SUPPLEMENTAL DATA

Metagenomic characterization of soil microbial communities in the Luquillo experimental forest (Puerto Rico) and implications for nitrogen cycling

Smruthi Karthikeyan[1], Luis H. Orellana[1], Eric R. Johnston[1], Janet K. Hatt[1], Frank Loeffler[3,4], Hector Ayala-del-Rio[4], Grizelle Gonzalez[5], Konstantinos T Konstantinidis[1,2]

Table S1: Physicochemical parameters of soils at the sampling sites

Description	Total_C	Total_N	Moisture	Class	pH	Chloride	NO_3-N	PO_4	Sulfate	NO_2-N	Chlorine	NH4-N
El Verde(0-5)	6.962	0.446	37.7%	Clay	4.42	35.49	< 0.190	< 1.0	6.64	< 0.210	1.50	15.37
El Verde(5-20)	3.698	0.246	33.8%	Clay	4.45	21.06	< 0.190	< 1.0	2.31	< 0.210	0.35	3.78
El Verde(20-30)	2.964	0.194	32.9%	Clay	4.50	14.48	< 0.190	< 1.0	< 0.850	< 0.210	0.30	2.39
Sabana 4(0-5)	3.960	0.281	33.1%	Clay	4.91	14.78	< 0.190	< 1.0	8.70	< 0.210	1.05	10.75
Sabana 4(5-20)	1.719	0.167	30.7%	Clay	4.72	11.68	< 0.190	< 1.0	3.43	< 0.210	0.40	2.39
Sabana 4(20-30)	0.980	0.081	29.2%	Clay	4.71	8.34	< 0.190	< 1.0	2.43	< 0.210	0.10	0.90
Palm Nido(0-5)	7.171	0.320	49.7%	Clay	4.86	71.2	0.388	< 1.0	11.34	< 0.210	0.33	7.12
Palm Nido(5-20)	6.604	0.280	48.9%	Clay Loam	4.97	82.2	< 0.190	< 1.0	10.08	< 0.210	0.15	5.52
Palm Nido(20-30)	5.633	0.255	44.7%	Clay Loam	4.93	10.54	< 0.190	< 1.0	< 0.850	< 0.210	0.15	3.26
Pico del Este (0-5)	12.091	0.428	58.7%	Loam	4.65	31.83	< 0.190	< 1.0	14.90	< 0.210	0.75	3.18
Pico del Este (5-20)	8.489	0.338	54.2%	Silty Clay Loam	4.67	59.84	< 0.190	< 1.0	6.46	< 0.210	0.55	1.16
Pico del Este (20-30)	3.001	0.131	43.8%	Clay	4.69	8.76	< 0.190	< 1.0	3.34	< 0.210	0.85	0.73

Total carbon and nitrogen are reported as % of dry weight; the remaining parameters measured are in mg/kg of soil at time of sampling (not dried).

Table S2: Summary of metagenome sample statistics

Metagenome ID	Location	Sampling depth	# Raw Reads	# Trimmed Reads	Avg. trimmed read length (bp)
A_13_1-34437631	El Verde	0-5cm	38,603,184	32,943,766	144.756
B_13_2-34462536	El Verde	5-20cm	38,969,952	33,517,824	143.034
C_13_3-34460523	El Verde	20-30cm	20,216,828	17,089,932	145.622
D_15_1-34460524	Sabana	0-5cm	34,703,794	28,887,538	145.49
E_15_2-3445656	Sabana	5-20cm	22,378,594	18,599,372	146.281
F_15_3-34451556	Sabana	20-30cm	39,659,410	34,311,032	142.323
G_16_1-34436613	Palm Nido	0-5cm	34,673,362	29,740,678	144.644
H_16_2-34437633	Palm Nido	5-20cm	51,852,508	45,179,536	143.475
I_16_3-34437634	Palm Nido	20-30cm	37,730,948	32,375,758	145.815
J_22_1-34456569	Pico del Este	0-5cm	29,380,762	24,855,902	146.252
K_22_2-34442602	Pico del Este	5-20cm	37,087,112	32,062,404	145.171
L_22_3-34452542	Pico del Este	20-30cm	37,532,494	32,453,186	144.221

Table S3: Key soil chemical parameters shaping the observed community diversity.

	NMDS1	NMDS2	r2	Pr(>r)	
Site	0.97015	0.24249	0.937	0.001	**Statistically significant**
Total_carbon	0.93321	0.35933	0.4486	0.078	
Total Nitrogen	0.83465	0.55078	0.0997	0.625	
pH	0.2074	0.97826	0.7203	0.003	**Statistically significant**
Sampling depth	-0.08845	-0.99608	0.1204	0.568	
Soil Moisture	0.97477	0.2232	0.8152	0.002	**Statistically significant**

Table S4: Distance-based redundancy analysis (dbRDA) of the impact of site/location and sampling depth on the microbial community diversity patterns observed among the sites.

	DF	Sum of Squares	F	Pr(>F)	Significance
Sampling depth	1	0.00597	0.9217	0.427	
Site	1	0.0469	7.2256	0.001	*** (0.001)
Sampling depth:Site	1	0.0045	0.6994	0.666	

Table S5: Assembly statistics for the co-assembled reads.

Site	N50	Sequences	Total Length
El Verde	1597	164848	240921112
Sabana	1352	123957	162083707
Palm Nido	1392	231188	307794698
Pico del Este	1547	174332	249155062

Table S6: Summary statistics for the MAGs recovered from each sample

Location	Completeness	Contamination
El Verde	99.14	2.04
El Verde	97.41	3.45
El Verde	77.9	1.72
Sabana	93.7	3.6
Palm Nido	84.7	4.0
Pico del Este	98.28	1.72

Only MAGs/bins >75% completeness and <5% contamination were used in downstream analyses.

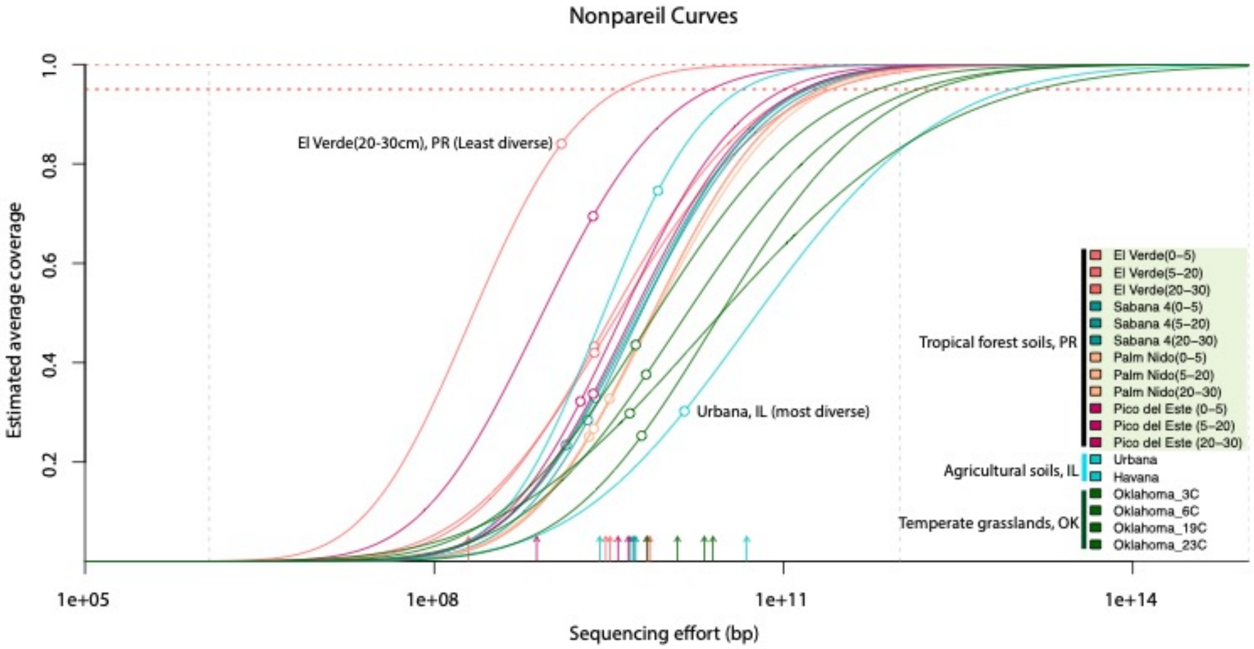


Fig. S1: Nonpareil sequencing coverage estimates of the soil microbial communities. Empty circles represent the estimated average coverage of the datasets obtained and projections based on model fitting to reach 95% and 99% coverage are indicated (horizontal dashed lines). Nonpareil curves representing the Puerto Rico tropical forest soil metagenomes (PR) as well those from Oklahoma temperate grasslands (OK) and Illinois agricultural soils (IL) are shown (see figure key). The arrows at the bottom represent sequencing effort required to achieve 50% coverage.

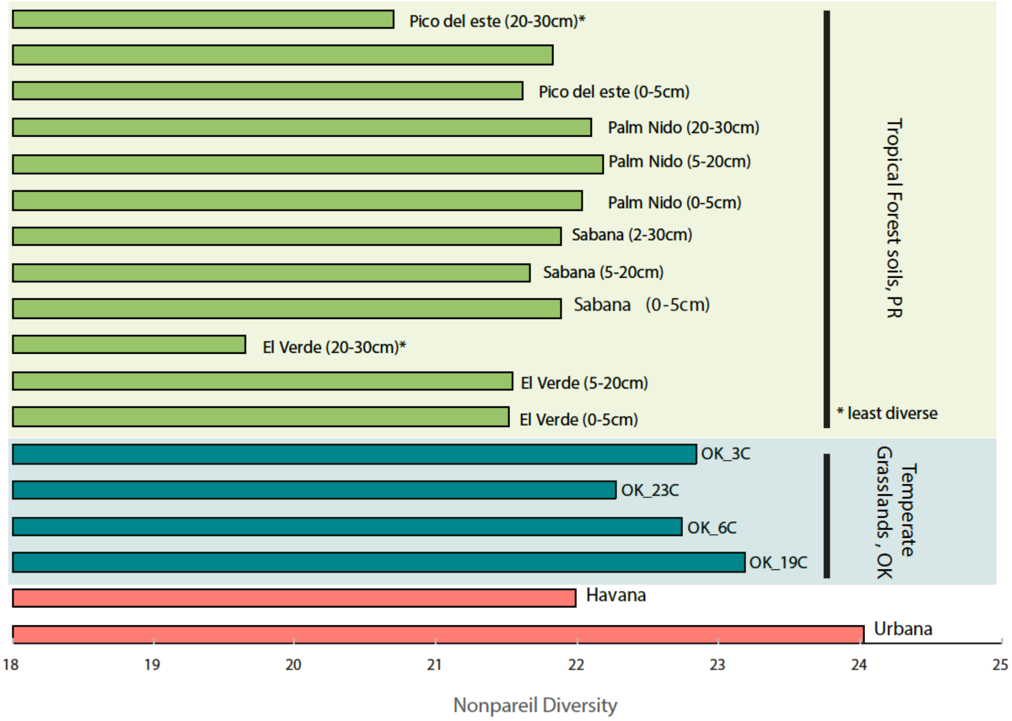


Fig. S2: A. Nonpareil diversity (N_d) values of PR, OK and IL metagenomic datasets. N_d represents a metric of α-diversity that takes into account both species evenness and richness as previously described (1).

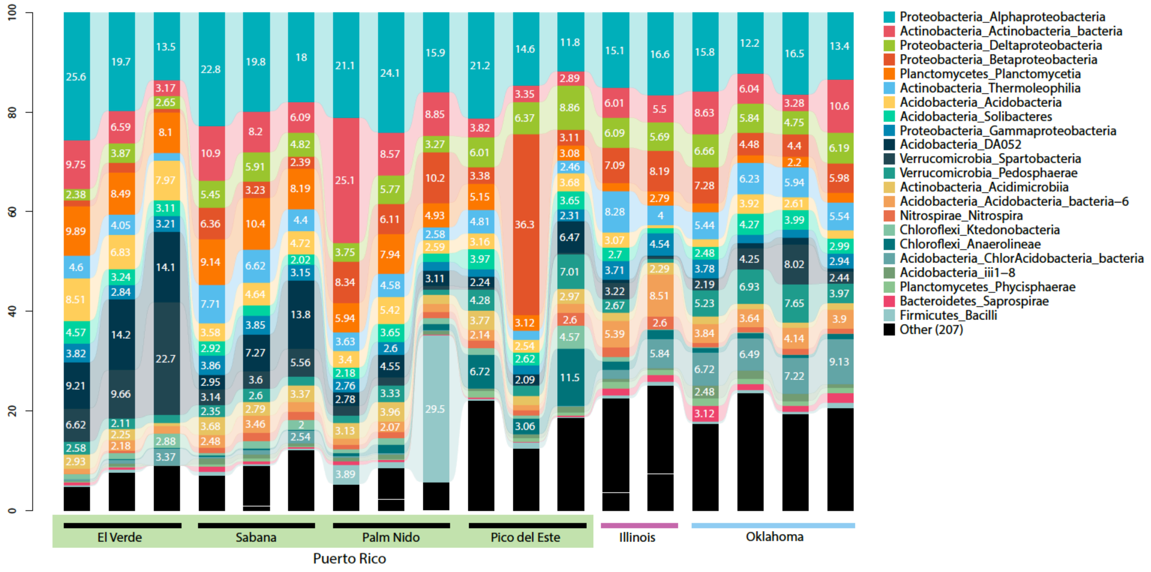


Fig. S2: B. Microbial community composition variation across the forest sites based on 16S rRNA gene fragments recovered in the metagenomes.

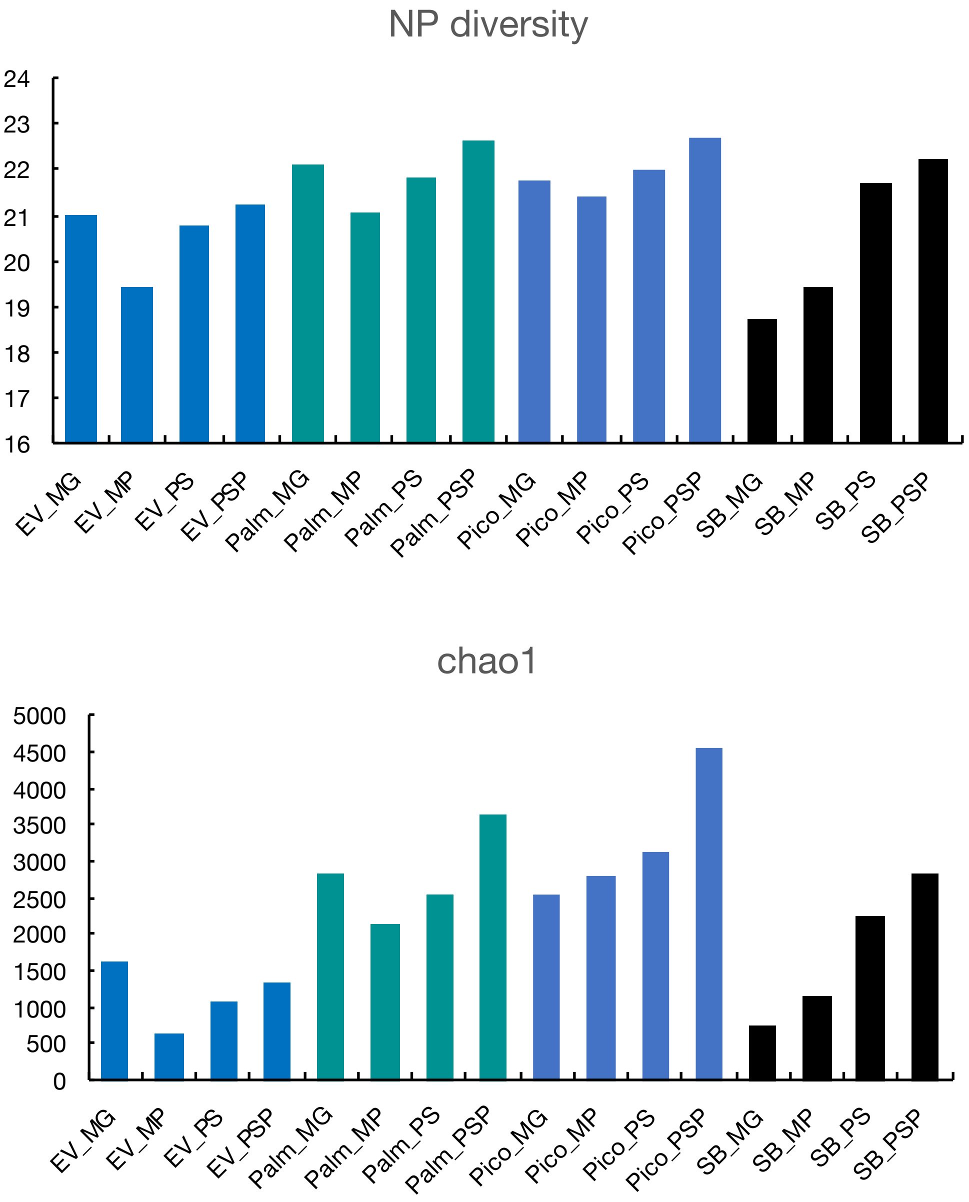


Fig S3. Comparison of alpha diversity estimates among four different DNA extraction methods. Upper Panel: Nonpareil diversity (N_d) estimates for the samples across the 4 sites and 4 different DNA extraction methods (N_d is given in log scale). Lower panel: Chao1 diversity estimates based on 16S rRNA gene-based OTUs for the samples across the 4 sites and 4 different DNA extraction methods. EV: El Verde, Palm: Palm Nido, Pico: Pico del Este, SB: Sabana, MG: Modified Griffith's protocol, MP: modified MP Bio FastDNA Spin kit protocol, PS: Qiagen PowerSoil kit, PSP: Qiagen PowerSoil Pro kit. Note: The MP method used in the samples reported in the main text provided similar diversity to other methods, especially for mid- and high-altitude samples, and similar trends across all samples in general. The Tsai-Olson extraction method (2) did not yield high quality DNA even after two column clean-up steps and hence, was not used in downstream analyses.

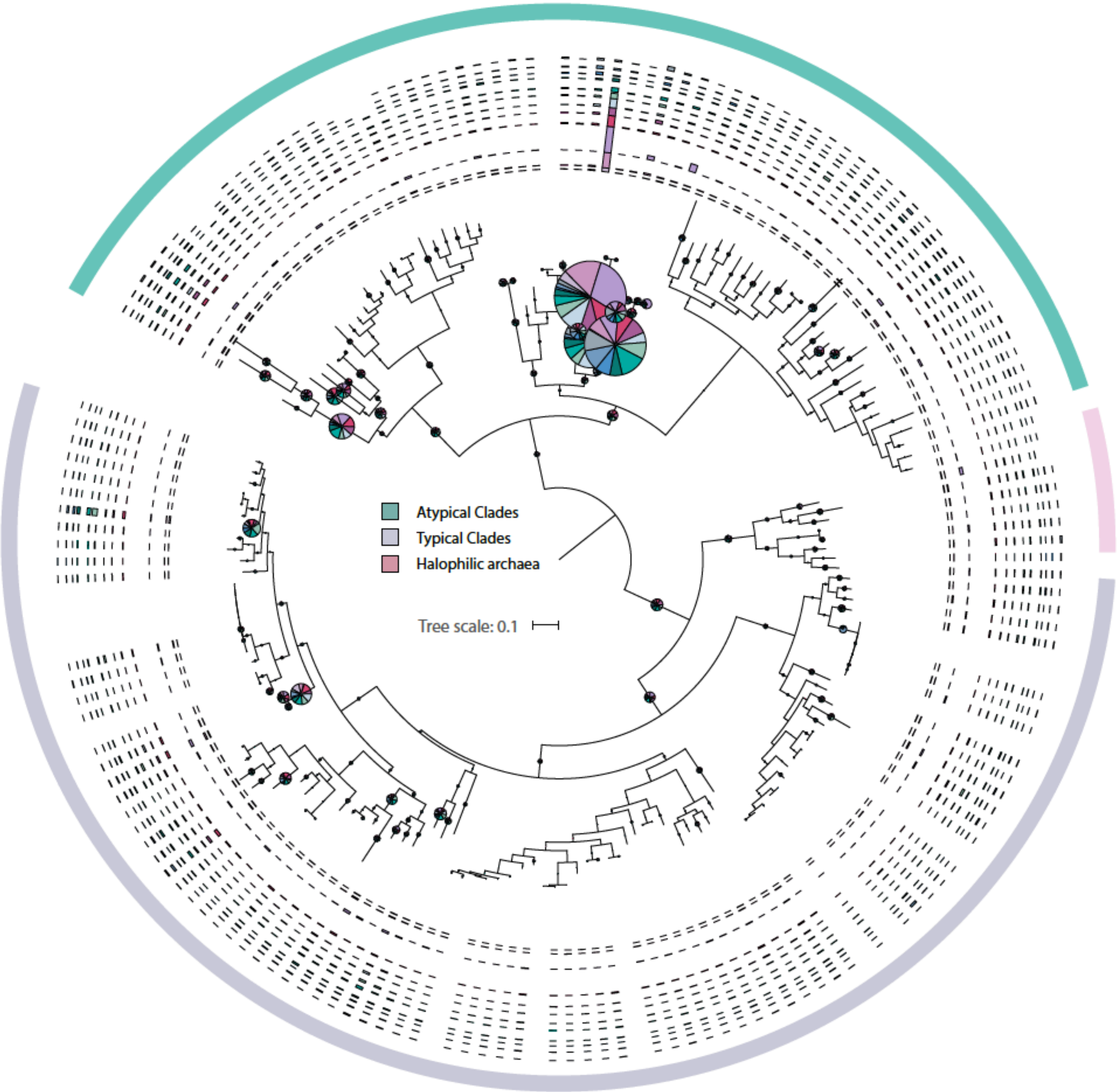


Fig. S4: ***nosZ*** **phylogeny for forest soils (PR).** Pie charts are proportional to the read placement and the bars represent the number of reads recruited by the corresponding subclades from each site (normalized by genome equivalents per single-copy gene; see Materials and Methods for further details). Scale bars are the same as shown in Fig. 3 of the main text.

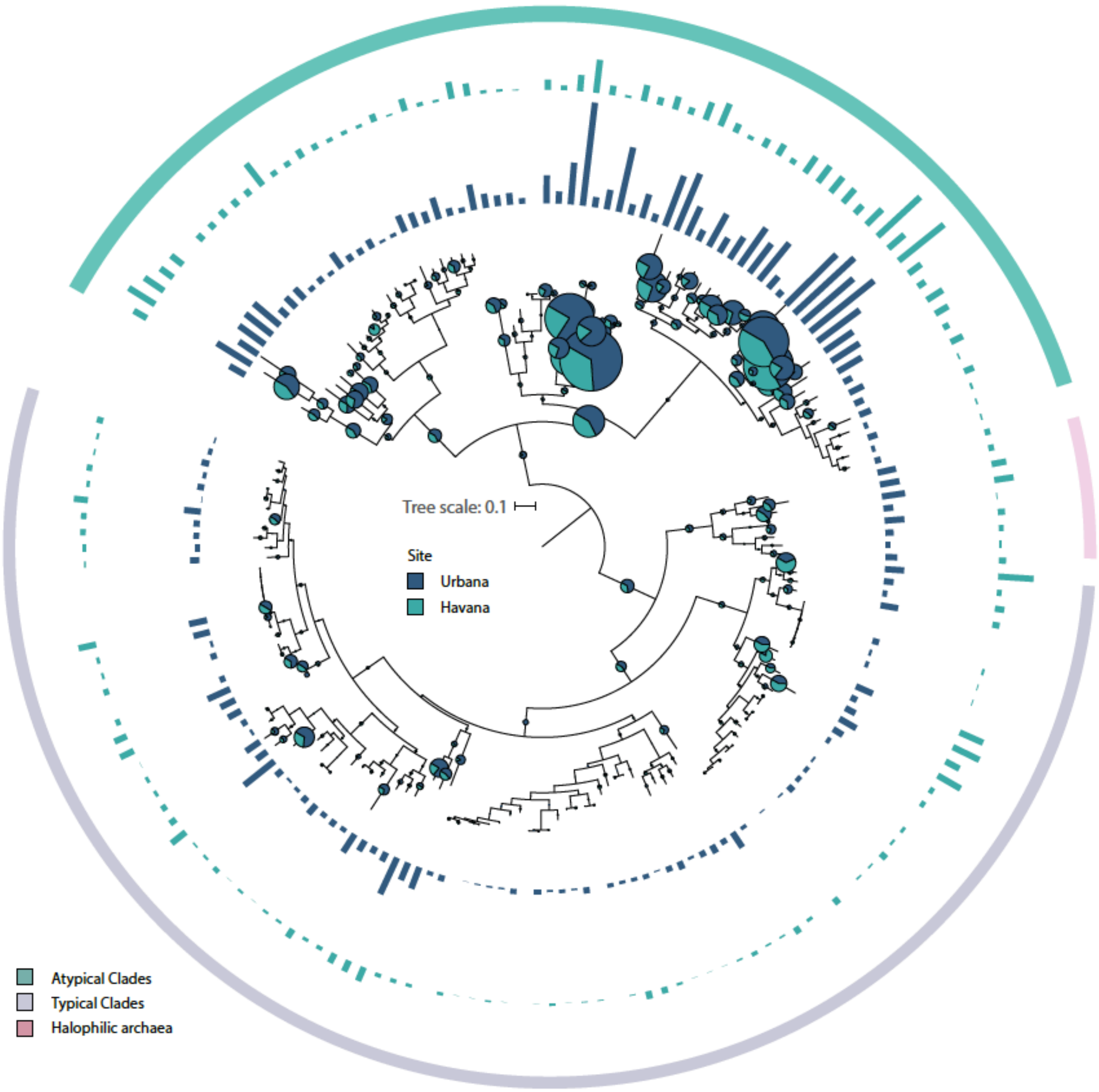


Fig. S5: *nosZ* phylogeny for agricultural soils (IL). Pie charts are proportional to the read placement and the bars represent the number of reads recruited by the corresponding subclades from each site (normalized by genome equivalents per single-copy gene; see Materials and Methods for further details). Scale bars are the same as shown in Fig. 3 of the main text.

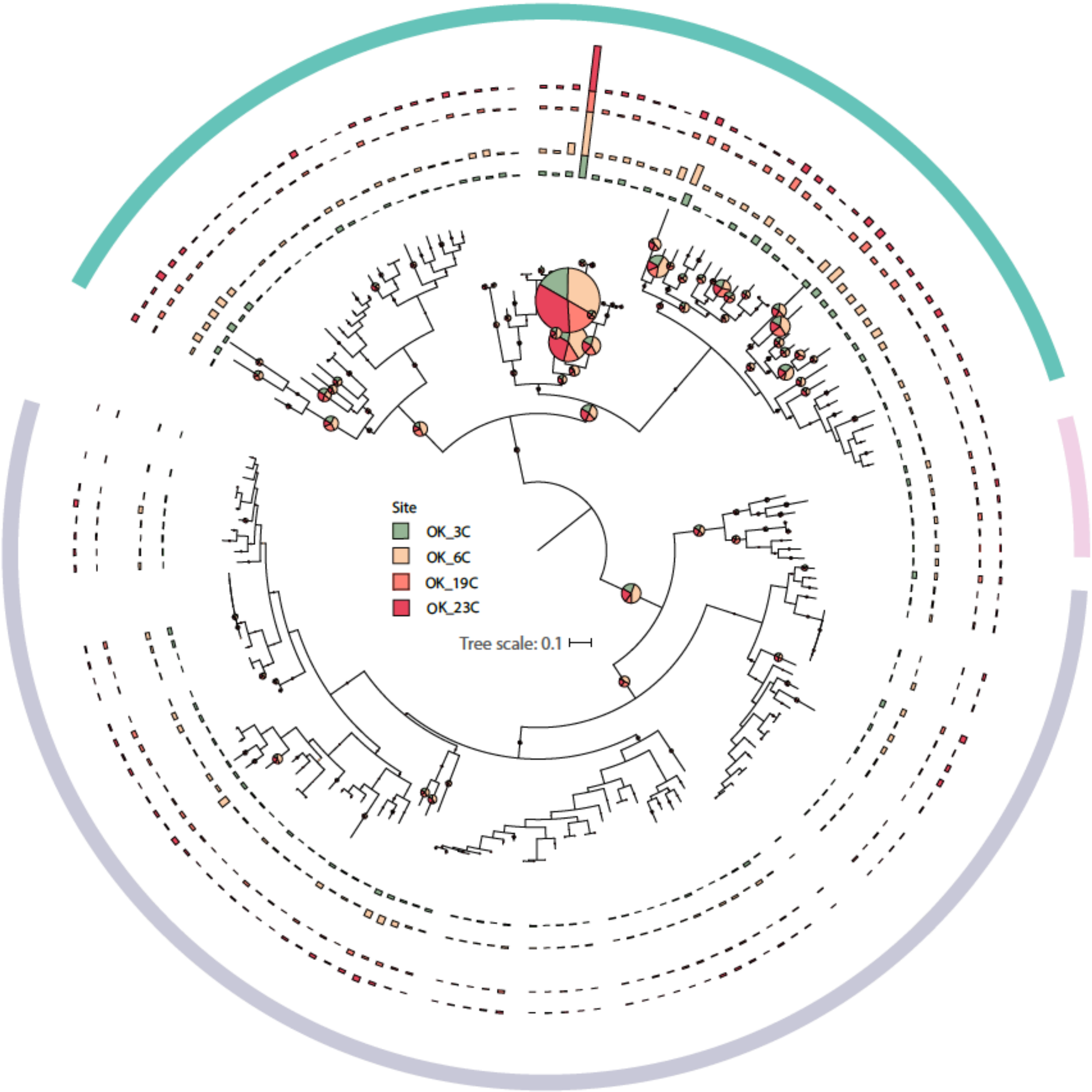


Fig. S6: nosZ phylogeny for restored grassland soils (OK). Pie charts are proportional to the read placement and the bars represent the number of reads recruited by the corresponding subclades from each site (normalized by genome equivalents per single-copy gene; see Materials and Methods for further details). Scale bars are the same as shown in Fig. 3 of the main text.

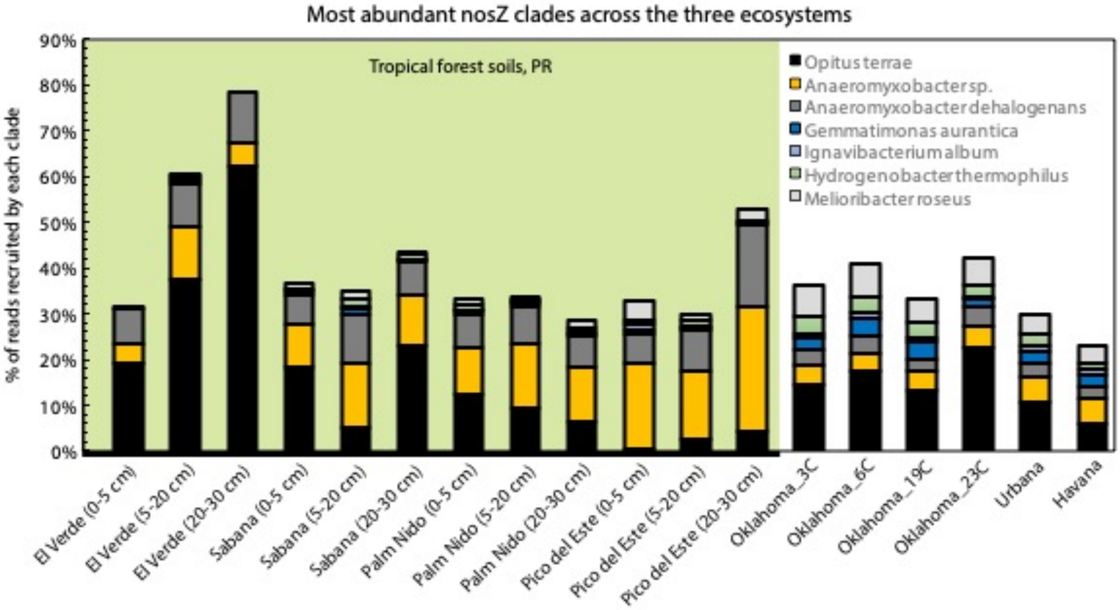
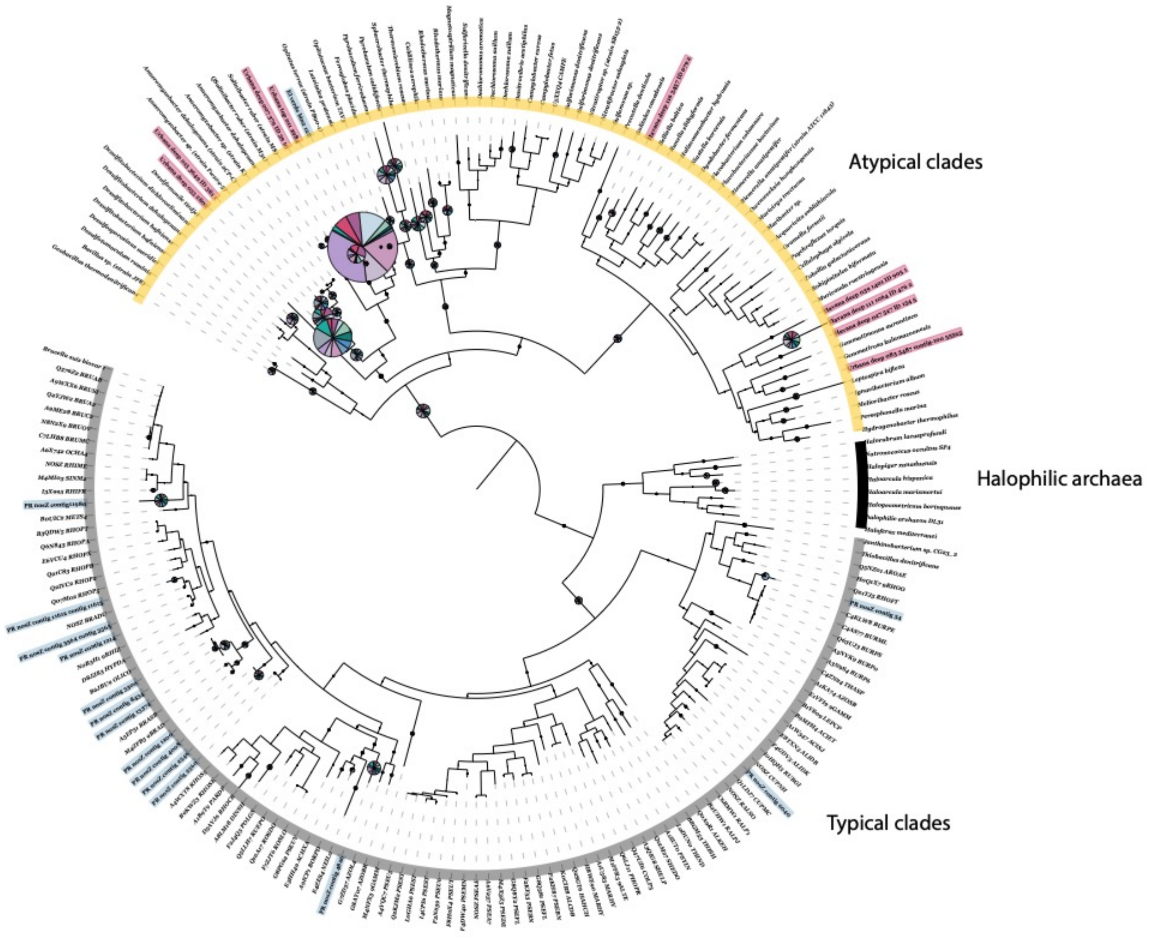


Fig. S7: Most abundant *nosZ*-encoding sub-clades and their distribution across the three ecosystems described in this study. The graph shows the relative abundances of *nosZ* OTUs defined at the 95% nucleotide sequence identity level.


Atypical clades
Halophilic archaea
Typical clades

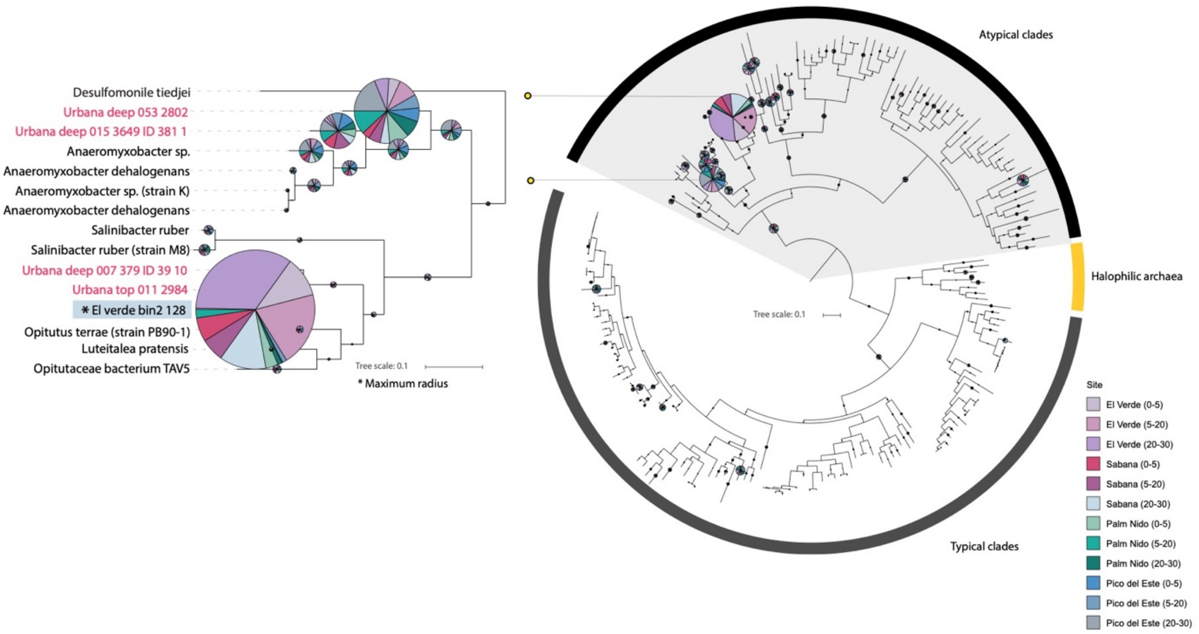

Desulfomonile tiedjei
Urbana deep 053 2802
Urbana deep 015 3649 ID 381 1
Anaeromyxobacter sp.
Anaeromyxobacter dehalogenans
Anaeromyxobacter sp. (strain K)
Anaeromyxobacter dehalogenans
Salinibacter ruber
Salinibacter ruber (strain M8)
Urbana deep 007 379 ID 39 10
Urbana top 011 2984
* El verde bin2 128
Opitutus terrae (strain PB90-1)
Luteitalea pratensis
Opitutaceae bacterium TAV5
Tree scale: 0.1
* Maximum radius
Atypical clades
Halophilic archaea
Typical clades
Tree scale: 0.1
Site
El Verde (0-5)
El Verde (5-20)
El Verde (20-30)
Sabana (0-5)
Sabana (5-20)
Sabana (20-30)
Palm Nido (0-5)
Palm Nido (5-20)
Palm Nido (20-30)
Pico del Este (0-5)
Pico del Este (5-20)
Pico del Este (20-30)

Fig. S8: *nosZ* phylogeny for forest soils (PR) with the updated reference tree including putative near full length *nosZ* sequences identified from assembled contigs/MAGs. Subclades highlighted with color in the top panel are sequences recovered form assemblies (i.e., not from isolate genomes). The lower panel shows an in-depth (zoomed) view of the Clade II (atypical) that recruited most of the short reads. Subclades highlighted as Urbana/Havana indicate sequences recovered from the IL soil assemblies. The sequence highlighted in blue was recovered from a PR soil MAG (closely relate to *Optitutus terrae*). Note that most of the short-reads are recruited by these subclades.

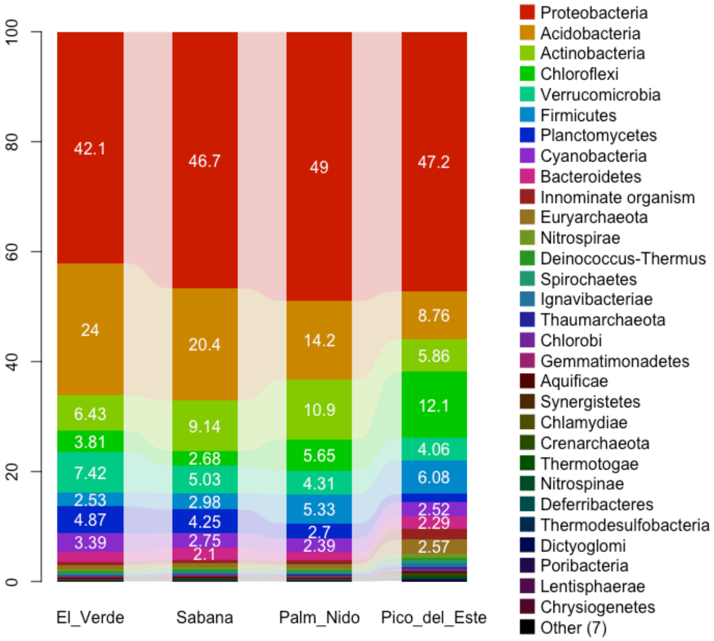


Fig S9. MyTaxa classification of the co-assembled reads showing the dominance of bacteria in the four forest sampling sites.

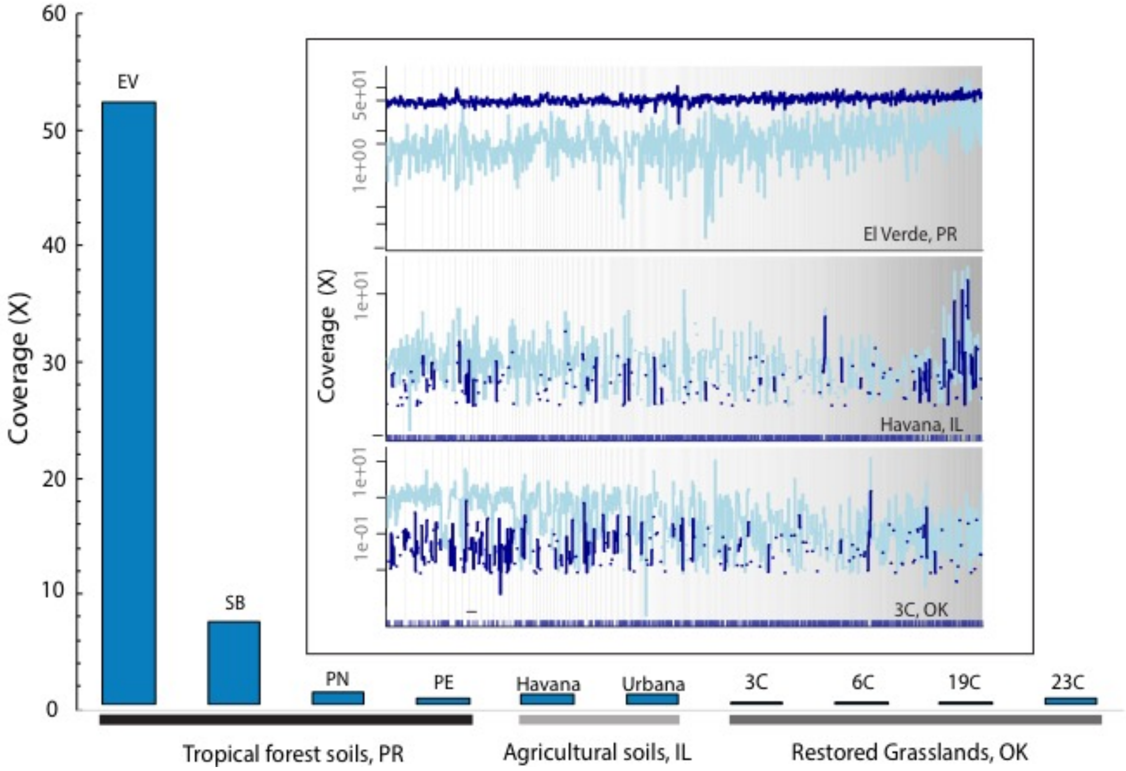


Fig S10. Abundance dynamics of population MAG 1 ***(Verrucomicrobia)*** **across the sampling sites**. Inset: Read recruitment plot showing the average coverage of the MAG across its genome, in 1,000bp windows, by the metagenomic reads from tropical (PR) and natural prairie (OK) sites (figure key). The dark blue histogram represents the coverage by reads matching the reference genome at ≥80bp in length and ≥95% nucleotide identity; light blue represents reads matching at <95% identity. The evenness of the coverage of the genome on the metagenomic datasets shows a sequence discrete population as described previously (3-4). Main graph shows the average coverage (single value) from the recruitment plots (y-axis) for each metagenome sample (x-axis)

References Cited

1. Rodriguez-R LM, Gunturu S, Tiedje JM, Cole JR, Konstantinidis KT. 2018. Nonpareil 3: fast estimation of metagenomic coverage and sequence diversity. mSystems 3:e00039-18. https://doi.org/10.1128/mSystems.00039-18.

2. Tsai YL, Olson BH. 1991. Rapid method for direct extraction of DNA from soil and sediments. Applied and Environmental Microbiology 57:1070-1074.

3. Caro-Quintero A, Konstantinidis KT. 2012. Bacterial species may exist, metagenomics reveal. Environ Microbiol 14:347-355.

4. Meziti, A., Rodriguez-R LM, Hatt JK, Peña-Gonzalez A, Levy K, Konstantinidis KT. How reliably do metagenome-assembled genomes (MAGs) represent natural populations? Insights from comparing MAGs against isolate genomes derived from the same fecal sample. Applied and Environmental Microbiology 87(6):e02593-20. doi: 10.1128/AEM.02593-20.