PNAS
www.pnas.org

Supplementary Information for:

Large-scale replicated field study of maize rhizosphere identifies heritable microbes

William A. Walters, Zhao Jin, Nicholas Youngblut, Jason G. Wallace, Jessica Sutter, Wei Zhang, Antonio González-Peña, Jason Peiffer, Omry Koren, Qiaojuan Shi, Rob Knight, Tijana Glavina del Rio, Susannah G. Tringe, Edward S. Buckler, Jeffery L. Dangl, and Ruth E. Ley

Corresponding Author: Ruth Ley
Email: rley@tuebingen.mpg.de

This PDF file includes:

Supplementary text
Figs. S1 to S6
Tables S1 to S6
References for SI reference citations

Supplemental Information Text

Maize rhizosphere soil collection and preparation

The planting, handling, collection, and DNA extraction of the 2010 samples were as described in

Peiffer et al (1), which included a subset (week 12) of the samples used in the current analysis (note that the

Peiffer study used 454 sequencing). The maize inbred cultivars sampled included: B73, B97, CML103,

CML228, CML247, CML277, CML322, CML333, CML52, CML69, Hp301, Il14H, Ki11, Ki3, Ky21,

M162W, M37W, Mo17, Mo18W, MS71, NC350, NC358, Oh43, Oh7B, P39, Tx303, and Tzi8.

 Rhizosphere samples were collected for weeks 1-15 and week 20 in the New York state fields of

Lansing and Aurora. The Ithaca field was sampled from weeks 2-15, and again on week 20. Urbana (state

of Illinois) and Columbia (state of Missouri), fields were sampled on week 12. Sampling was destructive, so

each sample represents an individual plant. Approximately 5 cm of root sample from the base of the plant,

along with adherent soil particles, were collected for each plant, placed on dry ice for transport and stored at

-80 °C until processing (more details can be found in (1)). Bulk soil was collected from between rows of

maize in the New York fields at weeks 10, 15, and 20. A total of 4,866 maize rhizosphere and bulk soil

samples were collected. Total genomic DNA was extracted using the MoBio Powersoil DNA isolation kits

(Mo Bio Laboratories, Carlsbad CA) in 2010. Ribosomal small subunit (SSU) amplicons were generated

using V4-region primers (2012-2013), with adapters and a random 0-3 base pair spacer added before the

barcode as described by Tremblay et al (2). The samples were sequenced (53 batches) on the MiSeq

Illumina system at the Joint Genome Institute in Walnut Creek, California.

As a partial replication study, five inbred maize lines (B73, CML277, HP301, IL14H, and Mo17)

were planted in the same Aurora, New York field as above on May 24, 2015. Bulk soil samples were

collected between plant rows immediately after maize was planted. Maize roots were collected from one

plant per row as in Peiffer et al (1), from weeks 4 to 12 of plant development. These samples were frozen at

-80°C and processed as a single batch in 2015, using the MoBio PowerSoil Kit to extract DNA. Samples were PCR-amplified with the SSU V4 hypervariable primers as described by Caporaso et al (3). Barcoded constructs were sequenced on the Illumina MiSeq platform (Cornell Biotechnology center).

16S rRNA amplicon sequence processing and analyses

2010 sequencing reads were processed using QIIME (3) 1.8.0 dev release (github SHA-1 hash e6e15f4aebb623313b5ff767b5e84b356b472294) to handle JGI interleaved format (extract_reads_from_interleaved_file.py). QIIME 1.9.0 was used for downstream processing. Paired end reads were stitched using fastq-join (4), and sequences were demultiplexed with default quality filtering settings. The total stitched reads were 627,638,736. A total of 567,013,569 reads then passed the default quality filtering of QIIME's split_library_fastq.py script. Primers were removed from the demultiplexed reads using a custom script (https://gist.github.com/walterst/ab88ae59a8900a2fa2da). 2015 data were stitched using fastq-join, and demultiplexed with QIIME using default parameters. All demultiplexed reads from 2010 and 2015 were clustered via open-reference OTU (Operational Taxonomic Unit, defined here at 97% identity to the centroid sequence of a cluster) picking against the Greengenes (5) August 2013 release. OTUs classified as mitochondria or chloroplast, or of less than 10 sequences, were removed. The number of samples remaining after these filtering steps were 4858, with a total sequence count of 453,674,044 and a total of 45,954 OTUs. Taxonomy plots were generated by filtering samples with less than 10000 sequences, and plotting the summarized genus level data in R (6) 3.3.2, using the packages tidyr (7) (0.6.0), ggplot2 (8) (2.2.0), and dplyr (9) (0.5.0), and with the QIIME summarize_taxa_through_plots.py script. Core microbiome tests to detect the presence of OTUs across samples were performed using the QIIME's compute_core_microbiome.py script on rhizosphere samples containing at least 10,000 sequences per sample.

Beta diversity and Adonis testing

Adonis (R vegan package) used unweighted and weighted UniFrac distance matrices on rarefied data (10,000 sequences per sample), with age and field location (for 2010 data) as factors. Beta diversity plots

were generated with the same 10,000 sequences per sample rarefied data as for Adonis testing, using the

QIIME beta_diversity_through_plots.py script and the default unweighted and weighted UniFrac metrics.

Genotype versus Beta diversity distances

NAM kinship distances were generated by querying NAM founders in maize Hapmap 2 (10), filtering

them for sites with calls in at least 10 of the founders, and using TASSEL (11) to make a genetic distance

matrix. R^2 values were calculated using the beta diversity distance versus NAM kinship distance for each

pair of samples.

Differential OTU abundance

For statistical testing with linear mixed models across OTUs for enrichment/depletion according to

maize group or inbred line, OTUs that were not present in 80% of samples were filtered out with QIIME's

filter_otus_from_otu_table.py script. Significance testing was done in R using the packages lme4 (12) (1.1-

12), lmerTest (13) (2.0-33), lsmeans (14) (2.25-5), vegan (15) (2.4-2), and MuMIn (16) (1.15.6). The model

used for the 2010 data was:

lme4::lmer(OTU_abundance~maize_type + maize_age + standardized_counts + (1|environment))

For 2015 data, this model was used:

lm(OTU_abundance~maize_type + maize_age + standardized_counts)

Where OTU_abundance is the cube-root transformed count data of OTUs, maize_type is the

classification of maize line among the major subpopulations ('tropical','stiff-stalk', etc.) or the particular

inbred line the maize belonged to (e.g. IL14H, Mo17, etc.), maize_age is the age in weeks of the sample,

standardized_counts is the standardized total read count per sample (see below), and environment is the

field the sample was taken from. Cube-root transformed OTU abundances were chosen because linear q-q

plots of residuals were obtained in most cases for this transformation. As sequence count data were many

orders of magnitude larger than OTU abundance, these were transformed with the vegan decostand

"standardize" function to yield the standardized_counts used in the above model. Pairwise tests with a

Tukey method p-value correction were determined with lsmeans:

lsmeans::lsmeans(model, pairwise~maize_type, mode = "kenward-roger")

Due to the number of samples involved, the default pbkrtest.limit of 3000 was increased via lsm.options(pbkrtest.limit = 10000). Resulting p-values were corrected with a Bonferroni approach for the 792 OTUs tested for the 2010 data and the 2557 OTUs tested for the 2015 data.

Alpha Diversity

Mean alpha diversity values for phylogenetic diversity (PD), Shannon index, and observed OTUs (full data, not the 80% shared data used above) were calculated with QIIME at 10,000 sequences per sample with 10x repeats. The model and significance testing was the same as for OTU abundance testing, however, the mean alpha diversity values were not transformed as the residuals showed linearity on q-q plots for the raw data.

Microbial responses to climate

To assess the impact of weather events on the rhizosphere microbiome, we obtained climate data (precipitation and air temperature) for the 2010 and 2015 season and used a linear mixed model to ascertain which taxa had relative abundances that correlated with rain or temperature. Precipitation data (rainfall amounts) were totalled for 1, 2 and 3 days and the direction of the effect (i.e., did rain correspond to an increase or decrease in the relative abundance of an OTU) was calculated. Climate data were queried for precipitation (inches of rain) and air temperature (degrees Fahrenheit)from the National Centers for Environmental Information (https://www.ncdc.noaa.gov/). The following station numbers were used, as these were the closest stations provided complete precipitation and temperature data: GHCND:USC00300331 Aurora NY (2010 and 2015); GHCND:USC00304174 Ithaca/Lansing NY (2010); GHCND:USC00231801 Columbia, MO (2010); GHCND:USW00094870 Champaign-Urbana, IL (2010).

The abundance of OTUs (80% shared data, as above) were correlated to climate variables with a linear model for the 2015 data:

lm(OTU_abundance~climate_measurement + week + standardized_counts)

For the 2010 data, a mixed model was used:

lmer(OTU_abundance~climate_measurement + week + standardized_counts + (1|environment))

Climate_measurement was either temperature or precipitation, and OTU relative abundance data

were transformed (cube root) as done for enrichment between maize lines. Raw sequences counts were transformed to the standardized_counts as for differential OTU abundance testing, above. The weather (temperature or precipitation) tested included the average air temperature for the same day that the sample was taken, the precipitation of the same day, the sum of precipitation for the last two days, or the sum of the precipitation for the last three days.

Heritability analysis

To calculate the broad-sense heritability of individual OTUs, the consolidated OTU table was subset to samples with at least 10,000 reads and to OTUs present across at least 80% of samples. OTUs were then normalized to fractional amounts within each sample. Relative abundances were log-transformed to more closely fit a normal distribution, and a mixed linear model was fit in R (6) using the lme4 (12) package. The fit model was:

$OTU_i \sim sample\_depth + week + location + inbred\_nested$

where OTU_i is the individual value for the ith OTU, *sample_depth* is the total sample depth across OTUs present in 80% or more of samples (as opposed to across all OTUs), *week* is the sampling week, *location* is the sampling field, and *inbred_nested* is the effect of each inbred line nested within each week-location combination. Sample_depth was fit as a fixed effect while week, location, and inbred_nested were all fit as random effects. Broad-sense heritability was estimated as the amount of variance explained by the inbred_nested term relative to the total variance for each model. Null distributions of heritability were calculated by randomly permuting OTU values across all samples 5000 times and running the same analysis on each permutation. All bioinformatics scripts for this analysis are available at: https://github.com/wallacelab/ley_rhizosphere_2017 . To test for sample size effects on heritability, a random subset of samples from the 2010 dataset (the same maize lines as 2015, 5 samples per week) were tested for heritability values. This was repeated ten times, giving H2 values in the range of 0.329 to 0.464.

Variance decomposition

Variance decomposition of the overall community structure was determined by running linear regression on the principal components across all sampling times and locations. A simple linear regression

was run in R with the following model:

$$PC_i \sim week * location * inbred$$

Where PC_i is the *i*th PC, *week* is the sampling week, *location* is the sampling location, *inbred* is the individual inbred line, and "*" indicates the inclusion of all main effects and all possible two- and three-way combinations. (Changing the order of model terms did not significantly change the results because the sampling data are largely balanced.) Variance was calculated as the sum-of-squares explained by each model component, including residuals. All bioinformatics scripts for this analysis are available at https://github.com/wallacelab/ley_rhizosphere_2017 .

To generate the phylogenetic tree of the 80% shared OTUs, with the subset of heritable OTUs labeled, first the OTUs that were a match to the Greengenes August 2013 release were selected. This was 753 of the total 792 OTUs in the 80% shared data (and 136 of the significant 143 heritable OTUs). The Greengenes August 2013 full length 97% OTUs tree then was filtered to retain these 753 tips, using the QIIME filter_tree.py script. Taxonomies for these OTUs were curated to retain the most specific named taxonomy for the OTUs. This tree, the heritable tips, and the taxa associated with each OTU were then uploaded into the Interactive Tree of Life tool (17) where the major phyla were colorized by hand.

High resolution *Pseudomonas* OTUs

To generate the Bray-Curtis dissimilarity PCoA plots for the 100% identity sequences within the three dominant 97% *Pseudomonas* OTUs, the OTU mapping files for the 2010 and 2015 data were filtered to only contain the sequences clustered into the OTUs 974121, 961783, and 646549. Next, demultiplexed sequences (i.e., pre-OTU picking sequences) were parsed with QIIME's filter_fasta.py script to retain only those sequences present in the filtered OTU mapping file. These sequences were then collapsed into 100% identity sequences using QIIME's pick_otus.py script with the options: -m prefix_suffix -p 1000. Beta diversity was then calculated for the resulting OTU table using bray-curtis dissimilarity, with even sampling at 1000 sequences per sample. To correlate the abundances of these 100% *Pseudomonas* OTUs from year to year, the data were first filtered to remove samples and OTUs that had less than 1000 sequences. Then the mean relative abundance of each OTU for each year was calculated. The top 10 most abundant OTUs and

the full set of OTU relative abundances were correlated in R with the Pearson product-moment, and corresponding R^2 values, calculated from cor.test().

Sequencing Data deposition

ENA (European Nucleotide Archive) study accession numbers:

2010 maize data: PRJEB21985

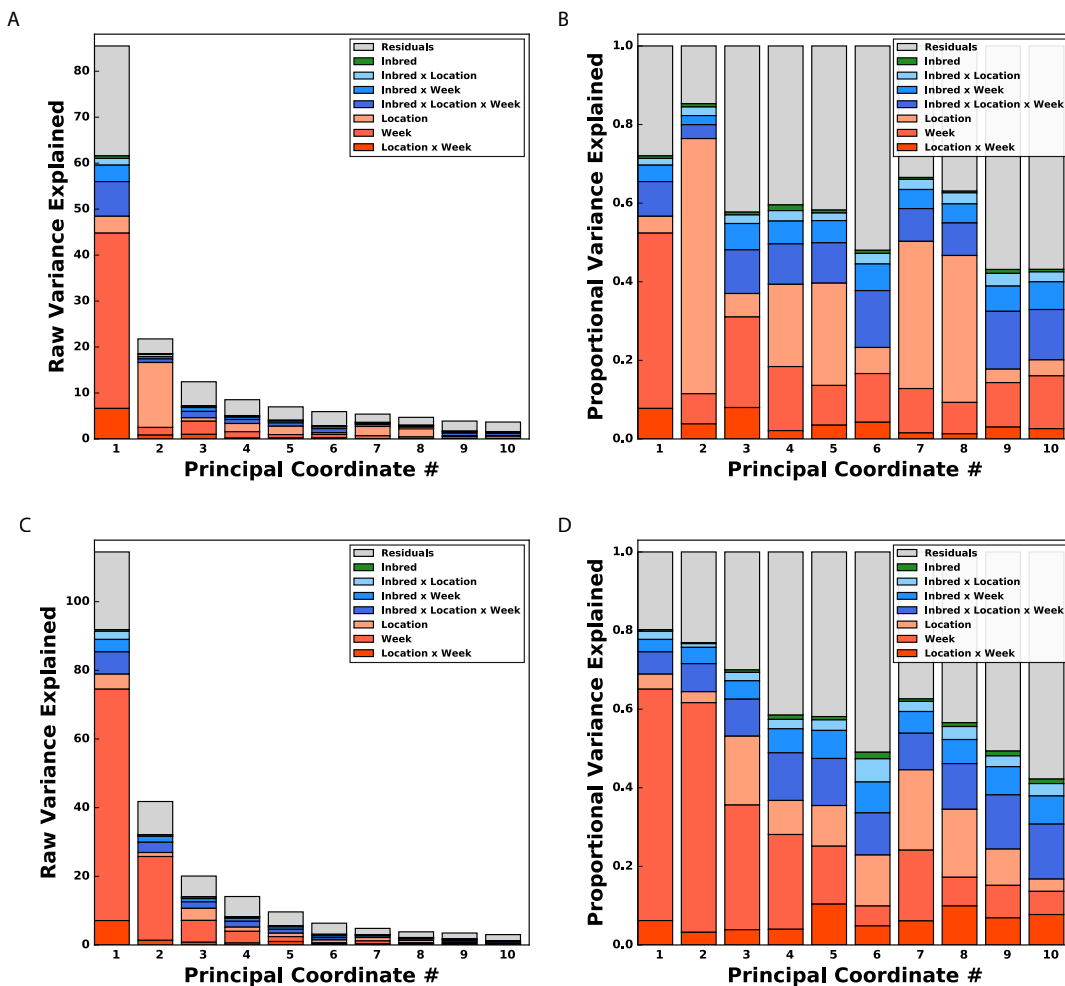2015 partial replication data: PRJEB21590

Supplemental Figures:

Figure S1. **Variance decomposition of principal coordinate analyses shown in Fig 1**. Variance decomposition for the first 10 principle coordinates of UniFrac-based ordination (shown in Fig 1) by week, location, inbred line, and the interactions between them, estimated with linear regression. Bars are colored based on whether proportional variance is attributed to genetic (green), environmental (orange), gene by environment (blue) or residual (gray) factors. Results for unweighted (A, B) and weighted (C, D) UniFrac are shown. Left panels (A, C) show both the fraction of variance explained by each factor as well as the total variance explained by the principal coordinate, while the right panels (B, D) show the proportional values of the factors normalized to 1.
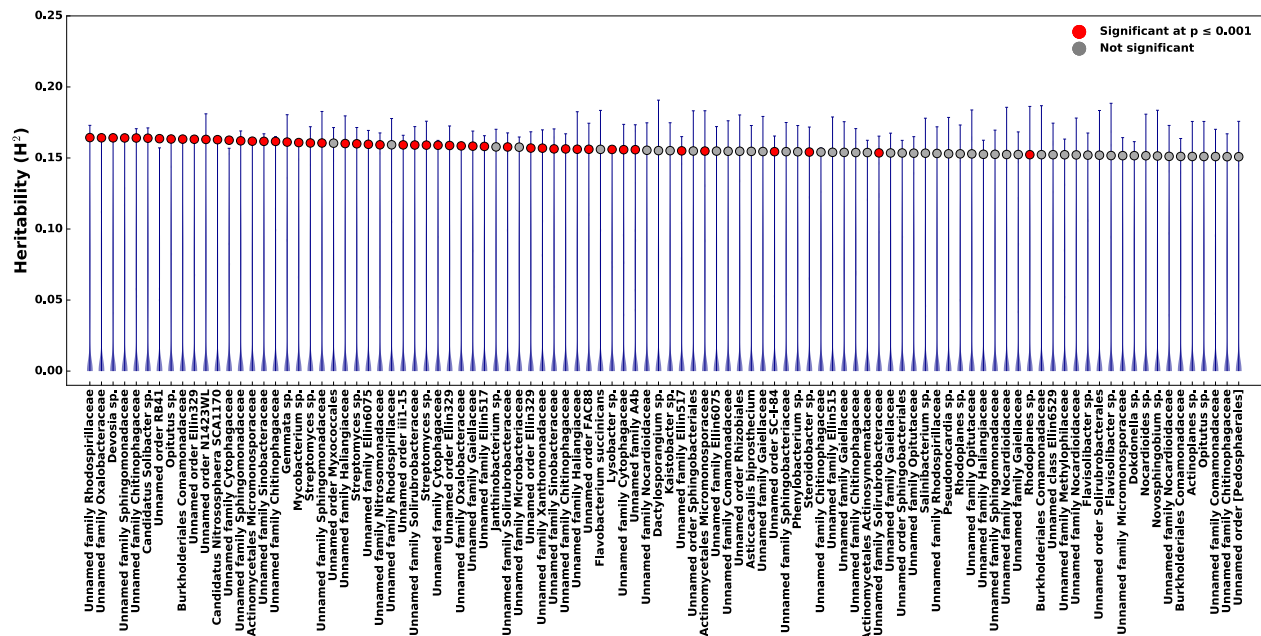
Figure S2. **Broad-sense heritability of individual OTUs for the 2010 field study.** The broad-sense heritability (H^2) is shown for the OTUs, sorted by H^2 values (these are OTUs 101-200 of the 200 OTUs with the highest H^2 values, a continuation of data from Fig 2). Circles show the actual H^2 values for each OTU in decreasing order and blue distributions show the corresponding H^2 values from 5000 permutations of the data. Red circles indicate OTUs with p-values ≤ 0.001. Taxonomies shown are most specific for each OTU.
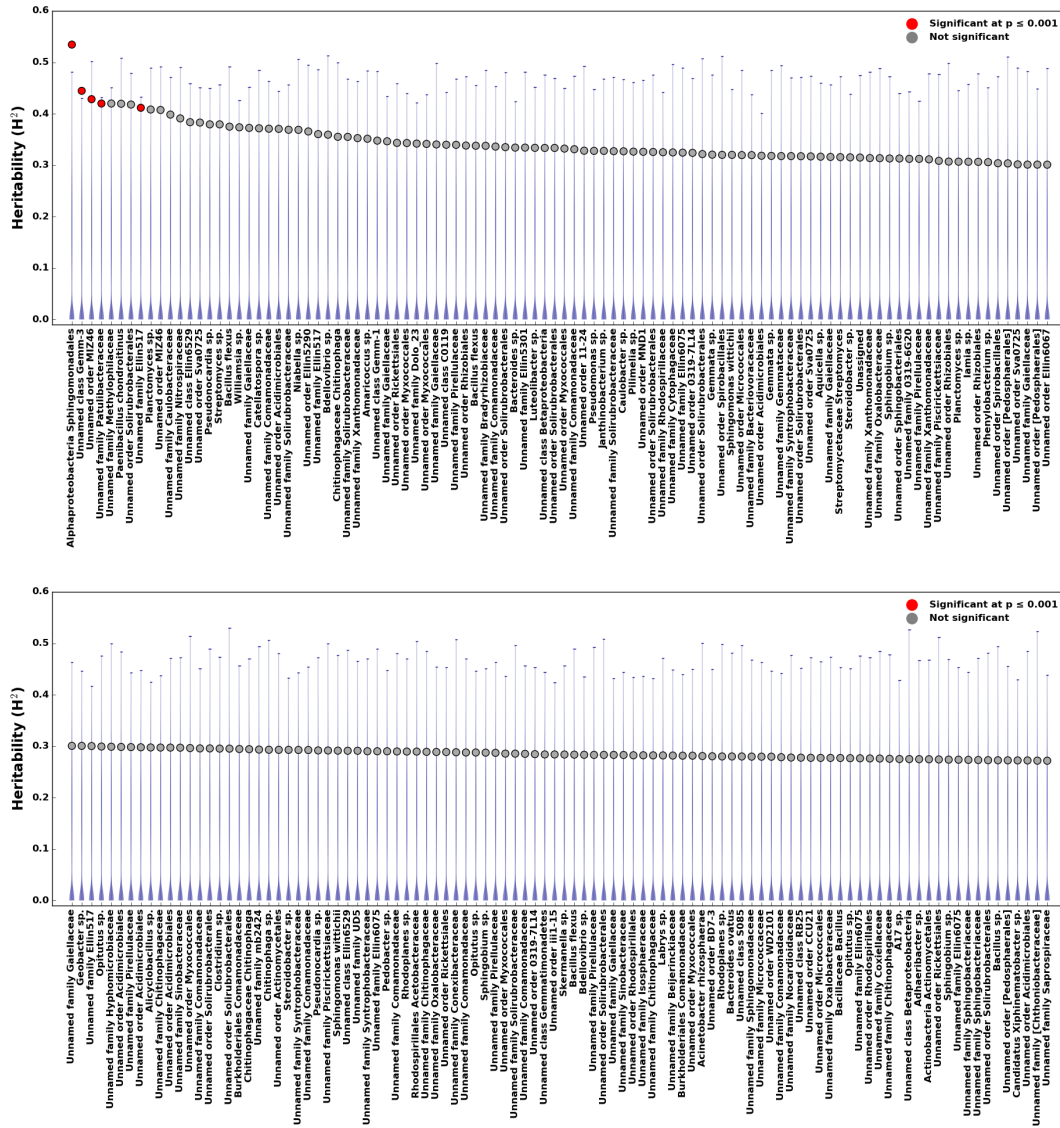
Figure S3. **Top 100 heritabilities for 2015 maize rhizosphere partial replication samples**. The broad-sense heritability (H^2) is shown for the 200 OTUs with highest H^2 values. Circles show the actual H^2 values for each OTU in decreasing order, and blue distributions show the corresponding H^2 values from 5000 random permutations of the data. Red circles indicate OTUs with empirical p-values ≤ 0.001. Five of these 200 OTUs show significantly higher heritability than expected by chance. Data also listed in Table S4.
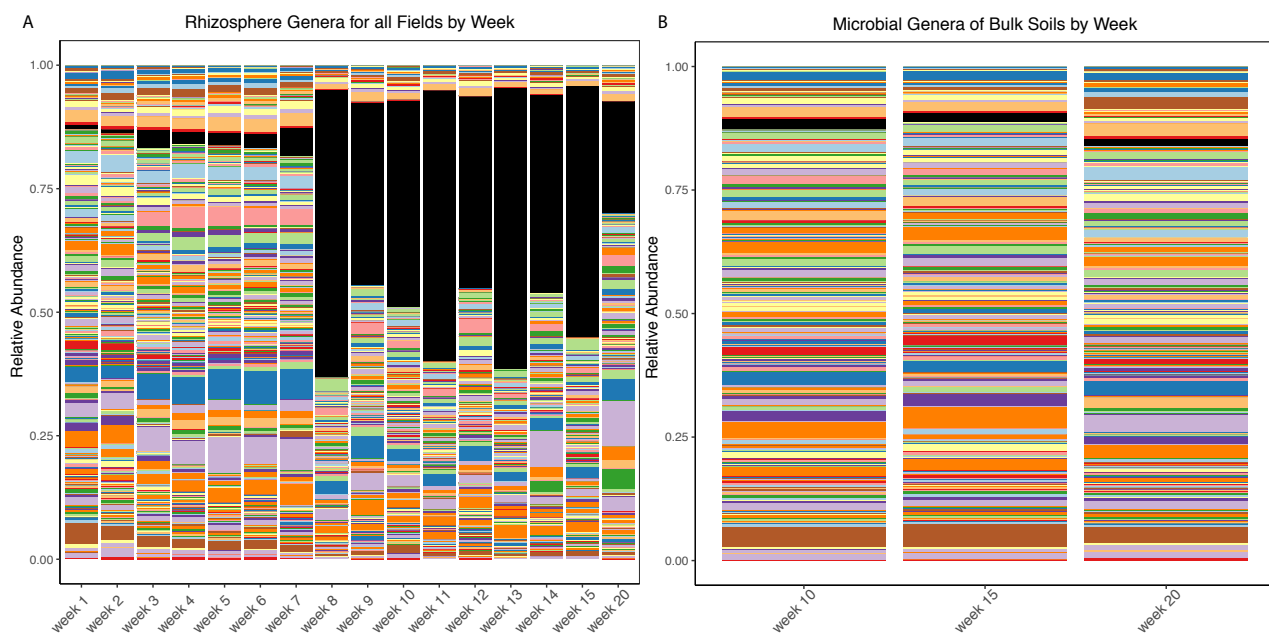
Figure S4. **Taxonomies of rhizospheres and bulk soils by sampling week.** A: Mean rhizosphere microbiota relative abundances for all fields and maize lines by week, colored by genera. *Pseudomonas* is colored black, and dominates the microbiota after the seventh week. Most genera are depressed in abundance in response to the Pseudomonas overgrowth, however at late time points, there is an increased abundance of *Pedobacter* and *Flavobacterium* spp. (lavender and dark green colors respectively) relative to pre-week eight abundances. B: Mean relative abundances of microbial genera for bulk soils by week, with all fields combined.
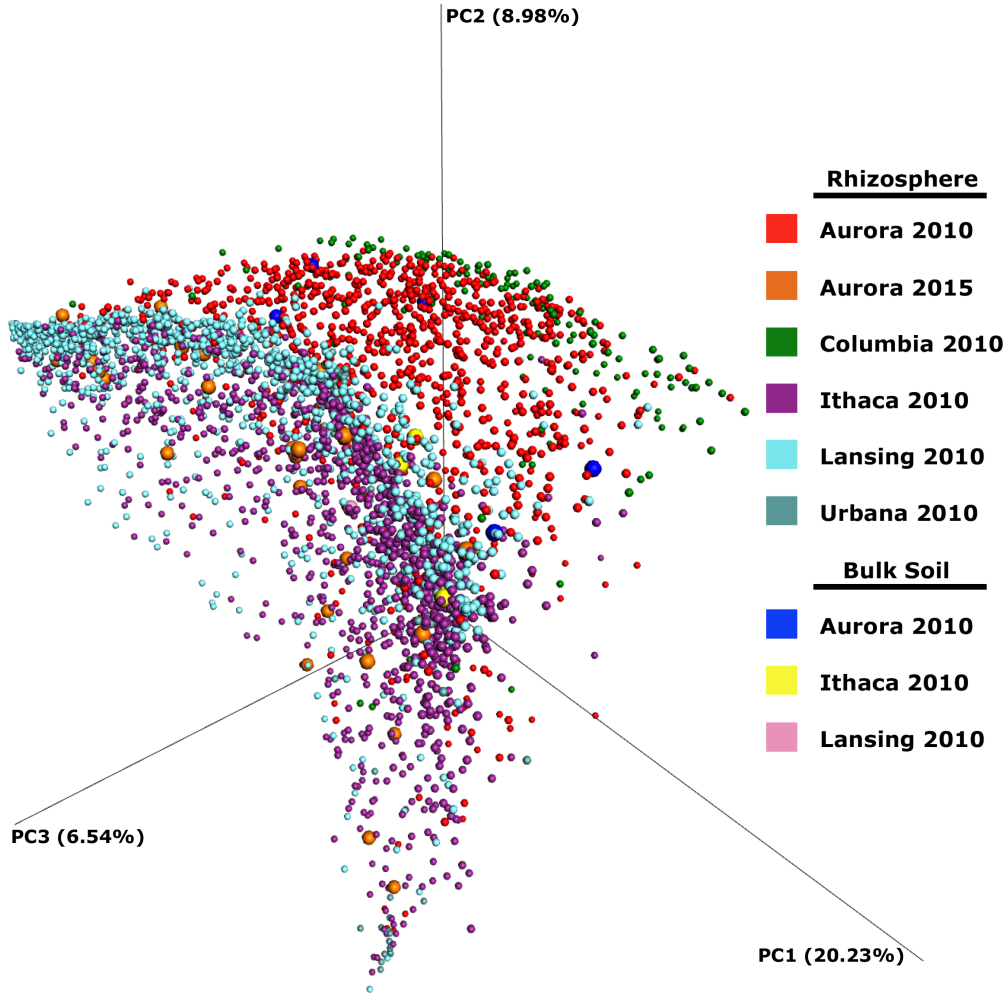
Figure S5. **Field-specific clustering by the three dominant *Pseudomonas* OTUs**. Clustering of samples by Bray-Curtis dissimilarity measures for unique sequences within the three largest *Pseudomonas* 97% OTUs from 2010 and 2015 data, colored by field and bulk soil categories. Several categories have few samples (or no samples in the case of 2015 bulk soils) after rarefaction due to low abundance of the three *Pseudomonas* OTUs in these categories. These samples have enlarged spheres for easier visualization. Percent of variation explained is shown by each principal coordinate (PC).
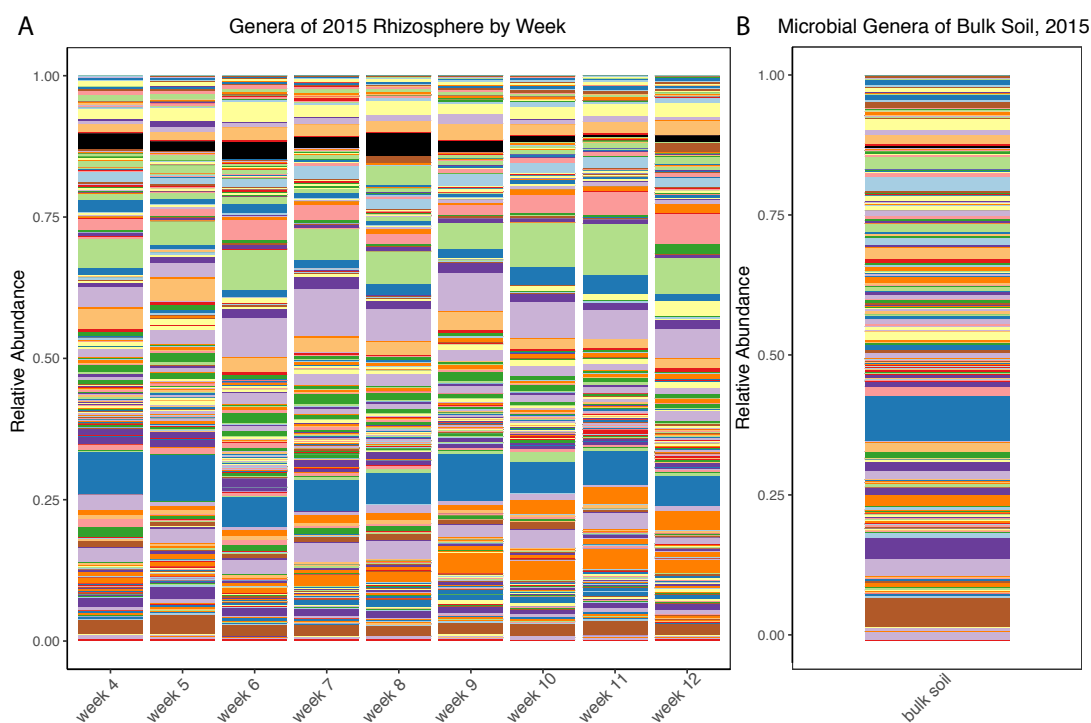
Figure S6. **Genera of 2015 maize rhizosphere and bulk soils, by week**. A. Microbial relative abundances for the 2015 Aurora field by week, colored by genera. B. Bulk soil for the 2015 Aurora field (average of three samples), colorized by genera. *Pseudomonas* is colored black.

Supplemental Tables:

Table S1. Core OTUs with Greengenes August 2013 taxonomies, detected in 100% and 95% of all maize rhizosphere samples.

Table S2. Pairwise differential abundance of OTUs with samples binned into broad maize groups and by inbred lines.

Table S3. Alpha diversity for the 2010 and 2015 samples. Only significant results (p<0.05) for differences between inbred lines or maize groups are shown. Values indicate mean alpha diversity measures for Observed OTUs, Shannon, and Faith's Phylogenetic Diversity.

Table S4. Heritability (H^2) and empirical p-values for 2010 and 2015 OTUs. OTUid shows Greengenes identifier or "New.Reference" for de novo OTUs. OTUs with the top 200 heritability scores are

shown

Table S5. Overlap between OTUs that are heritable and have pairwise differential abundance, for inbred maize lines and binned broad maize categories.

Table S6. Goodness of fit (R^2) between temperature or precipitation (same day, two day, or three day sum) and maize rhizosphere OTU abundances for 2010 and 2015.

References:

1.   Peiffer JA, et al. (2013) Diversity and heritability of the maize rhizosphere microbiome under field conditions. *Proc Natl Acad Sci* U S A 110(16):6548–6553.

2.   Tremblay J, et al. (2015) Primer and platform effects on 16S rRNA tag sequencing. *Front Microbiol* 6:771.

3.   Caporaso JG, et al. (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 7(5):335–336.

4.   Aronesty E (2013) Comparison of sequencing utility programs. Open Bioinforma J 7(1). Available at: https://benthamopen.com/contents/pdf/TOBIOIJ/TOBIOIJ-7-1.pdf.

5.   McDonald D, et al. (2012) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J* 6(3):610–618.

6.   R Core Team (2016) R: A Language and Environment for Statistical Computing. Available at: https://www.R-project.org/.

7.   Wickham H (2016) tidyr: Easily Tidy Data with `spread()` and `gather()` Functions. Available at: https://CRAN.R-project.org/package=tidyr.

8.   Wickham H (2009) ggplot2: Elegant Graphics for Data Analysis. Available at: http://ggplot2.org.

9.   Wickham H, Francois R (2016) dplyr: A Grammar of Data Manipulation. Available at: https://CRAN.R-project.org/package=dplyr.

10.   Chia J-M, et al. (2012) Maize HapMap2 identifies extant variation from a genome in flux. *Nat Genet* 44(7):803–807.

11.   Bradbury PJ, et al. (2007) TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23(19):2633–2635.

12.   Bates D, Mächler M, Bolker B, Walker S (2015) Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67(1):1–48.

13.   Kuznetsova A, Bruun Brockhoff P, Haubo Bojesen Christensen R (2016) lmerTest: Tests in Linear Mixed Effects Models. Available at: https://CRAN.R-project.org/package=lmerTest.

14.   Lenth RV (2016) Least-Squares Means: The R Package lsmeans. *Journal of Statistical Software* 69(1):1–33.

15.   Oksanen J, et al. (2012) vegan: Community Ecology Package Available at: http://CRAN.R-project.org/package=vegan.

16.   Bartoń K (2016) MuMIn: Multi-Model Inference. Available at: https://CRAN.R-project.org/package=MuMIn.

17.   Letunic I, Bork P (2016) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res* 44(W1):W242–5.