Metagenomic signatures of gut infection caused by different *E. coli* pathotypes

Peña-González, A., Soto-Girón, M.J., Smith S., Sistrunk J., Montero L., Páez M., Ortega E., Hatt J.K., Cevallos, W., Trueba G., Levy, K. and K.T. Konstantinidis

SUPPLEMENTARY MATERIAL

Fig. S1

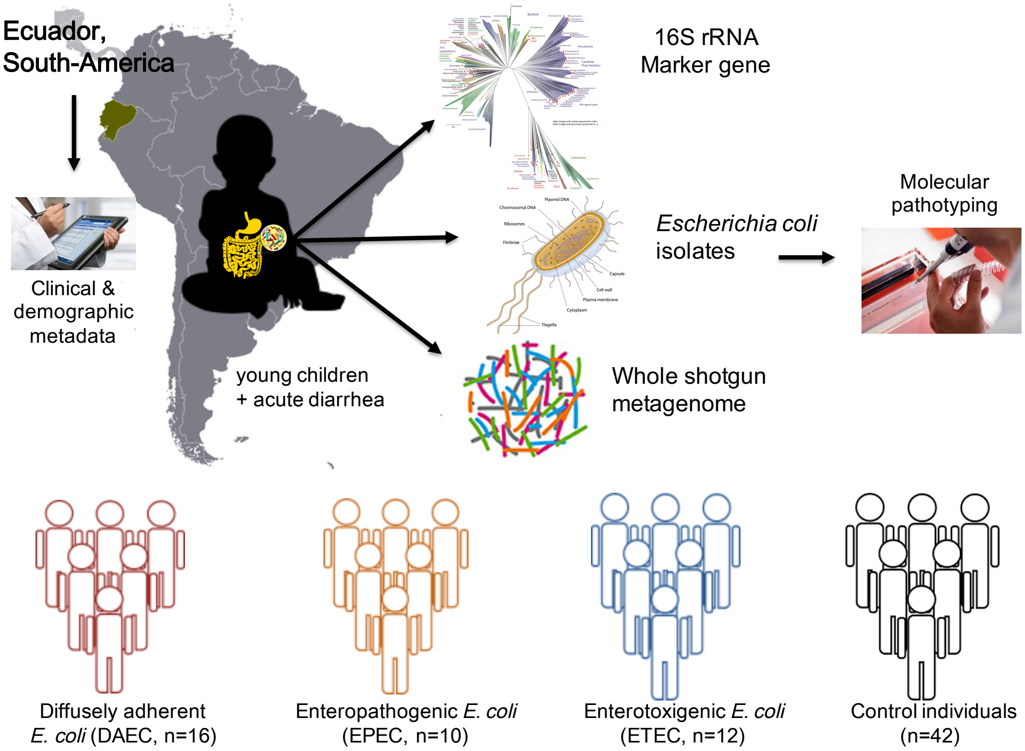

Fig. S1. Study design. A total of 80 young children between 1-6 years old were enrolled in this study, including 38 cases of acute diarrhea disease and 42 age-matched, control individuals. Cases included individuals visiting the clinic site with acute diarrhea (defined as three or more loose stools in a 24-hour period), and controls included individuals visiting the same clinic site for any other complaint, without diarrhea or vomiting in the prior seven days. From the cases of diarrhea, 16 individuals were PCR-positive for *afaB-I*, a virulence marker of DAEC (Diffusely Adherent *E. coli*), 10 were positive for *bfp*, a marker gene for typical EPEC (Enteropathogenic *E. coli*) and 12 were positive for *elt* and/or *sta*, marker genes of ETEC (Enterotoxigenic *E. coli*). Individuals included in control group were all PCR-negative for any of the pathotype markers characterized in this study. All 80 individuals were taxonomically screened by amplicon sequencing of the 16S rRNA marker gene, while all cases of diarrhea (n=38) and a subset of control samples (n=23) were further subjected to whole shotgun metagenomic sequencing. In addition, *E. coli* isolates isolated from stool samples in all cases of diarrhea were also sequenced for pathogenomic characterization and comparison.

Fig. S2

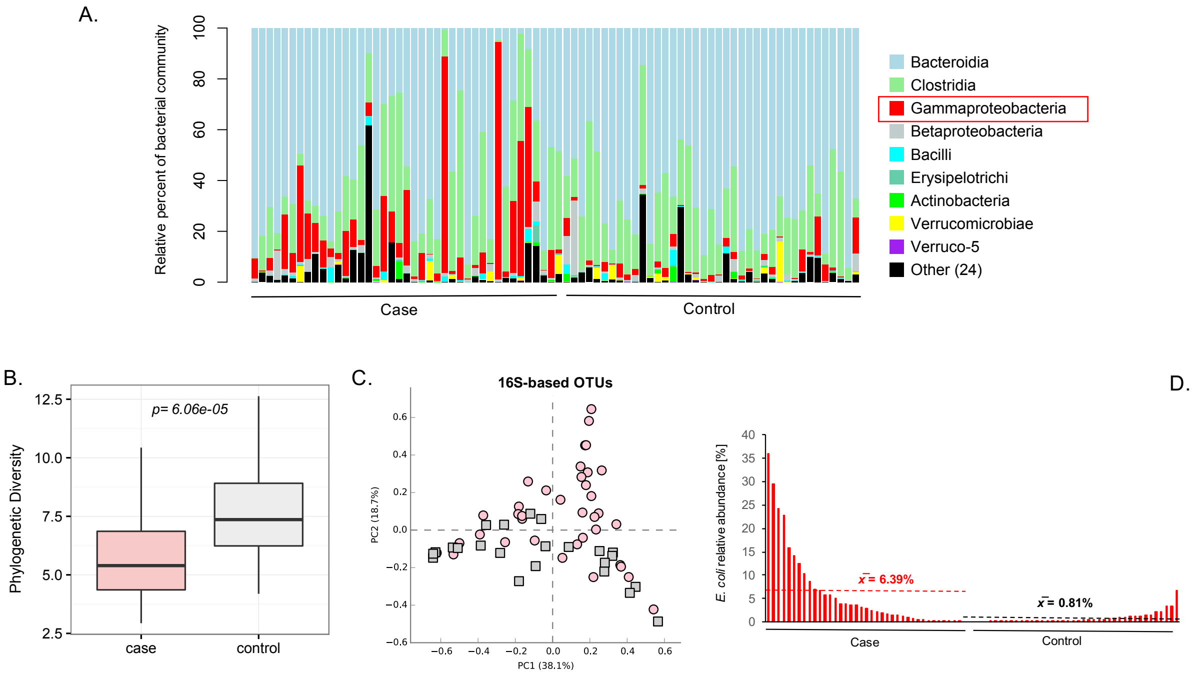

Fig. S2. 16S rRNA gene-based microbial community composition and diversity differences between diarrhea and control samples. Panel (A) shows the relative abundance of bacterial groups classified at the class level for diarrheal and control samples. Only the top ten most abundant phylogroups are displayed. Note that a higher abundance of *Gammaproteobacteria* was observed in diarrhea versus control groups. Panel (B) shows significant differences in Faith's phylogenetic diversity (PD) between diarrhea and control samples. Consistent with previous literature, diarrheal samples presented lower community diversity than control ones. Panel (C) represents the overall community dissimilarity based on the taxonomic composition at the genus level using Bray-Curtis dissimilarities matrix. Pink circles represent cases of diarrhea and gray squares represent control samples. Panel (D) shows the relative abundance of the 16S rRNA gene-based OTU (or sequence variant) taxonomically assigned to *E. coli*.

Fig. S3

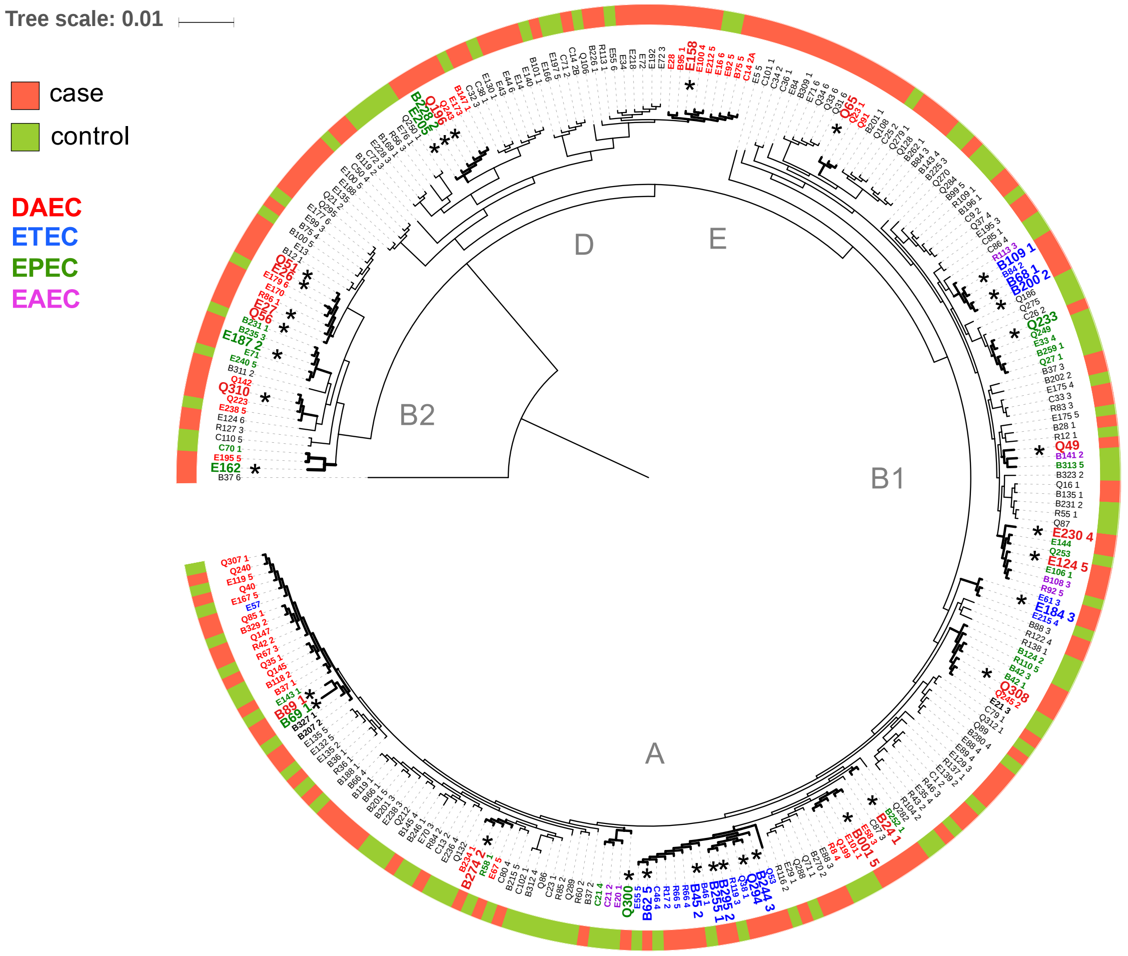

Fig. S3. Detection of clonal complexes in core genome phylogeny. The phylogenetic reconstruction of 263 *E. coli* strains circulating in rural and urban regions in northern Ecuador was calculated from the concatenated alignment of 1,200 core orthologous genes using FastTree 2.1.7 with the GTR model for nucleotide evolution and 1,000 SH-like local support replicates. The tree is also cross-referenced with metadata about clinical status (outer circle), showing isolates obtained from cases of diarrhea in red and control in green. Clonal complexes within the tree are bolded and the different isolate IDs are color-coded by pathotypes: red for DAEC, blue for ETEC, green fro EPEC and purple for EAEC. See Materials and Methods section on how clonal complexes were identified. In addition, the isolates that are part of this study are denoted with asterisks.

Fig. S4

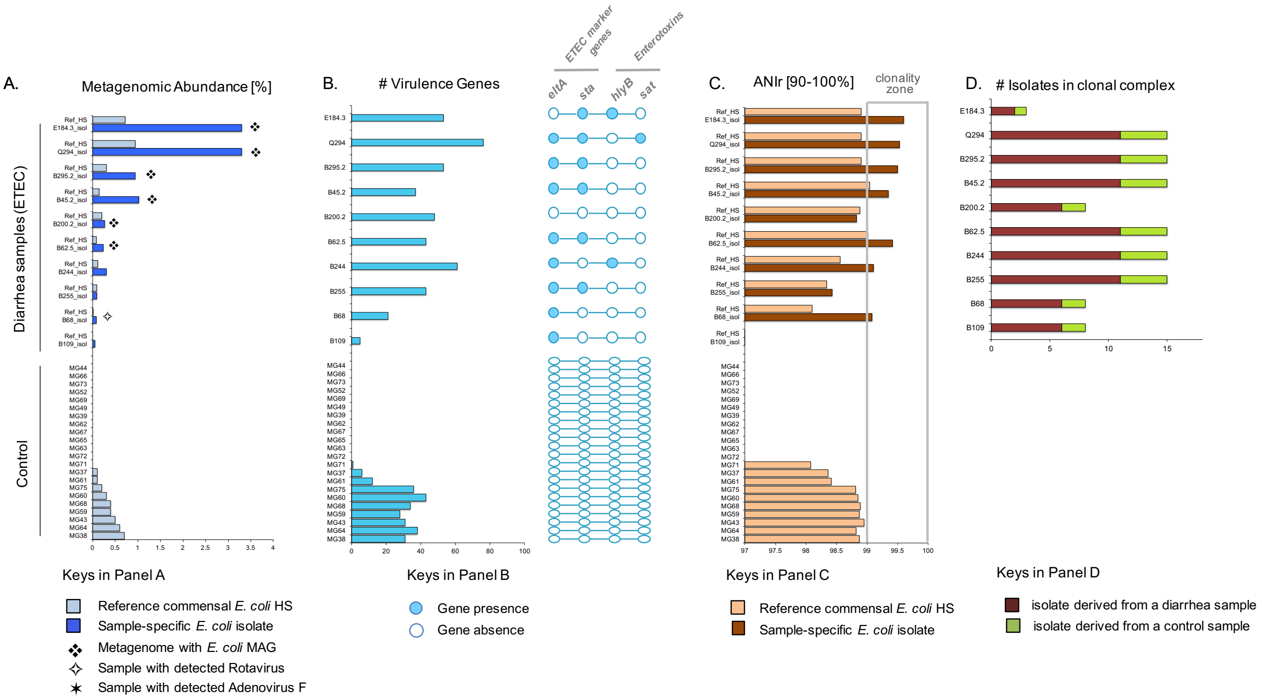

Fig. S4. Evaluation of Enterotoxigenic *E. coli* (ETEC) as causative agent of diarrhea. Panel (A) shows the estimated metagenomic abundance of the reference commensal *E. coli* (strain HS, in light blue), the ETEC isolate (in blue) recovered from the sample, along with the Elisa-based detection of Rotavirus (+) for each sample analyzed (rows). Samples where high-quality *E. coli* MAGs were recovered from the corresponding metagenome are denoted by a star. Panel (B) shows the number of total *E. coli* virulence genes observed in the metagenome and an array of four hallmark virulence factors including the two ETEC marker genes (*eltA* and *sta*) and two additional enterotoxins (the hemolysin subunit B (*hlyB*) and the secreted autotransporter toxin (*sat*)). Panel (C) shows the estimated *E. coli* intra-population diversity measured by ANIr of reads against the reference commensal strain HS (light orange) and the isolate obtained from the sample (dark brown). To avoid any potential bias by low *in-situ* abundance, only samples where the average sequence depth of the reference genome was ≥1X were evaluated for ANIr. Panel (D) shows the number of isolates that originated from cases of diarrhea (in red) vs. control samples (in green) and were assigned in the same core-genome-based phylogenetic clade as the isolate (epidemiology).

Fig. S5

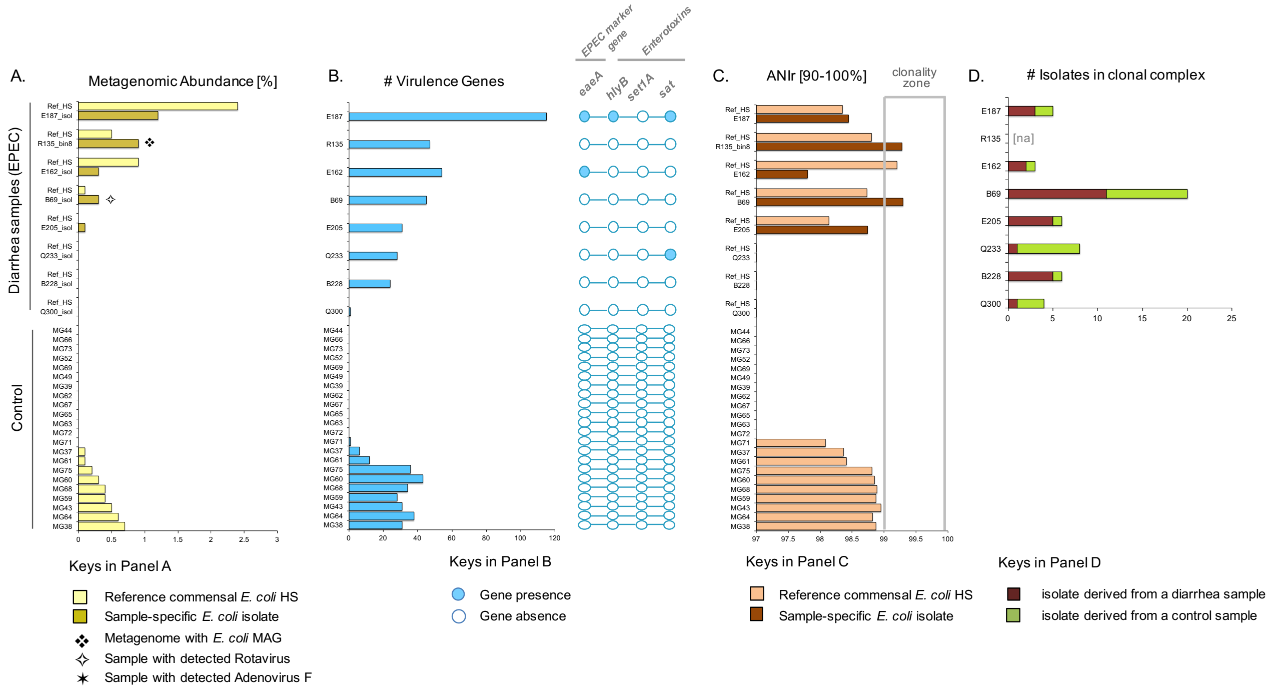

Fig. S5. Evaluation of Enteropathogenic *E. coli* (EPEC) as causative agent of diarrhea. Panel (A) shows the estimated metagenomic abundance of the reference commensal *E. coli* (strain HS, in light yellow), the EPEC isolate (in yellow) recovered from the sample, along with the Elisa-based detection of Rotavirus (+) for each sample analyzed (rows). Samples where high-quality *E. coli* MAGs were recovered are denoted by a star. Panel (B) shows the number of total *E. coli* virulence genes observed in the metagenome and an array of four hallmark virulence factors including the EPEC marker gene (*eaeA,* intimin) and three enterotoxins, i.e., the hemolysin subunit B (*hlyB*), the enterotoxin (*set1A*), and the secreted autotransporter toxin (*sat*). Panel (C) shows the estimated *E. coli* intra-population diversity measured by ANIr of reads against the reference commensal strain HS (light orange) and the isolate obtained from the sample (dark brown). To avoid any potential bias by low *in-situ* abundance, only samples where the average sequence depth of the reference genome was ≥1X were evaluated for ANIr. Panel (D) shows the number of isolates that originated from cases of diarrhea (in red) vs. control samples (in green) and were assigned in the same core-genome-based phylogenetic clade as the isolate (epidemiological data).

Fig. S6

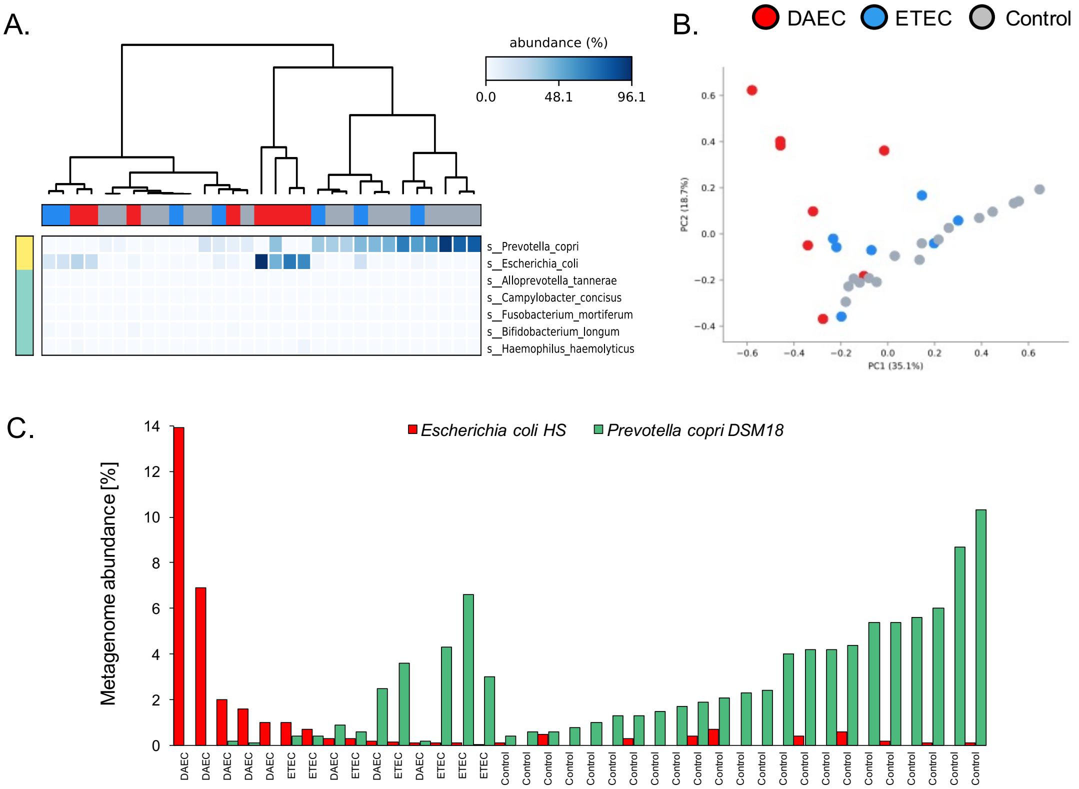

Fig. S6. Differentially abundant taxa between *E. coli* infectious diarrheal and control samples. Panel (A) shows a heatmap of the relative abundance of the seven significantly differentially abundant species between diarrhea and control groups for infections caused by different pathotypes based on the metagenomic data (Figures 3, S4 and S5; color-coded by pathotype as shown in the key of Panel B). Panel (B) shows a principal component analysis plot of the taxonomic relatedness of samples with pathotype infection and controls (see key). Taxonomic relatedness was assessed at the species level using clade-specific marker genes with MetaPhlAn2. Panel (C) shows the estimated metagenomic abundance of *E coli*, using strain HS as the reference commensal genome to recruit *E. coli* reads (NC_009800.1) and *Prevotella copri*, using as reference the genome of strain *DSM18205*, a gut fermentative microbe (NZ_ACBX00000000.2).

Table S1. Description of NGS 'omics' datasets used in this study. Y corresponds to 'yes' and N corresponds to 'no'. These conventions denote whether or not an individual sample was processed for the specific type of analysis shown. Note that no isolate was available for samples R135 and B64 (bolded).

Index	Sample ID	Metagenome	16S rRNA	Isolate	case/control	*E coli* Pathotype
1	B001	Y	Y	Y	case	DAEC
2	B24	Y	Y	Y	case	DAEC
3	B274	Y	Y	Y	case	DAEC
4	B89	Y	Y	Y	case	DAEC
5	E124	Y	Y	Y	case	DAEC
6	E158	Y	Y	Y	case	DAEC
7	E230	Y	Y	Y	case	DAEC
8	E26	Y	Y	Y	case	DAEC
9	E27	Y	Y	Y	case	DAEC
10	Q196	Y	Y	Y	case	DAEC
11	Q308	Y	Y	Y	case	DAEC
12	Q310	Y	Y	Y	case	DAEC
13	Q49	Y	Y	Y	case	DAEC
14	Q51	Y	Y	Y	case	DAEC
15	Q56	Y	Y	Y	case	DAEC
16	Q65	Y	Y	Y	case	DAEC
17	B228	Y	Y	Y	case	EPEC
18	B56	Y	Y	Y	case	EPEC
19	B69	Y	Y	Y	case	EPEC
20	E162	Y	Y	Y	case	EPEC
21	E187	Y	Y	Y	case	EPEC
22	E205	Y	Y	Y	case	EPEC
23	Q233	Y	Y	Y	case	EPEC
24	Q300	Y	Y	Y	case	EPEC
25	R126	Y	Y	Y	case	EPEC
26	R135	Y	Y	**N**	case	EPEC
27	B109	Y	Y	Y	case	ETEC
28	B200	Y	Y	Y	case	ETEC
29	B244	Y	Y	Y	case	ETEC
30	B255	Y	Y	Y	case	ETEC
31	B295	Y	Y	Y	case	ETEC
32	B45	Y	Y	Y	case	ETEC
33	B62	Y	Y	Y	case	ETEC
34	B64	Y	Y	**N**	case	ETEC

35	B68	Y	Y	Y	case	ETEC
36	E184	Y	Y	Y	case	ETEC
37	Q294	Y	Y	Y	case	ETEC
38	Q53	Y	Y	Y	case	ETEC
39	Q101	Y	Y	N	control	NEG
40	Q105	Y	Y	N	control	NEG
41	Q116	Y	Y	N	control	NEG
42	Q127	Y	Y	N	control	NEG
43	Q131	Y	Y	N	control	NEG
44	Q157	Y	Y	N	control	NEG
45	Q158	Y	Y	N	control	NEG
46	R0015	Y	Y	N	control	NEG
47	R0022	Y	Y	N	control	NEG
48	R0026	Y	Y	N	control	NEG
49	R0080	Y	Y	N	control	NEG
50	R0081	Y	Y	N	control	NEG
51	R0091	Y	Y	N	control	NEG
52	R0105	Y	Y	N	control	NEG
53	R0130	Y	Y	N	control	NEG
54	R0134	Y	Y	N	control	NEG
55	R124	Y	Y	N	control	NEG
56	R129	Y	Y	N	control	NEG
57	R131	Y	Y	N	control	NEG
58	R25	Y	Y	N	control	NEG
59	R29	Y	Y	N	control	NEG
60	R40	Y	Y	N	control	NEG
61	R97	Y	Y	N	control	NEG
62	B027	N	Y	N	control	NEG
63	B103	N	Y	N	control	NEG
64	B104	N	Y	N	control	NEG
65	B213	N	Y	N	control	NEG
66	B22	N	Y	N	control	NEG
67	E131	N	Y	N	control	NEG
68	E141	N	Y	N	control	NEG
69	E156	N	Y	N	control	NEG
70	E17	N	Y	N	control	NEG
71	E204	N	Y	N	control	NEG
72	E21	N	Y	N	control	NEG
73	E23	N	Y	N	control	NEG
74	E56	N	Y	N	control	NEG

75	E93	N	Y	N	control	NEG
76	Q143	N	Y	N	control	NEG
77	Q144	N	Y	N	control	NEG
78	Q168	N	Y	N	control	NEG
79	R0041	N	Y	N	control	NEG
80	R0128	N	Y	N	control	NEG

Table S2. Metagenomic yield, human content and read quality of diarrhea and control samples used in this study. Estimates for metagenomic yield are presented for only one paired-end read (PE1). 'HC' stands for human read cleaning. 'QC' stands for quality control. Samples R126 (EPEC) and Q53 (ETEC) in red were discarded from the analysis due to low metagenomic coverage or sequencing depth.

Index	**Sample**	**case/ ctl**	***E. coli* pathotype**	**Total PE read1**	**# after HC**	**% after HC**	**# after QC**	**% after QC**	**Final lib size (PE1)**
1	B001	case	DAEC	6,531,793	1,488,444	22.8	1,132,843	17.3	169M
2	B24	case	DAEC	13,156,535	12,952,566	98.4	10,623,704	80.7	1.5G
3	B274	case	DAEC	7,427,929	7,290,205	98.1	5,842,124	78.7	872M
4	B89	case	DAEC	10,171,921	1,089,781	10.7	908,926	8.9	126M
5	E124	case	DAEC	10,843,598	10,838,811	99.9	9,169,045	84.6	1.4G
6	E158	case	DAEC	8,418,618	7,154,333	85.0	5,874,743	69.8	878M
7	E230	case	DAEC	8,794,848	8,790,739	100.0	7,276,368	82.7	1.1G
8	E26	case	DAEC	5,327,047	5,325,794	100.0	4,079,691	76.6	609M
9	E27	case	DAEC	8,600,912	7,489,003	87.1	6,232,480	72.5	933M
10	Q196	case	DAEC	10,565,635	1,114,583	10.5	846,647	8.0	123M
11	Q308	case	DAEC	10,107,007	10,083,099	99.8	8,578,875	84.9	1.3G
12	Q310	case	DAEC	6,927,136	6,914,116	99.8	5,376,592	77.6	805M
13	Q49	case	DAEC	7,273,409	7,270,208	100.0	5,746,337	79.0	855M
14	Q51	case	DAEC	10,549,942	948,390	9.0	766,198	7.3	113M
15	Q56	case	DAEC	10,155,142	7,471,039	73.6	6,260,221	61.6	916M
16	Q65	case	DAEC	8,922,179	6,087,254	68.2	5,074,850	56.9	759M
17	B228	case	EPEC	9,257,000	9,220,639	100	7,785,343	84.1	1.2G
18	B56	case	EPEC	10,822,204	9,224,643	85	7,593,577	70.2	1.1G
19	B69	case	EPEC	7,542,022	7,518,431	100	6,113,656	81.1	909M
20	E162	case	EPEC	6,938,905	4,948,559	71	3,997,479	57.6	598M
21	E187	case	EPEC	9,032,039	8,537,949	95	7,158,475	79.3	1.1G
22	E205	case	EPEC	7,036,190	7,034,607	100	5,384,029	76.5	811M
23	Q233	case	EPEC	8,942,343	8,941,264	100	7,252,948	81.1	1.1G
24	Q300	case	EPEC	9,834,934	9,772,091	99	8,233,965	83.7	1.2G
25	**R126**	**case**	**EPEC**	**11,185,025**	**404,832**	**4**	**316,760**	**2.8**	**47M**

26	R135	case	EPEC	10,490,440	4,523,998	43	3,866,783	36.9	581M
27	B109	case	ETEC	9,192,633	9,189,530	99.9	7,675,507	83.5	1.2G
28	B200	case	ETEC	12,509,793	12,494,918	99.9	10,313,659	82.4	1.5G
29	B244	case	ETEC	8,295,416	7,134,222	86	5,775,817	69.6	862M
30	B255	case	ETEC	9,856,355	9,854,564	99.9	8,432,852	85.6	1.3G
31	B295	case	ETEC	7,378,814	7,376,611	99.9	5,930,153	80.4	888M
32	B45	case	ETEC	11,491,641	11,487,903	99.9	9,823,847	85.5	1.4G
33	B62	case	ETEC	8,850,080	8,845,711	99.9	7,358,733	83.1	1.1G
34	B64	case	ETEC	8,999,675	8,990,531	99.9	7,637,068	84.9	1.1G
35	B68	case	ETEC	8,218,792	7,986,605	97	6,509,243	79.2	964M
36	E184	case	ETEC	6,314,505	6,287,420	100	5,175,784	82.0	771M
37	Q294	case	ETEC	13,322,967	13,312,996	100	11,105,678	83.4	1.7G
38	**Q53**	**case**	**ETEC**	**1,286**	**1,213**	**94.32**	**388**	**32**	**49K**
39	Q105	control	NEG	5,377,446	5,375,319	99.96	4,016,518	74.69	1.2G
40	Q127	control	NEG	7,778,804	7,776,111	99.97	6,304,665	81.05	1.9G
41	Q158	control	NEG	7,502,883	7,501,382	99.98	5,800,986	77.32	1.7G
42	Q116	control	NEG	8,405,358	8,402,475	99.97	6,737,952	80.16	2G
43	Q157	control	NEG	6,500,061	6,495,920	99.94	5,136,050	79.02	1.5G
44	Q101	control	NEG	9,119,586	9,117,396	99.98	7,822,285	85.77	2.3G
45	Q131	control	NEG	7,899,998	7,897,301	99.97	6,472,827	81.93	1.9G
46	R15	control	NEG	8,506,033	8,442,959	99.26	6,998,463	82.28	2.1G
47	R22	control	NEG	7,069,379	7,066,763	99.96	5,426,039	76.75	1.6G
48	R26	control	NEG	7,603,692	7,601,120	99.97	6,394,649	84.10	1.9G
49	R80	control	NEG	7,055,082	7,052,399	99.96	5,731,005	81.23	1.7G
50	R81	control	NEG	6,062,506	6,060,167	99.96	5,026,868	82.92	1.5G
51	R91	control	NEG	9,315,213	9,313,429	99.98	7,569,775	81.26	2.2G
52	R105	control	NEG	9,295,732	9,292,607	99.97	7,799,973	83.91	2.3G
53	R130	control	NEG	9,161,454	9,161,060	99.99	7,756,239	84.66	2.3G
54	R134	control	NEG	9,271,545	9,270,609	99.99	7,793,289	84.06	2.3G
55	R124	control	NEG	8,928,121	8,925,587	99.97	7,258,528	81.30	2.1G
56	R129	control	NEG	7,789,512	7,788,783	99.99	6,659,108	85.49	2G
57	R131	control	NEG	7,470,866	7,459,793	99.85	6,145,847	82.26	1.8G
58	R25	control	NEG	10,159,250	10,157,817	99.99	8,966,203	88.26	2.7G
59	R29	control	NEG	7,547,198	7,546,293	99.99	6,386,418	84.62	1.9G
60	R40	control	NEG	8,606,454	8,606,023	99.99	7,255,722	84.31	2.2G
61	R97	control	NEG	10,513,556	10,486,995	99.75	8,803,096	83.73	2.6G

Table S3. General genome assembly statistics of *E. coli* isolates and MAGs recovered from diarrheal samples. 'PATH' denotes pathotype group. Size is measured in Mbp. 'CONTG' corresponds to the number of assembled contigs. 'GC' corresponds to GC percent. 'COMP' denotes the estimated completeness percent of the isolate or MAG and 'CONT' corresponds to contamination percent. Isolates B56 and R126 in red were discarded from further analysis due to high contamination.

SAMPLE ID	PATH	TYPE	SIZE	CONTG	GC	N50	COVERAGE	COMP	CONT
Q56_BIN2	DAEC	MAG	4.9	113	50.7	146837	82.4X	99.59	1.27
E230_BIN3	DAEC	MAG	5.2	149	50.5	131672	90.9X	99.12	1.01
Q196_BIN1	DAEC	MAG	4.9	541	51.8	20774	23.4X	97.1	1.85
E124_BIN9	DAEC	MAG	4.7	244	51.3	34664	16.7X	96.91	0.99
Q51_BIN1	DAEC	MAG	4.1	1516	50.4	1516	25X	70.3	3.6
Q65_BIN3	DAEC	MAG	4.6	286	50.8	36969	76.7X	96.24	2.78
E158_BIN1	DAEC	MAG	3.7	861	51.9	6609	123X	76.23	2.06
R135_BIN8	EPEC	MAG	4.9	961	51.1	8631	11X	95.21	2.84
B45_BIN12	ETEC	MAG	4.6	179	50.9	52598	18.6X	98.84	0.55
B295_BIN5	ETEC	MAG	5.1	465	51.2	26676	12.3X	97.99	3.78
E184_BIN4	ETEC	MAG	4.7	530	51.1	16705	32.3X	94.37	2.6
B62_BIN13	ETEC	MAG	4.3	1343	51.8	2835	6.3X	70.2	5
B200_BIN15	ETEC	MAG	4.4	720	51.3	10164	12.X	93.24	1.98
Q294_BIN5	ETEC	MAG	4.3	611	51.1	13230	71X	90.93	2.48
B001_5	DAEC	Isolate	5.1	257	50.5	61824	32X	99.1	0.99
B24_1	DAEC	Isolate	4.8	253	49.7	55219	162X	98.6	1.06
B274_2	DAEC	Isolate	4.9	254	50	59071	335X	98.87	0.63
B89_1	DAEC	Isolate	4.8	194	50.6	68348	89X	98.35	0.42
E124_5	DAEC	Isolate	5.1	204	50.53	86732	36X	99.47	1.1
E158	DAEC	Isolate	5.2	202	50.59	77037	27X	99.56	0.84
E230_4	DAEC	Isolate	5.8	1630	50	5050	40X	91.89	8.29
E26	DAEC	Isolate	5.3	206	50.44	105975	28X	99.45	1.62
E27	DAEC	Isolate	5.1	142	50.2	103201	87X	99.59	0.92
Q196	DAEC	Isolate	5.1	309	50.42	34191	17X	99.24	0.59
Q308	DAEC	Isolate	4.9	192	50.66	66199	31X	99.2	0.74
Q310	DAEC	Isolate	5	205	50.54	59022	27X	99.17	0.78
Q49	DAEC	Isolate	5	173	50.59	111283	28X	99.43	0.48
Q51	DAEC	Isolate	5.2	252	50.52	55381	17X	99.51	1.65
Q56	DAEC	Isolate	5	137	50.63	118040	22X	99.5	1.1
Q65	DAEC	Isolate	5.1	1223	50.07	6466	28X	92.74	4.6
B228_2	EPEC	Isolate	5.1	192	49.7	80361	428X	99.47	0.69
B56	**EPEC**	**Isolate**	**7.1**	**2675**	**50.8**	**3195**	**66X**	**88.3**	**61.3**
B69_1	EPEC	Isolate	4.8	168	449.6	66315	78X	98.82	0.23
E162	EPEC	Isolate	4.7	148	50.65	89603	23X	99.47	0.39

E187	EPEC	Isolate	5	298	50.3	42003	38X	98.52	0.57
E205	EPEC	Isolate	4.5	131	50.58	64835	27X	99.17	0.3
Q233	EPEC	Isolate	4.8	625	50.42	13401	12X	97.27	1.46
Q300	EPEC	Isolate	4.5	131	49.6	64835	64X	99.17	0.3
R126	**EPEC**	**Isolate**	**5.3**	**141**	**54.8**	**92131**	**35X**	**88.9**	**19.7**
B109_1	ETEC	Isolate	4.7	403	50.5	22016	17X	99.19	0.73
B200_2	ETEC	Isolate	4.8	245	50.63	42883	19X	99.34	0.37
B244_3	ETEC	Isolate	4.8	201	48.3	66123	289X	99.34	0.55
B255_1	ETEC	Isolate	4.7	246	48.7	49462	167X	98.64	0.39
B295_2	ETEC	Isolate	4.7	255	48.6	49462	118X	98.37	0.39
B45	ETEC	Isolate	4.8	318	50.6	30625	26X	98.34	0.66
B62_5	ETEC	Isolate	4.8	274	50.5	43727	140X	98.96	0.55
B68_1	ETEC	Isolate	4.7	200	49.5	64686	352X	99.38	0.74
E184_3	ETEC	Isolate	5	182	49.7	88848	46X	99.52	1.29
Q294	ETEC	Isolate	4.8	215	50.52	53151	25X	98.64	0.56
Q53	ETEC	Isolate	5.2	1611	50.1	43.03	18X	88.57	7.63