Science Advances

www.advances.sciencemag.org/cgi/content/full/1/3/e1500183/DC1

Supplementary Materials for

The microbiome of uncontacted Amerindians

Jose C. Clemente, Erica C. Pehrsson, Martin J. Blaser, Kuldip Sandhu, Zhan Gao, Bin Wang, Magda Magris, Glida Hidalgo, Monica Contreras, Óscar Noya-Alarcón, Orlana Lander, Jeremy McDonald, Mike Cox, Jens Walter, Phaik Lyn Oh, Jean F. Ruiz, Selena Rodriguez, Nan Shen, Se Jin Song, Jessica Metcalf, Rob Knight, Gautam Dantas, M. Gloria Dominguez-Bello

Published 17 April 2015, *Sci. Adv.* **1**, e1500183 (2015)
DOI: 10.1126/sciadv.1500183

The PDF file includes:

Fig. S6. ClonalFrame analysis of *Escherichia coli* strains based on the sequence of seven housekeeping genes.
Fig. S7. PCoA plot of Yanomami, Malawian, and Guahibo fecal samples.
Fig. S8. Fecal microbiome of non-*Bacteroides* OTUs in Yanomami, Malawian, Guahibo, and U.S. samples.
Fig. S9. Bacterial diversity as a function of age and body site.
Fig. S10. Procrustes analyses of shotgun metagenomic and 16*S* rRNA data for different subsets of samples.
Fig. S11. Comparison of PICRUSt versus sequenced shotgun metagenomes of fecal and oral samples with at least 10,000 mapped reads.
Fig. S12. Bacterial genome assembly.
References (*72–76*)

Other Supplementary Material for this manuscript includes the following:
(available at www.advances.sciencemag.org/cgi/content/full/1/3/e1500183/DC1)

Table S7 (Microsoft Excel format). List of all taxa significantly associated with each population in fecal samples as determined by LEfSe analysis.
Table S8 (Microsoft Excel format). List of all taxa significantly associated with each population in oral samples as determined by LEfSe analysis.
Table S9 (Microsoft Excel format). List of all taxa significantly associated with each population in skin samples as determined by LEfSe analysis.

Supplementary Materials

16S rRNA sequences

Quality-filtered, demultiplexed sequences from fecal samples were combined with those from a previous study that included a US population, and two semi-acculturated non-Western populations [11]. All sequences were trimmed at 100bp to avoid biases due to read length. We performed open-reference OTU picking [72] at 97% identity to identify both characterized and novel taxa, and taxonomy was assigned using Greengenes version 13_8 [73]. Oral and skin samples were processed in a similar manner combined with data from the American Gut project instead. In fecal samples, we obtained 768,221 sequences for Yanomami subjects (69,838 ± 6,630 seqs/sample), 913,44 for Guahibo (13,238 ± 6,338), 820,776 for Malawians (15,199 ± 3,152), and 4,510,511 for US (17,281 ± 4,459). In oral samples, we obtained 1,820,607 sequences for Yanomami (67,429 ± 16,553) and 890,367 for US (21,716 ± 7,388). In skin samples, we obtained 1,516,780 sequences for Yanomami (54,170 ± 9,571) and 872,475 for US (18,563 ± 9,380). Since the oral and skin sites sampled in the Yanomami (buccal mucosa, volar forearm) were physically close but not identical to the sites sampled in US subjects (tongue, hand palm), we compared the distances per body site between the populations against those obtained from sub-body sites (i.e. buccal mucosa versus tongue, hand versus forearm) within populations. We found that the distances between populations for both oral and skin samples are significantly larger than distances between sub-body sites within a population (t-test, $p < 0.01$; **Fig. S2F**), indicating that our results are robust.

Bacterial diversity in feces and skin, but not in the mouth, was significantly higher in Yanomami than in US subjects (**Fig. 1A, 1E, 1I**). Noticeably, fecal diversity was higher than in the semi-transculturated Guahibo Amerindians (ANOVA and Tukey's HSD test $p < 0.001$; **Fig. 1A**). PCoA of UniFrac distances reveals that bacterial communities from each body site segregate Westernized people from agrarian and hunter-gatherer groups (**Figs. 1B, 1F, 1J, and S1**).

Metagenome prediction and pathway analysis

We used PICRUSt 1.0.0 [18] with default parameters to predict metagenomic content from the 16S rRNA data. KEGG Ortholog (KO) counts were thus assigned to each sample, and KEGG pathways annotated to each KO. STAMP [43] was then utilized in combination with ANOVA with the Tukey-Kramer post-hoc test and Bonferroni correction for multiple comparisons to detect pathways differentially abundant across populations. Effect sizes were measured using η^2 (eta-squared).

Shotgun metagenomic sequencing and pre-processing

We used the Illumina Nextera® kit to perform shotgun metagenomic sequencing of 22 samples (8 fecal, 14 oral). 1 μL of metagenomic DNA from each sample was used as starting material. The dual-index multiplexed library was sequenced on an Illumina NextSeq (2x150bp). Shotgun reads were demultiplexed, allowing two mismatches per index, and trimmed with Trimmomatic v.0.30 [74] in paired-end palindrome mode to remove Nextera adapter sequences and low quality (<Q13) bases from the ends (minimum length 36 bases).

Computational filtering of assembled contigs from metagenomic libraries

From the sequenced Amerindian MDA NTCs, 2321 contigs assembled. By comparison to NCBI nt (August 15, 2013), all but five contigs aligned most closely to *Escherichia* species (including one in which *E. coli* was the second hit); the remaining contigs align most closely to cloning vectors, Taq polymerase, or rabbit hemoglobin, all of which are commonly used molecular biology tools, trace amounts of which may have been present in the MDA reagents.

Under our extremely stringent exclusion criteria (see Methods), most Amerindian metagenomic library-isolated genes from *E. coli* were excluded, as were many β-lactamases that aligned to a cloning vector on a NTC contig. The excluded contigs likely include some true AR genes present in the Amerindian microbiota, but we gave exclusion of false positives greater priority in this investigation due to the unique nature of the samples. We were also able to recover some of the true positives excluded by NTC-masking by separately interrogating cultured *E. coli* isolates from the Amerindian individuals.

Novel resistance gene with unknown function

The AR genes isolated from the Amerindian metagenomic library selection include a previously unrecognized gene conferring resistance to chloramphenicol, O23_CH_19:550-804. The 255bp ORF is unannotated, and the top blastn alignments in NCBI nt are *Neisseria lactamica* plasmids pNL14 and pNL871104, which have no annotation for that region (blastn: 98%, DQ229164.1 and DQ229166.1). The ORF is situated on a 4398bp contig (KJ910975) that contains a number of mobile genetic elements, including a~1.4kb region spanning an adjacent relaxase/mobilization domain and *mobC* that aligns with 95% nucleotide identity to *Neisseria gonorrhoeae* plasmids (FJ172221.1, CP003910.1). Although this gene is represented in NCBI, its resistance function was previously unrecognized.

Mobile genetic elements syntenic with antibiotic resistance genes

Of the AR genes isolated with functional selections, several are syntenic with mobile genetic elements, including six Amerindian genes (four isolated from metagenomic libraries and two from *E. coli* isolate libraries) and two Puerto Rican genes. O23_CH_19 is discussed above. The chloramphenicol acetyltransferases on F6_CH_2 are nearly identical and flank a plasmid mobilization element whose closest homolog in NCBI nr is Oscillibacter sp. KLE 1728 (WP_021751621.1; 74% local identity). They share 99-100% nucleotide identity with HMP stool assemblies and a plasmid from the fish pathogen *Aeromonas salmonicida* (GI:500229267). The AraC transcriptional regulators on F23_CH_11, Library_A_CH_8, and Library_B_CH_15 have no hits in NCBI nt, but blastx to NCBI nr indicates they are *ramA* homologs, and the top hit for all is *E. coli*. They are also adjacent to transposases that are 94-97% identical to *E. coli* (ref|WP_032270577.1|, ref|WP_000343728.1|, ref|WP_021499094.1|). Finally, the *tetX* variants on T1003_TE_12 and T1003_TG_1 are syntenic with integrases.

The ribosomal protection protein on F6_TE_1 is not adjacent to a mobile element, but is widespread in industrialized settings. The *tetW* variant aligns to HMP assemblies from stool, oral cavity, nasal cavity, skin, and vaginal sites and with ≥99% nucleotide identity to 32 species in NCBI, including gut commensals from two phyla and *Clostridium difficile* (**Fig. 3D**). *tetW* is also widespread in animal gut microbiota and the environment [75] and is frequently associated with mobile genetic elements. Future research could determine whether the Yanomami genes associated with mobile elements were ubiquitous prior to the antibiotic era, or whether the new antibiotic selective pressure drove their dissemination [76].

O23 CH 21 is also present in human pathogens, including *Haemophilus influenzae*, *Klebsiella pneumoniae*, and *Vibrio* spp. (**Fig. 3C**). The gene is often flanked by transposases that may contribute to its spread, although not in this context.

Alignment of functionally selected Amerindian and Puerto Rican AR genes

Seven of the Amerindian genes had homologs in fecal and oral metagenomic libraries from five Puerto Rican subjects, also identified by functional selections. Two pairs aligned with ≥95% nucleotide identity: the *tetW* variants from F6_TE_1 and 341_TE_1 and the *Neisseria* PBPs from O3_CZ_2_1/O3_CZ_2_2 and T1001_PITZ_5/T1001_CZ_11.

Alignment of functionally selected AR genes to the Human Microbiome Project and MetaHIT

All 108 AR genes were aligned to the HMP and MetaHIT assemblies. Excluding ribosomal protection proteins and tetracycline inactivation proteins, with a few exceptions, genes from the fecal microbiota align only to HMP stool assemblies and MetaHIT, and genes from the oral microbiota align only to the HMP oral assemblies. The exceptions are genes from the *E. coli* isolates and F5_PITZ_3:33-794 (100% amino acid identical to *ydeO* from *E. coli* str. K-12 substr. MG1655), which align to HMP subgingival plaque assemblies with >95% nucleotide ID, representing potential assembly contamination; oral-isolated ABC transporter 337_TE_6:2-1678, which aligns to HMP stool with only 75.5% nucleotide ID; and two β-lactamases from oral library 238_PE_3, whose top hits in NCBI nr are Enterobacteriaceae, that align to MetaHIT.

Table S1. Samples obtained from the 34 Amerindian subjects included in this study: skin from right (BD) or left (BI) forearm (n=28), oral (n=28) and fecal (n=12) specimens.

Subject ID	Age (y)	Gender	Sampled sites			
			Skin		ORAL	FECAL
			Left arm	Right arm		
24	4	m	BI24	BD24	O24	
20	7	f	BI20	BD20	O20	
23	7	f	BI23	BD23	O23	H23
21	8	f		BD21	O21	
1	10	m	BI1	BD1	O1	
3	11	m	BI3	BD3	O3	H3
11	11	m	BI11	BD11	O11	
9	12	f	BI9	BD9	O9	H9
2	12	m	BI2	BD2	O2	
19	13	m	BI19	BD19	O19	
15	14	f	BI15	BD15		
18	17	f	BI18	BD18	O18	
31	17	m				H31
29	18	m		BD29		
49	19	m	BI49	BD49	O49	H49
51	19	m	BI51	BD51	O51	
5	20	m	BI5		O5	H5
50	22	m	BI50	BD50	O50	H50
22	23	f	BI22	BD22	O22	
36	23	m				H36
26	30	m				H26
25	35	f	BI25	BD25	O25	
4	36	m	BI4	BD4	O4	
48	37	m				H48
41	39	m				H41
8	40	m	BI8	BD8	O8	
10	40	m	BI10	BD10	O10	
12	45	f	BI12	BD12	O12	
16	45	f	BI16	BD16	O16	
17	46	m	2 BI17	BD17	O17	
14	47	f	BI14	BD14	O14	
13	48	f	BI13	BD13	O13	
6	48	m	BI6		O6	H6
7	50	m	BI7	BD7	O7	
54	-	-	BI54	BD54	O54	

Table S2. Composition of Amerindian *E. coli* isolate genomic libraries.

	Subject IDs	**Strain IDs***	**Library size (GB)**
Library A	3, 5, 6, 23	15, 24, 30, 31, 55, 56, 122, 145, 188, 212, 217, 264, 293, 392, 393, 399, 402, 848, 849	4.50 – 4.95
Library B	9, 25, 31, 36, 48, 49, 50	335, 373, 378, 383, 408, 428, 429, 431, 451, 453, 478, 482, 502, 507, 525, 533, 564, 656, 1057	9.04
*Equal amounts of DNA were pooled from each isolate			

Table S3. Functional capture of antibiotic resistance genes from Amerindian *E. coli* isolate genomic libraries.

Antibiotic	**Library A** [†]	**Library B** [†]
Penicillin	ARG	ARG
Piperacillin*	ARG	ARG
Piperacillin-tazobactam*	ARG	ARG
Cefotaxime*	NO-ARG	NO-ARG
Ceftazidime	ARG	ARG
Cefepime*	NO-ARG	NO-ARG
Meropenem*	NO-ARG	NO-ARG
Aztreonam*	ARG	ARG
Chloramphenicol	ARG	ARG
Tetracycline*	ARG	ARG
Tigecycline	ARG	ARG
Gentamicin*	NO-ARG	NO-ARG
Ciprofloxacin*	NO-ARG	NO-ARG
Colistin	NO-ARG	NO-ARG

ARG = Resistance (*i.e.*, growth on selection plates in the absence of growth on negative control plates) was observed by functional screening of genomic library on corresponding antibiotic

NO-ARG = No resistance was observed by functional screening of genomic library on corresponding antibiotic

* All Amerindian *E. coli* isolates were screened for phenotypic resistance against these antibiotics, and were found to be susceptible in all cases.

[†] The composition of each library is defined in Table S2.

Table S4. Antibiotic resistance genes identified from functional selections of Amerindian *E. coli* isolate genomic libraries and Amerindian and Puerto Rican metagenomic libraries

Origin	Subject	Site	Contig	Accession	Gene coordinates	Antibiotic	Class	MG1655 gene	Length (bp)	Top hit by blastx to NCBI nr	%ID	Top hit accession
AE	A	Fecal	Library A_PE_9	KJ910913	2948-4114	PE	β-lactamase, Class C, *ampC*	ampC	1167	*Escherichia coli**	98.7	446841495
AE	B	Fecal	Library B_PE_6	KJ910933	2782-3948	PE	β-lactamase, Class C, *ampC*	ampC	1167	*Escherichia coli**	99.2	446841455
AE	A	Fecal	Library A_PI_1	KJ910915	445-1611	PI	β-lactamase, Class C, *ampC*	ampC	1167	*Escherichia coli**	99.2	446841495
AE	B	Fecal	Library B_PI_8	KJ910938	3221-3490	PI	β-lactamase	ampC	270	*Escherichia coli**	98.9	485961087
AE	B	Fecal	Library B_PI_8	KJ910938	3465-4388	PI	β-lactamase, Class C, *ampC*	ampC	924	*Escherichia coli**	99	446841495
AE	B	Fecal	Library B_PI_10	KJ910936	2960-3229	PI	β-lactamase	ampC	270	*Escherichia coli**	98.9	485961087
AE	B	Fecal	Library B_PI_10	KJ910936	3204-4127	PI	β-lactamase, Class C, *ampC*	ampC	924	*Escherichia coli**	99	446841495
AE	A	Fecal	Library A_PITZ_1	KJ910914	1989-2996	PITZ	β-lactamase, Class C, *ampC*	ampC	1008	*Escherichia coli**	99.7	485706785
AE	B	Fecal	Library B_PITZ_8	KJ910934	1924-3090	PITZ	β-lactamase, Class C, *ampC*	ampC	1167	*Escherichia coli**	99.5	446841455
AE	A	Fecal	Library A_CZ_1	KJ910910	22-1188	CZ	β-lactamase, Class C, *ampC*	ampC	1167	*Escherichia coli**	100	446841447
AE	B	Fecal	Library B_CZ_1	KJ910932	78-1244	CZ	β-lactamase, Class C, *ampC*	ampC	1167	*Escherichia coli**	99.5	446841447
AE	A	Fecal	Library A_AZ_1	KJ910904	54-860	AZ	β-lactamase	ampC	807	*Escherichia coli**	99.6	446841448
AE	A	Fecal	Library A_AZ_1	KJ910904	847-1221	AZ	β-lactamase	ampC	375	*Escherichia coli**	100	485961087
AE	B	Fecal	Library B_AZ_1	KJ910926	161-1294	AZ	β-lactamase, Class C, *ampC*	ampC	1134	*Escherichia coli**	100	485828396
AE	A	Fecal	Library A_CH_10	KJ910905	2957-4189	CH	MFS transporter	mdfA	1233	MULTISPECIES: Enterobacteriaceae	99.8 - 100	669336572
AE	B	Fecal	Library B_CH_4	KJ910931	3012-4244	CH	MFS transporter	mdfA	1233	*Escherichia coli* ISC7	99.5	571185026
AE	A	Fecal	Library A_TE_12	KJ910916	70-606	TE	MFS transporter *bcr/cflA*	bcr	537	*Shigella sonnei**	100	491278665
AE	A	Fecal	Library A_TE_12	KJ910916	623-1279	TE	MFS transporter *bcr/cflA*	bcr	657	*Escherichia coli**	99.5	559165244
AE	A	Fecal	Library A_TE_7	KJ910920	469-2250	TE	ABC transporter	mdlB	1782	*Escherichia coli**	99.5	487358840
AE	A	Fecal	Library A_TE_7	KJ910920	2243-4015	TE	ABC transporter	mdlA	1773	*Escherichia coli* 2-222-05_S1_C1	99.7	652048337
AE	A	Fecal	Library A_CH_4	KJ910907	935-1324	CH	AraC family transcriptional regulator, *marA*	marA	390	MULTISPECIES: Enterobacteriaceae*	100	445942659
AE	B	Fecal	Library B_CH_10	KJ910927	2188-2577	CH	AraC family transcriptional regulator, *marA*	marA	390	MULTISPECIES: Enterobacteriaceae*	100	445942659
AE	A	Fecal	Library A_PE_6	KJ910912	3-308	PE	AraC family transcriptional regulator	marA	306	MULTISPECIES: Enterobacteriaceae*	100	445942659
AE	A	Fecal	Library A_TE_4	KJ910917	1049-1438	TE	AraC family	marA	390	MULTISPECIES:	99.2	445942659

							transcriptional regulator, *marA*			Enterobacteriaceae*		
AE	B	Fecal	Library B_TE_7	KJ910941	1067-1456	TE	AraC family transcriptional regulator	marA	390	MULTISPECIES: Enterobacteriaceae*	96.9	445942659
AE	A	Fecal	Library A_TE_8	KJ910921	553-936	TE	AraC family transcriptional regulator, *marA*	marA	384	MULTISPECIES: Enterobacteriaceae*	99.2	445942659
AE	A	Fecal	Library A_TG_10	KJ910922	1093-1482	TG	AraC family transcriptional regulator, *marA*	marA	390	MULTISPECIES: Enterobacteriaceae*	99.2	445942659
AE	A	Fecal	Library A_CH_3	KJ910906	3520-4389	CH	AraC family transcriptional regulator	rob	870	MULTISPECIES: Enterobacteriaceae*	100	446293811
AE	B	Fecal	Library B_CH_20	KJ910930	157-1026	CH	AraC family transcriptional regulator	rob	870	MULTISPECIES: Enterobacteriaceae*	100	446293811
AE	A	Fecal	Library A_PE_2	KJ910911	631-1500	PE	AraC family transcriptional regulator	rob	870	MULTISPECIES: Enterobacteriaceae*	100	446293811
AE	A	Fecal	Library A_TE_6	KJ910919	2033-2902	TE	AraC family transcriptional regulator	rob	870	MULTISPECIES: Enterobacteriaceae*	100	446293811
AE	B	Fecal	Library B_TE_1	KJ910939	1855-2724	TE	AraC family transcriptional regulator	rob	870	MULTISPECIES: Enterobacteriaceae*	100	446293811
AE	A	Fecal	Library A_TG_6	KP063058	2116-2985	TG	AraC family transcriptional regulator	rob	870	MULTISPECIES: Enterobacteriaceae*	99.7	446293811
AE	A	Fecal	Library A_CH_6	KJ910908	476-799	CH	AraC family transcriptional regulator	soxS	324	MULTISPECIES: Enterobacteriaceae*	100	445941503
AE	B	Fecal	Library B_CH_17	KJ910929	139-462	CH	AraC family transcriptional regulator	soxS	324	*Escherichia coli*	99.1	585364975
AE	A	Fecal	Library A_TE_5	KJ910918	3-299	TE	AraC family transcriptional regulator	soxS	297	*Escherichia coli*	100	585364975
AE	B	Fecal	Library B_TE_6	KJ910940	1-198	TE	AraC family transcriptional regulator	soxS	198	*Shigella dysenteriae**	100	445941496
AE	A	Fecal	Library A_TG_1	KJ910923	2763-3086	TG	AraC family transcriptional regulator	soxS	324	*Escherichia coli*	99.1	585364975
AE	A	Fecal	Library A_TG_4	KJ910924	515-838	TG	AraC family transcriptional regulator	soxS	324	MULTISPECIES: Enterobacteriaceae*	99.1	445941503
AE	A	Fecal	Library A_CH_8	KJ910909	1691-2029	CH	AraC family	n/a	339	*Escherichia coli*	99.1	607731705

							transcriptional regulator			O121:H7 str. 2009C-3299		
AE	B	Fecal	Library B_CH_15	KJ910928	4-387	CH	AraC family transcriptional regulator	n/a	384	*Escherichia coli* O121:H7 str. 2009C-3299	91.5	607731705
AE	B	Fecal	Library B_PI_1	KJ910937	3-560	PI	AraC family transcriptional regulator	n/a	558	*Escherichia coli**	100	446161253
AE	B	Fecal	Library B_PITZ_9	KJ910935	1-690	PITZ	AraC family transcriptional regulator	n/a	690	*Escherichia coli**	100	446161253
AE	A	Fecal	Library A_TG_8	KJ910925	1-564	TG	AraC family transcriptional regulator	n/a	564	*Escherichia coli**	99.5	446161253
AM	5	Fecal	F5_PITZ_3	KJ910968	33-794	PITZ	AraC family transcriptional regulator	-	762	*Escherichia coli**	100	445982640
AM	6	Fecal	F6_PI_27	KJ910949	1276-2184	PI	β-lactamase, Class A, subclass *cblA*	-	909	*Bacteroides stercorirosoris*	95	647550201
AM	6	Fecal	F6_CT_1	KJ910947	1200-2108	CT	β-lactamase, Class A, subclass *cblA*	-	909	*Bacteroides stercorirosoris*	95.7	647550201
AM	6	Fecal	F6_CZ_1	KJ910948	1330-2238	CZ	β-lactamase, Class A, subclass *cblA*	-	909	*Bacteroides stercorirosoris*	94.7	647550201
AM	6	Fecal	F6_CP_1	KJ910946	308-1216	CP	β-lactamase, Class A, subclass *cblA*	-	909	*Bacteroides stercorirosoris*	95.7	647550201
AM	6	Fecal	F6_AZ_9	KJ910942	1352-2260	AZ	β-lactamase, Class A, subclass *cblA*	-	909	*Bacteroides stercorirosoris*	95.7	647550201
AM	3	Oral	O3_CZ_2_1	KJ910979	301-2049	CZ	Penicillin-binding protein	-	1749	*Neisseria mucosa**	100	647544871
AM	3	Oral	O3_CZ_2_2	KJ910980	301-2049	CZ	Penicillin-binding protein	-	1749	*Neisseria mucosa**	99.8	647544871
AM	5	Oral	O5_CZ_1	KJ910969	522-2270	CZ	Penicillin-binding protein	-	1749	*Neisseria perflava*	99.8	647544695
AM	5	Oral	O5_CZ_2	KJ910970	3-746	CZ	Penicillin-binding protein	-	744	*Neisseria mucosa*	99.6	647544835
AM	5	Oral	O5_CZ_4	KJ910971	287-2062	CZ	Penicillin-binding protein	-	1776	*Kingella kingae*	98	648255141
AM	23	Oral	O23_CZ_1	KJ910977	121-1902	CZ	Penicillin-binding protein	-	1782	*Kingella oralis**	80.8	489893834
AM	3	Fecal	F3_CH_11	KJ910965	1187-1852	CH	Chloramphenicol acetyltransferase	-	666	uncultured bacterium DCM009Cm07	61.1	602249281
AM	5	Fecal	F5_CH_2	KJ910967	271-807	CH	Chloramphenicol acetyltransferase	-	537	*Clostridium* sp. CAG:226*	96.6	514417560
AM	6	Fecal	F6_CH_2	KJ910943	1345-1971	CH	Chloramphenicol acetyltransferase	-	627	*Aeromonas salmonicida**	99.5	500229267
AM	6	Fecal	F6_CH_2	KJ910943	2946-3572	CH	Chloramphenicol	-	627	*Aeromonas*	100	50022926

							acetyltransferase			*salmonicida**		7
AM	6	Fecal	F6_CH_4	KJ910944	174-1193	CH	Chloramphenicol acetyltransferase	-	1020	uncultured bacterium DCM009Cm07	61.1	602249281
AM	6	Fecal	F6_CH_5	KJ910945	198-752	CH	Chloramphenicol acetyltransferase	-	555	*Prevotella copri**	90.4	493900592
AM	23	Fecal	F23_CH_11	KJ910951	1298-1636	CH	AraC family transcriptional regulator	-	339	*Escherichia coli* O121:H7 str. 2009C-3299	99.1	607731705
AM	23	Oral	O23_CH_21	KJ910976	997-1638	CH	Chloramphenicol acetyltransferase	-	642	MULTISPECIES: Gammaproteobacteria*	100	500075340
AM	23	Oral	O23_CH_19	KJ910975	550-804	CH	Unknown	-	255	*Neisseria bacilliformis**	61.3	494574895
AM	3	Fecal	F3_TE_1	KJ910966	770-2674	TE	Ribosomal protection protein	-	1905	*[Clostridium] saccharolyticum**	65.7	503038805
AM	6	Fecal	F6_TE_1	KJ910950	1651-3570	TE	Ribosomal protection protein	-	1920	MULTISPECIES: Bacteria*	99.8	488650015
AM	23	Fecal	F23_TE_1	KJ910953	69-1403	TE	MATE transporter	-	1335	*Phascolarctobacterium succinatutens**	99.1	496436881
AM	3	Oral	O3_TE_1	KJ910981	208-1638	TE	MFS transporter *emrB*	-	1431	*Kingella denitrificans**	96.2	489877768
AM	5	Oral	O5_TE_1	KJ910972	769-2208	TE	MFS transporter *emrB*	-	1440	*Neisseria sicca**	95	489857717
AM	5	Oral	O5_TE_3	KJ910973	2-703	TE	MFS transporter	-	702	*Kingella denitrificans**	97.3	489877768
AM	23	Oral	O23_TE_1	KJ910978	1-1845	TE	ABC transporter	-	1845	*Cardiobacterium valvarum**	97.1	494042102
PR	40	Fecal	220_PE_14	KJ911003	3-656	PE	β-lactamase	-	654	*Collimonas fungivorans**	37	503772503
PR	44	Oral	238_PE_3	KJ911002	446-805	PE	β-lactamase, Class A	-	360	*Escherichia coli**	98.3	487678966
PR	44	Oral	238_PE_3	KJ911002	777-1223	PE	β-lactamase, Class A	-	447	*Klebsiella pneumoniae*	99.3	344222647
PR	62	Oral	337_PE_1	KJ910989	390-1307	PE	β-lactamase, Class A	-	918	MULTISPECIES: Pasteurellaceae*	99	491823213
PR	62	Oral	337_PI_2	KJ910996	201-1118	PI	β-lactamase, Class A	-	918	MULTISPECIES: Pasteurellaceae*	99	491823213
PR	62	Oral	337_PITZ_2	KJ910995	282-1199	PITZ	β-lactamase, Class A	-	918	MULTISPECIES: Pasteurellaceae*	99.3	491823213
PR	62	Oral	337_CZ_1	KJ910983	703-1620	CZ	β-lactamase, Class A	-	918	MULTISPECIES: Pasteurellaceae*	100	491823213
PR	62	Oral	337_PE_13	KJ910985	5065-5652	PE	β-lactamase, Class B, subclass B3	-	588	uncultured bacterium	99.5	636669846
PR	62	Oral	337_PE_15	KJ910987	116-1045	PE	β-lactamase, Class B, subclass B3	-	930	uncultured bacterium	100	636669846
PR	62	Oral	337_PE_16	KJ910988	1046-1969	PE	β-lactamase, Class B, subclass B3	-	924	uncultured bacterium	99.7	636669325
PR	62	Oral	337_PE_24	KJ910991	837-1709	PE	β-lactamase,	-	873	uncultured bacterium*	100	63666984

							Class B, subclass B3					2
PR	62	Oral	337_PE_30	KJ910992	3280-3897	PE	β-lactamase, Class B, subclass B3	-	618	uncultured bacterium	98.5	636669508
PR	62	Oral	337_PE_34	KJ910993	562-1008	PE	β-lactamase, Class B, subclass B3	-	447	uncultured bacterium	91.3	636669343
PR	62	Oral	337_PE_11	KJ910984	464-1354	PE	β-lactamase, Class B, subclass B3, BJP	-	891	uncultured bacterium	100	636669512
PR	62	Oral	337_PE_14	KJ910986	712-1602	PE	β-lactamase, Class B, subclass B3, BJP	-	891	uncultured bacterium	99.3	636669830
PR	62	Oral	337_PE_34	KJ910993	1005-1445	PE	β-lactamase, Class B, subclass B3, BJP	-	441	uncultured bacterium	100	636669343
PR	62	Oral	337_PE_3	KJ910994	1205-2056	PE	β-lactamase, Class B, subclass B1, SPM	-	852	uncultured bacterium	99.3	636669520
PR	62	Oral	337_PE_13	KJ910985	4244-4666	PE	Penicillin-binding protein	-	423	uncultured bacterium	98.6	636669835
PR	T100	Oral	T1001_PITZ_7	KJ910963	2-862	PITZ	β-lactamase	-	861	*Ktedonobacter racemifer**	46.4	495198521
PR	T100	Oral	T1001_CZ_5	KJ910961	1-1212	CZ	β-lactamase	-	1212	*Methanosarcina mazei**	38.3	499345516
PR	T100	Oral	T1001_PITZ_5	KJ910962	66-1814	PITZ	Penicillin-binding protein	-	1749	*Neisseria macacae**	100	489872691
PR	T100	Oral	T1001_CZ_11	KJ910959	1532-3334	CZ	Penicillin-binding protein	-	1803	*Neisseria macacae**	99.6	489872691
PR	62	Oral	337_CH_1	KJ910982	349-1578	CH	MFS transporter	-	1230	*Acinetobacter lwoffii*	98.8	498013474
PR	40	Fecal	220_TE_2	KJ911006	6-1925	TE	Ribosomal protection protein	-	1920	MULTISPECIES: Firmicutes**	99.7 - 99.8	649571947
PR	40	Fecal	220_TE_6	KJ911008	118-2043	TE	Ribosomal protection protein	-	1926	MULTISPECIES: Prevotella*	100	490505649
PR	40	Fecal	220_TE_1	KJ911005	74-1234	TE	TetX	-	1161	MULTISPECIES: Bacteroidales**	99.2	598867507
PR	44	Fecal	242_TE_1	KJ911009	523-1683	TE	TetX	-	1161	*Bacteroides fragilis* str. S23L17*	99.5	598867507
PR	44	Fecal	242_TG_2	KJ911011	209-1369	TG	TetX	-	1161	MULTISPECIES: Bacteroidales*	99.7	495926503
PR	62	Fecal	341_TE_1	KJ911000	202-2121	TE	Ribosomal protection protein	-	1920	MULTISPECIES: Bacteria**	99.8 - 100	356668633
PR	62	Fecal	341_TE_2	KJ911001	837-1877	TE	Ribosomal protection protein	-	1041	MULTISPECIES: Bacteria**	98.8 - 100	446614390
PR	T100	Fecal	T1003_TE_12	KJ910954	588-1748	TE	TetX	-	1161	*Bacteroides fragilis**	99.7	492232865
PR	T100	Fecal	T1003_TG_1	KJ910958	1021-2181	TG	TetX	-	1161	*Bacteroides fragilis**	99.7	492232865

PR	62	Oral	337_TE_6	KJ910999	2-1678	TE	ABC transporter	-	1677	*Neisseria cinerea**	99.8	489773067
PR	62	Oral	337_TE_2	KJ910997	82-1290	TE	MFS transporter *bcr/cflA*	-	1209	*Neisseria cinerea**	98.5	489772787
PR	62	Oral	337_TE_4	KJ910998	156-1589	TE	MFS transporter *emrB*	-	1434	*Neisseria sicca**	98.7	489863199
PR	T100	Oral	T1001_TE_1	KJ910964	166-2085	TE	Ribosomal protection protein	-	1920	MULTISPECIES: Bacteria*	100	446614412

AE: Amerindian *E. coli* isolate, AM: Amerindian microbiota, PR: Puerto Rican microbiota. A and B are libraries created from the pooled genomic DNA of Amerindian *E. coli* isolates (Table S2). PE: penicillin, PI: piperacillin, PITZ: piperacillin-tazobactam, CT: cefotaxime, CZ: ceftazidime, CP: cefepime, AZ: aztreonam, CH: chloramphenicol, TE: tetracycline, TG: tigecycline. 'MG1655 gene' was identified through alignment to Escherichia coli str. K-12 substr. MG1655; all alignments ≥ 96% nucleotide ID. Investigated for *E. coli* isolate libraries only.

* The top hit comprises multiple accessions, and only the first is listed.

**The gene aligned equally well to multiple accessions, and only one is listed.

Table S5. Composition of metagenomic libraries from Amerindian and Puerto Rican fecal and oral microbiota.

Origin	**Individual**	**Sex**	**Age**	**Site**	**ID**	**Library size (GB)**
Amerindian	3	M	11	Fecal	F3	7.15
				Oral	O3	7.48
Amerindian	5	M	20	Fecal	F5	0.80
				Oral	O5	8.69
Amerindian	6	M	48	Fecal	F6	4.80
				Oral	O6	0.49
Amerindian	23	F	7	Fecal	F23	3.19
				Oral	O23	3.00
Puerto Rican	40	M	9	Fecal	220	1.08
Puerto Rican	62	M	13	Oral	337	4.09
				Fecal	341	5.77
Puerto Rican	70	M	20	Oral	381	1.41
Puerto Rican	T100	M	45	Oral	T100-1	8.51
				Fecal	T100-3	4.32
Puerto Rican	44	F	7	Oral	238	1.75
				Fecal	242	6.44

Table S6. Antibiotics used in genomic and metagenomic library selections for antibiotic resistance genes.

Cellular target	Antibiotic	Abbreviation	Source	Concentration (ug/mL)
Cell wall synthesis	Penicillin	PE	Natural	128
	Piperacillin	PI	Semi-synthetic	16
	Piperacillin-tazobactam	PITZ	Semi-synthetic	16/4
	Cefotaxime	CT	Semi-synthetic	8
	Ceftazidime	CZ	Semi-synthetic	16
	Cefepime	CP	Semi-synthetic	8
	Meropenem	ME	Semi-synthetic	16
	Aztreonam	AZ	Synthetic	8
Protein synthesis	Chloramphenicol	CH	Natural	8
	Tetracycline	TE	Natural	8
	Tigecycline	TG	Semi-synthetic	2
	Gentamicin	GE	Natural	16
DNA replication	Ciprofloxacin	CI	Synthetic	0.5
Cell membrane	Colistin	CL	Natural	8
Other	Nitrofurantoin	NI	Synthetic	32

Table S10. List of clusters at different k-mer and minimum contig length and their associated taxonomy.

cluster	k-mer size	threshold length	taxonomy
0	55	100	unidentified
1	55	100	unidentified
2	55	100	unidentified
3	55	100	unidentified
4	55	100	Mollicutes_bacterium Fusobacterium_ulcerans
5	55	100	unidentified
6	55	100	unidentified
7	55	100	Candidatus_Pelagibacter_ubique_SAR11
8	55	100	unidentified
9	55	100	Prevotella_sp._Oral_Taxon_472 Prevotella_ruminicola Prevotella_copri Prevotella_melaninogenica
10	55	100	Prevotella_pallens_Oral_Taxon_714 Prevotella_melaninogenica Prevotella_denticola
11	55	100	unidentified
12	55	100	unidentified
13	55	100	Prevotella_copri
14	55	100	unidentified
15	55	100	unidentified
16	55	100	unidentified
17	55	100	unidentified
18	55	100	Beggiatoa_sp.
19	55	100	Mollicutes_bacterium
20	55	100	unidentified
21	55	100	Prevotella_denticola Prevotella_melaninogenica
22	55	100	unidentified
23	55	100	Streptococcus_mitis
24	55	100	unidentified

cluster	k-mer size	threshold length	taxonomy
0	59	100	unidentified
1	59	100	Prevotella_melaninogenica Prevotella_denticola Prevotella_pallens_Oral_Taxon_714
2	59	100	Granulicatella_elegans
3	59	100	unidentified
4	59	100	unidentified
5	59	100	Streptococcus_sanguinis

			Streptococcus_oralis Streptococcus_mitis Streptococcus_sp. Streptococcus_pneumoniae Veillonella_atypica
6	59	100	unidentified
7	59	100	unidentified
8	59	100	unidentified
9	59	100	unidentified
10	59	100	unidentified
11	59	100	unidentified
12	59	100	unidentified
13	59	100	unidentified
14	59	100	Prevotella_copri
15	59	100	Streptococcus_sanguinis Streptococcus_oralis Streptococcus_sp. Streptococcus_pneumoniae Streptococcus_infantis
16	59	100	unidentified
17	59	100	unidentified
18	59	100	Burkholderia_mallei
19	59	100	unidentified
20	59	100	unidentified
21	59	100	unidentified
22	59	100	Streptococcus_sanguinis Streptococcus_dysgalactiae_equisimilis Streptococcus_sp. Streptococcus_pneumoniae Streptococcus_infantis
23	59	100	unidentified
24	59	100	unidentified
25	59	100	Fusobacterium_ulcerans Mollicutes_bacterium
26	59	100	unidentified
27	59	100	Streptococcus_mitis

cluster	k-mer size	threshold length	taxonomy
0	59	200	Granulicatella_elegans
1	59	200	unidentified
2	59	200	unidentified
3	59	200	Streptococcus_mitis
4	59	200	Neisseria_elongata_glycolytica

			Neisseria_mucosa
			Neisseria_subflava
			Neisseria_macacae
			Neisseria_sp._oral_taxon_014
			Neisseria_gonorrhoeae
			Neisseria_flavescens
5	59	200	unidentified
6	59	200	unidentified
7	59	200	unidentified
8	59	200	unidentified
9	59	200	Neisseria_lactamica
			Neisseria_elongata_glycolytica
			Neisseria_sicca
			Neisseria_mucosa
			Neisseria_subflava
			Neisseria_macacae
			Neisseria_meningitidis
			Neisseria_sp._oral_taxon_014
			Neisseria_cinerea
			Neisseria_gonorrhoeae
			Neisseria_flavescens
10	59	200	Prevotella_copri
11	59	200	Neisseria_mucosa
12	59	200	Prevotella_marshii
			Prevotella_melaninogenica
			Bacteroides_xylanisolvens
			Prevotella_ruminicola
			Chryseobacterium_gleum
			Zunongwangia_profunda
			Prevotella_sp._Oral_Taxon_472
			Bacteroides_intestinalis
			Bacteroides_uniformis
			Parabacteroides_johnsonii
			Bacteroides_cellulosilyticus
			Bacteroides_sp.
			Prevotella_buccalis
			Bacteroides_coprophilus
			Bacteroides_eggerthii
			Bacteroides_finegoldii
			Bacteroides_thetaiotaomicron
			Prevotella_buccae
			Alistipes_shahii

			Prevotella_sp._Oral_Taxon_299
			Riemerella_anatipestife
			Flavobacteriacea_bacterium
			Prevotella_timonensis
			Parabacteroides_sp.
			Bacteroides_stercoris
			Gramella_forsetii
			Bacteroides_dorei
13	59	200	unidentified
14	59	200	unidentified
15	59	200	unidentified
16	59	200	Fusobacterium_ulcerans
			Mollicutes_bacterium
17	59	200	unidentified
18	59	200	unidentified
19	59	200	Neisseria_sicca
			Neisseria_mucosa
20	59	200	unidentified
21	59	200	unidentified
22	59	200	Neisseria_meningitidis
			Neisseria_mucosa
			Kingella_denitrificans_Oral_Taxon_582
			Neisseria_sicca
			Neisseria_elongata_glycolytica
23	59	200	Lactobacillus_helveticus
			Streptococcus_sanguinis
			Streptococcus_gordonii
			Streptococcus_oralis
			Streptococcus_sp._Oral_Taxon_071
			Streptococcus_parasanguinis
			Streptococcus_sp.
			Streptococcus_salivarius
			Streptococcus_infantis
24	59	200	Streptococcus_sp.
25	59	200	Neisseria_mucosa
26	59	200	unidentified

cluster	k-mer size	threshold length	taxonomy
0	59	500	Lysinibacillus_fusiformis
			Bacillus_thuringiensis
			Streptococcus_pyogenes
			Enterococcus_casseliflavus_Oral_Taxon_801
			Lactobacillus_gasseri

Streptococcus_sp.
Streptococcus_uberis
Lactobacillus_coryniformis_coryniformis
Streptococcus_mutans
Lactobacillus_fermentum
Lactobacillus_antri
Staphylococcus_capitis
Streptococcus_sanguinis
Streptococcus_pneumoniae
Granulicatella_adiacens
Bacillus_coahuilensis
Granulicatella_elegans
Streptococcus_equinus
Bacillus_amyloliquefaciens_plantarum
Enterococcus_italicus_Oral_Taxon_803
Streptococcus_oralis
Lactobacillus_farciminis
Lactobacillus_ruminis
Streptococcus_dysgalactiae_equisimilis
Streptococcus_peroris_Oral_Taxon_728
Enterococcus_faecium
Eremococcus_coleocola
Leuconostoc_mesenteroides_cremoris
Aerococcus_viridans
Staphylococcus_aureus
Lactobacillus_coleohominis
Streptococcus_anginosus
Enterococcus_faecalis
Lactobacillus_iners
Enterococcus_gallinarum
Listeria_grayi
Streptococcus_sp._Oral_Taxon_07
Streptococcus_mitis
Lactobacillus_coryniformis_torquens
Staphylococcus_epidermidis
Lactobacillus_amylolyticus
Lactobacillus_reuteri
Lactobacillus_fructivorans
Weissella_paramesenteroides
Streptococcus_thermophilus
Anoxybacillus_flavithermus
Lactobacillus_animalis

			Streptococcus_infantis
			Streptococcus_salivarius
			Pediococcus_acidilactici
			Lactobacillus_buchneri
			Bacillus_subtilis_spizizenii
			Streptococcus_parasanguinis
			Bacillus_cereus
			Lactococcus_lactis_lactis
			Lactobacillus_salivarius
			Lactobacillus_johnsonii
			Lactobacillus_acidophilus
			Streptococcus_gordonii
			Streptococcus_suis
			Lactobacillus_rhamnosus
			Lactobacillus_delbrueckii_bulgaricus
			Bacillus_subtilis_subtilis
			Enterococcus_casseliflavus
			Gemmata_obscuriglobus
			Staphylococcus_aureus_aureus
			Streptococcus_australis
			Lactobacillus_paracasei
1	59	500	unidentified
2	59	500	Neisseria_mucosa
			Neisseria_flavescens
			Kingella_denitrificans_Oral_Taxon_582
			Neisseria_sicca
			Neisseria_elongata_glycolytica
			Neisseria_meningitidis

cluster	k-mer size	threshold length	taxonomy
0	63	500	unidentified
1	63	500	Streptococcus_sp.
			Streptococcus_mutans
			Streptococcus_sanguinis
			Streptococcus_pneumoniae
			Streptococcus_equinus
			Streptococcus_infantis
			Streptococcus_salivarius
			Streptococcus_oralis
			Streptococcus_dysgalactiae_equisimilis
			Streptococcus_peroris_Oral_Taxon_728
			Lactococcus_lactis_lactis
			Streptococcus_gordonii

			Streptococcus_anginosus
2	63	500	unidentified
3	63	500	unidentified
4	63	500	Fusobacterium_nucleatum_vincentii
			Fusobacterium_gonidiaformans
			Fusobacterium_sp.
5	63	500	unidentified

cluster	k-mer size	threshold length	taxonomy
0	67	500	Prevotella_copri
1	67	500	Streptococcus_sanguinis
			Granulicatella_elegans
			Streptococcus_equinus
			Streptococcus_oralis
			Streptococcus_dysgalactiae_equisimilis
			Streptococcus_peroris_Oral_Taxon_728
			Enterococcus_faecium
			Enterococcus_gallinarum
			Streptococcus_mitis
			Streptococcus_salivarius
			Lactococcus_lactis_lactis
			Streptococcus_australis
			Streptococcus_pyogenes
			Enterococcus_casseliflavus_Oral_Taxon_801
			Streptococcus_sp.
			Streptococcus_mutans
			Streptococcus_pneumonia
			Granulicatella_adiacens
			Streptococcus_anginosus
			Enterococcus_faecalis
			Streptococcus_cristatus
			Lactobacillus_animalis
			Streptococcus_infantis
			Streptococcus_gordonii
			Streptococcus_suis
2	67	500	unidentified
3	67	500	Fusobacterium_sp.
			Fusobacterium_nucleatum_nucleatum
			Fusobacterium_nucleatum_vincentii
			Fusobacterium_gonidiaformans
4	67	500	unidentified
5	67	500	unidentified

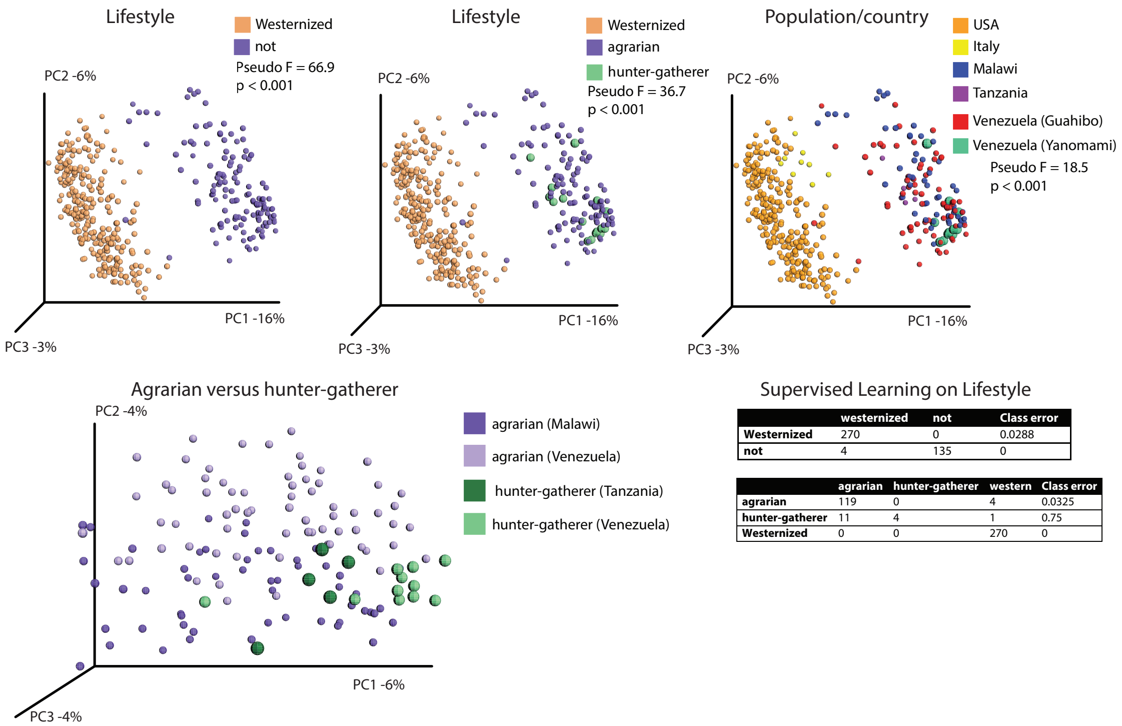
Table S11. Taxa of 329 bacterial strains cultured from feces of 12 isolated Amerindians.

Phylum	**Order**	**Genus**	**Species**	**Subjects**	**Isolates**
Actinobacteria	Actinomycetales	*Arthrobacter*	*sp* [a]	1	1
Bacteroidetes	Bacteroidales	*Bacteroides*	*dorei*	1	3
Bacteroidetes	Bacteroidales	*Bacteroides*	*eggerthii*	1	1
Bacteroidetes	Bacteroidales	*Bacteroides*	*ovatus*	2	2
Bacteroidetes	Bacteroidales	*Porphyromonas*	*sp* [a]	1	2
Bacteroidetes	Bacteroidales	*Porphyromonas*	*sp* [a]	1	1
Bacteroidetes	Bacteroidales	*Prevotella*	*sp* [a]	1	1
Bacteroidetes	Flavobacteria	*Capnocytophaga*	*ochracea*	1	1
Bacteroidetes	Flavobacteria	*Capnocytophaga*	*sp* [a]	1	2
Bacteroidetes	Flavobacteria	*Capnocytophaga*	*sputigena*	1	3
Firmicutes	Bacillales	*Bacillus*	*altitudinis*	1	1
Firmicutes	Bacillales	*Bacillus*	*sp* [a]	1	3
Firmicutes	Clostridiales	*Clostridium*	*clostridioforme*	5	8
Firmicutes	Clostridiales	*Clostridium*	*perfringens*	9	22
Firmicutes	Clostridiales	*Clostridium*	*sp* [a]	8	17
Firmicutes	Clostridiales	*Peptostreptococcus*	*asaccharolyticus*	3	4
Firmicutes	Clostridiales	*Peptostreptococcus*	*sp* [a]	5	7
Firmicutes	Lactobacillales	*Enterococcus*	*durans*	2	3
Firmicutes	Lactobacillales	*Enterococcus*	*faecium*	1	1
Firmicutes	Lactobacillales	*Enterococcus*	*fecalis*	2	12
Firmicutes	Lactobacillales	*Enterococcus*	*hirae*	5	61
Firmicutes	Lactobacillales	*Lactobacillus*	*salivarius*	1	1
Firmicutes	Lactobacillales	*Lactococcus*	*garvieae*	6	29
Fusobacteria	Fusobacteriales	*Fusobacterium*	*sp* [a]	1	1
Proteobacteria	Enterobacteriales	*Escherichia*	*coli*	11	131
Proteobacteria	Enterobacteriales	*Klebsiella*	*pneumoniae*	3	10
Proteobacteria	Neisseriales	*Neisseria*	*mucosa*	2	1

[a] Only defined to genus level

Table S12. Typing of 24 *E. coli* strains from 11 isolated Amerindians

Host			# Strain	Sequence type	MLST Cluster	Clonal Complex	Closest Reference Strain*	Reference strain host	Reference strain host species
ID	Age	Gender							
23	7	Female	55	3314	A	10	Ecor16	Animal	Captive leopard
23	7	Female	56	3315	A	10	Ecor10 + Ecor2	Human	Human
23	7	Female	145	94	B1	58	Ecor45	Animal	Domesticated pig
23	7	Female	402	94	B1	58	Ecor45	Animal	Domesticated pig
3	11	Male	212	3322	B1	641	Ecor69	Animal	Captive celebes ape
3	11	Male	217	3316	B1	Singleton	Ecor29	Animal	Wild kangaroo rat
3	11	Male	264	3316	B1	Singleton	Ecor29	Animal	Wild kangaroo rat
3	11	Male	392	3411	B1	641	Ecor69	Animal	Captive celebes ape
3	11	Male	848	3316	B1	Singleton	Ecor29	Animal	Wild kangaroo rat
9	12	Female	373	94	B1	58	Ecor45	Animal	Domesticated pig
31	17	Male	428	164	B1	Singleton	Ecor34	Animal	Domesticated dog
31	17	Male	429	3318	A	361	Various (Ecor18,19,20,21)	Animal	Captive celebes ape + Domesticated bovine
5	20	Male	24	3313	A	10	Ecor16	Animal	Captive leopard
5	20	Male	31	871	A	10	Ecor16	Animal	Captive leopard
5	20	Male	122	94	B1	58	Ecor45	Animal	Domesticated pig
5	20	Male	393	46	A	10	Ecor16	Animal	Captive leopard
50	22	Male	533	10	A	10	Various (Ecor14,12,11,1,9,8,5,25)	7 Human + 1 Animal	Human + Domesticated dog
36	23	Male	564	3415	B1	Singleton	Ecor27	Animal	Captive giraffe
26	30	Male	335	656	A	10	Ecor6	Human	Human
26	30	Male	383	115	D	115	Ecor46	Animal	Captive celebes ape
26	30	Male	507	3414	B1	58	Ecor45	Animal	Domesticated pig
48	37	Male	502	3413	B1	278	Ecor27	Animal	Captive giraffe
6	48	Male	293	3317	A	10	Ecor15	Human	Human
6	48	Male	399	3412	B1	278	Ecor27	Animal	Captive giraffe



Supervised Learning on Lifestyle

	westernized	not	Class error
Westernized	270	0	0.0288
not	4	135	0

	agrarian	hunter-gatherer	western	Class error
agrarian	119	0	4	0.0325
hunter-gatherer	11	4	1	0.75
Westernized	0	0	270	0

Figure S1. Meta-analysis of 16S V4 region fecal data. PCoA plots based on unweighted UniFrac distances colored by A) whether the population was Westernized or not, B) by subsistence mode, C) by country, and D) agrarian versus hunter-gatherer. D) Pseudo-F and p-values for permanova tests for each category in A-C. Rarefied at 1,000 seqs per sample. Data are from Schnorr *et al* (2014) and Yatsunenko *et al* (2012). Lifestyle effect is much bigger than study effect (in PC2).

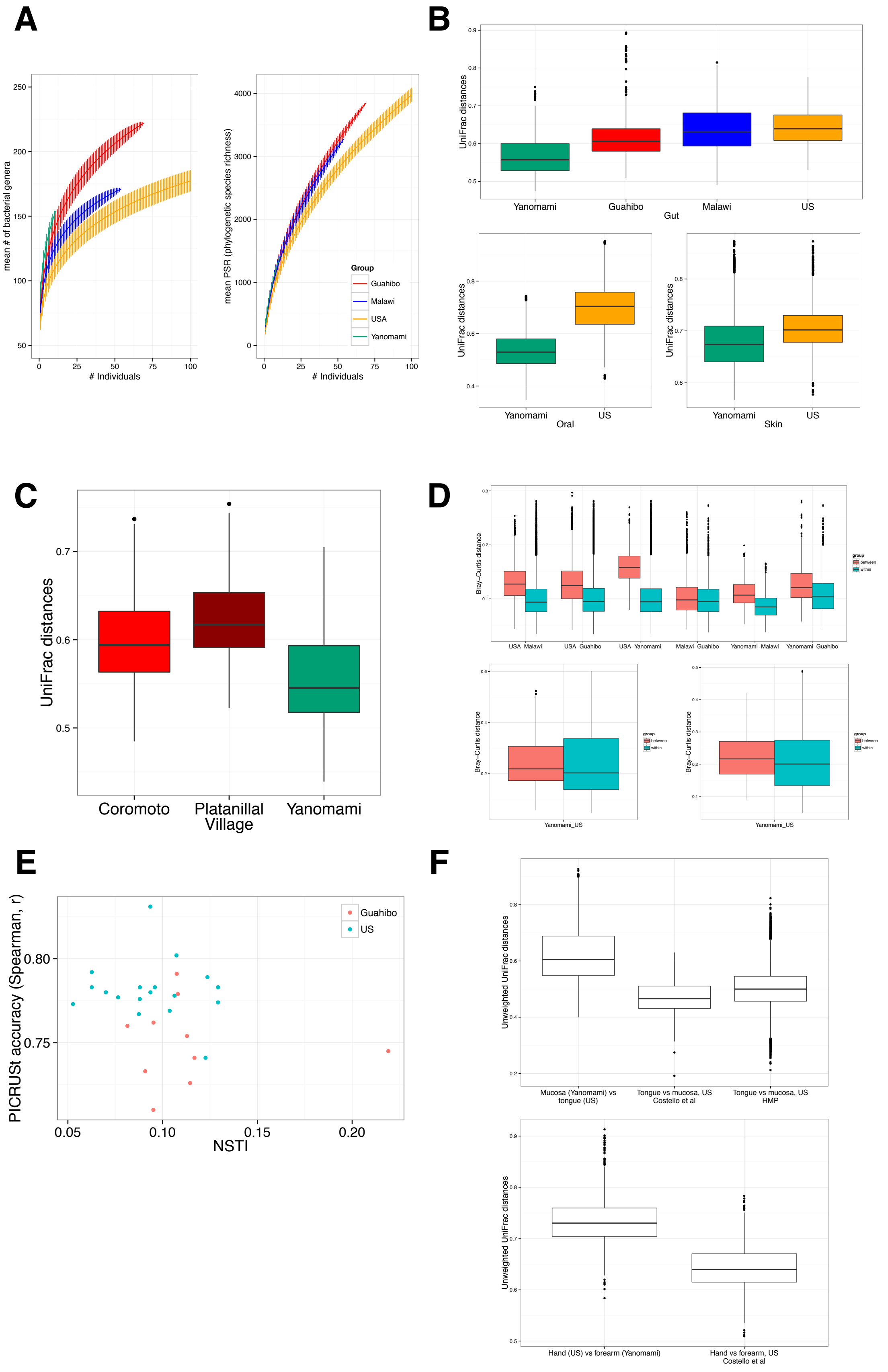

A
mean # of bacterial genera
250
200
150
100
50
0
25
50
75
100
Individuals
mean PSR (phylogenetic species richness)
4000
3000
2000
1000
0
Group
Guahibo
Malawi
USA
Yanomami
B
UniFrac distances
0.9
0.8
0.7
0.6
0.5
Yanomami
Guahibo
Malawi
US
Gut
0.8
0.6
0.4
Yanomami
US
Oral
0.8
0.7
0.6
Yanomami
US
Skin
C
UniFrac distances
0.7
0.6
0.5
Coromoto
Platanillal Village
Yanomami
D
Bray−Curtis distance
0.3
0.2
0.1
USA_Malawi
USA_Guahibo
USA_Yanomami
Malawi_Guahibo
Yanomami_Malawi
Yanomami_Guahibo
group
between
within
0.6
0.4
0.2
Yanomami_US
0.5
0.4
0.3
0.2
0.1
Yanomami_US
E
PICRUSt accuracy (Spearman, r)
0.80
0.75
0.05
0.10
0.15
0.20
NSTI
Guahibo
US
F
Unweighted UniFrac distances
0.8
0.6
0.4
0.2
Mucosa (Yanomami) vs tongue (US)
Tongue vs mucosa, US Costello et al
Tongue vs mucosa, US HMP
0.9
0.8
0.7
0.6
0.5
Hand (US) vs forearm (Yanomami)
Hand vs forearm, US Costello et al

Figure S2. Microbiome diversity between different human groups.

(A) Gamma diversity curves for fecal samples. (Left) Mean number of bacterial genera as a function of number of samples. All differences significant (Welch's t-test with n = 10 subjects, FDR corrected p-values) (Right) Mean phylogenetic species richness as a function of number of samples. Non-western (Yanomami, Guahibo, Malawians) versus US differences significant (Welch's t-test with n = 10 subjects, FDR corrected p-values) **(B) Unweighted UniFrac distances within population and body site.** All differences were significant (p < 0.001, PERMANOVA with 999 permutations) **(C) Unweighted UniFrac distances of fecal microbiota within villages in two Amerindian populations**. Distances within samples from the uncontacted Yanomami village are significantly smaller than distances within either of the Guahibo villages, Coromoto and Platanillal (p < 0.001, PERMANOVA with 999 permutations) **(D) Within versus between Bray-Curtis distances between all pairs of populations with metagenomic data. S**eparation between populations across body sites after randomly sub-sampling all populations to the same number of samples. Top: Fecal (p < 0.001, PERMANOVA with 999 permutations). Bottom, left: oral (p = 0.002, PERMANOVA with 999 permutations). Bottom, right: skin (p < 0.001, PERMANOVA with 999 permutations) **(E) Correlation between shotgun- and PICRUSt-predicted metagenomes in a US and Guahibo population**. Correlation measured as Spearman's *r* (vertical axis) versus mean closest reference genome measure as Nearest Sequenced Taxon Index (NSTI, horizontal axis). Mean NSTI of US subjects is not significantly lower than that of Guahibos (p = 0.1586, t-test), but Spearman's *r* is (p = 0.003, t-test) **(F) Unweighted UniFrac distances between microbiota in different sub-sites of the same body location**. (Top) Comparison between oral sub-sites. Left (this study): US oral mucosa vs Yanomami tongue; middle (Costello et al, Science 2009): US oral mucosa vs US tongue; right (HMP, Nature 2012): US oral mucosa vs US tongue. Differences between populations are larger than between sub-body sites (p < 0.001, PERMANOVA with 999 permutations) (Bottom) Comparison between cutaneous sub-sites. Left (this study): US hand vs Yanomami forearm; right (Costello et al, Science 2009): US hand vs US forearm. Differences between populations larger than between sub-sites (p < 0.001, PERMANOVA with 999 permutations)

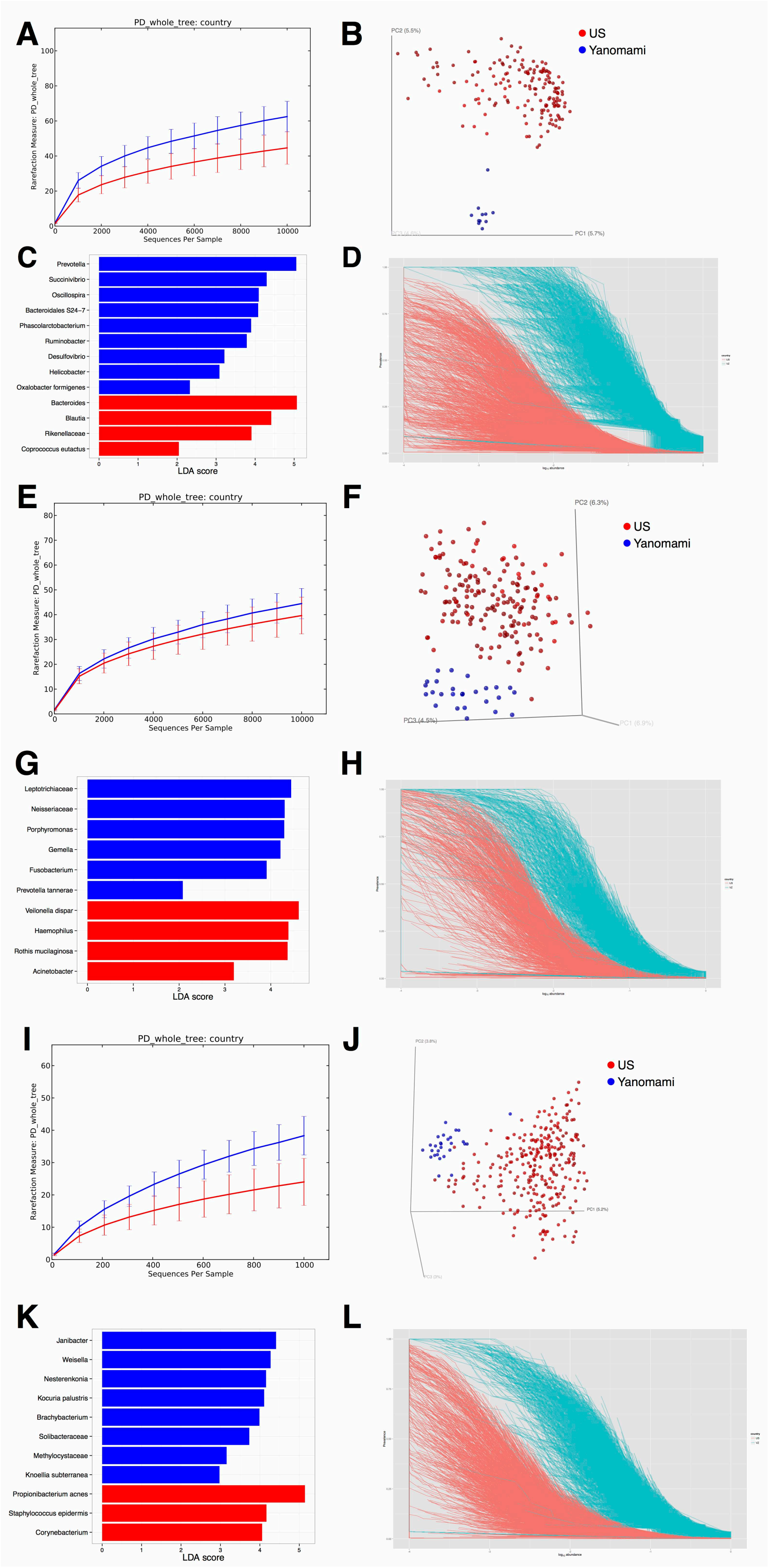

A
PD_whole_tree: country
Rarefaction Measure: PD_whole_tree
0
20
40
60
80
100
0
2000
4000
6000
8000
10000
Sequences Per Sample
B
PC2 (5.5%)
PC1 (5.7%)
PC3 (4.6%)
US
Yanomami
C
Prevotella
Succinivibrio
Oscillospira
Bacteroidales S24–7
Phascolarctobacterium
Ruminobacter
Desulfovibrio
Helicobacter
Oxalobacter formigenes
Bacteroides
Blautia
Rikenellaceae
Coprococcus eutactus
0
1
2
3
4
5
LDA score
D
Prevalence
1.00
0.75
0.50
0.25
0.00
-4
-3
-2
-1
0
log10 abundance
country
US
VZ
E
PD_whole_tree: country
Rarefaction Measure: PD_whole_tree
0
10
20
30
40
50
60
70
80
0
2000
4000
6000
8000
10000
Sequences Per Sample
F
PC2 (6.3%)
PC3 (4.5%)
PC1 (6.9%)
US
Yanomami
G
Leptotrichiaceae
Neisseriaceae
Porphyromonas
Gemella
Fusobacterium
Prevotella tannerae
Veilonella dispar
Haemophilus
Rothis mucilaginosa
Acinetobacter
0
1
2
3
4
LDA score
H
Prevalence
log10 abundance
country
US
VZ
I
PD_whole_tree: country
Rarefaction Measure: PD_whole_tree
0
10
20
30
40
50
60
0
200
400
600
800
1000
Sequences Per Sample
J
PC2 (3.8%)
PC1 (5.2%)
PC3 (3%)
US
Yanomami
K
Janibacter
Weisella
Nesterenkonia
Kocuria palustris
Brachybacterium
Solibacteraceae
Methylocystaceae
Knoellia subterranea
Propionibacterium acnes
Staphylococcus epidermis
Corynebacterium
0
1
2
3
4
5
LDA score
L
Prevalence
1.00
0.00
-4
-3
-2
-1
0
log10 abundance
country
US
VZ

Figure S3. Microbiota diversity in fecal, oral, and skin samples from uncontacted Yanomami in relation to US subjects using the V2 region of the 16S rRNA gene. (A) Faith's phylogenetic diversity (average ± standard deviation) of fecal samples from Yanomami and US subjects. OTU tables rarefied at 10,000 sequences/sample. Inter-population differences were significant ($p < 1.23e-06$, t-test) **(B)** PCoA plot based on UniFrac distances calculated on the OTU table of fecal samples rarefied at 10,000 sequences/sample. **(C)** Top discriminative bacteria among populations in fecal samples as determined by LEfSe analysis. **(D)** Normalized prevalence/abundance curves for all OTUs in fecal samples. **(E)** Faith's phylogenetic diversity (average ± standard deviation) of oral samples from Yanomami and US subjects. OTU tables rarefied at 10,000 sequences/sample. Inter-population differences were significant ($p = 0.0001$, t-test). **(F)** PCoA plot based on UniFrac distances calculated on OTU tables of oral samples rarefied at 10,000 sequences/sample. **(G)** Top discriminative bacteria among populations in oral samples as determined by LEfSe analysis. **(H)** Normalized prevalence/abundance curves for all OTUs in oral samples. **(I)** Faith's phylogenetic diversity (average ± standard deviation) of skin samples from Yanomami and US subjects. OTU tables rarefied at 1,000 sequences/sample. Inter-population differences were significant ($p < 1.418e-13$, t-test). **(F)** PCoA plot based on UniFrac distances calculated on OTU tables of skin samples rarefied at 1,000 sequences/sample. **(G)** Top discriminative bacteria among populations in skin samples as determined by LEfSe analysis. **(H)** Normalized prevalence/abundance curves for all OTUs in skin samples.

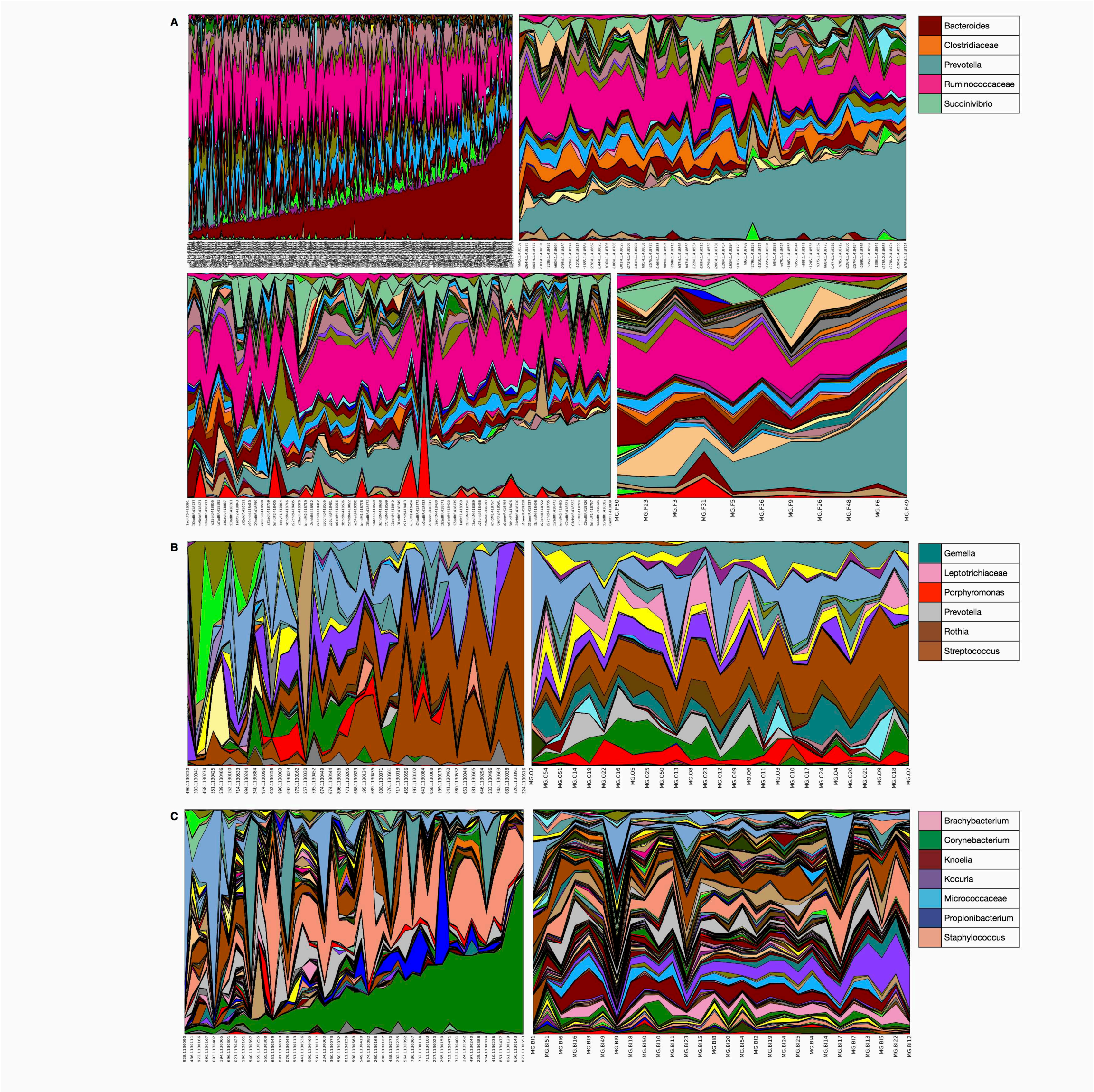


Figure S4. Taxonomic distribution across gut samples sorted by the dominant genera within each population. (A) From top to bottom, and left to right: Taxonomic abundance in fecal samples of US, dominated by Bacteroides; Malawi, Guahibo, and Yanomami, dominated by Prevotella. **(B)** Left to right: Taxonomic abundance in oral samples of US, dominated by Streptococcus and with significant abundance of Rothia and Yanomami, showing more even diversity. **(C)** Left to right: Taxonomic abundance in skin samples of US, dominated by Corynebacterium; Yanomami, showing more even diversity but also significant levels of Kocuria.

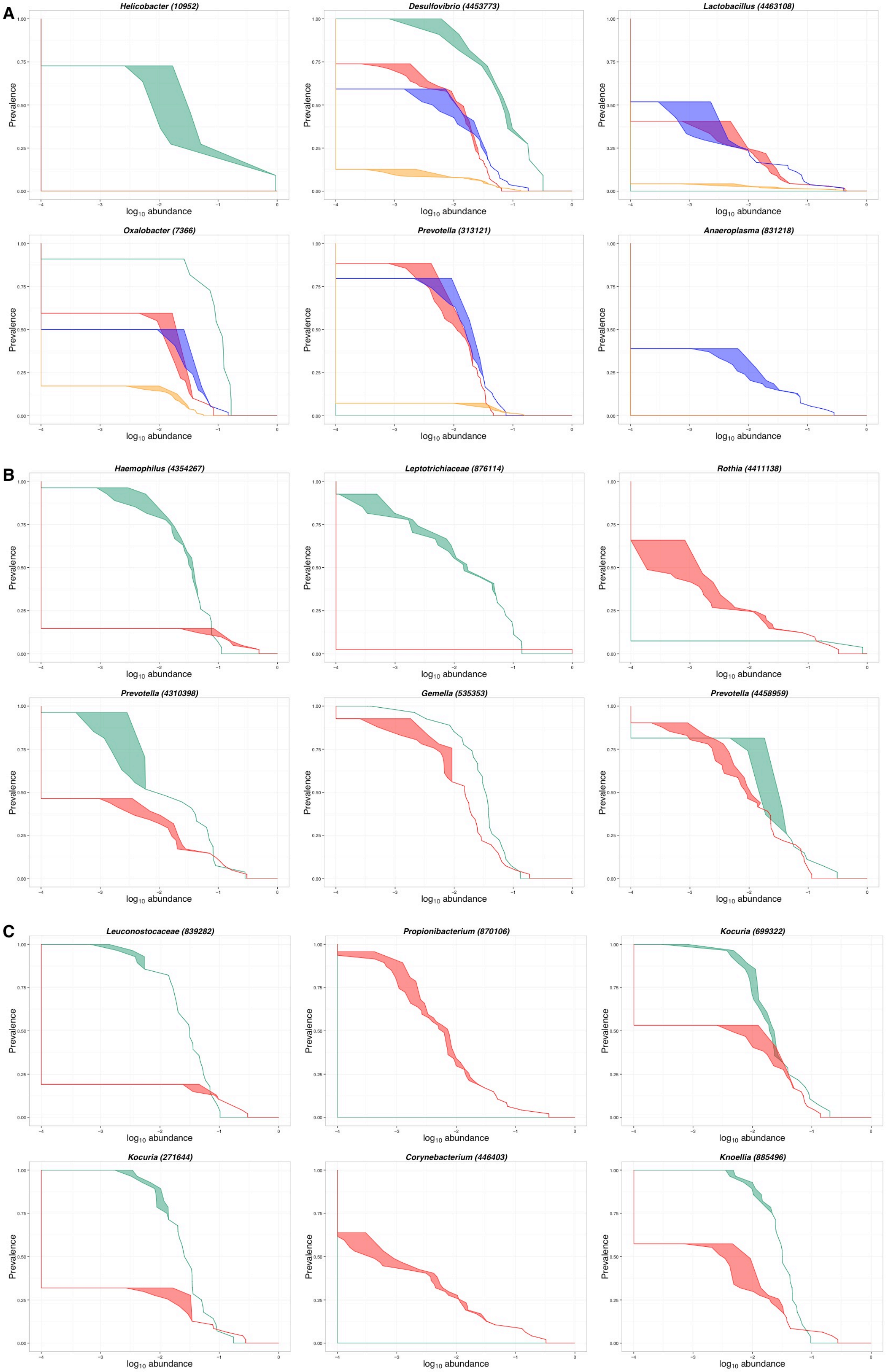

A
Helicobacter (10952)
Desulfovibrio (4453773)
Lactobacillus (4463108)
Oxalobacter (7366)
Prevotella (313121)
Anaeroplasma (831218)
B
Haemophilus (4354267)
Leptotrichiaceae (876114)
Rothia (4411138)
Prevotella (4310398)
Gemella (535353)
Prevotella (4458959)
C
Leuconostocaceae (839282)
Propionibacterium (870106)
Kocuria (699322)
Kocuria (271644)
Corynebacterium (446403)
Knoellia (885496)
Prevalence
1.00
0.75
0.50
0.25
0.00
-4
-3
-2
-1
0
$\log_{10}$ abundance

Figure S5. Prevalence/abundance curves of representative bacterial OTUs in feces (A), oral (B) and skin (C). Shaded areas encompass the mean and maximum curves generated over 10 independent rarefactions of the OTU table. The Greengenes OTU identifier is indicated in parentheses.

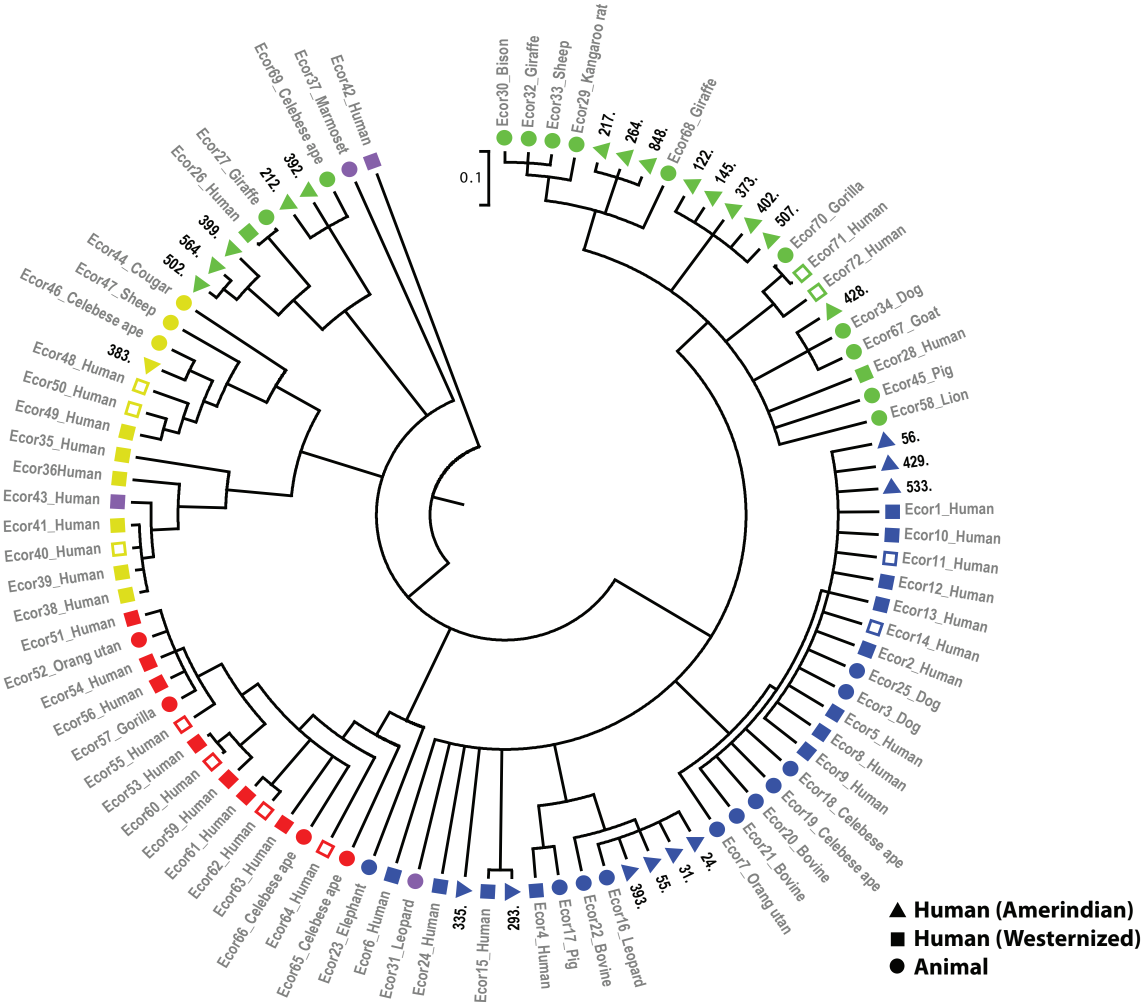


Figure S6. ClonalFrame analysis of *Escherichia coli strains* based on the sequence of 7 housekeeping genes. Strains included 72 reference collection *Escherichia coli* (ECOR) (circles, strains isolated from animals; squares, strains isolated from westernized humans) and 24 Amerindian isolates (triangles), the tree is rooted on *Escherichia fergusonii* (not shown). Non-pathogenic strains are indicated by a full symbol, whereas pathogenic strains are represented by an open symbol. Colors indicate the 5 major phylogenetic groups (blue, A; green, B1; red, B2; yellow, D; purple, E). For presentation purposes Amerindian isolates are indicated by black font, whereas the reference strains (ECOR) are indicated by grey font.

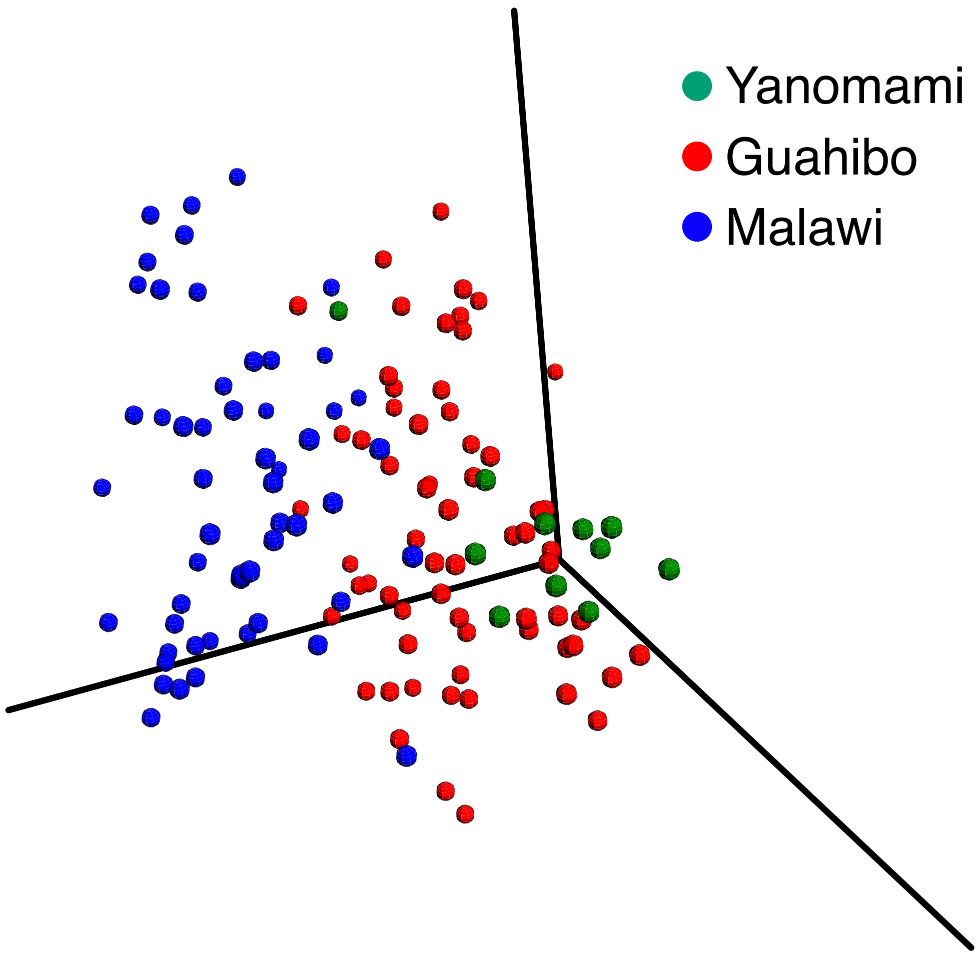


Figure S7. PCoA plot of Yanomami, Malawian, and Guahibo fecal samples. PCoA based on unweighted UniFrac distances on a table rarefied at 5,000 sequences per sample.

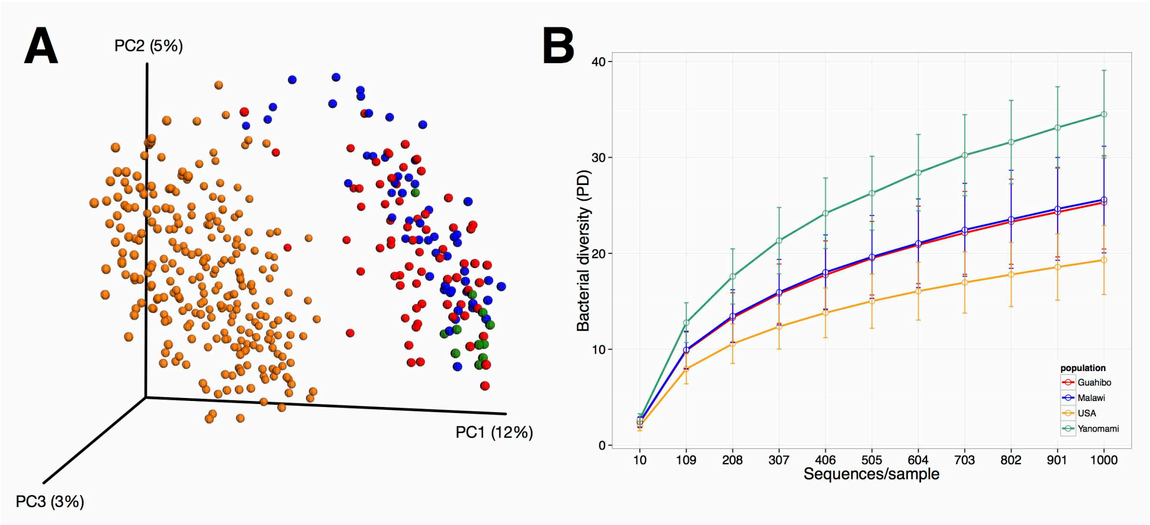


Figure S8. Fecal microbiome of non-Bacteroides OTUs in Yanomami, Malawian, Guahibo, and US samples. (A) PCoA plot based on unweighted UniFrac distances on a table rarefied at 1,000 sequences per sample. **(B)** Faith's phylogenetic diversity (average ± standard deviation) of oral samples from Yanomami and US subjects. OTU table rarefied at 1,000 sequences per sample.

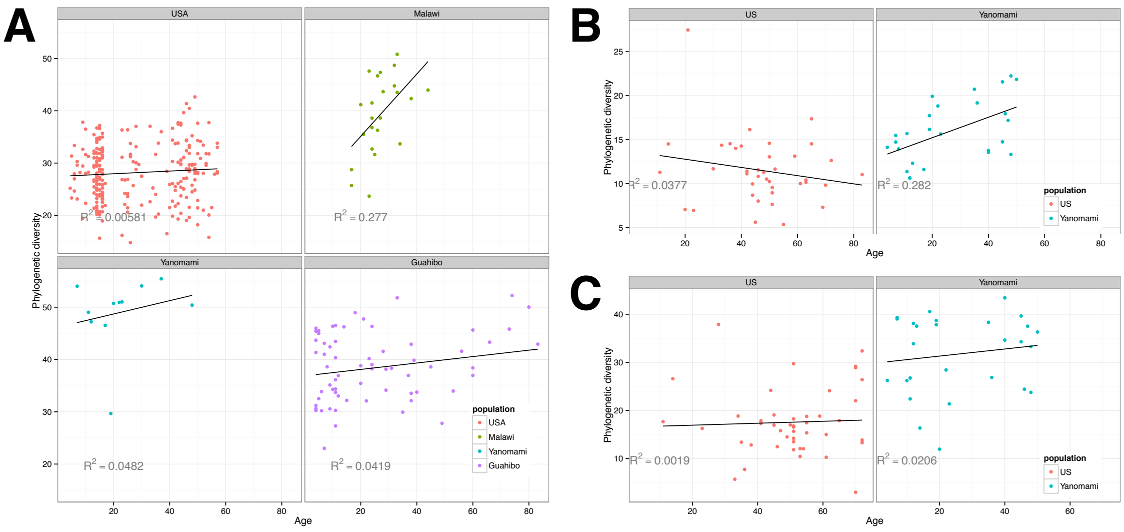


Figure S9. Bacterial diversity as a function of age and body site. (A) Average Faith's phylogenetic diversity of fecal samples from US, Malawi, Yanomami, and Guahibo subjects plotted against age. Samples rarefied at 5,000 sequences/sample. **(B)** Average Faith's phylogenetic diversity of oral samples from US and Yanomami subjects plotted against age. Samples rarefied at 1,500 sequences/sample. **(C)** Average Faith's phylogenetic diversity of skin samples from US and Yanomami subjects plotted against age. Samples rarefied at 1,500 sequences/sample.

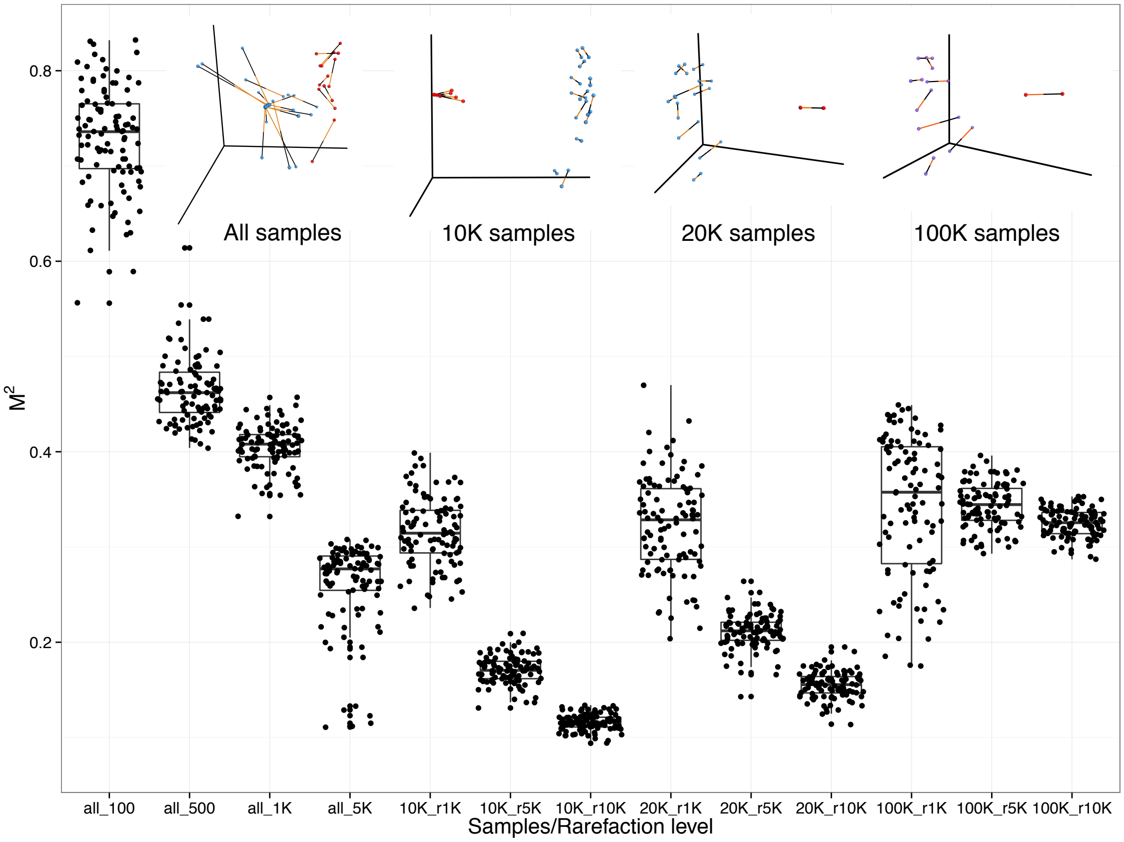


Figure S10. Procrustes analyses of shotgun metagenomic and 16S rRNA data for different subsets of samples. The vertical axis displays Procrustes M^2 between shotgun and 16S rRNA data. The horizontal axis presents subsets of samples ('all' = all samples; '10K' = samples with at least 10,000 reads mapped to IMG genoms; '20K' = samples with at least 20,000 reads mapped; '100K' = samples with at least 100,000 reads mapped) and rarefaction cutoffs (e.g. 'all_100' = all samples, rarefied at 100 seqs/sample; '10K_5K' = samples with 10K mapped reads, rarefied at 5,000 seqs/sample). A representative Procrustes plot is shown on top for each subset of samples (all samples, samples with >10K mapped reads, samples with >20K mapped reads, samples with >100K mapped reads).

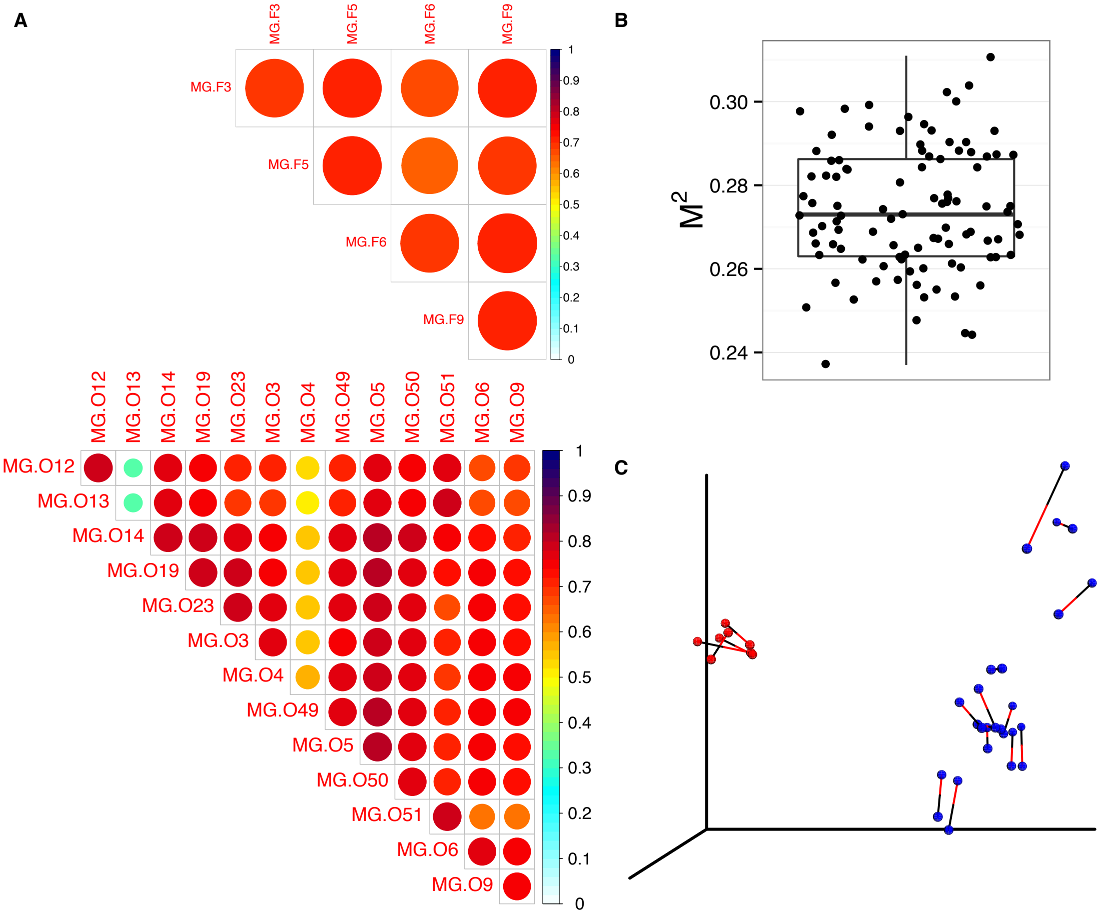


Figure S11. Comparison of PICRUSt versus sequenced shotgun metagneomes of fecal and oral samples with at least 10,000 mapped reads. (A) Pearson correlation coefficient between PICRUSt-predicted and sequenced metagenomes for each sample. Size and color of the circles indicate strength of the correlation. **(B)** Procrustes analysis between PICRUSt-predicted and sequenced metagenomes. 100 rarefied PICRUST-predicted and shotgun sequenced metagenomes were compared using Procrustes. The M^2 values of the comparisons, indicating goodness of fit (the lower the better), are presented ($p < 0.01$, Monte Carlo permutation test). **(C)** Representative Procrustes plot between PICRUSt and shotgun metagenomic tables.

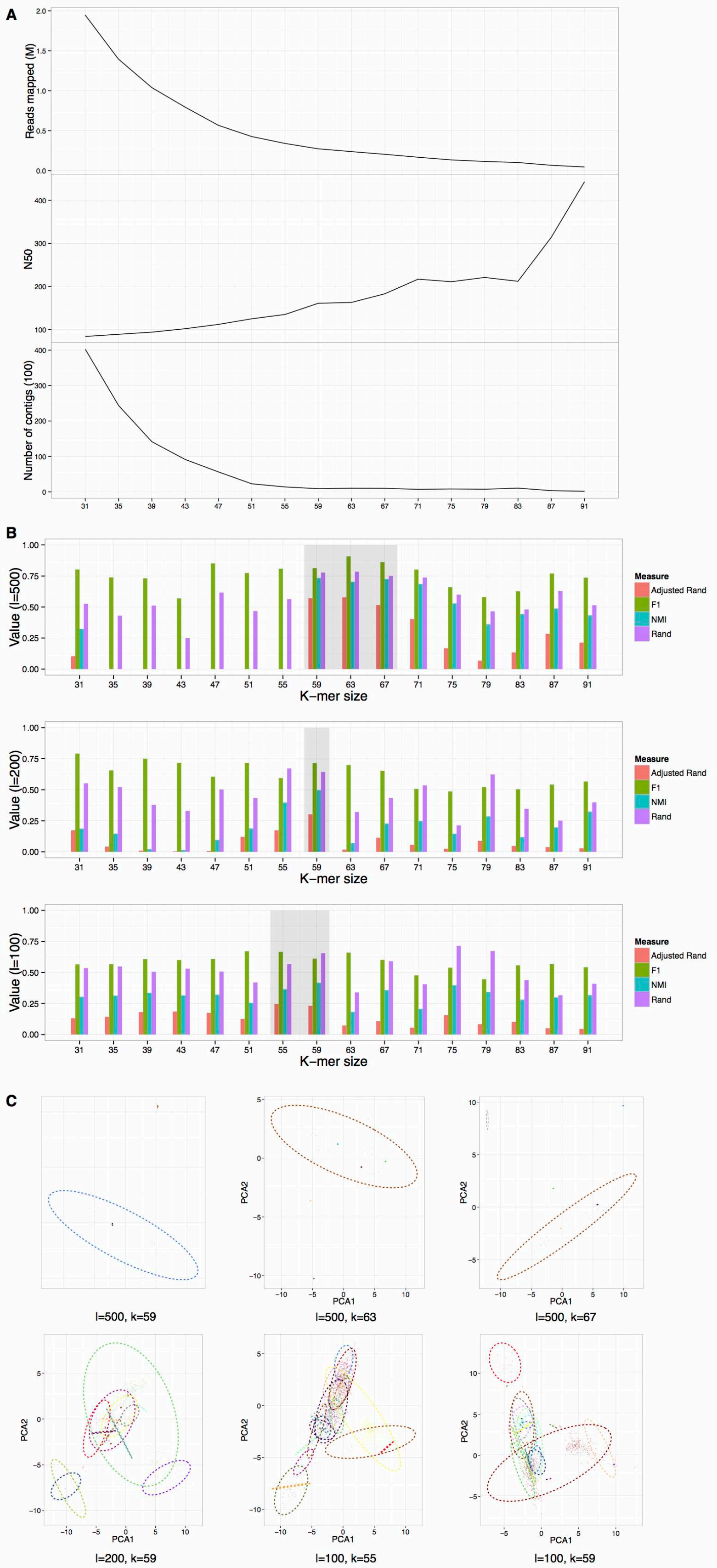

A
Reads mapped (M)
N50
Number of contigs (100)
31 35 39 43 47 51 55 59 63 67 71 75 79 83 87 91
B
Value (l=500)
Value (l=200)
Value (l=100)
K−mer size
Measure
Adjusted Rand
F1
NMI
Rand
C
PCA1
PCA2
l=500, k=59
l=500, k=63
l=500, k=67
l=200, k=59
l=100, k=55
l=100, k=59

Figure S12. Bacterial genome assembly.

(A) Shotgun metagenome assembly statistics per k-mer size. (Top) Number of reads mapped to contigs. (Middle) N50 statistic. (Bottom) Number of contigs. **(B) Quality metrics of metagenome assembly.** Statistics for cluster quality at different k-mer values (31 to 91) and three thresholds for contig length: All contigs above 500bp (top), all contigs above 200bp (middle), and all contigs above 100bp (bottom). Adjusted Rand = adjusted rand index; F1 measure =2*(precision * recall) / (precision + recall). NMI = Normalized Mutual Information. Rand = rand index. K-mers that resulted in best overall statistics are shaded (k=59, 63, 67 for >500bp contigs; k=59 for >200bp contigs; k=55, 59 for >100bp contigs). **(C) Contig clusters of genome assembly.** Contig clusters obtained with CONCOCT 0.3.3 using contigs of >500bp (k=59, 63, 67), >200bp (k=59), and >100bp (k=55, 59).