PNAS
www.pnas.org

Supplementary Information for

for

Plant genetic effects on microbial hubs impact host fitness in repeated 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[1,7$]

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
[7] New York University, New York, NY
*contributed equally to the work
[$] corresponding author: Joy Bergelson
Email: jb7684@nyu.edu

This PDF file includes:

Methods S1 to S5
Figures S1 to S12
Tables S1 to S6

Other supplementary materials for this manuscript include the following:

Datasets S1 to S5 in an excel files.
The legends for Datasets S1-S5 are provided at the end of this document.

Supplementary methods

Methods S1 : Sample collection and processing

The rosettes used to characterize the microbial community were harvested in the spring of 2012 and 2013 only a few days after the plants were exposed, following snow melt. We harvested 2 randomly selected 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 using a Geno/Grinder® (from Spex SamplePrep, USA, NJ) 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. Plates included a total of 50 empty wells distributed randomly. Although there was limited cross-contamination of common microbes, many of these were either heritable (see below) and/or identified in independent field isolates suggesting they were not external contaminants.

Methods S2 :DNA extraction

DNA extraction started with 2 enzymatic digestions to maximize yield from Gram-negative bacteria (42). 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 (43), 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.

Methods S3 :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, 44). For fungi, we amplified the ITS-1 region using the primers ITS1F (5'-CTTGGTCATTTAGAGGAAGTAA-3') (16) and ITS2 (5'-GCTGCGTTCTTCATCGATGC-3') (45). 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 (1). The appropriate linkers were chosen using the PrimerProspector program (2). 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 (3) 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 (4). The purified PCR products were quantified with Picogreen following the manufacturer's instruction (5) 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 (remaining 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 6 cycles of index reads twice), 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 (1). 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 (the final sequence for the R1 sequencing primer was 5'-TATGGTAATTTTAACMGGATTAGATACCCKGGTAGTCCACGC-3'). 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.

Methods S4 : *Bacteria sampling from wild* A. thaliana *plants.*

We collected 2 leaves from 10 plants at 5 locations in Sweden (Table S5) . 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 -20°C. Wild *A. thaliana* microbial isolates were collected using modified methods that were previously described (6). Briefly, the leaf and glycerol mixture was plated on six distinct media selected to capture a diverse set of bacterial isolates, including: R2A, Minimal media containing Methanol, Tryptic Soy Agar, Tryptone Yeast extract Glucose Agar, Yeast Extract Manitol Agar (6); 0.1 Tryptic Soy Agar (7). 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 250µl TES (10 mM Tris-HCl pH ~8, 1 mM EDTA, 100

mM NaCl), followed by the addition of 2 mg/mL Proteinase K in 250µl 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 (8).

Over 3900 isolates were cultured and identified using 16S and gyraseB sequencing. Matches to our experimental OTUs are indicated in Dataset S2. Of the isolates identified, we focused on the heritable hub, B38, which appears to contribute to seed-set in the field.

Methods S5 : 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 (9) and published in Long, Q. *et al.* (10). As pipelines evolved, 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) (11). 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) (12, 13). We used Samtools (v0.1.18) to convert file formats (14) and Sambamba (v0.6.3) to sort and index bam files (15). 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) (16, 17) 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 (18) used to genotype the RegMap panel (19) 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.

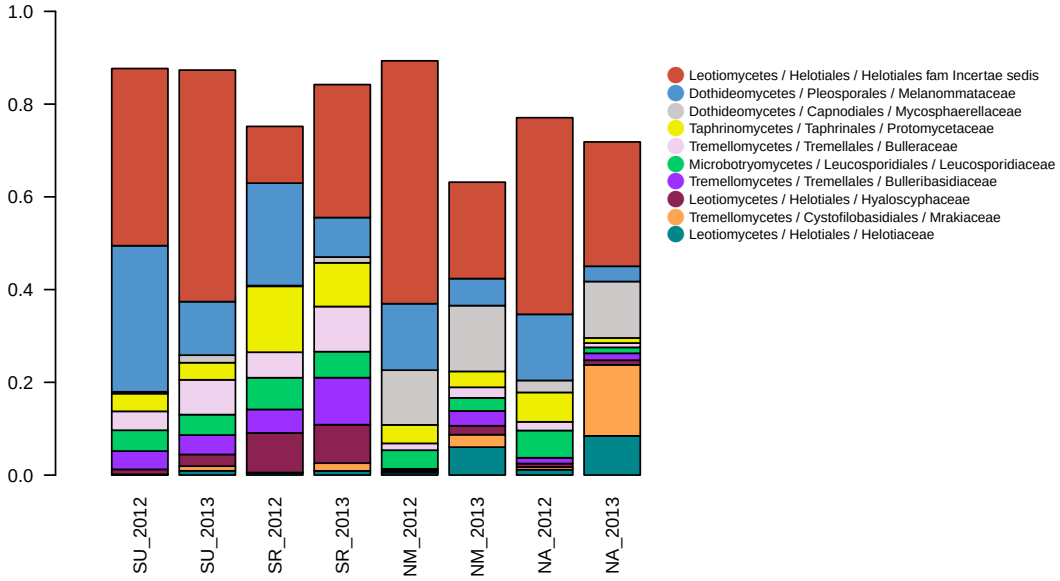
Preparation for use in genome-wide association analysis involved further filtering of individuals and SNPs using Plink1.9 (20, 21). Individuals not included in this study were removed and SNPs with over 5% missing data and with minor allele frequencies below 5% in our collection of accessions were removed.

Références

1. J. J. Kozich, S. L. Westcott, N. T. Baxter, S. K. Highlander, P. D. Schloss, 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).
2. W. A. Walters, *et al.*, PrimerProspector: De novo design and taxonomic analysis of barcoded polymerase chain reaction primers. *Bioinformatics* **27**, 1159–1161 (2011).
3. T. Samarakoon, S. Y. Wang, M. H. Alford, Enhancing PCR Amplification of DNA from Recalcitrant Plant Specimens Using a Trehalose-Based Additive. *Appl. Plant Sci.* **1**, 1200236 (2013).
4. N. Rohland, D. Reich, Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res.* **22**, 939–946 (2012).
5. J. G. Caporaso, *et al.*, Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* **6**, 1621–1624 (2012).
6. Y. Bai, *et al.*, Functional overlap of the Arabidopsis leaf and root microbiota. *Nature* **528**, 364–369 (2015).
7. A. E. McCaig, S. J. Grayston, J. I. Prosser, L. A. Glover, Impact of cultivation on characterisation of species composition of soil bacterial communities. *FEMS Microbiol. Ecol.* **35**, 37–48 (2001).
8. C. Bartoli, *et al.*, In situ relationships between microbiota and potential pathobiota in

Arabidopsis thaliana. *ISME J.* **12**, 2024–2038 (2018).
9. 1001 Genomes Consortium, 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell*, 481–491 (2016).
10. Q. Long, *et al.*, Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden. *Nat. Genet.* **45**, 884–890 (2013).
11. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10 (2011).
12. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
13. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv Prepr. ArXiv* **00**, 3 (2013).
14. H. Li, *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
15. A. Tarasov, A. J. Vilella, E. Cuppen, I. J. Nijman, P. Prins, Sambamba: Fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).
16. G. A. Van der Auwera, *et al.*, From fastq data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* (2013) https:/doi.org/10.1002/0471250953.bi1110s43.
17. M. A. DePristo, *et al.*, A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
18. K. Zhao, *et al.*, An Arabidopsis example of association mapping in structured samples. *PLoS Genet.* **3**, e4 (2007).
19. M. W. Horton, *et al.*, Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. *Nat. Genet.* **44**, 212–216 (2012).
20. S. Purcell, *et al.*, PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
21. C. C. Chang, *et al.*, Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* 4, 7 (2015).

Supplementary Figures



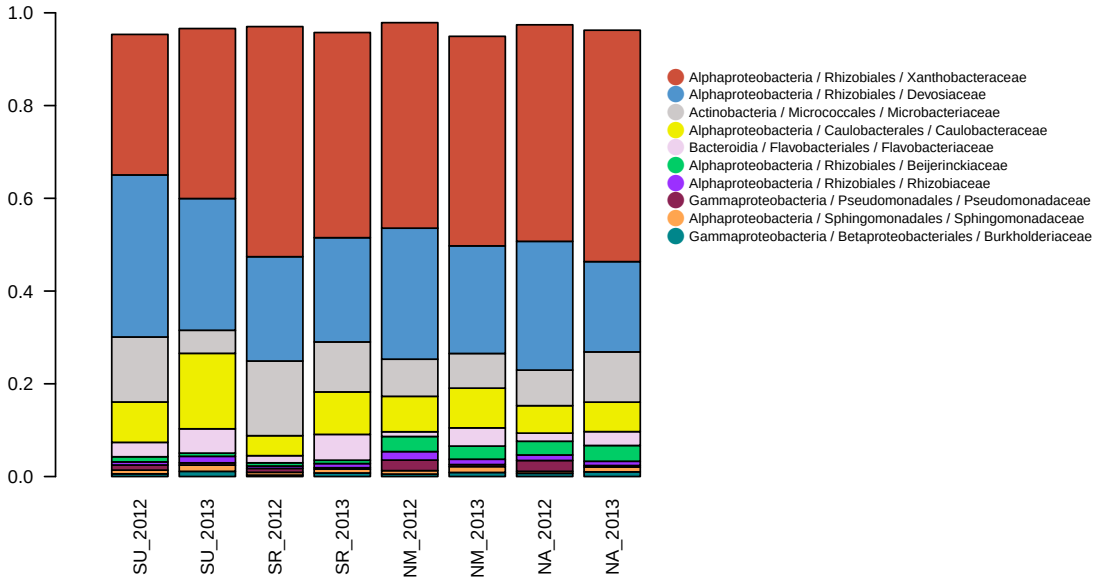


Figure S1 | Relative frequency of the 10 most frequent OTUs. Each stacked bar (x-axis) corresponds to a site/year combination. The y-axis gives the proportion of the 10 most frequent OTUs. The colors correspond to the taxonomic assignments of OTUs given in the legend (class / order / family).

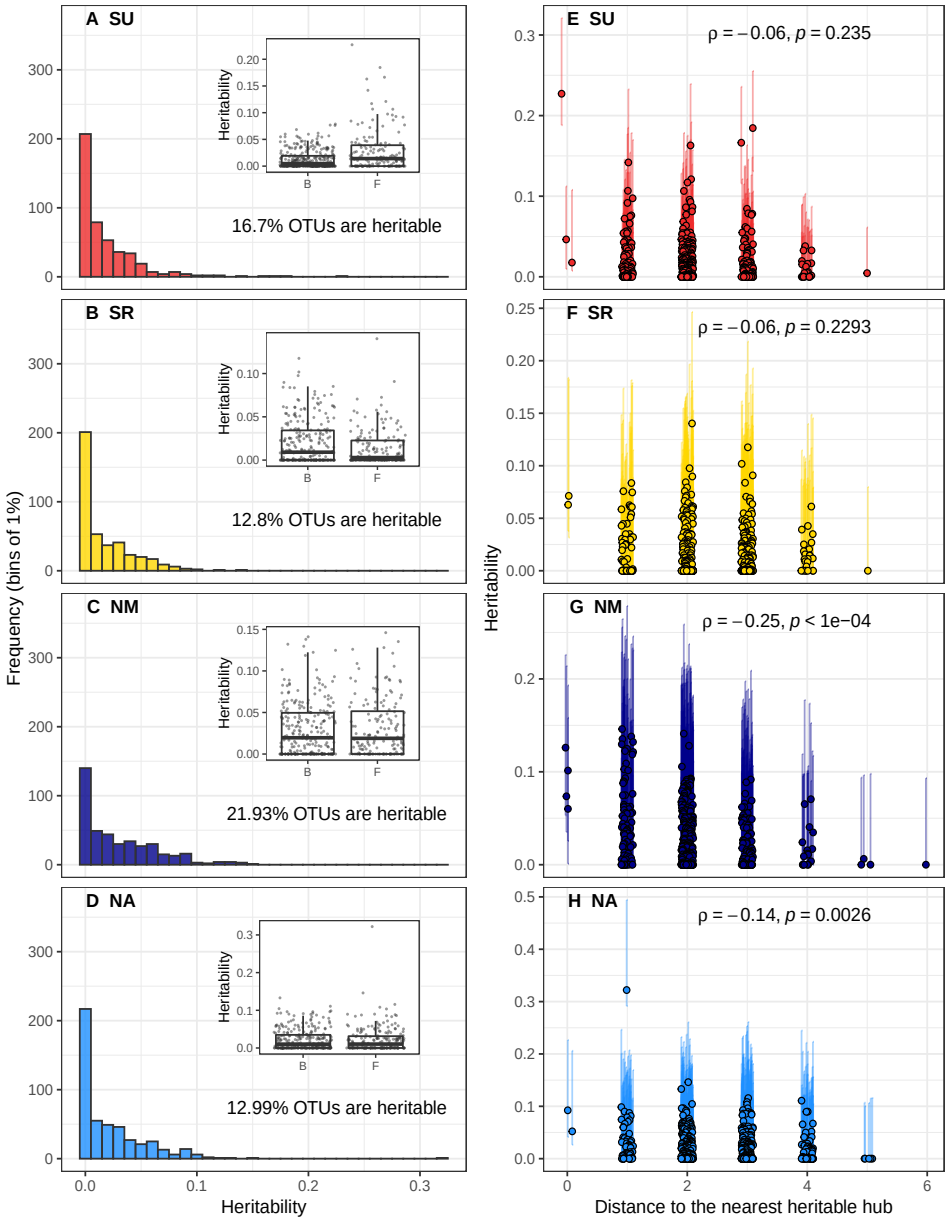


Figure S2 | The effect of host genetic variation on the microbial community targets relatively few OTUs and percolates through hubs. This figure corresponds to observations in the set of 4 experiments performed in 2012. The same figure is available for the 2013 experiments in the main text as Figure 2. **A-D:** Each frame presents the distribution of heritability estimates for individual OTUs in one site (SU, SR, NM, and NA). 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-H:** Relationship between heritability (y-axis) of individual OTUs and their distance to the nearest heritable hub in the sparse covariance networks. The heritable hubs are represented at 0 along the x-axis. The other points are OTUs connected to heritable hubs within the sparse covariance networks. OTUs with a distance of 1 are directly connected to one heritable hub. OTUs with distances above one are indirectly connected to a heritable hub. The correlation coefficients (τ) presented are Spearman rank correlations between heritability and distances to the heritable hub(s) (including 0s).

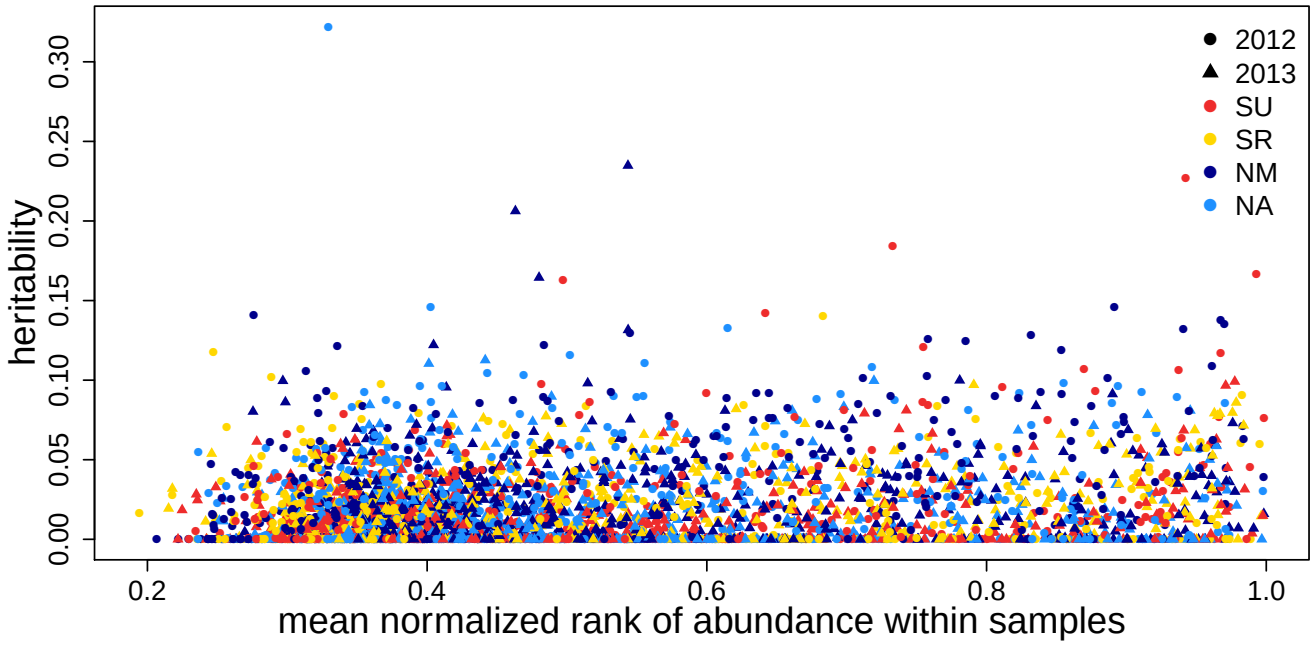


Figure S3 | Relationship between the mean per site / year combination of the normalized rank abundance of OTUs (x-axis, rank divided by the number of OTUs) in each sample, and heritability (y-axis). Colored points are heritable OTUs and the color and shape indicate the site and year, respectively. Normalized 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).

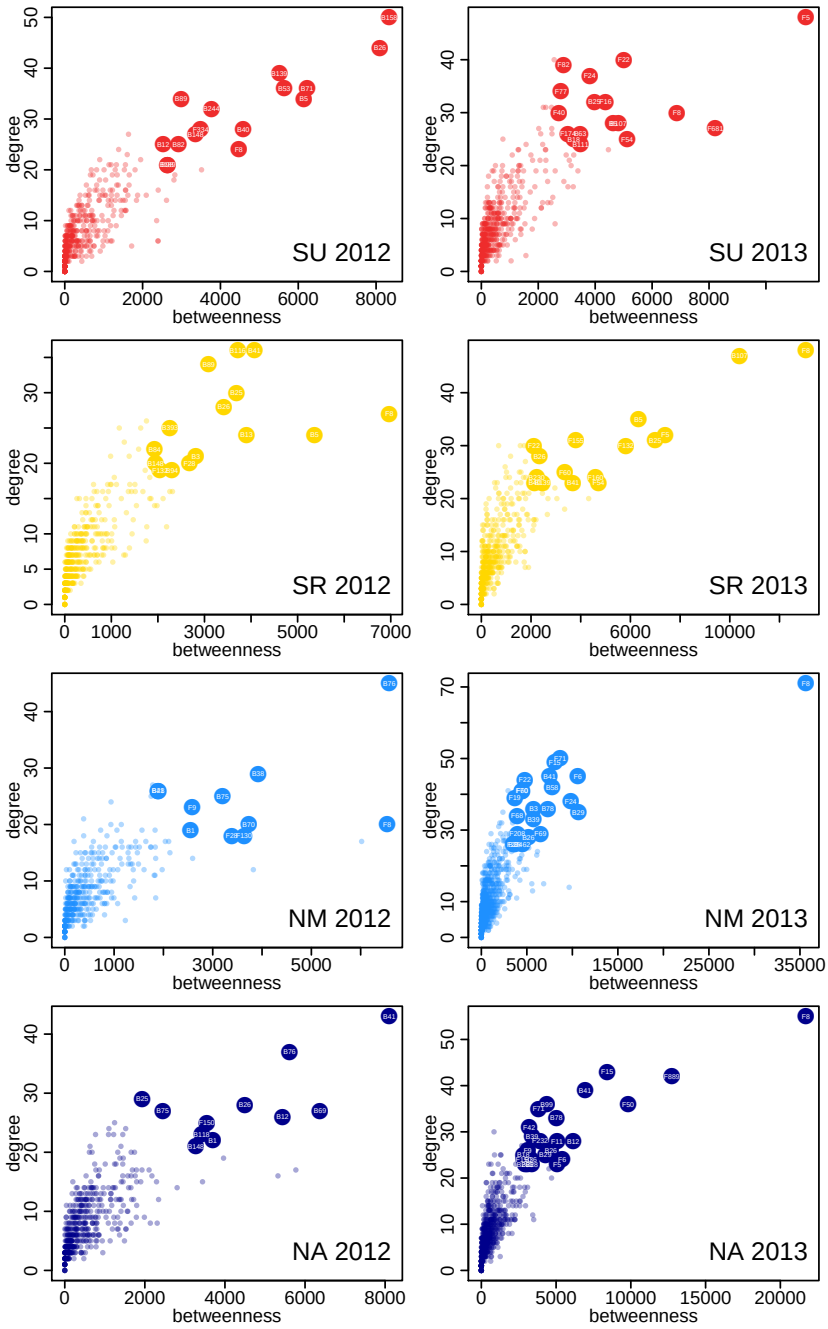


Figure S4 | Hubs in microbial networks. Each frame presents the relationship between degree and between-ness centrality for vertices in the networks computed for each site (SU, SR, NM and NA) and year (2012, 2013). Each dot represents an OTU (fungal or bacterial). The larger and labeled dots correspond to OTUs that have values of betweenness centrality and degree in the 5% tail of both statistics.

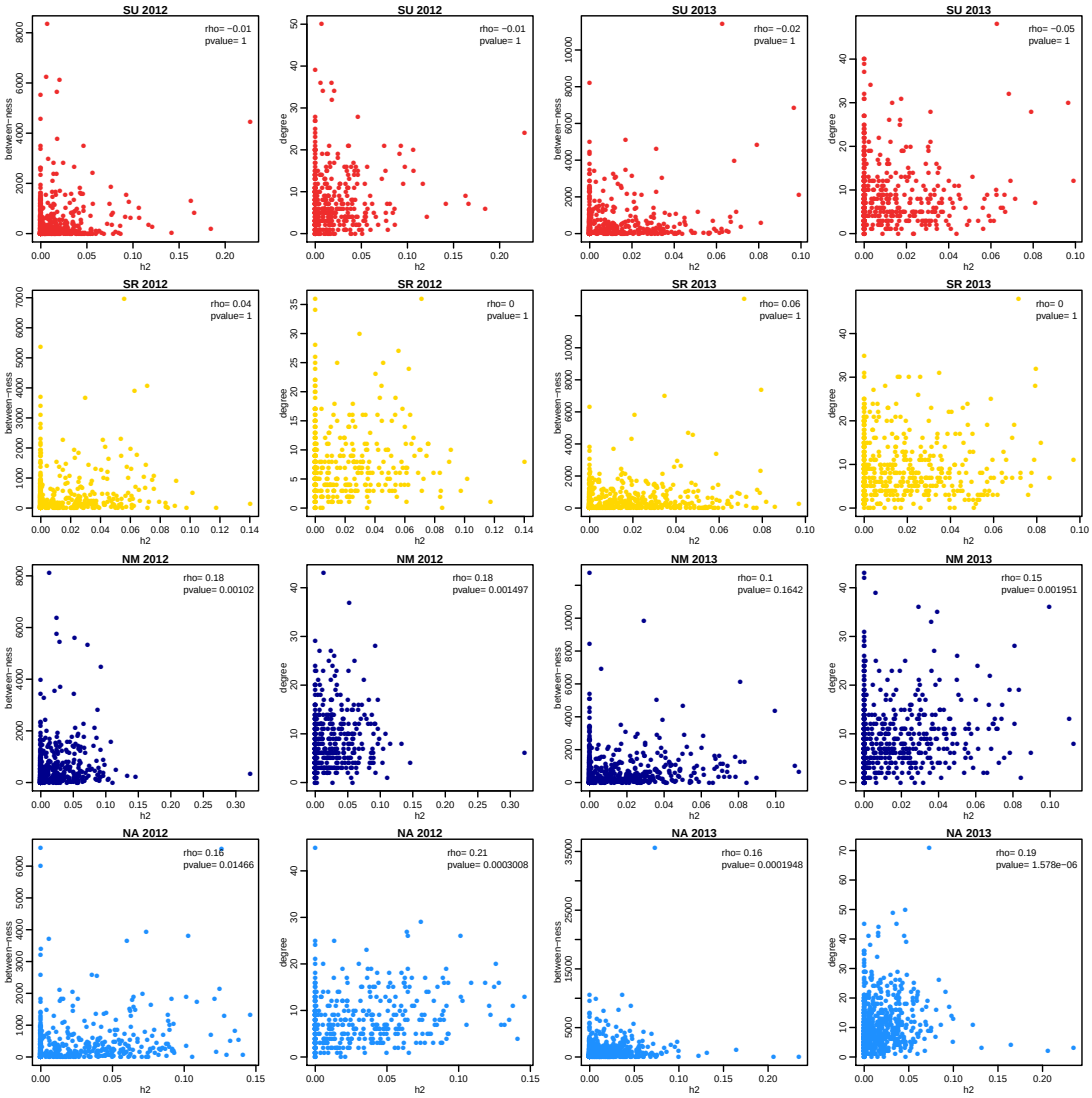


Figure S5 | Relationships between degree and between-ness in networks and heritability of OTUs. Each frame presents the relationship between degree or between-ness centrality and heritability of OTUs for each site (SU, SR, NM and NA) and year (2012, 2013). Each dot represents an OTU (fungal or bacterial). Spearman's correlation coefficients (rho) are reported in the top left corner of each panel, along with the associated *p*-values.

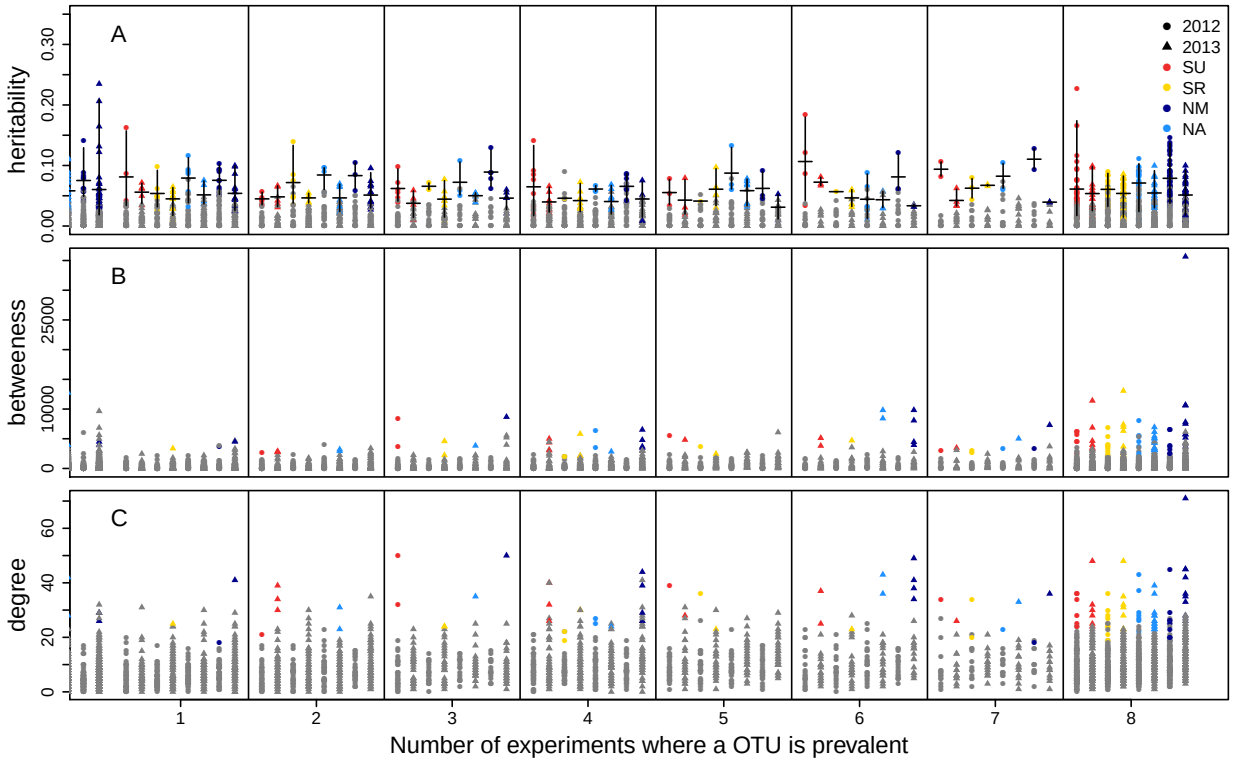


Figure S6 | Relationship between prevalence, heritability (A) , betweenness (B) and degree (C). We performed 8 independent experiments, over two years. For each experiment, we defined prevalent OTUs as those detected in over 50% of the plants. In the three panels, the x-axis represents the number of 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.

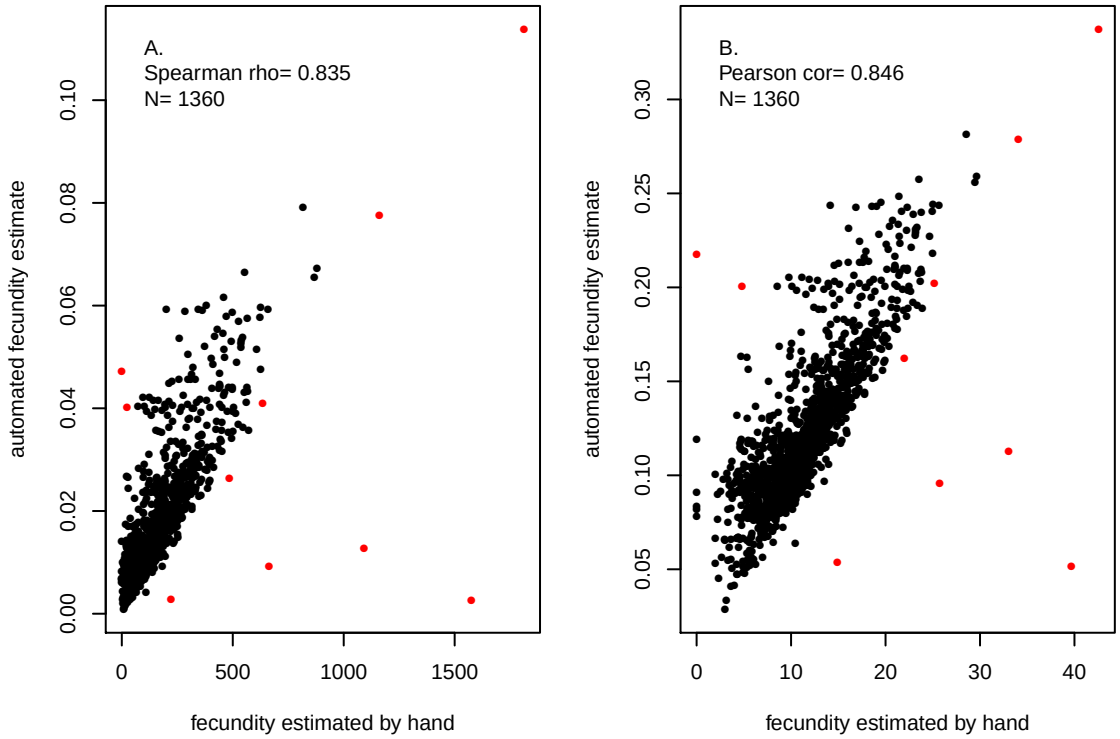


Figure S7 | Correlation between lifetime seed production (seed-set) estimates obtained by counting and measuring siliques (x-axis) versus automated seed-set estimates. A. Row data and Spearman rho rank correlation coefficient. B. Log transformed data and Pearson's correlation coefficient. In both panels, outliers are indicated in red.

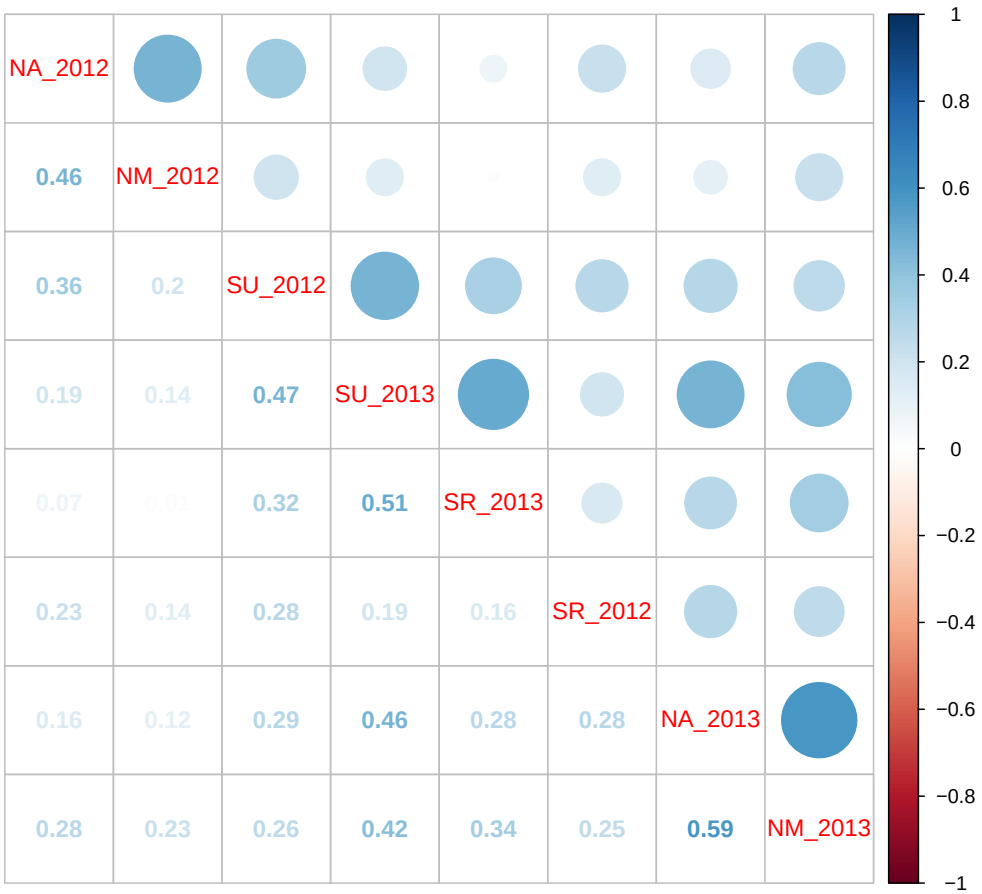


Fig. S8 | Positive correlations among genotype lifetime seed production (seed-set) estimates in different experiments. We measured seed-set, a major component of fitness in this autogamous selfing species, in four sites over two years for 200 Swedish accessions. This figure shows the pairwise correlations between accession effects on this fitness component estimated in the eight experiments.

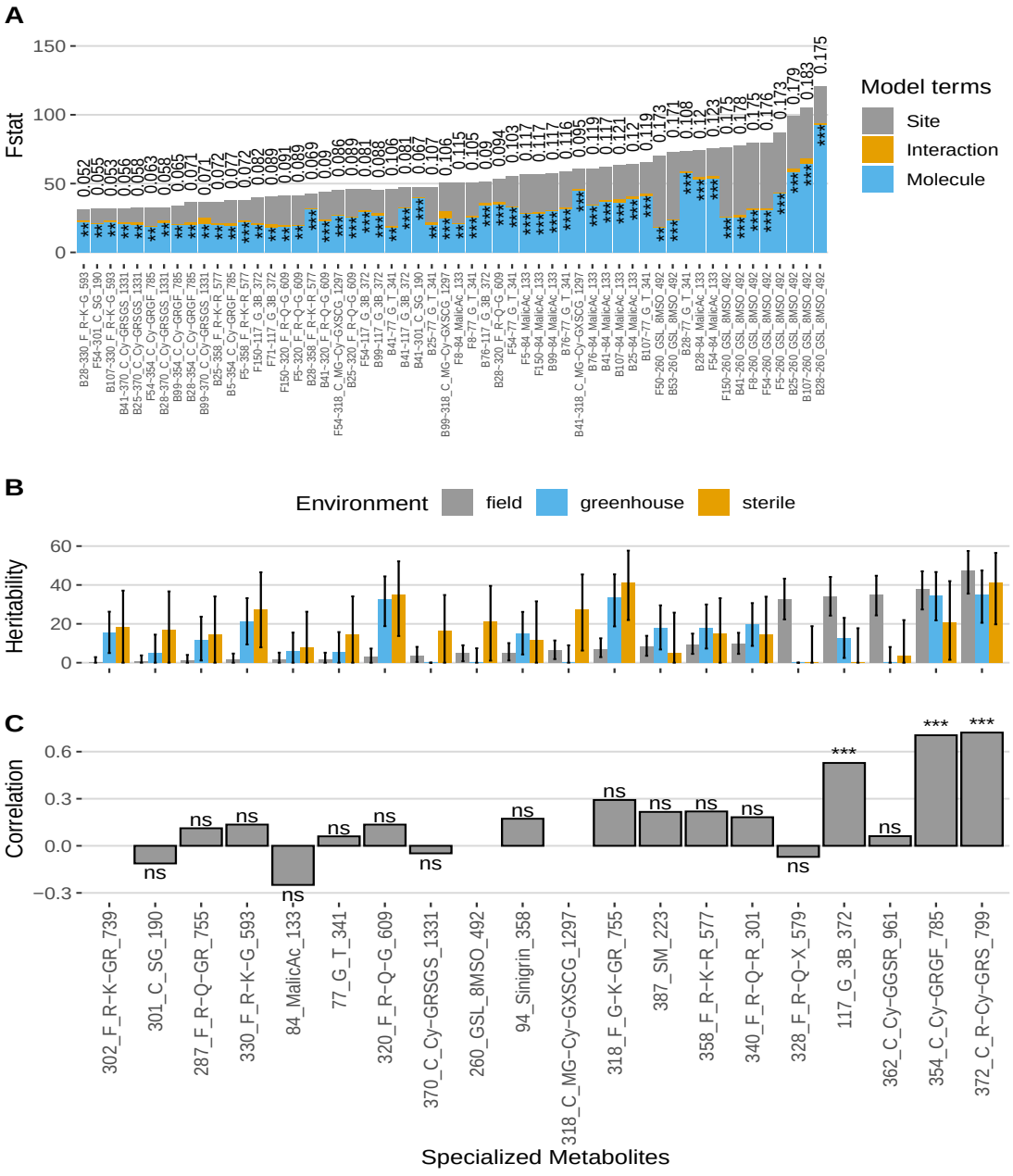

A
Fstat
Model terms
Site
Interaction
Molecule
B
Environment
field
greenhouse
sterile
Heritability
C
Correlation
Specialized Metabolites

Figure S9 | 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; none 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 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.

Supplementary Tables

Table S1 | Host variation has subtle impact on overall community variation.

community	site	year	Nh	VE	he
Fungal	SU	2012	9	58.72	5.22
		2013	9	56.25	0.81
	SR	2012	10	47.99	2.28
		2013	10	49.50	3.00
	NM	2012	10	60.07	2.80
		2013	8	42.54	1.14
	NA	2012	10	61.58	1.78
		2013	10	54.64	0.66
Bacterial	SU	2012	7	49.24	3.49
		2013	8	41.47	1.45
	SR	2012	10	57.18	1.18
		2013	10	45.90	1.99
	NM	2012	10	57.68	3.80
		2013	9	45.84	2.45
	NA	2012	9	55.19	1.28
		2013	10	51.78	1.64

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 proportion 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 and is calculated following:

$$he = 100 \times \sum_{i=1}^{Nh} (pv_i \times h_i)$$

where pv_i and h_i is the proportion of variance explained by component i the heritability of component i, respectively.

Table S2 | List of heritable hubs.

OTU	exp	year	h2	order	family	genus	species
B107	SU	2013	0.0792	Betaproteobacteriales	Burkholderiaceae	arcticum group	uncultured bacterium
B12	NA	2013	0.0809	Sphingomonadales	Sphingomonadaceae	Sphingomonas	Sphingomonas aquatilis
B13	SR	2012	0.0628	Betaproteobacteriales	Burkholderiaceae	NA	NA
B25	SU	2013	0.0685	NA	NA	NA	NA
B25	SR	2013	0.0348				
B26	SR	2013	0.0793	Rhizobiales	Beijerinckiaceae	Methylobacterium	uncultured bacterium
B26	NA	2013	0.0501				
B26	NA	2012	0.0923				
B28	NM	2012	0.1014	Betaproteobacteriales	Burkholderiaceae	Polaromonas	NA
B38	NM	2012	0.0736	Caulobacterales	Caulobacteraceae	Brevundimonas	Ambiguous taxa
B41	SR	2013	0.0111	Betaproteobacteriales	Burkholderiaceae	NA	NA
B41	SR	2012	0.0713				
B5	SU	2013	0.0315	Betaproteobacteriales	Burkholderiaceae	Variovorax	NA
B53	SU	2012	0.0177	Sphingomonadales	Sphingomonadaceae	Sphingomonas	uncultured soil bacterium
B76	NA	2012	0.0522	Corynebacteriales	Nocardiaceae	Nocardia	actinobacterium P23
B99	NA	2013	0.0996	Betaproteobacteriales	Burkholderiaceae	Rhizobacter	Ambiguous taxa
F130	NM	2012	0.0601	NA	NA	NA	NA
F150	NA	2013	0.0611	Taphrinales	Taphrinaceae	Taphrina	Taphrina tormentillae
F160	SR	2013	0.0479	Pleosporales	Phaeosphaeriaceae	Phaeosphaeria	Phaeosphaeria caricicola
F334	SU	2012	0.0464	Pleosporales	Pleosporaceae	Alternaria	Alternaria lolii
F5	SU	2013	0.0628	Capnodiales	Mycosphaerellaceae	Mycosphaerella	Mycosphaerella tassiana
F5	SR	2013	0.0796				
F50	NA	2013	0.0292	NA	NA	NA	NA
F54	SR	2013	0.0458	Sporidiobolales	Sporidiobolaceae	Sporobolomyces	Sporobolomyces roseus
F60	SR	2013	0.0588	Leucosporidiales	Leucosporidiaceae	Leucosporidium	Leucosporidium yakuticum
F60	NM	2013	0.0451				
F69	NM	2013	0.0083	NA	NA	NA	NA
F71	NM	2013	0.0463	Helotiales	Helotiaceae	Tetracladium	Tetracladium marchalianum
F71	NA	2013	0.0393				
F8	SU	2012	0.2271	Taphrinales	Protomycetaceae	Protomyces	Protomyces inouyei
F8	NM	2012	0.1260				
F8	SU	2013	0.0966				
F8	SR	2013	0.0716				
F8	NM	2013	0.0731				

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.

Table S3 | Hubs are enriched for interkingdom connections (edges).

site	year	edges	B_B	B_F	F_F	chisq	pval	adjpval
SU	2012	all	1148	136	555	106.43	<2e-05	1.60E-04
		withhubs	369	61	34			
SU	2013	all	1173	249	978	73.63	<2e-05	1.60E-04
		withhubs	143	86	276			
SR	2012	all	980	136	483	41.30	<2e-05	1.60E-04
		withhubs	276	40	50			
SR	2013	all	1228	249	991	32.46	<2e-05	1.60E-04
		withhubs	198	88	169			
NA	2012	all	1218	158	638	70.79	<2e-05	1.60E-04
		withhubs	240	29	23			
NA	2013	all	1433	288	1184	21.03	6e-05	4.80E-04
		withhubs	290	105	294			
NM	2012	all	1077	117	491	7.66	0.022	1.79E-01
		withhubs	174	27	58			
NM	2013	all	2049	358	1737	84.21	<2e-05	1.60E-04
		withhubs	273	130	406			

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.

Table S4 | Relationships between host genotype lifetime seed production and influence over microbial hubs.

year	site	terms	estimate	std error	t.value	p.value	significance
2012	SU	(Intercept)	-0.001	0.001	-1.079	2.82E-01	ns
		B53	0.055	0.015	3.595	4.12E-04	***
		F8	0.010	0.002	6.529	5.75E-10	***
		$F8^2$	0.004	0.002	2.211	2.82E-02	*
2013	SU	(Intercept)	0.000	0.001	0.310	7.57E-01	ns
		F8	0.018	0.004	4.845	2.56E-06	***
2012	SR	(Intercept)	0.000	0.001	-0.162	8.71E-01	ns
		B13	0.025	0.009	2.740	6.79E-03	**
		$B13^2$	0.123	0.078	1.572	1.18E-01	ns
2013	SR	(Intercept)	0.000	0.001	-0.004	9.97E-01	ns
		B25	0.013	0.009	1.492	1.37E-01	ns
		B26	0.010	0.004	2.378	1.84E-02	*
		B41	-0.054	0.026	-2.119	3.54E-02	*
		F160	0.006	0.004	1.409	1.60E-01	ns
		F5	-0.014	0.004	-3.632	3.64E-04	***
		F60	0.010	0.004	2.582	1.06E-02	*
		F8	0.013	0.004	3.071	2.45E-03	**
		$B41^2$	-1.308	0.715	-1.828	6.91E-02	.
		$F8^2$	0.033	0.016	2.102	3.69E-02	*
2013	NA	(Intercept)	-0.001	0.001	-1.004	3.16E-01	ns
		B26	0.013	0.008	1.749	8.19E-02	.
		$F150^2$	0.048	0.024	1.990	4.80E-02	*
2012	NM	(Intercept)	0.000	0.001	0.046	9.63E-01	ns
		$B28^2$	0.055	0.032	1.746	8.27E-02	.
2013	NM	(Intercept)	0.000	0.001	0.475	6.35E-01	ns
		F60	-0.028	0.008	-3.723	2.58E-04	***
		F69	-0.077	0.034	-2.281	2.36E-02	*
		F8	0.012	0.006	2.063	4.05E-02	*

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 ≥ *p*-value > 0.05, *: 0.05 ≥ *p*-value > 0.01, **: 0.01 ≥ *p*-value > 0.001, *** : *p*-value ≤ 0.001).

Table S5 | Geographical coordinates of Swedish collection sites for live microbial isolates.

Collection	Patch	Site Lat (N)	Site Long (E)	Sample ID	Date
Adal (NA)	Patch #1: 1-10	62.86216	18.33597	A	5-May-17
Varhallarna (near SR)	Patch #1: 1-5	55.58	14.334	Var	11-Apr-17
Varhallarna (near SR)	Patch #2: 6-10	55.58	14.334	Var	11-Apr-17
Ullstorp (near SU)	Patch #1: 1-5	56.0648	13.9707	Ull	10-Apr-17
Ullstorp (near SU)	Patch #2: 6-10	56.0648	13.9707	Ull	10-Apr-17
Tjor (inland) (near SR)	Patch #1: 1-10	58.041	11.683	TJ2	12-Apr-17
Tjor (beach) (near SR)	Patch #1: 1-5	58.041	11.683	Tjor	12-Apr-17
Tjor (beach) (near SR)	Patch #2: 6-8	58.041	11.683	Tjor	12-Apr-17
Tjor (beach) (near SR)	Patch #2: 9-10	58.041	11.683	Tjor	12-Apr-17

Table S6: 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 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.

ID*	Name	Category	Base structure	Glucose	Rhamnose	Xylose	Galactose	Sinapoyl	Malonyl	Coumaroyl	Feruloyl	RT	Mass (obs)	Mass (exp)
318_C_MG-**Cy-**GXSCG_1297		C	Cy	2		1		1	1	1		318	1341.3315	1341.3363
354_C_**Cy-**GRGF_785		C	Cy	1	1		1				1	354	931.248	931.25136
372_C_R-**Cy-**GRS_799		C	Cy	1	2			1				372	1052.22	
362_C_**Cy**-GGSR_961		C	Cy	1	1		1	1				362	961.2628	961.2608
370_C_**Cy**-GRSGS_1331		C	Cy	2	1			2				370	1331.353	
302_F_R-**K**-GR_739		F	K	1	2							302	739.2091	739.2091
318_F_G-**K**-GR_755		F	K	2	1							318	755.2027	755.204
330_F_R-**K**-G_593		F	K	1	1							330	593.15	593.1512
358_F_R-**K**-R_577	Kaempferitrin	F	K		2							358	577.1573	577.1563
287_F_R-**Q**-GR_755		F	Q	1	2							287	755.2055	755.204
320_F_R-**Q**-G_609		F	Q	1	1							320	609.1479	609.1461
328_F_R-**Q**-X_579		F	Q		1	1						328	579.1359	579.1355
340_F_R-**Q**-R_301		F	Q		2							340	593.1536	593.1512
260_GSL_8MSO_492	8-methylsulfinyloctyl	GSL		1								260	492.1039	492.1037
94_Sinigrin_358	Sinigrin	GSL		1								94	358.02	358.02
117_G_3B_372	3-butenyl	GSL		1								117	372.0442	372.0428
301_C_SG_190	Glucopyranosyl sinapate	O		1				1				301	385.1131	385.114
387_SM_223	Sinapoyl malate	O						1	1			387	339.07	338.26
84_MalicAc_133	Malic acid	O							1			84	133.03	134.08
77_G_T_341	Trehalose	O		2								77	341.1119	342.29

***Naming convention:** For flavonoids, molecule names in the ID column are formed as follows: RT_Category_CODE_MZ where RT is the retention time observed in seconds, Category can be Cyanidin, or F for flavonols, and MZ is the m/z for the ion which served as a diagnostic tag for the molecule. CODE is build as follows: Bold letters (Cy, K or Q) separated from other letters by dashes refer to the base structure of the molecule. Letters before the core refer to components on #5 carbon for cyanidins and #7 carbon for K or Q; and letters after the core refer to components on the #3 flavonol carbon.

Supplementary datasets

Dataset S1 | Natural accessions of *Arabidopsis thaliana* originating from Sweden and grown in 4 sites across Sweden.

Dataset S2 | Bacterial and Fungal OTUs detected. The table provides taxonomic assignments for the 581 Bacterial OTUs and 704 fungal OTUs. Columns “heritable”, “hubs”, “heritable hub” indicate the number of experiments (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 whether the OTU was part of the core microbiota defined in this study (1: yes, 0: no). The column “% match in field isolates” indicates the best 16S sequence match (98, 99 or 100%) detected when compared to 3906 cultured bacterial isolates from Sweden (see methods “Isolation, culture and identification of microbial hubs” section and extended data Table 5). Fungal OTUs were assigned “NA” as we did not attempt to culture them. Interestingly, the 151 bacterial OTUs from our experiments that match field isolates were all prevalent in at least one experiment. In addition, OTUs prevalent in all 8 experiments were over-represented among the OTUs for which we found a match to isolates from natural populations. Together this suggests that our prevalent microbes are from the field rather than contaminants from the greenhouse soil we used in our experiments.

Dataset S3 | 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.

Dataset S4 | 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.

Dataset S5 | Pathways significantly enriched among genes overlapping with QTLs for microbial hub variation. (See description Dataset S4).