## Supplemental Materials

**TABLE S1** Higher fungi/bacteria ratio and percent of AMF fungi are in rhizosphere sample (M1) than bulk sample (SB1).

| | % Bacteria | % Fungi | % AMF in Fungi | Fungi/Bacteria |
|---|---|---|---|---|
| SB1-SSU | 97.00% | 0.36% | 0.00% | 0.0037 |
| SB1-LSU | 96.75% | 0.42% | 0.00% | 0.0044 |
| M1-SSU | 92.38% | 2.60% | 0.18% | 0.0281 |
| M1-LSU | 92.94% | 2.48% | 0.18% | 0.0267 |

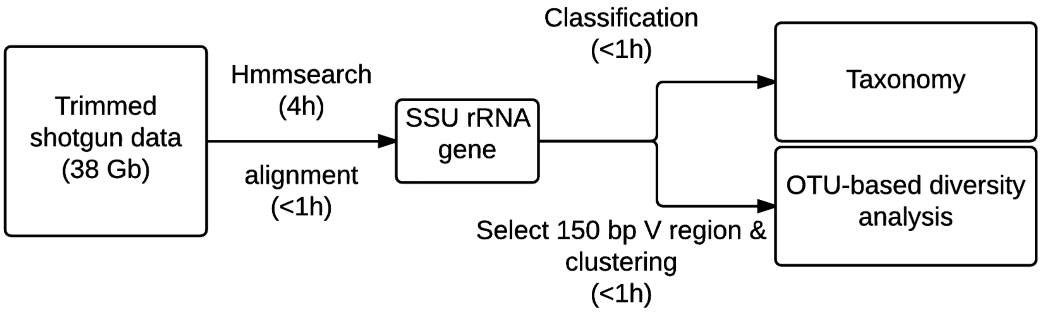
Trimmed shotgun data (38 Gb)
Hmmsearch (4h)
alignment (<1h)
SSU rRNA gene
Classification (<1h)
Taxonomy
Select 150 bp V region & clustering (<1h)
OTU-based diversity analysis

**FIG S1** Flowchart of SSUsearch pipeline. SSU rRNA gene fragments were retrieved by an hmmsearch and alignment step, which could be further used for reference-based (supervised) diversity analysis (taxonomy). Those fragments aligned to 150 bp of a variable region could be used for OTU-based (unsupervised) diversity analysis. SSU rRNA gene identification hmmsearch is the most time consuming steps. For 1 lane of trimmed HiSeq data (38 Gb) from Miscanthus rhizosphere sample (M1), SSU rRNA gene identification took about 4 hours with peak memory usage of about 4.5 Gb. In the analysis pipeline, where there was not a clear performance difference between tools, we mostly used Mothur including databases. For (de novo) OTU based analysis, SSU rRNA gene reference sequences with taxonomy information are required for classification. Another smaller set of aligned references is required to align the gene fragments from shotgun data. The SILVA database (the official one, not the one included with QIIME or Mothur) was used to build the HMM since the reference set from SILVA was more up to date. Two scripts in QIIME were used to cluster (UCLUST, the default) and pick representative sequences. Building the HMM is not part of the pipeline but using the built HMM is. We found that complete-linkage clustering is faster and requires less memory with McClust than with Mothur (dist.seqs and cluster). Additionally, we use two scripts in QIIME (pick_otus.py and pick_rep_set.py) to select representative sequences for building the HMM due to ease of use and the GreenGenes database is included for use with the Copyrighter copy number correction tool.

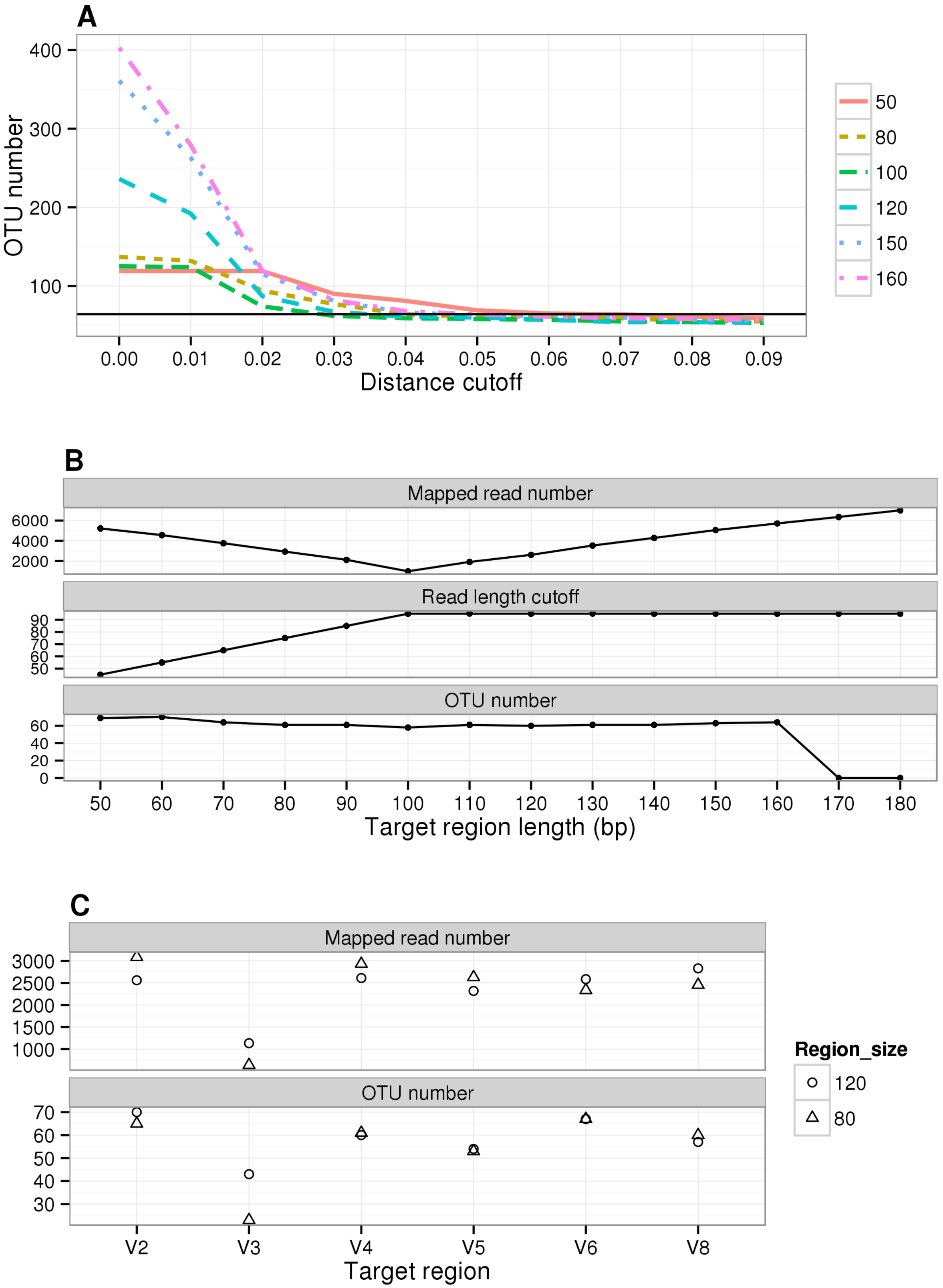
A
OTU number
400
300
200
100
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
Distance cutoff
50
80
100
120
150
160
B
Mapped read number
6000
4000
2000
Read length cutoff
90
80
70
60
50
OTU number
60
40
20
0
50
60
70
80
90
100
110
120
130
140
150
160
170
180
Target region length (bp)
C
Mapped read number
3000
2500
2000
1500
1000
OTU number
70
60
50
40
30
V2
V3
V4
V5
V6
V8
Target region
Region_size
○ 120
△ 80

**FIG S2** Testing the effect of target region size and variable region on clustering on a synthetic community with 64 species with read length at about 100 bp. Subfigure A shows a distance cutoff of 4% or 5% is proper for all regions sizes from 50 bp to 160 bp in V4 (OTU number approached the species number 64 as indicated by the black line). Subfigure B shows more details in the method used for subfigure A. Panel “Read Length cutoff” in B shows minimum read length was set to the target region size minus 5 bp if the region size was less than 100 bp, and 95 bp when the region size was longer than 100 bp. As a result, the number of reads aligned decreased as the target region size increased until 100 bp, and then the number of reads aligned increased with target region as shown in Panel “Mapped read number” in B. Panel “OTU number” in C shows OTU number at distance cutoff of 0.05 and our method works well from a 50 bp region to 160 bp region in V4. Subfigure C tests our unsupervised method on multiple hyper-variable regions (V2, V3, V4, V5, V6, V8) with region size of 120 bp (circle) and 80 bp (triangle). Panel “Mapped read number” in C shows the number of reads mapped to each chosen region. Panel “OTU number” in C shows the number of OTUs in each region at distance cutoff of 0.05. All regions have consistent mapped read number and OTUs except V3.

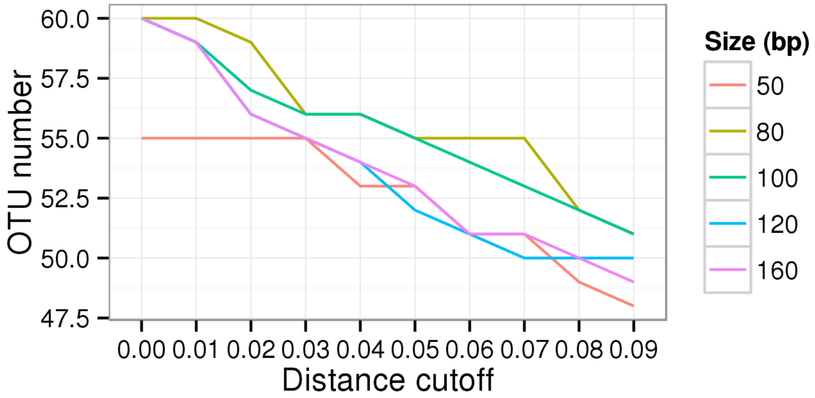

**FIG S3** Testing the effect of target region size on V4 of full-length SSU rRNA genes from a synthetic community with 64 species.

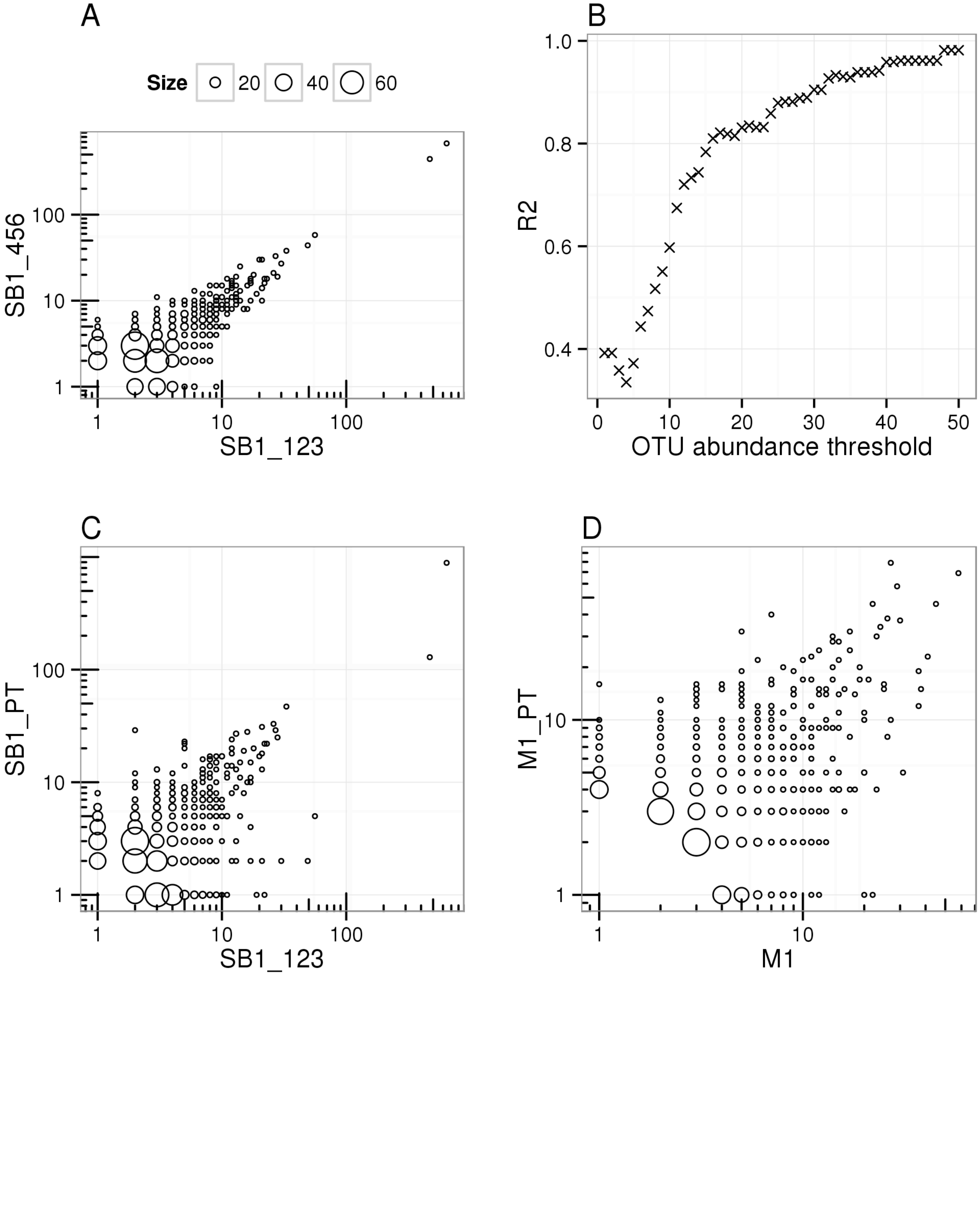
A
Size
20
40
60
SB1_456
1
10
100
SB1_123
1
10
100
B
R2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
OTU abundance threshold
C
SB1_PT
1
10
100
SB1_123
1
10
100
D
M1_PT
1
10
M1
1
10

**FIG S4** Technical reproducibility test of our unsupervised clustering and comparison of OTU abundances between paired shotgun and amplicon data. Subfigure A shows consistent OTU abundance profiles in two technical replicates (Pearson's correlation coefficient is 0.997). X axis shows number of reads in each OTU in replicate SB1_123, and y axis shows number of reads in each OTU in replicate SB1_456. The size of circle is proportional to number of OTUs at the same location in the plot (with the same counts in SB1_123 and also in SB1_456). The consistency of counts of each OTU in two replicates becomes better when the abundance of OTUs are higher. Subfigure B shows progressive dropout analysis of two technical replicates of shotgun data. There is significant correlation of counts of each OTU between technical replicates; X axis is the threshold of OTU abundance and y axis is the $R^2$ of linear regression of log transformed OTU abundances in two replicates. OTUs with lower abundance than the thresholds (x axis) were discarded before regression analysis. Subfigure C and D shows comparison of OTU abundance profile between paired shotgun and amplicon data in bulk soil sample (SB1) and rhizosphere sample (M1), respectively. There is inconsistency between shotgun data and amplicon data in both samples. X axis shows number of SSU rRNA gene fragments in shotgun data per OTU in log scale, and y axis shows number of amplicon sequences in each OTU in log scale. The OTU abundance in both amplicon and shotgun data were increased by 1 to avoid 0 counts that can be displaced in log scale. The size of circle is proportional to number of OTUs with the same abundance in both types of data. There are OTUs with significantly different abundances in the two types of data (circles deviate from diagonal line). Pearson's correlation between two types of data is 0.873 in SB1 and 0.581 in M1.

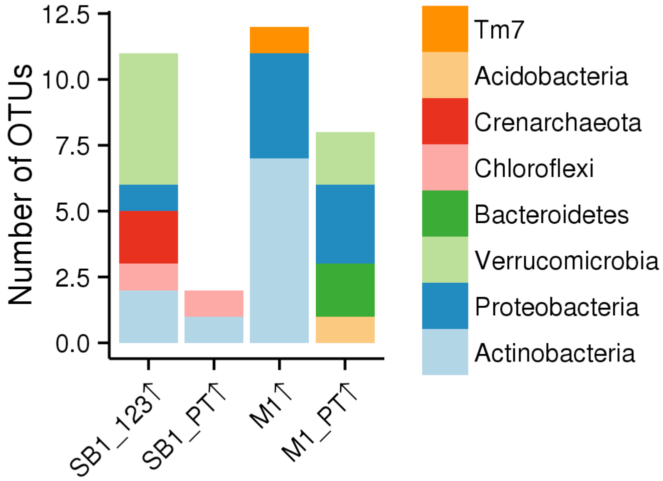

**FIG S5** Phyla of OTUs significantly different between shotgun data and amplicon data. SB1_123 are shotgun data and SB1_PT are amplicon data both from the same DNA from bulk soil sample. M1 are shotgun data and M1_PT are amplicon data both from the same DNA from Miscanthus rhizosphere sample. OTUs significantly different were defined as those with total abundance > 10 and fold change between two types of data > 5 or < 0.2. Verrucomicrobia was biased against in bulk soil sample amplified by V6-V8 primer (SB1_PT) but biased for in rhizosphere sample amplified with V4 primer (M1_PT). Actinobacteria was biased against in rhizosphere sample (M1).

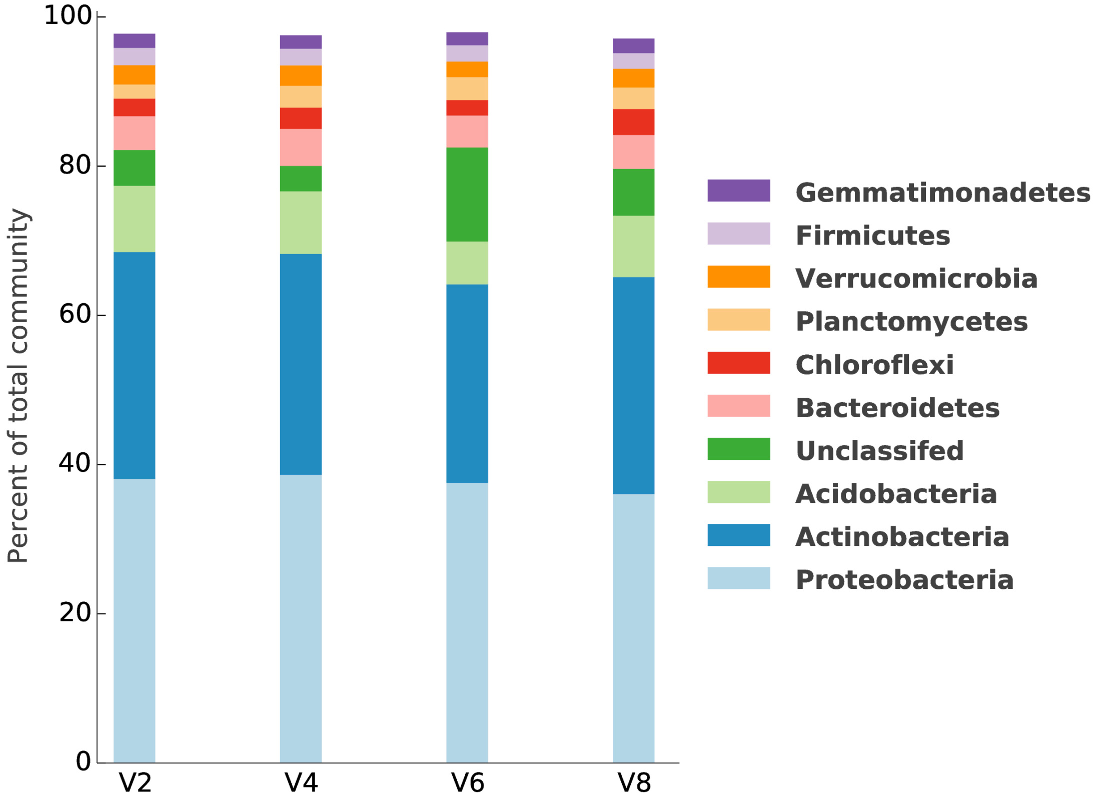

**FIG S6** Bacterial phylum profile comparison using different variable regions. Different variable regions have similar taxonomy profiles, except that V6 has more unclassified sequences. The minimum Pearson's correlation between the regions is 0.96. Classifications were done using SSU rRNA gene fragments from Miscanthus rhizosphere soil sample (M1) and SILVA database as reference.

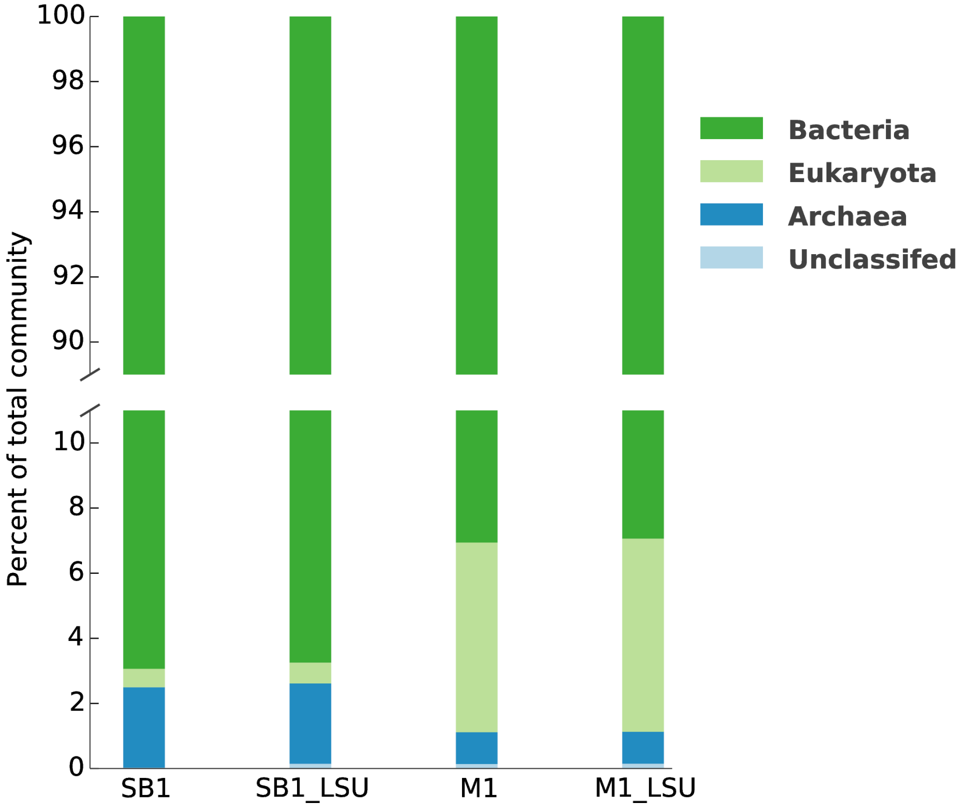

**FIG S7** Taxonomy profile comparison at domain level using SSU and LSU rRNA genes. For both the bulk soil sample (SB1) and rhizosphere sample (M1), SSU and LSU show consistent domain level taxonomy distribution (Pearson's correlation coefficient = 1). "_LSU" indicates taxonomy from LSU rRNA SILVA database and the rest are classified by SSU rRNA database.

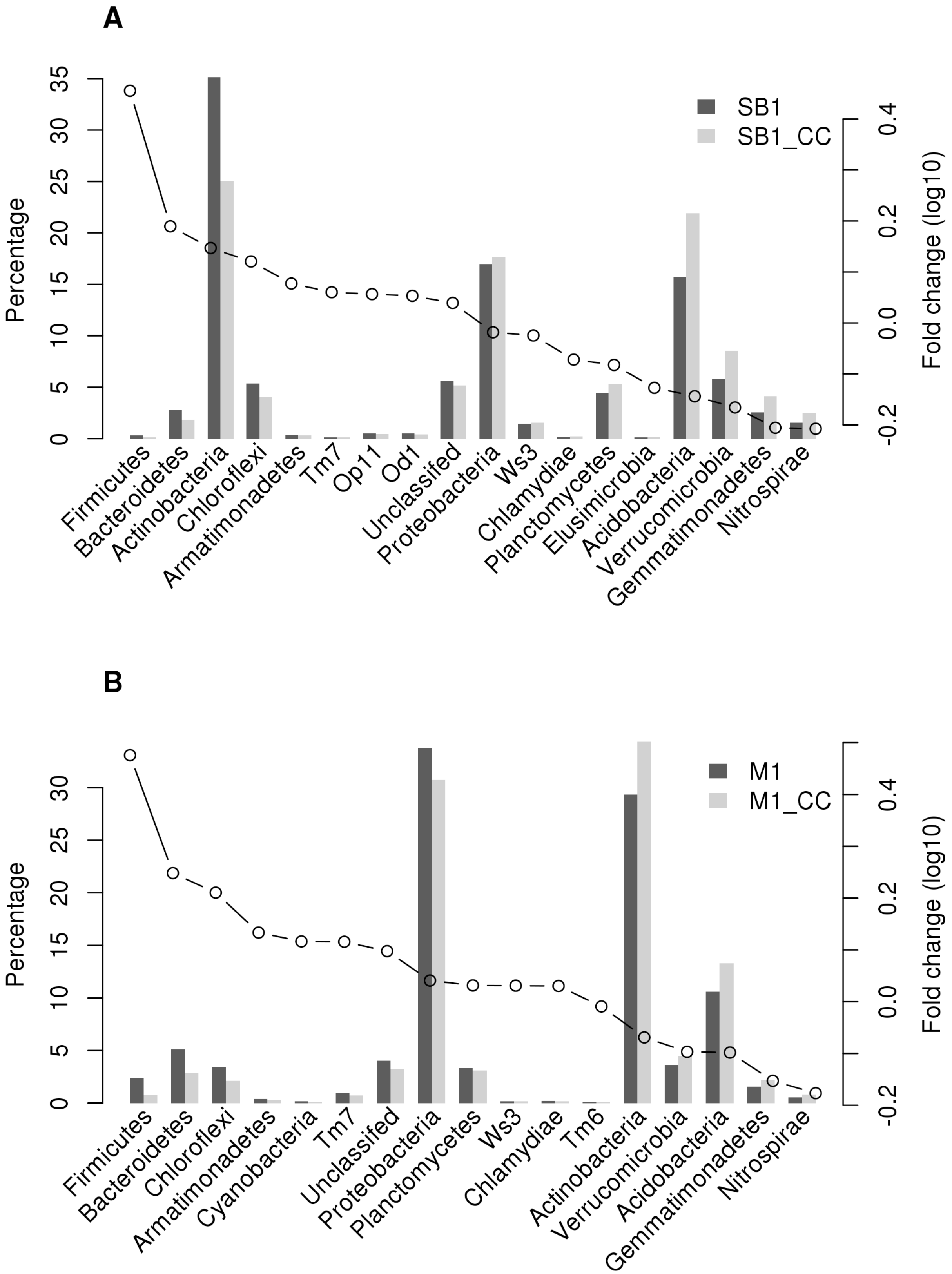
A
Percentage
0
5
10
15
20
25
30
35
SB1
SB1_CC
Fold change (log10)
-0.2
0.0
0.2
0.4
Firmicutes
Bacteroidetes
Actinobacteria
Chloroflexi
Armatimonadetes
Tm7
Op11
Od1
Unclassifed
Proteobacteria
Ws3
Chlamydiae
Planctomycetes
Elusimicrobia
Acidobacteria
Verrucomicrobia
Gemmatimonadetes
Nitrospirae
B
Percentage
0
5
10
15
20
25
30
M1
M1_CC
Fold change (log10)
-0.2
0.0
0.2
0.4
Firmicutes
Bacteroidetes
Chloroflexi
Armatimonadetes
Cyanobacteria
Tm7
Unclassifed
Proteobacteria
Planctomycetes
Ws3
Chlamydiae
Tm6
Actinobacteria
Verrucomicrobia
Acidobacteria
Gemmatimonadetes
Nitrospirae

**FIG S8** Bacterial phylum level taxonomy summary before and after SSU rRNA gene copy correction. Left vertical axis with bar plot shows percentage in total community, while right vertical axis with line plot shows fold change after copy number correction. Taxa with relative abundances of more than 0.1% before copy correction were chosen and were ordered based on fold change. Subfigure A is for bulk soil sample (SB1) and B is for Miscanthus rhizosphere sample (M1).

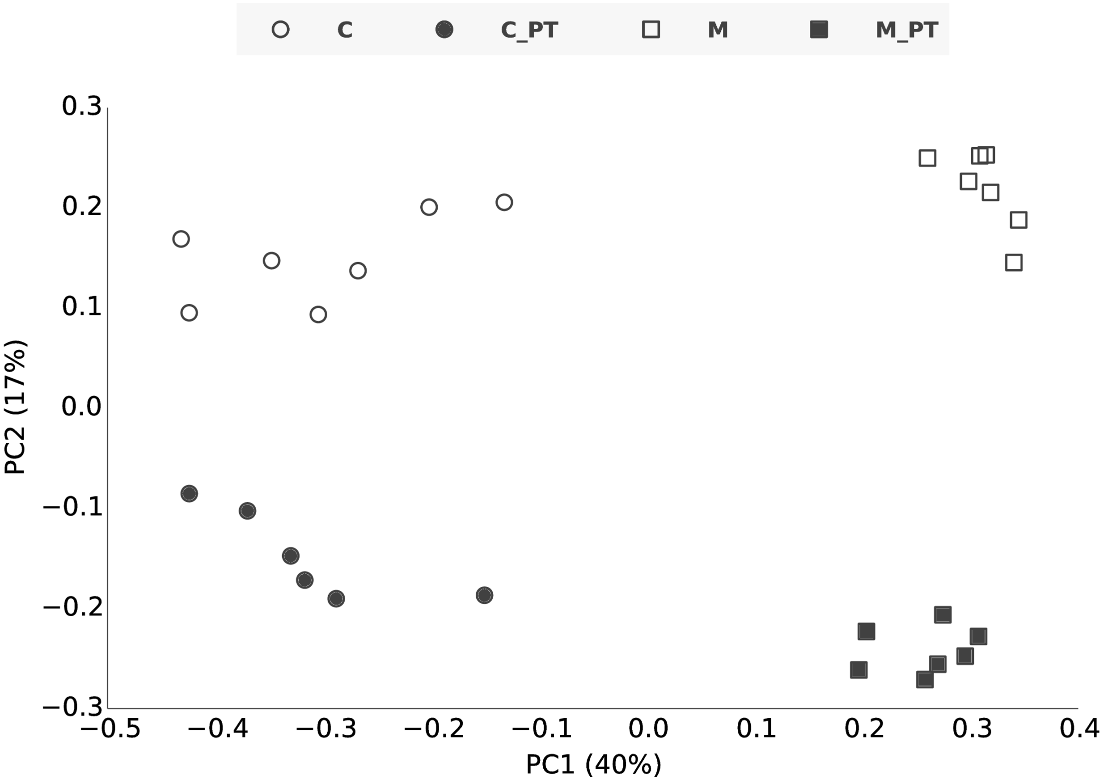

**FIG S9** Ordination plot of amplicon and shotgun derived data after copy correction. The plot is similar to the one without copy correction (Fig. 3). There were significant differences in amplicon and shotgun derived data (y-axis) and of corn and Miscanthus rhizosphere samples (x-axis), (AMOVA p-value < 0.001), after copy number correction. PCoA was applied to OTU table resulting from *de novo* clustering with shotgun data and amplicon data using 150bp of V4 region. The filled markers ("_PT") are amplicon data and the unfilled markers are shotgun data.

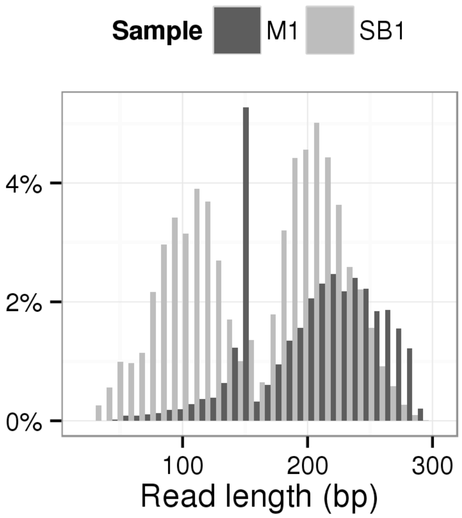

**FIG S10** Length distribution of trimmed reads after quality trimming and paired-end merging. SB1 is the bulk soil data and M1 is the rhizosphere data. The reads with >150 bp result from the merged paired ends, which benefits classification and clustering in downstream analyses. Reads less than 150 bp are also used in the analysis and come from unmerged paired reads.