Plant genes influence microbial hubs that shape beneficial leaf communities

Benjamin Brachi[1, 2], Daniele Filiault[3], Paul Darme[1], Marine Le Mentec[1], Envel Kerdaffrec[3], Fernando Rabanal[3], Alison Anastasio[1], Matthew Box[4], Susan Duncan[4], Timothy Morton[1], Polina Novikova[3], Matthew Perisin[1,5,6], Takashi Tsuchimatsu[3], Roderick Woolley[1], Man Yu[4], Caroline Dean[4], Magnus Nordborg[3], Svante Holm[7], Joy Bergelson[1]

Affiliations:
[1] Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637
[2] BIOGECO, INRA, Univ. Bordeaux, 33620 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]U.S. Army Research Laboratory, 2800 Powder Mill Road, Adelphi, Maryland 20783, USA
[6]Oak Ridge Associated Universities, 4692 Millennium Drive, Suite 101. Belcamp, Maryland 21017, USA
[7] Mid Sweden University, Sundsvall, Sweden

Abstract:

Although the 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. We characterized the leaf microbiota within 200 clonal accessions in eight field experiments and detected effects of both local environment and host genotype on community structure. Within environments, hosts' genetics preferentially associate with a core of ubiquitous microbial hubs that, in turn, structure the community. These microbial hubs correlate with host performance, and a GWAS revealed strong candidate genes for the host factors impacting heritable hubs. Our results reveal how selection may act to enhance fitness through microbial associations and bolster the possibility of enhancing crop performance through these host factors.

Text:

Hosts harbor complex microbial communities that are thought to impact health and development[1]. This is best studied in human hosts for which the 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 resident populations [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 recently 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]. That said, microbial communities are complex entities that are influenced by the combined impact of host factors, environment and microbe-microbe interactions [11]. As a consequence, the extent to which host plants can control microbial communities to their advantage, especially in a natural context, is unclear.

Here, we combine large scale field experiments of plant genotypes grown in their natural environments, extensive microbial community analysis, and genome-wide association mapping to (i) disentangle how the influence of the host is distributed among microbial community members, and thus how host variation shapes the microbiota, (ii) propose plant genes and functions that correlate with variation in the microbiota across environmental conditions, and (iii) examine how key microbial associates impact plant fitness. Our motivation is to further the goal of generating plants with an enhanced ability to host beneficial microbial communities.

Snapshot of microbial community variation

We performed a set of field experiments that included genetically fixed inbred lines of *Arabidopsis thaliana* (hereafter "accessions") originally collected throughout Sweden, mainly in two climatically contrasted regions of the country (Supplementary Table 1); *A. thaliana* in the north of Sweden experiences harsh, long winters on the south facing slopes of rocky cliffs whereas *Arabidopsis* populations in the south of Sweden are typically associated with agricultural or disturbed fields that experience a much milder climate. We established identical

experiments in four representative *Arabidopsis* sites, two each in the North (sites N1 and N2) and South (sites S1 and S2). Experiments were repeated across two years, for a total of eight experiments. Each experiment was organized in a complete randomized block design including 24 replicates of 200 re-sequenced accessions [12], established as seedlings in a mixture of native and potting soil and timed to coincide with local germination flushes in late summer. 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 [13–15]. The sequences obtained were clustered into Operational Taxonomic Units (OTUs) using Swarm to generate community matrices [16]. After filtering, we considered 6656 samples and 1381 OTUs for the fungal community and 6793 samples and 990 OTUs for the bacterial community. The frequency distributions of OTUS were highly skewed, with the top ten most common OTUs contributing over 78% of the reads in each experiment (Extended Data Fig. 1).

Although the same accessions were grown in each site and year, the microbial communities differed across the experiments. We performed constrained coordinate analyses on Bray-Curtis distances to capture dimensions discriminating locations and years (Extended Data Fig. 2) [17]. The first components from these analyses primarily captured differences between Northern and Southern sites, explaining 11 and 6% of the overall diversity in fungal and bacterial communities, respectively. The second components captured differences between the two consecutive years, and explained 5 and 4% of the overall diversity in the fungal and bacterial communities, respectively. The limited variation explained by these components suggests that only a small fraction of the microbial OTUs contribute to this differentiation. To identify OTUs

that are present across locations and years, we calculated the number of experiments in which each OTU was prevalent, which we defined as occurring in at least 50% of the plant samples from each experiment. As reported above, a large fraction of OTUs are at very low frequencies. Nevertheless, many of the 990 bacterial OTUs and 1381 fungal OTUs were locally prevalent multiple times. We could therefore define a core microbiota comprised of 278 fungal and bacterial OTUs that were prevalent in all 8 experiments (Fig. 1).

Heritability of the microbiota

Our experiments provided the opportunity to investigate associations between genetic variation among hosts and their resident microbiomes within the context of natural environmental variation across time and space. First, restricting attention to those OTUs that accounted for more than 0.01% of filtered reads per site and year (as for all following analyses), we performed simple, unconstrained principal coordinates analysis (PCoA) within each experiment and computed the proportion of variance explained by the host (hereafter heritability or H^2). Heritability of these components of the bacterial and of the fungal communities varied widely, from 0% to 16%, with 3 - 9 of the 10 components revealing significant heritability depending on the community, site and year (Extended Data Table 1). These results indicate that genetic variation in the host impacts at least a fraction of the microbiota, as has been observed in previous studies [8–10]. The heritable component, however, did not explain a large part of the overall variation in microbial community structure because components that are heritable were not necessarily those that explained the greatest variation in community structure. We found that host genetic variation explains on average 1.93[min=1, max=3.83] and 2.13[min=0.607, max=5.25]% of the variation in bacterial and fungal OTUs, respectively (Extended Data Table

1), revealing a detectable, but subtle, impact of the host on β-diversity of the microbial community.

Hosts could, in principle, shape microbial community structure by exerting weak control over a large number of community members or by targeting a few microbes that influence the rest of the microbiota through microbe-microbe interactions. We found that between 4.59 and 12.5% of all OTUs revealed significant genotype effects (with the 95% confidence interval of heritability not overlapping 0), depending on location and year. Thus, evidence of host control is focused on relatively few OTUs. There was no consistent difference in the heritability of bacterial versus fungal OTUs (Extended Data Fig. 3).

We explored the ecological importance of these heritable OTUs by computing networks of ecological microbe-microbe interactions for each experiment. We applied the SPIEC-EASI pipeline, which gains power to detect true interactions by assuming that interactions are relatively rare (sparse method) and by using the inverse covariance to capture interactions conditional on variation of the other members of the community [18]. Although our networks included both fungal and bacterial OTUs, most microbe-microbe interactions occurred within each kingdom, with an average of only 9.77 [4.89, 15.52]% of edges connecting fungal and bacterial OTUs. We quantified the ecological importance of OTUs using two common characteristics of nodes in a network. “Degree” is defined as the number of connections between a node and all others. “Between-ness centrality” is defined as the number of shortest paths between all nodes that traverse through a given node[11]. We defined ecologically important hubs as OTUs in the 95% tail of both of these statistics in each network. We identified a total of 122 hubs, representing 71 unique OTUs across all 8 experiments; these hubs are connected to an average of 19.73[min=14.56, max=25.24]% of the edges in the networks, indicating that they are

likely important in structuring the microbial community. In addition, hubs tend to be involved in interactions between fungi and bacteria more often than expected by chance; this suggests that they act as gatekeepers between the two communities (Extended Data Table 2).

Twenty-one of the 122 hubs that we identified (corresponding to 13 unique OTUs, 10 of which are hubs in at least two experiments) were significantly heritable (Table 1). This represents a significant enrichment of hubs amongst heritable OTUs (χ^2 = 11.54, df = 1, *p*-value <0.0007), suggesting that host effects on the microbiota may favor ecologically important microbes. We reasoned that if heritable hubs structure the broader microbial community, we would expect to see a decrease in the heritability of individual OTUs with distance from the heritable hubs in the microbial network. Figure 2 presents the relationships between heritability and the distance from the closest heritable hub. In 7 out of 8 experiments, we observed a significant negative relationship between heritability and the distance to the nearest heritable hub. Thus, host genetic variation most strongly associates with a few microbial hubs that then influence the microbes with which they interact.

Not only do heritable hubs have an impact that percolates through the microbial community, they also tend to be widely distributed among accessions, spatial locations and year. We found that OTUs that are heritable hubs at least once were over-represented in the core microbiota (χ^2=34.814, df=1, *p*-value<1e-6), demonstrating that the ecologically important OTUs with greatest affinity to host genotype are unusually ubiquitous. Host control of the fungal OTU #8 (hereafter F8) is especially important; this OTU was heritable in 5 out of the 7 experiments in which it was a hub (Table 1), suggesting that natural variation in *A. thaliana* influences its microbiota with some consistency across environments. These results suggest the exciting possibility that variation at particular host genes associates with these hubs across time

and space, and thus has broad influence on microbiota.

Finding host genes influencing the microbiota

Which host factors underlie the association between *Arabidopsis thaliana* genotypes and their heritable hubs? An identification of genetic factors enabling control of the microbiota would facilitate genomic selection and breeding of varieties with increased yield. We performed genome-wide association analysis on all 21 heritable hub OTUs using approximately 1.3 million SNPs from the 1001 genomes project [19,20] in a classical, single trait mixed model framework [21]. We investigated candidate genes within ~10kb windows around SNPs with association scores ($-\log_{10}$(*p*-values)) above 5. Pseudo-heritabilities (or SNP based heritabilities) varied from 0 to 18.92%, with an average of 6.95% (N=21). These estimates are consistent with previous β-diversity based estimates [9] and further suggest that host genetics influence hubs in microbial communities. Mapping of individual hubs yielded few if any genome-wide significant peaks. The best association score observed was 7.37 (just above genome-wide significance, which is 7.31 after Bonferroni correction), for F60 in site S2-2013, for a SNP located at position 8637774 on chromosome 5 that has a minor allele frequency of 0.261 (Extended Data Fig. 4). F60 was a heritable hub in both N1-2013 and S2-2013, and belongs to the genus *Leucosporidiella*. The SNP is located within the gene *RRC1* (AT5G25060), a putative splicing factor that has been shown to be involved in phytochrome B-mediated alternative splicing [22]. Interestingly, *RRC1* itself is alternatively spliced, with the more stable variant increasing upon sucrose treatment as well as in both red and far-red light [23]. This association with *RRC1* therefore suggests a possible link between the composition of the microbial community and alternative splicing, especially in response to host energy balance and/or light sensitivity.

We investigated whether associations were shared among the 21 heritable hubs. With a threshold of -log(*p*-value) ≥ 5, only 2 SNPs were associated with 2 different heritable hubs. However, the heritable hubs sharing associations were detected within the same site and location, suggesting that the common association could be due, at least in part, to microbe-microbe interactions rather than direct host effects.

To increase our power to ascertain whether a subset of heritable hubs is influenced by variation at the same (or genetically linked) host genes, we estimated genetic correlations among heritable hubs (*i.e.* pairwise correlations between predicted hub abundance in each host genotype). We found a group of 10 correlated heritable hubs, including hubs detected in different sites and years (Fig. 3). The fact that these correlated hubs did not co-occur in space and time suggests that direct host effects, rather than microbe-microbe interactions, were responsible for their correlations. This observation prompted us to look for host genetic associations that are consistent within this group. As mentioned previously, inspecting the highest associations for the different hubs did not show overlap. To retrieve moderate to strong associations shared among multiple heritable hubs from this cluster, we instead combined *p*-values from the association scans from these 10 OTUs (using Fisher's method for combining *p*-values)[24]. Using a stringent significance threshold -log(combined *p*-value) ≥ 8, this strategy yielded 64 associated SNPs (out of over 1 million) in 10 loci of about 20kb located on three chromosomes. Supplementary Table 2 lists the 47 genes within a 20kb window around associated SNPs.

In addition to *RRC1*, we found promising candidate genes in these significantly associated regions. On chromosome two, we detected significant associations just upstream of *CINNAMATE 4-HYDROXYLASE* (*C4H*, AT2G30490, Extended Data Fig. 5). A wounding-inducible enzyme in the phenylpropanoid pathway [25], *C4H* has been shown to control the amount

and composition of lignin produced [26]. Lignin, and other phenylpropanoids such as anthocyanins and salicylic acid, are important components of plant response to pathogens [27–29]. We also found significantly associated variants on chromosome four just upstream of AT4G14940 (*AMINE OXIDASE 1*, *AtAO1*, Extended Data Fig. 6). This gene encodes a cell wall copper amine oxidase[30] induced by jasmonic acid [31], a hormone whose role in defense responses is well established. Cell wall amine oxidases, and the hydrogen peroxide they produce, have been implicated in processes such as cell wall strengthening, wound healing, and both programmed and hypersensitive cell death [32], all of which could play a part in mediating plant/microbiome interactions. Therefore our analysis provides a short list of candidates that can be tested for their effects on the abundance of heritable hubs and on the microbial communities that these hubs structure.

Heritable hubs influence plant performance

A long-term goal in agriculture is to shape the microbial community to enhance plant performance [6], although examples of positive relationships between microbial community composition and plant fecundity remain rare [7,33,34]. We therefore asked whether we could detect any association between microbial communities and plant performance in our experiments. Independent replicates of all accessions in each of the 8 experiments were left to flower and mature in the field, and mature stems were harvested in early summer. We used high-throughput image analysis to estimate fecundity by measuring the size of mature stems. This method provides estimates highly correlated (Pearson's correlation coefficient=0.85, p-value$\leq 2.2X10^{-16}$) with manual estimates of the total length of silique produced, a common proxy for seed production in *A. thaliana* [35].

First, we tested genetic correlations between log ratios of the relative abundance of heritable hubs and estimates of fecundity, a major fitness component in this annual selfing plant. Among the 21 heritable hubs, 6 show significant correlations (after FDR correction for multiple testing) with this estimator of seed production (Fig. 4). The three strongest correlation coefficients are positive and involve hub F8 in each of 3 Southern experiments. Second, we confirmed that these relationships remain positive and significant in linear mixed models controlling for population structure (Extended Data Table 3). Finally, to explore the fitness impact of the larger microbial community (rather than just the heritable hubs) we fit random forest regression models capable of capturing nonlinear relationships and interactions among all heritable OTUs and patterns of genetic relatedness. These models explained large amounts of variation in 3 out of the 4 southern experiments (between 22 and 28%), and included the heritable hub F8 among the most important variables explaining fecundity breeding-values (Extended Data Fig. 7). Indeed, in the model for site S1 in 2012, which explains 24.48% of fecundity breeding-value variation, F8 is the most important variable, surpassing the effect of local adaptation captured by the first component of pairwise genetic distance (which discriminates northern and southern accessions and captures much of the background host genetic variation)[36]. Thus, the microbial hubs influenced by host genetic variation detected in our study can have extensive effects on a major fitness component, which goes well beyond the effect of the genetic background and locally adaptive variation.

Conclusion

Our findings that widespread, hub species are especially heritable, influences the relative abundance of other microbial community members, correlate with plant fitness and are

influenced by host factors that can be mapped using genome-wide association mapping, opens the door to shaping microbial communities to enhance the performance of agricultural and wild species. While these results encourage optimism for improving productivity, orchestrating targeted changes in the microbial community through control of particular hubs is still challenging. Furthermore, crops selected in the context of intensive chemical agriculture that reduces soil microbial diversity may have diminished capacity to interact efficiently with microorganisms colonizing their tissues [37,38]. Nevertheless, it is clear that targeted manipulation of host factors that shape the microbiome holds promise for a novel means of enhancing productivity, even in realistic and diverse natural settings.

References

1. Opstal, E. J. v. & Bordenstein, S. R. Rethinking heritability of the microbiome. *Science (80-.).* **349,** 1172–1173 (2015).
2. Vétizou, M. *et al.* Anticancer immunotherapy by CTLA-4 blockade relies on the gut microbiota. *Science* **350,** 1079–84 (2015).
3. Abdul-Aziz, M. A., Cooper, A. & Weyrich, L. S. Exploring relationships between host genome and microbiome: New insights from genome-wide association studies. *Front. Microbiol.* **7,** 1–9 (2016).
4. Goodrich, J. K. *et al.* Human genetics shape the gut microbiome. *Cell* **159,** 789–799 (2014).
5. Pamer, E. G. Resurrecting the intestinal microbiota to combat antibiotic-resistant pathogens. *Science (80-.).* **352,** 535–538 (2016).
6. FAO. Sustainable agriculture for biodiversity/biodiversity for sustainable agriculture. 48 (2016). doi:FAO, 2016- I6602EN/1/12.16
7. Santhanam, R. *et al.* Native root-associated bacteria rescue a plant from a sudden-wilt disease that emerged during continuous cropping. *Proc. Natl. Acad. Sci. U. S. A.* **112,** E5013-20 (2015).
8. Wagner, M. R. *et al.* Host genotype and age shape the leaf and root microbiomes of a wild perennial plant. *Nat. Commun.* **7,** 12151 (2016).
9. Horton, M. W. *et al.* Genome-wide association study of *Arabidopsis thaliana* leaf microbial community. *Nat. Commun.* **5,** 5320 (2014).
10. Peiffer, J. a *et al.* Diversity and heritability of the maize rhizosphere microbiome under field conditions. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 6548–53 (2013).
11. Agler, M. T. *et al.* Microbial hub taxa link host and abiotic factors to plant microbiome variation. *PLoS Biol.* **14,** e1002352 (2016).
12. Long, Q. *et al.* Massive genomic variation and strong selection in Arabidopsis thaliana

lines from Sweden. *Nat. Genet.* **45,** 884–890 (2013).
13. Chelius, M. K. & Triplett, E. W. The diversity of archaea and bacteria in association with the roots of *Zea mays L. Microb. Ecol.* **41,** 252–263 (2001).
14. Gardes, M. & Bruns, T. D. *ITS* primers with enhanced specificity for basidiomycetes--application to the identification of mycorrhizae and rusts. *Mol. Ecol.* **2,** 113–118 (1993).
15. Horton, M. W. *et al.* Genome-wide association study of Arabidopsis thaliana leaf microbial community. *Nat. Commun.* **5,** (2014).
16. Mahé, F., Rognes, T., Quince, C., de Vargas, C. & Dunthorn, M. Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ* **2,** e593 (2014).
17. Anderson, M. J. & Willis, T. J. Canonical analysis of principal coordinates□: a useful method of constrained ordination for ecology. *Ecology* **84,** 511–525 (2003).
18. Kurtz, Z. D. *et al.* Sparse and compositionally robust inference of microbial ecological networks. *PLOS Comput. Biol.* **11,** e1004226 (2015).
19. Weigel, D. & Mott, R. The 1001 genomes project for *Arabidopsis thaliana*. *Genome Biol.* **10,** 107 (2009).
20. Klapwijk, M. J., Csóka, G., Hirka, A. & Björkman, C. Forest insects and climate change: long-term trends in herbivore damage. *Ecol. Evol.* **3,** 4183–4196 (2013).
21. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44,** 821–4 (2012).
22. Shikata, H. *et al.* The RS domain of Arabidopsis splicing factor RRC1 is required for phytochrome B signal transduction. *Plant J.* **70,** 727–738 (2012).
23. Hartmann, L. *et al.* Alternative splicing substantially diversifies the transcriptome during early photomorphogenesis and correlates with the energy availability in arabidopsis. *Plant Cell* **28,** tpc.00508.2016 (2016).
24. Sokal, R. R. & Rohlf, F. J. *Biometry. The principles and practice of statistics in biological research*. *San Francisco:: WH Freeman and company* (1969). doi:10.1126/science.167.3915.165
25. Bell-Lelong, D. A., Cusumano, J. C., Meyer, K. & Chapple, C. Cinnamate-4-hydroxylase expression in Arabidopsis. Regulation in response to development and the environment. *Plant Physiol.* **113,** 729–38 (1997).
26. Schilmiller, A. L. *et al.* Mutations in the cinnamate 4-hydroxylase gene impact metabolism, growth and development in Arabidopsis. *Plant J.* **60,** 771–782 (2009).
27. Nicholson, R. L. & Hammerschmidt, R. Phenolic compounds and their role in disease resistance. *Annu. Rev. Phytopathol.* **30,** 369–389 (1992).
28. Lu, H. Dissection of salicylic acid-mediated defense signaling networks. *Plant Signal. Behav.* **4,** 713–717 (2009).
29. Lev-Yadun, S. & Gould, K. S. in *Anthocyanins* (eds. Winefield, C., Davies, K. & Gould, K.) 22–28 (Springer New York, 2008). doi:10.1007/978-0-387-77335-3_2
30. Møller, S. G. & McPherson, M. J. Developmental expression and biochemical analysis of the Arabidopsis atao1 gene encoding an H2O2-generating diamine oxidase. *Plant J.* **13,** 781–791 (1998).
31. Ghuge, S. A. *et al.* The apoplastic copper AMINE OXIDASE1 mediates jasmonic acid-induced protoxylem differentiation in Arabidopsis roots. *Plant Physiol.* **168,** 690–707 (2015).
32. Cona, A., Rea, G., Angelini, R., Federico, R. & Tavladoraki, P. Functions of amine oxidases in plant development and defence. *Trends Plant Sci.* **11,** 80–88 (2006).

33. Timm, C. M. *et al.* Two Poplar-associated bacterial isolates induce additive favorable responses in a constructed plant-microbiome system. *Front. Plant Sci. Front. Plant Sci* **7,** 4973389–497 (2016).
34. Lau, J. A. & Lennon, J. T. Rapid responses of soil microorganisms improve plant fitness in novel environments. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 14058–14062 (2012).
35. Roux, F., Gasquez, J. & Reboud, X. The dominance of the herbicide resistance cost in several *Arabidopsis thaliana* mutant lines. *Genetics* **166,** 449–460 (2004).
36. Horton, M. W. *et al.* Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. *Nat. Genet.* **44,** 212–216 (2012).
37. Finlay, R. D. Ecological aspects of mycorrhizal symbiosis: With special emphasis on the functional diversity of interactions involving the extraradical mycelium. *J. Exp. Bot.* **59,** 1115–1126 (2008).
38. Farrar, K., Bryant, D. & Cope-Selby, N. Understanding and engineering beneficial plant-microbe interactions: Plant growth promotion in energy crops. *Plant Biotechnology Journal* **12,** (2014).

Figures

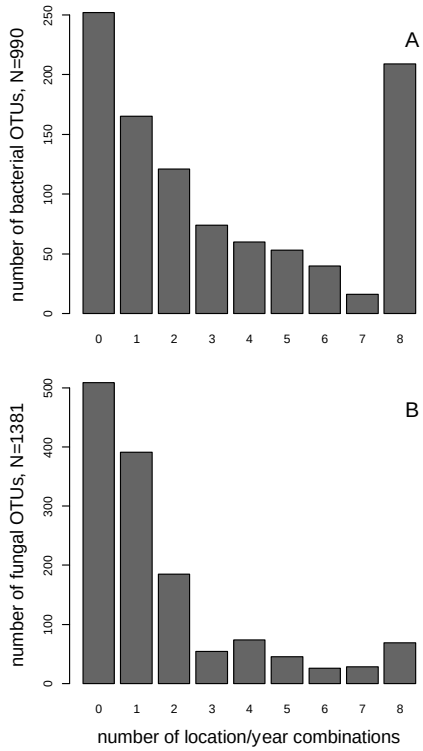

Fig. 1: Microbial communities include a core of microbes prevalent in all locations in both years. The x-axis indicates the number of location/year combinations in which OTUs were prevalent and the y-axis the number of Bacterial OTUs (A) and Fungal OTUs (B). Here we considered all OTUs with more than 10 reads in 5 samples (2371 OTUs). A majority of OTUs were never prevalent, but across both fungal and bacterial communities, a set of 278 fungal and bacterial OTUs were prevalent in the 8 independent experiments we performed.

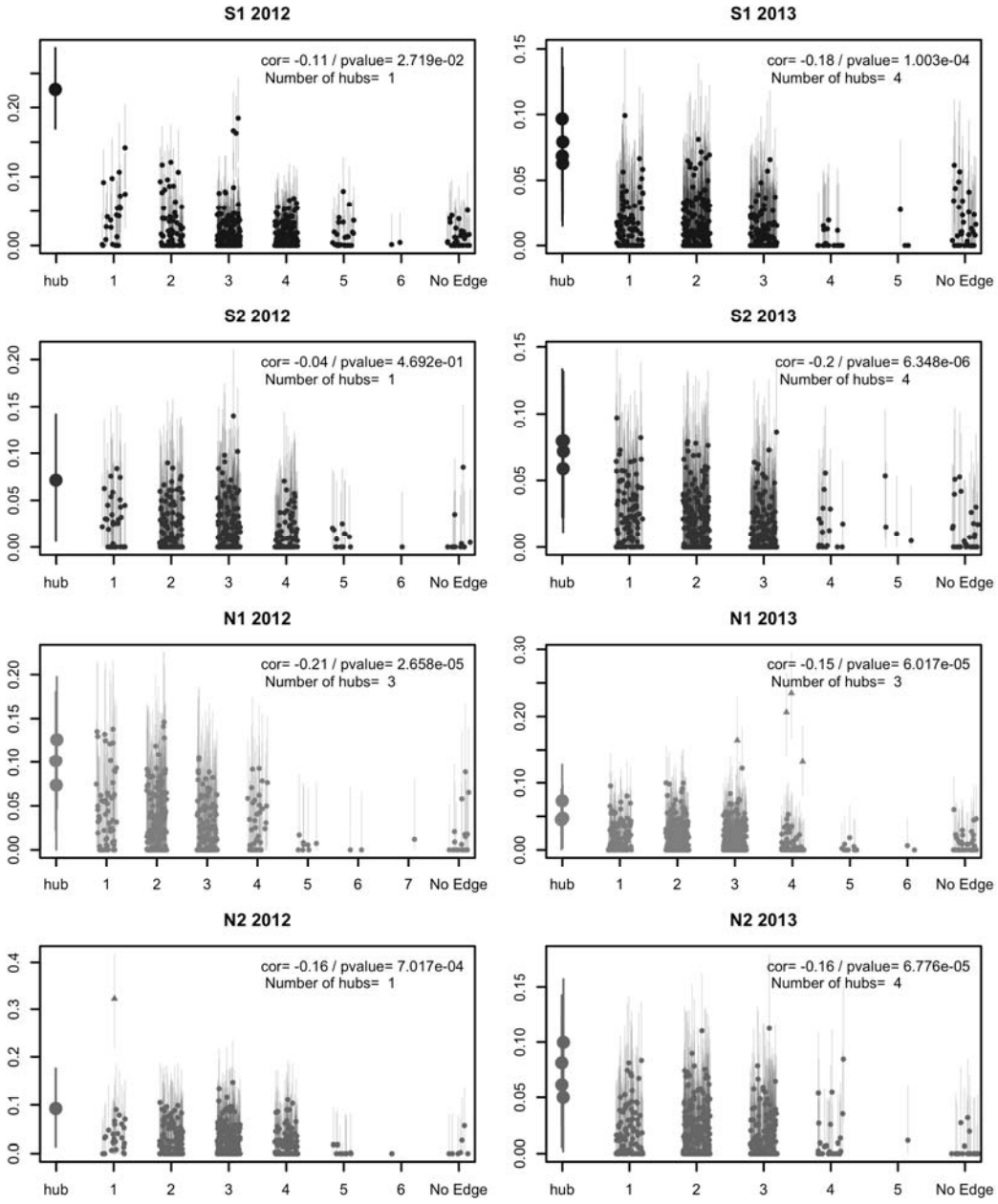

Fig. 2: The effect of host genetic variation on the microbial community percolates through hubs. The heritable hubs are represented by large dots, at a distance of 0 (hub). The other OTUs are represented by smaller dots and the x-axis represents their distance to the nearest heritable hub(s) within the networks of microbe-microbe interactions. The number of heritable hubs detected in each experiment is indicated in the legend. The correlation coefficients presented are Spearman rank correlations calculated for OTUs with a distance to the heritable hub(s) above 0. In N2-2012 and N1-2013, the few OTUs represented by triangles are outliers: their heritabilities are higher than the upper limit of the confidence interval of the heritability of the hub to which they are connected.

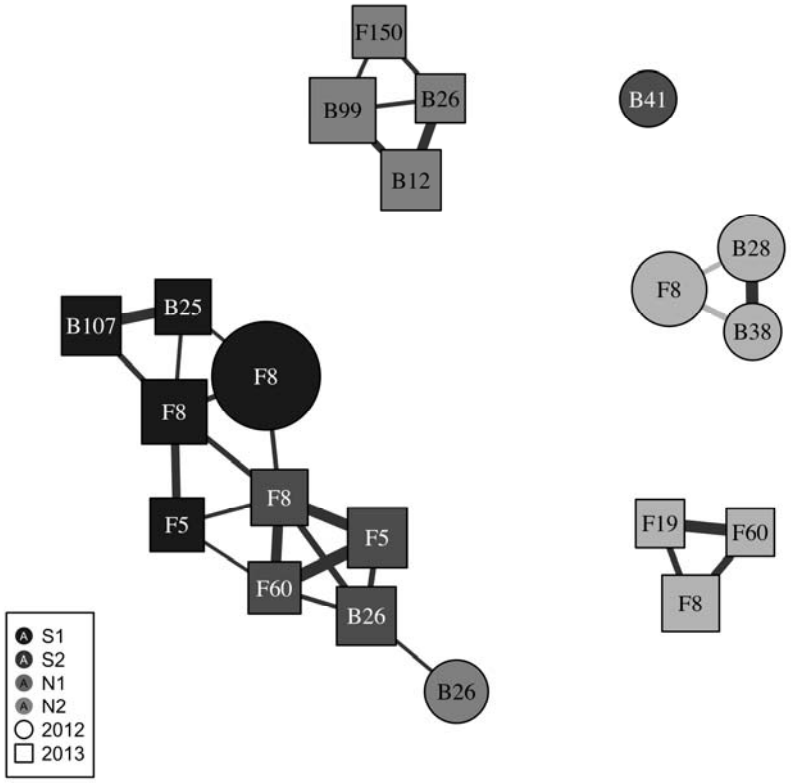

Fig. 3: Genetic correlations among heritable hubs. Each node is a heritable hub, in one year and one experiment, shaped and shaded according to the legend. Dark grey and lighter grey edges correspond to positive and negative pairwise Pearson correlation coefficients between breeding value, respectively. The vertices are sized proportionally to the broad sense heritability estimates of heritable hubs.

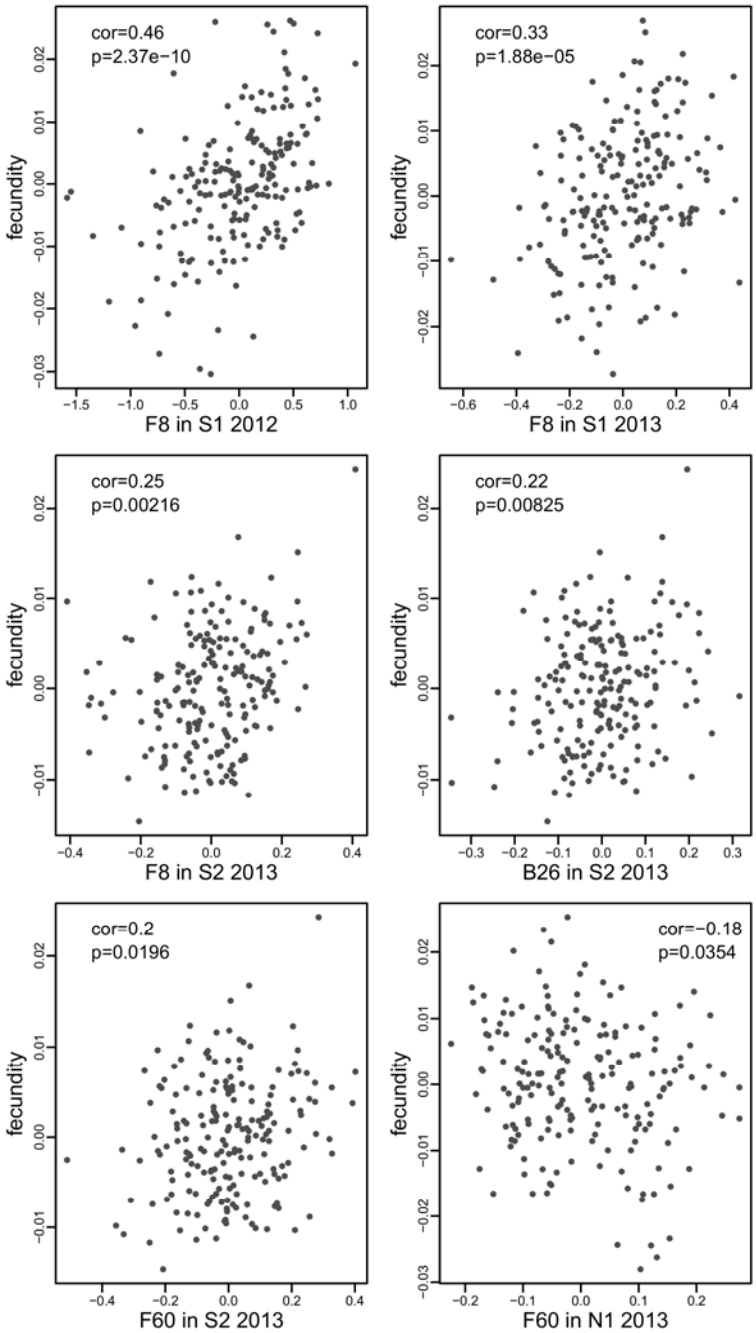

Fig. 4: Effect of heritable hubs on plant performance. Each panel displays 1 of the 6 significant genetic correlations between plant fecundity (accession BLUPS, y-axis) and relative abundance of a heritable hub (accession BLUPs, x-axis). We report Pearson's correlation coefficient (cor) in each panel, along with the *p*-value (p) adjusted for false discovery rate (fdr) for 21 tests (corresponding to the 21 heritable hubs identified).

Tables

OTU	loc	year	H2	CI-low	CI-high	order	family	genus
F8	S1	2012	0.227	0.169	0.286	Pleosporales	unclassified	unclassified
F8	N1	2012	0.126	0.047	0.198	Pleosporales	unclassified	unclassified
B28	N1	2012	0.101	0.022	0.181	Burkholderiales	Comamonadaceae	Variovorax
B99	N2	2013	0.100	0.044	0.157	Burkholderiales	Comamonadaceae	Aquabacterium
F8	S1	2013	0.097	0.042	0.151	Pleosporales	unclassified	unclassified
B26	N2	2012	0.092	0.014	0.176	Lactobacillales	Streptococcaceae	Streptococcus
B12	N2	2013	0.081	0.024	0.142	Sphingomonadales	Sphingomonadaceae	Sphingomonas
F5	S2	2013	0.080	0.026	0.132	Capnodiales	Davidiellaceae	Davidiella
B26	S2	2013	0.079	0.023	0.133	Lactobacillales	Streptococcaceae	Streptococcus
B107	S1	2013	0.079	0.024	0.136	Burkholderiales	Oxalobacteraceae	uncultured
B38	N1	2012	0.074	0.000	0.143	Caulobacterales	Caulobacteraceae	Brevundimonas
F8	N1	2013	0.073	0.023	0.127	Pleosporales	unclassified	unclassified
F8	S2	2013	0.072	0.022	0.121	Pleosporales	unclassified	unclassified
B41	S2	2012	0.071	0.007	0.142	Burkholderiales	Comamonadaceae	Ambiguous taxa
B25	S1	2013	0.068	0.019	0.124	Burkholderiales	Comamonadaceae	uncultured
F5	S1	2013	0.063	0.015	0.115	Capnodiales	Davidiellaceae	Davidiella
F150	N2	2013	0.061	0.006	0.121	Taphrinales	Taphrinaceae	unclassified
F60	S2	2013	0.059	0.011	0.114	Leucosporidiales	Leucosporidiaceae	Leucosporidiella
B26	N2	2013	0.050	0.002	0.108	Lactobacillales	Streptococcaceae	Streptococcus
F19	N1	2013	0.047	0.003	0.096	Pleosporales	family Incertae sedis	Phoma
F60	N1	2013	0.045	0.001	0.090	Leucosporidiales	Leucosporidiaceae	Leucosporidiella

Table 1: List of heritable hubs. Each hub is defined by an OTU, a location (loc) and a year. "CI-low" and "CI-high" indicate the lower and upper 95% confidence interval of the heritability estimate (H2) computed over 500 bootstraps. The columns order family and genus provide taxonomic assignments.

Materials and Methods:

Field experiments

This study uses a set of 200 diverse accessions (inbred lines, Supplementary Table 1) that were previously re-sequenced [12]. 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 locations in which the experiments were installed:

- S1: Ullstorp (lat: 56.067, long: 13.945)
- S2: Ratchkegården (lat: 55.906, long: 14.260)
- N1: Ramsta (lat: 62.85, long: 18.193)
- N2: Ådal (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 University and Lund University, in the North and in the South, respectively). Trays were watered as needed and missing seedlings were transplanted between pots within blocks and then thinned to one per pot 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 across the most obvious environmental gradient

(exposition, shading, slope…). The pierced bottom of the wells allowed the roots to grow through and reach the soil, as was verified upon harvest. The same protocol was followed in 2011 and 2012.

Sample collection and processing

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

DNA extraction

DNA extraction started with 2 enzymatic digestions to maximize yield from Gram-negative bacteria [39]. First, we added 250µl of TES with 50 units.μl^{-1} of Lysozyme (Ready-Lys

Lysozyme, Epicenter) to each well. The plates were then shaken using the Geno-Grinder for 2 min at 1750 rpm, briefly spun and incubated 30 min at room temperature. Second, we 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 [40], 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') [13,15]. For fungi, we amplified the ITS-1 region using the primers ITS1F (5'-CTTGGTCATTTAGAGGAAGTAA-3') [14] and ITS2 (5'-GCTGCGTTCTTCATCGATGC-3') [41]. 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 [42]. The appropriate linkers were chosen using the PrimerProspector program from the Qiime package. 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 [43] 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 extension step (68˚C for 40"). A final extension step at 68˚C was performed for 7' before storing the samples at 4˚C. For each plate the PCRs were performed in triplicates, pooled, and purified using 90 µl of a magnetic bead solution prepared and used following [44]. The purified PCR products were quantified with Picogreen following the manufacturer's instruction [45] and pooled into an equimolar mix. Between 5 and 7 plates (480 to 672 samples) were pooled in each MiSeq run. If the bioanalyzer traces for pooled libraries showed only one dominant peak, they were sequenced directly following the standard MiSeq library preparation protocols for amplicons. In cases where the bioanalyzer trace presented peaks for smaller fragments (left over primers, primer dimers, small PCR products), the libraries were first concentrated 20X on a speedvac (55˚C for 2 to 3 hours), purified with 0.9 volume of magnetic bead solution, and/or size selected using a Blue Pippin (range mode between 300 and 800 bp).

The sequencing was performed using MiSeq 500 cycle V2 kits (251 cycles per read and twice 6 cycles of index reads), using a loading concentration of 12.5pM for *ITS* fragments and 8pM for *16S* fragments following the standard Illumina protocol. Sequencing primers were designed and spiked in following [42]. The sequencing primer for the first read of *16S* fragments was prolonged into the conserved beginning of the fragment amplified to reach a sufficient melting temperature. This primer modification produced no change in the Blast results of the

primers against the GreenGene database. A total of 11 sequencing runs were performed for each of the fungal and bacterial communities.

Sequence processing and clustering

The demultiplexed fastq files generated by MiSeq reporter for the first read of each run were quality filtered and truncated to remove potential primer sequences and low quality basecalls using the program cutadapt [46]. The reads were then further filtered and converted to fasta files using the FASTX-Toolkit (-q 30 -p 90 -Q33). The fasta files for each run were then dereplicated using AWK code provided in the swarm git repository (https://github.com/torognes/swarm). The resulting dereplicated fasta files were filtered for PCR chimeras using the vsearch uchime_denovo command (https://github.com/torognes/vsearch). The dereplicated fasta files for each run were then combined and further dereplicated at the study level. The fasta files were then used as input for OTU clustering using swarm (-t 4 -c 20000). The clustering identified 150 412 and 251 065 OTUs for the fungal and bacterial communities, respectively. The output files were combined into two separate community matrices using a custom python script. The taxonomy of each OTU was determined by blasting the representative sequences for the fungal and bacterial OTUs to the UNITE and SILVA database, respectively [47,48].

Count table filtering

The count tables obtained for both the bacterial and fungal communities were filtered in successive steps by removing:

1) samples corresponding to empty wells and additional genotypes present in the experiments sampled by mistake (leaving 7476 and 7240 samples for the fungal and bacterial count tables, respectively).
2) samples with less than 1000 reads (leaving 6678 and 6819 samples for the fungal and bacterial count tables, respectively)
3) OTUs represented by less than 10 read in 5 samples (leaving 1381 and 993 OTUs for the fungal and bacterial count tables, respectively)
4) for the bacterial community, OTUs assigned to plant mitochondria (leaving 990 OTUs in the bacterial count table)
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 990 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 is 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 [49] and following [17]. 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 distance

matrix among samples (instead of Euclidean distance in the case of rda). This allows identification of the dimension that maximize the variance explained by components, while discriminating groups of samples, here sites, years and their interaction [17].

Core microbiota

In order to define a core microbiota, we counted, for each OTU, the number of location/year combinations in which it was prevalent. We defined "prevalent" as being present in at least 50% of the samples in a given location/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 be represented by at least 10 reads in 5 samples across all site/year combinations (see "Count table filtering") and finally to have non-zero counts in more than 50% of the samples within all site/year combinations.

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 effect capturing the identity of the accessions present in the experiment (block effects had limited impact on H^2 estimates). 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 $\geq 1e^{-10}$).

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 [50,51]. H^2 estimates and confidence intervals were computed for individual OTUs using the method described in the previous paragraph. For mapping and investigating genetic correlations, we used the Best Linear Unbiased Predictions (BLUPs) computed from the random accession effect of the linear model.

Microbe–microbe interaction networks

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 pipeline [18]. 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)[52].

Network statistics

The inferences of microbe-microbe ecological interactions inferred using SPIEC-EASY were passed to the igraph package [53], which was used for enforcing simplicity of graphs (no loops or duplicated edges), computing degree and between-ness 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 between-ness 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 connected to heritable hubs were assigned a distance equal to one more than the maximum distance observed for OTUs connected to heritable hubs.

Genome-wide association mapping

Single polymorphism calling and filtering

Single nucleotide polymorphisms (SNP) used in this study were generated from the sequences generated in the context of the 1001genome project [19] and published in [12]. As pipelines constantly 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) [46]. 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) [54,55]. We used Samtools (v0.1.18) to convert file formats [54] and Sambamba (v0.6.3) to sort and index bam files [56]. 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) [57,58] 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 [59] used to genotype the RegMap panel [36] 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, *a prior* probability of 15, 4 as maximum number of Gaussian distributions, 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 [60,61]. 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 21 heritable hubs transformed to centered log-ratios. Association analysis were performed using a classical one trait mixed model accounting for genetic relatedness among accessions (kinship)[21]. In order to only model one error term throughout our analysis, we didn't first compute BLUPs or means per accession before running association analysis. Instead we considered phenotypes for individual plants, thus only modeling phenotypes with SNP genotypes and the genome-wide kinship. We investigated candidate genes within 5kb on each side of SNPs with –log(*p*-value) above 5.

Genetic correlations and shared genetics

To investigate genetic correlations we computed BLUPs for each heritable hub and estimated Pearson's correlation coefficient for each pair of hubs. Representation in the form of a graph of significant pairwise correlations (p-value$\leq$0.01) revealed a cluster of 10 genetically correlated heritable hubs. This cluster includes heritable hubs detected in multiple locations and years suggesting variation is shaped by shared genetics. In order to identify consistently shared genetics across those 10 heritable hubs, we computed a combined *p*-value using Fisher's method

[24]. SNPs with combined association scores $\log_{10}$(combined *p*-value)≥8, a threshold corresponding to the top 64 for SNPs out of 1,004,654 included this analysis. Candidate genes were investigated within 10kb on each side of associated SNPs.

Estimation of seed production

The experiments each included 8 replicates per block per accession (24 replicates per experiment). While we harvested 2 replicates per block (6 replicates per experiment) for microbiota analysis, the remaining plants were left to grow, flower and produce seeds in the field. We harvested the mature stems of all remaining plants at the end of the spring, when all plants had finished flowering and siliques were mature, and stored them flat in individual paper envelops. We estimated fecundity by the size of the mature stems. After removing remaining traces of roots and rosettes, each mature plant was photographed on a black background, using a DSLR camera (Nikon 60D) mounted on a copy-stand and equipped with a 60mm macro lens (Nikon 60mm). The photographs were segmented (using custom scripts in R based on the EBimage package [62] to isolate plants from the image background and estimate the total surface of the image they occupied.

We validated this method with mature plants harvested from a previous experiment that was planted in N1 in fall 2010, and that included the 200 accessions used in this study. We counted siliques and estimated the average silique size for 1607 mature stems that were also photographed. The total silique length produced per plant (number * average size) was highly correlated with our size estimates based on image analysis (Spearman's rho=0.84) and displayed a clear linear relationship.

Heritable hubs and fecundity

To investigate the relationship between heritable hubs and fecundity in each experiment, we computed BLUPs (breeding values) per accessions for both heritable hubs and square-root transformed fecundity estimates. We then performed three different analyses. First we calculated Pearson's correlation coefficients, which is a classical measure of genetic correlations among normally distributed estimates of breeding values. Second, to account for potential confounding that could arise from population structure, we used the same BLUPs in a linear mix-model accounting for the genetic relatedness among accessions (pairwise kinship matrix) using the function lmekin the R package coxme [63]. Third, in order to account both for non-linear relationships and population structure, we constructed random-forest models [64] for each experiment aiming to explain fecundity breeding values per accessions with breeding values for each heritable OTU and 3 components resulting from the multi-dimensional scaling of a 1-kinship distance matrix among accessions. For each experiment, we computed 10000 trees, sampling 1 third of the heritable OTUs.

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://bitbucket.org/bbrachi/microbiota.git The sequencing data used in this study is available for download from MG-RAST.

References:

39. Morgan, J. L., Darling, A. E. & Eisen, J. A. Metagenomic sequencing of an in vitro-simulated microbial community. *PLoS One* **5,** (2010).
40. Amani, J., Kazemi, R., Abbasi, A. R. & Salmanian, A. H. A simple and rapid leaf genomic DNA extraction method for polymerase chain reaction analysis. *Iran. J. Biotechnol.* **9,** 69–71 (2011).

41. White, T. J., Bruns, S., Lee, S. & Taylor, J. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. *PCR Protocols: A Guide to Methods and Applications* 315–322 (1990). doi:citeulike-article-id:671166
42. Kozich, J. J. *et al.* Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the miseq Illumina sequencing platform. *Appl. Environ. Microbiol.* **79,** 5112–5120 (2013).
43. Samarakoon, T., Wang, S. Y. & Alford, M. H. Enhancing pcr amplification of dna from recalcitrant plant specimens using a trehalose-based additive. *Appl. Plant Sci.* **1,** 1200236 (2013).
44. Rohland, N. & Reich, D. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res.* **22,** 939–946 (2012).
45. Caporaso, J. G. *et al.* Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* **6,** 1621–1624 (2012).
46. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10 (2011).
47. Kõljalg, U. *et al.* Towards a unified paradigm for sequence-based identification of fungi. *Mol. Ecol.* **22,** 5271–5277 (2013).
48. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41,** D590–D596 (2013).
49. Oksanen, J. *et al.* Package 'vegan'. *R Packag. ver. 2.0–8* 254 (2013). doi:10.4135/9781412971874.n145
50. Lê Cao, K.-A. K.-A., González, I., Déjean, S. & González, I. Unravelling 'omics' data with the R package mixOmics. *HAL* (2012).
51. Aitchison, J. The Statistical Analysis of Compositional Data. *J. R. Stat. Soc. Ser. B* **44,** 365–374 (1982).
52. 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.* **13,** 1059–1062 (2012).
53. Csárdi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal, Complex Syst.* **1695,** 1–9 (2006).
54. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
55. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv Prepr. arXiv* **0,** 3 (2013).
56. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: Fast processing of NGS alignment formats. *Bioinformatics* **31,** 2032–2034 (2015).
57. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–498 (2011).
58. Van der Auwera, G. 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
59. Zhao, K. *et al.* An Arabidopsis Example of Association Mapping in Structured Samples. *PLoS Genet.* **3,** e4 (2007).
60. Purcell, S. *et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
61. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer

datasets. *Gigascience* **4,** 7 (2015).
62. Pau, G., Fuchs, F., Sklyar, O., Boutros, M. & Huber, W. EBImage--an R package for image processing with applications to cellular phenotypes. *Bioinformatics* **26,** 979–981 (2010).
63. Therneau, T. coxme: mixed effects Cox models. R package version 2.1-3. (2011).
64. Breiman, L. Random Forests. *Mach. Learn.* **45,** 5–32 (2001).

Supplementary Information is linked to the online version of the paper.

Acknowledgements: Many thanks to Timothée Flutre and Talia Karasov for helpful discussions; 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. 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).

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, TM, PN, TT, RW took part in fieldwork. MY generated stem images used for fecundity estimation and manual fecundity estimates. FR computed the SNP data used for association analysis. BB, PD, MLM, RW produced the microbiota sequence 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, TM, MP, CD, MN provided comments on the manuscript.

Author Information

The authors declare no competing financial interests.

Correspondence and requests for materials should be addressed to jbergels@uchicago.edu