nature microbiology

SUPPLEMENTARY INFORMATION

ARTICLE NUMBER: 16198 | DOI: 10.1038/NMICROBIOL.2016.198

Genome reduction in an abundant and ubiquitous soil bacterium, '*Candidatus* Udaeobacter copiosus'

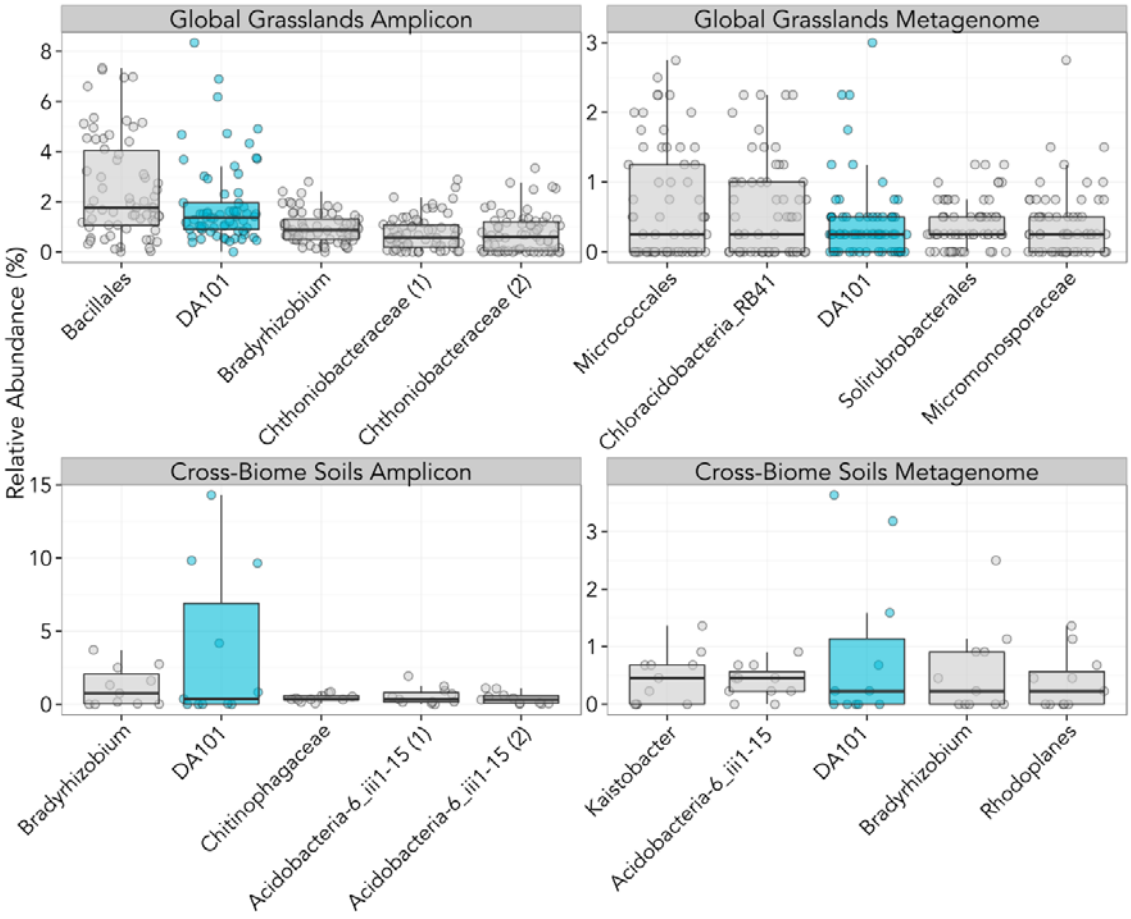
Tess E Brewer[1, 2], Kim M Handley[3], Paul Carini[1], Jack A Gilbert[4, 5], Noah Fierer[1, 6, *]

[1]Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO 80309; [2]Department of Molecular, Cellular, and Developmental Biology, University of Colorado, Boulder, CO 80309; [3]School of Biological Sciences, The University of Auckland, Auckland 1142, New Zealand; [4]Department of Ecology and Evolution, The University of Chicago, Chicago, IL 60637; [5]Argonne National Laboratory, Institute for Genomic and Systems Biology, Argonne, IL 60439; [6]Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, CO 80309

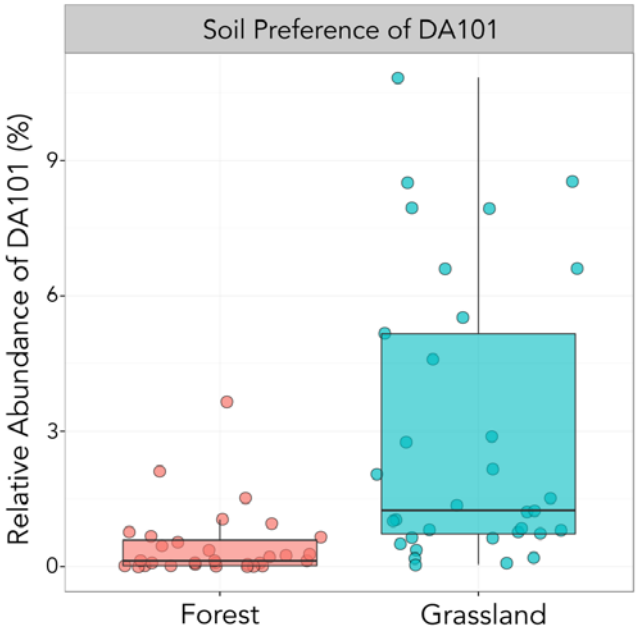
Corresponding author:
Noah Fierer
noah.fierer@colorado.edu

Supplementary figures 1-5 and supplementary tables 1-4

DOI: 10.1038/NMICROBIOL.2016.198

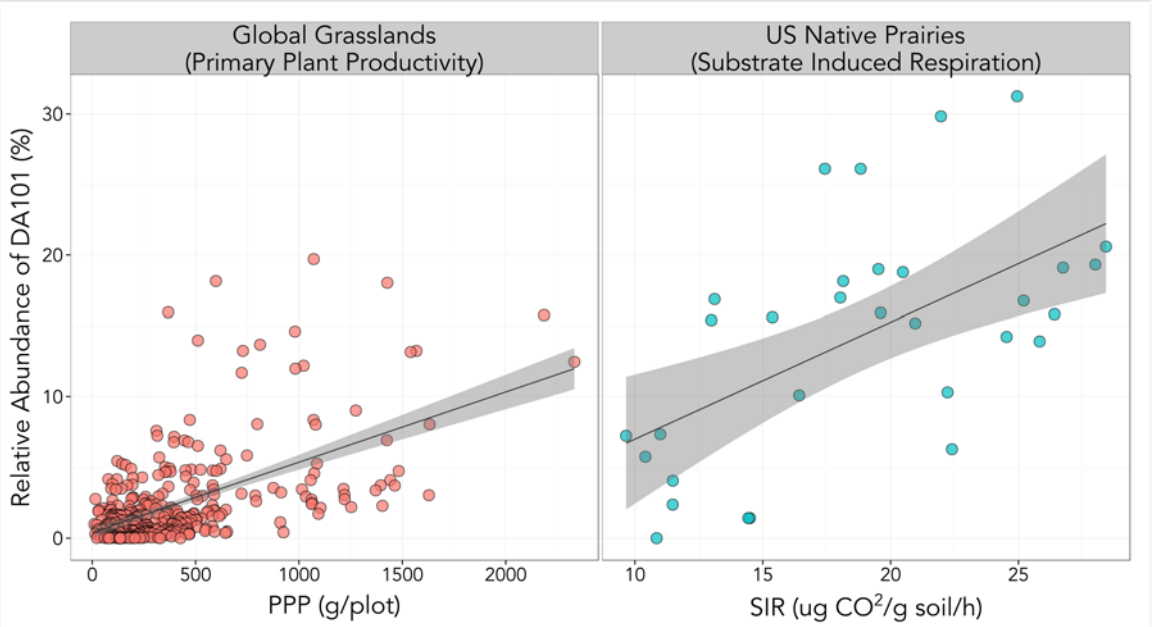


Supplementary Figure 1: DA101 rank is similar in amplicon and metagenomic data. The top 5 phylotypes from two matched amplicon and metagenomic datasets (Global Grasslands = Leff et al. 2015[1], Cross-Biome Soils = Fierer et al. 2012[2]) are shown in order of decreasing median rank. Each point represents one sample within the corresponding dataset (Not all samples in the global grasslands dataset had metagenomic sequencing). DA101's position is highlighted with blue while all other phylotypes are grey. The top and bottom of each box represents the 25th and 75th percentile, the mid line represents the 50th percentile/median, and the whiskers represent the range of points excluding outliers. Further details on these studies are provided in Supplementary Table 1.

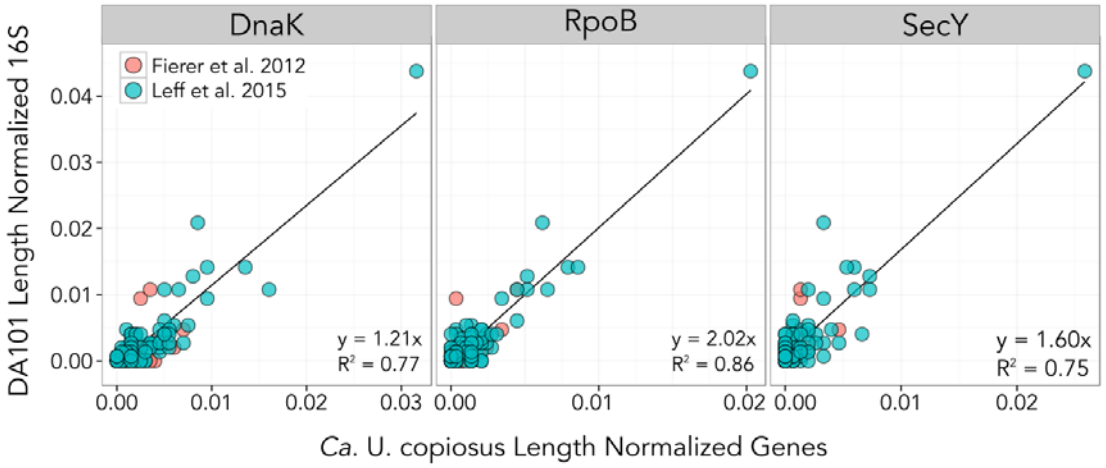


Supplementary Figure 2: Phylotype DA101 is more abundant in grasslands than forests. ($p<0.0001$, $n=64$, Mann-Whitney test) The top and bottom of each box represents the 25th and 75th percentile, the mid line represents the 50th percentile/median, and the whiskers represent the range of points excluding outliers. Data is from Crowther et al. 2014[3]. Further details on this study are provided in Supplementary Table 1.

DOI: 10.1038/NMICROBIOL.2016.198

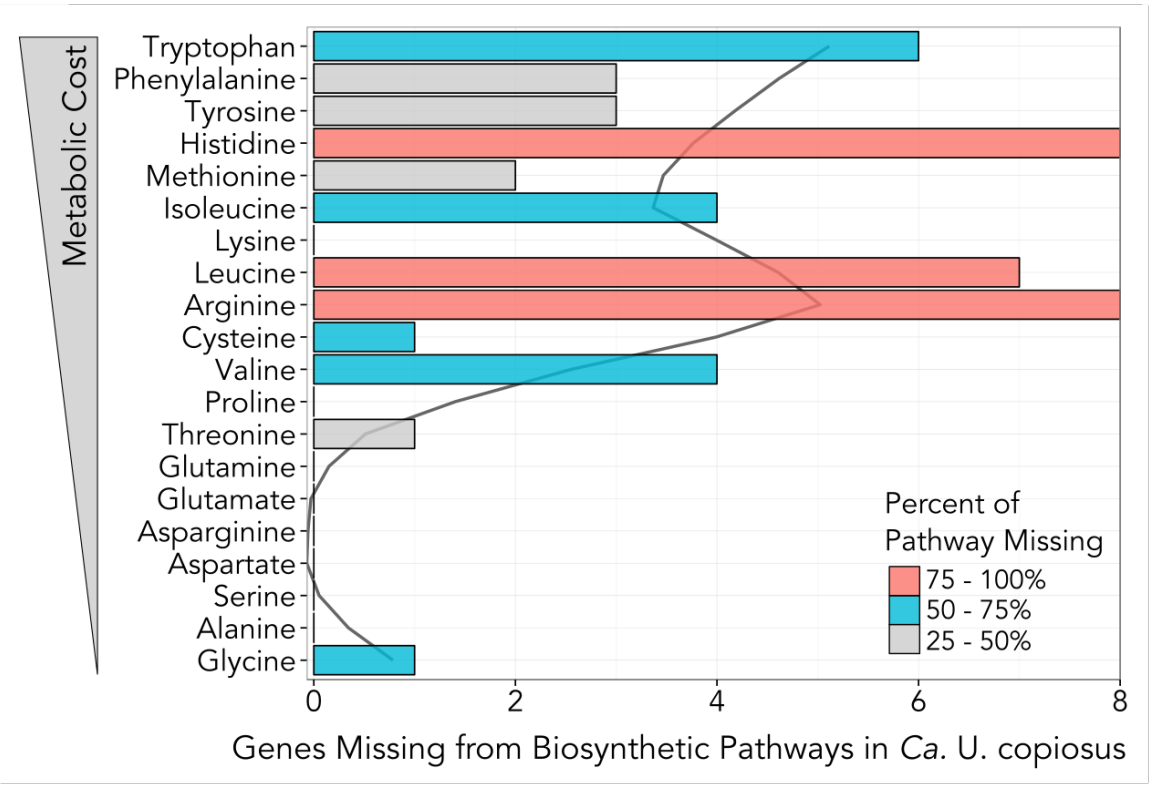


Supplementary Figure 3: The abundance of the DA101 amplicon correlates with measures of microbe and plant biomass. (Primary plant productivity $p<0.0001$ $\rho=0.47$ $n=366$, and substrate induced respiration $p<0.001$ $\rho=0.57$ $n=31$, Spearman correlations). The shaded region represents the 95% confidence interval of the trend line. Global Grasslands = Leff et al. 2015[1] and US Native Prairies = Fierer et al. 2013[4].



Supplementary Figure 4: The abundances of housekeeping gene sequences from the *Ca.* U. copiosus genome and DA101 16S rRNA gene sequences are well correlated across soil metagenomes ($P < 0.0001$, $\rho > 0.87$, n=102, Pearson correlation). We extracted matches to three *Ca*. U. copiosus housekeeping genes using blastN[5] and compared the length-normalized abundance of these fragments to the length-normalized abundance of DA101 16S rRNA gene sequences extracted using Metaxa2. We counted genes as a match to *Ca*. U. copiosus if the percent identity was greater than 85% and *Ca*. U. copiosus was the best hit. Our blastN database included the corresponding housekeeping genes from all named verrucomicrobial genomes in IMG. We chose 85% identity as our cutoff for several reasons: *i*) Protein coding genes are inherently more variable than rRNA genes; *ii*) the intraspecies percent identity variation for these genes has been reported to be as low as 87.7%[6]; *iii*) there are no other representatives of this genus with a sequenced genome to permit direct comparisons.

DOI: 10.1038/NMICROBIOL.2016.198



Supplementary Figure 5: Pathways to synthesize several expensive amino acids are underrepresented in the *Ca.* U. copiosus genome. 34 unique genes are currently missing from the *Ca.* U. copiosus genome that would enable to synthesis of all 20 amino acids. The cost of each of amino acid was estimated in *E. coli* by number of high-energy phosphate bonds hydrolyzed[7]. The number of genes missing in each pathway was calculated from KEGG metabolic pathways. The line simply represents the general trend through the metabolic cost gradient.

Supplementary Table 1: Descriptions of Datasets Used in this Study

	Fierer et al. 2012	Fierer et al. 2013	Crowther et al. 2014	Ramirez et al. 2014	Leff et al. 2015	Table Mountain, CO
REGION DESCRIPTION	Global cross-biome	U.S. native prairies	U.S. matched forest/grasslands	Central park, NYC	Global grasslands	CO Terrace
# SAMPLES	15	31	64	595	367	29
# SITES	11	31	11	1	25	1
METAGENOMIC DATA	15	-	-	-	87	-
ITS DATA	×	×	✓	✓	✓	×
18S DATA	×	×	×	✓	×	×
pH RANGE	4.1-9.5	5.8-7.9	4.0-8.1	3.9-8.4	4.4-8.2	-
MAT RANGE (°C)	-19-25	3.7-18.9	-3.2-22.8	13	0.3-18.4	11
MAP RANGE (mm)	100-4000	503-1148	287-3460	1016	262-1898	525
RAREFACTION DEPTH	15000	942	4000	5000	18000	11000

Supplementary Table 2: Samples used to construct EMIRGE phylogeny

SAMPLE	DATASET	LOCATION	LATITUDE	LONGITUDE	DESCRIPTION
NTP21	Fierer et al. 2013	Hayden, IA	43.26	-92.23	Native prairie
NTP28	Fierer et al. 2013	Glynn Prairie, MN	44.15	-95.41	Native prairie
NN1182	Leff et al. 2015	Val Mustair, Switzerland	46.63	10.37	Alpine grassland
NN772	Leff et al. 2015	Msunduzi Municipality, South Africa	-29.67	30.40	Mesic grassland
TM25	New data set	Table Mountain, CO	40.01	-105.5	Grassland terrace
GG14	New data set	Gordon Gulch, CO	40.12	-105.2	Meadow

DOI: 10.1038/NMICROBIOL.2016.198

Supplementary Table 3: '*Candidatus* Udaeobacter copiosus' genome has 34/36 single copy housekeeping genes

IMG GENE ID	COG/PFAM PRODUCT NAME	SEQUENCE LENGTH (BP)	COG/PFAM
2653240560	arginyl-tRNA synthetase	1737	COG0018
2653239845	DNA-directed RNA polymerase subunit alpha	1044	COG0202
2653239819	DNA-directed RNA polymerase subunit beta	2910	COG0085
2653240331	histidyl-tRNA synthetase	1239	COG0124
2653239579	Ribosome-binding ATPase GTP1/OBG family	1125	COG0012
2653241119	isoleucyl-tRNA synthetase	2748	COG0060
2653239815	large subunit ribosomal protein L1	702	COG0081
2653239810	large subunit ribosomal protein L11	429	COG0080
2653239936	large subunit ribosomal protein L13	438	COG0102
2653241207	large subunit ribosomal protein L14	366	COG0093
2653241203	large subunit ribosomal protein L16	426	COG0197
2653241213	large subunit ribosomal protein L18	375	COG0256
2653241195	large subunit ribosomal protein L3	726	COG0087
2653241210	large subunit ribosomal protein L5	573	COG0094
2653241212	large subunit ribosomal protein L6	540	COG0097
2653239619	leucyl-tRNA synthetase	2412	COG0495
2653242017	N6-L-threonylcarbamoyladenine synthase	1086	COG0533
2653241883	phenylalanyl-tRNA synthetase alpha chain	282	pfam01409
2653241216	preprotein translocase subunit SecY	1509	COG0201
2653241215	ribosomal protein L15	636	COG0200
2653241201	ribosomal protein L22	480	COG0091
2653241222	small subunit ribosomal protein S11	630	COG0100
2653241191	small subunit ribosomal protein S12	435	COG0048
2653241221	small subunit ribosomal protein S13	396	COG0099
2653241789	small subunit ribosomal protein S15	177	pfam00312
2653241206	small subunit ribosomal protein S17	297	COG0186
2653239969	small subunit ribosomal protein S2	702	COG0052
2653241202	small subunit ribosomal protein S3	690	COG0092
2653241223	small subunit ribosomal protein S4	612	COG0522
2653241214	small subunit ribosomal protein S5	564	COG0098
2653241192	small subunit ribosomal protein S7	474	COG0049
2653241211	small subunit ribosomal protein S8	399	COG0096
2653239935	small subunit ribosomal protein S9	396	COG0103
2653241990	valyl-tRNA synthetase	816	pfam00133

Supplementary Table 4: '*Candidatus* Udaeobacter copiosus' has a small genome compared to other heterotrophic soil *Verrucomicrobia*

GENOME NAME	GENOME SIZE (ASSEMBLED)	PREDICTED FINAL GENOME SIZE	ESTIMATED COMPLETENESS	16S rRNA COPY #	IMG TAXON ID
Candidatus Udaeobacter copiosus	2.66	2.81	94.4	Likely 1	2651869889
Chthoniobacter flavus Ellin428	7.85	8.07	97.2	1	642791618
Opitutus terrae PB90-1	5.96	5.96	100.0	1	641522643
Pedosphaera parvula Ellin514	7.41	7.85	94.4	1	645058870

Supplemental references:

1. Leff, JW et al. Consistent responses of soil microbial communities to elevated nutrient inputs in grasslands across the globe. *Proc Natl Acad Sci USA* **112**, 10967-72 (2015).

2. Fierer N et al. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proc Natl Acad Sci USA* **109**, 21390–5 (2012).

3. Crowther, TW et al. Predicting the responsiveness of soil biodiversity to deforestation: a cross-biome study. *Glob Chang Biol* **20**, 2983-94 (2014).

4. Fierer N et al. Reconstructing the microbial diversity and function of pre-agricultural tallgrass prairie soils in the United States. *Science* **342**, 621-4 (2013).

5. Camacho C, et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

6. Lee I-M, Bottner-Parker KD, Zhao Y, Davis RE, Harrison NA. Phylogenetic analysis and delineation of phytoplasmas based on secY gene sequences. *Int J Syst Evol Microbiol* **60**, 2887-97 (2010).

7. Akashi H, and Gojobori T. Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. *Proc Natl Acad Sci USA* 99, 3695-3700 (2002).