Supplementary Information:

Detecting macroecological patterns in bacterial communities across independent studies of global soils

Authors: Kelly S Ramirez*[+1], Christopher G. Knight[+2], Mattias de Hollander[1], Francis Q. Brearley[3], Bede Constantinides[4], Anne Cotton[5], Si Creer[6], Thomas W. Crowther[1,7], John Davison[8], Manuel Delgado-Baquerizo[9], Ellen Dorrepaal[10], David R. Elliott[3,11], Graeme Fox[3], Rob Griffiths[12], Chris Hale[13], Kyle Hartman[14], Ashley Houlden[15], David L. Jones[6], Eveline J. Krab[10], Fernando T. Maestre[16], Krista L. McGuire[17], Sylvain Monteux[10], Caroline H. Orr[18], Wim H van der Putten[1,19], Ian S. Roberts[15], David A. Robinson[20], Jennifer D. Rocca[21], Jennifer Rowntree[3], Klaus Schlaeppi[14], Matthew Shepherd[22], Brajesh K. Singh[23], Angela L. Straathof[2], Jennifer M. Bhatnagar[24], Cécile Thion[25], Marcel G.A. van der Heijden[14,26,27], and Franciska T. de Vries[2]

Supplementary Table 1: Description of all datasets and samples within data used in the analyses. See 'summary_datsets.csv'.

Supplementary Table 2: Primer bias by primer pair. 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.

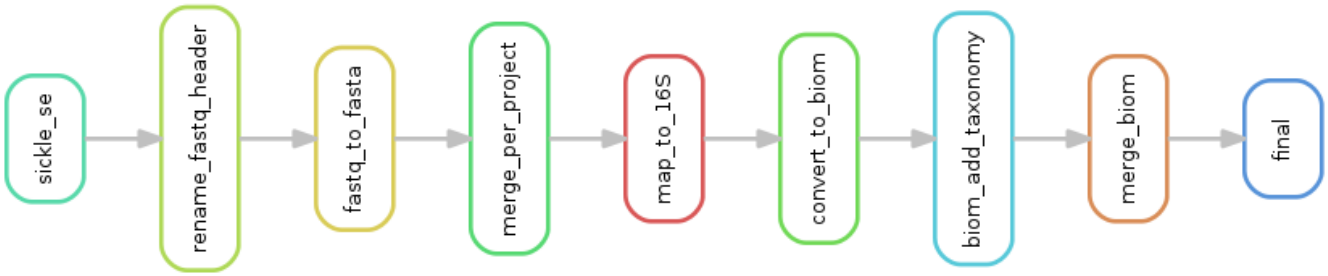
	Primer Names									
	341F-806R	341F-518R	27F-338R	66F-518R	341F-805R	799F-1193R	341F-907R	357F-926R	515F-806R	577F-926R
	Percentage Coverage of Taxonomic Group									
Archaea	1%	0%	0%	-	66%	-	0%	0%	94%	51%
Bacteria	93%	94%	81%	28%	94%	78%	94%	94%	94%	95%
Unclassified	28%	29%	36%	14%	30%	22%	29%	29%	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 of observed and permuted data. Diversity was alculated 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.

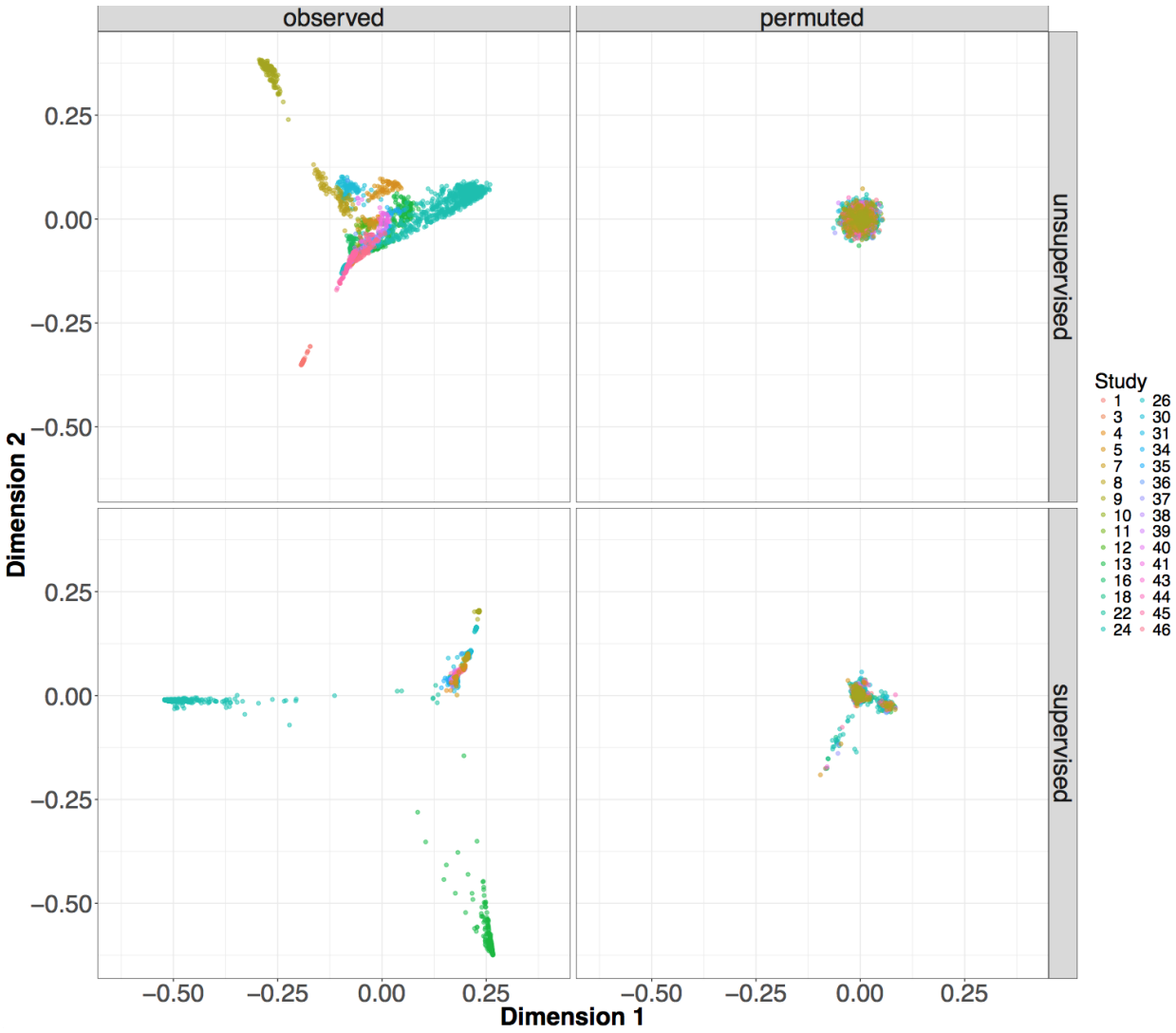
	Alpha	Beta	Gamma
Observed data	4.73 ± 0.004	0.947 ± 0.015	5.68 ± 0.022
	(114± 0.021)	(2.58 ± 0.870)	(293± 4.8)
Permutated data	4.80 ± 0.003	0.909 ± 0.017	5.71 ± 0.022
	(121± 0.022)	(2.48 ± 0.943)	(301± 5.50)

Supplementary Table 4: Taxa importance for separating communities and studies. See Ramirez_etal_data.csv

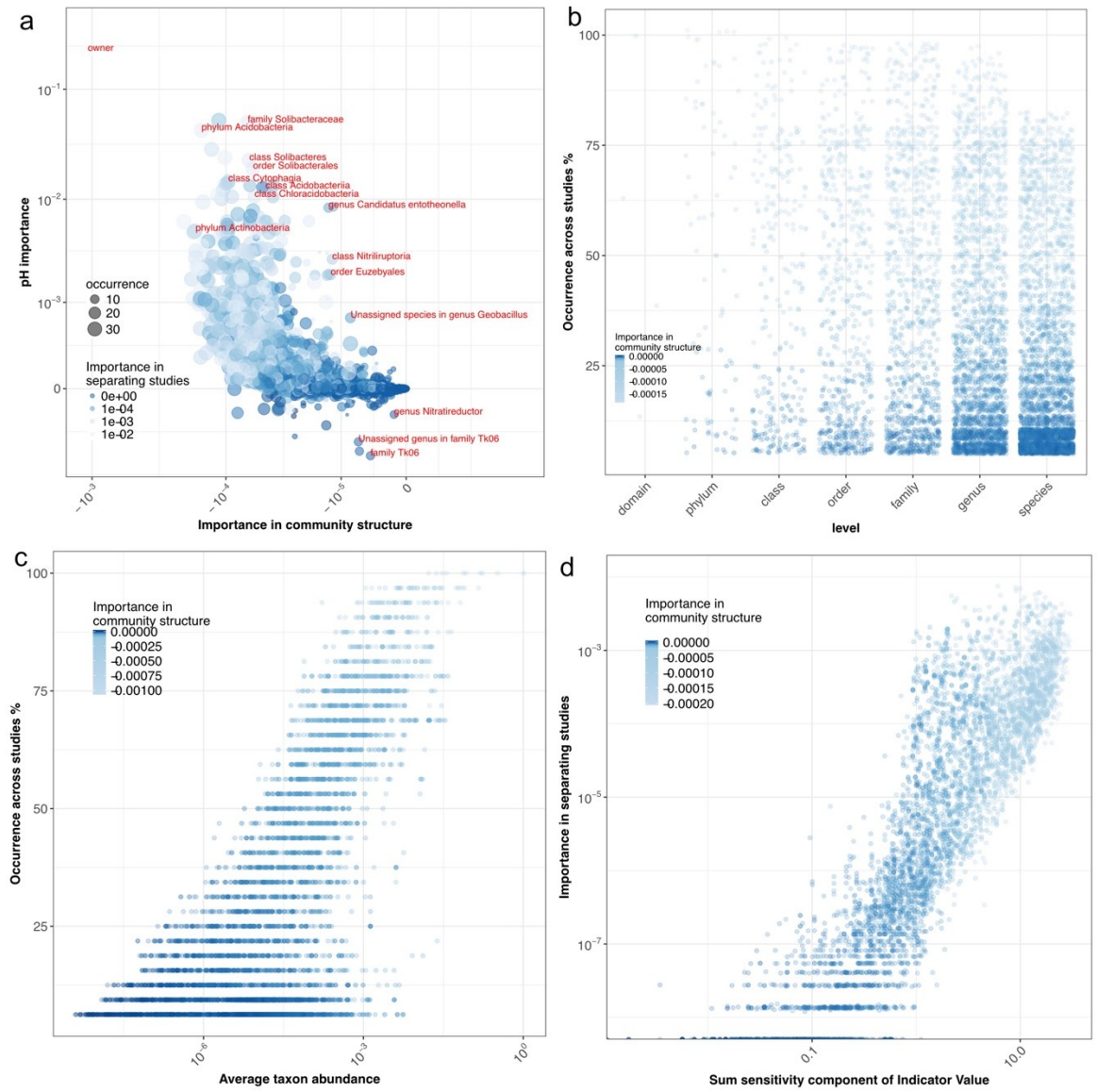
Supplementary Figures



Supplementary Figure 1: Workflow to merge raw sequence data ((De Hollander 2016).

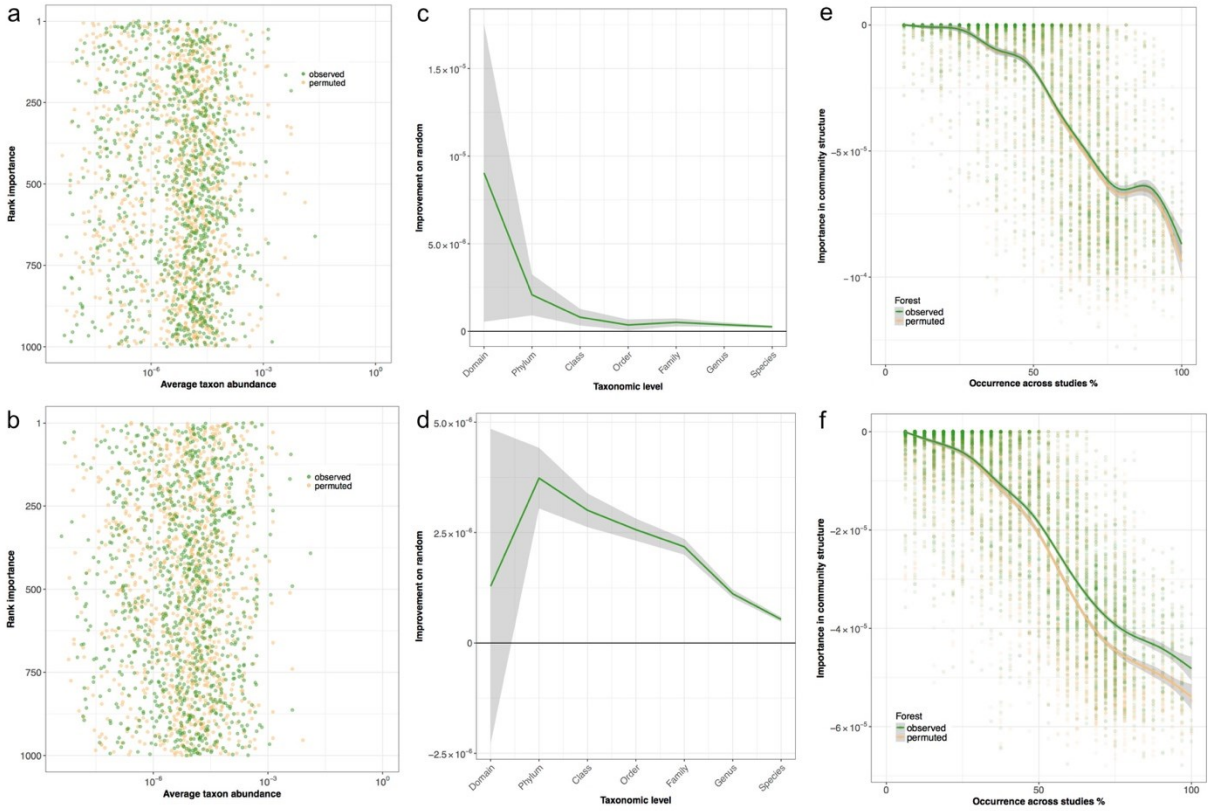


Supplementary Figure 2: Two-dimensional multi-dimensional scaling (MDS) plots for both observed and permuted data. MDS was applied to the proximity matrices derived from the unsupervised (community structure) and the supervised (separating studies) Random Forest analyses. Colored by study number.

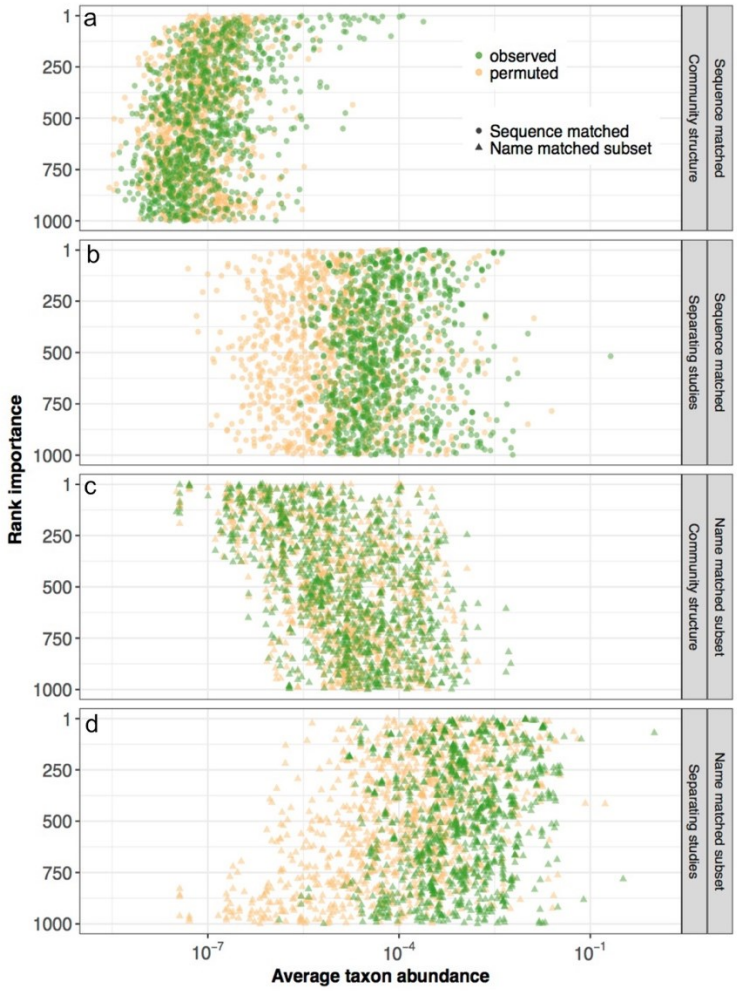


Supplementary Figure 3: a.) A supervised Random Forest model was fitted to predict pH from taxa and technical variables (in the same way as the supervised model separating studies described in the Methods). The importance of taxa and technical variables in this model is plotted against their importance 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

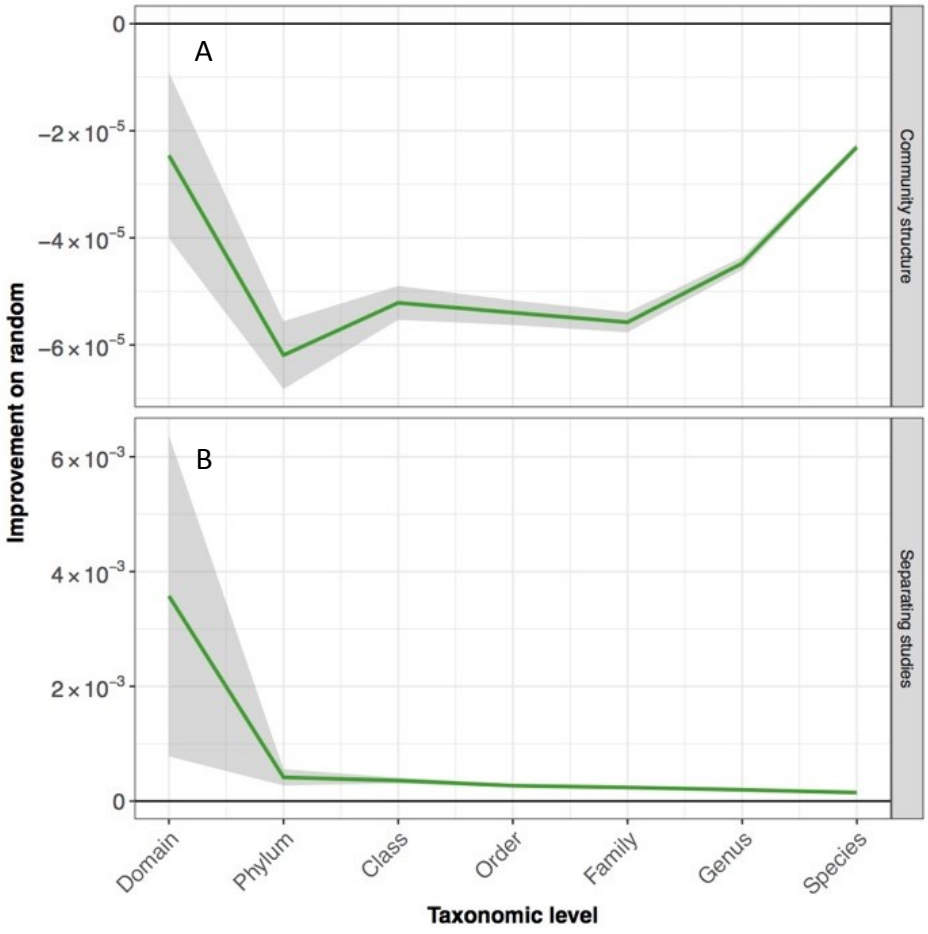
lower taxonomic rank tend to be detected in fewer studies (ρ = 0.3). Similarly, **c.)** low abundance taxa tend to be detected in fewer studies (ρ = 0.59). Finally, **d.)** the importance for separating studies given by the supervised Random Forest model correlates closely with the sensitivity component of the indicator value of a given taxon (ρ = 0.89). In b-d, darker colors indicate taxa more important in the model of community structure.



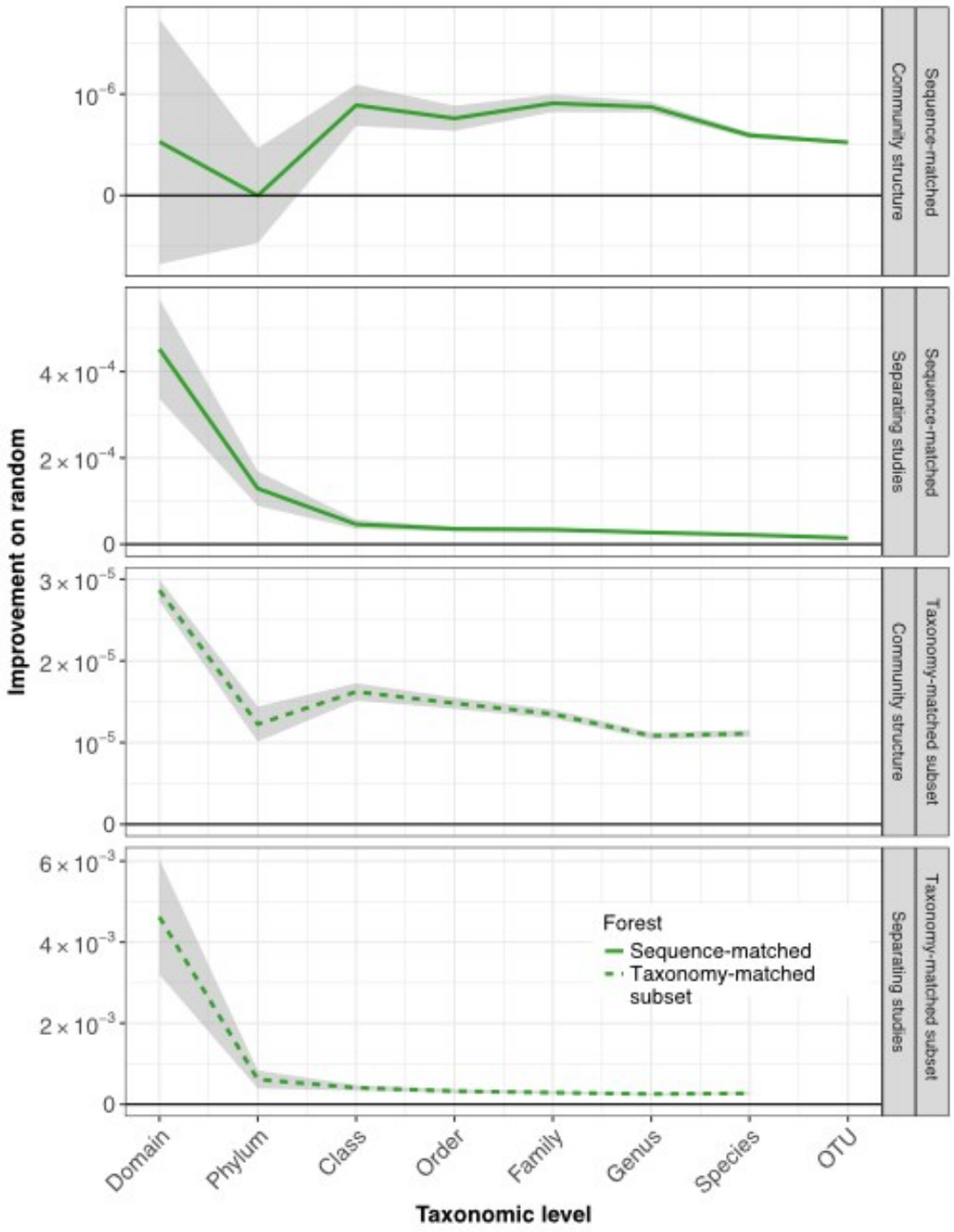
Supplementary Figure 4: Assessment of the community structure of two of the largest individual studies within the wider dataset: from Central Park, NYC encompassing 594 samples (study #24) (*top panels*) and a global dataset encompassing 103 samples (study #30) (*bottom panels*) demonstrates that there 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).



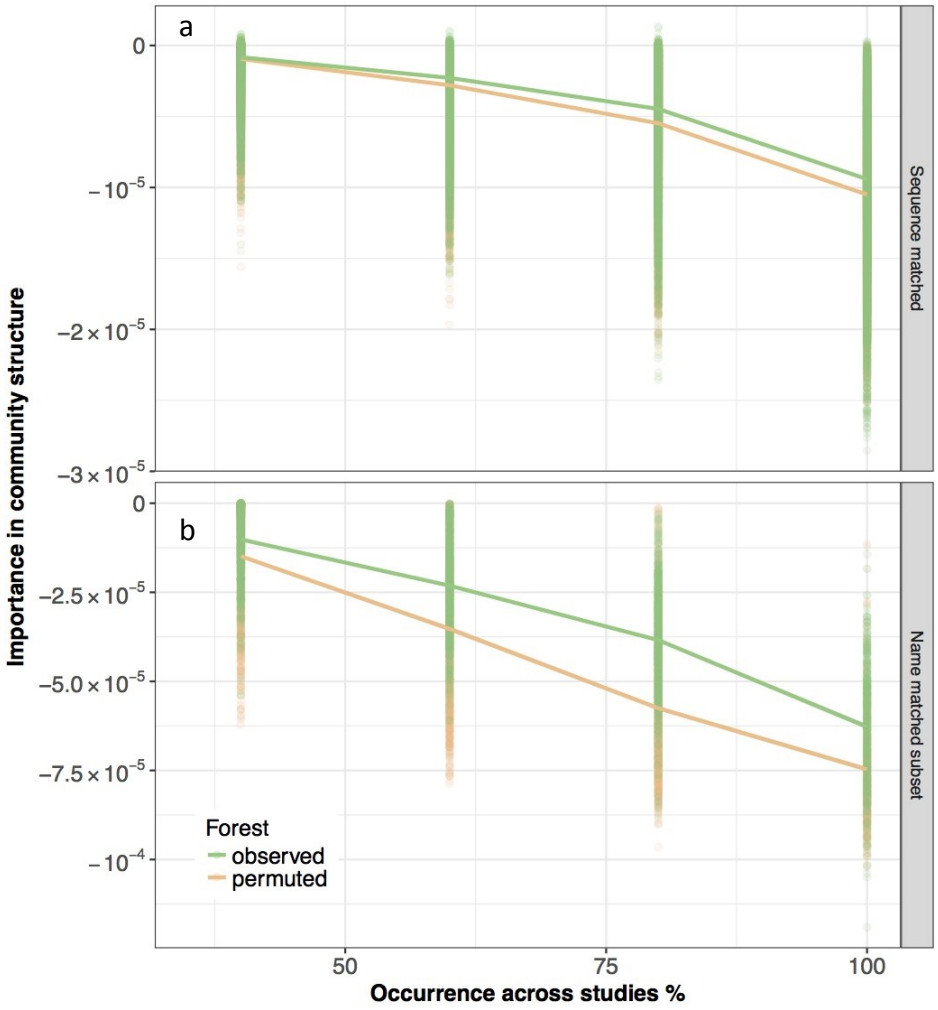
Supplementary Figure 6. The importance of bacterial taxa classified at different taxonomic ranks when considering only presence/absence data (i.e. without abundance information). While lower taxonomic resolution is more important for separating studies (b) it is still possible to conclude that there 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.



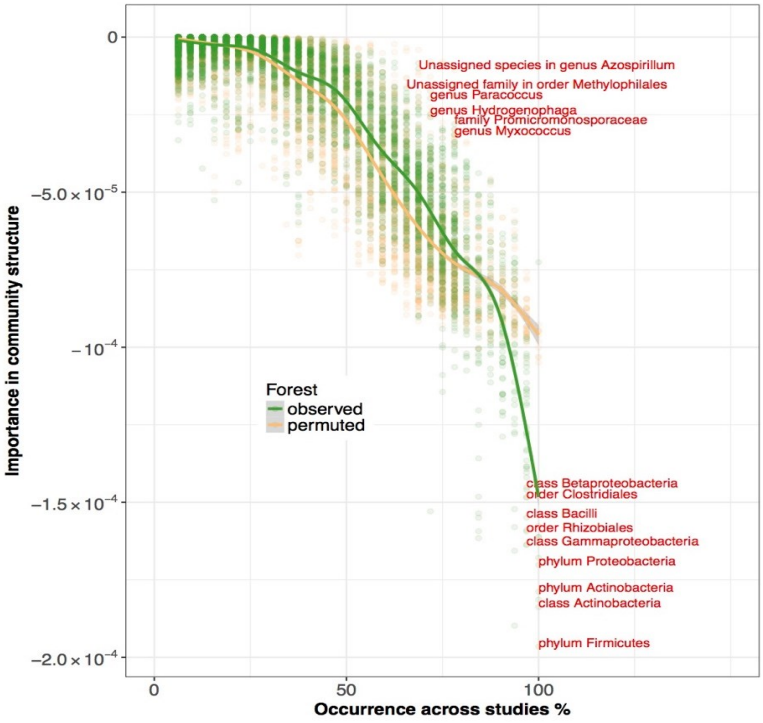
96

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

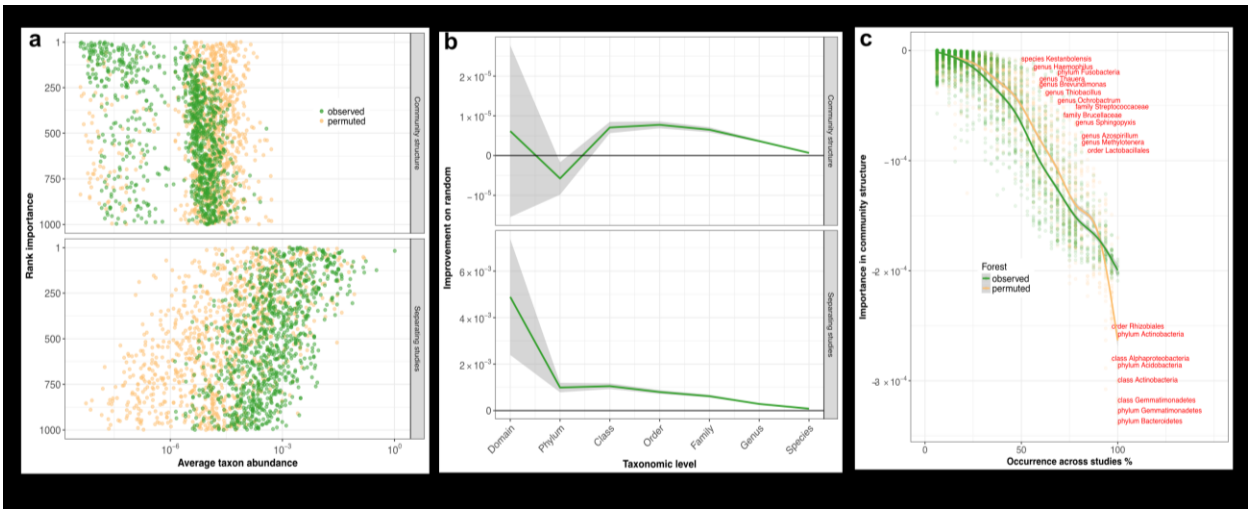
100



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).