
Ramirez, Kelly S.; Knight, Christopher G.; de Hollander, Mattias; Brearley, Francis Q.; Constantinides, Bede; Cotton, Anne; Creer, Si; Crowther, Thomas W.; Davison, John; Delgado-Baquerizo, Manuel; Dorrepaal, Ellen; Elliott, David R.; Fox, Graeme; Griffiths, Robert I.; Hale, Chris; Hartman, Kyle; Houlden, Ashley; Jones, David L.; Krab, Eveline J.; Maestre, Fernando T.; McGuire, Krista L.; Monteux, Sylvain; Orr, Caroline H.; van der Putten, Wim H.; Roberts, Ian S.; Robinson, David A.; Rocca, Jennifer D.; Rowntree, Jennifer; Schlaeppi, Klaus; Shepherd, Matthew; Singh, Brajesh K.; Straathof, Angela L.; Bhatnagar, Jennifer M.; Thion, Cécile; van der Heijden, Marcel G.A.; de Vries, Franciska T.. 2018. **Detecting macroecological patterns in bacterial communities across independent studies of global soils.** *Nature Microbiology*, 3 (2). 189-196. https://doi.org/10.1038/s41564-017-0062-x

SUPPLEMENTARY INFORMATION:

List of scripts and data:

1. Summary of datasets script - CleanAndCombineEnv_Final_JKR.R
2. ‘Sequence-matched’ sequence merging (De Hollander 2016): https://gitlab.bioinf.nioo.knaw.nl/amplicon-metagenomics/meta-16S
3. Taxonomy-based OTU table –
4. Sequence-matched OTU table –
5. Summary Datasets – summary_datsets.csv
6. Taxa list - importance for separating community and studies – Supplement_table3.csv
7. Figure generation code – Ramirez_etal.R
8. Figure generation data – Ramirez_etal.csv

Methods:

Primer Biases

It has long been well understood that different primers vary in their biases for amplifying members of the bacterial community[1,2]. To demonstrate this bias, the likelihood of significant differences in primer biases for the ten pairs of primers used in the studies analysed were determined by *in silico* analysis. Sequences of primer pairs were compared to all 16S rRNA gene sequences in the SILVA non-redundant reference database (SSURef NR) release 128[3] using TestPrime v1.0 (as described in[4]). The percentages of sequences of each bacterial phyla that matched both primers (with a one base pair mismatch allowance at least 1bp from the 3’ end of the primers) were calculated to compare predicted differences in primer coverage of different bacterial taxa.

1. Suzuki, M. T. & Giovannoni, S. J. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Appl. Environ. Microbiol.* **62,** 625–30 (1996).
2. Sipos, R. *et al.* Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. *FEMS Microbiol. Ecol.* **60,** 341–350 (2007).
3. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41,** D590-6 (2013).
4. Klindworth, A. *et al.* Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.* **41,** e1 (2013).
5. Jost, L. PARTITIONING DIVERSITY INTO INDEPENDENT ALPHA AND BETA COMPONENTS. *Ecology* **88,** 2427–2439 (2007).

Supplementary Table 1: See summarydatsets.csv for full table.

set	collection_da	OwnerRefNu	country	location	sequencing_	primers	seq_region	gene	processing_r	taxonomy_da
1	4/3/13	1	the netherlar	NA	454	577f/926r	NA	16s	mothur	silva
3	5/9/12	3	austria	odenwinkelkees	illumina	341f/806r	v3	16s	mothur	silva
3	9/9/12	3	switzerland	damma	illumina	341f/806r	v3	16s	mothur	silva
4	7/12/13	4	switzerland	zurich	illumina	799f/1193r	v5_v7	16s	qiime	greengenes
5	29/04/2014	5	uk	hertfordshire	illumina	515f/806r	v4	16s	macqiime	greengenes
6	16/05/12	6	uk	manchester	454	66f/518r	v1_v3	16s	amplicon noi:	greengenes
7	21/09/2009	7	uk	nafferton farm	454	357f/926r	v3_v5	16s	qiime	rdp
8	15/07/2007	8	usa	cheyenne	illumina	515f/806r	v4	16s	macqiime,ucl	greengenes
10	15/08/2009	10	usa	sagwonhills	illumina	515f/806r	v4	16s	macqiime,ucl	greengenes
11	1/17/11	11	uk	lincolnshire	454	27f/338r	v1_v2	16s	ampliconnois	greengenes
12	1/10/12	12	uk	manchester	454	27f/338r	v1_v2	16s	ampliconnois	greengenes
13	1/6/13	13	uk	wales	illumina	515f/806r	NA	16s	qiime	greengenes
16	1/3/12	16	botswana	kalahari	454	341f/907r	v3	16s	uclust	greengenes
18	6/7/10	18	uk	holme moss	454	341f/907r	v3	16s	uparse	greengenes
22	30/07/2015	22	malaysia	pasoh	illumina	515f/806r	v4	16s	uparse	rdp
24	23/07/2012	24	usa	Central Park, NYC	illumina	515f/806r	v4	16s	qiime	greengenes
26	14/11/2013	26	uk	South West Peninsula	454	NA	v1_v3	16s	NA	greengenes
30	NA	30	argentina	Lucas Cuesta	illumina	341f/805r	v3	16s	qiime	greengenes
30	NA	30	australia	Nevertire	illumina	341f/805r	v3	16s	qiime	greengenes
30	NA	30	chile	Choros_P1	illumina	341f/805r	v3	16s	qiime	greengenes
30	NA	30	iran	Sokeh	illumina	341f/805r	v3	16s	qiime	greengenes
30	NA	30	mexico	lvaro Obreg‹n	illumina	341f/805r	v3	16s	qiime	greengenes
30	NA	30	morocco	Saka2	illumina	341f/805r	v3	16s	qiime	greengenes
30	NA	30	spain	Barrax_CSA	illumina	341f/805r	v3	16s	qiime	greengenes
30	NA	30	tunisia	Tataouine	illumina	341f/805r	v3	16s	qiime	greengenes
30	NA	30	usa	EPES_3	illumina	341f/805r	v3	16s	qiime	greengenes
30	NA	30	venezuela	Tocuyo_P2	illumina	341f/805r	v3	16s	qiime	greengenes
30	NA	30	israel	IL_LH_6	illumina	341f/805r	v3	16s	qiime	greengenes
30	NA	30	australia	JM100	illumina	341f/805r	v3	16s	qiime	greengenes
31	23/07/14	31	sweden	suorooaivi (abisko)	illumina	341f/518r	v3	16s	qiime	greengenes
34	NA	34	china	NA	illumina	515f/806r	v4	16s	NA	NA
35	NA	35	uk	Scotland	illumina	515f/806r	v4	16s	NA	NA
36	NA	36	india	NA	illumina	515f/806r	v4	16s	NA	NA
37	NA	37	usa	NA	illumina	515f/806r	v4	16s	NA	NA
41	NA	41	panama	NA	illumina	515f/806r	v4	16s	NA	NA
43	1/4/08	43	uk	NA	454	27f/338r	NA	16s	NA	silva
46	2010	NA	usa	Harvard Forest	illumina	515f/806r	v4	16s	rdp	rdp
46	2010	NA	usa	Cedar Creek	illumina	515f/806r	v4	16s	rdp	rdp
46	2012	NA	usa	Konza Prarie	illumina	515f/806r	v4	16s	rdp	rdp
46	2011	NA	usa	Kellogg Biological Station	illumina	515f/806r	v4	16s	rdp	rdp
46	2012	NA	usa	Hawaii Experimental Tropical Forest	illumina	515f/806r	v4	16s	rdp	rdp
46	2010	NA	usa	Andrews Experimental Forest	illumina	515f/806r	v4	16s	rdp	rdp
46	2011	NA	usa	Hawaii Experimental Tropical Forest	illumina	515f/806r	v4	16s	rdp	rdp
46	2010	NA	usa	Konza Prarie	illumina	515f/806r	v4	16s	rdp	rdp
46	2011	NA	usa	Cedar Creek	illumina	515f/806r	v4	16s	rdp	rdp
46	2010	NA	usa	Bonanza Creek	illumina	515f/806r	v4	16s	rdp	rdp
46	2012	NA	usa	Cedar Creek	illumina	515f/806r	v4	16s	rdp	rdp
46	2011	NA	usa	Harvard Forest	illumina	515f/806r	v4	16s	rdp	rdp
46	2010	NA	usa	Bonanza Creek	illumina	515f/806r	v4	16s	rdp	rdp
46	2011	NA	usa	Konza Prarie	illumina	515f/806r	v4	16s	rdp	rdp
46	2011	NA	usa	Andrews Experimental Forest	illumina	515f/806r	v4	16s	rdp	rdp
46	2010	NA	usa	Hubbard Brook	illumina	515f/806r	v4	16s	rdp	rdp
46	2012	NA	usa	Harvard Forest	illumina	515f/806r	v4	16s	rdp	rdp
46	2010	NA	usa	Kellogg Biological Station	illumina	515f/806r	v4	16s	rdp	rdp
46	2010	NA	usa	Kellogg Biological Station	illumina	515f/806r	v4	16s	rdp	rdp
46	2012	NA	usa	Kellogg Biological Station	illumina	515f/806r	v4	16s	rdp	rdp
46	2012	NA	usa	Kellogg Biological Station	illumina	515f/806r	v4	16s	rdp	rdp
46	2011	NA	usa	Kellogg Biological Station	illumina	515f/806r	v4	16s	rdp	rdp
46	2011	NA	usa	Niwot Ridge	illumina	515f/806r	v4	16s	rdp	rdp
46	2012	NA	usa	Niwot Ridge	illumina	515f/806r	v4	16s	rdp	rdp
46	2010	NA	usa	Hawaii Experimental Tropical Forest	illumina	515f/806r	v4	16s	rdp	rdp
46	2012	NA	usa	Luquillo LTER	illumina	515f/806r	v4	16s	rdp	rdp
46	2012	NA	usa	Luquillo LTER	illumina	515f/806r	v4	16s	rdp	rdp
46	2011	NA	usa	Luquillo LTER	illumina	515f/806r	v4	16s	rdp	rdp
46	2010	NA	usa	Luquillo LTER	illumina	515f/806r	v4	16s	rdp	rdp
46	2012	NA	usa	Hawaii Experimental Tropical Forest	illumina	515f/806r	v4	16s	rdp	rdp
46	2011	NA	usa	Luquillo LTER	illumina	515f/806r	v4	16s	rdp	rdp
46	2012	NA	usa	Andrews Experimental Forest	illumina	515f/806r	v4	16s	rdp	rdp
46	2011	NA	usa	Andrews Experimental Forest	illumina	515f/806r	v4	16s	rdp	rdp
46	2011	NA	usa	Hubbard Brook	illumina	515f/806r	v4	16s	rdp	rdp
46	2010	NA	usa	Cedar Creek	illumina	515f/806r	v4	16s	rdp	rdp
46	2010	NA	usa	Hubbard Brook	illumina	515f/806r	v4	16s	rdp	rdp
46	2011	NA	usa	Bonanza Creek	illumina	515f/806r	v4	16s	rdp	rdp
46	2011	NA	usa	Cedar Creek	illumina	515f/806r	v4	16s	rdp	rdp
46	2012	NA	usa	Hubbard Brook	illumina	515f/806r	v4	16s	rdp	rdp
46	2010	NA	usa	Coweta	illumina	515f/806r	v4	16s	rdp	rdp
46	2010	NA	usa	Andrews Experimental Forest	illumina	515f/806r	v4	16s	rdp	rdp
46	2010	NA	usa	Coweta	illumina	515f/806r	v4	16s	rdp	rdp
46	2012	NA	usa	Andrews Experimental Forest	illumina	515f/806r	v4	16s	rdp	rdp
46	2011	NA	usa	Niwot Ridge	illumina	515f/806r	v4	16s	rdp	rdp
46	2012	NA	usa	Niwot Ridge	illumina	515f/806r	v4	16s	rdp	rdp
46	2012	NA	usa	Coweta	illumina	515f/806r	v4	16s	rdp	rdp
46	2010	NA	usa	Niwot Ridge	illumina	515f/806r	v4	16s	rdp	rdp
46	2011	NA	usa	Hawaii Experimental Tropical Forest	illumina	515f/806r	v4	16s	rdp	rdp
46	2012	NA	usa	Hubbard Brook	illumina	515f/806r	v4	16s	rdp	rdp
46	2012	NA	usa	Harvard Forest	illumina	515f/806r	v4	16s	rdp	rdp
46	2012	NA	usa	Konza Prarie	illumina	515f/806r	v4	16s	rdp	rdp
46	2011	NA	usa	Konza Prarie	illumina	515f/806r	v4	16s	rdp	rdp
46	2010	NA	usa	Konza Prarie	illumina	515f/806r	v4	16s	rdp	rdp
46	2012	NA	usa	Bonanza Creek	illumina	515f/806r	v4	16s	rdp	rdp
46	2011	NA	usa	Bonanza Creek	illumina	515f/806r	v4	16s	rdp	rdp
46	2010	NA	usa	Hawaii Experimental Tropical Forest	illumina	515f/806r	v4	16s	rdp	rdp
46	2011	NA	usa	Coweta	illumina	515f/806r	v4	16s	rdp	rdp
46	2012	NA	usa	Coweta	illumina	515f/806r	v4	16s	rdp	rdp
46	2010	NA	usa	Luquillo LTER	illumina	515f/806r	v4	16s	rdp	rdp
46	2011	NA	usa	Coweta	illumina	515f/806r	v4	16s	rdp	rdp
46	2010	NA	usa	Niwot Ridge	illumina	515f/806r	v4	16s	rdp	rdp
46	2011	NA	usa	Hubbard Brook	illumina	515f/806r	v4	16s	rdp	rdp
46	2010	NA	usa	Harvard Forest	illumina	515f/806r	v4	16s	rdp	rdp
46	2011	NA	usa	Harvard Forest	illumina	515f/806r	v4	16s	rdp	rdp

Supplementary Table 2: Results of *in silico* analysis to determine primer biases of primer pairs used to produce the analyzed study data. Percentages of sequences predicted to be amplified by the primers (allowing for a one base pair mismatch at least 1bp from the 3' end of the primers) by comparison to 16S RRNA gene sequences in the SILVA database are given for each domain and phylum.

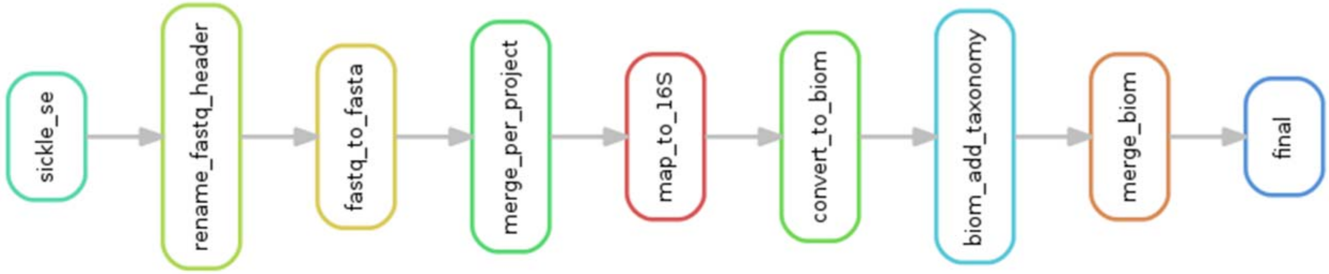
	Primer names									
	341F 806R	341F 518R	27F 338R	66F 518R	341F 805R	99F 1193R	341F 907R	357F 926R	515F 806R	577F 926R
	Percentage coverage of taxonomic group									
Archaea	1%	0%	0%	-	66%	-	0%	0%	94%	51%
Bacteria	93%	94%	81%	28%	94%	78%	94%	94%	94%	95%
Unclassified	28%	29%	36%	14%	30%	22%	29%	29%	31%	30%
Acidobacteria	96%	98%	86%	2%	96%	46%	97%	97%	96%	97%
Actinobacteria	86%	94%	77%	1%	95%	93%	96%	96%	85%	96%
Aquificae	92%	93%	10%	22%	95%	71%	90%	90%	95%	93%
Armatimonadetes	32%	33%	54%	0%	28%	28%	32%	32%	95%	95%
Bacteroidetes	95%	96%	85%	70%	95%	80%	95%	95%	95%	95%
Caldiserica	97%	75%	68%	-	99%	76%	99%	99%	94%	99%
Chlamydiae	68%	66%	4%	-	72%	36%	69%	69%	94%	98%
Chlorobi	95%	95%	93%	-	95%	86%	95%	95%	96%	98%
Chloroflexi	82%	88%	52%	1%	81%	29%	87%	87%	87%	94%
Chrysiogenetes	100%	100%	50%	-	100%	100%	78%	78%	100%	89%
Deferribacteres	96%	98%	89%	3%	96%	93%	97%	97%	96%	96%
Deinococcus-Thermus	97%	97%	84%	0%	96%	72%	97%	97%	96%	98%
Dictyoglomi	100%	100%	33%	-	100%	-	89%	89%	89%	89%
Elusimicrobia	98%	99%	94%	3%	97%	74%	96%	96%	98%	94%
Fibrobacteres	95%	96%	82%	2%	95%	83%	93%	93%	96%	94%
Fusobacteria	94%	93%	64%	1%	94%	93%	91%	91%	93%	93%
Gemmatimonadetes	95%	98%	89%	1%	94%	90%	96%	96%	94%	96%
Lentisphaerae	86%	87%	77%	1%	94%	5%	87%	87%	94%	91%
Planctomycetes	33%	33%	30%	1%	90%	10%	33%	33%	94%	96%
Proteobacteria	96%	97%	83%	55%	96%	84%	96%	96%	96%	96%
Spirochaetes	87%	93%	82%	0%	94%	86%	94%	94%	87%	96%
Synergistetes	96%	98%	91%	1%	92%	18%	98%	98%	94%	97%
Tenericutes	93%	94%	84%	0%	94%	56%	82%	82%	96%	88%
Thermodesulfobacteria	100%	98%	71%	2%	100%	90%	100%	100%	100%	98%
Thermotogae	96%	93%	60%	1%	95%	59%	97%	97%	94%	97%
Verrucomicrobia	92%	95%	24%	1%	92%	27%	90%	90%	93%	92%
Acetothermia	100%	100%	57%	-	96%	56%	72%	72%	96%	72%
Aminicenantes	95%	96%	87%	2%	94%	0%	96%	96%	96%	95%
Atribacteria	100%	100%	100%	4%	97%	87%	100%	100%	100%	100%
BRC1	94%	96%	80%	1%	97%	2%	96%	96%	95%	98%
candidate division WPS-1	30%	29%	15%	-	66%	1%	30%	30%	93%	96%
candidate division WPS-2	2%	2%	4%	1%	93%	2%	2%	2%	92%	96%
candidate division ZB3	98%	100%	94%	9%	98%	44%	100%	100%	98%	100%
Candidatus Calescamantes	100%	100%	100%	-	100%	-	100%	100%	100%	100%
Candidatus Saccharibacteria	95%	93%	87%	2%	95%	6%	4%	4%	95%	95%
Cloacimonetes	95%	96%	88%	1%	92%	43%	94%	94%	90%	91%
Cyanobacteria/Chloroplast	93%	94%	80%	2%	92%	0%	94%	94%	94%	96%
Firmicutes	95%	95%	85%	2%	94%	84%	95%	95%	94%	94%
Hydrogenedentes	90%	96%	7%	5%	91%	19%	94%	94%	94%	98%
Ignavibacteriae	93%	95%	89%	1%	92%	94%	95%	95%	95%	98%
Latescibacteria	97%	96%	89%	1%	97%	37%	98%	98%	95%	96%
Marinimicrobia	89%	91%	86%	6%	93%	66%	90%	90%	95%	98%
Microgenomates	-	18%	6%	-	-	-	-	-	49%	76%
Nitrospinae	99%	99%	88%	4%	99%	2%	100%	100%	98%	98%
Nitrospirae	95%	96%	83%	6%	95%	83%	96%	96%	94%	95%
Omnitrophica	100%	100%	75%	-	83%	44%	100%	100%	100%	100%
Parcubacteria	70%	31%	63%	-	96%	-	65%	65%	52%	90%
Poribacteria	89%	87%	42%	-	89%	24%	31%	31%	87%	29%
SR1	91%	93%	74%	1%	93%	-	-	-	96%	-
unclassified_Bacteria	78%	77%	74%	5%	81%	43%	76%	76%	89%	92%

Supplementary Table 3 Shannon diversity calculated within (alpha) and between (beta) all samples and overall (gamma) according to (Jost 2007)[5]. Values given with Standard errors (calculated using 100 bootstrap replicates), with number equivalents in parentheses below.

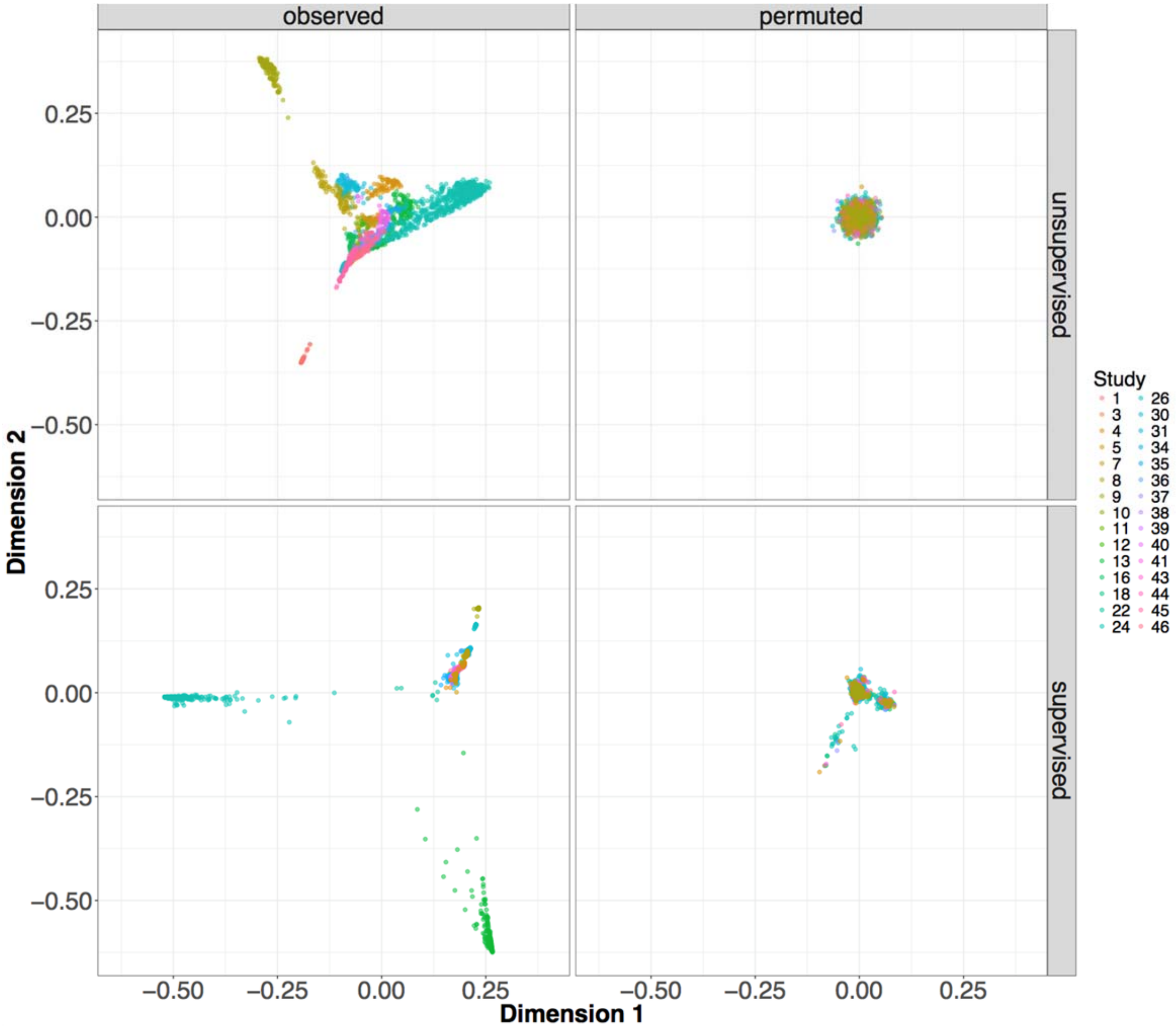
	Alpha	Beta	Gamma
Observed data	4.73 ± 0.004	0.947 ± 0.015	5.68 ± 0.022
	(114± 0.021)	(2.58 ± 0.870)	(293± 4.8)
Permutated data	4.80 ± 0.003	0.909 ± 0.017	5.71 ± 0.022
	(121± 0.022)	(2.48 ± 0.943)	(301± 5.50)

Supplementary Table 4: Taxa list - importance for separating community and studies - Stable3.docx

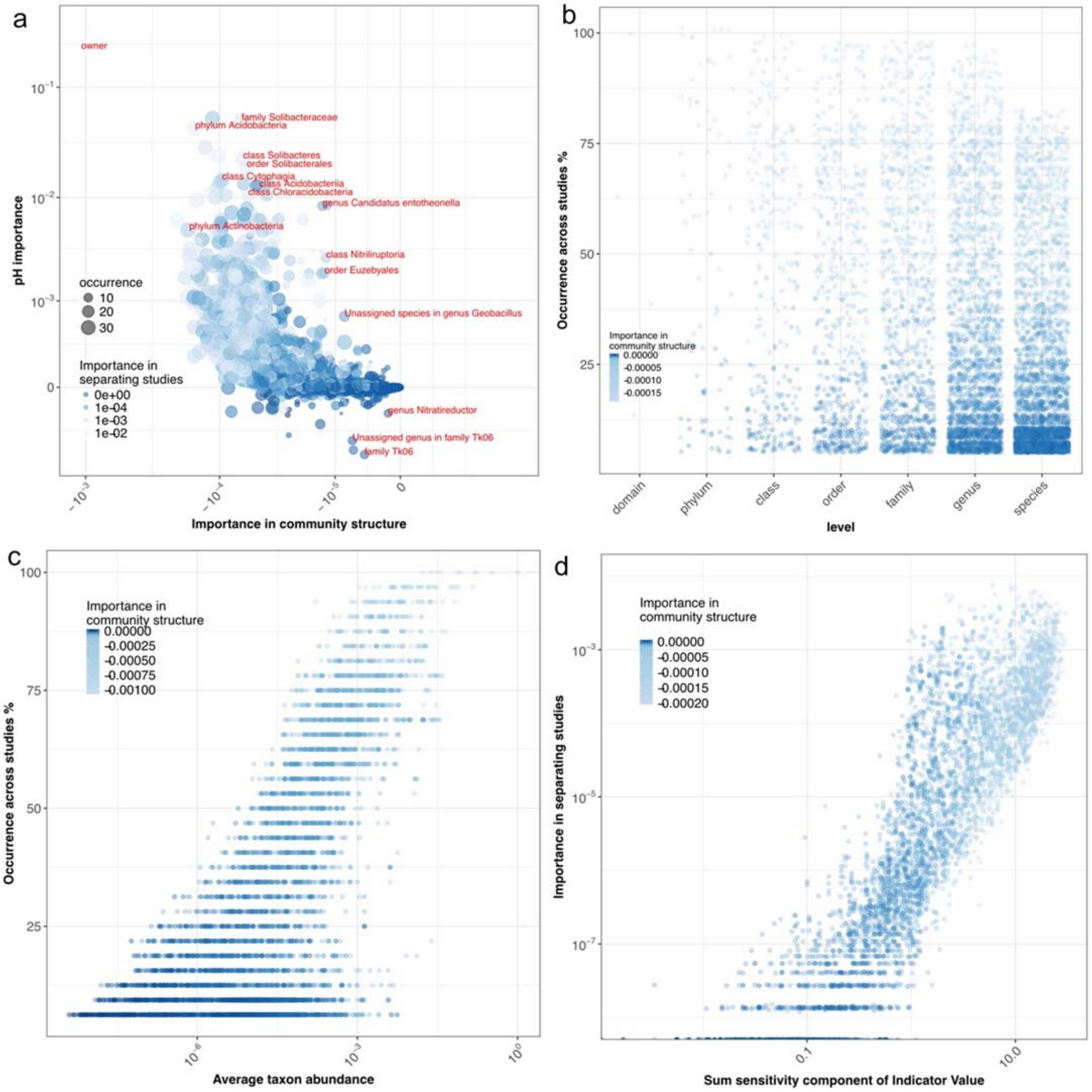
SUPPLEMENTARY FIGURES



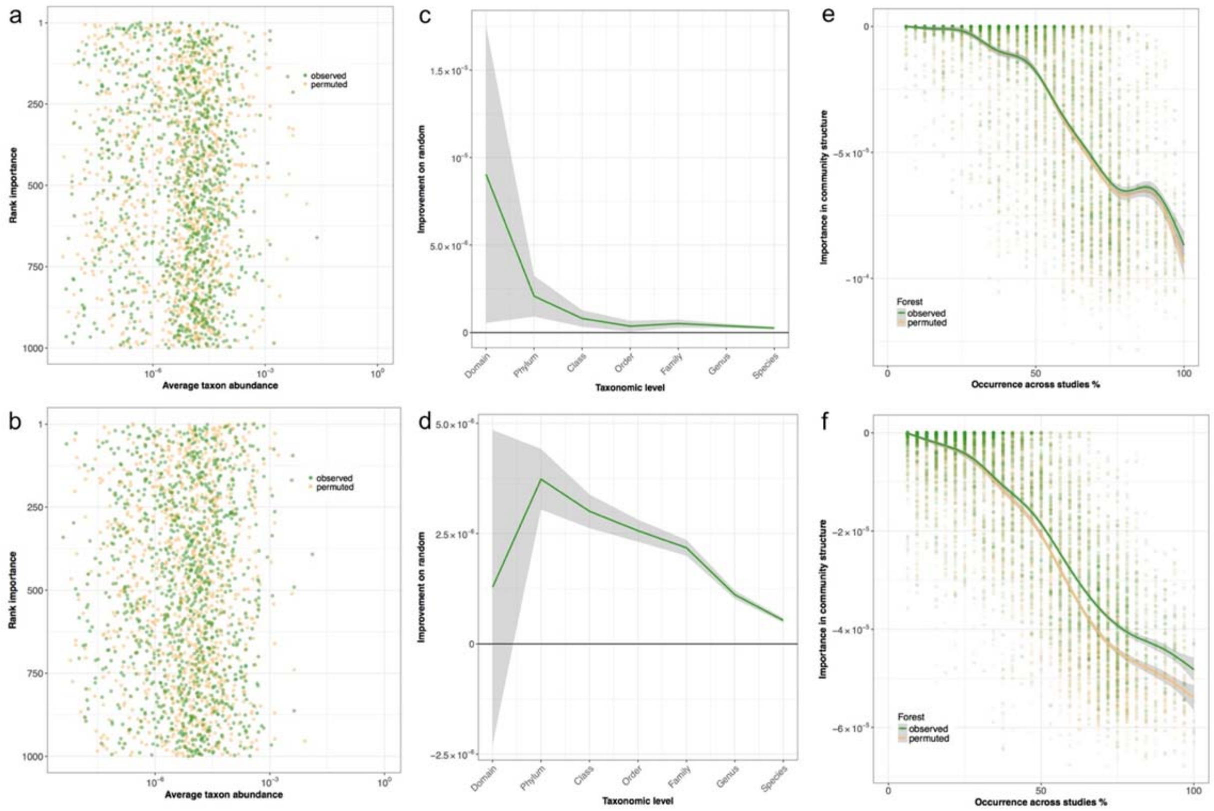
Supplementary Figure 1: Workflow to merge raw sequence data:



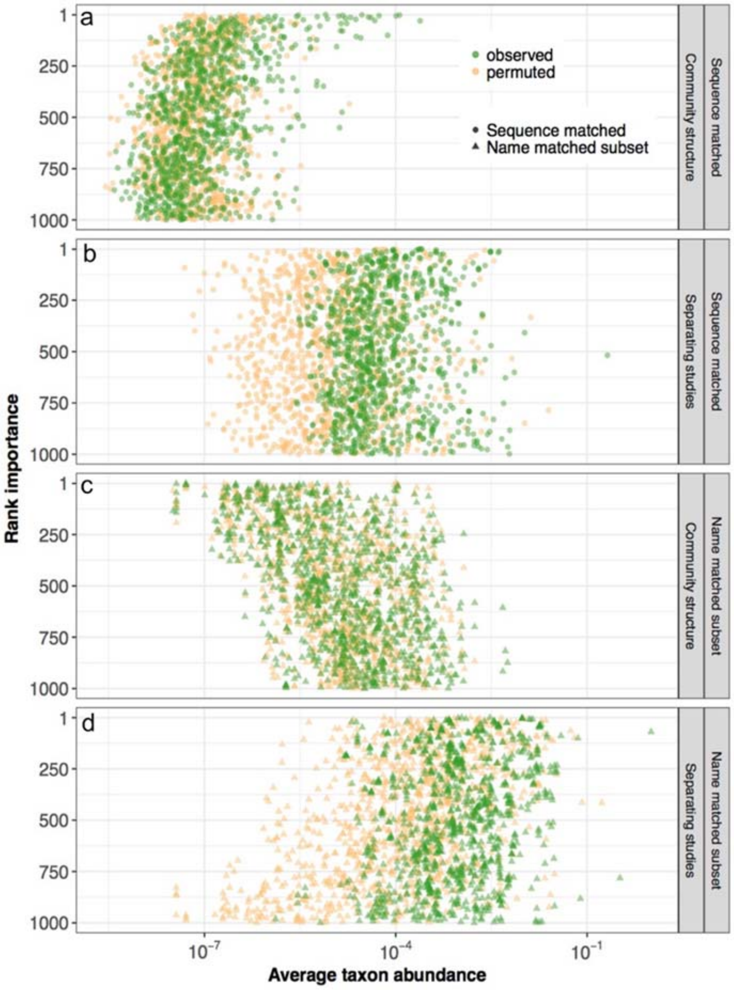
Supplementary Figure 2: Two-dimensional multi-dimensional scaling (MDS) plots for both observed and permuted data. MDS was applied to the proximity matrices derived from the unsupervised (community structure) and the supervised (separating studies) Random Forest analyses. Colored by study number.



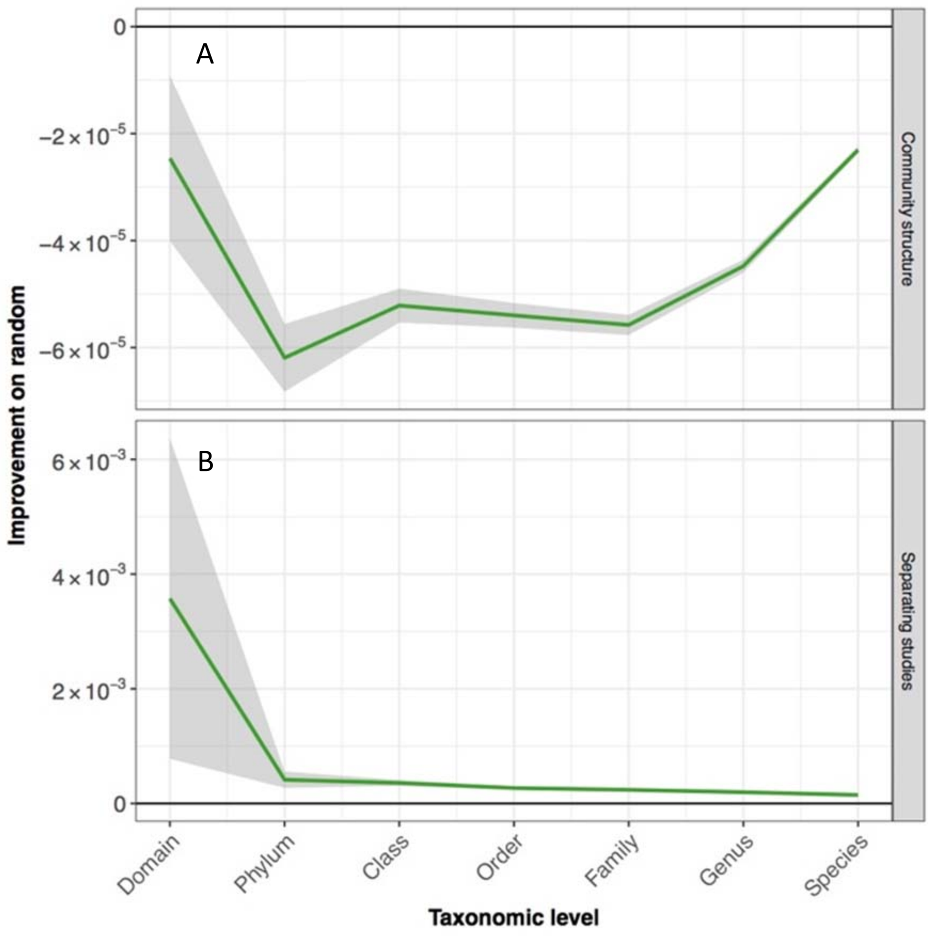
Supplementary Figure 3: a.) A supervised Random Forest model was fitted to predict pH from taxa and technical variables (in the same way as the supervised model separating studies described in the Methods). The importance of taxa and technical variables in this model is plotted against their importance for community structure, colored such that taxa confounded with technical variables (important for separating studies) are paler than those with low association with particular studies. ‘owner’ predicts pH the best and the phylum Acidobacteria is second best at separating studies. However, neither strongly associated with community structure. **b.)** Taxa of lower taxonomic rank tend to be detected in fewer studies ($\rho = 0.3$). Similarly, **c.)** low abundance taxa tend to be detected in fewer studies ($\rho = 0.59$). Finally, **d.)** the importance for separating studies given by the supervised Random Forest model correlates closely with the sensitivity component of the indicator value of a given taxon ($\rho = 0.89$). In b-d, darker colors indicate taxa more important in the model of community structure.



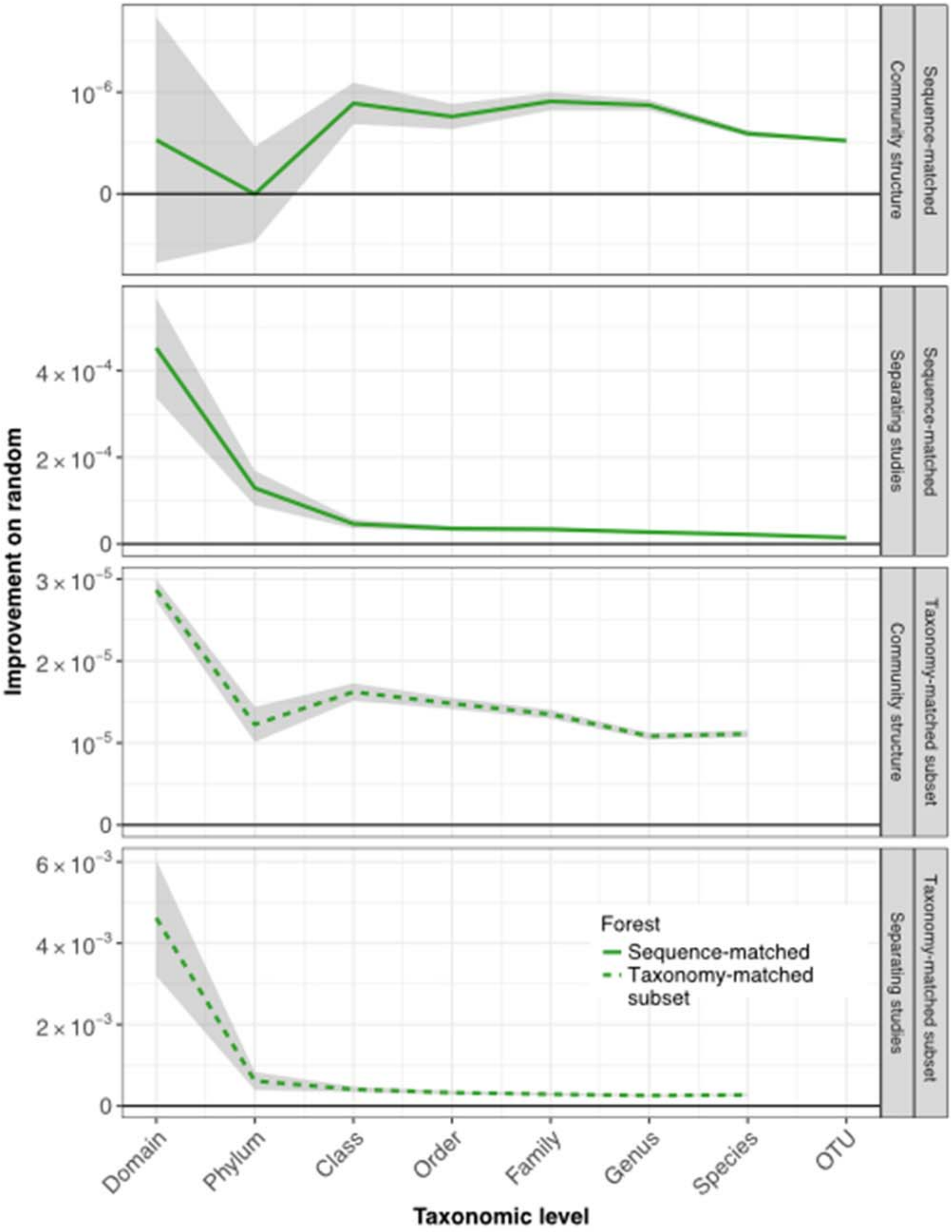
Supplementary Figure 4: Assessment of the community structure of two of the largest individual studies within the wider dataset: from Central Park, NYC encompassing 594 samples (study #24) (*top panels*) and a global dataset encompassing 103 samples (study #30) (*bottom panels*) demonstrates that there is **a,b)** no power to see associations of community structure with low abundance taxa, **c,d)** the relative importance of different taxonomic levels varies both among studies and from the analysis across studies (Figure 4) and **e,f)** there is power to separate observed from permuted data, but this is less than observed across the full dataset (Figure 5) and the stable 'core' soil taxa of high taxonomic level and high abundance identified in the full dataset (Figure 5) is not visible in the individual datasets. These analyses were completed as described for Figures 3, 4 and 5 in the main text.



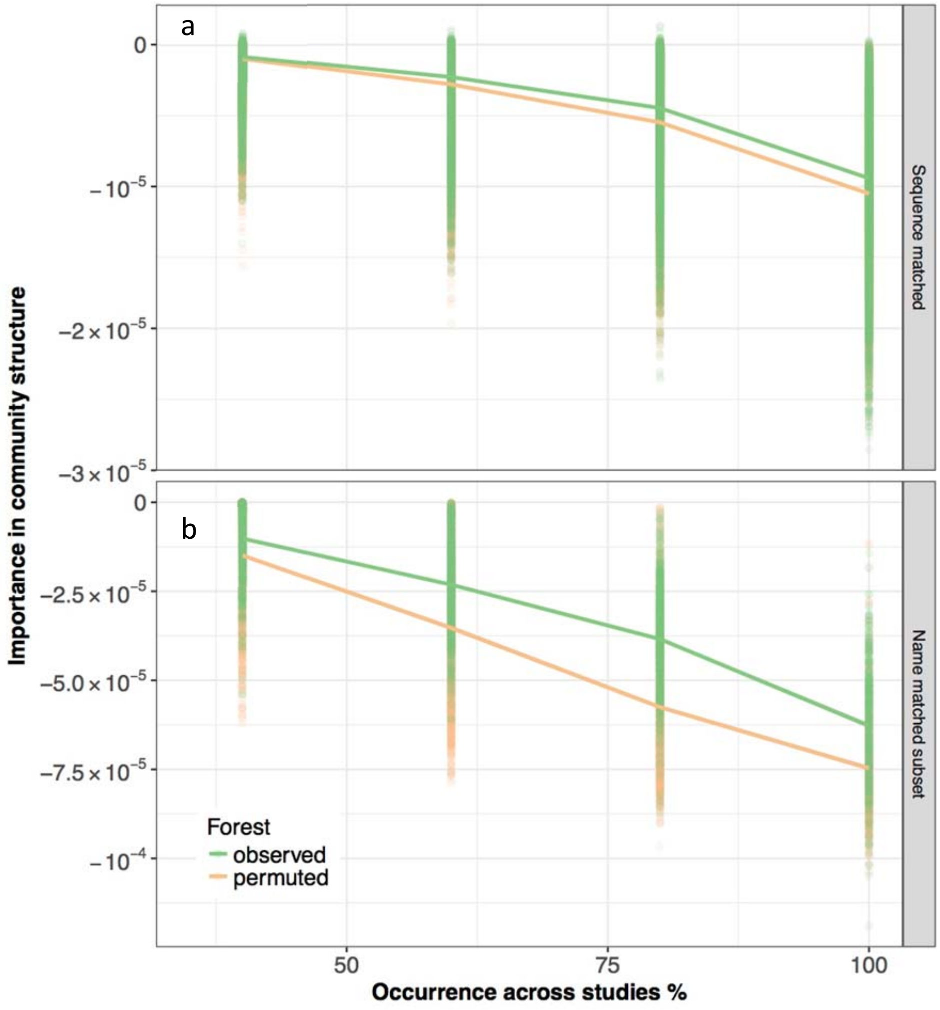
Supplementary Figure 5. The average abundance of the 1000 most important taxa in the analysis of the sequence-matched sequence dataset (**a b**) and of equivalent analyses of the same 5 studies when name-matched (**c, d**). While, the results look similar to the full dataset (Figure 3) for the models separating studies (b and d) there is no distinction between observed and permuted data in the community structure models (a and c). We see very comparable patterns between sequence-matched and name-matched datasets (a and b versus c and d).



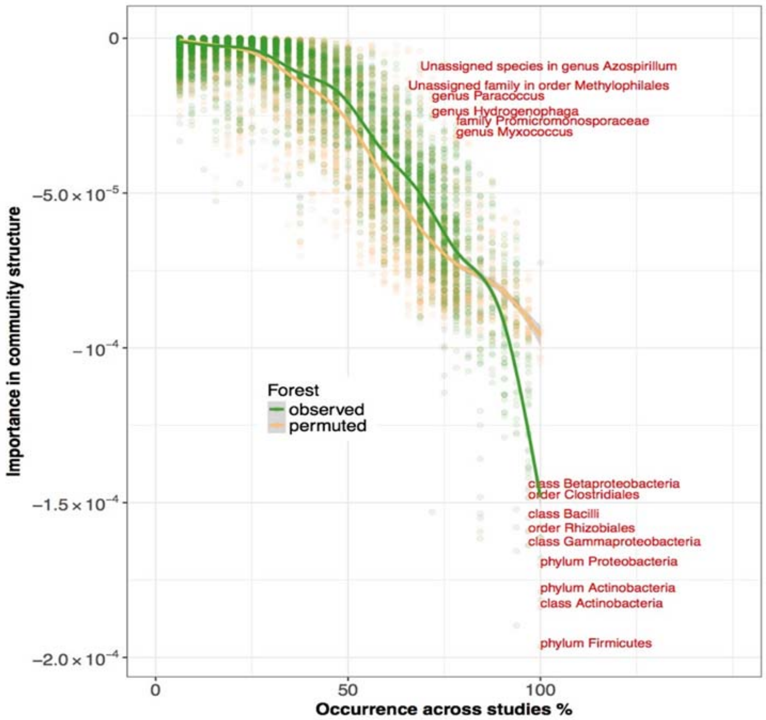
Supplementary Figure 6. The importance of bacterial taxa classified at different taxonomic ranks when considering only presence/absence data (i.e. without abundance information). While lower taxonomic resolution is more important for separating studies (b) it is still possible to conclude that there is a stable core soil microbiome and the most stable taxonomic level is phylum (a). The lines and grey ribbons show the mean and standard error respectively of these values across taxa at each taxonomic level considered.



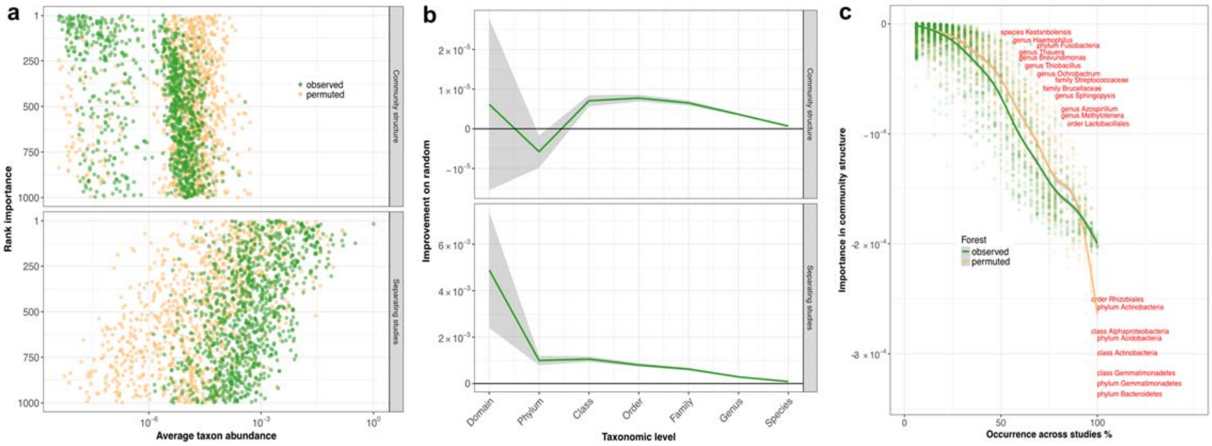
Supplementary Figure 7. The importance of bacterial taxa classified at different taxonomic ranks As shown in Figure 4 of the main text, but here **a,b)** the sequence-matched data and **c,d)** equivalent analyses of the same 5 studies when name-matched.



Supplementary Figure 8. As shown in Figure 5, but here **a)** the sequence-matched data shown in comparison to **b)** equivalent analysis of the same 5 studies when name-matched. Lines connect mean values, confidence intervals not visible outside the lines.



Supplementary Figure 9: A filtered subset of the data where only taxa present at above 0.003% in any given sample were included in this analysis. Other aspects equivalent to Figure 5 of the main text.



Supplementary Figure 10. Equivalent analyses to Figures 3, 4 and 5 (respectively **a**, **b**, and **c**) on a dataset in which all taxa unclassified at any level were removed (see Methods). The results are similar to analysis of the full dataset (see the main text figures for details).