Characterisation of bacterial and archeal diversity in soils from Gordon Square with 16S amplicon suggest *Rhizobiales* and *Saprospirae* abundance as markers for the level of decay.

Abstract
Soil micro-organisms for long have been indicated to interact with plants to provide protection or causing diseases. Advances in metagenomics enabled direct assessment of biological diversity to facilitate the study of micro-plant interaction. Here we report Rhizobiales and Saprospirae abundance as potential markers for the metagenomic status of the sample, and make the hypothesis that their abundances were related to the level of decay of the corresponding soil sample.

Introduction
Metagenomics is a recent methodology based on high throughput next generation sequencing, which enabled direct characterisation of millions of sequences simultaneously[1]. 16S RNA sequencing represents the consensus method to characterise the bacterial/archeal diversity and has been tested on mock community to for its sensitivity and consistency [2–4], whereas 18S ITS sequencing is the equivalent for fungi[5]. Metagenomics offer several advantages including access to genomes that are not cultivable in laboratory and direct assessment of relative abuandances[6].

Soil microbiome has been shown to interact and affect plant health, as well as being actively shaped[7]. The existence of specific microbes such as, could inhibit a variety of pathogens directly and indirectly. It is also reported that plant, actively shape its microbiome via enrichment of various plant exduates. As a result, vicinity of plant roots (rhizosphere) usually associates with a more dense but less diverse microbiome.

We measured genetic diversity of soil samples collected at Gordon Square. To fully characterise the diversity, we took multiple samples from different locations on site, characterised their metagenomic profiles, and applied statistical methods to infer the source of observed variability.

Material and Method
Soil Collection and Numbering:
Eight soil samples were collected from Gordon Square. Approximately 10mL (~20g) of Soil was collected with an auger at 10cm±2cm. Each sample is assigned with a sample ID. (e.g. “2017_1” means the sample is collected in 2017). GPS coordinates and a photo of the site of collection were recorded (see appendix, data incomplete).

Soil characterisation:
Each soil sample was measured for pH and water content (data incomplete). 2.0g of soil was dissolved in 10mL of distilled H2O and it pH was measured with a pH strip. 2.0g of soil is microwaved for 5 minutes and reweighed, and its water content estimated as follows:

$$water\_content = \frac{dried\_mass - original\_mass}{original\_mass} \cdot 100\%$$

(Equation 1)

Genomic DNA(gDNA) extraction:
Each gDNA extraction was prepared from a 0.25g subsample of soil sample using MoBio® PowerSoil DNA isolation Kit (Catalog No.12888). Three gDNA extractions were prepared for each soil sample. The manufacture protocol with centrifugation was followed exactly. The DNA extractions were subject to a 0.5% agarose gel electrophoresis for quality control.

PCR amplification and primer design:
F515/R806 primers[3] with Golay barcodes were used to amplify the ~250bp V4 region of 16S ribosomal DNA, that is 533-786 in *E. coli* strain 83972 sequence (greengenes acc. prokMSA_id:470367). The primers were 16-fold_ degenerate to account for sequence diversity. Each soil sample was associated a unique barcode. The total length of amplicon is 384bp. Detailed primer design is outlined in appendix. Primers were ordered as SIGMA NGSO-1.

Three 25uL PCR reactions were prepared for each DNA extraction with Bioline® BioMix (catalog No.) , with Mg2+ ions adjusted to 3.0mM using 50mM $MgCl_2$ stock solution. The 5’ and 3’ primers were both added to 50nM final concentration. The DNA extraction was added as PCR template with a 25-fold dilution (1uL in 25uL).

The prepared mixtures were loaded onto a 96-well plate and amplified using a standard PCR machine. The machine was programmed to run 35 cycles at

94°C for 15sec, 54°C for 45sec, 72°C for 30sec. Additionally there was a 1-minute initial denaturation at 94°C and a 5-minute final extension at 72°C. Three PCR products for the same gDNA extraction were pooled and the amplicons were subject to a 1.5% agarose gel electrophoresis for quality control.

The pooled PCR products were purified for amplicons with QIAGEN® QIAquick PCR cleanup kit (catalog No. 28014/28106). The manufacture protocol was followed with a 70uL starting volume and no pH indicator.

Double-stranded DNA quantification:
The purified amplicons were measured for dsDNA concentration using SpectraMax® Quant™ AccuClear™ Nano dsDNA Assay Kit. Manufacture protocol was followed and standard solution for dsDNA concentration at 25, 10, 3 and 1ng/uL were prepared.

Sequencing:
A 20uL sample at an amplicon concentration of 20nM is prepared from each amplicon solution. Samples from the same year are pooled together and sequenced on an Illumina® MiSeq platform using a 500-cycle MiSeq Reagent Kit v2, with 250 cycles for each read.(catalog No. MS-102-2003).

Sequence mapping
All data analyses were performed with QIIME[8]. Raw sequence reads were mapped to soil samples according to their barcodes and scanned for OTUs on the LEGION cluster at UCL, with the 16S rDNA database from greengenes(v13_8) as reference alignment and phylogenetic tree and uclust_ref as the method[9]. Only the 5' read is processed due to low quality of the 3' read. The resultant OTU table is merged with OTU table from earlier experiments in 2016 (bioc3301_) for subsequent analyses.

Alpha-diveristy
The total phylogenetic distance for the whole tree (PD_whole_tree) is used to measure alpha-diversity. The OTU table is filtered to include OTU's with a relative abudance above a threshold. PD_whole_tree was calculated for 10 abundance threshold from 0.001% to 0.01%.

Beta-diversity
Taxono mic summaries at from at phylum, class, order and family level were produced for the sample. To perform valid PCoA, OTU tables from 8 other studies were downloaded from EBI metagenomics according to their project ID (appendix_)

Both the sample dataset and the combined dataset were subject to PCoA analyses. For each dataset, five PCoA's were performed with different settings. The first four were with Bray-Curtis distance[10], at phylum, class, order, family level. The final PCoA is performed with unweighted unifrac distance[11]. All PCoA plots were visulised with Emperor[12].

PC1 coordinates from each run were incorporated into the mapping file. Each OTU/taxonomic-group is then tested for association with the PC1 coordinate at corresponding level using Spearman correlaion. P-value is generated both with bootstrapping and Fisher transform. Hits with a p-value<0.005 from both methods are recorded.

Distributions of *Bacteroidetes* and of *Alphaproteobacteria were* extracted from Class level and Order level summary for visualisation, respectively.

Results
Genomic DNA extraction was successful for all soil samples (see appendix_). Most extractions show strong a diffuse band at high molecular weight corresponding to large genomic DNA fragments.

All amplicon solutions show a distinct band at ~400bp (appendix), corresponding to the desired 384bp amplicon. All amplicons show extra additional diffuse bands at 1650bp and at 600bp. These unexpected amplicons were inferred as daisy-chain products due to over-amplification after depletion of primers. This caused problem later in sequencing.

The first (5') read was of good quality (Q30 unknown), whereas the second (3') read was of poorer quality due to over-clustering. Only 75% of bases in 3' read associate with a Q-score>=30 (99.9% probability to be correct). Due to its particular low quality, the 3'-read is excluded for subsequent data analysis.

The mean OTU number is 86300 (min-max:59800-109400, n=8). We pooled these OTU data with archived OTU data of 4 soil samples from 2016 [13]. The 2016 data is picked from paired-end reads and a more fluctuated OTU number (mean=887000,min-max:37200-1427000,n=4). Overall, the 2017 data seems to be more consistent.

4. Sampling depth and alpha diversity.

Earlier study of 16S metagenomics on Illumina 454 indicated that apparent alpha-diversity is dependent on the filtering threshold[3]. We confirm that the alpha-diversity is a function of filtering threshold θ (R^2=1.00, see figure 2). Moreover, α-diversity can be expressed in :

$\alpha = a \cdot ln\theta + b$ (Equation 2)

, where a is the slope and b is the intercept (Appendix). We used the normalised parameter b (Table 1) as the α-diversity in the following analysis.

SampleID	a(slope)	b(intercept)	rsquare
2016_1	-37.75	282.76	1.00
2016_3	-34.25	250.95	1.00
2016_2	-37.71	282.12	1.00
2016_4	-33.51	245.32	1.00
2017_2	-42.29	321.21	1.00
2017_1	-38.25	285.19	1.00
2017_3	-33.57	242.17	1.00
2017_4	-40.41	305.47	1.00
2017_5	-36.65	270.83	1.00
2017_6	-31.31	226.15	1.00
2017_7	-35.87	265.92	1.00
2017_8	-36.89	273.82	1.00

Table 1:Normalisation of PD_whole_tree with lograthmic regression. "b" represents the normalised value.

5. Beta diversity (Between-sample diversity): Acroess samples, alpha diversity is steady across the samples (mean±s.d.=271±27.2), whereas the phylum compositions are similar except for 2017_3 (figure 1). This sample shows strong enrichment in Bacteroidetes and Proteobacteria, and depletion in Acidobacteria.

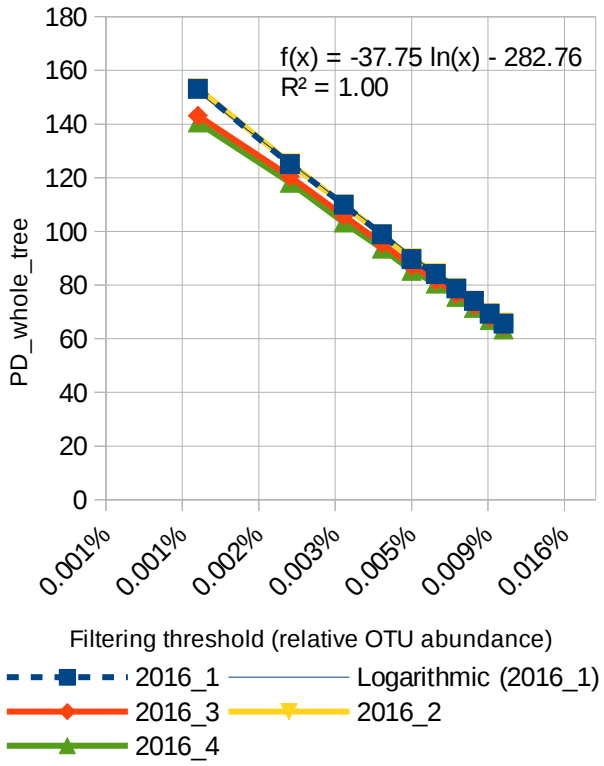


Figure 2: Alpha diversity (PD_whole_tree, y-axis) is dependent on filtering. Only 2016 data is shown for neatness. See Appendix_ for more data

To quantify to what extent this composition change matters, external datasets were imported. To illustrate their necessity, principal component analyses (PCoA) are performed using both the sample dataset only and using combined dataset (figure 3). In the sample-only plot, the within-environment diversity is displayed and the data has no principal orientation, with principal components of similar strengths. In the

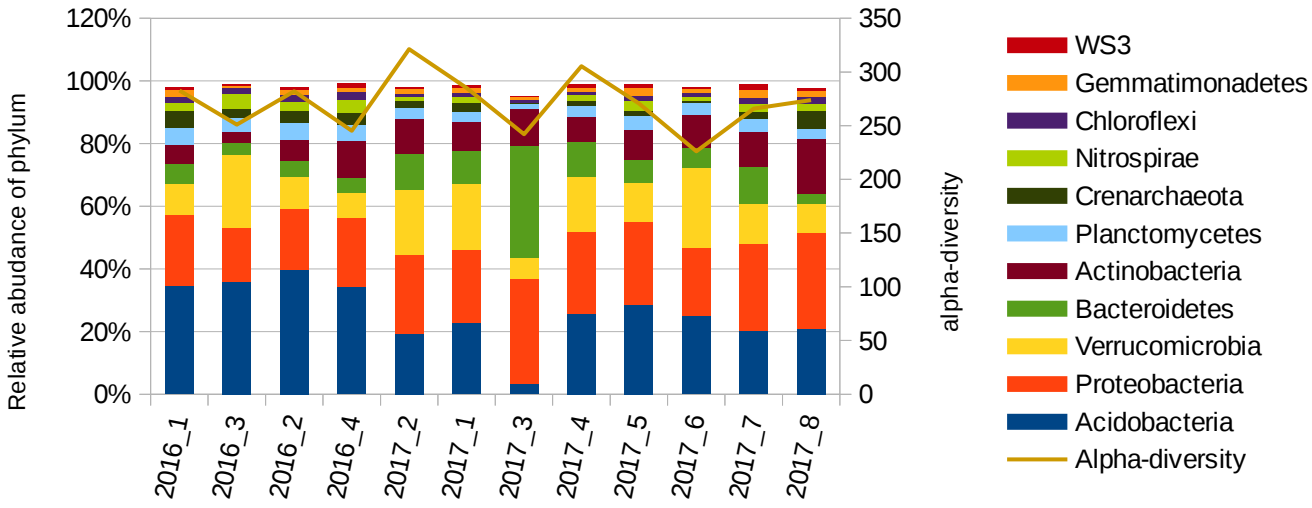


Figure 1: Taxonomic composition at phylum level

combined plot, however, a clear division is resulted between soil-based samples and other samples. Importantly, 2017_3 are apparent deviating from the major “sample” cluster, and this deviation is captured by “PC1” in the combined plot. We then ask what does “PC1” mean in terms of sample composition.

This division specified by “PC1” can be recaptured in PCoA plots using Bray-Curtis distance at different taxonomic levels. (Appendix). This division makes its appearance first at Class level, from which increasing taxa information only causes compaction of points and some minor divisions. Thus the division is inferred be visible at the Class level information

We then set order level “PC1” as a parameter for the sample and performed correlation tests that whether any order is associated with this parameter using Spearman correlation. Orders with a p-value<0.005 is recorded. This is repeated for the phylum level “PC1”(Table 2). We note an increase in “PC1” is associated with enrichment in Bacteroidetes and depletion in Alphaproteobacteria. The other hits are not

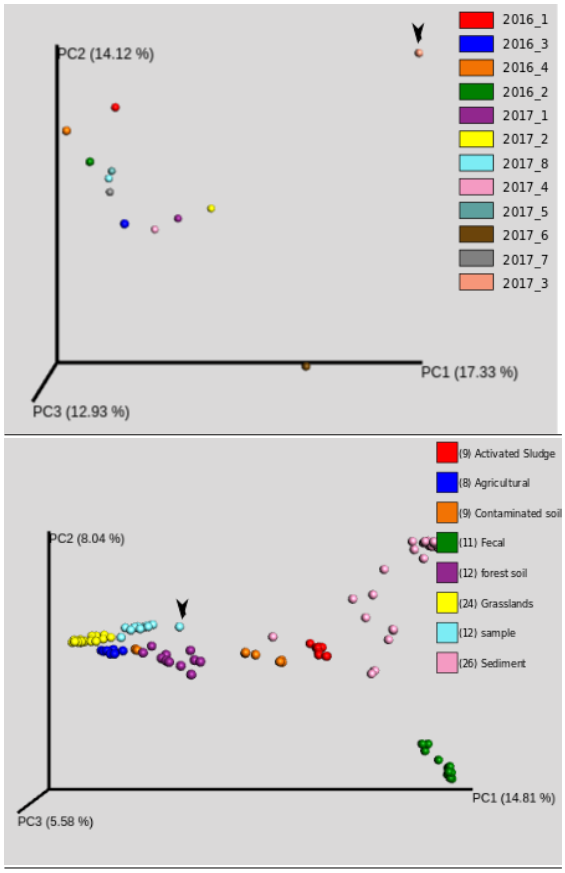


Figure 3: PCoA plots for sample-only (top) and combined (bottom) datasets. Arrowhead:2017_3

L2(Phylum)		bootstrapped
taxonomy	Test stat.	pval
k__Bacteria;p__GAL15	-0.8169216656	0
k__Bacteria;p__Bacteroidetes	0.7832167832	0.005
L3(Order)		bootstrapped
taxonomy	Test stat.	pval
k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria	-0.8601398601	0.001
k__Bacteria;p__Chloroflexi;c__TK10	-0.7832167832	0.002
k__Bacteria;p__Chloroflexi;c__Ktedonobacteria	-0.8041958042	0.003

Table 2: Taxonomic units whose abundance are statistically associated with the PC1 coordinate.

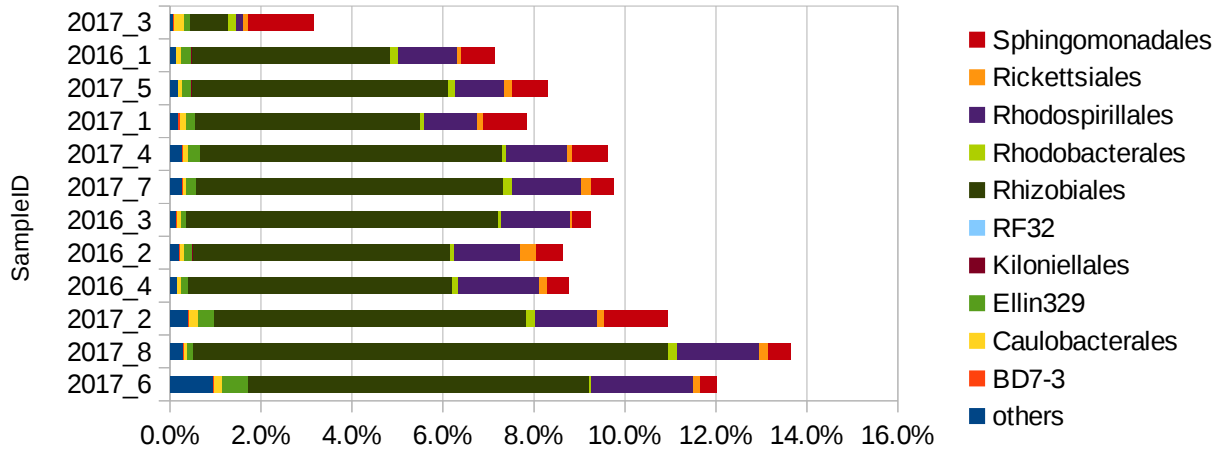


Figure 4: Detailed Alphaproteobacteria composition. SampleID ordered in descending PC1 (max at top)

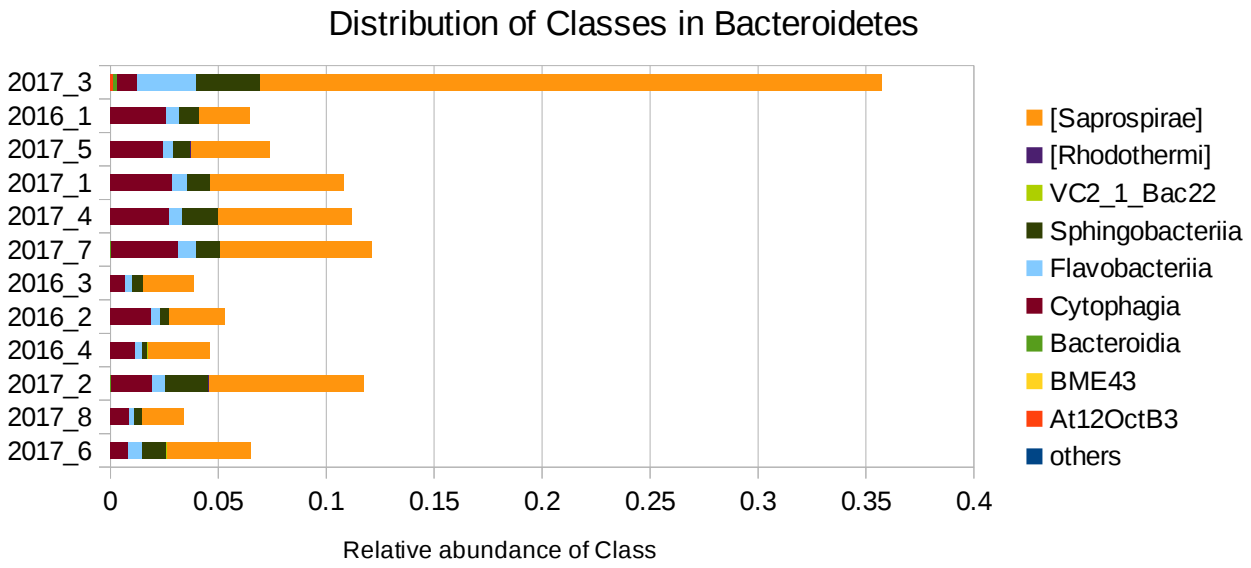


Figure 5: Detailed Bacteroidetes distribution. SampleID ordered in descending PC1 (max at top)

further analysed due to their low relative abundance (<0.01).

On a closer look, depletion of *Alphaproteobacteria* is mainly due to reduced *Rhizobiales* (figure 4), whereas increase in *Bacteroidetes* is mainly due to *Saprospirae* (figure 5).

Discussion

We identified an abnormal soil sample with *Rhizobiales* depleted and *Saprospirae* elevated. Studies have indicated a predator role for *Saprospirae* [14], whereas *Rhizobiales* is renowned for its symbiotic relation to plants due to its N-fixing ability[15–17]. There is also evidence rhizosphere bacteria composition is shifted up-regulated in presence of saprotrophic fungi[18].

These facts combined with the statistical analysis in the study warrant a hypothesis that the observed transition in taxonomic composition is associated with the decay of plants[19]. Several experiments are suggested: (1) 18S ITS fungal metagenomics to test whether saprotrophic fungi are elevated in 2017_3. (2) Analysing multiple samples on a spatial coordinate around the original collection site of 2017_3 to acquire more samples and to determine the influence radius of the decaying source, with the help of PCR assay using specific markers *Saprospirae* and *Rhizobiales*, to avoid expensive sequencing. (3) A more in-depth analysis of the metabolic network within 2017_3. (4) Correlate soil characteristic with level of decay.

Additionally, we confirm that the apparent alpha-diversity is linear to the logarithm of filtering threshold, which result in two robust parameters. It is an open question to determine their usefulness.

Appendix

Contain tables, figures.

Code for analysis can be found at: https://github.com/shouldsee/feng_metagenomics_2017

References

1. *Metzker, M. L. APPLICATIONS OF NEXT-GENERATION SEQUENCING Sequencing technologies - the next generation. Nat. Rev. Genet.* ***11,*** *31–46 (2010).*
2. Shakya, M. *et al.* Comparative metagenomic and rRNA microbial diversity characterization using archaeal and bacterial synthetic communities. *Environ. Microbiol.* **15,** 1882–1899 (2013).
3. Caporaso, J. G. *et al.* Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci. U. S. A.* **108 Suppl,** 4516–22 (2011).
4. Brooks, J. P. *et al.* The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. *BMC Microbiol.* **15,** 66 (2015).

5. Schoch, C. L. *et al.* Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 1–6 (2012).

6. Chen, K. & Pachter, L. Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS Comput. Biol.* **1,** 0106–0112 (2005).

7. Mendes, R. *et al.* Deciphering the Rhizosphere Microbiome for Disease-Suppressive Bacteria. *Science (80-.).* **332,** 1097–1100 (2011).

8. Caporaso, J. G. *et al.* correspondence QIIME allows analysis of high- throughput community sequencing data Intensity normalization improves color calling in SOLiD sequencing. *Nat. Publ. Gr.* **7,** 335–336 (2010).

9. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26,** 2460–2461 (2010).

10. Clarke, K. R., Somerfield, P. J. & Chapman, M. G. On resemblance measures for ecological studies, including taxonomic dissimilarities and a zero-adjusted Bray-Curtis coefficient for denuded assemblages. *J. Exp. Mar. Bio. Ecol.* **330,** 55–80 (2006).

11. Lozupone, C. & Knight, R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* **71,** 8228–8235 (2005).

12. Vázquez-Baeza, Y., Pirrung, M., Gonzalez, A. & Knight, R. EMPeror: a tool for visualizing high-throughput microbial community data. *Gigascience* **2,** 16 (2013).

13. Baron, M. *BIOC3301 is a UCL course undertaking regular metagenomics experiments*.

14. Pasternak, Z. *et al.* By their genes ye shall know them: genomic signatures of predatory bacteria. *ISME J.* **7,** 756–69 (2013).

15. Ivanova, E. G. *et al.* Facultative and obligate aerobic methylobacteria synthesize cytokinins. *Mikrobiologiia* **69,** 764–769 (2000).

16. Verginer, M. *et al.* Monitoring the plant epiphyte Methylobacterium extorquens DSM 21961 by real-time PCR and its influence on the strawberry flavor. *FEMS Microbiol. Ecol.* **74,** 136–145 (2010).

17. Erlacher, A. *et al.* Rhizobiales as functional and endosymbiontic members in the lichen symbiosis of Lobaria pulmonaria L. *Front. Microbiol.* **6,** 1–9 (2015).

18. De Boer, W. *et al.* Antifungal rhizosphere bacteria can increase as response to the presence of saprotrophic fungi. *PLoS One* **10,** 1–15 (2015).

*19. Rinta-Kanto, J. M. et al. Natural decay process affects the abundance and community structure of Bacteria and Archaea in Picea abies. FEMS Microbiol. Ecol. **92,** 1–10 (2016).*

Appendix:
Code used to process and analyse the data can be found at the following github repository:
https://github.com/shouldsee/feng_metagenomics_2017

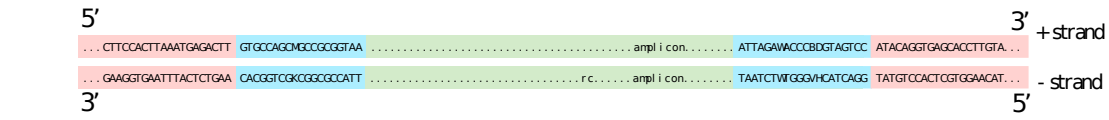
1.Soil characterisation
a. photo of collection site(incomplete)
2017_7:

b. table of soil details(incomplete)

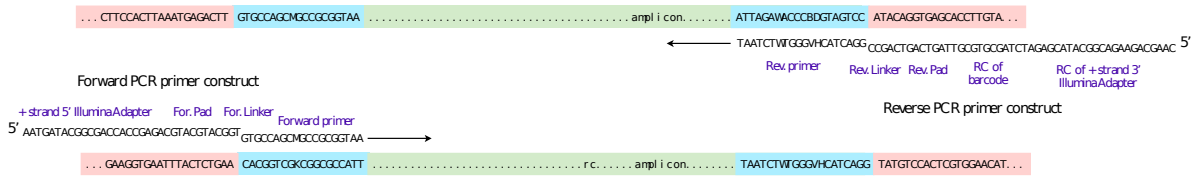
#SampleID	SampleType	Year	Month	Day	Latitude	Longitude	Description	Humidity	pH	Temperature
2016_1	soil	2016	1	18	51.524735	-0.13108	bioc3101.2016_1	0	0	0
2016_2	soil	2016	1	18	51.524392	-0.13112	bioc3101.2016_2	0	0	0
2016_3	soil	2016	1	18	51.523861	-0.1306	bioc3101.2016_3	0	0	0
2016_4	soil	2016	1	18	51.524309	-0.130543	bioc3101.2016_4	0	0	0
2017_1	soil	2017	2	6	51.524531	-0.1309062	bioc3301.2017_1	0	0	0
2017_2	soil	2017	2	6	51.524102	-0.130661	bioc3301.2017_2	0	7.5	5
2017_3	soil	2017	2	6	51.524341	-0.130825	bioc3301.2017_3	0	0	0
2017_4	soil	2017	2	6	51.524586	-0.131266	bioc3301.2017_4	0	8	6
2017_5	soil	2017	1	30	51.524074	-0.130824	bioc3301.2017_5	85	5.5	8
2017_6	soil	2017	1	30	51.524073	-0.130863	bioc3301.2017_6	0	0	0
2017_7	soil	2017	1	30	51.524371	-0.130757	bioc3301.2017_7	0	0	0
2017_8	soil	2017	1	30	51.524185	-0.130664	bioc3301.2017_8	0	0	0

2. Primer design and table.

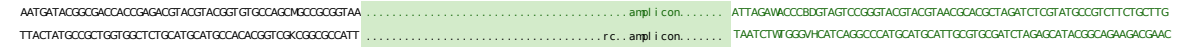
Target gene:



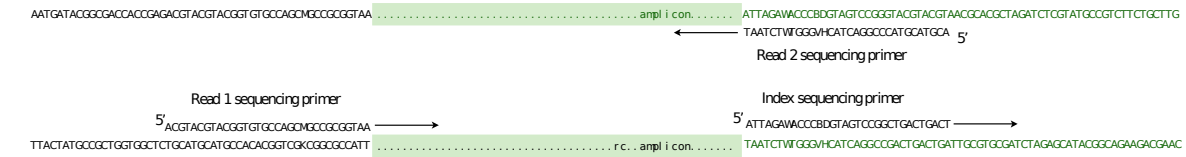
Amplif cation primers with annealing sites:



Amplif cation products:



Sequencing primers with annealing sites:



(adpated from [3])

Our primer design is similar to that depicted by the figure except that our barcode is inserted in 5' forward primer whereas in the figure it is inserted in 3'-backward primer.

The forward primer takes the form of 5'-(Illumina 5' Adaptor)-(Golay Barcode)-(Pad)-(Linker)-(5' Annealing Primer)-3':

5'-(AATGATACGGCGACCACCGAGATCTACACGCT)-(XXXXXXXXXXXX)-(TATGGTAATT)-(GT)-(GTGYCAGCMGCCGCGGTAA)-3'

The backward primer takes the form 5'-(RC of Illumina 3' adaptor)-(RC of Pad)-(RC of Linker)-(RC of 3' annealing primer)-3':

5'-(CAAGCAGAAGACGGCATACGAGAT)-(AGTCAGTCAG-(CC)-(GGACTACNVGGGTWTCTAAT)-3'

#SampleID	BarcodeSequence
2016_1	TACGAGACTGAT
2016_2	TGCTGTACGGAT
2016_3	TATCACCAGGTG
2016_4	TTGGTCAACGAT
2017_1	AGCCTTCGTCGC
2017_2	TCCATACCGGAA
2017_3	AGCCCTGCTACA
2017_4	CCTAACGGTCCA
2017_5	CGCGCCTTAAAC
2017_6	TATGGTACCCAG
2017_7	TACAATATCTGT
2017_8	AATTTAGGTAGG

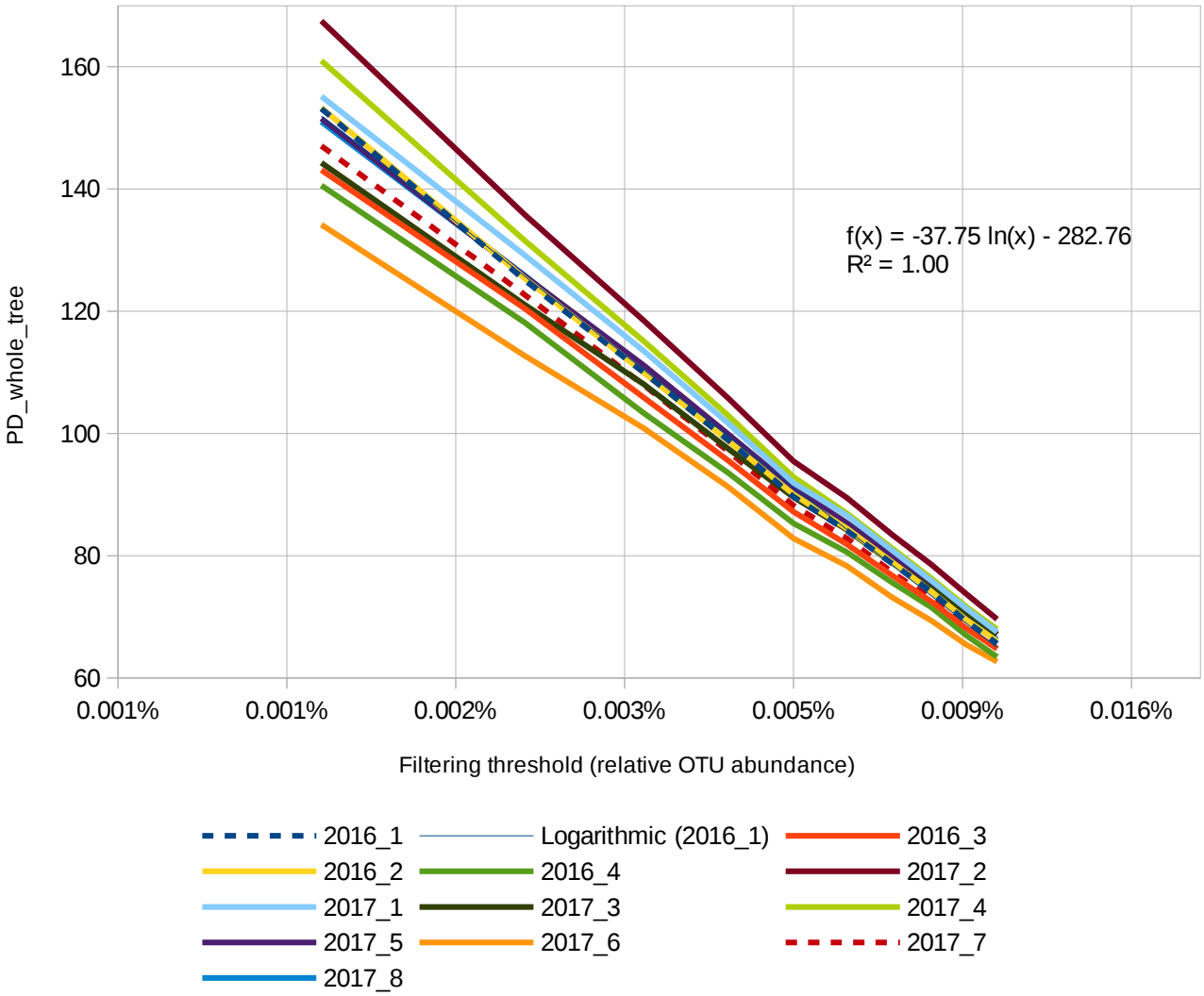
3. gDNA gel

The intensive band at high molecular weight indicate most gDNA is intact. The diffuse extended band indicate DNA fragements.

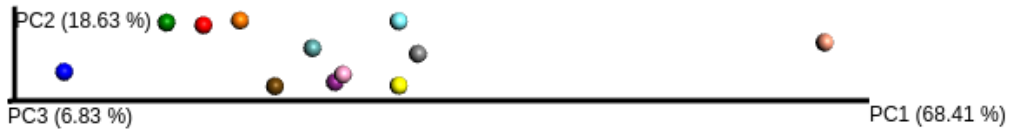
4. PCR gel

The number are in units of kb. The desired amplicon at ~400bp is clear intensified. However, in many most lanes, there is also band at 600bp, 700bp, and even at 1.0kb(2017_8). These bands are indicative of daisy-chain product after overamplification.

5. aApha diversity plot
alpha density is measured in PD_whole_tree. The x-axis is in lograthmic scale. Note -282.76 should actually be 282.76 due to fault in LibreOffice.



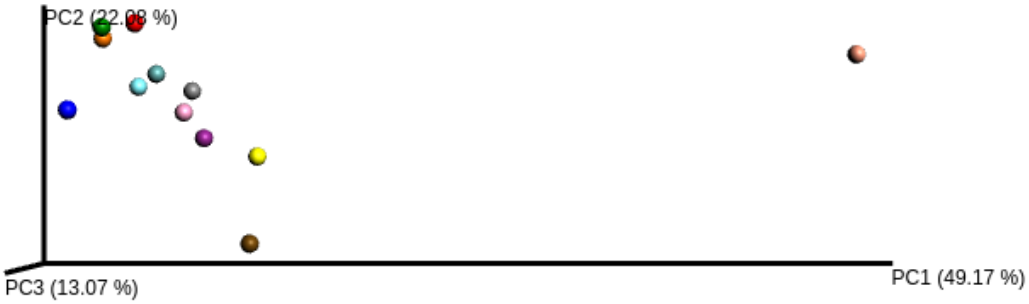
6. PCoA with bray-curtis distance
a)Sample-only dataset
L2_Phylum:

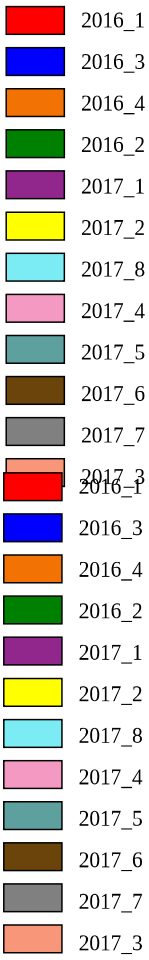


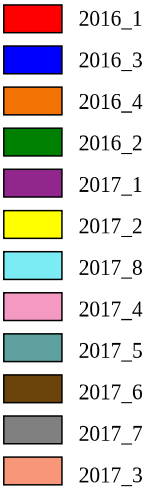
L3_Class:

L4_Order:

L5_Family:

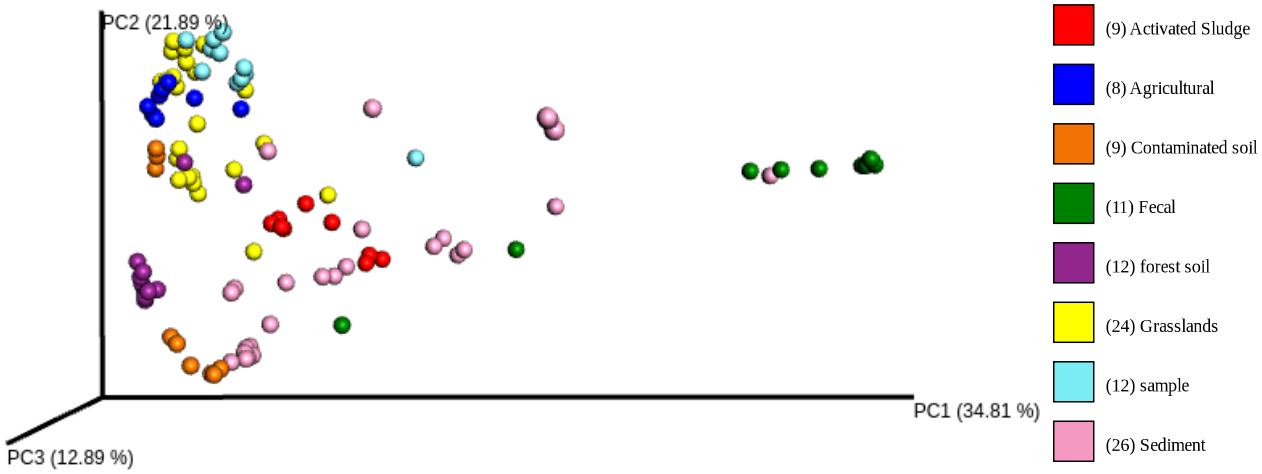




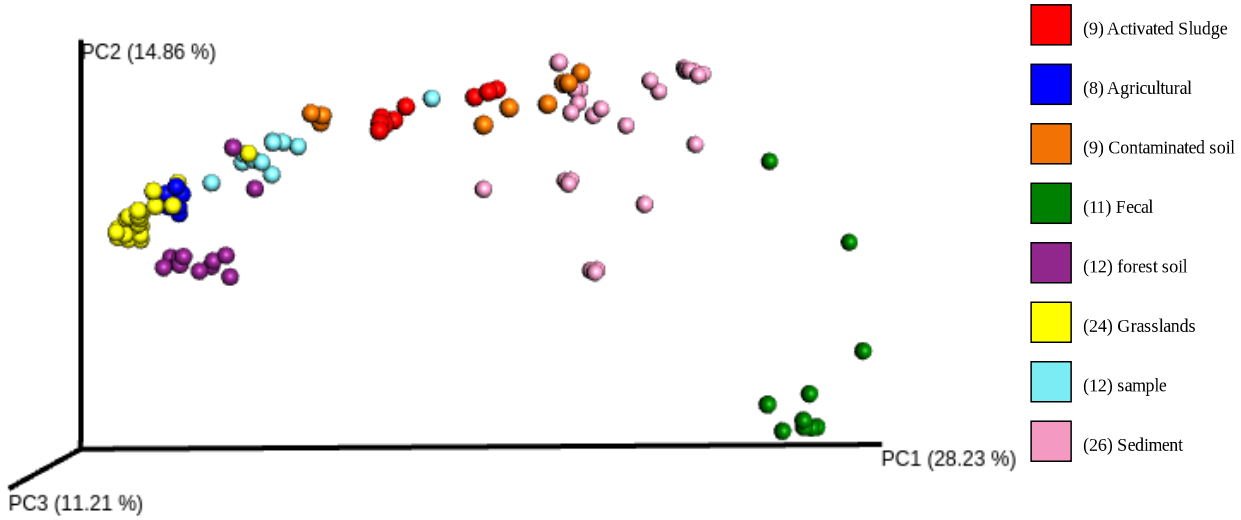


b)Combined dataset:

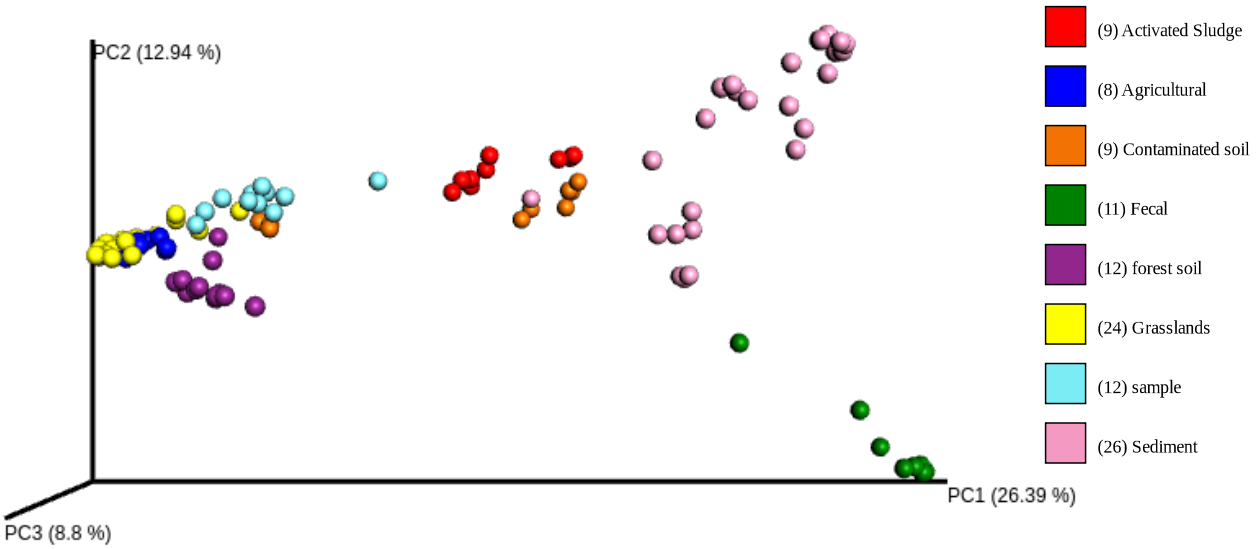
L2_Phylum:



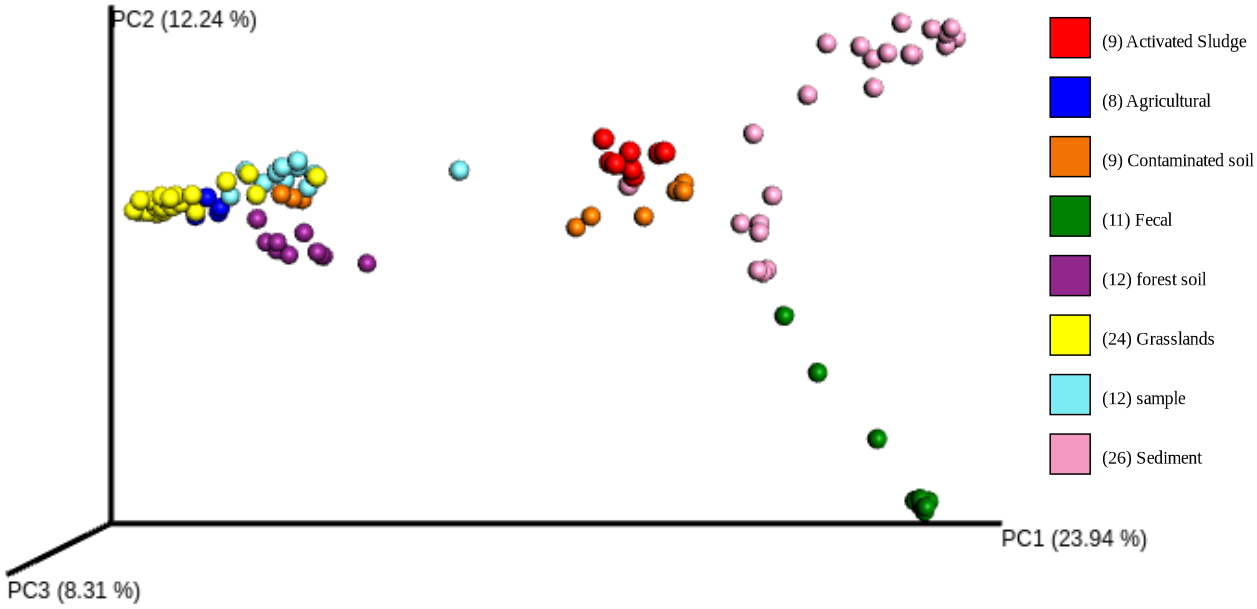
L3_Class:



L4_Order:



L5_Family:


PC2 (12.24 %)
PC1 (23.94 %)
PC3 (8.31 %)
(9) Activated Sludge
(8) Agricultural
(9) Contaminated soil
(11) Fecal
(12) forest soil
(24) Grasslands
(12) sample
(26) Sediment