets were coarsely ground, and distributed in 18 96-well plates with two ceramic grinding beads per well (10mg per well +/- 2mg). Samples were randomized across all plates to limit confounding of biological effects. In addition, each plate included 16 random samples (1⁄6) from each experimental unit (greenhouse, sterile, and the 4 field experiments).

Specialized metabolite extraction and LC-MS analysis

The extraction protocol was designed to extract polar compounds such as glucosinolates and flavonoids. Samples in plates were ground using a Geno/Grinder (SPEX SamplePrep 2010) at 1750 rpm for two minutes. The extraction buffer (70% methanol, 30% water, internal standard: quercetin, 0.0708 mM) was added using a Tecan pipetting robot

35

(100 µl per milligram of dry material). Samples were shaken at room temperature for two hours and filtered on 96-well filter plates (0.45µm) on a vacuum manifold. The flow-through was collected in 96-well plates and stored at 4°C.

Samples were auto injected through a Zorbax SB-C18 2.1 × 150 mm, 3.5 µm column on an Agilent Q-TOF LC–MS with dual ESI (Agilent 6520) with the following parameters: 325 °C gas temperature, 6 L min−1 drying gas, 35 eV fixed collision energy, 35 psig nebulizer, 68 V skimmer voltage, 750 V OCT 1 RF Vpp, 170 V fragmentor, and 3500 V capillary voltage. Mass accuracy was within 2–5 ppm. Samples were eluted with 0.1% formic acid in water (A) and 100% acetonitrile (B) using the following separation gradient: 95% A injection followed by a gradient to 90% A at 1 min, 45% A at 6 min, 100%B at 6.5 min with 4 min hold and 3 min equilibration. An external standard (sinigrin, 1mM) was run 4 times before each plate and one time every 20 samples to monitor and maintain run quality. Compounds were characterized using retention times and fragmentation patterns of chromatograms with automatic agile integration in Agilent Mass Hunter Software (Qualitative Analysis B6 2012) and fragments were compared to online databases, massbank (massbank.jp) and plantCyc (plantcyc.org). The XCMS package for peak detection in R (cran.r-project.org) was used to align chromatograms, adjust retention times, and group the peaks. For every molecule, a “barcode” peak was chosen to have a unique retention time and mass to charge ratio (m/z) combination. The size of these peaks relative to the internal standard, Quercetin, was used to quantify each molecule in every sample.

Statistical analysis.

The peaks intensities relative to the internal standard were used to capture molecule concentration variation. Standardized intensities were square-root transformed before

36

analysis. Heritability of individual compounds in the three conditions were performed using random intercept models identical to those used to estimate OTU heritability. A fixed "site" effect was added for the field samples. In the greenhouse and sterile conditions, a simple random accession term was used to quantify heritability and estimate accession effects (blups). Those accession effects were used to estimate genetic correlation between specialized metabolites field and greenhouse. We used Pearson's correlation coefficient and corrected the corresponding p-values for false discovery rate (FDR, N=20).

For the field samples we modeled the relationships between the relative abundances of 19 microbial hubs and the relative intensity of 20 compounds (Extended Data Table 6) using a linear models following:

$$H_i \sim \beta 1_s.S_{si} + \beta 2.M_i + \beta 3_s.S_{si}.M_i + \varepsilon_i$$

where H_i are the log-ratio transformed counts of one of the 19 microbial hubs used for mapping, and $\beta 1_s$ are the four site effects, S_{si} is the design matrix assigning sample i to site s, and $\beta 2$ is the effect of one of the 20 molecules identified in our untargeted screen, M_i is the relative intensity of the molecules measured in sample i. $\beta 3_s$ are site specific regression coefficients (interactions between the site and molecule effects). We fitted 380 models (19 hubs and 20 molecules) and used F-tests to estimate term significance. All *p*-values corresponding to the molecule effect $\beta 2$ were corrected for False Discovery Rate (N=380).

37

References:

1. Opstal EJ v., Bordenstein SR. Rethinking heritability of the microbiome. Science. 2015;349: 1172–1173. doi:10.1126/science.aab3958
2. Vétizou M, Pitt JM, Daillère R, Lepage P, Waldschmitt N, Flament C, et al. Anticancer immunotherapy by CTLA-4 blockade relies on the gut microbiota. Science. 2015;350: 1079–84. doi:10.1126/science.aad1329
3. Abdul-Aziz MA, Cooper A, Weyrich LS. Exploring relationships between host genome and microbiome: New insights from genome-wide association studies. Front Microbiol. 2016;7: 1–9. doi:10.3389/fmicb.2016.01611
4. Goodrich JK, Waters JL, Poole AC, Sutter JL, Koren O, Blekhman R, et al. Human genetics shape the gut microbiome. Cell. 2014;159: 789–799. doi:10.1016/j.cell.2014.09.053
5. Pamer EG. Resurrecting the intestinal microbiota to combat antibiotic-resistant pathogens. Science. 2016;352: 535–538. doi:10.1126/science.aad9382
6. FAO. Sustainable agriculture for biodiversity/biodiversity for sustainable agriculture. 2016; 48. doi:FAO, 2016- I6602EN/1/12.16
7. Santhanam R, Luu VT, Weinhold A, Goldberg J, Oh Y, Baldwin IT. Native root-associated bacteria rescue a plant from a sudden-wilt disease that emerged during continuous cropping. Proc Natl Acad Sci U S A. 2015;112: E5013-20. doi:10.1073/pnas.1505765112
8. Wagner MR, Lundberg DS, Del Rio TG, Tringe SG, Dangl JL, Mitchell-Olds T. Host genotype and age shape the leaf and root microbiomes of a wild perennial plant. Nat Commun. 2016;7: 12151. doi:10.1038/ncomms12151
9. Horton MW, Bodenhausen N, Beilsmith K, Meng D, Muegge BD, Subramanian S, et al. Genome-wide association study of *Arabidopsis thaliana* leaf microbial community. Nat Commun. 2014;5: 5320. doi:10.1038/ncomms6320
10. Peiffer JA, Spor A, Koren O, Jin Z, Tringe SG, Dangl JL, et al. Diversity and heritability of the maize rhizosphere microbiome under field conditions. Proc Natl Acad Sci. 2013;110: 6548–6553. doi:10.1073/pnas.1302837110
11. Agler MT, Ruhe J, Kroll S, Morhenn C, Kim S-T, Weigel D, et al. Microbial hub taxa link host and abiotic factors to plant microbiome variation. Waldor MK, editor. PLOS Biol. 2016;14: e1002352. doi:10.1371/journal.pbio.1002352
12. Rochefort A, Briand M, Marais C, Wagner M-H, Laperche A, Vallée P, et al. Influence of environment and host plant genotype on the structure and diversity of the *Brassica napus* seed microbiota. Phytobiomes J. 2019;3: 326–336. doi:10.1094/PBIOMES-06-19-0031-R
13. Veach AM, Morris R, Yip DZ, Yang ZK, Engle NL, Cregger MA, et al. Rhizosphere microbiomes diverge among *Populus trichocarpa* plant-host genotypes and chemotypes, but it depends on soil origin. Microbiome. 2019;7: 76. doi:10.1186/s40168-019-0668-8
14. Long Q, Rabanal FA, Meng D, Huber CD, Farlow A, Platzer A, et al. Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden. Nat Genet. 2013;45: 884–890. doi:10.1038/ng.2678
15. Gardes M, Bruns TD. *ITS* primers with enhanced specificity for basidiomycetes–application to the identification of mycorrhizae and rusts. Mol Ecol. 1993;2: 113–118. doi:10.1111/j.1365-294X.1993.tb00005.x
16. Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M. Swarm: robust and fast clustering method for amplicon-based studies. PeerJ. 2014;2: e593. doi:10.7717/peerj.593
17. Beilsmith K, Perisin M, Bergelson J. Natural bacterial assemblages in *Arabidopsis thaliana* tissues become more distinguishable and diverse during host development. Ecology; 2020 Mar. doi:10.1101/2020.03.04.958165
18. Bulgarelli D, Rott M, Schlaeppi K, van Themaat E, Ahmadinejad N, Assenza F, et al. Revealing structure and assembly cues for Arabidopsis root-inhabiting bacterial

915 microbiota. Nature. 2012;488: 91–95. doi:10.1038/nature11336
916 19. Bergelson J, Mittelstrass J, Horton MW. Characterizing both bacteria and
917 fungi improves understanding of the Arabidopsis root microbiome. Sci Rep. 2019;9: 1–
918 11. doi:10.1038/s41598-018-37208-z
919 20. Deng S, Caddell D, Yang J, Dahlen L, Washington L, Coleman-Derr D.
920 Genome wide association study reveals plant loci controlling heritability of the
921 rhizosphere microbiome. bioRxiv. 2020; 2020.02.21.960377.
922 doi:10.1101/2020.02.21.960377
923 21. Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. Sparse
924 and compositionally robust inference of microbial ecological networks. von Mering C,
925 editor. PLOS Comput Biol. 2015;11: e1004226. doi:10.1371/journal.pcbi.1004226
926 22. Sokal RR, Rohlf FJ. Biometry. The principles and practice of statistics in
927 biological research. 1969. doi:10.1126/science.167.3915.165
928 23. Roux F, Gasquez J, Reboud X. The dominance of the herbicide resistance
929 cost in several *Arabidopsis thaliana* mutant lines. Genetics. 2004;166: 449–460.
930 doi:10.1534/genetics.166.1.449
931 24. Bai Y, Müller DB, Srinivas G, Garrido-Oter R, Potthoff E, Rott M, et al.
932 Functional overlap of the Arabidopsis leaf and root microbiota. Nature. 2015;528: 364–
933 369. doi:10.1038/nature16192
934 25. Farrar K, Bryant D, Cope-Selby N. Understanding and engineering beneficial
935 plant-microbe interactions: plant growth promotion in energy crops. Plant Biotechnol J.
936 2014;12: 1193–1206. doi:10.1111/pbi.12279
937 26. Bonhomme M, Fariello MI, Navier H, Hajri A, Badis Y, Miteul H, et al. A local
938 score approach improves GWAS resolution and detects minor QTL: application to
939 *Medicago truncatula* quantitative disease resistance to multiple *Aphanomyces*
940 *euteiches* isolates. Heredity. 2019;123: 517–531. doi:10.1038/s41437-019-0235-x
941 27. Mueller LA, Zhang P, Rhee SY. AraCyc: A Biochemical Pathway Database
942 for Arabidopsis. Plant Physiol. 2003;132: 453–460. doi:10.1104/pp.102.017236
943 28. Schläpfer P, Zhang P, Wang C, Kim T, Banf M, Chae L, et al. Genome-Wide
944 Prediction of Metabolic Enzymes, Pathways, and Gene Clusters in Plants. Plant
945 Physiol. 2017;173: 2041–2059. doi:10.1104/pp.16.01942
946 29. Simillion C, Liechti R, Lischer HEL, Ioannidis V, Bruggmann R. Avoiding the
947 pitfalls of gene set enrichment analysis with SetRank. BMC Bioinformatics. 2017;18:
948 151. doi:10.1186/s12859-017-1571-6
949 30. Charmpi K, Ycart B. Weighted Kolmogorov Smirnov testing: an alternative for
950 Gene Set Enrichment Analysis. Stat Appl Genet Mol Biol. 2015;14: 279–293.
951 doi:10.1515/sagmb-2014-0077
952 31. Brachi B, Faure N, Horton M, Flahauw E, Vazquez A, Nordborg M, et al.
953 Linkage and association mapping of *Arabidopsis thaliana* flowering time in nature.
954 PLoS Genet. 2010;6: e1000940. doi:10.1371/journal.pgen.1000940
955 32. Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, et al.
956 Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines.
957 Nature. 2010;465: 627–631. doi:10.1038/nature08800
958 33. Donoso R, Leiva-Novoa P, Zúñiga A, Timmermann T, Recabarren-Gajardo G,
959 González B. Biochemical and Genetic Bases of Indole-3-Acetic Acid (Auxin
960 Phytohormone) Degradation by the Plant-Growth-Promoting Rhizobacterium
961 Paraburkholderia phytofirmans PsJN. Parales RE, editor. Appl Environ Microbiol.
962 2017;83: e01991-16, e01991-16. doi:10.1128/AEM.01991-16
963 34. Ganin H, Kemper N, Meir S, Rogachev I, Ely S, Massalha H, et al. Indole
964 Derivatives Maintain the Status Quo Between Beneficial Biofilms and Their Plant Hosts.
965 Mol Plant-Microbe Interactions®. 2019;32: 1013–1025. doi:10.1094/MPMI-12-18-0327-
966 R
967 35. Farré-Armengol G, Filella I, Llusia J, Peñuelas J. Bidirectional Interaction
968 between Phyllospheric Microbiotas and Plant Volatile Emissions. Trends Plant Sci.
969 2016;21: 854–860. doi:10.1016/j.tplants.2016.06.005

970 36. Fahey JW, Zalcmann AT, Talalay P. The chemical diversity and distribution of
971 glucosinolates and isothiocyanates among plants. Phytochemistry. 2001;56: 5–51.
972 doi:10.1016/S0031-9422(00)00316-2
973 37. Huang AC, Jiang T, Liu Y-X, Bai Y-C, Reed J, Qu B, et al. A specialized
974 metabolic network selectively modulates *Arabidopsis* root microbiota. Science.
975 2019;364: eaau6389. doi:10.1126/science.aau6389
976 38. Castrillo G, Teixeira PJPL, Paredes SH, Law TF, Lorenzo L de, Feltcher ME,
977 et al. Root microbiota drive direct integration of phosphate stress and immunity. Nature.
978 2017;543: 513–518. doi:10.1038/nature21417
979 39. Finkel OM, Castrillo G, Herrera Paredes S, Salas González I, Dangl JL.
980 Understanding and exploiting plant beneficial microbes. Curr Opin Plant Biol. 2017;38:
981 155–163. doi:10.1016/j.pbi.2017.04.018
982 40. Foster KR, Schluter J, Coyte KZ, Rakoff-Nahoum S. The evolution of the host
983 microbiome as an ecosystem on a leash. Nature. 2017;548: 43–51.
984 doi:10.1038/nature23292
985 41. Morgan JL, Darling AE, Eisen JA. Metagenomic sequencing of an in vitro-
986 simulated microbial community. PLoS ONE. 2010;5. doi:10.1371/journal.pone.0010209
987 42. Amani J, Kazemi R, Abbasi AR, Salmanian AH. A simple and rapid leaf
988 genomic DNA extraction method for polymerase chain reaction analysis. Iran J
989 Biotechnol. 2011;9: 69–71.
990 43. Chelius MK, Triplett EW. The diversity of archaea and bacteria in association
991 with the roots of *Zea mays L.* Microb Ecol. 2001;41: 252–263.
992 doi:10.1007/s002480000087
993 44. White TJ, Bruns S, Lee S, Taylor J. Amplification and direct sequencing of
994 fungal ribosomal RNA genes for phylogenetics. 1990.
995 45. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. Development
996 of a dual-index sequencing strategy and curation pipeline for analyzing amplicon
997 sequence data on the miseq Illumina sequencing platform. Appl Environ Microbiol.
998 2013;79: 5112–5120. doi:10.1128/AEM.01043-13
999 46. Walters WA, Caporaso JG, Lauber CL, Berg-Lyons D, Fierer N, Knight R.
1000 PrimerProspector: De novo design and taxonomic analysis of barcoded polymerase
1001 chain reaction primers. Bioinformatics. 2011;27: 1159–1161.
1002 doi:10.1093/bioinformatics/btr087
1003 47. Samarakoon T, Wang SY, Alford MH. Enhancing PCR Amplification of DNA
1004 from Recalcitrant Plant Specimens Using a Trehalose-Based Additive. Appl Plant Sci.
1005 2013;1: 1200236. doi:10.3732/apps.1200236
1006 48. Rohland N, Reich D. Cost-effective, high-throughput DNA sequencing
1007 libraries for multiplexed target capture. Genome Res. 2012;22: 939–946.
1008 doi:10.1101/gr.128124.111
1009 49. Caporaso JG, Lauber CL, Walters W a, Berg-Lyons D, Huntley J, Fierer N, et
1010 al. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and
1011 MiSeq platforms. ISME J. 2012;6: 1621–1624. doi:10.1038/ismej.2012.8
1012 50. Martin M. Cutadapt removes adapter sequences from high-throughput
1013 sequencing reads. EMBnet.journal. 2011;17: 10. doi:10.14806/ej.17.1.200
1014 51. Kõljalg U, Nilsson RH, Abarenkov K, Tedersoo L, Taylor AFS, Bahram M, et
1015 al. Towards a unified paradigm for sequence-based identification of fungi. Mol Ecol.
1016 2013;22: 5271–5277. doi:10.1111/mec.12481
1017 52. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The
1018 SILVA ribosomal RNA gene database project: improved data processing and web-
1019 based tools. Nucleic Acids Res. 2013;41: D590–D596. doi:10.1093/nar/gks1219
1020 53. Dixon P. VEGAN, a package of R functions for community ecology. J Veg Sci.
1021 2003;14: 927–930. doi:10.1111/j.1654-1103.2003.tb02228.x
1022 54. Anderson MJ, Willis TJ. Canonical analysis of principal coordinates : a useful
1023 method of constrained ordination for ecology. Ecology. 2003;84: 511–525.
1024 doi:10.1890/0012-9658(2003)084[0511:CAOPCA]2.0.CO;2

55. Bates D, Maechler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Softw. 2015. doi:10.18637/jss.v067.i01
56. Aitchison J. The Statistical analysis of compositional data. J R Stat Soc Ser B. 1982;44: 365–374. doi:10.2307/2345821
57. Lê Cao K-AK-A, González I, Déjean S, González I. Unravelling “omics” data with the R package mixOmics. HAL. 2012.
58. Zhao T, Liu H, Roeder K, Lafferty J, Wasserman L. The huge Package for High-dimensional Undirected Graph Estimation in R. J Mach Learn Res JMLR. 2012;13: 1059–1062.
59. Csárdi G, Nepusz T. The igraph software package for complex network research. InterJournal Complex Syst. 2006;1695: 1–9.
60. Pau G, Fuchs F, Sklyar O, Boutros M, Huber W. EBImage–an R package for image processing with applications to cellular phenotypes. Bioinformatics. 2010;26: 979–981. doi:10.1093/bioinformatics/btq046
61. McCaig AE, Grayston SJ, Prosser JI, Glover LA. Impact of cultivation on characterisation of species composition of soil bacterial communities. FEMS Microbiol Ecol. 2001;35: 37–48. doi:10.1111/j.1574-6941.2001.tb00786.x
62. Wang K, Sipilä TP, Overmyer K. The isolation and characterization of resident yeasts from the phylloplane of Arabidopsis thaliana. Sci Rep. 2016;6: 39403. doi:10.1038/srep39403
63. Collado J, Platas G, Paulus B, Bills GF. High-throughput culturing of fungi from plant litter by a dilution-to-extinction technique. FEMS Microbiol Ecol. 2007;60: 521–533. doi:10.1111/j.1574-6941.2007.00294.x
64. Bartoli C, Frachon L, Barret M, Rigal M, Huard-Chauveau C, Mayjonade B, et al. In situ relationships between microbiota and potential pathobiota in *Arabidopsis thaliana*. ISME J. 2018;12: 2024–2038. doi:10.1038/s41396-018-0152-7
65. Karasov TL, Almario J, Friedemann C, Ding W, Giolai M, Heavens D, et al. *Arabidopsis thaliana* and Pseudomonas pathogens exhibit stable associations over evolutionary timescales. Cell Host Microbe. 2018;24: 168-179.e4. doi:10.1016/J.CHOM.2018.06.011
66. Bradski G. The OpenCV Library. Dr Dobbs J Softw Tools. 2000.
67. Alonso-Blanco C, Andrade J, Becker C, Bemm F, Bergelson J, Borgwardt KM, et al. 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. Cell. 2016;166: 481–491. doi:10.1016/j.cell.2016.05.063
68. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25: 1754–1760. doi:10.1093/bioinformatics/btp324
69. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv Prepr ArXiv. 2013;00: 3. doi:arXiv:1303.3997 [q-bio.GN]
70. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079. doi:10.1093/bioinformatics/btp352
71. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: Fast processing of NGS alignment formats. Bioinformatics. 2015;31: 2032–2034. doi:10.1093/bioinformatics/btv098
72. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From fastq data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Curr Protoc Bioinforma. 2013. doi:10.1002/0471250953.bi1110s43
73. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43: 491–498. doi:10.1038/ng.806
74. Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C, et al. An Arabidopsis example of association mapping in structured samples. PLoS Genet. 2007;3: e4. doi:10.1371/journal.pgen.0030004
75. Horton MW, Hancock AM, Huang YS, Toomajian C, Atwell S, Auton A, et al.

Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. Nat Genet. 2012;44: 212–216. doi:10.1038/ng.1042

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

77. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015;4: 7. doi:10.1186/s13742-015-0047-8

78. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 2012;44: 821–4. doi:10.1038/ng.2310

79. Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, Cunningham JM, et al. Genome partitioning of genetic variation for complex traits using common SNPs. Nat Publ Group. 2011;43: 519–525. doi:10.1038/ng.823

Acknowledgements: Many thanks to Mia Holm for her hospitality and wonderful diners after hard work in the field as well as help during harvesting; to Einar Holm for helping with field work and taking photos of harvested plants; to Torbjörn Säll for assistance with sampling and providing greenhouse space in Lund; and finally to the Kleen family, the Öhman family, Nils Jönsson and the Rathckegården farm for allowing us to install our experiments on their land. Thanks to Timothée Flutre and Talia Karasov for helpful discussions on previous versions of the manuscript. Thanks to Man Yu from the Dean lab who helped generate stem images used for seed production estimates and manual seed production estimates. This work was funded by a grant from the National Health Institute (R01 GM 083068) to JB, MN and CD, by a Dropkin Foundation Fellowship to BB and with support from the University of Chicago to JB. BB has received the support of the EU in the framework of the Marie-Curie FP7 COFUND People Programme, through the award of an AgreenSkills/AgreenSkills+ fellowship (under grant agreement n° 267196). Computing resources and storage were provided by: the Center for Research Informatics, funded by the Biological Sciences Division at the University of Chicago with additional funding provided by the Institute for Translational Medicine, CTSA grant number UL1 TR000430 from the National Institutes of Health; the genotoul bioinformatics platform Toulouse Occitanie, France (Bioinfo Genotoul, https://doi.org/10.15454/1.5572369328961167E12); and Bordeaux Bioinformatics Center (CBiB) at the University of Bordeaux, France.

Author Contributions: BB, DF, SH, JB, MN and CD designed the field trials. BB, DF and SH coordinated fieldwork. BB, DF, EK, FR, AA, MB, SD, TCM, PN, TT, RW took part in fieldwork. HW, with assistance from FH and RF isolated B38, produced the genome sequence and assessed its growth promoting effect in controlled conditions. FR computed the SNP data used for association analysis. BB, PD, MLM, RW produced the microbiota sequence data. PLG, TCM and BB generated and analyzed the metabolomics data. BB and JB conceived of analyses, and BB analyzed the data. BB and JB wrote the paper. MP helped develop the methods for microbiota sequencing. DF, TCM, MP, CD, MN provided comments on the manuscript.

Competing interests: The authors declare no competing interests.

42

Repeatability of analysis and data availability

All scripts used to performed the analyses presented in this paper, as well as non-essential but complementary figures, are available in the repository

https://forgemia.inra.fr/bbrachi/microbiota_paper.git

The ITS and 16S amplicons and B38 sequence data is available under bioproject PRJNA707473

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Materials & Correspondence

Reprints and permissions information is available at www.nature.com/reprints
The authors declare no competing financial interests.
Correspondence and requests for materials should be addressed to jbergels@uchicago.edu

43

Figures:

Fig. 1 | Plants grown in different environments have different microbial communities. The plots represent the projection of each sample on the plane defined by the first two constrained components of the fungal and bacterial communities, describing variation among sites and years. The percentages in parentheses are the proportion of the total inertia (square root of the Bray-Curtis dissimilarity) explained by each component. The colors of the points indicate the site from which samples were collected. Experiments from the South are represented in red (SU) and yellow (SR), and experiments from the North in blue (NR) and dark blue (NA). All points from 2012 and 2013 are encircled by a dark and lighter grey line respectively.

Fig. 2 | The effect of host genetic variation on the microbial community targets relatively few OTUs and percolates through hubs. This figure corresponds to observations in the set of four experiments sampled in 2013, see Extended Data Figure 3 for experiments performed in 2012. **A-D:** Each frame presents the distribution of heritability estimates for individual OTUs in one site. In each frame, the inset graph is a box and whiskers plot contrasting the heritability (y-axis) of bacterial (B) and fungal (F) OTUs. **E-F**: The heritable hubs are represented by large dots, at a distance of 0 (hub). The other OTUs are represented by smaller dots and the x-axis represents their distance to the nearest heritable hub(s) within the sparse covariance networks. The number of heritable hubs detected in each experiment is indicated in the legend. The correlation coefficients presented are Kendall rank correlations calculated for OTUs with a distance to the heritable hub(s) above 0. NE stands for "no edge".

Fig. 3 | Relationship between host genotype seed production and influence on microbial hubs across sites and years. A. Proportion of heritable hub relative counts explained by host effects across all sites and years. **B**. Coefficients for the linear regression explaining lifetime see production variation among accession with accession effects on microbial hubs across experiments (after model selection).

Extended Data:

Extended Data Fig. 1 | Relative frequency of the 10 most frequent OTUs. Each stacked bar (x-axis) corresponds to a site/year combination. The y-axis gives the proportion of the 10 most frequent OTUs. The colors correspond to the taxonomic assignments of OTUs given in the legend (class / order / family).

Extended Data Fig. 2 | The effect of host genetic variation on the microbial community targets relatively few OTUs and percolates through hubs. This figure corresponds to observations in the set of 4 experiments performed in 2012. The same figure is available for the 2013 experiments in Figure 1. **A-D:** Each frame presents the distribution of heritability estimates for individual OTUs in one site. In each frame, the inset graph is a box and whiskers plot contrasting the heritability (y-axis) of bacterial (B) and fungal (F) OTUs. **E-F:** The heritable hubs are represented by large dots, at a distance of 0 (hub). The other OTUs are represented by smaller dots and the x-axis represents their distance to the nearest heritable hub(s) within the sparse covariance networks. The number of heritable hubs detected in each

44

experiment is indicated in the legend. The correlation coefficients presented are Kendall rank correlations calculated for OTUs with a distance to the heritable hub(s) above 0.

Extended Data Fig. 3 | Relationship between the mean per site / year combination of the normalized rank abundance of OTUs (x-axis, rank divided by the number of OTUs) in each sample, and heritability (y-axis). Colored points are heritable OTUs and the color and shape indicate the site and year, respectively. Normalized rank abundance of OTUs displays a positive weak but significant relationship with heritability which has an adjusted *r*-squared of 0.04674 (Fstat=205.8, df=4176, p-value: < 2.2e-16).

Extended Data Fig. 4 | Hubs in microbial networks. Each frame presents the relationship between degree and betweenness centrality for vertices in the networks computed for each site (SU, SR, NM and NA) and year (2012, 2013). Each dot represents an OTU (fungal or bacterial). The larger and labeled dots correspond to OTUs that have values of betweenness centrality and degree in the 5% tail of both statistics.

Extended data Fig. 5 | Relationship between prevalence, heritability (A) , betweenness (B) and degree (C). We performed 8 independent experiments, over two years. For each experiment, we defined prevalent OTUs as those detected in over 50% of the plants. In the three panels, the x-axis represents the number of experiments (from 1 to 8) in which an OTU was prevalent, with years distinguished by shape and sites distinguished by color. In A, the y-axis indicates heritability of OTU relative abundance (i.e. variance explained by a random accession effect) estimated within experiments. Colored points represent OTUs with significant heritability. In B and C, the y-axis indicates betweenness and degree of OTU in networks computed for each experiment and colors points are OTUs defined as hubs.

Extended Data Fig. 6 | Correlation between lifetime seed production (LSP) estimates obtained by counting and measuring siliques (x-axis) versus automated LSP estimates. A. Row data and Spearman rho rank correlation coefficient. B. Log transformed data and Pearson's correlation coefficient. In both panels, outliers are indicated in red.

Extended Data Fig. 7 | Positive correlations among genotype lifetime seed production (LSP) estimates in different experiments. We measure LSP, a major component of fitness in this autogamous selfing species, in four sites over two years for 200 Swedish accessions. This figure shows the pairwise correlations between accession effects on this fitness component estimated in the eight experiments.

Extended Data Fig. 8 | Abundant plant specialized metabolites contribute to shaping the relative abundance of microbial hubs. A. Relationships between specialized metabolites and microbial hubs across experiments. Each bar corresponds to an F-statistic for the effects of the site (grey), the molecule (blue) and the interaction between the two (orange) in a model following the formula HUB ~ Molecule + Site + Molecule *Site (in the form HUB ~ Molecule along the x-axis). The stars associated with each bar indicate the level of significance of the Molecule effect (after FDR correction for 623 tests, only models with *p*-value <0.01 for the molecule effects are shown). Site effects were large for all hubs but the interactions between site and molecule were always small and generally not significant (33 significant in 623 tests without FDR correction; only one significant with FDR correction). **B. Heritability estimates of the molecules** in the field (grey bars) and in the greenhouse (blue bars), and in sterile conditions (orange bars) for each molecule. The vertical segments are

45

95% confidence intervals obtained with 500 bootstraps for heritability estimates. **C. Genetic correlations for specialized metabolites between accessions grown in the field and in the greenhouse.** Each bar represents a Pearson's correlation coefficient between field and greenhouse estimates of accession effects (blups) and significance is given by the stars (after FDR correction for 17 tests). Missing bars correspond to molecules with no heritability in the greenhouse and/or the field. B and C share the x-axis labels.

Extended Data Table 1 | Host variation has a subtle impact on overall community variation. The first 3 columns indicate the community, site and year for which the analyses were performed. Nh stands for the number of principal coordinate components with significant broad sense heritability estimates (95% confidence intervals not overlapping 0). A total of 10 components were computed for each community/site/year combination. "VE" indicates the total amount of microbial community variation captured by the first 10 components and "he" provides an estimated proportion of total variation explained by the identity of host accessions (over the i heritable components for each site/year combination). The overall host effects reported in the main text reflect the distribution of VE*he in this table.

Extended Data Table 2 | List of heritable hubs. Hub OTUs detected in each site and year. H2 is the point heritability estimate for each hub. The columns order, family and genus provide taxonomic assignments.

Extended Data Table 3 | Hubs are enriched for interkingdom connections (edges). For each site (first column) and year (second column), the table presents the results from a $\chi 2$ testing for enrichment in interkingdom edges (third column) when considering all edges, or edges involving at least one hub. B_B, B_F, F_F give the number of edges between 2 bacterial OTUs, a bacterial and a fungal OTU, and 2 fungal OTUs, respectively. The following columns are chi-square values, *p*-values and FDR adjusted *p*-values for 8 tests.

Extended Data Table 4 | Relationships between host genotype lifetime seed production and influence over microbial hubs. For each experiment, we computed a multiple linear regression aimed at explaining variation in lifetime seed production among accessions as a function of variation in the effects of accessions on heritable microbial hubs (as well as their squared values indicated by "^{2}", for example F8 and F8^2) . The table summarizes the results for each site and year, giving the number of accessions used and the adjusted r^2 for each model after forward/backward model selection. The column "selected terms" indicate the microbial hubs included in the final model, the sign of the effect (-, +) with the significance in the last column (ns: p-value ≥ 0.1, . : $0.1 \geq p$-value > 0.05, *: $0.05 \geq p$-value > 0.01, **: $0.01 \geq p$-value > 0.001, *** : p-value ≤ 0.001).

Extended Data Table 5 | Geographical coordinates of Swedish collection sites for live microbial isolates.

Extended Data Table 6 | Secondary metabolites detected in this study. "ID" refers to the identifier assigned to each molecule. "Name" indicates the putative names for the molecules if identified. "Category" describes the type of metabolite: C stands for cyanidin, F stands for flavonoid; GSL stands for glucosinolate; O stands for other. "Base structure" describes the flavonol core of the flavonoids: C stands for Cyanidin, K for Kaempferol and Q for Quercetin. The next eight columns indicate the numbers of different saccharides or the

46

chemical groups that enter in the structure of molecules. “RT” stands for retention time (in second). “mass” indicates the molecular weight: “(obs)” stands for observed and “(exp)” for expected according to the formula.

Supplementary information:

Supplementary Table 1 | Natural accessions of *Arabidopsis thaliana* originating from Sweden and grown in 4 sites across Sweden.

Supplementary Table 2 | Bacterial and Fungal OTUs detected. The table provides, for the 581 Bacterial OTUs and 704 fungal OTUs, the taxonomic assignations. in addition column “heritable”, “hubs”, “heritable hub” indicate the number of experiment (0 to 8) in which OTUs were significantly influenced by host genotype, a hub in the community and both, respectively. Column “Nexp” indicates the number of experiments in which each OTU was prevalent. “Core microbiota” indicates if the OTU was part of the core microbiota defined in this study (1: yes, 0: no).

Supplementary Table 3 | QTLs associated with host effects on hubs and our fitness estimate across experiments. The columns “chromosome”, “start”, and “stop” indicate the genomic coordinates for each QTL. The columns “Nqtl” indicates the number of overlapping associated loci identified by the local score approach which were merged into the QTL. The column “repres” provides a representative SNP for each associated loci aforementioned. Representative SNPs are chosen to have the largest absolute effect on the phenotype for each associated loci. The following column describes which traits display associations with each QTL. For example on line 2, the QTL region overlaps with a loci associated with B41 (value =1) and is an exact match for the loci associated with B99 (value =2). The column “Ntraits” simply counts the number of traits with associations in a QTL region and the column “sizes” is simply the difference between “start” and “stop” and measures QTLs sizes in base pairs.

Supplementary Table 4 | Biological processes significantly enriched among genes overlapping with QTLs for microbial hub variation. “trait” simply indicates the trait for which we detected significant enrichment. The columns “name”, “description” and “databases refer to GO terms identification. and “pathway” is the pathway description. “size” refers to the number of genes annotated with the corresponding terms, “setRank” is the setRank statistic characterizing the importance of a gene set, i.e. how much it overlaps with other gene sets, “pSetRank” expresses the probability of observing a gene set with the same setRank value in a random network with the same number of nodes and edges as the observed gene set network. “correctedPValue” is the enrichment *p*-value accounting for overlapping gene sets and “adjustedPValue” is the same probability but adjusted for multiple testing. “enr” and “pv” are the enrichment and associated *p*-value for the method accounting for linkage disequilibrium and non-random distribution of terms along the genome.

Supplementary Table 5 | Pathways significantly enriched among genes overlapping with QTLs for microbial hub variation. (See description Supplementary Table 4).

47