BMC Microbiology

RESEARCH

Open Access

Analysis of an Indian colorectal cancer faecal microbiome collection demonstrates universal colorectal cancer-associated patterns, but closest correlation with other Indian cohorts

Mayilvahanan Bose[1*], Henry M. Wood[2*], Caroline Young[2], International C. R. C. Microbiome Network (AMS/CRUK), Philip Quirke[2] and Ramakrishan Ayloor Seshadri[1]

Abstract

It is increasingly being recognised that changes in the gut microbiome have either a causative or associative relationship with colorectal cancer (CRC). However, most of this research has been carried out in a small number of developed countries with high CRC incidence. It is unknown if lower incidence countries such as India have similar microbial associations.

Having previously established protocols to facilitate microbiome research in regions with developing research infrastructure, we have now collected and sequenced microbial samples from a larger cohort study of 46 Indian CRC patients and 43 healthy volunteers.

When comparing to previous global collections, these samples resemble other Asian samples, with relatively high levels of *Prevotella*. Predicting cancer status between cohorts shows good concordance. When compared to a previous collection of Indian CRC patients, there was similar concordance, despite different sequencing technologies between cohorts.

These results show that there does seem to be a global CRC microbiome, and that some inference between studies is reasonable. However, we also demonstrate that there is definite regional variation, with more similarities between location-matched comparisons. This emphasises the importance of developing protocols and advancing infrastructure to allow as many countries as possible to contribute to microbiome studies of their own populations.

Keywords Colorectal cancer, Cancer microbiome, Indian microbiome

*Correspondence:
Mayilvahanan Bose
Mayilvahanago2mayil@gmail.com
Henry M. Wood
h.m.wood@leeds.ac.uk
[1] Cancer Institute (WIA), Chennai, India
[2] Pathology and Data Analytics, Leeds Institute of Medical Research at St. James's University Hospital, University of Leeds, Leeds LS9 7TF, UK

Importance

Colorectal cancer is increasing in many countries, thought to be partly due to the interaction between gut bacteria and changing diets. While it is known that populations in different parts of the world have very different gut microbiomes, the study of their role in colorectal cancer is almost exclusively based in the USA and Europe. We have previously shown that there is overlap between the colorectal cancer microbiome in multiple different countries, establishing robust protocols in the

process. Here we expand that into a new Indian cohort. We show that while there are similarities between countries, by concentrating on one country, we can uncover important local patterns. This shows the value of sharing expertise and ensuring that work of this nature is possible wherever this disease occurs.

Introduction

Colorectal cancer (CRC) is the second biggest cause of cancer-related deaths, globally. Although incidence has been higher in more developed countries, cases are rising in countries with traditionally lower rates [1]. Over recent years, there has been a growing focus on the associations between the gut microbiome and the development of CRC, whether this is in the form of specific bacteria or bacterial toxins thought to have a role in carcinogenesis [2–4] or as an association with the overall gut flora, that could be informative as to the overall biology of the cancer, or useful in screening [5–7].

A major obstacle to a comprehensive understanding of the role of the microbiome in CRC biology (and microbiome studies in general) is the lack of diversity of sample populations. A recent meta-analysis of 444,829 publicly available microbiome samples from 2,592 studies found that, where origin was known, 71% of samples were from North America or Europe [8]. The South Asian subcontinent was particularly under-represented, making up just 1.8% of samples, despite the region accounting for around a quarter of the global population, and an increasing CRC incidence [1]. Studies focusing on Indian microbiomes have demonstrated a marked difference with Western populations [9–11], with taxa such as *Prevotella* being noticeably more abundant in Indian samples. This means that published associations between CRC and the microbiome may not be relevant to Indian (or other under-represented) patients. Studies of Indian populations have mainly sought to understand the microbiome of healthy individuals, or are group-specific studies comparing rural with urban, tribal or other localised population patterns. Sample numbers of studies examining the CRC microbiome in under-represented populations have mostly been limited, and have not been compared to global patterns, in order to better understand the similarities and differences between countries with different CRC incidences.

As part of an effort to address this imbalance, we have previously established an International CRC Microbiome Network seeking to advance the study of the microbiome of CRC in under-represented countries (namely India, Vietnam, Argentina and Chile in the initial phase). As an extension of the process of data generation, we participated in knowledge and expertise exchange. We also optimised cost-effective sample collection and storage protocols alongside sequencing and analysis strategies that are robust enough to be reproducible in countries which may not have the infrastructure of better-resourced facilities. We demonstrated that storing faecal samples on guaiac faecal occult blood (gFOBT) cards followed by 16S V4 rRNA sequencing allows long-term room-temperature storage and shipping of samples prior to processing, and produced data showing country specific microbial patterns as well as an international CRC bacterial signature [12]. This work built on our previous study which showed that 16S data produced from gFOBT faecal storage was similar to that from frozen samples, with no changes to floral composition or abundance [13].

This previous work was characterised by regional cancer centres in the aforementioned countries collecting samples, followed by shipping to the UK for processing. In the current study, we present the next logical step in the International CRC Microbiome Network aims of developing local expertise. We collected stool samples from a new cohort of 90 Indian CRC patients and healthy volunteers, and processed and sequenced them in India. We compared them to our previous data and to external cohorts to demonstrate the infrastructural and technical compatibility of the Indian institutions in carrying out microbial research into their own local cohorts, thereby demonstrating the value of country-specific analyses in globally significant fields such as CRC microbiome research.

Methods

All research was carried out at The Cancer Institute (WIA), Chennai, India. Wherever possible, we used protocols developed during our previous work to establish a global CRC microbiome network [12], adapted for local use.

Patient and healthy volunteer recruitment

Samples used in the study were collected between October 2018 and November 2019. Cases include CRC patients attending the Cancer Institute (WIA) and controls including healthy volunteers who were asymptomatic and unrelated to any gastrointestinal tract / genitourinary tract cancer patients. Inclusion criteria were: Subject age of 18 or over at time of enrolment; Subject must be able to provide signed and dated informed consent and be able to attend scheduled follow-up appointments; Cases must have a confirmed diagnosis of CRC, no previous history of cancer, and be willing to provide stool specimens; Controls must have not previous history of cancer and be willing to provide stool specimens. Exclusion criteria were: Subjects are unwilling or unable to provide informed consent; Use of any systemic antibiotic, antifungal, antiviral or antiparasitic drugs in

the previous six months; Use of oral, intravenous, intramuscular, nasal or inhaled corticosteroids in the previous six months; Use of methotrexate or immunosuppressive cytotoxic agents in the previous six months; Use of large doses of commercial probiotics (greater than or equal to 10^8 cfu or organisms per day) in the previous six months (including products were probiotics are the primary component, but not ordinary dietary components such as fermented beverages, yoghurts etc.); Unstable dietary history defined by major changes in diet in the previous month, where the subject has eliminated or significantly increased a major food group in the diet; Positive test for HIV, HBV or HCV; Severe co-morbidity conditions such as uncontrolled diabetes or hypertension; Systemic autoimmune disorders; Major surgery to the gastro-intestinal tract, with the exception of cholecystectomy and appendectomy, in the previous five years; Any major bowel resection at any time.

As there was no intervention in this study, no randomisation was carried out.

Ethical approval

This study was performed in accordance with the Declaration of Helsinki with the approval from Institutional ethics committee (IEC/2018/01) at Cancer Institute (WIA), Chennai and Indian Council of Medical Research, study reference number (2018–0337). All patients and healthy volunteers gave appropriate informed consent.

Sample collection

Patients and healthy volunteers each provided a stool which was applied to all windows of a gFOBT card, and developer solution added within six hours. Once dry, cards were stored in sealed bags at ambient temperature until batch processing.

DNA extraction

Three faecally loaded squares were excised using sterile scalpels, and DNA was extracted from gFOBT cards using a modified version of the QIAamp DNA mini kit (Qiagen, Germany) with additional Buffer ASL (Qiagen, Germany) as previously described [12].

16S rRNA library preparation and sequencing

We used the Earth Microbiome Project (EMP) 16S Illumina library preparation protocols [14] with Illumina 16S V4 primer constructs 515F (Parada)-806R (Apprill) [15], as previously described [12].

All libraries were pooled and sequenced by MedGenome Labs (Bangalore, India) on a single run of an Illumina MiSeq, using 2×250 bp paired-end reads.

Data processing

Reads were stripped of Illumina adapters using cutadapt [16]. Further processing was carried out using QIIME2 (version 2019.10) [17]. Reads were trimmed to 220 bp to remove poor quality base calls, before denoising, pair merging and representative sequence assignment using DADA2 [18]. Taxa were assigned to the SILVA database (version 132) [19] using BLAST + [20], implemented within the QIIME2 q2 feature classifier plugin [21].

Within QIIME2, sequences were rarified to the level of the lowest-depth sample (43,000 QC-passing reads) before diversity analyses. Shannon index alpha diversity [22] of each sample was calculated, as well as Bray–Curtis beta diversity distances [23] between samples. Adonis PERMANOVA tests [24] were used to compare Bray–Curtis distances to sample metadata. Bray–Curtis distances were visualised using principle coordinate plots.

Taxa tables, and alpha and beta diversity measures were exported from QIIME2 for further analysis and visualisation in R (version 4.0.5). Random forest models [25] were built using randomForest [26] and validated using pROC [27]. Taxa significantly associated with metadata were called using LEfSe [28].

Comparisons to external data

We compared this dataset to our previous dataset [12] by merging QIIME2 tables, before processing as described above. We also compared to a metagenomic faecal dataset of Indian CRC samples [29], by examining genus-level taxonomic assignments of the two groups of samples.

Results

Sample numbers and sequencing quality

We recruited 47 CRC patients and 43 control healthy volunteers. Of these, one CRC patient failed sequencing QC, leaving 46 patients and 43 controls. Age and gender, comorbidities and information on the consumption of tobacco, meat and alcohol was collected for all participants, with additional, tumour-specific metadata available for the CRC patients, summarised in Table 1.

Between 43,443 and 252,721 (median 120,789) sequences were called per sample, following pair merging and denoising by DADA2. This is a higher number than the Indian samples from our previous study [12], (minimum 58,984, maximum 128,124, median 106,324). However, feature number is mostly a reflection of sequencing depth. When all samples were rarefied to a similar depth and Shannon diversity compared (supplementary figure S1), the two datasets were comparable. The samples from the current study had a slightly higher diversity, but this was not significant (Mann–Whitney $p = 0.059$). The

Table 1 Patient and healthy volunteer demographics and metadata

Sample details	Controls	CRC
Total	43	46
Gender		
Male	29	26
Female	24	20
Age		
Range	23–76	21–76
Median	38	51
Comorbidities		
Diabetes	6	9
Hypertension	1	6
Personal habits		
Smoker	2	4
Alcohol	5	7
Meat eater	28	41
Cancer specific data		
Site of cancer	Right sided colon	6
	Left sided colon	15
	Rectum	25
Grade	Grade 1	1
	Grade 2	34
	Grade 3	11
Stage	Stage 1	0
	Stage 2	8
	Stage 3	37
	Stage 4	1

slightly higher measured diversity could be a function of the sequencing technology used. The current study was performed using an Illumina MiSeq with 2×250 bp reads to measure a 240 bp PCR product, whilst the previous study used an Illumina HiSeq with 2×150 bp reads. The MiSeq data deteriorated less over the span of the read, and combined with the longer read length might have made a range of representative sequences easier to reliably detect. Alternatively, the small sample numbers of the previous data (20 Indian samples) might have resulted in a slightly unrepresentative sample set.

Adonis PERMANOVA analysis of Bray–Curtis distances between the current study and the previous Indian samples, showed that sample status (cancer vs healthy volunteer) was associated with the largest proportion of measured differences, followed by which dataset the sample belonged, then gender and age (supplementary figure S2).

Comparison with previous data

We first compared the current data with our previous global comparison using principle coordinate plots of Bray–Curtis distances (Fig. 1). As we saw with our previous study, there appeared to be more geographic separation than separation by cancer status. Although the plot is now dominated by Indian samples, so biased, the South American samples are all in the left side of the plot, and Vietnamese samples mostly in the top half. Adonis PERMANOVA analysis indicated that sequencing run was associated with 3.4% of variation, country of origin with 6.7%, and cancer status with 2.3%, all of which were significant ($p < 0.001$).

Next, we compared the most prevalent taxa of the two datasets (Fig. 2). As with the principle coordinate plot, there was clear geographical separation. The Asian cohorts were characterised by far greater abundance of *Prevotella*, and relatively less *Bacteroides*. All the cancer cohorts had more *Escherichia/Shigella* than their respective control cohorts. We tested for taxa differentiating between cancer and healthy volunteer for a merged dataset of all countries from both cohorts, and again with just Indian samples, using LEfSe [28]. This allowed us to examine taxa outside of the most common genera. (Supplementary figure S3). For both comparisons, known CRC taxa such as *Akkermansia*, *Fusobacterium* and *Ruminococcus* were evident, reiterating our previous findings that there appears to be a consistent CRC-associated faecal flora. Taxonomic assignments are given in full in table S1.

We also compared alpha diversity of cancer versus healthy volunteer for the different countries (Fig. 3). As we had previously showed, Indian (and Vietnamese) samples have lower diversity than South American samples. It was previously seen that cancer samples had higher diversity than healthy volunteers for all the countries, although the small sample numbers for the respective countries made this difficult to claim with certainty. This trend can now be confirmed for Indian samples, given the increased sample number (Mann–Whitney $p = 0.0005$).

Prediction of cancer status using previous data

As well as visually inspecting similarities between our previous global dataset and our current study, we sought to ascertain if the previous data could be used to predict cancer status in the current samples, and vice versa.

For both the Young et al. cohort and our new cohort, we generated random forest models and used them to predict cancer status in their own samples, and then validated them in the other dataset (Fig. 4). There was very good concordance. The previous samples had an area under the curve (AUC) of 0.77 when used to predict their own samples, which only dropped to 0.76 when our current study was used as the original model. Our current study had an internal AUC of 0.86, which dropped to 0.85 when cancer status was predicted using the model from the Young et al. cohort. It is interesting

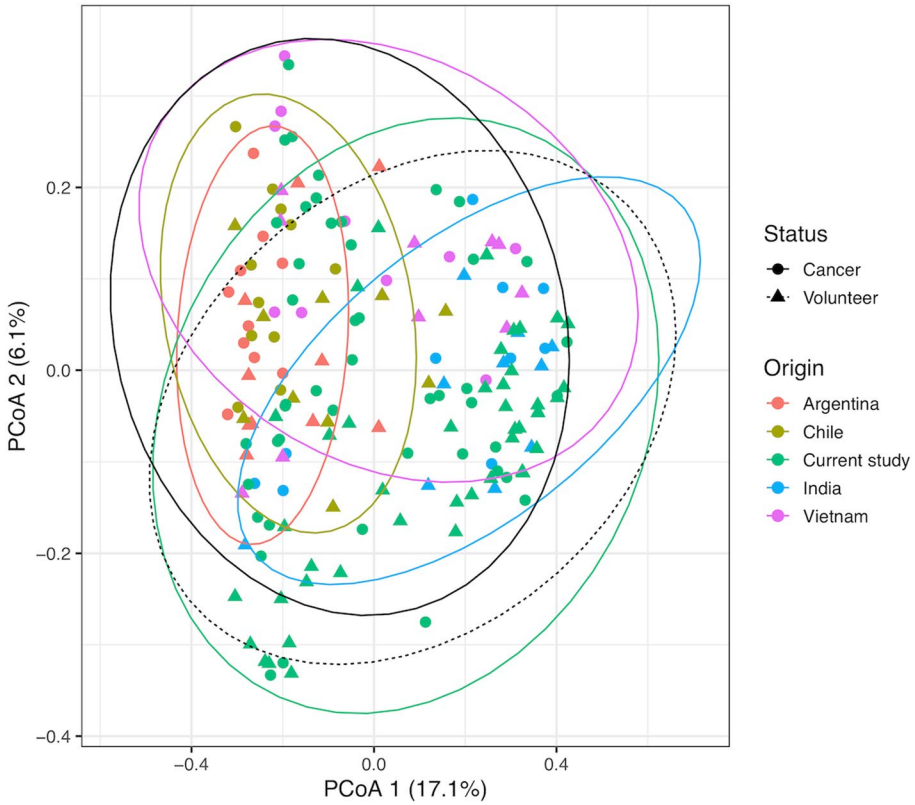

Fig. 1 Principle coordinate plot of Bray–Curtis distances of the current study, compared to global samples from Young et al. "India" refers to previous Indian samples. 95% confidence intervals for sample groupings are displayed as ellipses

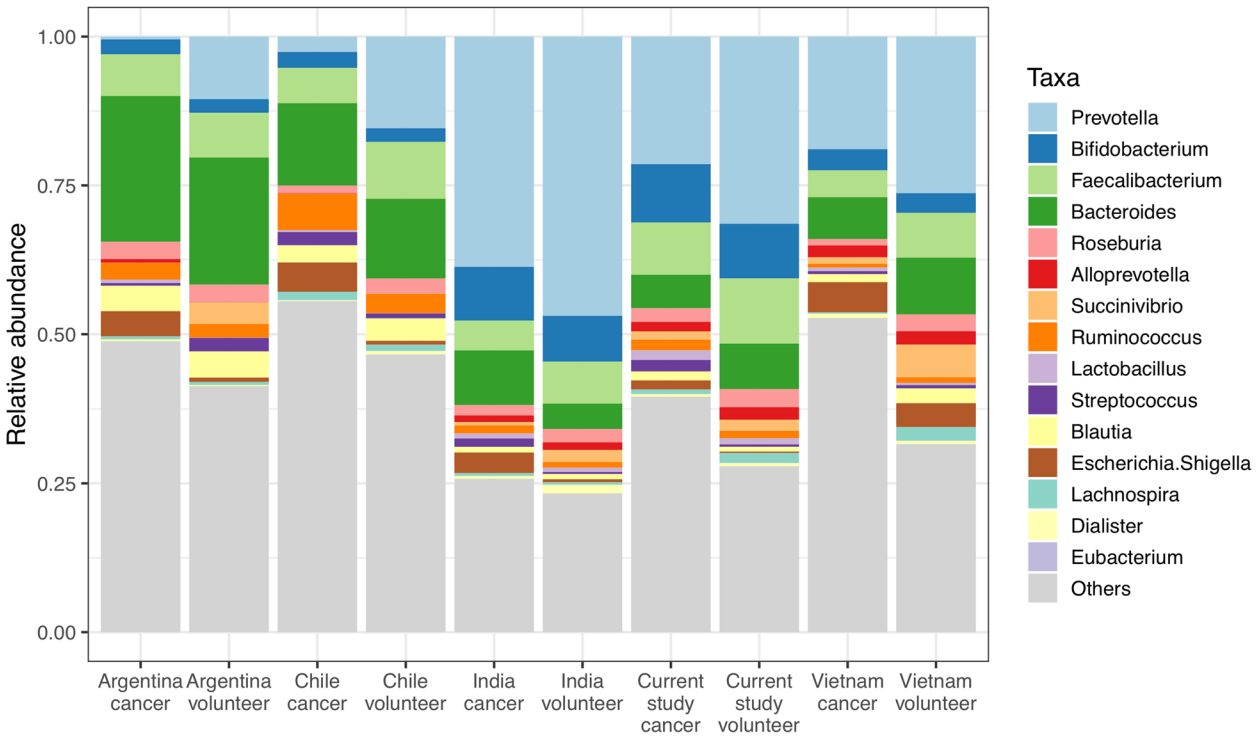

Fig. 2 Taxa abundance of the current study and Young et al. The top 15 genera across all cohorts are displayed

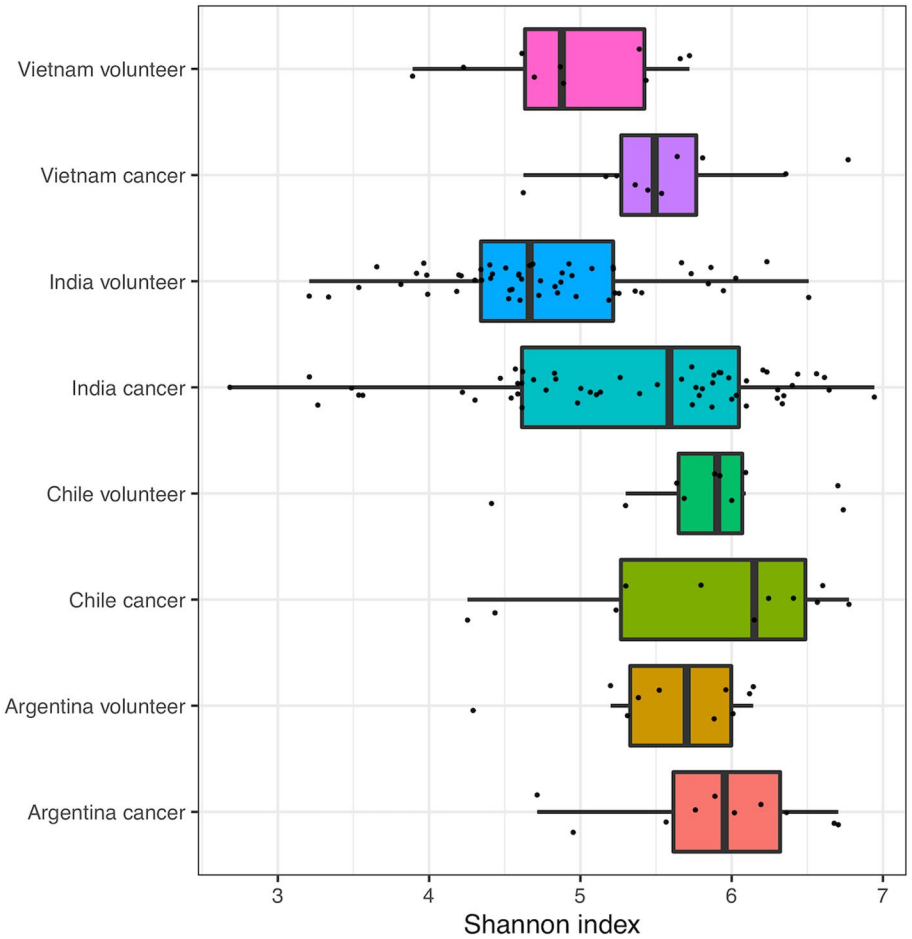

Fig. 3 Shannon index alpha diversity of the current study and Young et al. Indian samples from the two studies have been merged

that prediction status of our current dataset based on the model of our previous samples performs better than our previous cohort predicting the status of its own samples. Perhaps using samples from only one country gives a more homogenous dataset, or maybe Indian samples are easier to predict, with more dramatic differences in levels of predictive taxa such as *Prevotella* between cancer patients and healthy volunteers. The theory that Indian samples may be easier to predict is corroborated by meta-analysis carried out within the Young et al. paper. Ten studies, including that of Young et al. and the Gupta et al. cohort if Indian CRC and healthy volunteers. A random forest model was built of each cohort and validated against every other cohort. The validation using the Indian Gupta cohort had the highest AUC value for five of the nine models (including the model based on Young et al.) as well as the highest mean AUC.

These results repeat our previous findings that there appears to be a consistent, global pattern of changes in the faecal microbiome of CRC patients, and that if this is adequately catalogued, can be used to predict the cancer status of new samples.

Comparisons with other Indian datasets

There are a limited number of studies profiling Indian faecal microbiomes. We visually compared our data to three studies profiling healthy Indian healthy volunteers [9–11]. Whilst these studies had different aims, largely based on profiling the variation across India and between Indian and non-Indian populations, they were all characterised by the dominance of *Prevotella* amongst a large proportion of their samples, as we observed in both our healthy volunteers and cancer patients.

We were only aware of one other study comparing the microbiomes of CRC patients to healthy volunteers in Indian samples, based on individuals from Bhopal and Kerala [29].

First, we compared the alpha diversity of CRC and volunteer samples of the Gupta metagenomic dataset, both in the original species calls, and converted into genus calls, to match our 16S genus calls.

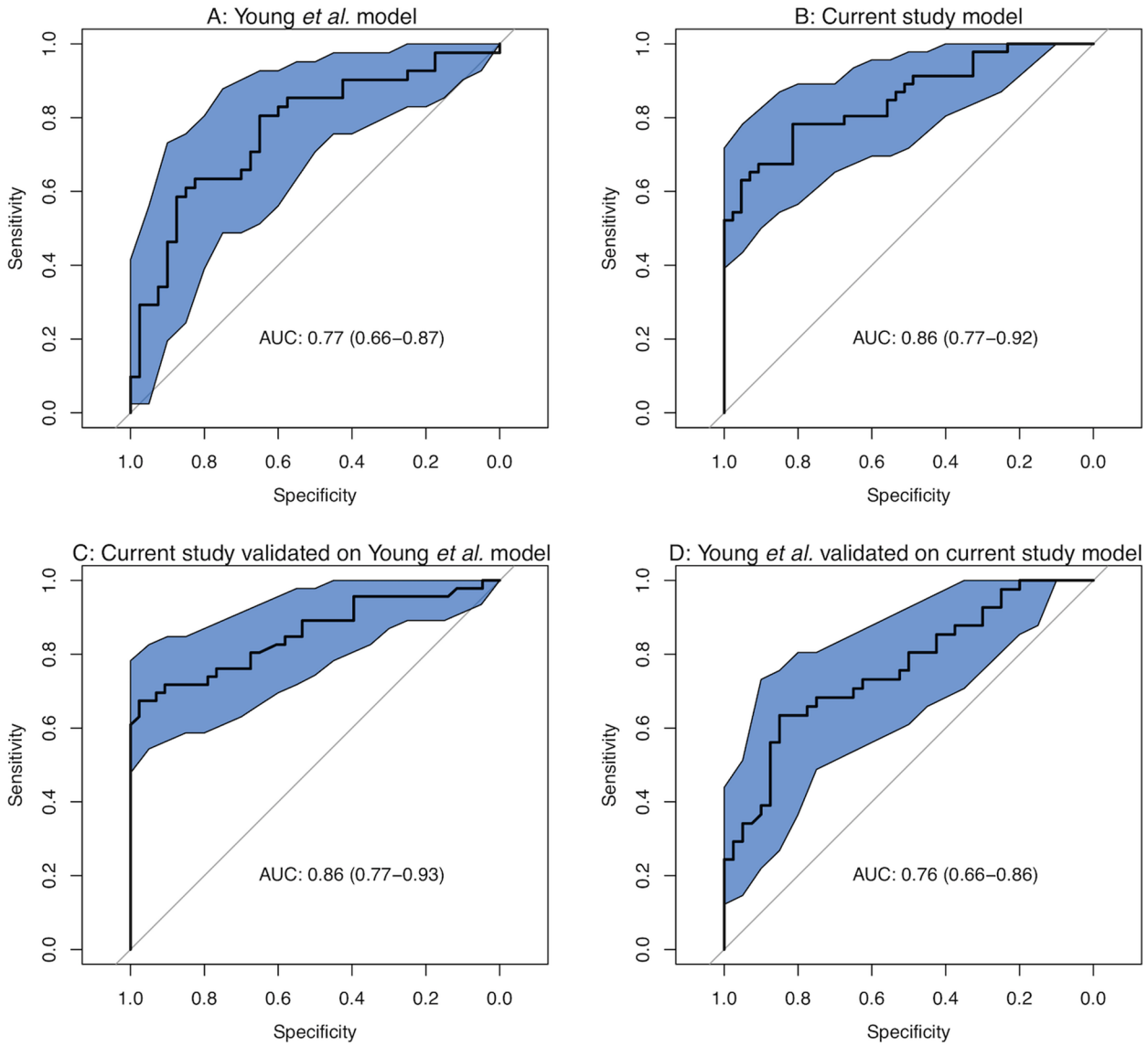

Fig. 4 Using random forest models of **A** our previous global dataset and **B** our current study to predict cancer status in their own samples and **C**, **D** to predict cancer status in samples from the other study

Both the species and genus calls of the Gupta dataset repeated the observation in our data that CRC alpha diversity was significantly higher than volunteer samples, compared to CRC diversity being lower in more commonly studied countries. This is displayed in Fig. 5. We are unable to ascertain if the higher diversity in our dataset compared to the Gupta cohort is a technical artefact of sequencing strategy, or a real, geographical difference.

We then compared the CRC and healthy volunteer taxa in both groups of samples.

Visually, whilst there were differences, some patterns were consistent between the datasets (Fig. 6). *Prevotella* was higher in control populations for both groups while *Escherichia/Shigella* was higher in CRC samples. When comparing using LEfSe (Supplementary figure S3), there were several similarities. Out of 43 CRC-associated taxa in the current study, 13 were amongst the 28 CRC-associated taxa from Gupta et al. Out of control-associated taxa, two were shared between four taxa called from the current study and six from Gupta et al. The main discrepancy was that *Bacteroides* was associated with healthy controls in the current study and CRC in the Gupta cohort, possibly due to differences between 16S and whole metagenomic sequencing strategies, but more likely due to using different cohorts from different regions of India. Different populations may have different

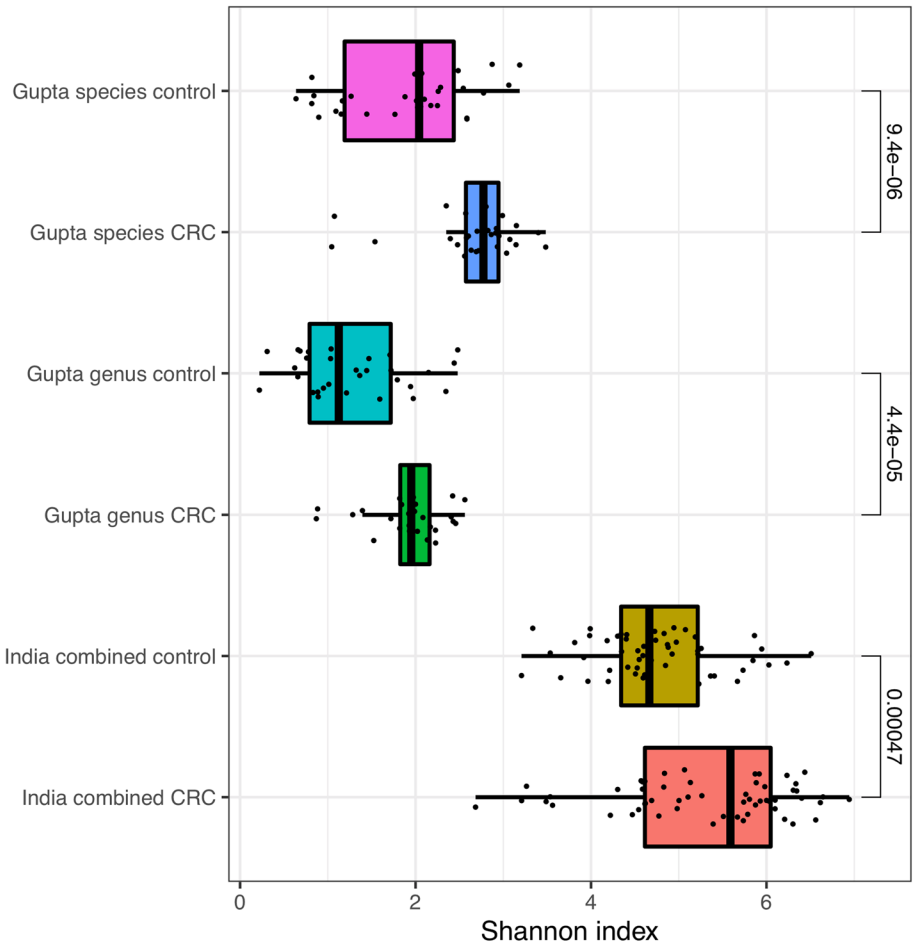

Fig. 5 Shannon index alpha diversity of the Indian samples from our cohorts (India combined), and the genus and species level calls of the Gupta et al. dataset. Mann–Whitney p-values are displayed above each CRC/control comparison

prevalent species of *Bacteroides*, with some relatively benign and some containing CRC-associated strains [30].

Adonis PERMANOVA analysis of location and cancer status of the merged datasets suggested that location within India was associated with 18% of the variation, whilst cancer status was associated with 4.4%. However, it is impossible to know how much of the apparent location-associated variation is in fact a product of sequencing strategy. By removing our samples, thus keeping sequencing strategy uniform, location was associated with 7.8% of variation and cancer status with 10.8%. All associations in both comparisons were significant ($p < 0.001$).

As well as simply counting taxa which fell into different subgroups, we built random forest models of our data and that of Gupta et al. and validated them with the other dataset (Fig. 7). Again, there was good concordance. The cancer status of both datasets were predicted better using their own samples than the models built with the alternative dataset, but with overlapping confidence intervals. This demonstrates again that taxa calls from one dataset can reliably predict cancer status in another, despite completely different sequencing strategies being used.

Associations with metadata

Lastly, and acknowledging the relatively small size of our cohort, we examined whether any of the different patient groupings of our CRC samples were associated with changes in the microbiome.

When comparing different metadata categories using adonis PERMANOVA of Bray–Curtis distances, only anatomical site and sex were significant. We compared anatomy, age and sex on a principle coordinate plot which reiterated these findings. Although there was overlap, all the right-sided colon samples were in the top half of the plot, and almost all the female samples in the left half (Fig. 8).

When comparing taxa for the different metadata categories (Fig. 9), a number of differences could be seen, most noticeably in the relative abundances of the most

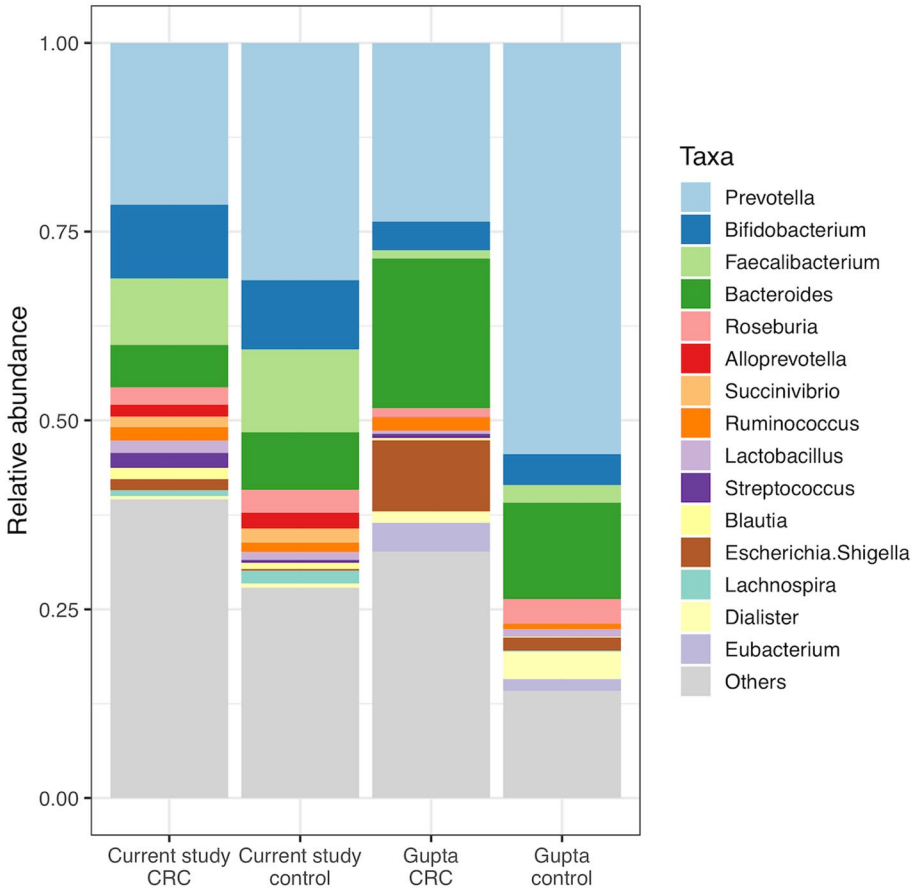

Fig. 6 Taxa abundance of the current study and Gupta et al. The top 15 genera across all cohorts are displayed

common genera *Prevotella*, *Bifidobacterium*, *Faecalibacterium* and *Bacteroides.* It should be noted that some sample numbers of some groupings are very small. For instance, there were only six right-sided colon samples. Therefore, we are unwilling to analyse this aspect of the data too deeply, and present it more as an example of what kind of analyses would be possible with larger sample numbers.

Discussion

Given the known paucity of microbial datasets from outside of Europe and North America [8], it is vital for the community to strive to increase the number and size of available datasets for a better understanding of the role of the microbiome in health and disease. In particular, it is important to gain an understanding of which groups of taxa or specific components of the microbiome such as toxins [4] are important for non-invasive CRC diagnosis, or as potential therapeutic targets. As the field of CRC microbiome studies advances, it is important to know which developments are universal, and which are confined to limited geographical regions. This is particularly evident when examining the limited number of Indian microbial datasets [9–11, 29] which consistently report that *Prevotella* is the dominant genus present, in contrast with most other countries examined. In this context, it becomes important to know if putative oncomicrobes or potential screening targets are of value in Indian populations. The next most common genera were *Bifidobacterium* and *Faecalibacterium*. Both these taxa are associated with good gut health, and were the most important genera associated with healthy controls in a combined random forest analysis of ten CRC/control datasets from multiple countries [7]. If they offer some level of protection against CRC development, or are associated with a healthy bowel environment, then the high levels of these taxa in India may be associated with the relatively low incidence of CRC.

Having previously set up an international network of institutions from countries with under-developed microbial research infrastructure, we have been developing robust protocols and producing CRC microbiome data from those countries, demonstrating marked differences

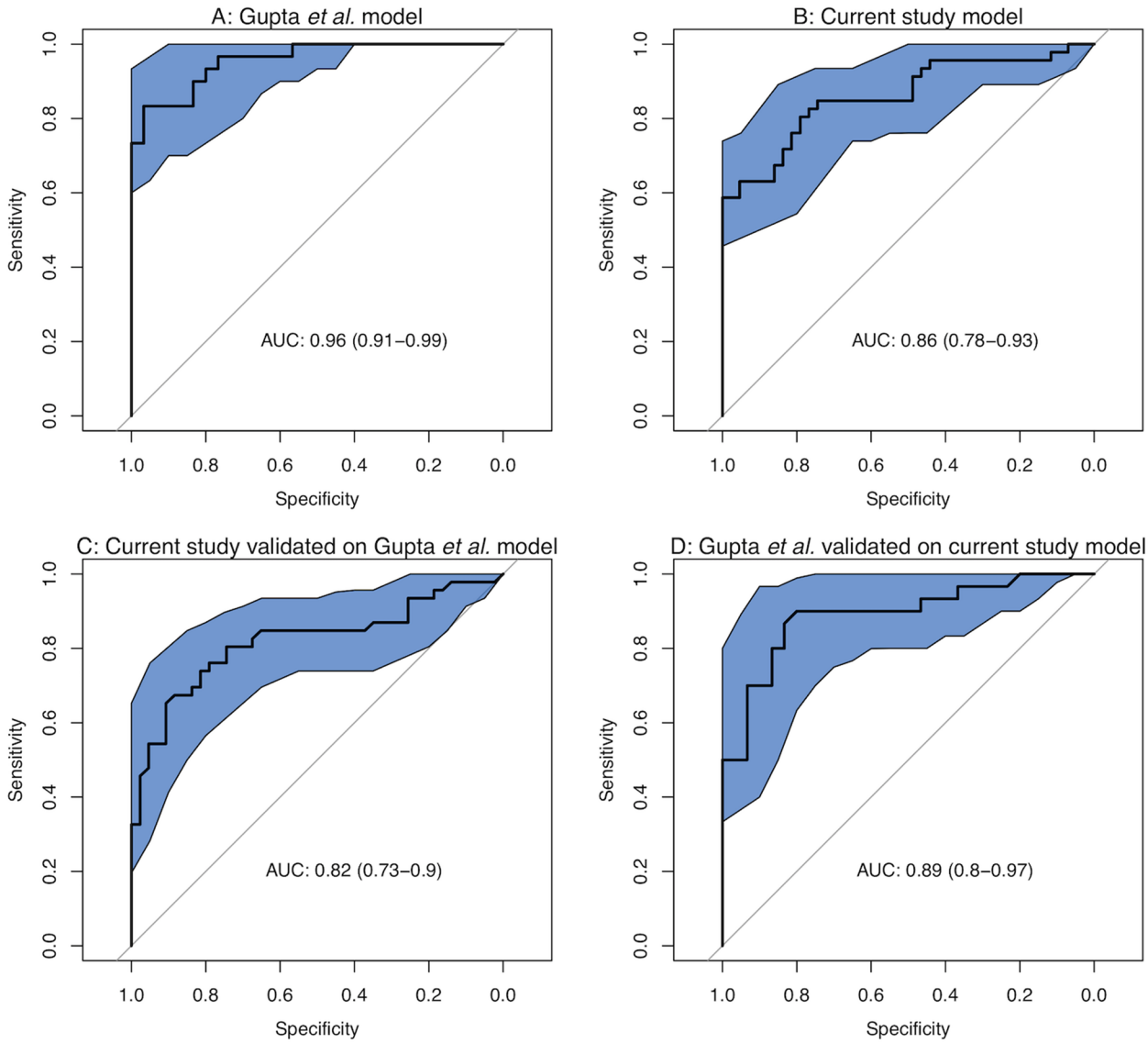

Fig. 7 Using random forest models of (**A**) Gupta et al. and (**B**) our current study to predict cancer status in samples in their own samples and (**C**, **D**) to predict cancer status in samples from the other study

between regions, yet with promising evidence of a universal CRC faecal microbiome [12].

As a critical output resulting from the establishment of that knowledge transfer network, we are now able to present a new cohort of CRC and control microbiomes, wholly collected and processed in India. We compared these samples to the Indian samples from our previous collection, sequenced in the UK, and found them to be largely comparable. The new data was of slightly higher quality, and separated from the old data in diversity analyses. We attribute this to the sequencing technology used, with less quality deterioration across the length of sequencing reads observed in the MiSeq reads than in the HiSeq reads of the previous study. Therefore, we have demonstrated that Indian institutions can advance their microbiome research associated with healthcare benefits, in a systematic manner comparable with that of the datasets produced elsewhere. While this study may have a modest sample size of 46 CRC samples and 43 controls, any dataset in populations which are under-represented in microbial research is to be welcomed. By utilising standard sample collection and sequencing strategies, we are able to combine this dataset with our previous study, or future work which has yet to be carried out. All conclusions are based on statistical tests which take sample size into account.

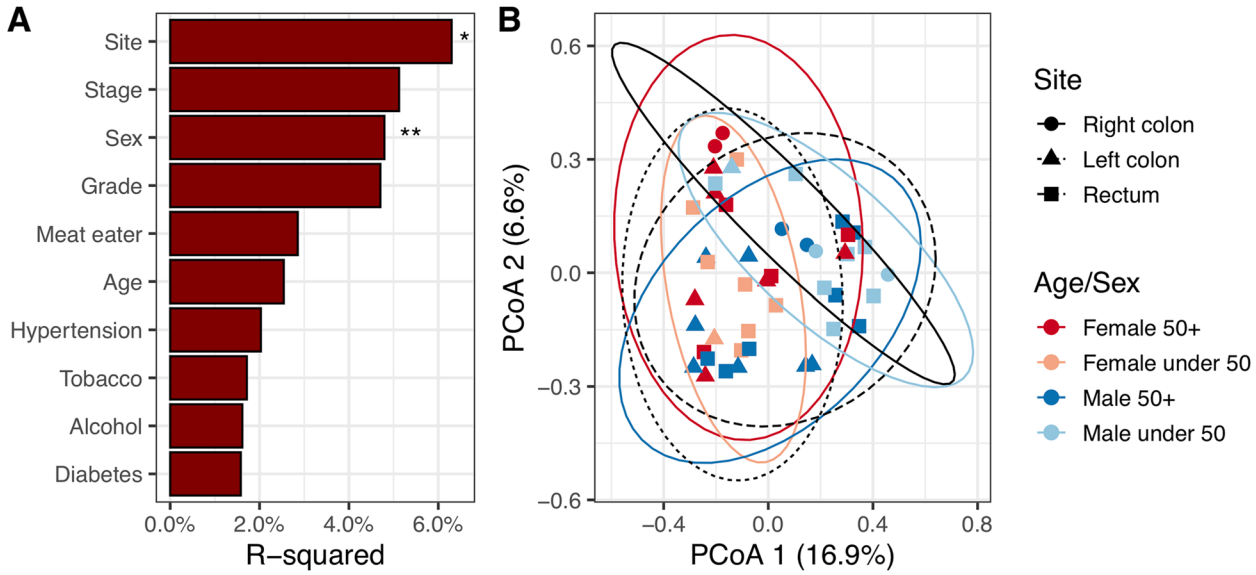

Fig. 8 Beta diversity analyses of current study cancer samples. **A** Adonis PERMANOVA comparison with metadata. R-squared refers to amount of Bray–Curtis variation associated with each category. *P*-value is indicated by: **—$p <= 0.01$; *—$p <= 0.05$. **B** Principle coordinate plot of Bray–Curtis distances. Shape of points refers to anatomical site, while colour refers to age and sex. 95% confidence intervals for sample groupings are displayed as ellipses

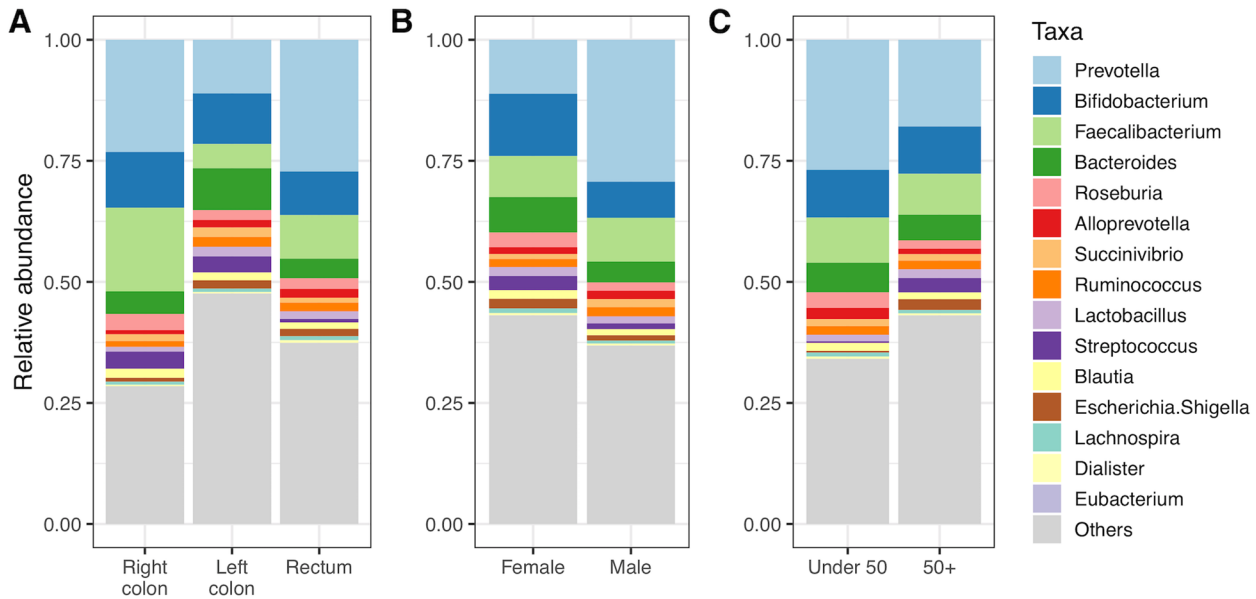

Fig. 9 Taxa abundances of the cancer samples split by (**A**) anatomical site, (**B**) sex and (**C**) age

Moving from logistics to biology, we compared our new dataset to all the samples from our international network and external studies. As seen before, our Indian samples were characterised by a high abundance of *Prevotella*, especially in the healthy volunteers, and an increase in alpha diversity amongst the cancer samples. This pattern of higher CRC diversity was also seen in the Gutpa et al. dataset, which was generated using metagenomic sequencing rather than 16S methodology. This is contrary to the pattern seen in European and North American samples, such as our own UK cohort [7]. As we have seen the same pattern in two Indian cohorts with different methodology, it is likely that this is a real phenomenon and shows again the value of studying under-represented

populations. It has previously been suggested that the dominance of *Prevotella* in healthy Indian samples leads to low diversity, and that the relative decrease in *Prevotella* in Indian CRC samples allows more taxa to flourish, resulting in increased diversity [29].

There were differences between our dataset and that of Gupta et al. Despite being the dominant genus, *Prevotella* was less abundant in our samples. This could be a bias of sequencing strategy, or it could be differences between Indian regions. Comparisons of the datasets with and without mixed sequencing strategies suggested that both factors had measurable effects. Our samples were all from Chennai region, whereas the Gupta samples were taken from Bhopal and Kerala, several hundred kilometres away, with different local conditions and prevalent diets. It needs to be highlighted at this juncture that the socio-economic differences, and regional/cultural diversity will have a major effect on the dietary patterns prevalent in Indian studies and will influence the health status of individuals. Even within India, these differences will affect microbial patterns associated with the gut microbiome in general, and CRC in particular. Previous comparisons of healthy Indian samples have shown considerable regional variation in *Prevotella* and other genera, thought to be largely due to diet, specifically levels of meat intake [9–11]. It is not possible to make exact comparisons with diet between our study and the Gupta cohort, which has no dietary information recorded. A historical study across India [31] suggested that Chennai (21%) has levels of vegetarianism intermediate to that of Bhopal (45%) and Kerala (6%). 22% of our cohort were vegetarians, although the small sample size means that we cannot infer that this is reflective of the population as a whole.

When we used models from our previous and new datasets to predict the cancer status of each other, there was good concordance. Only slight drops of predictive success were observed between internal and external validation. This pattern was repeated when we compared our new dataset with that of Gupta et al., despite the Gupta dataset using a completely different sequencing strategy (metagenomic shotgun sequencing compared to 16S rRNA amplicon sequencing, each of which introduce different biases). Together, these comparisons show that while there does seem to be a universal CRC-microbiome, and that findings can be generalised across continents, there are more similarities between datasets collected within countries, and that efforts should continue to improve the number of samples available from under-represented regions. It may well be that using local control datasets may allow us to improve the predictive value of microbiome studies in bowel cancer screening.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12866-023-02805-0.

Additional file 1: Supplementary table S1. genus-level taxonomic calls for each sample in this study.

Additional file 2: Supplementary figure S1. Number of features per sample called by DADA2 (A) and Shannon index alpha diversity (B) for the current study and the Indian samples from our previous study, Young et al. **Supplementary figure S2.** Adonis PERMANOVA comparison of the current study and Indian samples from our previous study. R-squared refers to amount of Bray-Curtis variation associated with each metadata category. Status is cancer vs healthy volunteer. Study refers to the current study vs the previous work. P-value is indicated by: *** - $p <= 0.001$; ** - $p <= 0.01$; * - $p <= 0.05$;' – $p <= 0.1$. **Supplementary figure S3.** LEfSe results comparing (A) cancer versus volunteer for merged datasets of Young et al with the current study, (B) just the Indian samples of Young et al and the current study, (C) just the current study, and (D) metagenomic samples from Gupta et al.

Acknowledgements
Not applicable.

International CRC Microbiome Network (AMS/CRUK)
Pham Van Nang & Mai Van Doi, Can Tho University of Medicine and Pharmacy, Can Tho, Vietnam; Carlos Vaccaro, Tamara Alejandra Piñero & Julieta Arguero, Instituto de Medicina Traslacional e Ingeniería Biomédica (IMTIB)—CONICET—Instituto Universitario del Hospital Italiano (IUHI), Hospital Italiano de buenos Aires (HIBA), Buenos Aires, Argentina; Luis Contreras Melendez & Camilo Tapia Valladares, Universidad de los Andes, Santiago, Chile.

Authors' contributions
MB, HW, RAS, and PQ: study design and supervision. MB and RAS: acquisition of data and samples. MB, ICMN and CY: sample processing. HW, MB, CY: data analysis. MB, HW, and RAS: drafting of the manuscript. MB, HW, RAS, CY, ICMN and PQ: critical revision of the manuscript. CY, RAS, and PQ: fundraising for the study. All authors read and approved the final version of the manuscript.

Funding
Cancer Research UK Grand Challenge Initiative (OPTIMISTICC C10674/A27140), Academy of Medical Sciences Global Challenges Research Fund Networking Grant (GCRFNG\100433), Pathological Society of Great Britain & Ireland "Visiting Fellowship" (2234). PQ is a National Institute of Health Research Senior Investigator. MB and RAS would like to acknowledge HCL Foundation for their generous funding support in upgrading the bioinformatics division in the Department of Molecular Oncology, Cancer Institute (WIA), Chennai, Tamilnadu, India. The funders had no role in study design, data collection, analysis, and interpretation, or in the writing of the report.

Availability of data and materials
Raw sequence and sample metadata are available from the European Nucleotide Archive, accession number PRJEB53415 (http://www.ebi.ac.uk/ena/data/view/PRJEB53415).
A STORMS (Strengthening The Organizing and Reporting of Microbiome Studies) [32] checklist 1.03 is available at doi: https://doi.org/10.5281/zenodo.6839159.

Declarations

Ethics approval and consent to participate
This study was performed in accordance with the Declaration of Helsinki with the approval from Institutional ethics committee (IEC/2018/01) at Cancer Institute (WIA), Chennai and Indian Council of Medical Research, study reference number (2018–0337). All patients and healthy volunteers gave appropriate informed consent.

Consent for publication
Not applicable.

Competing interests
None.

Received: 1 November 2022 Accepted: 22 February 2023
Published online: 02 March 2023

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71(3):209–49.
2. Bullman S, Pedamallu CS, Sicinska E, Clancy TE, Zhang X, Cai D, et al. Analysis of Fusobacterium persistence and antibiotic response in colorectal cancer. Science. 2017;358(6369):1443–8.
3. Wu S, Rhee KJ, Albesiano E, Rabizadeh S, Wu X, Yen HR, et al. A human colonic commensal promotes colon tumorigenesis via activation of T helper type 17 T cell responses. Nat Med. 2009;15(9):1016–22.
4. Pleguezuelos-Manzano C, Puschhof J, Rosendahl Huber A, van Hoeck A, Wood HM, Nomburg J, et al. Mutational signature in colorectal cancer caused by genotoxic pks(+) E. coli. Nature. 2020;580(7802):269–73.
5. Janney A, Powrie F, Mann EH. Host-microbiota maladaptation in colorectal cancer. Nature. 2020;585(7826):509–17.
6. Sears CL, Garrett WS. Microbes, microbiota, and colon cancer. Cell Host Microbe. 2014;15(3):317–28.
7. Young C, Wood HM, Fuentes Balaguer A, Bottomley D, Gallop N, Wilkinson L, et al. Microbiome analysis of more than 2,000 NHS bowel cancer screening programme samples shows the potential to improve screening accuracy. Clin Cancer Res. 2021;27(8):2246–54.
8. Abdill RJ, Adamowicz EM, Blekhman R. Public human microbiome data are dominated by highly developed countries. PLoS Biol. 2022;20(2):e3001536.
9. Das B, Ghosh TS, Kedia S, Rampal R, Saxena S, Bag S, 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(1):10104.
10. Dhakan DB, 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. Gigascience. 2019;8(3):giz004.
11. Dubey AK, Uppadhyaya N, Nilawe P, Chauhan N, Kumar S, Gupta UA, et al. LogMPIE, pan-India profiling of the human gut microbiome using 16S rRNA sequencing. Sci Data. 2018;5:180232.
12. Young C, Wood HM, Seshadri RA, Van Nang P, Vaccaro C, Melendez LC, et al. The colorectal cancer-associated faecal microbiome of developing countries resembles that of developed countries. Genome Med. 2021;13(1):27.
13. Taylor M, Wood HM, Halloran SP, Quirke P. Examining the potential use and long-term stability of guaiac faecal occult blood test cards for microbial DNA 16S rRNA sequencing. J Clin Pathol. 2017;70(7):600–6.
14. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, et al. A communal catalogue reveals Earth's multiscale microbial diversity. Nature. 2017;551(7681):457–63.
15. Parada AE, Needham DM, Fuhrman JA. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. Environ Microbiol. 2016;18(5):1403–14.
16. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal. 2011;17(1):10–2.
17. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol. 2019;37(8):852–7.
18. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. DADA2: high-resolution sample inference from Illumina amplicon data. Nat Methods. 2016;13(7):581–3.
19. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013;41(Database issue):D590-6.
20. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.
21. Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, et al. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome. 2018;6(1):90.
22. Shannon CE. The mathematical theory of communication. 1963. MD Comput. 1997;14(4):306–17.
23. Bray JR, Curtis JT. An ordination of the upland forest communities of Southern Wisconsin. Ecol Monogr. 1957;27(4):325–49.
24. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, et al. vegan: Community Ecology Package. 2020.
25. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
26. Liaw A, Weiner M. Classification and regression by randomForest. R News. 2002;2(3):18–22.
27. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12:77.
28. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, et al. Metagenomic biomarker discovery and explanation. Genome Biol. 2011;12(6):R60.
29. Gupta A, Dhakan DB, Maji A, Saxena R, P KV, Mahajan S, et al. Association of Flavonifractor plautii, a flavonoid-degrading bacterium, with the gut microbiome of colorectal cancer patients in India. mSystems. 2019;4(6):e00438-19.
30. Boleij A, Hechenbleikner EM, Goodwin AC, Badani R, Stein EM, Lazarev MG, et al. The Bacteroides fragilis toxin gene is prevalent in the colon mucosa of colorectal cancer patients. Clin Infect Dis. 2015;60(2):208–15.
31. Bhattacharya M. A historical exploration of Indian diets and a possible link to insulin resistance syndrome. Appetite. 2015;95:421–54.
32. Mirzayi C, Renson A, Genomic Standards C, Massive A, Quality Control S, Zohra F, et al. Reporting guidelines for human microbiome research: the STORMS checklist. Nat Med. 2021;27(11):1885–92.

Publisher's Note