Metagenomics Reveals Impact of Urbanisation in Central India on the Human Gut Microbiome and its Antimicrobial Resistance Profiles

Tanya Monaghan (tanya.monaghan@nottingham.ac.uk)
University of Nottingham https://orcid.org/0000-0001-7622-3997

Tim J. Sloan
University of Nottingham School of Life Sciences

Stephen R. Stockdale
APC Microbiome

Adam M. Blanchard
University of Nottingham

Richard D. Emes
University of Nottingham

Mark Wilcox
University of Leeds

Rima Biswas
Central India Institute of Medical Sciences

Rupam Nashine
Central India Institute of Medical Sciences

Sonali Manke
Central India Institute of Medical Sciences

Jinal Gandhi
Central India Institute of Medical Sciences

Pratishtha Jain
Central India Institute of Medical Sciences

Shrejal Bhotmange
Central India Institute of Medical Sciences

Shrikant Ambalkar
Sherwood Forest Hospitals NHS Foundation Trust

Ashish Satav
Mahatma Gandhi Tribal Hospital

Lorraine A. Draper
APC Microbiome

Colin Hill

APC Microbiome
Rajpal Singh Kashyap
Central India Institute of Medical Sciences

Research

Keywords: Gut microbiome, antibiotic resistome, virome, diarrhoea, Clostridioides difficile, Central India

Posted Date: November 14th, 2019

DOI: https://doi.org/10.21203/rs.2.17205/v1

Abstract

Background The impact of the rapid urbanisation of low- and middle-income countries on the human gut microbiome remains grossly understudied. Whilst the effect of urbanisation on the bacterial populations of the human gut microbiome have been documented, little is known about the influence of diet and antibiotics on the bacteriome, its virome, and antibiotic resistome. Here, we use shotgun metagenomics to comprehensively characterise the bacterial and viral fractions of the human gut microbiome, and their encoded functions, from two divergent Central Indian populations (rural agriculturalists from Melghat and an urban population in Nagpur). Additionally, we investigate cohorts with and without diarrhoea, and the potential burden of Clostridioides difficile, associated with widespread unregulated use of antibiotics in India.

Results We observed distinct rural-urban differences in the gut microbiome, including viral diversity and composition, with geography exhibiting a greater influence than diarrhoeal status. Urban microbiomes were enriched in metabolic pathways responsible for degradation of drugs and organic compounds, which were predicted to relate to replacement of rural-enriched Prevotella spp. and fermentative Clostridiales with Enterobacteriaceae and Bacteroides spp. By linking phages present in the microbiome to their bacterial hosts through CRISPR spacers, a shift from Prevotella- and Eubacterium-infecting phages to Bacteroides- and Parabacteroides-infecting phages was observed in rural and urban populations, respectively. Additionally, the auxiliary metabolic potential of rural-associated phage populations was enriched for carbon and amino acid energy harvesting potential, compared to urban-associated phages. A core set of antimicrobial resistance genes was identified in both populations, particularly those conferring resistance to macrolides, tetracyclines and 1stgeneration cephalosporins, with the majority also showing evidence of resistance to fluoroquinolones, aminoglycosides and sulphonamides. In a subgroup of urban subjects with diarrhoea and high antibiotic exposure, most of whom tested positive for C. difficile toxin, evidence of resistance to fosfomycin, glycopeptides, daptomycin, 3rd generation cephalosporins and carbapenems was widespread.

Conclusions We report distinct differences in antimicrobial resistance gene profiles as well as a marked variation in the burden of C. difficile disease between rural and urban populations. The key drivers of variation in urban and rural Indian microbiomes are geography, diet, industrial and healthcare exposures.

Background

The human gut houses a complex microbial ecosystem referred to as the microbiome, which includes prokaryotic, eukaryotic and viral components. While the bacterial components of the microbiome have received considerable attention, comparatively little is known about the composition and physiological significance of human gut-associated bacteriophage populations, otherwise known as the phageome [1]. Moreover, despite the growing global burden of antibiotic resistance to modern health care, very few studies have directly [2-3] or indirectly,(through analysing urban sewage) [4] examined the antibiotic resistomes of human faecal metagenomes. Such paucity of data prevents a complete understanding of

the global burden and transmission of antimicrobial resistance (AMR), which is essential to support national and global priority setting, public health actions, and treatment decisions. Although recent years have seen an explosion of gut microbiome studies in rural pre-industrialised societies such as hunter-gatherer and other geographically diverse populations [5-10], little is known about microbial variability and its implications for health and disease in other underrepresented populations in South America, Africa, and regions in Asia, particularly India, where there is a scarcity of microbiome data in diarrhoeal and other populations. Diarrhoeal diseases are a major cause of morbidity and mortality in India, making identification of aetiological agents of utmost importance.

In India, there is tremendous opportunity to study highly diverse communities with varied geographic distribution, dietary habits and socioeconomic stratification. Some of these communities, including a large tribal population, remain dependent on hunting, agriculture and fishing with their own culture, tradition, dietary habits, language and genetic make-up. Recently, studies have begun to explore the Indian gut microbiome including that of the country's scheduled tribes, principally using 16S rRNA gene amplicon sequencing methods to profile mainly gut bacterial diversity in rural and urban healthy populations [11-14] with only a few reports employing whole-genome shotgun metagenomic sequencing approaches [15-16]. Whilst the majority of the aforementioned studies have analysed small population cohorts from Northern, Southern and Western Indian territories, there is a dearth of information characterising the gut microbiomes of Central Indian populations. Furthermore, little is known about the burden of *Clostridioides difficile* infection (CDI) in India, the leading worldwide cause of antibiotic-associated diarrhoea in hospitalised and community populations [17-21] and its impact on Indian metagenomes. Profligate, unregulated antibiotic use and inappropriate prescribing suggest that CDI could be widespread in India, the world's largest consumer of antibiotics.

Via a pre-existing research partnership between the University of Nottingham and the Central India Institute of Medical Sciences (CIIMS), we were able to define the gut bacteriome, antibiotic resistome and virome in understudied rural and urban diarrhoeal and control populations in Central India. CIIMS has established multisite links with several hospital laboratories in the surrounding district of Nagpur, as well as a satellite laboratory in the Mahatma Gandhi Tribal hospital, Melghat, home to the Korku tribe of agriculturalists. We also concentrated on the pathogen *Clostridiodies difficile*, and assessed its prevalence as well as impact on the gut microbiome.

Our results indicate that the rural habitants of Melghat show a *Prevotella*-dominant microbiome compared with the urban population of Nagpur, which is enriched with *Bacteroides spp*. Urbanisation is associated with functional enrichment of genes involved in xenobiotic and lipid metabolism. Although a core set of AMR genes are detectable in the Korku population, Nagpurian urbanites display a much higher burden of AMR overall. Viral diversity and composition is more influenced by geography than diarrhoeal status, with urban- and rural-specific phage populations linked to bacterial hosts through CRISPR spacer identification. *C. difficile* is principally detected in the urban and peri-urban exposed antibiotic populations, many of which carry AMR genes to virtually every class of antibiotic.

Results

Cohort Characteristics

We initially screened 1222 Central Indian urban (Nagpur inpatients n=340; outpatients n=331), peri-urban (outpatients n=51) and rural (Melghat non-hospital exposed n=500) participants presenting with diarrhoea for the presence of toxigenic *C. difficile* (Supplementary Table 1). Our results clearly demonstrate higher *C. difficile* toxin positivity rates in the urban inpatient (8.5%) and outpatient (4.5%) as well as peri-urban outpatient groups (7.8%) compared to that seen in the rural Melghat population (0.2%), a finding that likely reflects less frequent exposure to antibiotics and hospitalisation in the latter. We also observed GDH positive toxin negative cases across all groups, with highest rates of non-toxigenic *C. difficile* carriage seen in the Nagpurian peri-urban outpatient (15.6%), urban hospitalised (7%) and urban outpatient participants (3.9%), compared to the rural community population (0.6%).

For our faecal metagenome study in which we were comparing urban vs rural microbiome profiles and assessing impact of diarrhoea and CDI, we analysed faecal samples collected from 105 Central Indian participants comprising 35 rural (12 with diarrhoea) and 70 urban (46 with diarrhoea) participants from Melghat and Nagpur districts, respectively (Supplementary Table 2). We selected an enriched set of faecal DNA samples derived from diarrhoeal samples that had previously tested positive in our aforementioned diagnostic *C. difficile* immunoassays for whole-genome shotgun sequencing (WGS). Of these diarrhoeal samples, 63% (29/46; urban) and 25% (3/12; rural) had tested positive for toxigenic *C. difficile* in the C. DIFF QUIK CHEK assay.

In addition to testing for *C. difficile*, and in considering the emergence of diarrhoeagenic *Escherichia. coli* in India, diarrhoeal samples were also assayed for the presence of pathogenic *E. coli* with a multiplex PCR assay targeting the following pathotypes; Enteropathogenic (EPEC, *eae* and *bfpA* genes), Enterohemorrhagic (EHEC, *hlyA*), Enterotoxigenic (ETEC, *elt*), Enteroaggregative (EAEC, *CVD432*) and Enteroinvasive (EIEC, *est*). Overall 23/46 (50%) urban and 8/12 (75%) rural diarrhoeal subjects tested positive for at least one pathotype (Supplementary Table 3) with EHEC being the most commonly identified (15/31) followed by ETEC (12/31), and EAEC (9/31).

Stool samples received centrally by CIIMS were collected at recruitment over 13 months from the 1st of March 2017 to 30th April 2018 from participants resident at 48 sites in Nagpur district (Figure 1A-C) and 19 participating rural villages in Melghat (Figure 2A-C), 3 of which were very small villages and are not marked on Google maps. The mean duration of diarrhoea for urban diarrhoeal group (n=34) was 5.2 days (SD 2.7 days). The mean age of participants was greater for urban (42 years) versus rural (35.6 years) participants, p=0.01, with a lower percentage of females represented in the urban and rural control groups compared to the diarrhoeal groups which did not reach statistical significance. Mean body mass index (BMI) [weight (kg)/height (m) squared] was also higher in the urban (21.8) compared with rural (19.3) participants group, p<0.0001). It was noteworthy that one third of participants in the urban non-diarrhoeal control group had received antibiotics in the three months prior to recruitment, although none were taking

antibiotics when sampled. The vast majority of participant housing in the rural areas was deemed to be of poor quality based on a lack of piped water supply (water tank only), no access to latrines, limited electricity supply (<18 hours/day) and small living space (Figure 2C), whereas just over half of the urban cohort resided within housing of good quality, as reflected in access to Corporation tap water, longer duration electricity supply (>18 hours/day) and larger living quarters (Figure 1B). A higher proportion of rural participants kept domestic animals within their living quarters (cattle, goats, chickens) compared with their urban counterparts.

Overall, significant confounding associations were observed between geographic location and several other study variables. Consequently, we focussed our analyses primarily on geographic location, with the understanding this accounts for both subject specific and environmental factors.

Dietary Information for Sampled Cohorts

As donor participants provided samples to CIIMS from geographically dispersed sites across Nagpur (Figure 1A) and Melghat (Figure 2A), it was not possible to systematically administer customised and standardised food frequency questionnaires to each participant. However, it was possible to elicit the major constituents of the inpatient diets based on knowledge of the principal foods provided within the hospital sector. A typical oral dietary hospital regime consisted of a morning beverage (tea, coconut water, fruit juice or lemonade), a lunchtime choice of oatmeal, rice porridge (semi-solid preparation of rice and cumin seed with coriander or basil leaves), semolina, curd rice or dal khichdi (1:2 proportion of pulses and rice added to water with salt and tumeric), an evening meal of black tea with Sago Kheer (sweet pudding made with tapioca pearls or sabudana and milk), and a late dinner which replicated the lunch menu. Beyond the hospital environment, the typical Nagpurian diet is vegetarian predominant and consists of a diverse mix of fruits, vegetables, grains, non-saturated fats and proteins. In contrast, the dietary repertoire of the rural Korku tribal participants is considerably narrower and typically consists of locally available plant-based foods rich in carbohydrates and high in fibre but low in protein content such as jowar ki roti made from millet flour and water in combination with various types of vegetable chutneys containing garlic, salt and green chillies. They generally feed twice daily, rarely consuming milk or meat, and usually eat the leftover food from the previous day.

Rural subjects have a distinct microbiome when compared with urban subjects

Principal coordinates analysis was performed on a Bray-Curtis Dissimilarity matrix of the species-level taxonomic profiles (n=105), excluding viral taxa. Urban (n=70) and rural (n=35) subjects separated well along the 1st principal component (Figure 3A) but diarrhoeal status (control n=47 vs. diarrhoeal n=58) did not appear to have as much influence on sample clustering. This observation was confirmed by PERMANOVA which indicated that geographic location (urban vs rural) accounted for 7.7% of the

variation between samples (F=8.67, p=0.001) while diarrhoeal status accounted for a further 1.7% (F=1.94, p=0.028). Including *C. difficile* toxin status and recent antibiotic exposure in the model accounted for an additional 2.1% (F=2.48, p=0.005) and 1.4% (F=1.62, p=0.09) of variation respectively. Considering other demographic variables of interest, including age, gender, BMI, housing quality and animal ownership when combined with geography, only age (2.1%, F=2.41, p=0.008) contributed significantly to the residual variation explained, reflecting the strong association of these variables with study location.

Sample alpha diversity was calculated using the Inverse Simpson Index for the taxonomic abundances at species level and compared between control and diarrhoeal subjects from either an urban or rural location (Figure 3B). Rural diarrhoeal subjects had the lowest diversity (n=12, mean 3.66 +/- 2.5) which was significantly lower than urban control subjects who had the highest diversity (n=24, 6.75 +/- 3.5, p=0.05).

Individual taxonomic profiles showed a high level of heterogeneity at genus level both within and between study groups (Figure 3C). Overall, profiles from urban areas tended to be dominated by *Bacteroides spp.* while profiles from rural areas had much lower abundance of *Bacteroides spp.* but *Prevotella spp.* were predominant, particularly in control subjects.

Analysing the species-level taxonomic abundances using generalized linear models yielded 26 taxa which differed significantly between rural and urban control subjects, and 16 taxa which differed significantly between control and diarrhoeal subjects (Figure 3D, Supplementary Tables 4 & 5). A direct comparison was also made between diarrhoeal subjects testing positive and negative for *C. difficile* toxin, yielding 18 taxa which differed significantly (Supplementary Table 6).

Antimicrobial resistance is more prevalent in urban areas

Antimicrobial resistance gene profiles were compiled from the faecal metagenomes of all subjects in the study using ARIBA. Individual gene counts were aggregated by antibiotic class to identify broad trends between subjects according to geographic location and antibiotic exposure (Figure 4A). Genes conferring resistance to beta-lactam antibiotics, tetracyclines and macrolides, lincosamides and streptogramins (MLS) were identified in virtually all subjects. Hierarchical clustering analysis identified a group of largely rural subjects with low antibiotic exposure whose faecal metagenomes had few other antibiotic resistance genes identified apart from for these 3 classes (Figure 4A). Conversely, a group of urban subjects with high antibiotic exposure, many of whom were also *C. difficile* toxin positive, were carrying antibiotic resistance genes to virtually every class of antibiotic. This included resistance to glycopeptides (predominantly *vanA* genes) and two classes from the World Health Organisation essential medicines reserve group; fosfomycin and lipopeptides (daptomycin). Compared with other antibiotic classes, metronidazole resistance was rare and only detected in a single subject.

Beta lactam antibiotics are widely used in clinical practice and resistance to broad spectrum beta lactam antibiotics, particularly carbapenems, is of significant public health concern. Individual beta lactam gene clusters were analysed in more detail by subject to identify differences between study variables, particularly geography (Figure 4B). Resistance mechanisms included production of beta-lactamases (Ambler class A to D), alteration of penicillin binding proteins (PBPs) and mutation of outer membrane porins in Gram negative bacteria. The *cfx* gene cluster, encoding an Ambler class A beta-lactamase, was the most prevalent cluster detected, identified in 94 of 105 subjects. Other beta lactam gene clusters varied markedly between urban and rural areas, with 9 clusters identified as differing significantly in abundance between groups (Figure 4C). These clusters included *ctx* genes (rural gene abundance 1.54 +/- 3.1 vs urban 10.7 +/- 15.2, p.corr=9.5 x 10^{-6}) which encode extended spectrum beta-lactamases (ESBLs), and the gene encoding the Ambler class B metallo-beta-lactamase NDM-1 which was detected in only 1 of 35 rural subjects but 32/70 urban subjects (rural 0.03 +/- 0.2 vs urban 0.56 +/- 0.67, p.corr=1.13 x 10^{-4}).

Microbiota variations between groups are predicted to drive functional shifts in metabolic pathways

Differentially abundant metabolic pathways between urban and rural subjects and their predicted taxonomic contributions were identified with FishTaco (Figure 5). A total of 28 pathways were enriched in urban subjects, with the majority (24/28) in the following categories; xenobiotics biodegradation and metabolism (16/28), lipid metabolism (6/28) and amino acid metabolism (2/28). Several *Bacteroides spp.*, *Parabacteroides distasonis, Klebsiella pneumoniae* and *E. coli* were identified as potential contributors to the enrichment of these pathways in urban subjects.

Of the 33 pathways enriched in rural subjects, 13/33 related to metabolism of amino acids, 4/33 to carbohydrate metabolism and 4/33 to metabolism of cofactors and vitamins. *Prevotella copri*, *Prevotella stercorea* and several members of the *Firmicutes* phylum, including *Ruminococcus bromii*, *Eubacterium rectale* and *Faecalibacterium prausnitzii*, were identified as potentially important contributors to the enrichment of these pathways in rural subjects, counterbalanced by the presence of *Parabacteroides distasonis* in urban subjects.

As the contribution of each taxa to the functional shifts had been inferred based on a comparison of taxonomic abundance to gene abundance across all samples, we sought further evidence based on the genomic content of related reference genomes to corroborate these findings. KEGG orthology copy number data for the top 10 urban and rural enriched metabolic pathways were obtained for 4 representative rural and urban genomes (Supplementary Tables 7 & 8). Several pathways relating to xenobiotics biodegradation and metabolism enriched in urban subjects were encoded at high copy number by the *Klebsiella pneumoniae* and *E. coli* reference genomes but were absent or encoded at low copy number by representative rural species, particularly *Prevotella copri*. For the rural enriched pathways, most were encoded at high copy number across all 8 representative rural and urban species, consistent

with the more balanced FishTaco profiles for these pathways. Although copy number by species for rural enriched pathways tended to be slightly higher for the urban representative species, their overall contribution may be offset by their relative abundance as a proportion of the total microbiota per subject.

Although no differences were identified in pathway enrichment between *C. difficile* positive and negative diarrhoeal subjects, 54 pathways were enriched in control non-diarrhoeal subjects when compared with diarrheal subjects. These included multiple pathway categories relating to amino acid metabolism (14/54), carbohydrate metabolism (10/54), cofactors and vitamins (8/54) and energy metabolism (6/54).

Indian faecal viromes differ by geographic location

A total of 8,746 non-redundant viral sequences were detected in the whole community metagenomic sequencing data for 105 Indian faecal samples. These viruses group into 1,344 Viral Clusters (VCs), which are concordant with viral genera [22]. Network visualisation of the shared protein clusters between VCs shows the majority of Indian faecal viruses identified are connected to previously described *Caudovirales* (Figure 6A). Several *Microviridae*, *Inoviridae*, and archaeal viruses of the *Rudiviridae* and *Bicaudaviridae* families, were also detected. Unknown viruses were observed which did not share protein clusters with previously characterised viruses.

As viruses were identified in whole community metagenomic data, and not specifically targeted using viral isolation and sequencing protocols, it is expected that rare viruses are poorly represented in the final Indian faecal virome. Therefore, for diversity comparisons between cohorts, the Inverse Simpson's index was employed as it is less sensitive to rare taxa. No difference in viral diversity was observed between diarrhoeal and control subjects within specific residence locations. However, a difference in the Inverse Simpson's index was detected between the rural and urban cohorts (rural mean 58.00 +/- 37.53 versus urban mean 46.01 +/- 25.36, p adj=0.002; Figure 6B).

The unique composition of Indian faecal viromes were assessed through PCoA. The greatest variance is attributable to geographical residence, with 7.8% of the data explained by urban or rural location (F=8.67, p=0.001; Figure 6C). The interaction of geographical residence and the diarrhoeal status of subjects accounts for a further 2.1% of the observed viral differences (F=2.36, p=0.012). Amongst the urban and rural Indian cohorts that were suffering from diarrhoea, the *C. difficile* status of individuals only accounted for an additional 0.6% of the PCoA variation (F=0.63, p=0.897). The impact of antibiotic usage with the geographical residence or diarrhoeal status of subject explains 1.0% and 1.4% of the calculated differences, respectively (F=1.13, p=0.315 and F=1.64, p=0.071, respectively). Additional recorded variables were tested for their effect on the Indian faecal virome. However, in combination, age, gender, BMI, and housing condition, only accounted for 1.3% of the Indian faecal virome dissimilarities (F=1.50, p=0.09).

Specific VCs were strongly associated with distinct geographical locations and diarrhoeal status. The relative abundance differences observed for the 50 VCs that had the greatest fold change by geographical location demonstrates that specific VCs are also associated with controls (Figure 6D). Particular VCs associated with urban residing subjects were also clearly associated with diarrhoea.

CRISPR spacers were used to link VCs to their potential bacterial hosts. The relative abundance of VCs and the number of CRISPR spacers against specific VCs demonstrates that urban subjects contain a greater abundance of phages targeting *Bacteroides*, *Parabacteroides*, *Bifidobacterium* and *Escherichia spp.*, while there are trends towards more *Eubacterium* and *Prevotella*-infecting VCs amongst rural-residing individuals (Figure 6E). CRISPR spacers against predicted *Klebsiella*-infecting phages were also more frequent in urban individuals, which could reflect the greater abundance of urban individuals recruited with diarrhoeal symptoms.

Virome-associated auxiliary metabolic functions

While the Indian faecal virome composition analysis was conducted on VCs present in 2 or more individuals, all viral-associated auxiliary metabolic functions were assessed on VCs present in 10 or more individuals. These criteria were implemented in order to focus on the functions associated with the most abundant Indian faecal viruses. There were 723 VCs shared by 10 or more individuals. Of these VCs, the majority (419/723 VCs, 57.95%) are detectable amongst both rural and urban habiting individuals (Figure 7A). However, urban and rural-specific VCs were also observed (240 and 64 VCs, respectively).

The functions associated with the largest representative sequence of each VC was predicted. As expected for virome analyses, the most abundant functional predictions corresponded to eggNOG category S: ‘Function unknown’ and category L: ‘Replication, recombination, and repair’ (Figure 7C). The presence/absence similarity between VC-encoded functions associated with an individual’s virome were compared using PCoA. The variation of the virome-associated auxiliary metabolic functions were better explained by geography than diarrheal status (7.1% versus 2.2%, p=0.001 and 0.025, respectively; Figure 7B).

In order to assess the energy harvesting metabolic potential or urban and rural viral communities, eggNOG categories E and G (‘Amino acid transport and metabolism’, and ‘Carbohydrate transport and metabolism’, respectively) were investigated. The rurally abundant VCs encode at statistically higher frequency genes involved in amino acid and carbohydrate transport and metabolism (Figure 7D-E).

Discussion

The composition of the gut microbiome in the context of health and to a much lesser extent, disease, in Indian populations is not well understood. This study is the first to utilise shotgun metagenomics sequencing to comprehensively characterise the gut bacteriome, resistome and virome of rural and urban

diarrhoeal and control populations without diarrhoea living in two geographically and culturally distinct regions of Central India, Nagpur and Melghat. To improve our understanding of CDI epidemiology in Central India, we initially focused our efforts on studying for the first time the prevalence of CDI in peri-urban, urban and rural populations in Nagpur and Melghat, respectively. We report important new epidemiologic data for Central India highlighting that CDI is an emerging but as yet under-recognised healthcare-associated infection associated mainly with urbanisation and antibiotic exposure. Although there is very limited data on the incidence and epidemiology of CDI in India as a whole, a handful of reports mainly conducted in hospitalised patients, indicate detection rates in the range of 6-15.7% [19-21], which is in line with *C. difficile* toxin positivity rates described herein. These findings highlight the need to enhance awareness of and testing of subjects with diarrhoea for *C. difficile* in India, particularly in high-risk individuals with recent or ongoing antibiotic exposure or hospitalisation.

In our follow-on faecal metagenome study in which we also sought to characterise the impact of *C. difficile*, we selected an enriched set of faecal DNA samples derived from diarrhoeal samples testing positive in diagnostic *C. difficile* immunoassays for whole-genome shotgun sequencing (WGS). The taxonomic profiles revealed geographically distinct gut microbiota signatures. As compared with the urban population of Nagpur district, the rural villagers of the Korku tribe in Melghat were observed to have a significantly higher abundance of *Prevotella spp*, particularly in the control subjects, and an underrepresentation of common members of urban-industrial gut microbiomes (e.g., *Bacteroides spp*.). *Prevotella* has been reported as the most prevalent genus associated with the healthy Indian population in previous microbiome studies [11, 14-16] and has also been observed as the dominant genus in Mongolian, Amerindian and Malawian groups [11], indicating the occurrence of Enterotype 2 as proposed by Arumugam et al., 2011 [23]. *Prevotella* predominance may reflect the diet of the Korku tribe, which is rich in carbohydrates and dietary fibres. In contrast, Nagpur samples were associated with enterotype-1, which were driven by *Bacteroides* and may be again explained by this population's dietary habits, which typically consists of rice, with some meat and fish. Interestingly, multivariate analysis revealed that geographic location actually accounted for most of the variation in gut microbial communities with diarrhoeal status, including *C. difficile* toxin positivity and antibiotics contributing to a lesser extent. Consistent with recent findings from a large-scale clinical microbiome study which surveyed over 7000 individuals across 14 districts within the Guangdong province in China [21], inter-individual differences in the composition of the gut microbiome could be overwhelmingly explained by an individual's geographic location. Nevertheless, it is also now accepted that ethnicity strongly selects for specific taxa, although it is unclear what aspects of ethnicity, whether culturally related activities or genetics, underlie its observed association with the microbiota [24-25].

The misuse and overuse of antibiotics in veterinary, agricultural and clinical applications is rampant in India, fuelling antimicrobial resistance. Inadequate public health infrastructure, poor sanitation, and infection control practices in the primary healthcare system increase demand for parallel markets and further contribute to the overuse of antibiotics. Antibiotic resistance is also being driven environmentally by untreated urban waste, sewage effluent from Indian hospitals [26] and pharmaceutical pollution of waterways [27]. Indiscriminate use of beta-lactam antibiotics in both the community setting and

hospitals has given rise to the presence of antibiotic-resistant *Enterobacteriaceae* in healthy human faecal samples in North India [28]. Our faecal resistome data has corroborated recent shotgun metagenomics data indicating the widespread presence of AMR genes in virtually all subjects irrespective of geographic location and is consistent with that reported in Chinese, Hazda hunter-gatherer and resource-limited Latin American faecal microbiotas [2, 3, 7]. However, although genes conferring resistance to beta-lactam antibiotics, tetracyclines and macrolides, lincosamides and streptogramins (MLS) appeared to be common throughput Nagpur district and Melghat habitats, rural subjects from the Korku tribe generally reported lower exposure to antibiotics and thus displayed a lower abundance of other AMR genes compared with the urban Nagpur participants. In this latter group, those individuals with *C. difficile* infection on antibiotics were carrying AMR genes to virtually every antibiotic class.

The co-occurrence of pathogens and AMR genes for critically important antibiotics offers increased opportunities for unwanted horizontal gene transfer events [26]. Perhaps of most concern, the Ambler class B metallo-beta-lactamase NDM which was detected in only 1 of 35 rural subjects was found in 32/70 urban subjects, and also supports clinical data detecting carbapenemase producing pathogens from Mumbai [29] and another recent study showing that NDM-1 is also common in hospital effluent from Delhi [30]. Our findings suggest that improving sanitation, health, and education as part of the UN Sustainable Development Goals as well as the consideration of new legislative measures for curtailing environmental pollution may be effective strategies for limiting the burden of AMR in India and globally.

Analysis of taxon-level shift contribution profiles in the Nagpurian population suggested that distinct bacteria such as *Bacteroides spp., Parabacteorides distasonis, Klebsiella pneumonia* and *E. coli* may potentially possess xenobiotic, lipid and amino acid metabolising capabilities. In support of these observations, *Parabacteroides distasonis* has recently been shown to transform bile acids which have lipid-digestive and absorptive functions, and enhances the level of succinate in the gut. *Bacteroides spp.* are also dominant in amino acid metabolism in the large intestine [31]. In addition, different species of *Klebsiella* appear to have substantial potential for the biodegradation of diverse pollutants, such as halogenated aromatic and nitroaromatic compounds [32]. This result is in line with previous evidence, which suggest that individuals belonging to different geographies have microbiota with distinct xenobiotic metabolising capacities [33]. Our analysis of taxa associated shifts in metabolic function could also reflect diet and/or the higher exposure of these urban habitants to industrial/agricultural chemicals such as pesticides, fertilisers, antibiotics and other pharmaceuticals.

In rural subjects, metagenome-based abundance of pathways comprising amino acid and carbohydrate metabolism and metabolism of cofactors and vitamins correlated with the abundances of *Prevotella spp.* and several Firmicutes. These observations are consistent with previous evidence indicating that *Prevotella spp.* show capacity to digest complex carbohydrates and display enzymatic potential to break down cellulose and xylan from foods [34]. A specific strain, *Prevotella copri,* is one of the strongest driver species associated with branched chain amino acid biosynthesis in the gut and insulin resistance [35], and vitamin A and β-carotene from bananas and mangos can stimulate the growth of both *P. copri* and *P. stercorea* [36]. Furthermore, the faecal metagenomes of the rural subjects were also enriched in genes

associated with thiamine metabolism. It is feasible that thiamine deficiency, which is likely to be prevalent in the Korku, may be leading to a host driven compensatory increase in thiamine producing microbiota in the gut.

Ecological studies of macro-organisms consistently demonstrate the importance of predators within environments. Nonetheless, the majority of human microbiome studies only consider its bacterial fraction and do not concomitantly study this ecosystem's predators, viruses. In this study, we identified and analysed 8,746 viral sequences grouped into 1,344 putative genera termed Viral Clusters (VCs). Similar to previous studies of the human faecal virome, the vast majority of viruses detected are tailed phages of the order *Caudovirales* that infect bacteria (Figure 6A).

Phage predation has been proposed to modulate bacterial populations within ecosystems through various predator-prey interactions [37-38]. The faecal virome diversity of Central India rural inhabitants was greater than their urban counterparts (Figure 6B). A similar observation is described by Rampelli *et al* (2017), whereby two hunter-gatherer communities also had a higher faecal viral diversity compared to two Western society cohorts [39].

The changes in the relative abundance of VCs demonstrates specific viruses are strongly associated with urban and rural communities, and also with diarrhoeal status (Figure 6D). The identification of VCs' host bacteria through CRISPR spacers is in agreement with the bacterial analysis of Indian faecal microbiomes. The relative abundance of viruses targeting *Bacteroides* and *Parabacteroides* is greater amongst urban residing individuals, while viruses targeting *Eubacterium* and *Prevotella* are more abundant amongst rural inhabitants (Figure 6E).

The abundance of unique proteins associated with VC representative sequences demonstrates the majority of functions are shared between urban and rural viruses (Figure 7A), with geography best explaining the observed differences (Figure 7B). The most abundant functional annotations associated with Indian faecal viromes correspond to 'function unknown' and 'replication, recombination and repair' (Figure 7C). However, recent studies have highlighted the auxiliary metabolic potential of phages. Oceanic virome studies have demonstrated phages enhance the fitness of infected bacteria through augmenting their photosynthetic capability and energy production [37, 40]. Therefore, we investigated the energy harvesting potential encoded by human gut viruses. Specific pathways for amino acid and carbohydrate transport and metabolism are more abundant in rural VCs (Figure 7D & E). The increased abundance in rural associated VCs may be attributed to a narrower repertoire of encoded functions.

There were several limitations to this study. Co-morbidity data were unknown and we were unable to capture BMIs for all participants. Detailed dietary information was not available using a standard FFQ approach. Further, the control population comprised mainly hospitalized patients without diarrhoea and thus do not represent healthy controls. Due to lack of diagnostic facilities, we were unable to determine the etiological cause of acute diarrhoea or in the case of *C. difficile* positive samples, undertake further strain characterisation studies. Finally, due to limitations related to specimen collection and preparation,

we were unable to assess other components of the microbiome, including RNA viruses and intestinal parasites.

Conclusions

Here we report the most comprehensive study to date that has simultaneously examined the enteric bacteriome, DNA virome and antibiotic resistome in divergent populations in Central India, a region of the world that has been grossly understudied. Together, these data suggest that not all rural traditional societies display a healthy gut microbiota as exemplified by a lack of significant difference in bacterial diversity between our rural and urban cohorts and the presence of a core set of AMR genes. Our findings will help assess progress towards meeting the goals of global and national action plans to tackle AMR and the burden of infectious diarrhoea in India, including CDI. These results may also be useful in laying the foundations for implementing culturally acceptable One Health-inspired interventions to improve healthcare outcomes in this region of the world.

Methods

Experimental design and aim of study

The main aim of this observational cohort study was to use shotgun metagenomics to characterise the gut bacteriome, DNA virome and antibiotic resistome of two highly divergent populations in Central India; rural agriculturalists in Melghat and an urban population in Nagpur. We also sought to investigate comparative differences in microbiome profiles in subjects with and without diarrhoea, including the impact of CDI.

Human participants

Inclusion and Exclusion Criteria

During participant selection, inclusion criteria were (i) adults aged from 18 to 70 years who could provide written or thumb-print acknowledged informed consent, (ii) HIV, hepatitis B or C negative, and (iii) not pregnant or breast-feeding.

For the diarrhoeal group, a presumptive diagnosis of infective diarrhoea was defined as 3 or more loose stools in a 24-hour period accompanied by other gastrointestinal symptoms such as nausea, vomiting, abdominal cramps, tenesmus, bloody stools, or fever (oral temperature $\geq 38^{o}$C). All subjects in the *C. difficile*-infected group had diarrhoea and a positive stool *C. difficile* (enzyme immunoassay) for toxin.

The exclusion criteria for this group were (i) any individual with a known non-infectious cause of diarrhoea such as inflammatory bowel disease, (ii) those unable to provide a stool sample, (iii) or if the sample is formed stool. For the non-diarrhoeal control group, the exclusion criteria were (i) presence of acute diarrhoea at the time of or within 2 weeks of recruitment or (ii) those unable to provide a stool

sample. It was acknowledged that such individuals could be recruited from the in- or outpatient population and could have been exposed to antibiotics in the recent past (within 3 months of recruitment), although ideally not at the time of recruitment.

Immunosuppression was defined as those on prednisolone (>5mg/d), immunomodulators (azathioprine, methotrexate, calcineurin inhibitor) or biologics.

Human Geography - Nagpur

Nagpur is the third largest city of the Indian state of Maharashtra and the 13th largest city by population (2.5M) in India. It is located at the exact centre of the Indian peninsula (zero milestone) and enjoys a tropical savannah climate where temperatures can reach in excess of 48 ^{o}C in the summer months. Hinduism is the main religion followed closely by Buddhism and Islam, with smaller contributions from Christianity, Jainism and Sikhism.

Nagpur is an emerging metropolis attracting significant commercial inward investment and is a major education hub in Central India. It is also home to the Central Indian Institute of Medical Sciences (CIIMS). Nagpur was declared open defecation free in January 2018 and is one of the cleanest and most livable cities in India, as a leader in healthcare, green spaces and public transportation. The majority of households have good drinking water and sanitation facilities, and use clean fuel for cooking.

Human Geography - Melghat

Melghat Tiger Reserve, with its diverse flora and fauna, is located in Amaravati district of Maharashtra and is home to approximately 250,000 members of the Korku tribe spread across two talukas, Dharni and Chikaldhara and 300 villages, and extends across 4,000 square km. By road, it is approximately 250 km from Nagpur.

All rural Melghat subjects within the Melghat Tiger Reserve of Maharashtra identify as members of the Korku Scheduled Tribe and practice Hinduism mixed with ancestral worship. The Korku are an Adivasi ethnic group, speak Korku dialect, and are primarily an agriculturalist community of low socioeconomic status, high rates of illiteracy and malnutrition and possess poor access to medical and educational facilities. They live in small huts typically made of mud, grass and bamboo frames which lack an electricity or running water supply or proper sanitation systems and possess unique and distinct cultural knowledge, beliefs, and customs. MAHAN Trust is a non-governmental organisation which provides medical facilities to the tribal population of the Melghat region through its charitable Mahatma Gandhi Tribal Hospital.

Metadata collection (C. difficile prevalence study)

Site-specific project coordinators were assigned to review health records form each participant. Basic demographic details including age, gender, geographic location, hospitalisation exposure, antibiotic usage during and before (within 3 months) of study recruitment, and toxigenic (GDH^+ toxin -) and non-toxigenic (GDH^+ $toxin^-$) *C. difficile* detection rates were recorded for peri-urban outpatient, urban in-and outpatients and a rural population.

Metadata collection (Metagenome study)

In addition to the metadata collected for the *C. difficile* prevalence study, site-specific coordinators also recorded BMI, immunosuppression status, and environmental details: type and location of home dwelling, number in family, drinking water supply, hygiene practices and number and type of domestic animals

Faecal Sample Collection and Storage

All specimens were anonymised and assigned a study code number linked to participant demographic details. Human faecal samples were collected from urban participants with and without diarrhoea that were either in- or outpatients from the Central Indian Institute of Medical Sciences (CIIMS), Nagpur or from other hospitals within a 20 km radius of CIIMS. Similarly, faecal samples were also collected from participants with and without diarrhoea in Melghat with the assistance of research fellows based at the Mahatma Gandhi Tribal Hospital, which hosts a CIIMS satellite laboratory and other neighbouring hospitals within Melghat. Suitable recruits were identified by the research fellows who interacted daily with village healthcare workers to facilitate participant recruitment and sample collection. Up to two samples (3-5 grams each) were collected in UV sterilised dry plastic containers at the time of recruitment from each participant and placed in a cool box. As per the standard operating procedures, all stool specimens were stored at 4^oC immediately after collection to avoid enzymatic degradation prior to detection of toxigenic *C. difficile* and genomic DNA extraction which were performed within 24 hours of sample collection.

Detection of Clostridioides difficile GDH antigen and free toxin in diarrheal stool samples

All diarrhoeal samples in the metagenome study (58/105) were tested for *Clostridioides difficile* infection (detection of glutamate dehydrogenase antigen and toxins A/B) using the C. DIFF QUIK CHEK COMPLETE-enzyme immunoassay (QCC; TechLab, Blacksburg, VA, USA) in accordance with the manufacturers' instructions, including the use of appropriate controls as specified in the package insert.

Briefly, ~25 ml of stool sample was added to a tube containing the diluent and conjugate and the mixture was transferred to the device sample well. After incubation for 15 min at room temperature, the wash buffer followed by the substrate were added to the reaction window. The results were read after 10 min. The GDH antigen and/or toxins were reported as positive if a clear visible band was seen on the antigen and toxin side of the device display window, respectively, confirming the presence of toxigenic *C. difficile* as per manufacturer guidelines.

Detection of diarrhoeagenic E. coli virulence genes in diarrhoeal stool samples

Multiplex polymerase chain reaction assays (HiMedia Laboratories Pvt. Ltd., Mumbai, India) were used to detect diarrheagenic *E.coli* (DEC) 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) as per the manufacturers' instructions, including the use of 2 different species-specific primer sets and appropriate controls as specified in the package insert (see Supplementary Methods for further details).

Faecal DNA extraction

DNA was extracted from 1 to 1.5g of feces and homogenised in lysis buffer (Tris HCl, EDTA, NaCl and SDS). The content was centrifuged at 7,000 *x g* for 10 min. The supernatant was then transferred to a 1.5mL tube containing a mixture of Isopropanol and Sodium acetate (5M) and incubated at -20^{o}C for 30 min. Following removal of the supernatant the pellet was dried for about an hour. The pellet was suspended in 1X Tris EDTA buffer (pH 8) and incubated at 65^{o}C for 15 min. An approximate equal volume (0.5- 0.7 ml) of Phenol: Chloroform- Isoamyl alcohol (24:1) was added, mixed thoroughly and centrifuged for 10 min at 12,000 *x g*. The aqueous viscous supernatant was carefully transferred to a new 1.5mL tube. An equal volume of Chloroform-Isoamyl alcohol (1:1) was added, followed by centrifugation for 10 min at 12,000 *x g*. The supernatant was mixed with 0.6x volume of Isopropanol to aid precipitation. The precipitated nucleic acids were washed with 75% ethanol, dried and re-suspended in 50µL of TE buffer.

Whole-Genome Shotgun (WGS) Sequencing

Sequencing was carried out by Source Biosciences (Nottingham, U.K.). High quality genomic DNA was quantified using Qubit Broad Range (Invitrogen, U.K.) and prepared for Illumina paired end sequencing following the TruSeq DNA Nano manufacturers protocol (Rev D, June 2015) (Illumina Inc, San Diego, U.S.A.). The DNA was sequenced using a standard HiSeq 4000 150bp PE flowcell. Raw data has been

submitted to the European Nucleotide Archive under the accession number https://www.ncbi.nlm.nih.gov/bioproject/PRJNA564397

Generation of taxonomic, resistome and functional profiles from metagenomic shotgun data

Raw Fastq files (average 13,410,735 reads per sample) were assessed for quality using skewer [41], trimming adaptor reads and regions of quality below a phred of 30. The filtered reads (average 10,635,653 reads per sample) were then assessed for taxonomic assignments using Metaphlan2 [42] and for the presence of antimicrobial resistance genes using ARIBA [43] with the MegaRes database [44].

Functional analysis was performed using MOCAT2 (v2.1.3) [45]. Briefly, trimmed and filtered reads were assembled into contigs with SOAPaligner (v2.21). These contigs are initially corrected for indels and chimeric reads using BWA (v0.7.5a-r16) and screened against the human hg19 reference to filter out reads which originated from the host using USEARCH (v5/v6). Genes were predicted using Prodigal (v2.60). Single copy marker genes are extracted using fetchMG (v1.0) and clustered using CD-HIT (v4.6). The gene catalogues were annotated using DIAMOND (v0.7.9.58) against multiple functional databases including eggNOG [46] and KEGG [47]. The abundance of genes annotated to specific KEGG orthologs (KO) was determined using the insert mm dist among unique norm setting in MOCAT2, normalising by read length and sequencing depth and allowing for multiple mappers.

Analysis of taxonomic contributions to functional shifts

Functional shifts between groups and predicted taxonomic contributions were calculated using the FishTaco package [48], taking the species-level taxonomic table produced by Metaphlan2 and the normalised KO abundance table from MOCAT2 as inputs. Only 49 taxa which exceeded a minimum proportional abundance of greater than 0.1 in any single sample were included in the final model. Enriched pathways were identified using the Wilcoxon rank sum test at FDR corrected $p<0.05$. Taxonomic contributions were predicted by *de novo* inference in FishTaco, inferring genomic content through a permutation-based approach and performing a total of 50 permutations per differentially abundant pathway.

For comparison of gene copy number for enriched metabolic pathways, KO gene copy numbers for 8 gut-associated annotated reference genomes were obtained from the Integrated Microbial Genomes and Microbiomes (IMG) database [49] as follows; *Prevotella stercorea* DSM 18206 (IMG: 2513237318), *Prevotella copri* CB7 DSM 18205 (IMG: 2562617166), *Eubacterium rectale* DSM 17629 (IMG: 650377936), *Ruminococcus bromii* L2-63 (IMG: 650377966), *Escherichia coli* UM147 (IMG: 2728369554), *Klebsiella pneumoniae* YH43 (IMG: 2687453226), *Bacteroides vulgatus* mpk (IMG: 2687453192),

Parabacteroides distasonis 2b7A (IMG: 2660238380). KO gene copy numbers associated with each enriched metabolic pathway were aggregated to yield overall pathway gene counts.

Detecting viruses in whole community metagenomic shotgun data

Sequencing reads were processed using Trimmomatic (version 0.36) [50], to remove Illumina adaptors and prune sequences where the Phred score dropped below 30 across a 4bp sliding window. All surviving reads less than 70bp were discarded. Fastq reads were assess pre- and post-processing using fastqc [51]. Both the paired and unpaired, forward and reverse reads from samples were assembled individually using metaSPAdes (version 3.11.1) [52]. Only contigs greater than 1,000bp were examined further.

Two approaches were employed to find viruses within whole community metagenomic assemblies. A standard reference-based similarity search was performed to detect sequence relatedness to known viruses, while a reference-independent approach was undertaken by searching for sequences which encode a high density of viral proteins. For the reference-based search, nucleotide sequences were queried locally using BLAST (version 2.6.0+) [53] against the viral RefSeq database (version 89; E-value 1E-10) [54], the complete Reference Viral Database (C-RVDB version 14.0; E-value 1E-05) [55], and 249 crAss-like phages previously described as the human gut's most abundant virus (E-value 1E-05) [56].

For the reference-independent approach, proteins for all contigs were predicted using Prodigal (version 2.6.3) [57] with the 'meta' option enabled for small contigs and Shine-Dalgarno training bypassed. Proteins were subsequently queried against the prokaryote Viral Orthologous Groups database (pVOGs) [58] using HMMER (version 3.1b2) [59], with a minimum score requirement of 15. Putative reference-independent discovered viruses needed to fulfil three basic requirements: (i) ≥1.5kb, (ii) encode 2 distinct proteins with similarity to 2 unique pVOGs, and (iii) encode ≥2 pVOGs per 10kb-equivalent genome length. Additional stringent dynamic filtering was applied to contigs based on their actual genome length. For contigs <5kb, it was required that there were at least ≥5 distinct pVOG hits; contigs ≥5kb and <10kb, ≥6 pVOG hits; contigs ≥10kb and <20kb, ≥7 pVOG hits; contigs ≥20kb and <40kb, ≥8 pVOG hits; contigs ≥40kb and <60kb, 9 pVOG hits; and contigs ≥60kb, 10 pVOG hits.

All putative viral contigs detected using the reference-dependent and -independent methods were pooled and made non-redundant as follows: following a BLASTn all-v-all, the larger of two contigs were retained when the blast identity and coverage between two sequences exceeded 90%. Subsequently, any putative virus encoding a ribosomal protein (BLASTp, E-value 1E-10) was removed from further analysis. This was performed for stringency despite recent research showing specific viruses can encode ribosomal proteins [60]. In addition, any contig encoding a protein with similarity to all available Pfam sequences (version 32.0) plasmid replication proteins PF01051, PF01446, PF01719, PF04796, PF05732, and PF06970, were removed (HMMER, score 15).

Viral contigs were grouped into Viral Clusters (VCs) using vContact2 (version 0.9.8) [22], implemented through the CyVerse Discovery Environment. Protein clusters were identified amongst VCs using default settings (Diamond, E-value 0.0001), and with the inclusion of known viruses (Bacterial and Archaeal Viral RefSeq 85, with ICTV and NCBI taxonomy). Following vContact2, only viral clusters that contain viral sequences from two or more of the study's complete cohort (n=105) were analysed further. This was designed to remove singleton and spurious viral sequences that may be transiently associated with diet, but are not abundant or stable components of the faecal microbiome. The final Indian faecal virome was visualised as a network through Cytoscape (version 3.7.1) [61], with viral sequences as nodes and shared protein clusters as edges. The edge distance between connected viruses is calculated by vContact2 as their 'interaction'.

Discerning differences in virome diversity and abundance

Quality filtered reads, both paired and unpaired, were mapped onto the final Indian faecal virome using bowtie2 in 'end-to-end' mode (version 2.3.4.1) [62]. The read alignment outputs were converted to sorted bam files through samtools (version 1.7) [63]. The abundance and breadth of coverage of reads mapping to each contig was determined using the bedtools coverage function (version 2.26.0) [64]. Subsequently, in order to determine if a viral sequence was indeed present in a faecal virome, a breadth of coverage filtering was applied. This was designed to remove viruses where potentially 100s of reads could map onto a single conserved region. Therefore, for viral sequences ≤5kb, 75% of the genome needed to be covered by aligned reads; sequences >5kb and ≤50kb, 50% of the genome needed to be covered; and >50kb, 25% of the genome needed to be covered.

In addition to 105 faecal metagenomes, two negative control samples (water) were sequenced. While these samples contributed no contigs to the final Indian faecal virome, the breadth of coverage of sequencing reads from these samples was used to remove potential contaminant sequences. Any viral sequence, from any sample, which 'passed' the breadth of coverage filtering using reads derived from either water sample were removed from further analysis.

Any viral sequence from a faecal microbiome sample which failed the breadth of coverage filtering was recorded as zero reads, while if the filtering step was passed, the observed number of reads aligned were used to populate the read count matrix. Due to differences in sequencing depth between samples, the read count matrix was normalised per sample using the DESeq2 ratio of means method [65]. The reads aligned to individual viral sequences were aggregated by their vContact2 determined VCs. DESeq2 was subsequently used to calculate the VC changes between cohorts. The normalised VC read count matrix was used to determine the diversity and statistical differences observed between Indian faecal microbiome cohorts (see 'Statistical Analyses' below).

Determining phage-host pairs and viral encoded functions

CRISPR spacers from bacterial contig assemblies were predicted using PILER-CR (version 1.06) [66]. Putative CRIspr spacer predictions <20bp and >100bp were discarded. The CRISPR spacers were queried locally using BLASTn against all individual viral sequences which formed the Indian faecal virome VCs. Due to the use of short nucleotide sequences, only CRISPR spacers with an E-value ≤0.001 and ≤1 mismatch were considered as significant. In order to determine the taxonomy of the original assembled bacterial contigs, or the pre-assembled contigs from the Pasolli *et al*. (2019) study [67], contig kmer MinHash sketches were queried against JGI taxonomy server using the BBMap sendsketch function (version 38.44) [68].

The functions associated with Indian faecal viruses were determined using eggNOG-mapper v1 (online submission portal) using the eggNOG 4.5.1 database [46]. For each VC, the largest viral sequence was chosen as a representative of that VC. In order to avoid the confounding effect of viral abundance fluctuations within the faecal microbiome, the relative abundance of VCs observed at the specific sampling time-point were not taken into consideration. Only the overall presence-absence and abundance of viral-encoded functions were considered. The similarity between virome-encoded functions, with respect to presence-absence, were assessed through PCoA using the Jaccard index. The abundance of specific metabolic genes were compared between cohorts, with statistical difference determined by the Mann-Whitney U test with Bonferroni correction using the 'ggpubr' compare means function in R.

Statistical Analyses and Graphic Generation

All statistical analyses were conducted in R (64-bit, version 3.6.0; Foundation for Statistical Computing, Vienna). The package 'vegan' was used for measures of taxonomic diversity including alpha diversity (Inverse Simpson Index) and beta diversity (Principal Coordinates Analysis with Bray Curtis Dissimilarity and Jaccard Similarity). Differences in alpha diversity between study groups was assessed by ANOVA with Tukey's honest significance test. The contribution of categorical variables to beta diversity was tested for using the Adonis function (PERMANOVA) in vegan. Generalised linear models assuming a negative binomial distribution were used to identify differentially abundant taxa between study groups as implemented in the R package 'mare'. Hierarchical clustering of resistance gene abundances and heatmap generation was performed with the package 'heatmap3'. For comparison of resistance gene and metabolic pathways counts between groups, the Mann-Whitney U test was used. All p values obtained from testing with multiple comparisons were corrected for false discovery rate (FDR, Benjamini-Hochberg). The fold changes observed in the relative abundances of VCs across geographical and diarrhoeal status cohorts were calculated using the 'gtools' package in R. Using the same package, the fold changes were converted to log ratios (base 10). All graphical images were generated using 'ggplot2'.

Abbreviations

CRISPR: Clustered Regularly Interspaced Short Palindromic Repeats

AMR: Antimicrobial Resistance

CDI: *Clostridioides difficile* infection

CDT: *C. difficile* toxin

CIIMS: Central India Institutes of Medical Sciences

BMI: Body Mass Index

PBP: Penicillin Binding Protein

ESBL: Extended Spectrum Beta-lactamases

VCs; Viral Clusters

ICTV: International Committee on Taxonomy of Viruses

MLS: Maximum Length Sequence

NDM-1: New Delhi Metallo-Beta-Lactamase 1 Enzyme

RNA: Ribonucleic Acid

DNA: Deoxyribonucleic Acid

rRNA: Ribosomal RNA

HCl: Hydrochloric Acid

NaCl: Sodium Chloride

EDTA: Ethylenediaminetetraacetic Acid

SDS: Sodium Dodecyl Sulphate Reagent

Tris: Tris[hydroxymethyl]aminomethane

WGS: Whole-Genome Sequencing

NCBI: National Center for Biotechnology Information

Declarations

Ethics approval and consent to participate

This study was approved by the Faculty of Medicine and Health Sciences Research Ethics Committee at the University of Nottingham (REC No. 199-1901) and the Ethical Committee of the Central India Institute of Medical Sciences, Nagpur.

Availability of data and material

Metagenomic sequencing datasets generated and analysed during the current study are available in the European Nucleotide Archive under accession number:
[https://www.ncbi.nlm.nih.gov/bioproject/PRJNA564397]

All sequencing reads that map to the human reference genome have been removed from the sequencing files.

Competing interests

TMM is a Consultant advisor for CHAIN Biotechnology. MHW has received consulting fees from Actelion, Astellas, bioMerieux, Da Volerra, Merck, Meridian, Pfizer, Sanofi-Pasteur, Seres, Singulex, Summit, Synthetic Biologics, Valneva, Vaxxilon & VenatoRx; lecture fees from Alere, Astellas, Merck, Pfizer & Singulex; and grant support from Actelion, Alere, Astellas, bioMerieux, Da Volterra, Merck, MicroPharm, Morphochem, AG, MotifBio, Paratek, Sanofi-Pasteur, Seres, Summit & Tetraphase. All other authors declare no competing interests.

Funding

This work was supported by a University of Nottingham Anne McLaren Fellowship to Tanya Monaghan and supplemented by the National Institute for Health Research (NIHR) Nottingham Digestive Diseases Biomedical Research Centre based at Nottingham University Hospitals NHS Trust and University of Nottingham, as well as research funding for *C. difficile* diagnostic assays provided by Mark Wilcox, University of Leeds. The funders had no involvement in study design, writing the manuscript or decision for publication.

Author contributions

T.M.M., T.J.S., S.S., and A.B. designed the study, analyzed the data and wrote the paper. T.M.M., R.S.K., A.S., developed the clinical sample cohorts and R.B., R.N., S.M., J.G., and P.J. managed sample and metadata collection, DNA extraction and quantification. A.B., T.J.S., S.S., analysed the WGS data. T.M.M., T.J.S., and S.S. performed the statistical analyses. S.A., R.D.E., M.W., L.A.D., and C.H., in addition to all other co-authors, reviewed the manuscript, provided feedback, and approved the manuscript in its final form.

Acknowledgements

We are grateful to the participants that have made this research possible. We thank Melanie Lingaya and Yirga Falcone for their technical assistance in sample preparation; to Guru Aithal and the Nottingham Digestive Diseases Centre who provided financial assistance with travel and sample transportation costs of faecal nuclei acid, to Dr Lokendra Singh, Director of CIIMS, for providing access to the laboratory facilities at CIIMS and granting approval of the study, and to Teresa Coughlan at Source BioScience for help in sequence production and sample management.

References

1. Shkoporov AN, Hill C. Bacteriophages of the Human Gut: The “Known Unknown” of the Microbiome. Cell Host & Microbe. 2019; *25*, 195–209.
2. Hu, Y, Yang X, Qin J, Lu N, Cheng G, Wu N, et al. Metagenome-wide analysis of antibiotic resistance genes in a large cohort of human gut microbiota. Nat. Commun. 2013; *4*, 2151.
3. Pehrsson EC, Tsukayama P, Patel S, Mejita-Bautista M, Sosa-Soto G, Navarrete, KM, et al. Interconnected microbiomes and resistomes in low-income human habitats. Nature. 2016; *533*, 212-6.
4. Hendriksen, RS, Munk P, Njage P, van Bunnik B, McNally L, Lukjancenko O, et al. Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. Nat. Commun. 2019; *10*, 1124.
5. Yatsunenko T, Rey FE, Manary MJ. Trehan I, Dominguez-Bello MG, Contreras M, et al. Human gut microbiome viewed across age and geography. Nature. 2012; *486*, 222-7.
6. Schnorr SL, Candela M, Rampelli S, Centanni M, Consolandi M, Basaglia G, et al. Gut microbiome of the Hazda hunter-gatherers. Nat. Commun. 2014; *5*, 3654.
7. Rampelli S, Schnorr SL, Consolandi C, Turroni S, Severgnini M, Peano C, et al. Metagenome Sequencing of the Hazda Hunter-Gatherer Gut Microbiota. Curr. Biol. 2015; *25*, 1682-93.
8. Obregon-Tito AJ, Tito RY, Metcalf J, Sankaranarayanan K, Clemente JC, Ursell LK, et al. Subsistence strategies in traditional societies distinguish gut microbiomes. Nat. Commun. 2015; *6*, 6505.
9. Clemente JC, Pehrsson EC, Blaser MJ, Sandhu K, Gao Z, Wang B, et al. The microbiome of uncontacted Amerindians. Scientific Adv. 2015; *1*, e1500183.

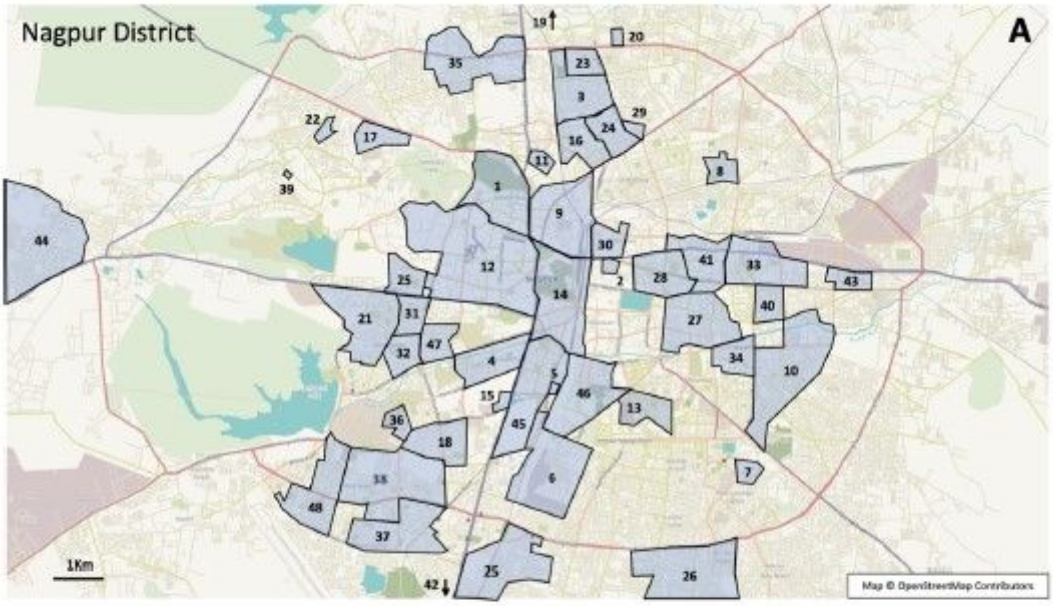
10. Martinez I, Stegen JC, Maldonado-Gomez MZ, Eren AM, Siba PM, Greenhill AR, et al. The gut microbiota of rural papua new guineans: composition, diversity patterns, and ecological processes. Cell Rep. 2015; *11*, 527-38.
11. Dehingia M, Devi KT, Talukdar NC, Talukdar R, Reddy N, Mande SS, et al. Gut bacterial diversity of the tribes of India and comparison with the worldwide data. Sci Rep. 2015; *5*, 18563.
12. Ramadass B, Sandya Rani B, Pugazhendhi S, John KR, Ramakrishna BS. Faecal microbiota of healthy adults in south India: Comparison of a tribal & a rural population. Indian J Med Res. 2017; *145*, 237-246.
13. Das B, Ghosh TS, Kedia S, Rampal R, Saxena S, Bag SM et al. Analysis of the Gut Microbiome of rural and urban Healthy Indians Living in Sea Level and High Altitude Areas. Sci Rep. 2018; *8*, 10104.
14. Kulkarni AS, Kumbhare SV, Dhotre DP, Shouche YS. Mining the Core Gut Microbiome from a Sample Indian Population. Indian J Microbiol.2019; *59*, 90-95.
15. Kushugulova A, Forslund SK, Costea PI, Kozhakhmetov S, Khasssenbekova Z, Urazova M, et al. Metagenomic analysis of gut microbial communities from a Central Asian population. BMJ Open. 2018; *8*, e021682.
16. Dhakan DD, Maji A, Sharma AK, Saxena R, Pulikkan J, Grace T, et al. The unique composition of Indian gut microbiome, gene catalogue, and associated fecal metabolome deciphered using multi-omics approaches. GigScience. 2016; *8*, 1-20.
17. Forrester JD, Cai LZ, Mbanje C, Rinderknecht TN, Wren SM. Clostridium difficile infection in low- and middle-human development index countries: a systematic review. Trop Med Int Health. 2017; *10*, 1223-1232.
18. Roldan RS, Cui AX, Pollock NR. Assessing the Burden of Clostridium difficile Infection in Low- and Middle-Income Countries. J Clin Microbiol. 2018; *56*, e01747-17.
19. Chaudhry R, Sharma N, Gupta N, Kant K, Behadur T, Shende T, et al. Nagging Presence of Clostridium difficile Associated Diarrhoea in North India. J Clin Diagn Res 2017; 11 (9): DC06-DC09
20. Singh M, Vaishnavi C, kochhar R, Mahmood S. Toxigenic Clostridium difficile isolates from clinically significant diarrhoea in patients from a tertiary care centre. Indian J Med Res 2017; 145 (6): 840-846
21. Vaishnavi C, Singh M, Mahmood S, Kochhar R. Prevalence and molecular types of Clostridium difficile isolates from faecal specimens of patients in a tertiary care centre. J Med Microbiol 2015; 64: 1297-304.
22. Bin Jang H, Bolduc B, Zablocki O, Kuhn JH, Roux S, Adriaenssens EM, et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nat Biotechnol. 2019. *37*, 632–639.
23. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, et al. Enterotypes of the human gut microbiome. Nature. 2011; *473*, 174-80.
24. He Y, Wu W, Zheng HM, Li P, McDonald D, Sheng HF, et al. Regional variation limits applications of healthy gut microbiome reference ranges and disease models. Nat Med. 2018; *24*, 1532-1535.

25. Gaulke CA, Sharpton, TJ. The influence of ethnicity and geography on human gut microbiome composition. Nat Med. 2018; 24, 1495-1496.
26. Marathe NP, Berglund F, Razavi M, Pal C, Droge J, Samant S, et al. Sewage effluent from an Indian hospital harbors novel carbapenemases and integron-borne antibiotic resistance genes. Microbiome. 2019; *7*, 97.
27. Bomboy A, Barneoud L. Recipe for disaster. NewScientist. 2019; *242*, 42-45.
28. Gupta M, Didwal G, Bansal S, Kaushal K, Batra N, Gautam V, et al. Antibiotic-resistnat Enterobacteriaceae in healthy gut flora: A report from north Indian semiurban community. Indian J Med Res. 2019; *149*, 276-280.
29. Kazi M, Drego L, Nikam C, Ajbani K, Soman R, Shetty A, et al. Molecular characterization of carbapenem-resistant Enterobacteriaceae at a tertiary care laboratory in Mumbai. Eur J Clin microbiol Infect Dis. 2015; *34*, 467-472.
30. Lamda, M, Graham DW, Ahammad SZ. Hospital wastewater releases of carbapenem-resistance pathogens and genes in urban India. Environ Sci Technol. 2017; *51*, 13906-13912.
31. Ma N, and Ma X. Dietary Amino Acids and the Gut-Microbiome-Immune Axis: Physiological Metabolism and Therapeutic Prospects. Comprehensive Reviews in Food Science and Food Safety. 2019; *18*, 221-242.
32. Rajkumari J, Singha LP, Pandey P. Genomic insights of aromatic hydrocarbon degrading Klebsiella pneumoniae AWD5 with plant growth promoting attributes: a paradigm of soil isolate with elements of biodegradation. 3 Biotech. 2018; *8*, 118.
33. Das A, Srinivasan M, Ghosh TS, Mande SS. Xenobiotic Metabolism and Gut Microbiomes. PLoS ONE. 2016; *11*, e0163099.
34. Dubois G, Girard V, Lapointe FJ, Shapiro BJ. The Inuit gut microbiome is dynamic over time and shaped by traditional foods. Microbiom*e.* 2017; *5* (1), 151.
35. Pedersen, H.K., Gudmundsdottir, V., Nielsen, H.B., Hyotylainen, T., Nielsen, T., Jensen, B.A., Forslund, K., Hildebrand, F., Prifti, E., Falony, G., et al. (2016). Human gut microbes impact host serum metabolome and insulin sensitivity. Nature *573*, 376-81.
36. Nakayama J, Yamamoto A, Palermo-Conde LA, Higashi K, Sonomoto K, Tan J, et al. Impact of westernized diet on gut microbiota in children on Leyte Island. Front Microbiol. 2017; *8*, 197.
37. Breitbart M, Bonnain C, Malki K, and Sawaya NA. Phage puppet masters of the marine microbial realm. Nat Microbiol. 2018; *3*, 754–766.
38. Hsu BB, Gibson TE, Yeliseyev V, Liu Q, Lyon L, Bry L, et al. Dynamic Modulation of the Gut Microbiota and Metabolome by Bacteriophages in a Mouse Model. Cell Host & Microbe 2019; *25*, 803-814.e5.
39. Rampelli S, Turroni S, Schnorr SL, Soverini M, Quercia S, Barone, M, et al. Characterization of the human DNA gut virome across populations with different subsistence strategies and geographical origin: Human DNA gut virome in different populations. Environ Microbiol. 2017; *19*, 4728–4735.

40. Mann NH, Cook A, Millard A, Bailey S, and Clokie, M. Bacterial photosynthesis genes in a virus. Nature. 2003; *424*, 741–741.
41. Jiang H, Lei R, Ding SW, Zhu S. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC Bioinformatics. 2014; *15*, 182.
42. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nature Methods. 2015; *12*, 902–903.
43. Hunt M, Mather AE, Sánchez-Busó L, Page AJ, Parkhill J, Keane JA, et al. ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. Microbial Genomics. 2017; 1–11.
44. Lakin SM, Dean C, Noyes NR, Dettenwanger A, Ross AS, Doster E, et al. MEGARes: An antimicrobial resistance database for high throughput sequencing. Nucleic Acids Research. 2017; *45*, D574–D580.
45. Kultima JR, Coelho LP, Forslund K, Huerta-Cepas J, Li SS, Driessen M, et al. MOCAT2: A metagenomic assembly, annotation and profiling framework. Bioinformatics. 2016; *32*, 2520–2523.
46. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, et al. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. 2016; *44*, D286–D293.
47. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research. 1999; *27*, 29–34.
48. Manor O, and Borenstein E. Systematic characterization and analysis of the taxonomic drivers of functional shifts in the human microbiome. Cell Host Microbe. 2017; *21*(2), 254-267.
49. Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, Pillay M, et al. IMG 4 version of the integrated microbial genomes comparative analysis system. Nucleic Acids Res. 2014; *42*, D560-567.
50. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014; *30*, 2114–2120.
51. Andrews S. FastQC: a quality control tool for high throughput sequence data: Available: http://www.bioinformatics.babraham.ac.uk.
52. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 2017; *27*, 824–834.
53. McGinnis S, and Madden TL. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Research. 2004; *32*, W20–W25.
54. Pruitt KD. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research. 2004; *33*, D501–D504.
55. Goodacre N, Aljanahi A, Nandakumar S, Mikailov M, Khan AS. A Reference Viral Database (RVDB) To Enhance Bioinformatics Analysis of High-Throughput Sequencing for Novel Virus Detection. MSphere. 2018; *3*, e00069-18, /msphere/3/2/mSphere069-18.atom.
56. Guerin E, Shkoporov A, Stockdale SR, Clooney AG, Ryan FJ, Sutton TDS, et al. Biology and Taxonomy of crAss-like Bacteriophages, the Most Abundant Virus in the Human Gut. Cell Host & Microbe. 2018; *24*, 653-664.e6.

57. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser, LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010b; *11*, 119.
58. Grazziotin AL, Koonin EV, Kristensen DM. Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. Nucleic Acids Res. 2017; *45*, D491–D498.
59. Finn R.D, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Research. 2011; *39*, W29–W37.
60. Mizuno CM, Guyomar C, Roux S, Lavigne R, Rodriguez-Valera F, Sullivan MB, et al. Numerous cultivated and uncultivated viruses encode ribosomal proteins. Nat Commun. 2019; *10*, 752.
61. Shannon P. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Research. 2013; *13*, 2498–2504.
62. Langmead B, and Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat. Methods. 2012; *9*, 357–359.
63. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; *25*, 2078–2079.
64. Quinlan AR, and Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; *26*, 841–842.
65. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014; *15*, 550.
66. Edgar RC. PILER-CR: Fast and accurate identification of CRISPR repeats. BMC Bioinformatics. 2007; *8*, 18.
67. Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, et al. Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. Cell. 2019; *176*, 649-662.e20.
68. Bushnell B. (2014). BBMap: A Fast, Accurate, Splice-Aware Aligner.

Figures



1. Sadar
2. Panchpaoli
3. Jaripatka
4. Ramdaspeth
5. Congress Nagar
6. Ajni
7. Sakhardara
8. Vaishali Nagar
9. Mohan Nagar
10. Nandanvan
11. Chhaoni
12. Civil Lines
13. Hanuman Nagar
14. Sitabuldi
15. Rahate Colony
16. Mecosabagh
17. Gittikhadan
18. Laxminagar
19. Om Nagar
20. Kushi Nagar
21. Ram Nagar
22. Friends Colony
23. Kukreja Nagar
24. Bezonbagh
25. Narendra Nagar
26. Manewada
27. Mahal Nagar
28. Gandhibagh
29. Indora
30. Mominpura
31. Gokulpeth
32. Shivaji Nagar
33. Lakadganj
34. Ganesh Nagar
35. Jafar Nagar
36. Abhyankar Nagar
37. Khamla
38. Pratap Nagar
39. Krishna Nagar
40. Bagadganj
41. Itwari
42. Manish Nagar
43. Wardhaman Nagar
44. Duttawadi
45. Dhantoli
46. Rambagh
47. Dharampeth
48. Trimurtee Nagar

Figure 1

Nagpur District. (A) Mapped locations of study participant home residences in Nagpur district. (B) Modern dwelling in Nagpur City. (C) Downtown Nagpur City street.

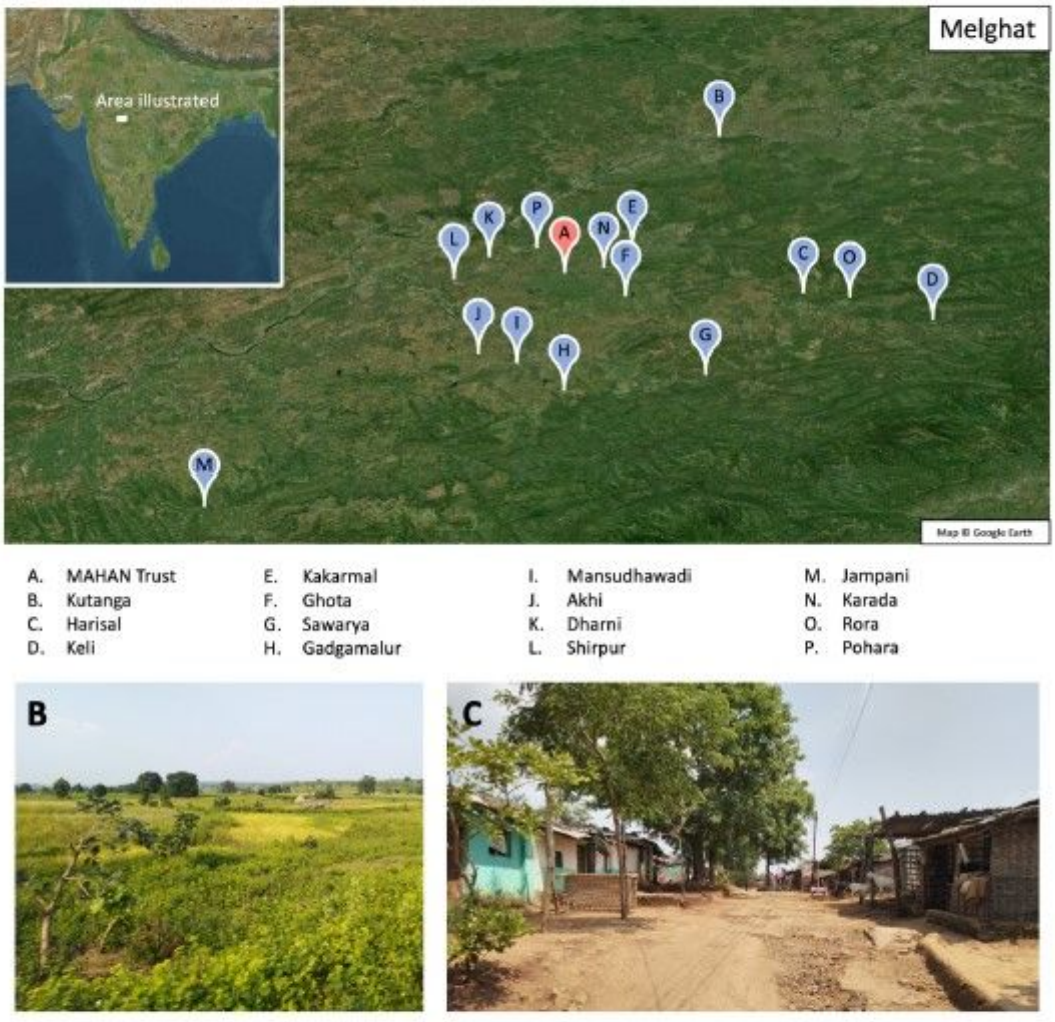


Figure 2

Melghat Region. (A) Mapped locations of Melghat villages participating in this study. (B) Geographic landscape in Melghat. (C) Traditional Melghat village.

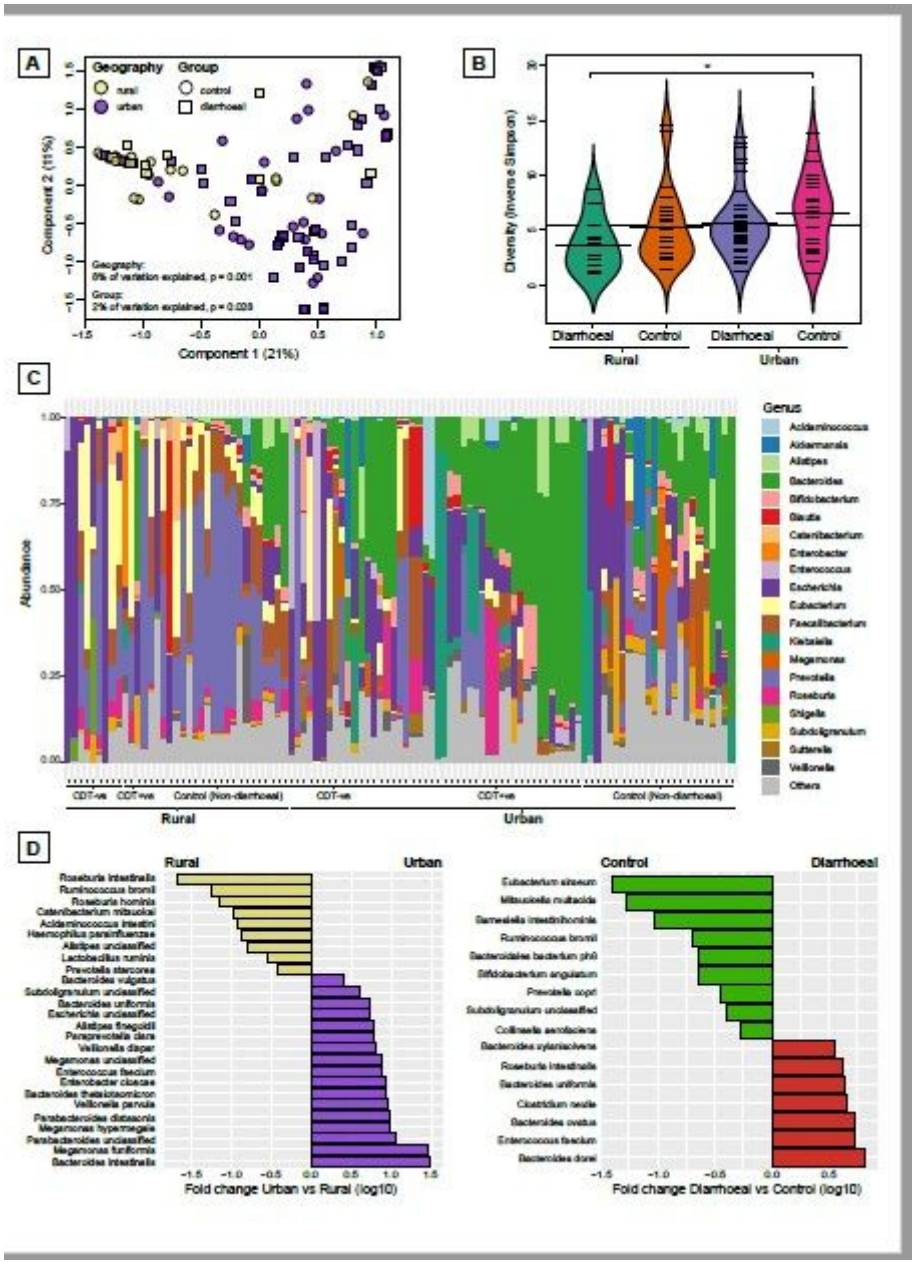


Figure 3

. Variations in the gut microbiota by geographic location and diarrhoeal status. (A) Principal coordinates analysis (PCoA) of microbiota profiles based on Bray-Curtis Dissimilarity of species-level taxonomic abundance. Subject profiles vary by both geographic location and diarrhoeal status. (B) Comparison of microbial diversity between diarrhoeal and non-diarrhoeal control subjects from both rural and urban geographic locations. * p=0.05. (C) Summary of genus-level taxonomic profiles by subject. Subjects are grouped by geographic location and diarrhoeal status, with diarrhoeal subjects further subdivided into C. difficile toxin positive (CDT +ve) and negative (CDT –ve). Bacteroides dominant profiles are more frequent in urban subjects, while Prevotella dominant profiles are more frequent in rural subjects. (D) Differentially abundant taxa at species-level based on either geographic location (left, rural vs urban control subjects) or diarrhoeal status (right, non-diarrhoeal controls vs diarrhoeal). All taxa shown are significantly different between groups based on generalized linear models with FDR corrected p<0.05.

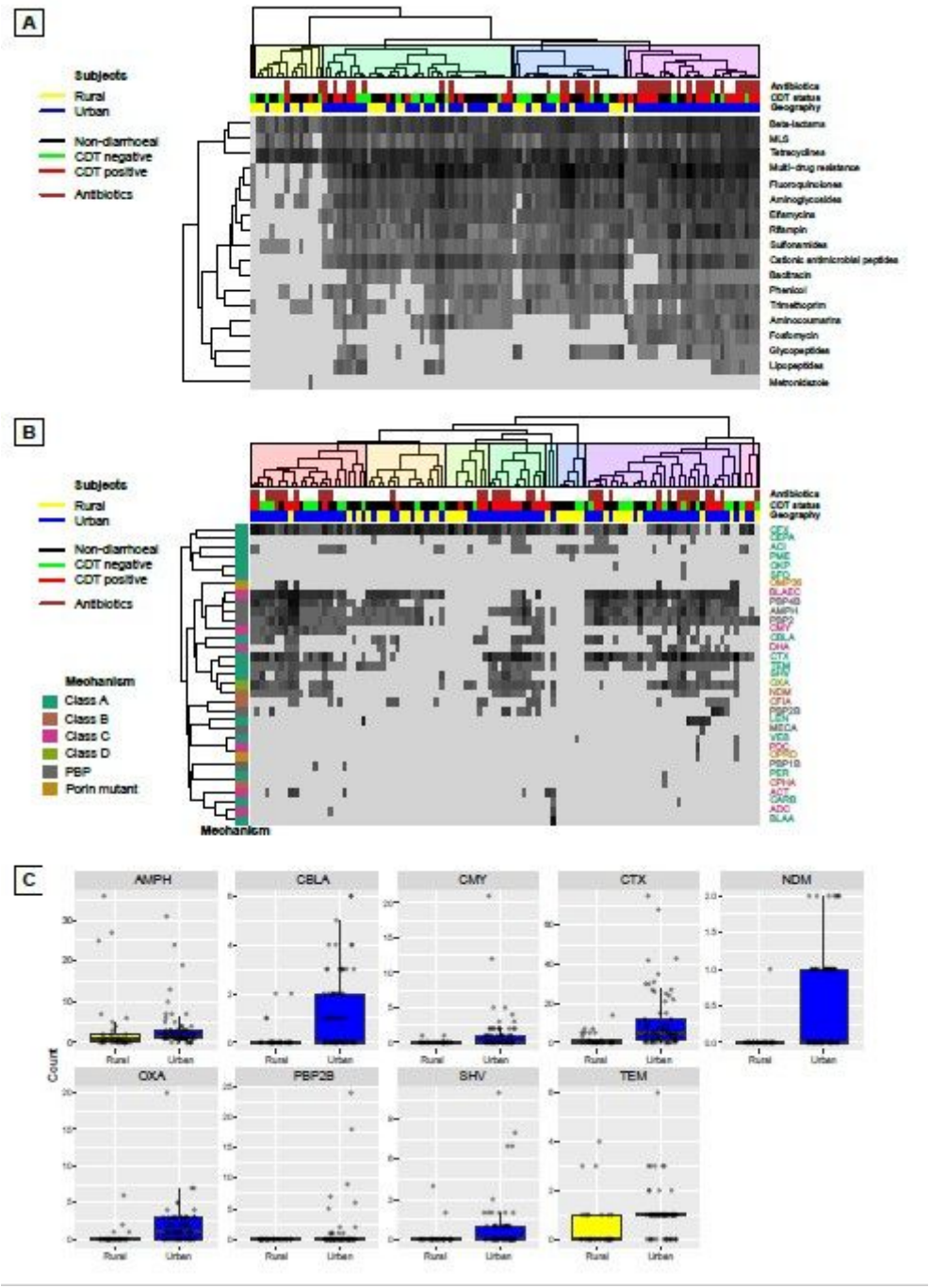


Figure 4

Analysis of antimicrobial resistance gene carriage by gut microbiota. (A) Heatmap of antimicrobial resistance (AMR) gene abundance aggregated by antibiotic class. MLS = Macrolides, Lincosamides and Streptogramins. (B) Heatmap of antimicrobial resistance gene cluster abundance for Beta-lactam antibiotics. Beta-lactam resistance mechanisms for each gene cluster are indicated; Ambler class A to D, PBP (Penicillin Binding Protein) or Porin mutant. (C) Comparison of the presence of Beta-lactam resistance genes between rural and urban subjects.

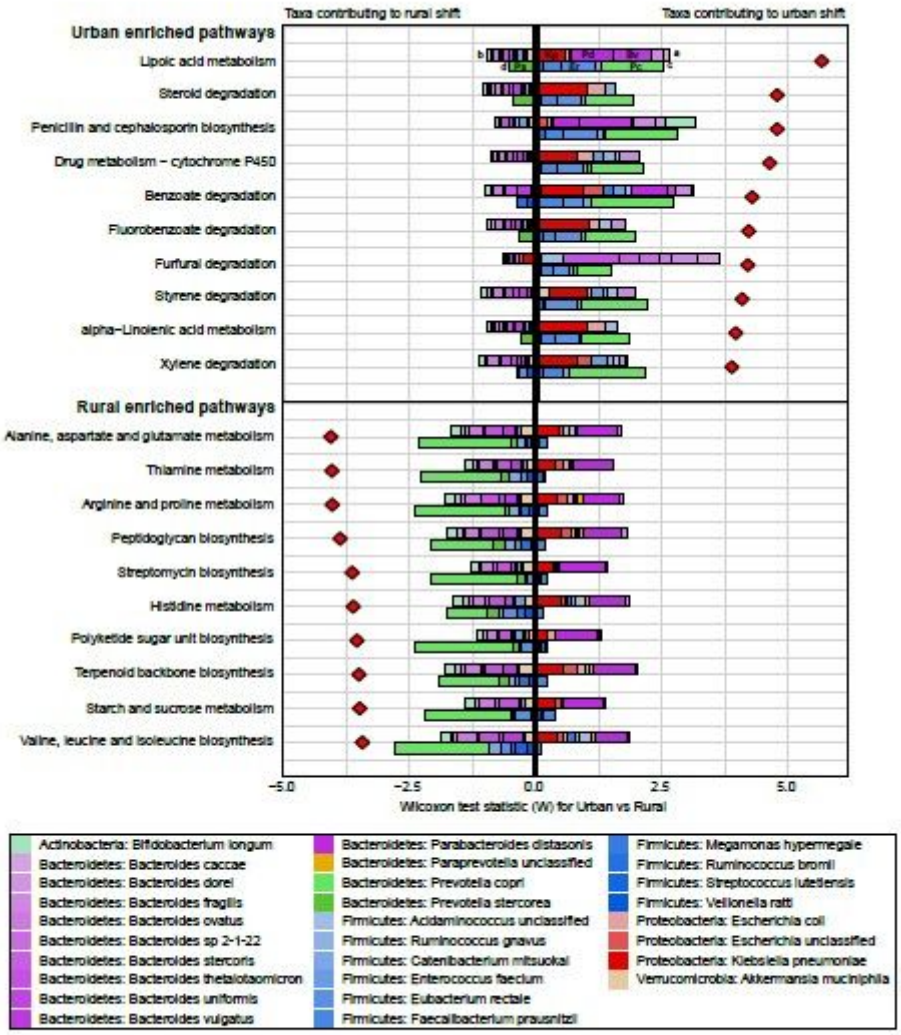


Figure 5

Taxonomic contributions to differentially enriched metabolic pathways. The top 10 pathways enriched in either urban or rural subjects are shown with the predicted contribution of individual taxa to the overall pathway variance (red diamonds). For each pathway, the top and bottom bars indicate urban and rural associated taxa respectively, displaying the predicted contribution of each taxon to enrichment in either group; urban (positive) or rural (negative). For example, enrichment of Lipoic acid metabolism in urban subjects is associated with the positive contribution (a) of Klebsiella pneumoniae (Kp), Parabacteroides distasonis (Pd) and Bacteroides vulgatus (Bv), with only minor negative contributions from multiple other species (b). Rural associated taxa contributing to enrichment in urban subjects (c), most likely because they encode the function sparsely, include Prevotella copri (Pc) and Eubacterium rectale (Er). Prevotella stercorea (Ps) is predicted to enrich this pathway in rural subjects (d), acting against the total observed shift.

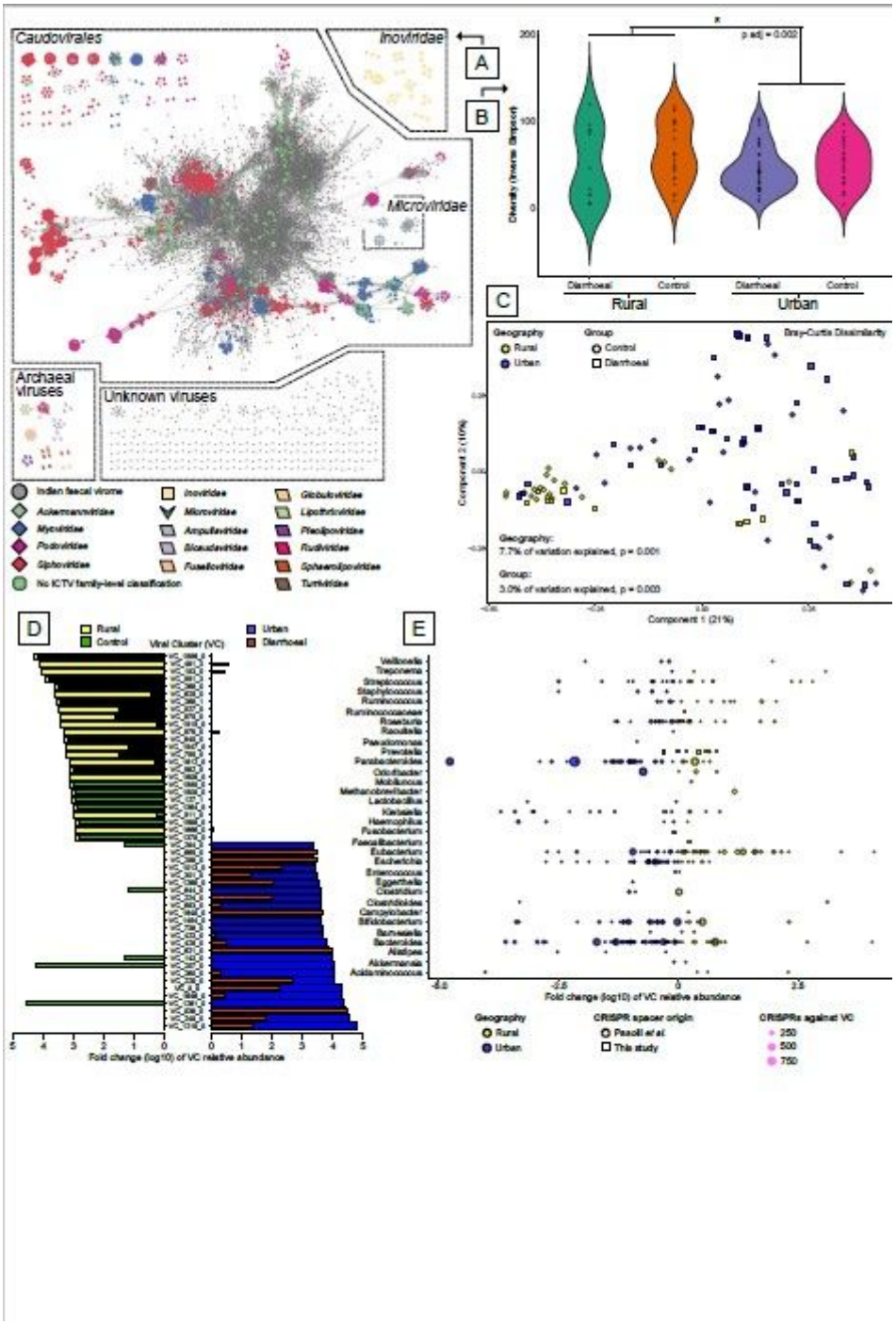


Figure 6

Contrasting faecal viromes by geographic location and diarrhoeal status. (A) Network visualisation of viral clustering.Viral clusters (VCs) containing previously characterised viral sequences (viral RefSeq 85) are coloured by International Committee on Taxonomy of Viruses (ICTV) family-level taxonomic assignments. While Microviridae VCs are connected to Caudovirales through shared protein clusters, these taxa are unrelated. (B) Inverse Simpson diversity comparisons of subjects by diarrhoeal status and geographic location. (C) Principal coordinate analysis of VC profiles based on Bray-Curtis Dissimilarity. (D) The fold change (log10) of the top 25 most abundant rural and urban VCs, with superimposition of the same VC's association with either health or diarrhoeal status. (E) The fold change (log10) of all VCs

relative abundance that are targeted by CRISPR spacers from identifiable bacterial genera. Each point represents a VC, with size representing the aggregate number of CRISPR spacers targeting individual viruses within a cluster.

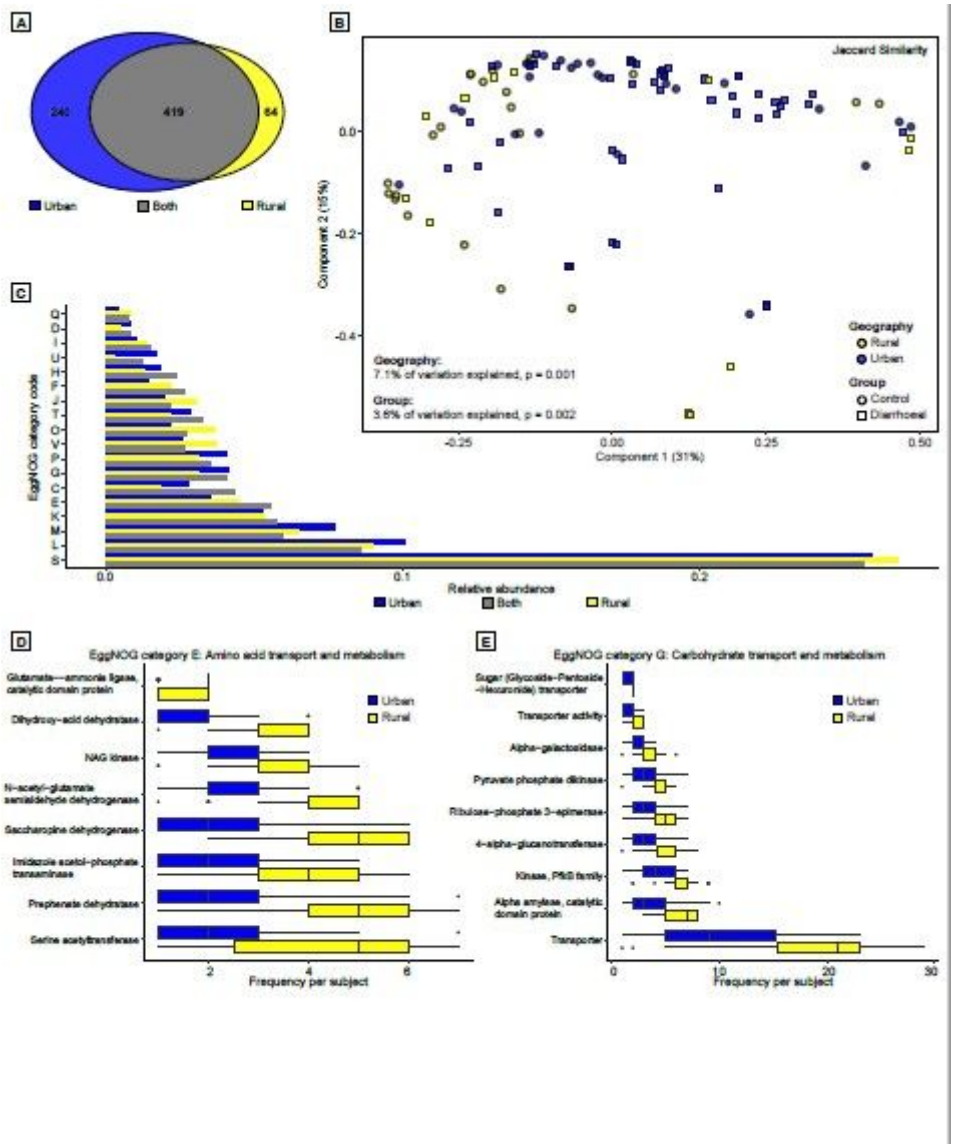


Figure 7

Examination of the auxiliary metabolic potential of human faecal viruses. (A) Shared proteins encoded by Viral Clusters (VCs) shared amongst 10 or more individuals within this study. (B) The VC-encoded metabolic functions were determined per individual virome, with the similarities between subjects visualised by principal coordinate analysis using the Jaccard index. (C) Relative abundance comparisons of the protein categorical-function predictions of VCs by residence. (D & E) The observed frequency of amino acid transport and metabolism functions, and carbohydrate transport and metabolism functional predictions encoded by individual virome VCs. Only statistically significant EggNOG functional predictions are displayed (Mann-Whitney U test with Bonferroni correction, p adj = 0.05).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Monaghan2019supplementarymaterial2.xlsx
- Monaghan2019supplementarymaterial1.docx