SUPPLEMENTAL INFORMATION

The interconnected rhizosphere: High network complexity dominates rhizosphere assemblages

Shengjing Shi*, Erin Nuccio*, Zhou Jason Shi, Zhili He, Jizhong Zhou, and Mary Firestone

* Equal contribution

Supplemental Methods

Illumina MiSeq sequencing of 16S rRNA gene amplicons. Soil microbial DNA was extracted from 0.5 g of rhizosphere or bulk soil samples using a phenol-chloroform purification protocol (Griffiths et al. 2000) and quantified by PicoGreen (Invitrogen, Carlsbad, CA, USA). The V4 region of 16S rDNA was amplified with primer set F515 and R806 (Caporaso et al. 2012) and sequenced on an Illumina MiSeq 2.0 platform at the Institute for Environmental Genomics, University of Oklahoma. PCR assays contained 5 ng of DNA template, 0.5 units AccuTaq polymerase (Invitrogen, Carlsbad, CA, USA), 2.5 µl 10X PCR reaction buffer II (including dNTPs), and 0.4 µM of each primer in 50 µl final volume. The reverse primer had a unique barcode for each sample, and the linkers had varying lengths to increase sequencing diversity (Wu et al. 2015). Samples were amplified in triplicate using the following thermocycler protocol: 94°C for 1 min, 30 cycles of 94°C for 20 s, 53°C for 25S and 68°C for 45 s, with a final extension step at 68°C for 10 min. PCR products from all three amplifications were pooled at equal molality, purified by QIAquick Gel Extraction Kit (QIAGEN Sciences, Germantown, MD, USA), and re-quantified with PicoGreen. Finally, sequencing libraries were prepared from purified PCR products diluted to 2 nM (Caporaso et al. 2012). Libraries were prepared according to the MiSeq protocol provided by Illumina and were sequenced using 250 bp paired-end sequencing.

Sequencing Analysis. Sequence data were processed using an in-house pipeline at the University of Oklahoma built on the Galaxy platform. The raw data was evaluated with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) prior to analysis. Bases with quality scores of <20 were trimmed using Btrim (Kong 2011), and paired-end reads were merged using FLASH (Magoč and Salzberg 2011). Merged sequences were discarded if they were <251 bp or >256 bp or contained ambiguous residues. Chimeric sequences were discarded based on prediction by UCHIME (Edgar et al. 2011) using the reference database mode (RDP Database) (Cole et al. 2014). Sequences were clustered into OTUs using UCLUST (Edgar 2010) at a threshold of 97% similarity and singletons were discarded. Taxonomies of 16S OTUs were annotated according to the RDP 16S rRNA classifier (version 2.5) (Wang et al. 2007) using the RDP database. Samples were randomly resampled to the same sequence depth (11,914 sequences per sample), and in total 153,504 OTUs were created for 288 samples. Rarefaction curves before and after resampling are presented in Figure S4.A and S4.B, respectively.

Network Analysis. Various network approaches have been used to construct molecular ecological networks for microbial communities, including differential equation-based network methods, Bayesian network methods, and relevance network methods. However, correlation-

based relevance network methods, such as those generated by random matrix theory (RMT), are most commonly used due to their simple calculation procedures and tolerance to noise (Deng et al. 2012). In addition, while most studies involving relevance network analysis use arbitrary thresholds that create subjective rather than objective networks, the RMT approach is able to avoid this pitfall by automatically defining a threshold for the pairwise similarity coefficient cutoff. Therefore, we used an RMT-based approach to construct networks, and their topological properties were characterized for this study (Deng et al. 2012). The RMT pipeline used in this study (MENA) is freely available for use at the University of Oklahoma's Institute for Environmental Genomics web server (http://129.15.40.240/MENA/).

RMT network construction network construction includes four steps: data collection, data transformation/standardization, pair-wise similarity matrix calculation, and the adjacent matrix determination, where the last two steps are key steps (Deng et al. 2012). For the key steps, after the pair-wise similarity matrix is constructed, an initial similarity threshold is set to determine which covariations are part of the network and define the adjacency matrix. The RMT network procedure then determines whether the adjacency matrix contains significant non-random patterns by evaluating whether the spacing of the eigenvalue distribution follows a Gaussian or Poisson distribution. If the data is random, the similarity threshold is increased until an optimal similarity threshold is found where significant non-random patterns are detected in the network (i.e., spacing of eigenvalues follows Poisson distribution). If the data are truly random, no adjacency matrix will be selected at any similarity threshold. After the adjacency matrix is defined, nodes and edges can be drawn in an indirect network graph. Extensive evaluations of the RMT-based approach indicate that it is a reliable, sensitive and robust tool for identifying transcriptional networks for analyzing high-throughput genomics data for modular network identification and gene function prediction (Luo et al. 2006, Luo et al. 2007).

After the adjacency matrix is selected, modules are identified within the network. In molecular ecological networks, a module is a group of OTUs/genes that are highly connected within the group, but have few connections with OTUs/genes outside the group. Several methods can be used to define modules, including the short random walk method, the leading eigenvector of the community matrix method, the simulated annealing method, and the greedy modularity optimization method. We used the greedy modularity optimization method to identify modules in this study because a previous analysis showed that this method was more effective and sensitive at separating a complex network into modules compared to other methods (Deng et al., 2012).

In this study, we constructed each network from the bacterial community relative abundances derived from 16 biological replicates (i.e., 16 independent microcosms). The MENA pipeline was used with the following settings: OTUs were required to be present in 10 out of 16 replicates; for missing data; fill 0.01 in blanks if data have paired valid values; do not take logarithm; Pearson Correlation Coefficient similarity matrix; calculate by decreasing the cutoff from the top; and for speed selection, regress Poison distribution only. Pipeline was also conducted using logarithm-normalized data (Figure S5). A large number of replicates are required to create a single network because this process assesses covariation patterns between OTUs across replicates. Figure S6 demonstrates the data that underlies a single network link between two OTUs in this analysis (left graph shows positive covariation across 16 replicates; right graph shows negative covariation across 16 replicates). Without sufficient replication, there will not be

enough data to generate robust covariations across the replicates, and the RMT-method will not detect significant non-random patterns in the network.

References

Caporaso, J. G., C. L. Lauber, W. A. Walters, D. Berg-Lyons, J. Huntley, N. Fierer, S. M. Owens, J. Betley, L. Fraser, M. Bauer, N. Gormley, J. A. Gilbert, G. Smith, and R. Knight. 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. The ISME Journal **6**:1621-1624.

Cole, J. R., Q. Wang, J. A. Fish, B. Chai, D. M. McGarrell, Y. Sun, C. T. Brown, A. Porras Alfaro, C. R. Kuske, and J. M. Tiedje. 2014. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Research **42**:D633-642.

Deng, Y., Y.-H. Jiang, Y. Yang, Z. He, F. Luo, and J. Zhou. 2012. Molecular ecological network analyses. BMC Bioinformatics **13**:113.

Edgar, R. C. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics **26**:2460-2461.

Edgar, R. C., B. J. Haas, J. C. Clemente, C. Quince, and R. Knight. 2011. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics **27**:2194-2200.

Griffiths, R. I., A. S. Whiteley, A. G. O'donnell, and M. Bailey. 2000. Rapid method for coextraction of DNA and RNA from natural environments for analysis of ribosomal DNA- and rRNA-based microbial community composition. Applied and Environmental Microbiology **66**:5488-5491.

Kong, Y. 2011. Btrim: a fast, lightweight adapter and quality trimming program for next-generation sequencing technologies. Genomics **98**:152-153.

Luo, F., Y. Yang, J. Zhong, H. Gao, L. Khan, D. K. Thompson, and J. Zhou. 2007. Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory. BMC Bioinformatics **8**:299.

Luo, F., J. Zhong, Y. Yang, R. H. Scheuermann, and J. Zhou. 2006. Application of random matrix theory to biological networks. Physics Letters A.

Magoč, T., and S. L. Salzberg. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics **27**:2957-2963.

Wang, Q., G. M. Garrity, J. M. Tiedje, and J. R. Cole. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and Environmental Microbiology **73**:5261-5267.

Wu, L., C. Wen, Y. Qin, H. Yin, Q. Tu, J. D. Van Nostrand, T. Yuan, M. Yuan, Y. Deng, and J. Zhou. 2015. Phasing amplicon sequencing on Illumina Miseq for robust environmental microbial community analysis. BMC Microbiology **15**:125.

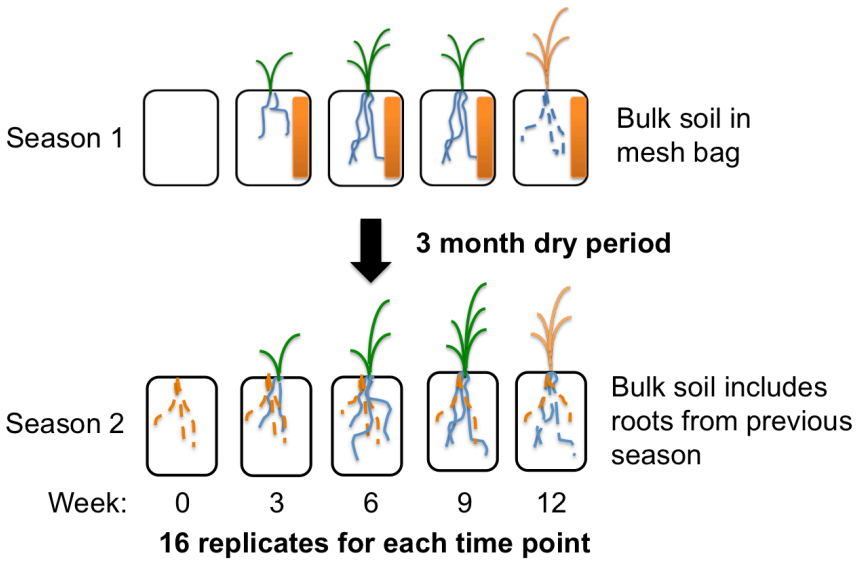


Figure S1. Experimental design and definition of soil types used in this study. Microcosms containing *Avena fatua* were destructively harvested at 10 time points over two seasons (16 replicates per time point, 160 microcosms). Samples were harvested at the following vegetative stages: pre-planted (0 weeks), seedling (3 weeks), vegetative (6 weeks), flowering (9 weeks), and senescence (12 weeks). For season 2, intact microcosms were replanted following a 3-month dry period to simulate a dry California summer. During the first season, bulk soil was collected from 1µm mesh bags. During season 2, bulk soil was collected after removing fresh roots and the residual roots remaining from the previous season.

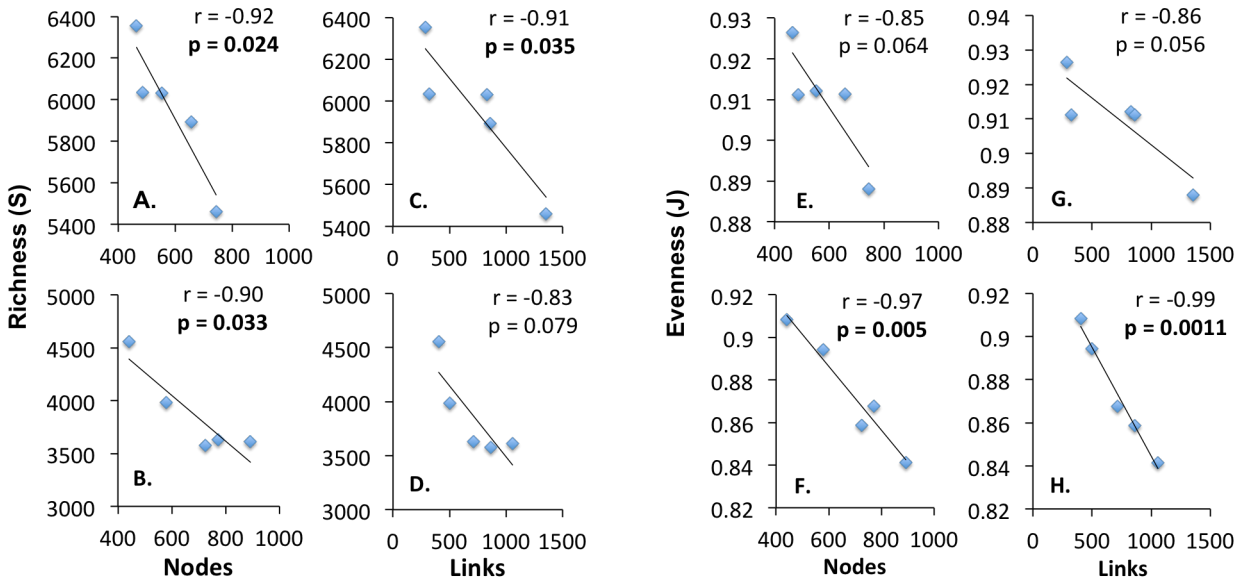
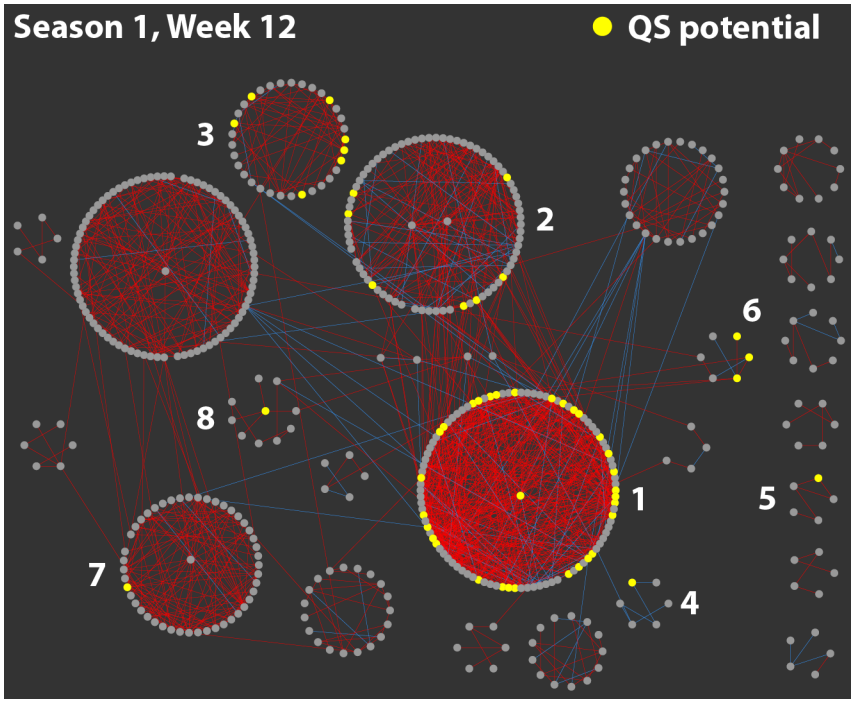


Figure S2. Pearson's correlation analyses correlating network size (number of nodes) and connectivity (number of links) with rhizosphere diversity for Season 1 (top row: A, C, E, G) and Season 2 (bottom row: B, D, F, H). Diversity was assessed by the following metrics: richness (A – D); evenness (E – H). Pearson's product-moment correlation coefficients are indicated by r values. The p values in bold text are significant ($p < 0.05$), and p values in regular text are marginally significant ($p < 0.1$).

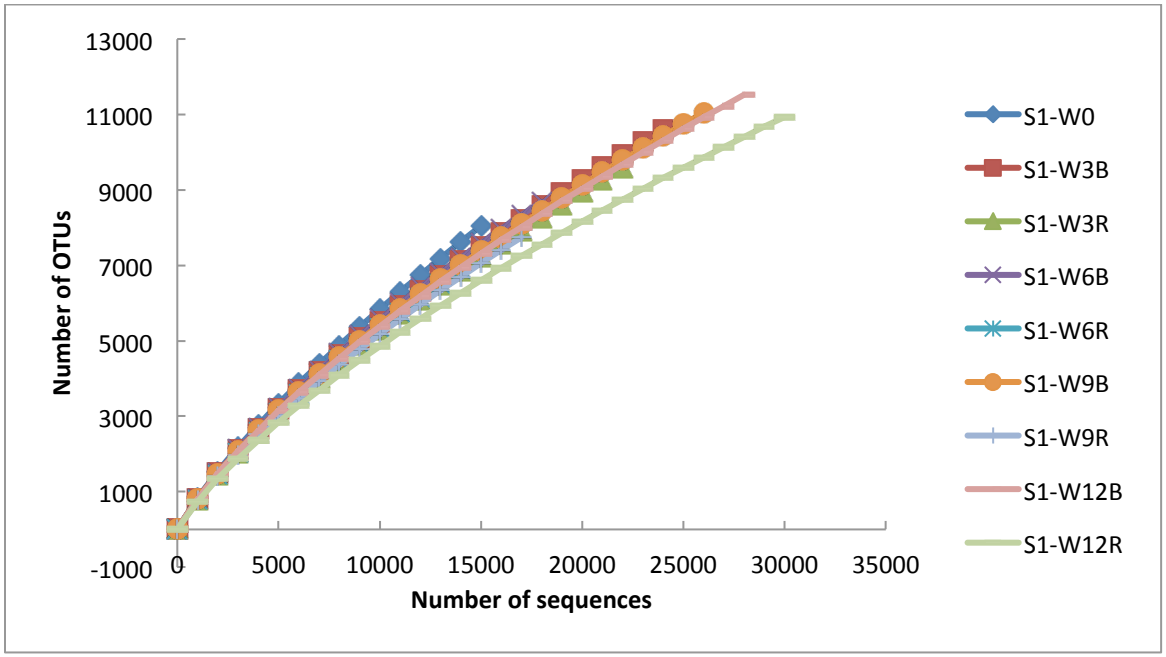


Module ID	# Nodes	Nodes w/ QS potential	% QS
1	98	32	33%
2	77	7	9%
3	33	7	21%
4	6	1	17%
5	5	1	20%
6	6	3	50%
7	47	1	2%
8	11	1	9%

Figure S3: Nodes highly similar to isolates with previously demonstrated quorum sensing (QS) capabilities for Season 1, Week 12. Numbers indicate module IDs, which corresponds to the table. Yellow nodes were >97% similar to isolates where QS activity was previously detected by whole-cell biosensors (DeAngelis 2006); gray nodes were <97% similar to isolates. Table indicates the corresponding number of nodes with QS potential per module. Module hubs are positioned in the center of modules.

DeAngelis (2006). Microbial community ecology and bacterial quorum sensing as control points in rhizosphere nitrogen cycling. Ph.D. thesis. University of California, Berkeley

A).



B).

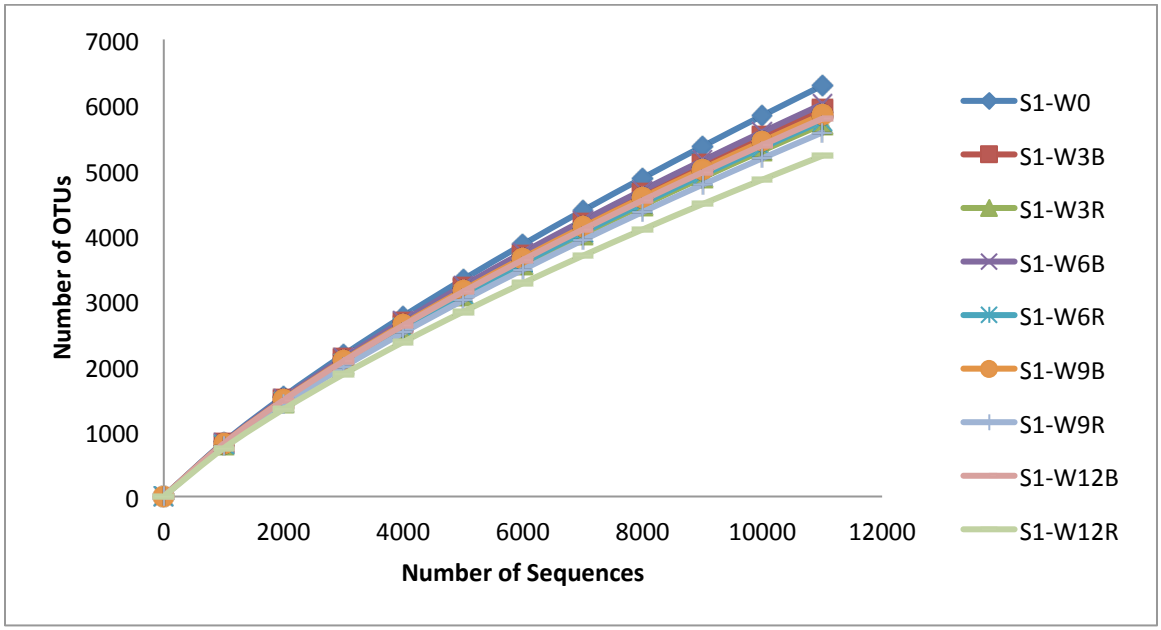


Figure S4. Mean rarefaction curves of the bulk and rhizosphere bacterial communities sampled at different growth stages of *Avena fatua*. Rarefaction curves are presnted (a) before and (b) after re-sampling to the same depth of 11914 sequences per sample (n=16).

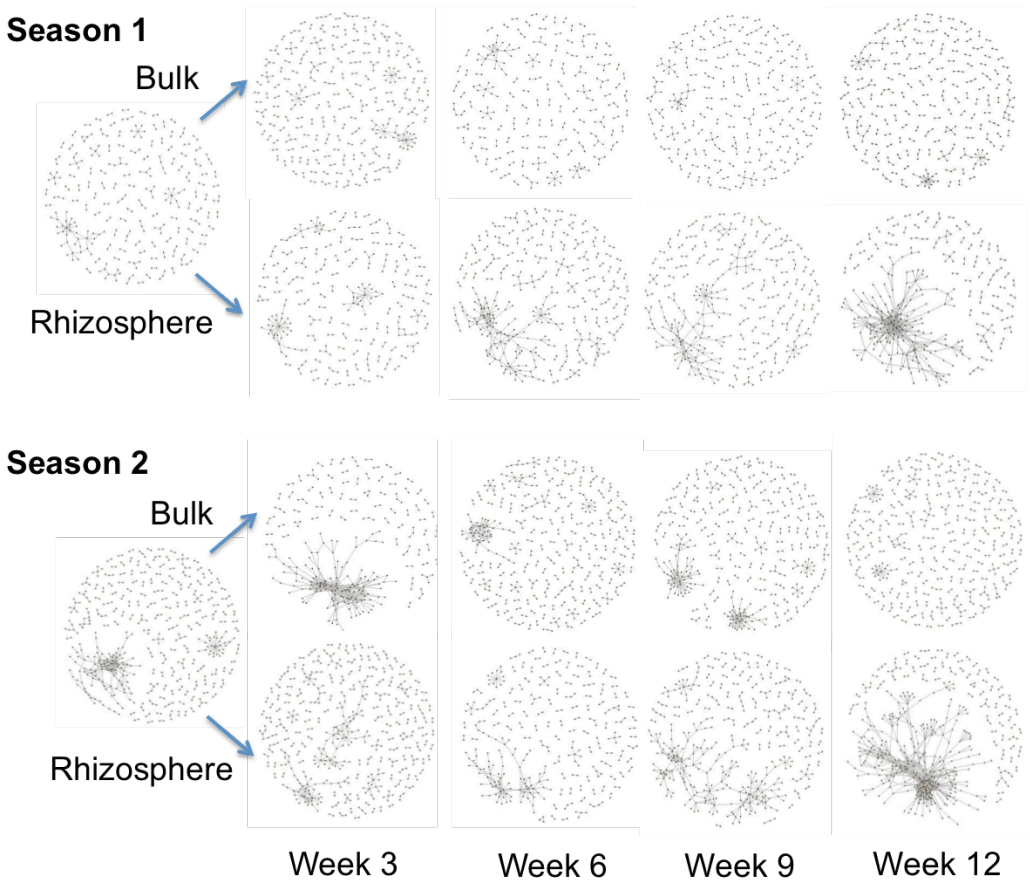


Figure S5. Succession of networks in rhizosphere and bulk soil over two seasons using log10 normalized data. Bulk soil from season 2 contains residual litter debris leftover from season 1.

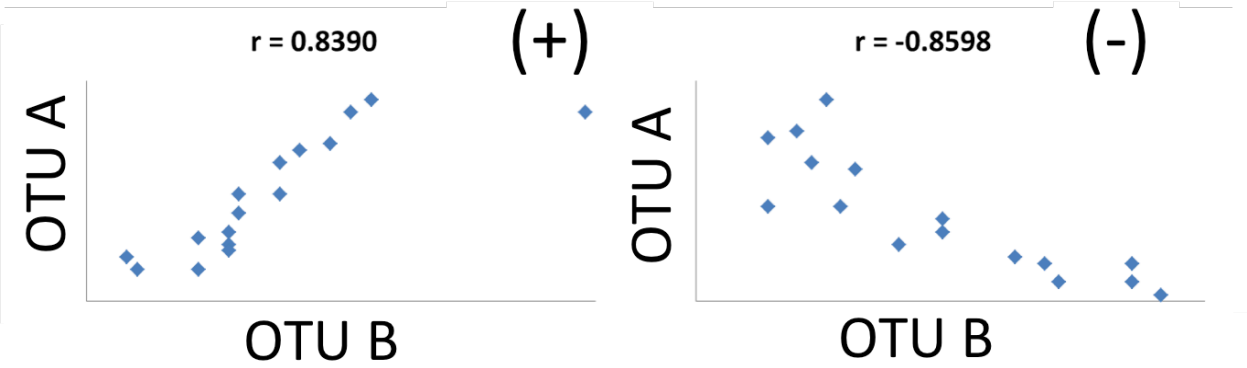


Figure S6: Example of pairwise correlations used to determine positive covariations (left graph) and negative covariations (right graph) between OTUs in this study (n=16).

Table S1. Topological properties of the empirical networks of pre-planted soil and rhizosphere microbial communities at different plant growth stages.

season		season 1					season 2				
soil		Pre-planted	Rhizosphere				Pre-planted	Rhizosphere			
sampling time		Week 0	Week 3	Week 6	Week 9	Week 12	Week 0	Week 3	Week 6	Week 9	Week 12
Commonly present OTU No.		1870	1773	1676	1655	1466	2039	1933	1805	1746	1633
Empirical Network	Total nodes	465	487	553	658	744	441	579	772	725	892
	Total links	286	323	829	861	1354	404	500	713	862	1057
	Average degree (avgK)	1.23	1.326	2.998	2.617	3.64	1.832	1.727	1.847	2.378	2.37
	Average clustering coefficient (avgCC)	0.024	0.046	0.131	0.124	0.179	0.062	0.071	0.091	0.115	0.098
	Harmonic geodesic distance (HD)	299.25	238.85	27.906	24.109	13.58	137.27	120.09	156.5	27.603	21.163
	Similarity threshold	0.82	0.82	0.82	0.81	0.81	0.83	0.82	0.8	0.81	0.79
	R^2 of power-law	0.921	0.964	0.867	0.93	0.906	0.735	0.972	0.946	0.937	0.919
	Modularity	0.99	0.98	0.79	0.795	0.731	0.800	0.930	0.944	0.852	0.861

Random networks*	Average clustering coefficient ± SD	0.003 ± 0.001	0.002 ± 0.001	0.017 ± 0.004	0.008 ± 0.002	0.016 ± 0.003	0.010 ± 0.003	0.002 ± 0.002	0.002 ± 0.002	0.004 ± 0.002	0.004 ± 0.002
	Average Harmonic geodesic distance ± SD	276.070 ± 8.884	166.840 ± 23.342	5.515 ± 0.207	6.487 ± 0.229	4.726 ± 0.120	12.220 ± 0.934	17.600 ± 1.233	15.070 ± 0.803	7.714 ± 0.260	7.684 ± 0.223

*Random networks were generated by rewiring all of the links of a corresponding empirical network with the identical nodes and links. Data were generated from 100 random runs and SD indicates the standard deviation from the 100 runs.

Table S2. Topological properties of the empirical networks of pre-planted and bulk/residual soil microbial communities at different plant growth stages.

season		season 1					season 2				
soil		pre-planted	Bulk				pre-planted	Residual			
sampling time		Week 0	Week 3	Week 6	Week 9	Week 12	Week 0	Week 3	Week 6	Week 9	Week 12
Commonly present OTU No.		1870	1925	1952	1939	1906	2039	2074	2108	2113	2132
Empirical Network	Total nodes	465	426	500	415	487	441	490	454	465	496
	Total links	286	250	297	302	313	404	476	351	358	295
	Average degree (avgK)	1.23	1.174	1.188	1.455	1.285	1.832	1.943	1.546	1.54	1.19
	Average clustering coefficient (avgCC)	0.024	0.007	0.014	0.048	0.01	0.062	0.057	0.06	0.063	0.009
	Harmonic geodesic distance (HD)	299.25	302.93	336.75	200.79	252.87	137.27	81.029	141.68	104.01	331.07
	Similarity threshold	0.82	0.82	0.82	0.83	0.82	0.83	0.83	0.83	0.83	0.82
	R^2 of power-law	0.921	0.99	0.964	0.94	0.96	0.735	0.855	0.927	0.931	0.989
	Modularity	0.990	0.991	0.992	0.961	0.986	0.800	0.822	0.939	0.926	0.986

Random networks*	Average clustering coefficient ± SD	0.003 ± 0.001	0.003 ± 0.000	0.005 ± 0.001	0.002 ± 0.002	0.003 ± 0.002	0.010 ± 0.003	0.008 ± 0.003	0.002 ± 0.002	0.002 ± 0.002	0.004 ± 0.001
	Average Harmonic geodesic distance ± SD	276.070 ± 8.884	286.690 ± 7.117	337.380 ± 4.527	41.740 ± 5.677	248.250 ± 11.097	12.220 ± 0.934	10.632 ± 0.776	30.070 ± 2.729	29.100 ± 2.390	326.830 ± 5.398

*Random networks were generated by rewiring all of the links of a corresponding empirical network with the identical nodes and links. Data were generated from 100 random runs and SD indicates the standard deviation from the 100 runs.

Table S3: F table for most parsimonious two-way ANCOVA model analyzing the relationship between the number of nodes, time, sample type (rhizosphere vs. bulk), and season after model simplification.

Model: nodes ~ time * sample_type * season - time:sample_type:season - time:season

Terms	Df	Sum Sq	Mean Sq	F value	*p*	Significance
time	1	89776	89776	45.8	9.07E-06	***
sample_type	1	140616	140616	71.7	7.02E-07	***
season	1	15401	15401	7.86	0.0141	*
time:sample_type	1	68807	68807	35.1	3.71E-05	***
sample_type:season	1	10080	10080	5.14	0.0397	*
Residuals	14	27449	1961			

Abbreviations: Df (Degrees of freedom); Sum Sq (Sum of Squares); Mean Sq (Mean Square);
*** $p < 0.001$, * $p < 0.05$

Table S4: F table for most parsimonious two-way ANCOVA model analyzing the relationship between the number of links, time, sample type (rhizosphere vs. bulk), and season after model simplification. Season was not significant and was removed during model simplification.

Model: links ~ time * sample_type

Terms	Df	Sum Sq	Mean Sq	F value	*p*	Significance
time	1	422714	422714	43.3	6.29E-06	***
sample_type	1	743822	743822	76.3	1.75E-07	***
time:sample_type	1	522580	522580	53.6	1.72E-06	***
Residuals	16	156053	9753			

Abbreviations: Df (Degrees of freedom); Sum Sq (Sum of Squares); Mean Sq (Mean Square);
*** $p < 0.001$

Table S5. Number of modules (with nodes > 4) present in surrounding and rhizosphere networks.

	Soil network	pre-planted	rhizosphere				bulk/residual			
	Time points	Week 0	Week 3	Week 6	Week 9	week 12	Week 3	Week 6	Week 9	week 12
Module	Season 1	13	8	13	23	23	6	10	9	16
	Season 2	4	11	36	23	36	12	10	9	11
Nodes	Season 1	74	76	265	359	494	35	63	76	119
	Season 2	57	189	388	411	623	143	107	110	56
Links	Season 1	66	92	829	693	1210	32	55	115	106
	Season 2	192	282	489	681	906	284	164	162	66