Supplementary Material for:

Title: Natural product biosynthetic gene diversity in geographically distinct soil microbiomes.

Authors: Boojala Vijay B. Reddy, Dimitris Kallifidas, Jeffrey H. Kim, Zachary Charlop-Powers, Zhiyang Feng and Sean F. Brady*

Author affiliation: Laboratory of Genetically Encoded Small Molecules, Howard Hughes Medical Institute, The Rockefeller University, 1230 York Avenue, New York, NY 10065.

***Corresponding Author:** Sean F. Brady

Contact: Laboratory of Genetically Encoded Small Molecules
The Rockefeller University
1230 York Avenue
New York, NY 10065
Phone: 212-327-8280
Fax: 212-327-8281
Email: sbrady@rockefeller.edu

Running Title: Biosynthetic gene diversity in soil microbiomes

Supplementary Table 1: Number of reads remaining after each 454 data processing step.

Processing Step	**AD**			**KS**			**KSα**		
Soil→	**AB**	**AZ**	**UT**	**AB**	**AZ**	**UT**	**AB**	**AZ**	**UT**
454 Reads	63,403	60,761	79,059	23,511	36,910	29,855	21,928	37,653	62,976
RDPP Trim	25,631	33,893	45,046	12,833	23,460	18,977	14,864	27,549	34,837
No Chimes	25,550	33,740	44,909	12,765	23,376	18,936	14,581	27,327	34,514
Non-redundant	21,880	30,190	38,433	7,233	14,573	10,840	6,992	13,640	21,302
Ref-Homologs	18,943	26,514	33,155	6,231	13,097	9,484	6,511	12,785	20,094
Ref-Hom-Redun	22,597	30,023	39,559	11,651	21,780	17,460	14,072	26,328	33,057
97% Unique	11,308	16,324	15,369	1,835	4,671	2,453	572	600	432

Supplementary Table 2: Number of 16S reads remaining after each 454 data processing step.

Processing Step	**Soil→**AB	AZ	UT
454 Reads	89,525	83,680	103,434
RDPP Trim	59,521	53,869	67,077
No Chimes	57,865	52,045	62,452
Non-redundant	23,191	18,930	22,054
Ref-Homologs	23,169	18,909	21,999
NoEcoli	21,239	15,111	20,875
Ref-Hom-Redun	28,081	28,617	49,637
97% Unique	3,103	3,491	5,237

Supplementary Table 3: 16S sequence classification into major phyla based on RDP classifier.

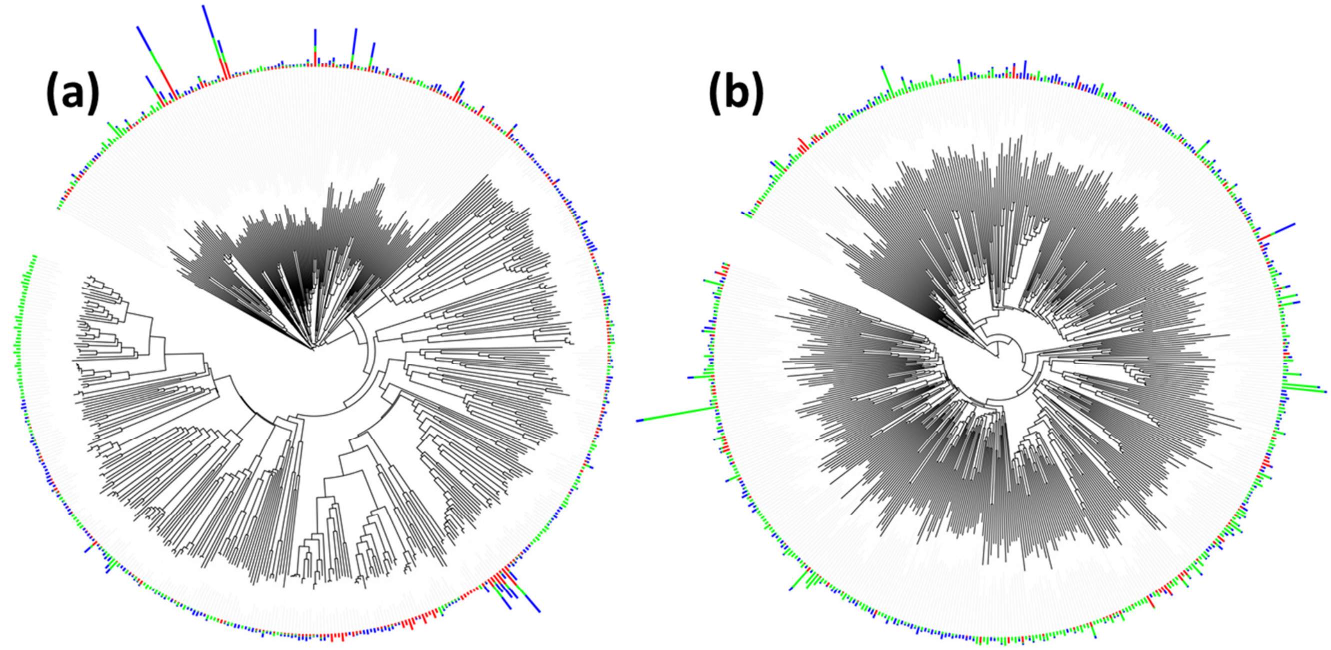
Phylum	**AB**	**AZ**	**UT**
Unclassified	545	626	1024
Proteobacteria	904	952	1421
Actinobacteria	778	564	480
Verrucomicrobia	153	649	934
Acidobacteria	139	285	487
Bacteroidetes	174	125	397
Planctomycetes	221	124	134
Gemmatimonadetes	117	128	248
Cyanobacteria	31	0	20
Firmicutes	17	14	32
Chloroflexi	9	3	7
Deinococcus-Thermus	5	0	2
BRC1	4	4	15
OD1	4	4	18
TM7	2	0	0
Chlamydiae	1	1	2
OP10	0	1	4
Nitrospira	0	3	8
WS3	0	1	1
Bacteria_incertae_sedis	0	0	1
Archaea	0	7	2
TOTAL	3103	3491	5237

Supplementary Table 4: Number of clades formed when sequences from each library are clustered at various percent identities.

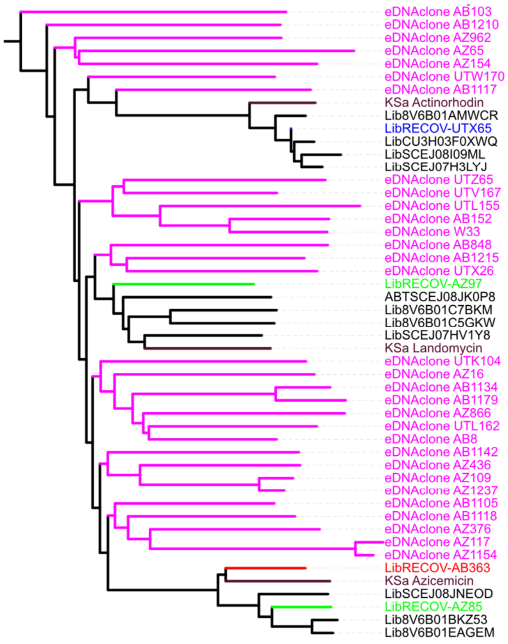
Gene or domain	Percent identity used for clustering	Total clusters from all libraries	Number of total clusters from individual libraries			Number of clusters with sequences from each designated category						
			AB	AZ	UT	AB only	AZ only	UT only	AB&AZ	AB&UT	AZ&UT	AB&AZ&UT
AD	97.00%	42,706	11,308	16,324	15,369	11,071	16,117	15,146	74	149	133	16
	95.00%	36,409	9,992	14,304	12,963	9,484	13,823	12,315	138	306	276	67
	90.00%	25,947	7,665	10,935	9,214	6,546	9,898	7,872	288	594	510	239
	85.00%	21,085	6,470	9,330	7,533	5,105	8,033	6,014	413	636	566	318
KS	97.00%	8,805	1,835	4,673	2,472	1,735	4,554	2,349	44	48	67	8
	95.00%	8,029	1,711	4,264	2,323	1,562	4,077	2,132	67	71	109	11
	90.00%	6,790	1,561	3,638	2,099	1,303	3,268	1,744	120	105	217	33
	85.00%	5,826	1,425	3,167	1,927	1,080	2,668	1,457	151	122	276	72
KSα	97.00%	1,518	572	610	454	472	531	407	61	29	8	10
	95.00%	1,148	456	508	328	341	411	261	68	38	20	9
	90.00%	753	331	384	237	191	254	140	71	38	28	31
	85.00%	515	236	294	186	110	163	89	56	22	27	48

Supplementary Table 5: Phylogenetic specificities of the sequences identified in the NCBI-NT database using primer-based search.

Kingdom/Phyla	Genera	AD	KS	KSα
Actinobacteria	Streptomyces	63	894	60
	Rhodococcus	32	0	1
	Amycolatopsis	30	33	1
	Actinoplanes	12	15	0
	Nocardia	3	56	0
	Mycobacterium	6	42	0
	Micromonospora	3	31	3
	Other Genera	57	90	12
Proteobacteria	Pseudomonas	30	3	0
	Burkholderia	29	0	0
	Polyangium	1	19	0
	Chondromyces	0	5	0
	Others Genera	17	2	0
Fungi	Aspergillus	0	12	0
	Other Genera	2	13	0
Uncultured		49	88	2
Total		334	1303	79



Supplementary Figure 1: Phylogenetic trees of (a) AD and (b) KS library sequences constructed using multiple sequence alignments of representative sequences found in OTUs generated when the *97% Unique* data sets were clustered at 85% sequence identity. The color bars show relative number of sequences from each sample library, AB (red), AZ (green) and UT (blue). Only representative sequences from the 500 most populated OTUs are shown.



Supplementary Figure 2: Phylogenetic tree of eDNA derived KSα sequences having less than 85% sequence identity to any functionally characterized sequences along with the sequences shown in Fig 5B.