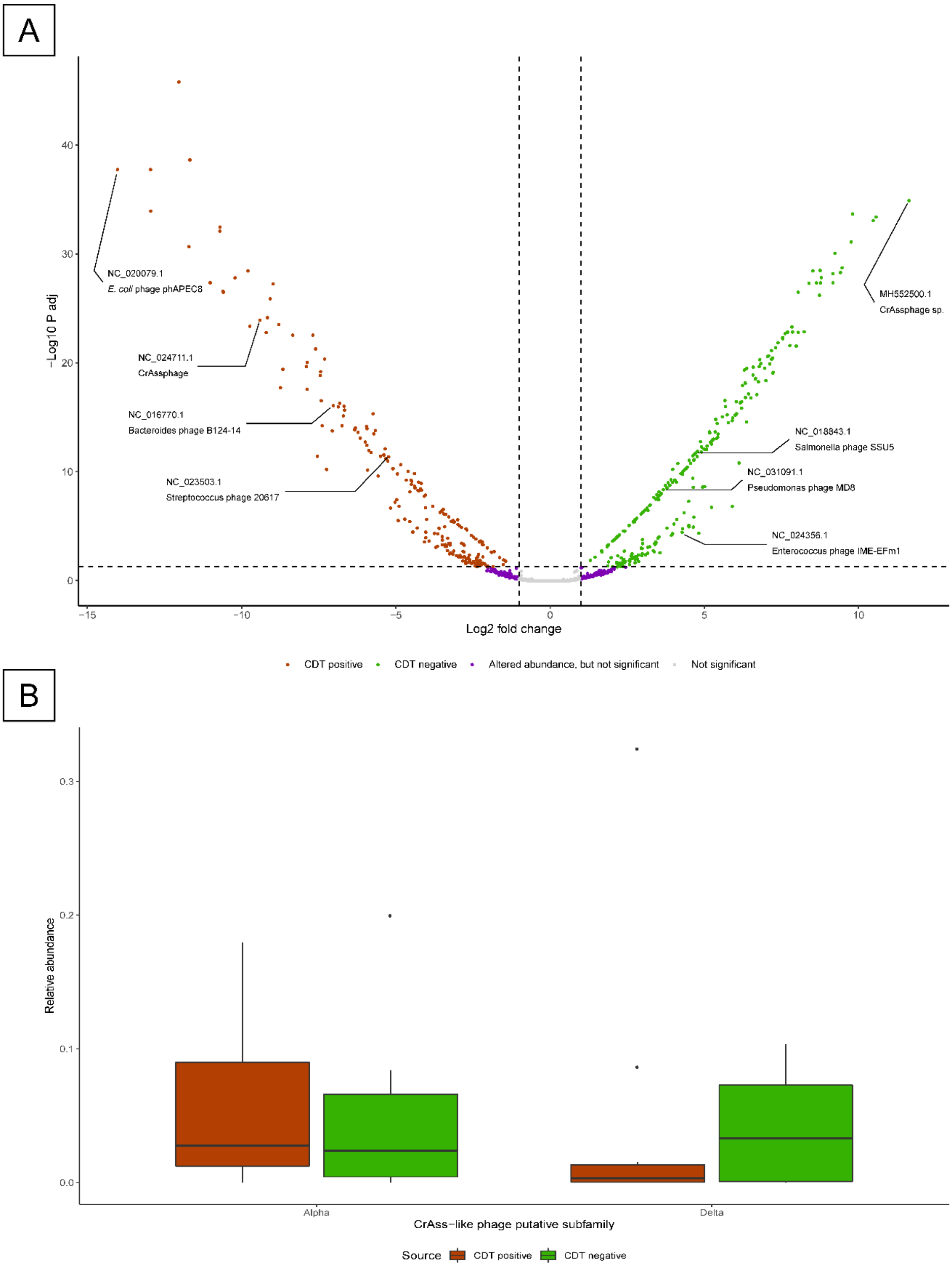
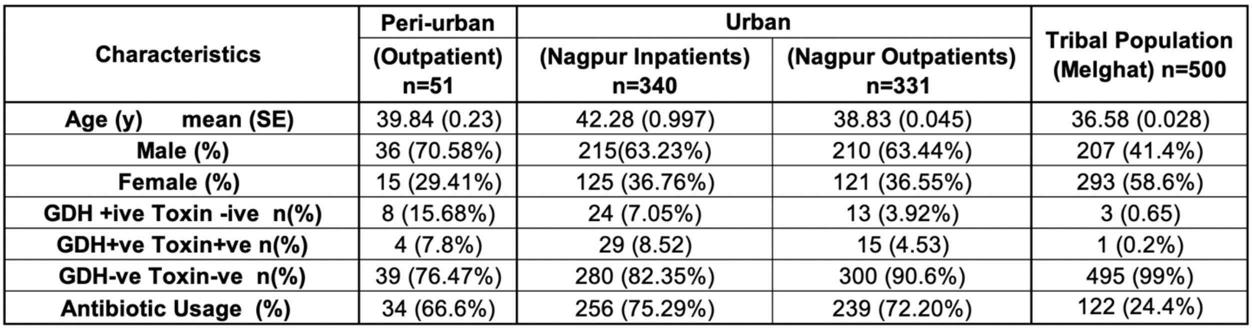
Supplementary Material

Supplementary Figure 1. Virome differences between diarrhoeal sufferers, which tested positive or negative for *C. difficile* toxin (CDT).



(A) Volcano plot of the fold change versus statistical significance of Viral Clusters (VCs). The top BLAST hit, and accession number, for several VCs are highlighted. **(B)** Examination of the relative abundance of putative crAss-like phage subfamilies alpha and delta, predicted to infect *Bacteroides* and *Prevotella*, respectively. No statistical differences were observed between crAss-like phage abundances by CDT status or geographical residence (Kruskal-Wallis test; CDT status: p-values 0.6 and 0.47, respectively; Geography: p-values 0.76 and 0.059, respectively).

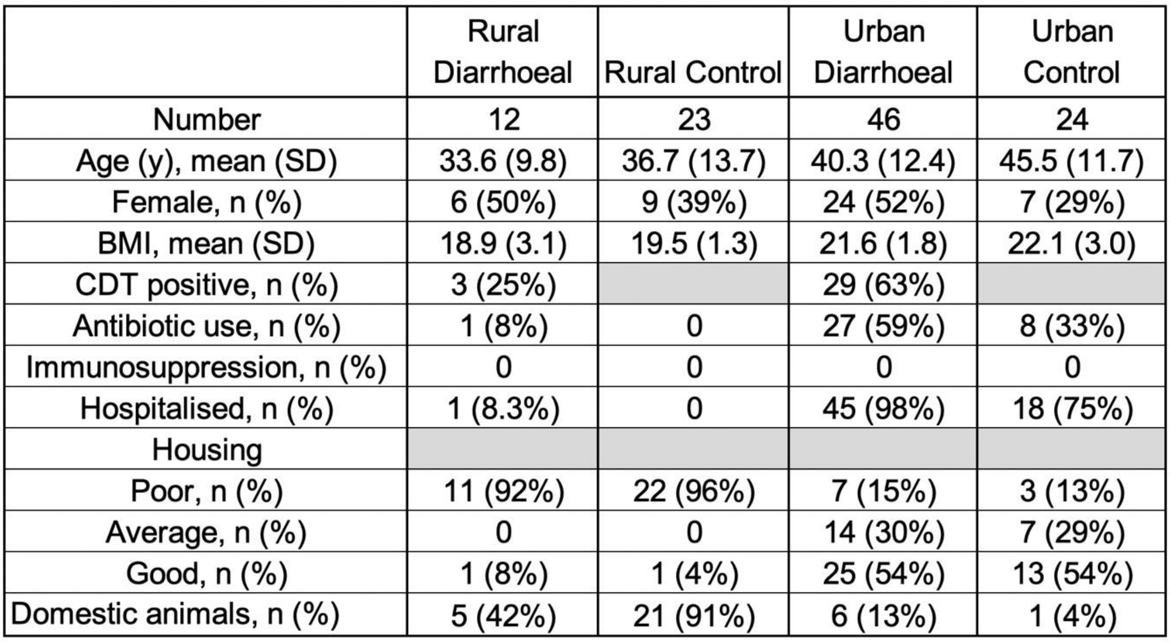
SUPPLEMENTARY TABLE 1



Characteristics	Peri-urban	Urban		Tribal Population (Melghat) n=500
	(Outpatient) n=51	(Nagpur Inpatients) n=340	(Nagpur Outpatients) n=331	
Age (y) mean (SE)	39.84 (0.23)	42.28 (0.997)	38.83 (0.045)	36.58 (0.028)
Male (%)	36 (70.58%)	215(63.23%)	210 (63.44%)	207 (41.4%)
Female (%)	15 (29.41%)	125 (36.76%)	121 (36.55%)	293 (58.6%)
GDH +ive Toxin -ive n(%)	8 (15.68%)	24 (7.05%)	13 (3.92%)	3 (0.65)
GDH+ve Toxin+ve n(%)	4 (7.8%)	29 (8.52)	15 (4.53)	1 (0.2%)
GDH-ve Toxin-ve n(%)	39 (76.47%)	280 (82.35%)	300 (90.6%)	495 (99%)
Antibiotic Usage (%)	34 (66.6%)	256 (75.29%)	239 (72.20%)	122 (24.4%)

Baseline Cohort Characteristics for *C. difficile* Prevalence Study

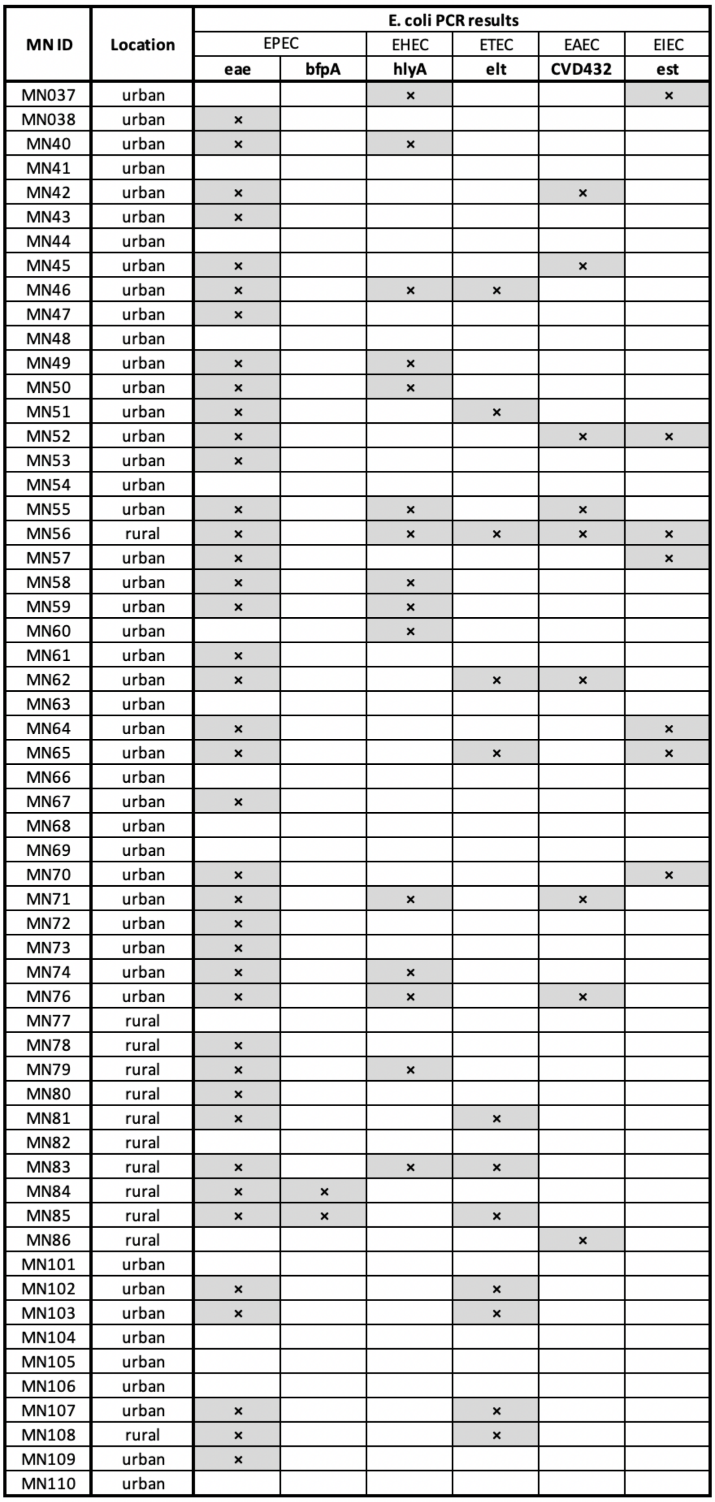
SUPPLEMENTARY TABLE 2



	Rural Diarrhoeal	Rural Control	Urban Diarrhoeal	Urban Control
Number	12	23	46	24
Age (y), mean (SD)	33.6 (9.8)	36.7 (13.7)	40.3 (12.4)	45.5 (11.7)
Female, n (%)	6 (50%)	9 (39%)	24 (52%)	7 (29%)
BMI, mean (SD)	18.9 (3.1)	19.5 (1.3)	21.6 (1.8)	22.1 (3.0)
CDT positive, n (%)	3 (25%)		29 (63%)	
Antibiotic use, n (%)	1 (8%)	0	27 (59%)	8 (33%)
Immunosuppression, n (%)	0	0	0	0
Hospitalised, n (%)	1 (8.3%)	0	45 (98%)	18 (75%)
Housing				
Poor, n (%)	11 (92%)	22 (96%)	7 (15%)	3 (13%)
Average, n (%)	0	0	14 (30%)	7 (29%)
Good, n (%)	1 (8%)	1 (4%)	25 (54%)	13 (54%)
Domestic animals, n (%)	5 (42%)	21 (91%)	6 (13%)	1 (4%)

Baseline Cohort Characteristics for Metagenome Cohort

SUPPLEMENTARY TABLE 3



MN ID	Location	E. coli PCR results					
		EPEC		EHEC	ETEC	EAEC	EIEC
		eae	**bfpA**	**hlyA**	**elt**	**CVD432**	**est**
MN037	urban			×			×
MN038	urban	×					
MN40	urban	×		×			
MN41	urban						
MN42	urban	×				×	
MN43	urban	×					
MN44	urban						
MN45	urban	×				×	
MN46	urban	×		×	×		
MN47	urban	×					
MN48	urban						
MN49	urban	×		×			
MN50	urban	×		×			
MN51	urban	×			×		
MN52	urban	×				×	×
MN53	urban	×					
MN54	urban						
MN55	urban	×		×		×	
MN56	rural	×		×	×	×	×
MN57	urban	×					×
MN58	urban	×		×			
MN59	urban	×		×			
MN60	urban			×			
MN61	urban	×					
MN62	urban	×			×	×	
MN63	urban						
MN64	urban	×					×
MN65	urban	×			×		×
MN66	urban						
MN67	urban	×					
MN68	urban						
MN69	urban						
MN70	urban	×					×
MN71	urban	×		×		×	
MN72	urban	×					
MN73	urban	×					
MN74	urban	×		×			
MN76	urban	×		×		×	
MN77	rural						
MN78	rural	×					
MN79	rural	×		×			
MN80	rural	×					
MN81	rural	×			×		
MN82	rural						
MN83	rural	×		×	×		
MN84	rural	×	×				
MN85	rural	×	×		×		
MN86	rural					×	
MN101	urban						
MN102	urban	×			×		
MN103	urban	×			×		
MN104	urban						
MN105	urban						
MN106	urban						
MN107	urban	×			×		
MN108	rural	×			×		
MN109	urban	×					
MN110	urban						

Distribution of Diarrhoeagenic *E. coli* (DEC) Pathotypes Among the Patients with Diarrhoea in the Metagenome Study

EPEC = enteropathogenic *Escherichia coli*; EHEC; enterohaemorrhagic *Escherichia coli*; ETEC = enterotoxigenic *Escherichia coli*; EAEC = enteroaggregative *Escherichia coli*; EIEC = enteroinvasive *Escherichia coli*

SUPPLEMENTARY TABLE 4

Taxa	**Fold Change Urban vs Rural**	**FDR p.corr**
Roseburia intestinalis	0.02	1.9×10^{-7}
Bacteroides intestinalis	29.26	1.7×10^{-5}
Megamonas funiformis	28.58	0.0001
Roseburia hominis	0.07	0.0007
Ruminococcus bromii	0.06	0.0014
Parabacteroides unclassified	11.50	0.0034
Lactobacillus ruminis	0.28	0.0083
Parabacteroides distasonis	9.23	0.0084
Catenibacterium mitsuokai	0.11	0.0125
Bacteroides thetaiotaomicron	8.42	0.0125
Veillonella parvula	8.94	0.0186
Enterobacter cloacae	8.28	0.0192
Subdoligranulum unclassified	3.98	0.0192
Prevotella stercorea	0.38	0.0225
Acidaminococcus intestini	0.12	0.0225
Megamonas hypermegale	9.28	0.0242
Haemophilus parainfluenzae	0.13	0.0252
Alistipes unclassified	0.15	0.0252
Escherichia unclassified	5.43	0.0342
Veillonella dispar	6.15	0.0422
Bacteroides vulgatus	2.49	0.0422
Alistipes finegoldii	6.01	0.0422
Paraprevotella clara	6.04	0.0422
Megamonas unclassified	7.32	0.0422
Bacteroides uniformis	5.31	0.0422
Enterococcus faecium	7.45	0.0422

Bacterial taxa enriched in Urban vs Rural subjects

SUPPLEMENTARY TABLE 5

Taxa	**Fold Change** **Diarrhoeal vs Non-diarrhoeal**	**FDR p.corr**
Eubacterium siraeum	0.04	1.81×10^{-8}
Mitsuokella multacida	0.05	1.44×10^{-7}
Barnesiella intestinihominis	0.09	0.0002
Collinsella aerofaciens	0.54	0.0017
Bacteroides ovatus	5.35	0.0050
Prevotella copri	0.35	0.0050
Bacteroides dorei	6.45	0.0069
Bacteroides uniformis	4.39	0.0132
Enterococcus faecium	5.48	0.0132
Ruminococcus bromii	0.20	0.0132
Bifidobacterium angulatum	0.23	0.0205
Bacteroidales bacterium ph8	0.22	0.0223
Subdoligranulum unclassified	0.40	0.0246
Clostridium nexile	4.64	0.0275
Roseburia intestinalis	4.16	0.0417
Bacteroides xylanisolvens	3.62	0.0448

Bacterial taxa enriched in Diarrhoeal vs Non-diarrhoeal subjects

SUPPLEMENTARY TABLE 6

Taxa	**Fold Change** **CDT Positive vs Negative**	**FDR p.corr**
Coprobacillus unclassified	35.23	1.72×10^{-7}
Bacteroides ovatus	17.28	6.08×10^{-6}
Lachnospiraceae bacterium 2-1-58FAA	22.90	1.42×10^{-5}
Megamonas unclassified	0.05	0.0006
Catenibacterium mitsuokai	0.07	0.0009
Bacteroides fragilis	3.59	0.0017
Eubacterium eligens	0.09	0.0019
Enterococcus faecium	0.10	0.0046
Eubacterium rectale	0.32	0.0046
Barnesiella intestinihominis	0.11	0.0049
Bifidobacterium adolescentis	0.10	0.0060
Bifidobacterium breve	8.40	0.0113
Ruminococcus gnavus	7.71	0.0141
Alistipes indistinctus	8.35	0.0149
Bacteroides eggerthii	7.58	0.0190
Parabacteroides distasonis	5.85	0.0296
Dialister succinatiphilus	0.14	0.0335
Bacteroides intestinalis	5.95	0.0493

Bacterial taxa enriched in *C. difficile* toxin (CDT) positive subjects

SUPPLEMENTARY TABLE 7

Pathway	Rural Associated Taxa				Urban Associated Taxa				Total Rural	Total Urban
	Ps	**Pc**	**Er**	**Rb**	**Ec**	**Kp**	**Bv**	**Pd**		
alpha-Linolenic acid metabolism	0	0	1	0	5	6	0	1	1	12
Benzoate degradation	1	2	5	3	6	33	3	4	11	46
Drug metabolism - cytochrome P450	0	0	1	0	8	11	1	1	1	21
Fluorobenzoate degradation	0	0	0	0	2	8	0	0	0	10
Furfural degradation	0	0	0	0	0	0	0	0	0	0
Lipoic acid metabolism	2	0	0	0	3	3	2	3	2	11
Penicillin and cephalosporin biosynthesis	0	0	0	0	0	0	2	0	0	2
Steroid degradation	0	0	0	0	1	1	0	0	0	2
Styrene degradation	0	0	0	0	0	3	0	0	0	3
Xylene degradation	0	0	2	1	0	9	1	0	3	10

Total kegg orthology (KO) gene copy numbers per pathway enriched in urban subjects for representative taxa from each group.

Ps = *Prevotella stercorea*, Pc = *Prevotella copri,* Er = *Eubacterium rectale*, Rb = *Ruminococcus bromii,* Ec = *Escherichia coli,* Kp = *Klebsiella pneumoniae*, Bv = *Bacteroides vulgatus,* Pd = *Parabacteroides distasonis*.

SUPPLEMENTARY TABLE 8

Pathway	**Rural Associated Taxa**				**Urban Associated Taxa**				**Total Rural**	**Total Urban**
	Ps	**Pc**	**Er**	**Rb**	**Ec**	**Kp**	**Bv**	**Pd**		
Alanine, aspartate and glutamate metabolism	19	29	23	19	27	41	29	36	90	133
Arginine and proline metabolism	11	22	24	13	41	67	24	35	70	167
Histidine metabolism	5	9	11	12	10	16	13	15	37	54
Peptidoglycan biosynthesis	18	18	17	17	23	26	15	20	70	84
Polyketide sugar unit biosynthesis	5	4	4	2	5	10	4	4	15	23
Starch and sucrose metabolism	18	34	36	20	36	49	31	49	108	165
Streptomycin biosynthesis	7	7	6	4	20	18	10	13	24	61
Terpenoid backbone biosynthesis	14	15	13	10	16	17	13	17	52	63
Thiamine metabolism	10	12	14	10	13	16	10	8	46	47
Valine, leucine and isoleucine biosynthesis	1	11	13	11	17	20	9	15	36	61

Total kegg orthology (KO) gene copy numbers per pathway enriched in rural subjects for representative taxa from each group.

Ps = *Prevotella stercorea*, Pc = *Prevotella copri,* Er = *Eubacterium rectale*, Rb = *Ruminococcus bromii*, Ec = *Escherichia coli,* Kp = *Klebsiella pneumoniae*, Bv = *Bacteroides vulgatus*, Pd = *Parabacteroides distasonis*.

Supplementary Methods

PCR amplification

Extracted faecal DNA samples were used as a template for amplification of 5 pathotypes of diarrheagenic *E.coli* (DEC) targeting the virulence genes including *eae* and *bfpA* for Enteropathogenic *E. coli (*EPEC), *hlyA* for Enterohaemorrhagic E. coli (EHEC), *elt* for Enterotoxigenic *E. coli* (ETEC), *CVD432* for Enteroaggregative *E. coli* (EAEC) and *est* for Enteroinvasive *E. coli* (EIEC) by 2 different species-specific primer sets available in the commercial *E.coli* detection kit (HiMedia laboratories Pvt. Ltd, India). The PCR reaction mixture was prepared as described below in Supplementary table 9. After a quick 10 sec spin at 6000 rpm place, the reaction mixtures were subjected to the recommended PCR program cycling conditions below (Supplementary Table 10). The 1st primer set targets *eae* (482 bp), *bfpA* (300) bp and *CDV432* (194) bp and 2nd primer set targets *hlyA*, *elt* and *est* at 534bp, 322bp and 170bp respectively. Additionally, two non-template controls were included. Amplification reactions were carried out using proflex 96- Well Thermal Cycler (Applied Biosystems) and the amplified targets were detected by agarose gel electrophoresis (Supplementary Figure 2).

Supplementary Table 9

Components	**Tube 1 (µL)**	**Tube 2 (µL)**
2X PCR Taq Mixture	12.5	12.5
Primer set 1	6	-
Primer set 2	-	7.5
Molecular Biology grade water	1.5	-
Template DNA	5	5
Total Volume	25	25

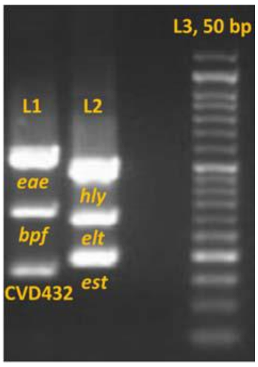
PCR mastermix method for detection of diarrheagenic *E. coli*

Supplementary Table 10

Initial denaturation	94˚C for 10 min	
Denaturation	94˚C for 45 sec	30 Cycles
Annealing	56˚C for 30 sec	
Extension	72˚C for 30 sec	
Final Extension	94˚C for 10 min	
Hold	4˚C for infinity	

Thermocycler protocol

Supplementary Figure 2



Lane no. 1 – Primer Set 1
Lane no. 2 – Primer Set 2
Lane no. 3 – 50 bp ladder

Primer Set 1
eae – 482 bp
bfpA – 300 bp
CVD432 – 194 bp

Primer Set 2
hlyA – 534 bp
elt – 322 bp
est – 170 bp

Amplified diarrhoeagenic *E. coli* virulence gene targets by agarose gel electrophoresis