Title: Comparing faster evolving *rplB* and *rpsC* versus SSU rRNA for improved microbial community resolution

Authors: Jiarong Guo[a], James R. Cole[a], C. Titus Brown[b], James M. Tiedje[a]

Author affiliation:

[a]Center for Microbial Ecology, Michigan State University

[b]Department of Population Health and Reproduction, University of California, Davis

Corresponding author:

James M. Tiedje

tiedjej@msu.edu

Author contributions: J.G. performed the analyses under the supervision of J.M.T., C.T.B. and J.R.C.. All also helped with the analysis approaches and writing of the paper.

Abstract:

Amplicon sequencing of the SSU rRNA gene is standard for microbial ecology but has several drawbacks including limited resolving power for taxa below the level of genus and variable multiplicity presenting difficulties in quantifying different organisms. Many conserved protein-coding core genes are single copy and evolve faster than the SSU rRNA gene but their use has been precluded by the lack of universal primers for amplicon sequencing. Recent advances in gene targeted assembly methods for large shotgun metagenomes make their use feasible. To evaluate this approach, we compared the variation of two single copy ribosomal protein genes, *rplB* and *rpsC,* with the SSU rRNA gene for all completed bacterial genomes in NCBI RefSeq. As expected, among pairwise comparisons of all species that belong to the same genus, 94.9% and 91.0% of the pairs of *rplB* and *rpsC,* respectively, showed more variation than did their SSU rRNA sequences. To circumvent primer bias and lack of universal primer issues of amplicon methods, we used a gene targeted assembler, Xander, to assemble *rplB* and *rpsC* from shotgun metagenomic data. When tested on rhizosphere samples of three crops -- corn, an annual, and *Miscanthus* and switchgrass, both perennials -- both genes separated all three communities while SSU rRNA gene could only separate the annual from the two perennial communities in ordination analyses. Furthermore, the Xander assemblies of *rplB* and *rpsC* yielded significantly higher numbers of OTUs (alpha diversity) than SSU rRNA gene recovered from short reads and from amplicon data. These results confirm the better resolving ability of these faster evolving marker genes for comparative microbiome studies.

Importance:

High resolution marker genes are central to determining diversity of communities and differences between or among communities. Many ecologically determinative features occur at genetic levels not resolved by the relatively conserved SSU rRNA gene; hence marker genes are needed with finer community resolution. Further, if they were single copy, counting would be more accurate than for the variable copy SSU rRNA genes. The rapid advancement of shotgun sequencing and metagenome assembly has enabled us to avoid the need for and the inevitable bias of primers, to recover single-copy protein-coding genes directly from shotgun metagenomes. Targeting a few genes for assembly, like those coding for ribosomal proteins, samples more organisms and speeds the analysis over using whole genome assemblies for this purpose.

Introduction:

Shaped by 3.5 billion years of evolution, microorganisms are estimated to comprise up to one trillion species and the majority of genetic diversity in the biosphere (1). However, our understanding of this diversity is limited because of this huge number, and that the majority are yet to be cultured and their physiology or functions characterized. Since the pioneering work of Carl Woese in the late 1970s, the SSU rRNA gene has been the dominant marker used in microbial community structure analyses (2–5). While it has been extremely useful to advancing understanding of the microbial world, it does have important limitations, namely that it is highly conserved and that there are usually multiple copies and some with intra-genomic variations making this gene problematic for taxonomic identification at species and ecotypes levels and incapable of reflecting community distinctions at ecologically meaningful levels (6–8).

With the accelerated accumulation of microbial genomes in NCBI in recent years (9), whole genome-based comparison is now feasible and a more accurate method for species and strain identification (9–15). However, whole genome-based comparison is computationally more expensive compared to marker gene comparison, and it is not yet possible to reliably obtain genome sequences of many members of natural microbial communities. Hence, marker gene analysis remains useful. Single copy protein coding housekeeping genes stand out as the best candidates. First, their single copy status provides more accurate species and strain counting, identification and OTU clustering than the SSU rRNA gene. Second, they are present in virtually all members of the three domains of life. Third, protein coding genes evolve faster than rRNA genes not only because rRNA genes are more conserved due to their critical role in ribosome function (16), but also because of the redundancy in the genetic code, especially at the third codon position (6).

Here, we evaluate two single copy protein coding genes, *rplB* (50S ribosomal large subunit protein L2), and *rpsC* (30S ribosomal small subunit protein S3) as potential housekeeping genes for phylogenetic markers for microbial community analyses. Earlier studies showed the potential of protein coding genes over SSU rRNA genes as higher resolution phylogenetic markers for microbial diversity analyses using both genomic data (111 genomes) and metagenomic data (< 6 Gbp by Sanger sequencing) (6, 8). We revisited this comparison with the now much larger data set - all completed bacterial genomes (~4500 with one contig) and then tested the resolving power of these two genes versus SSU rRNA gene among different crop rhizospheres using large shotgun metagenomic data (~1TB). The novelty of our analyses is the application of gene targeted assembly to recover single copy protein coding genes from shotgun metagenomic data (17) and the use of *de novo* OTU-based diversity analyses, commonly used in microbial diversity analyses, rather than just taxonomic identification as previous studies (6, 8).

Methods:

Bacterial genome assembly information from NCBI (`ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/assembly_summary.txt`) was used to construct the link to download each genome based on the instructions described in this link (`http://www.ncbi.nlm.nih.gov/genome/doc/ftpfaq/#allcomplete`). Command line "wget" was then used to retrieve the genome sequences with links obtained from the above step.

For extracting genes from genomes, the SSU rRNA gene HMM (Hidden Markov Model) from SSUsearch (18) was used to recover rRNA genes. Aligned *rplB* and *rpsC* nucleotide sequences of the "training set" retrieved from the RDP FunGene database (19) were used to build the HMM models using hmmbuild command in HMMER (version 3.1b2) (20). The nhmmer command in HMMER was then used to identify SSU rRNA, *rplB* and *rpsC* sequences from bacteria genomes obtained from NCBI using score cutoff (-T) of 60. Next, nhmmer hits of least 90% of the length of the HMM model were accepted as the target gene. For the purpose of comparing SSU rRNA and *rplB* and *rpsC* gene distances, one copy of the SSU rRNA gene was randomly picked from each genome. Pairwise comparison among gene sequences was done using vsearch (version 1.1.3) with "--allpairs_global --acceptall" (21). Three species of environmental interest, *Rhizobium leguminosarum*, *Pseudomonas putida* and *Escherichia coli*, were chosen for closer comparison of *rplB* and *rpsC* pairwise distances and SSU rRNA gene distances.

The shotgun data are from DNA from seven field replicates of rhizosphere samples of three biofuel crops: corn (C) *Zea maize*, switchgrass (S) P*anicum virgatum*, and *Miscanthus* x gigantus

(M) that had been grown for 5 years. Shotgun sequence data for the 21 samples were downloaded from the JGI web portal (http://genome.jgi.doe.gov/); JGI Project IDs are listed in Table S1. Raw reads were quality trimmed using fastq-mcf in EA-Utils (verison 1.04.662) (http://code.google.com/p/ea-utils) "-l 50 -q 30 -w 4 -k 0 -x 0 --max-ns 0 -X". Overlapping paired-end reads were merged by FLASH (version 1.2.7) (22) with "-m 10 -M 120 -x 0.20 -r 140 -f 250 -s 25" described in (18).

SSU rRNA gene amplicon data (JGI project ID: 1025756) from the same DNA used for shotgun sequence were trimmed the same way as shotgun data (described above). Paired ends were joined by FLASH (-m 10 -M 150 -x 0.08 -p 33 -r 200 -f 300 -s 25) (22) and primer sequences were removed by cutadapt (-f fasta --discard-untrimmed) (23). For community analyses, the open reference OTU picking method in QIIME was used for clustering and Bray-Curtis index was used for beta-diversity index (24).

For SSU rRNA gene analyses with shotgun data, SSU rRNA gene fragments and those aligned to the V4 region (*E. coli* position: 577 - 727) of each sample were identified using the SSUsearch pipeline (18) and clustered using RDP's McClust tool (25) at a distance of 0.05 and minimal overlap of 25 bp, following the tutorial in SSUsearch (http://microbial-ecology-protocols.readthedocs.io/en/latest/SSUsearch/overview.html).

Both *rplB* and *rpsC* sequences were assembled using Xander with "MAX_JVM_HEAP=500G, FILTER_SIZE=40, K_SIZE=45, genes = *rplB* and *rpsC*, MIN_LENGTH=150, THREADS=9" (17). Data for each crop were assembled separately. The assembled *rplB* or *rpsC* sequences (nucleotide and protein) from the three crops were pooled and clustered using RDP's McClust tool (25). For each gene, a table of OTU counts of each sample was made based on mean k-mer coverage of the representative sequence of each OTU (provided in "*_coverage.txt" output file from Xander). Further, diversity analyses were done with the vegan package in R using functions “rda” for ordination and “diversity” for Shannon diversity index, respectively, from the OTU (count) tables. An implementation of this pipeline is publicly available at https://doi.org/10.5281/zenodo.1438073.

To assess how many potential target gene reads of *rplB* and *rpsC* were assembled by Xander, we did a six-frame translation of the short reads (nucleotide sequences) into protein sequences by transeq in EMBOSS tool (26). We then searched HMMs against the protein sequences and the hits with bit score > 40 (e-value < $6.2 * 10^{-6}$) were treated as reads from the target gene. Meanwhile, “*_match_reads.fa”, a collection of reads that share a k-mer (k=45) with assembled sequences, output from Xander, provided the reads assembled by Xander. Then we compared the fold coverage of reads found by hmmsearch and reads used by Xander, by estimating fold coverage of each read with median kmer coverage using khmer package (27, 28).

Results:

A total of 4,457 of complete bacteria genomes defined as one sequence were downloaded. SSU rRNA gene copy number ranged from 1 to 16 with a mean of 4 and 99.9% of genomes have single copies of *rplB* and *rpsC* (Table S2). Both of these genes were present in 4,440 of the complete genomes. When evaluating intra-genomic variation among copies of SSU rRNA genes in completed genomes of *R. leguminosarum*, *P. putida* and *E. coli*, *E. coli* had the largest variation with a minimum of 95.4% identity (Fig. S1). For the pairwise comparison between genomes, one copy of each gene was randomly picked as a representative for genomes with multiple copies.

For the selected taxa, Rhizobiales, Pseudomonadales, *Rhizobium*, and *Pseudomonas*, *rplB* and *rpsC* had similar variations and both had larger variation among the genomes than SSU rRNA genes within their corresponding order (among genera), and genus (among species) (Fig. 1

and 2). When comparing all species of completed genomes that belong to the same genus, we found SSU rRNA gene has an identity range of 63.2% to 100.0% and a median of 95.2%, *rplB* has an identity range of 43.2% to 100.0% and a median of 87.2%, *rpsC* has an identity range of 46.0% to 100.0% and a median of 90.3%. Between *rplB* and SSU rRNA gene, 88,993 pairs (94.9% of total) has larger variation in *rplB*, 3,573 pairs have larger variation in SSU rRNA gene, and 1,167 pairs have the same variation (Fig. 3A); 77,885 pairs (91.0% of total) has larger variation in *rpsC*, 6,074 pairs have larger variation in SSU rRNA gene, and 1,622 pairs have the same variation (Fig. 3B); 54,755 pairs (63.7%) has larger variation in *rplB*, 28,393 pairs have larger variation in *rpsC* gene, and 2,808 pairs have the same variation for *rplB* and *rpsC* (Fig. 3C).

We compared SSU rRNA genes with *rplB and rpsC* to test the ability of shotgun data to resolve community differences among plant rhizospheres. We chose these two genes as they had a suitable length for Xander assembly, were long enough for resolving power, and had HMMs that were both specific and sensitive for fragment recovery due to their uniqueness in sequence as parts of the ribosome, and both have been used as phylogenetic marker in other shotgun metagenomic studies (29, 30). On average, 0.04% of total reads were identified as SSU rRNA gene fragments and 0.004% of total reads aligned to the 150 bp of V4 region of the gene with SSUsearch (18). Another 0.01% and 0.008% of total reads were identified as *rplB* and *rpsC,* respectively, by Xander (Table S3). To test the sensitivity of Xander, we found that the number of potential *rplB* and *rpsC* reads assembled were 49.5% and 47.9%, respectively, of those defined by hmmsearch with bit score cutoff of 40 (Table S4) and have much higher fold coverage than the rest of reads (excluding shared reads) in hmmsearch hits (Figure S2).

Beta diversity analyses of all three genes showed that the rhizosphere communities of the annual crop, corn, were different from those of the two perennial grasses, *Miscanthus* and switchgrass, but only *rplB* and *rpsC* distinguished the communities of the two perennial grasses (Fig. 4). This was true whether the analysis was at the nucleotide or protein level. The alpha diversity of the corn rhizosphere communities was significantly lower than those of *Miscanthus* and switchgrass rhizospheres by all three measures except for Chao1 index with *rpsC* and SSU rRNA gene (Fig. 5). When comparing among genes, the numbers of OTUs from *rplB* and *rpsC* are also significantly higher than SSU rRNA gene (Fig. 5). Since SSUSearch returns shorter fragments than Xander assembled genes, we also evaluated whether the longer fragments of SSU rRNA from amplicon data ~ 250 to 300 bp, could distinguish the two perennial grass communities, and they could not (Fig. 3E).

Discussion:

We confirmed the advantages of *rplB* and *rpsC* over the SSU rRNA gene as a more resolving phylogenetic marker using updated large genomic data (~4500 complete genomes) (Fig. 1, 2 and 3). We also demonstrated that *rplB* and *rpsC* can be assembled from large shotgun metagenomes and showed that they provided higher community resolution by separating *Miscanthus* and switchgrass rhizosphere samples while the SSU rRNA gene did not (Fig. 4). The two perennial grasses would be expected to have more similar microbiome than the annual since the latter is re-established each year while the fibrous perennial grass roots are more similar and not physically disturbed annually and thus do not have full regrowth at a new random site each year.

In large genomic data analyses, *rplB* and *rpsC* show advantages in following three aspects:

First, SSU rRNA gene, a multiple copy gene, poses difficulties for interpreting species abundance, while *rplB* and *rpsC* do not have the same issue as it is single copy genes in > 99.9% of complete genomes (Table S2). Additionally, variations among multiple SSU copies can cause multiple OTUs (sequence clusters) from the same species (Figure S1) and thus leads to overestimation of species richness (31). Since a single copy of the *rplB* and *rpsC* genes is contained in every cell in a community, the relative abundance of *rplB* and *rpsC* gene sequences provides a reference for estimating the fraction of organisms possessing other genes.

Second, *rplB* and *rpsC* are better able to differentiate closely related species based on their lower sequence similarities compared to the SSU rRNA gene in pairwise comparisons among genomes (Fig. 1, 2, and 3). This is consistent with the crucial role SSU rRNA plays in translation (ensuring translation accuracy) (16), also confirmed by another study showing SSU rRNA genes (along with LSU rRNA genes, tRNA and ABC transporter genes) to be the most conserved genes (32).

Third, SSU rRNA genes in genomes are also more prone to assembly errors (chimera) than single copy genes due to their higher overall nucleotide identity and the presence of highly conserved regions interspersed in SSU rRNA genes. Note that these erroneous sequences might be further collected by databases and used as references for taxonomy, alignment, and chimera detection, and thus have an impact on common microbial ecology diversity analyses. Switching to a single copy gene that is less prone to assembly error can mitigate the above problem.

Finally, this method provides for higher resolution community diversity analyses in large shotgun metagenomes, leveraging a scalable gene targeted assembler, Xander. Assembly is desirable for short read data to correctly identify the gene and provide enough length for resolving power, a major objective in ecology studies. Assembly misses the rarer species that do not have enough sequencing depth in metagenomes, confirmed by the higher fold coverage of reads used in assemblies compared to the other reads in hmmsearch hits (Fig. S2). We did find that the number of reads used in assemblies are about half of the reads identified as the targeted genes by hmmsearch (Table S4). The hmmsearch though could also have recovered some false positives due to mistaken short-read identification and thus overestimated the total gene number.

However, *rplB* and *rpsC* yield significantly higher alpha diversity (Fig. 5) than SSU rRNA gene despite missing rare members. Thus they reveal more diversity among abundant members than SSU rRNA gene, which offsets and exceeds the diversity of the rare members that are not assembled, further confirming their higher resolution.

We choose two protein coding genes to be sure our results were not gene specific, and both gave very similar results at both the nucleotide and protein levels. At least from extensive completed genomes, most of these two genes are single copy making quantitative (ratio) comparisons with other genes more consistent. For future use, *rplB* might have slight advantage over *rpsC* since it is longer, about 830 bp on average vs 660 bp of *rpsC*, providing a bit more resolving power, which is consistent with results in genome comparisons showing *rplB* has lower median sequence identity than *rpsC* (Fig 3).

It is of course possible to find in reference databases the best match to the assembled sequence of these marker genes and potentially have finer taxonomic resolution than provided by SSU rRNA. But, the reference database is only from sequenced genomes and hence is very unbalanced and incomplete compared to 16S rRNA databases (17) so this use is not generally beneficial at this time.

Although sequencing depth needed varies depending on community diversity, we estimate it based on our rhizosphere soil samples as a practical guide. The reads from *rplB* are around 0.01% of total (Table S3). Assuming a fold coverage of 3000 of *rplB* for each sample, to be comparable to 3000 amplicons in planning amplicon-based studies, one needs about 25 Gbp (3000 * 830 / 0.01%) of shotgun metagenome (830 bp is the average gene length of *rplB*). The major requirement for using this method beyond sufficient shotgun sequence depth is an access to a high performance computer since large memory (> 250 Gb recommended for soil samples) is needed to run Xander.

Conclusion:

We demonstrated that *rplB* and *rpsC*, single copy protein coding genes can provide finer resolution of taxa and hence better distinguish among communities than the more commonly used SSU rRNA gene and also provide finer scale *de novo* (OTU) diversity analysis. This method does require shotgun sequence of sufficient depth, so is currently more costly than amplicon based analyses, but as sequencing costs decline, capacity and access increase, read length grows, and genome reference databases grow, single copy protein coding genes such *rplB* and *rpsC* have the potential to complement or even replace the SSU rRNA gene as a phylogenetic marker and better reflect ecology of communities.

Acknowledgement:

We thank Ribosomal Database Project (RDP), Institute for Cyber-Enabled Research (iCER) and High Performance Computing Center (HPCC) at Michigan State University for technical support. Support for this research was provided by the U.S. Department of Energy, Office of Science, Program of Biological and Environmental Research (Awards DE-FC02-07ER64494, DE-FG02-99ER62848 and DE-SC0010715), and by the National Science Foundation Long-term Ecological Research Program (DEB 1637653) at the Kellogg Biological Station, and by Michigan State University AgBioResearch.

References:

1. Locey KJ, Lennon JT. 2016. Scaling laws predict global microbial diversity. Proc Natl Acad Sci USA 113:5970–5975.

2. Woese CR, Fox GE. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci USA 74:5088–5090.

3. Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, Pace NR. 1985. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proc Natl Acad Sci USA 82:6955–6959.

4. Huse SM, Dethlefsen L, Huber JA, Mark Welch D, Welch DM, Relman DA, Sogin ML. 2008. Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. PLoS Genet 4:e1000255.

5. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M, Gormley N, Gilbert JA, Smith G, Knight R. 2012. Ultra-high-

throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J 6:1621–1624.

6. Case RJ, Boucher Y, Dahllof I, Holmstrom C, Doolittle WF, Kjelleberg S. 2007. Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. Appl Env Microbiol 73:278–288.

7. Wu M, Eisen JA. 2008. A simple, fast, and accurate method of phylogenomic inference. Genome Biol 9:R151.

8. Roux S, Enault F, Bronner G, Debroas D. 2011. Comparison of 16S rRNA and protein-coding genes as molecular markers for assessing microbial diversity (Bacteria and Archaea) in ecosystems. FEMS Microbiol Ecol 78:617–628.

9. Land M, Hauser L, Jun SR, Nookaew I, Leuze MR, Ahn TH, Karpinets T, Lund O, Kora G, Wassenaar T, Poudel S, Ussery DW. 2015. Insights from 20 years of bacterial genome sequencing. Funct Integr Genomics 15:141–161.

10. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. 2007. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol 57:81–91.

11. Luo C, Walk ST, Gordon DM, Feldgarden M, Tiedje JM, Konstantinidis KT. 2011. Genome sequencing of environmental Escherichia coli expands understanding of the ecology and speciation of the model bacterial species. Proc Natl Acad Sci USA 108:7200–7205.

12. Varghese NJ, Mukherjee S, Ivanova N, Konstantinidis KT, Mavrommatis K, Kyrpides NC, Pati A. 2015. Microbial species delineation using whole genome sequences. Nucleic Acids Res 43:6761–6771.

13. Rodriguez-R LM, Castro JC, Kyrpides NC, Cole JR, Tiedje JM, Konstantinidis KT. 2018. How Much Do rRNA Gene Surveys Underestimate Extant Bacterial Diversity? Appl Environ Microbiol 84:e00014-18.

14. Scortichini M, Marcelletti S, Ferrante P, Firrao G. 2013. A Genomic Redefinition of Pseudomonas avellanae species. PLOS ONE 8:e75794.

15. Rodriguez-R LM, Gunturu S, Harvey WT, Rosselló-Mora R, Tiedje JM, Cole JR, Konstantinidis KT. 2018. The Microbial Genomes Atlas (MiGA) webserver: taxonomic and gene diversity analysis of Archaea and Bacteria at the whole genome level. Nucleic Acids Res 46:W282–W288.

16. Carter AP, Clemons WM, Brodersen DE, Morgan-Warren RJ, Wimberly BT, Ramakrishnan V. 2000. Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics. Nature 407:340–348.

17. Wang Q, Fish JA, Gilman M, Sun Y, Brown CT, Tiedje JM, Cole JR. 2015. Xander: employing a novel method for efficient gene-targeted metagenomic assembly. Microbiome 3:32.

18. Guo J, Cole JR, Zhang Q, Brown CT, Tiedje JM. 2015. Microbial community analysis with ribosomal gene fragments from shotgun metagenomes. Appl Environ Microbiol AEM.02772-15.

19. Fish JA, Chai B, Wang Q, Sun Y, Brown CT, Tiedje JM, Cole JR. 2013. FunGene: the functional gene pipeline and repository. Front Microbiol 4.

20. Eddy SR. 2009. A new generation of homology search tools based on probabilistic inference. Genome Inf 23:205–211.

21. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: a versatile open source tool for metagenomics. PeerJ 4:e2584.

22. Magoc T, Salzberg SL. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27:2957–2963.

23. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17:10–12.

24. Kuczynski J, Stombaugh J, Walters WA, Gonzalez A, Caporaso JG, Knight R. 2012. Using QIIME to analyze 16S rRNA gene sequences from microbial communities. Curr Protoc Microbiol Chapter 1:Unit 1E.5.

25. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, Tiedje JM. 2014. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res 42:D633–642.

26. Rice P, Longden I, Bleasby A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet 16:276–277.

27. Crusoe MR, Alameldin HF, Awad S, Boucher E, Caldwell A, Cartwright R, Charbonneau A, Constantinides B, Edvenson G, Fay S, Fenton J, Fenzl T, Fish J, Garcia-Gutierrez L, Garland

P, Gluck J, González I, Guermond S, Guo J, Gupta A, Herr JR, Howe A, Hyer A, Härpfer A, Irber L, Kidd R, Lin D, Lippi J, Mansour T, McA'Nulty P, McDonald E, Mizzi J, Murray KD, Nahum JR, Nanlohy K, Nederbragt AJ, Ortiz-Zuazaga H, Ory J, Pell J, Pepe-Ranney C, Russ ZN, Schwarz E, Scott C, Seaman J, Sievert S, Simpson J, Skennerton CT, Spencer J, Srinivasan R, Standage D, Stapleton JA, Steinman SR, Stein J, Taylor B, Trimble W, Wiencko HL, Wright M, Wyss B, Zhang Q, zyme en, Brown CT. 2015. The khmer software package: enabling efficient nucleotide sequence analysis. F1000Research.

28. Brown CT, Howe A, Zhang Q, Pyrkosz AB, Brom TH. 2012. A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data. ArXiv12034802 Q-Bio.

29. Sharon I, Kertesz M, Hug LA, Pushkarev D, Blauwkamp TA, Castelle CJ, Amirebrahimi M, Thomas BC, Burstein D, Tringe SG, Williams KH, Banfield JF. 2015. Accurate, multi-kb reads resolve complex populations and detect rare microorganisms. Genome Res 25:534–543.

30. Hug LA, Castelle CJ, Wrighton KC, Thomas BC, Sharon I, Frischkorn KR, Williams KH, Tringe SG, Banfield JF. 2013. Community genomic analyses constrain the distribution of metabolic traits across the Chloroflexi phylum and indicate roles in sediment carbon cycling. Microbiome 1:22.

31. Sun DL, Jiang X, Wu QL, Zhou NY. 2013. Intragenomic heterogeneity of 16S rRNA genes causes overestimation of prokaryotic diversity. Appl Env Microbiol 79:5962–5969.

32. Isenbarger TA, Carr CE, Johnson SS, Finney M, Church GM, Gilbert W, Zuber MT, Ruvkun G. 2008. The most conserved genome segments for life detection on Earth and other planets. Orig Life Evol Biosph 38:517–533.

Figures

Figure 1: Pairwise comparisons among all genomes of Rhizobiales (panels A, C, E) and of all *Rhizobium* (panels B, D, F) using the SSU rRNA gene, *rplB* and *rpsC*. SSU rRNA gene identities are higher than *rplB* and *rpsC*, and *rplB* and *rpsC* have similar sequence identities in most genome pairs. The dashed line is y = x. Data below y=x line indicate the gene on X axis is more conserved. The dot size indicates the number of pairwise comparisons with those values.

Figure 2: Pairwise comparison among all genomes of Pseudomonadales (panels A, C, E) and of all *Pseudomonas* (panels B, D, F) using the SSU rRNA gene, *rplB* and *rpsC*. SSU rRNA gene identities are higher than *rplB* and *rpsC* in most genome pairs, and *rplB* and *rpsC* have similar sequence identities. The dashed line is y = x. Data below y=x line indicate gene on X axis is more conserved. The dot size indicates the number of pairwise comparisons with those values.

Figure 3: Pairwise comparison among all completed genomes of species in the same genus using the SSU rRNA gene, *rplB,* and *rpsC*. SSU rRNA gene identities are larger than *rplB* in most genomes. The diagonal dashed line is y = x and data below the line indicates the gene on X axis is more conserved. The dot intensity is the number of comparisons with those values. Subplot A, B, and C are comparisons of *rplB* and SSU rRNA gene (93,733 pairwise comparisons), *rpsC* and SSU rRNA gene (85,581 pairwise comparisons), *rplB* and *rpsC* (85,956 pairwise comparisons).

Figure 4: Comparison of the SSU rRNA gene, *rpsC* and *rplB* in beta diversity analyses (ordination) using large soil metagenome sequences from seven field replicates. All genes show that the microbial community of the corn (C) rhizosphere is significantly different from those of *Miscanthus* (M) and switchgrass (S) while *rplB* and *rpsC* at both the nucleotide (n) and protein (p) levels separate microbial communities of *Miscanthus* and switchgrass. The SSU rRNA gene does not separate *Miscanthus* and switchgrass with either shotgun (SSU.sg) or amplicon (SSU.am) data. “**” indicates p < 0.01 in PERMANOVA test.

Figure 5: Comparison of the SSU rRNA gene, *rplB*, and *rpsC* in alpha diversity analyses (Chao1, Shannon, and OTU number by protein, "_p" and nucleotide, "_n") using large soil metagenome sequence. All genes show that the microbial community of the corn (C) rhizosphere has significantly less alpha diversity than those of *Miscanthus* (M) and switchgrass (S) except for *rpsC* and SSU rRNA gene with Chao1 index ($p < 0.01$). Wilcoxon test was used to compare SSU_sg against each of the other genes including rplB_n, rplB_p, rpsC_n, rpsC_p and SSU_am. For Chao1 index, rplB_n and rpsC_n show significantly higher abundance than SSU_sg in *Miscanthus* and switchgrass; For Shannon index and OTU number, rplB_n and rpsC_n show significantly higher abundance than SSU_sg in all three crops ("***" is $p < 0.001$, '**' is $p < 0.01$, '*' is $p < 0.05$, "ns" is $p > 0.05$).

Figures

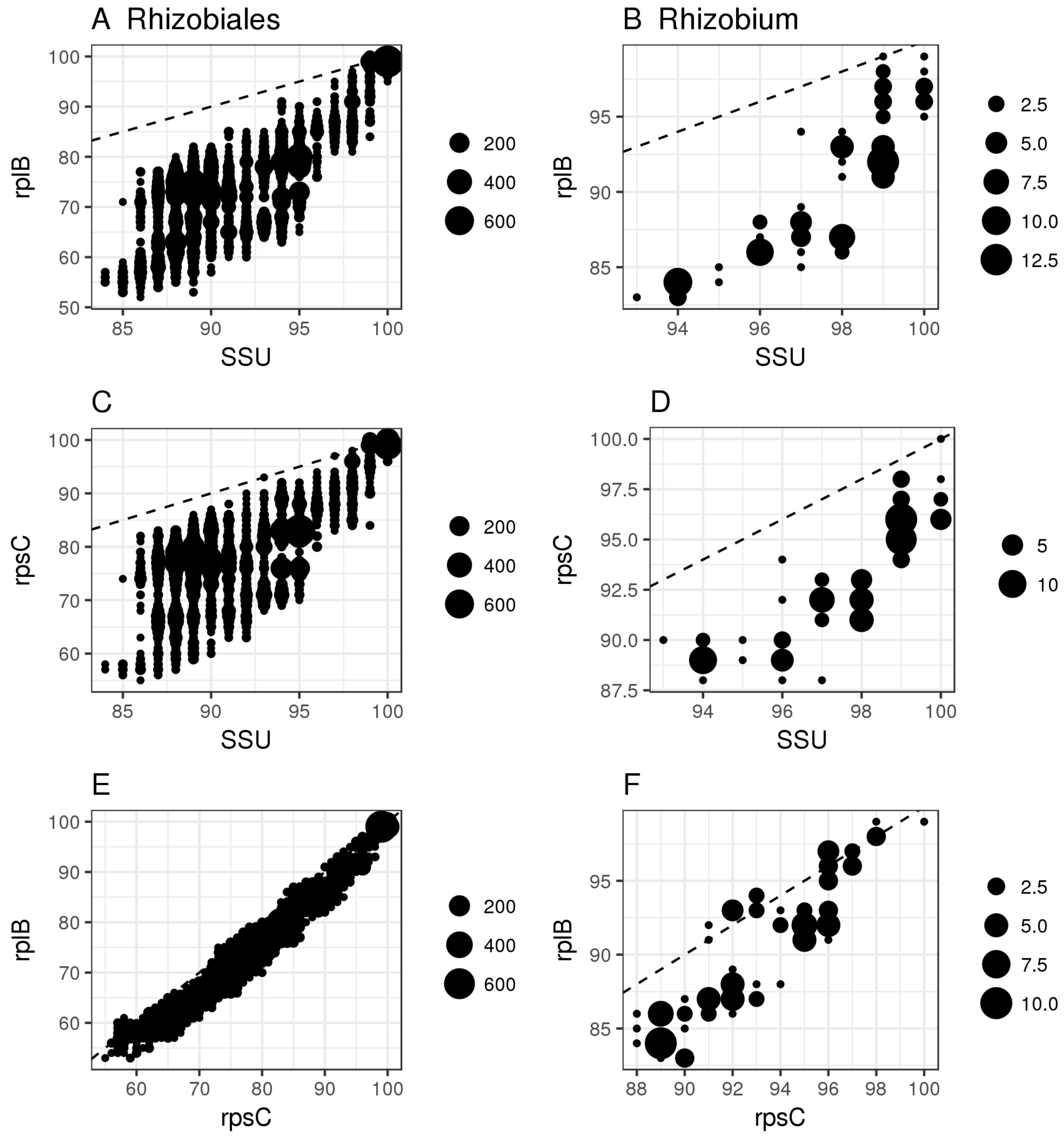
A Rhizobiales
rplB
100
90
80
70
60
50
85
90
95
100
SSU
200
400
600
B Rhizobium
rplB
95
90
85
94
96
98
100
SSU
2.5
5.0
7.5
10.0
12.5
C
rpsC
100
90
80
70
60
85
90
95
100
SSU
200
400
600
D
rpsC
100.0
97.5
95.0
92.5
90.0
87.5
94
96
98
100
SSU
5
10
E
rplB
100
90
80
70
60
60
70
80
90
100
rpsC
200
400
600
F
rplB
95
90
85
88
90
92
94
96
98
100
rpsC
2.5
5.0
7.5
10.0

Figure 1: Pairwise comparisons among all genomes of Rhizobiales (panels A, C, E) and of all *Rhizobium* (panels B, D, F) using the SSU rRNA gene, *rplB* and *rpsC*. SSU rRNA gene identities are higher than *rplB* and *rpsC*, and *rplB* and *rpsC* have similar sequence identities in most genome pairs. The dashed line is y = x. Data below y=x line indicate the gene on X axis is more conserved. The dot size indicates the number of pairwise comparisons with those values.

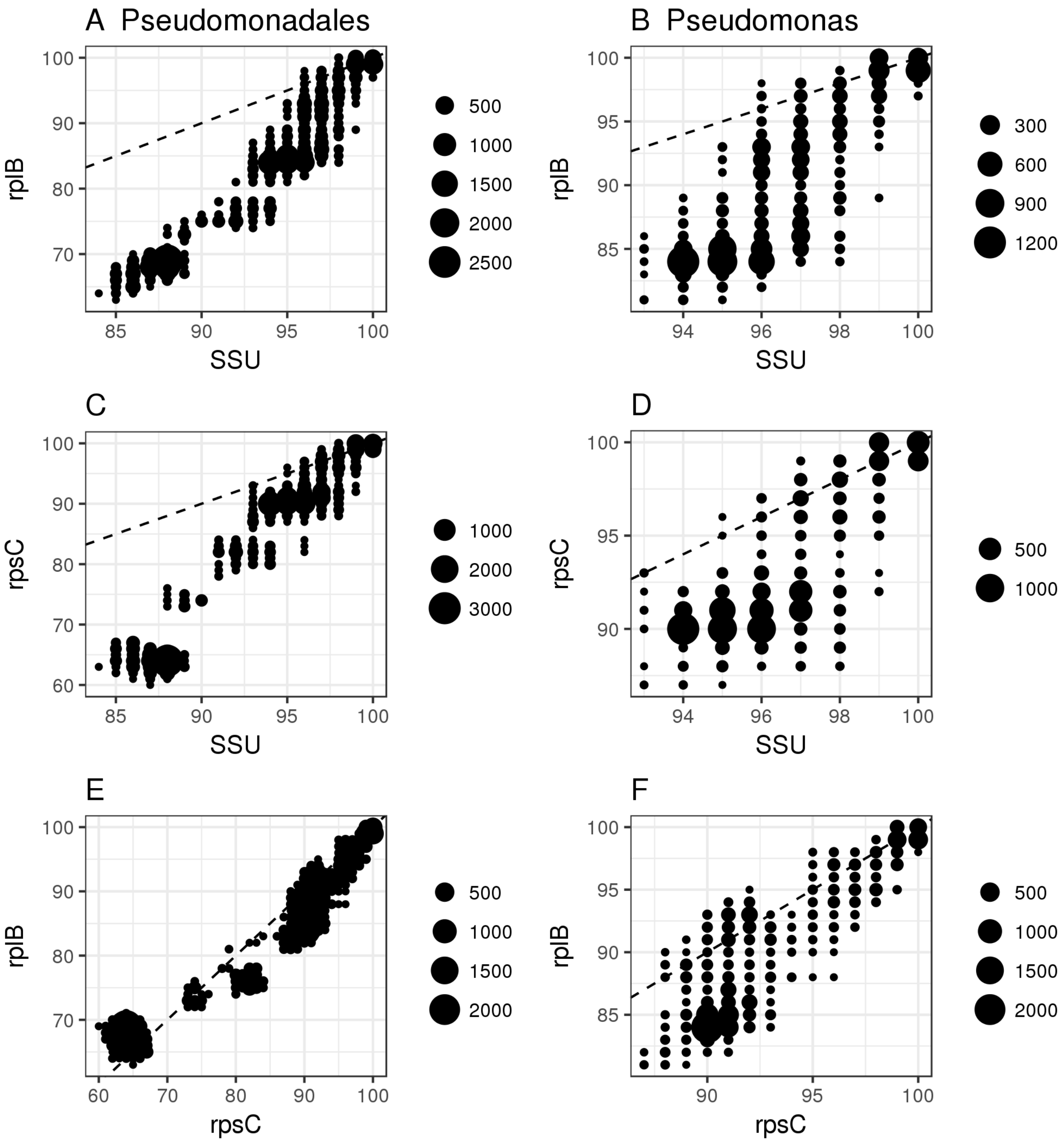
A Pseudomonadales
rplB
100
90
80
70
85
90
95
100
SSU
500
1000
1500
2000
2500
B Pseudomonas
rplB
100
95
90
85
94
96
98
100
SSU
300
600
900
1200
C
rpsC
100
90
80
70
60
85
90
95
100
SSU
1000
2000
3000
D
rpsC
100
95
90
94
96
98
100
SSU
500
1000
E
rplB
100
90
80
70
60
70
80
90
100
rpsC
500
1000
1500
2000
F
rplB
100
95
90
85
90
95
100
rpsC
500
1000
1500
2000

Figure 2: Pairwise comparison among all genomes of Pseudomonadales (panels A, C, E) and of all *Pseudomonas* (panels B, D, F) using the SSU rRNA gene, *rplB* and *rpsC*. SSU rRNA gene identities are higher than *rplB* and *rpsC* in most genome pairs, and *rplB* and *rpsC* have similar sequence identities. The dashed line is y = x. Data below y=x line indicate gene on X axis is more conserved. The dot size indicates the number of pairwise comparisons with those values.

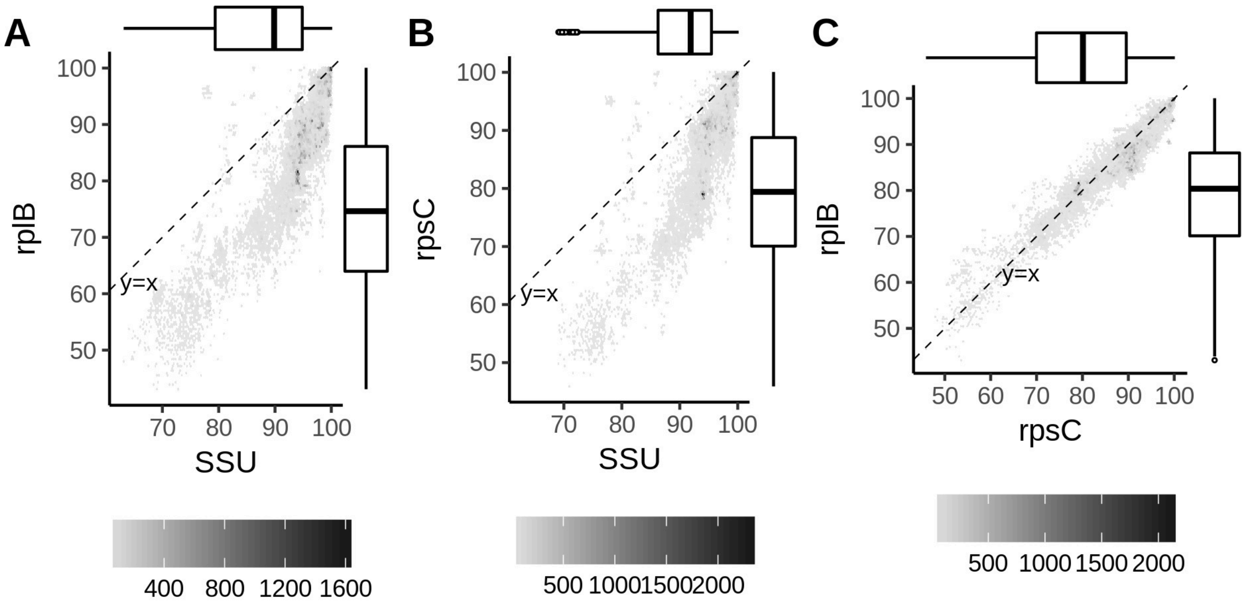

Figure 3: Pairwise comparison among all completed genomes of species in the same genus using the SSU rRNA gene, *rplB,* and *rpsC*. SSU rRNA gene identities are larger than *rplB* in most genomes. The diagonal dashed line is y = x and data below the line indicates the gene on X axis is more conserved. The dot intensity is the number of comparisons with those values. Subplot A, B, and C are comparisons of *rplB* and SSU rRNA gene (93,733 pairwise comparisons), *rpsC* and SSU rRNA gene (85,581 pairwise comparisons), *rplB* and *rpsC* (85,956 pairwise comparisons).

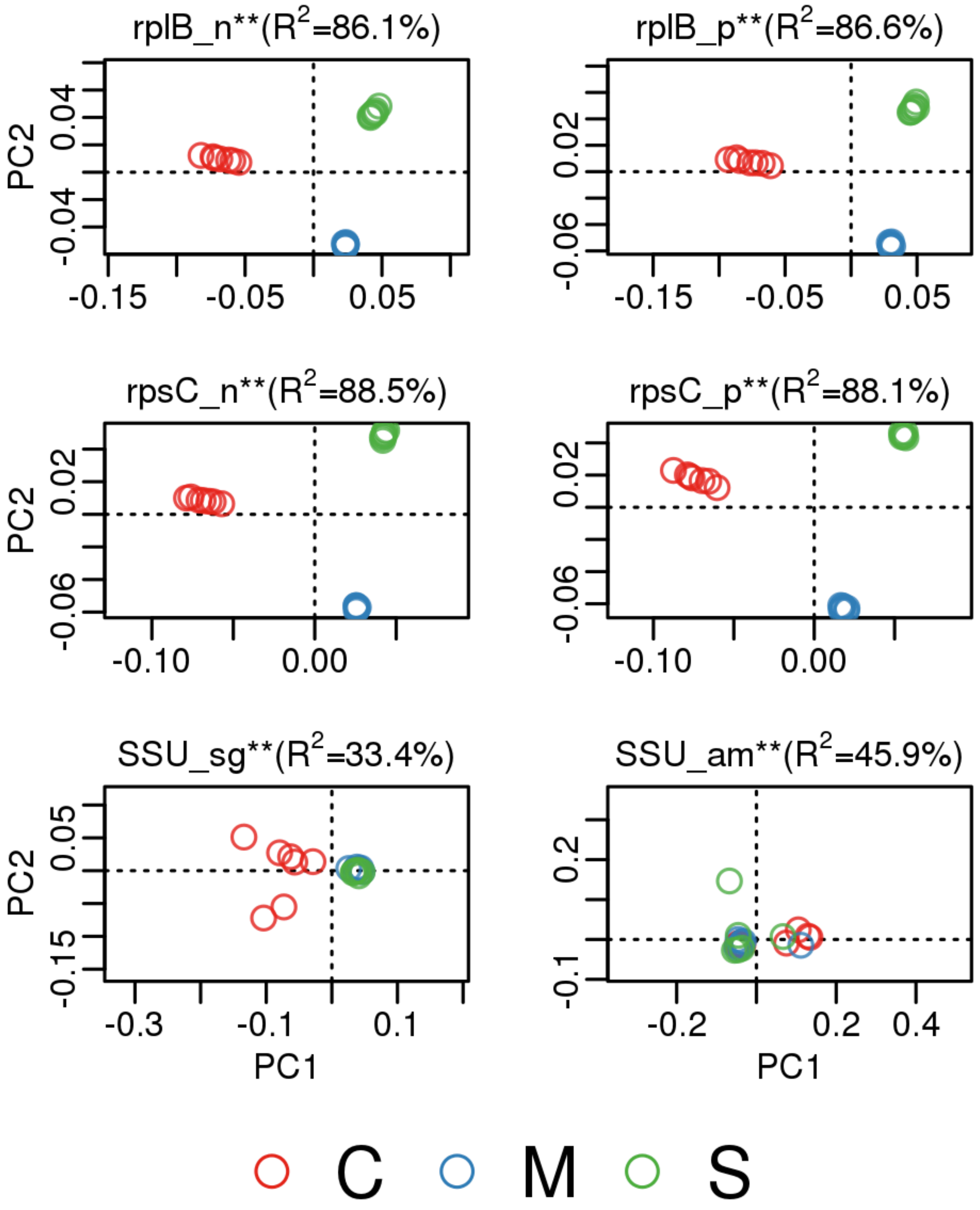
rplB_n**(R^2=86.1%)
PC2
0.04
0.0
-0.04
-0.15
-0.05
0.05
rplB_p**(R^2=86.6%)
0.02
0.0
-0.06
-0.15
-0.05
0.05
rpsC_n**(R^2=88.5%)
PC2
0.02
0.0
-0.06
-0.10
0.00
rpsC_p**(R^2=88.1%)
0.02
0.0
-0.06
-0.10
0.00
SSU_sg**(R^2=33.4%)
PC2
0.05
-0.15
-0.3
-0.1
0.1
PC1
SSU_am**(R^2=45.9%)
0.2
-0.1
-0.2
0.2
0.4
PC1
○ C
○ M
○ S

Figure 4: Comparison of the SSU rRNA gene, *rpsC* and *rplB* in beta diversity analyses (ordination) using large soil metagenome sequences from seven field replicates. All genes show that the microbial community of the corn (C) rhizosphere is significantly different from those of *Miscanthus* (M) and switchgrass (S) while *rplB* and *rpsC* at both the nucleotide (n) and protein (p) levels separate microbial communities of *Miscanthus* and switchgrass. The SSU rRNA gene does not separate *Miscanthus* and switchgrass with either shotgun (SSU.sg) or amplicon (SSU.am) data. "**" indicates $p < 0.01$ in PERMANOVA test.

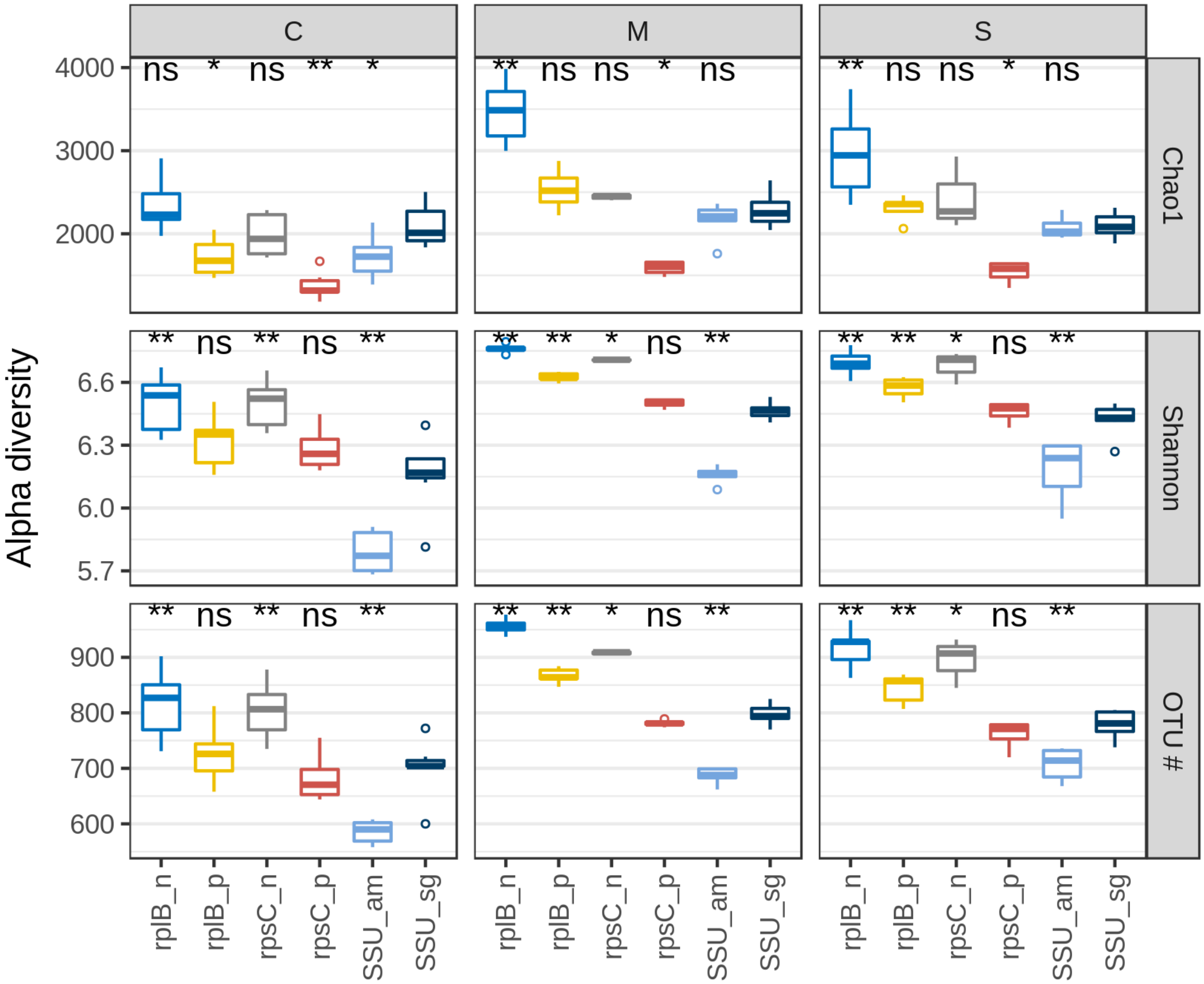
C
M
S
Chao1
Shannon
OTU #
Alpha diversity
4000
3000
2000
6.6
6.3
6.0
5.7
900
800
700
600
ns * ns ** *
** ns ns * ns
** ns ns * ns
** ns ** ns **
** ** * ns **
** ** * ns **
** ns ** ns **
** ** * ns **
** ** * ns **
rplB_n
rplB_p
rpsC_n
rpsC_p
SSU_am
SSU_sg

Figure 5: Comparison of the SSU rRNA gene, *rplB*, and *rpsC* in alpha diversity analyses (Chao1, Shannon, and OTU number by protein, "_p" and nucleotide, "_n") using large soil metagenome sequence. All genes show that the microbial community of the corn (C) rhizosphere has significantly less alpha diversity than those of *Miscanthus* (M) and switchgrass (S) except for *rpsC* and SSU rRNA gene with Chao1 index ($p < 0.01$). Wilcoxon test was used to compare SSU_sg against each of the other genes including rplB_n, rplB_p, rpsC_n, rpsC_p and SSU_am. For Chao1 index, rplB_n and rpsC_n show significantly higher abundance than SSU_sg in *Miscanthus* and switchgrass; For Shannon index and OTU number, rplB_n and rpsC_n show significantly higher abundance than SSU_sg in all three crops ("***" is $p < 0.001$, '**' is $p < 0.01$, '*' is $p < 0.05$, "ns" is $p > 0.05$).