

ARTICLE OPEN



Tumour microbiomes and *Fusobacterium* genomics in Vietnamese colorectal cancer patients

Hoang N. H. Tran^{1,11}, Trang Nguyen Hoang Thu^{1,11}, Phu Huu Nguyen², Chi Nguyen Vo^{2,3}, Khanh Van Doan⁴, Chau Nguyen Ngoc Minh¹, Ngoc Tuan Nguyen², Van Ngoc Duc Ta², Khuong An Vu², Thanh Danh Hua², To Nguyen Thi Nguyen¹, Tan Trinh Van¹, Trung Pham Duc¹, Ba Lap Duong², Phuc Minh Nguyen², Vinh Chuc Hoang², Duy Thanh Pham^{1,5}, Guy E. Thwaites^{1,5}, Lindsay J. Hall^{6,7,8}, Daniel J. Slade⁹, Stephen Baker¹⁰, Vinh Hung Tran² and Hao Chung The¹✉

Perturbations in the gut microbiome have been associated with colorectal cancer (CRC), with the colonic overabundance of *Fusobacterium nucleatum* shown as the most consistent marker. Despite its significance in the promotion of CRC, genomic studies of *Fusobacterium* is limited. We enrolled 43 Vietnamese CRC patients and 25 participants with non-cancerous colorectal polyps to study the colonic microbiomes and genomic diversity of *Fusobacterium* in this population, using a combination of 16S rRNA gene profiling, anaerobic microbiology, and whole genome analysis. Oral bacteria, including *F. nucleatum* and *Leptotrichia*, were significantly more abundant in the tumour microbiomes. We obtained 53 *Fusobacterium* genomes, representing 26 strains, from the saliva, tumour and non-tumour tissues of six CRC patients. Isolates from the gut belonged to diverse *F. nucleatum* subspecies (*nucleatum*, *animalis*, *vincentii*, *polymorphum*) and a potential new subspecies of *Fusobacterium periodonticum*. The *Fusobacterium* population within each individual was distinct and in some cases diverse, with minimal intra-clonal variation. Phylogenetic analyses showed that within four individuals, tumour-associated *Fusobacterium* were clonal to those isolated from non-tumour tissues. Genes encoding major virulence factors (Fap2 and RadD) showed evidence of horizontal gene transfer. Our work provides a framework to understand the genomic diversity of *Fusobacterium* within the CRC patients, which can be exploited for the development of CRC diagnostic and therapeutic options targeting this oncobacterium.

npj Biofilms and Microbiomes (2022)8:87; <https://doi.org/10.1038/s41522-022-00351-7>

INTRODUCTION

Colorectal cancer (CRC) is the second leading cause of cancer mortality worldwide, contributing to an estimate of 850,000 deaths and ~1.8 million new cases in 2018^{1,2}. The majority of CRC cases are sporadic, with well-established lifestyle risk factors attributed to obesity, alcohol consumption and a diet enriched with red or processed meat³. The gut microbiome is an integral part of human health, and act as an important interface mediating the interactions between environmental cues, host biology, and CRC^{4,5}. Research on CRC gut microbiome has consistently underlined the abundances of certain marker bacteria, among which *Fusobacterium nucleatum* has been most widely reported and intensively studied^{6–10}.

The Gram-negative rod-shaped *F. nucleatum* is a common anaerobic member of the human oral microbiome, and it is currently composed of four subspecies (*nucleatum*, *vincentii*, *animalis*, and *polymorphum*)¹¹. Mechanistic studies have demonstrated that *F. nucleatum* possesses several virulence factors, most notably FadA and Fap2, which enable the bacteria to potentiate colonic tumorigenesis. The adhesin FadA binds to E-cadherin in CRC cells and activates the β -catenin-dependent oncogenic pathways¹², while the lectin Fap2 further facilitates *F. nucleatum* invasion into CRC cells by specifically binding to the tumour-enriched carbohydrate Gal-GalNAc¹³. Such interaction triggers the secretions of the pro-inflammatory (IL-8) and

pro-metastatic (CXCL-1) cytokines, creating a tumour environment conditioned for accelerated growth and migratory tendency¹⁴. Recent studies have further highlighted that the bacteria could induce DNA damage in oral and colorectal cancerous cells^{15,16}. As a result, enrichment of *F. nucleatum* in CRC microbiomes has been associated with more severe prognosis and poorer overall survival, particularly in a subset of patients with mesenchymal tumours^{17–19}. Preclinical research demonstrated that *F. nucleatum* elimination by antibiotics reduced colorectal tumour proliferation in mice²⁰. These evidences strongly support for the utilization of *F. nucleatum* as a target for CRC diagnosis and therapy, but current translational potential is hampered by the lack of insights into *F. nucleatum* diversity and its genomic characteristics in CRC patients.

The majority of microbiome studies were conducted in high-income countries, and such data are sparse regarding populations in developing settings, where host factors, diet and lifestyle could greatly influence the gut microbiome composition and function. Vietnam has an increasing ageing population adopting a more 'Westernized' diet and sedentary lifestyle²¹, where CRC incidence is predicted to climb and rank as among the top three cancers by 2025²². Therefore, CRC microbiome studies in Vietnam are necessary to establish the basis for the implementation of microbiome-oriented strategies for CRC prevention, diagnosis, prognosis and therapy. We set out to investigate the microbiome signatures of Vietnamese CRC patients, by applying

¹Oxford University Clinical Research Unit, Ho Chi Minh City, Vietnam. ²Binh Dan Hospital, Ho Chi Minh City, Vietnam. ³Tan Tao University, Long An, Vietnam. ⁴Department of Oral Biology, Yonsei University College of Dentistry, Seoul, Korea. ⁵Centre for Tropical Medicine and Global Health, Nuffield Department of Clinical Medicine, University of Oxford, Oxford, United Kingdom. ⁶Quadram Institute Biosciences, Norwich Research Park, Norwich, United Kingdom. ⁷Norwich Medical School, University of East Anglia, Norwich Research Park, Norwich, United Kingdom. ⁸Intestinal Microbiome, School of Life Sciences, ZIEL - Institute for Food & Health, Technical University of Munich, Freising, Germany. ⁹Department of Biochemistry, Virginia Tech, Blacksburg, VA 24061, USA. ¹⁰Department of Medicine, Cambridge Institute of Therapeutic Immunology and Infectious Diseases (CITIID), University of Cambridge, Cambridge, United Kingdom. ¹¹These authors contributed equally: Hoang N. H. Tran, Trang Nguyen Hoang Thu. ✉email: haoc@oucru.org

16S-rRNA gene profiling on the saliva and gut tissues collected from patients with CRC and non-cancerous colorectal polyps. Additionally, different from prior studies, we used anaerobic culturing and whole genome sequencing (WGS) to study the genomic diversity of *Fusobacterium* colonising these CRC patients, allowing an in-depth and high-resolution examination of these bacterial populations.

RESULTS

Gut mucosal, but not salivary, microbiomes differ significantly between CRC and controls

We enrolled 43 CRC patients (cases) and 25 patients with colorectal polyps (controls) between December 2018 and January 2020. 16S rRNA microbiome profiling was performed for all the saliva and gut tissue samples collected from the participants, including those originating from the diseased (CRC tumour or polyps) and the adjacent normal sites. Since the majority of bacterial biomass in gut tissues originates from the mucosa, the terms mucosal and tissue-associated microbiomes were used interchangeably. To limit the scope of this study, we selected participants with tumours/polyps detected in the distal colon or rectum. The patients' demographic and clinical data were summarized in Table 1, which showed that there were no significant differences between the two groups. All polyps showed not more than low-grade dysplasia (i.e. non-cancerous), demonstrating the validity of our control group. No CRC patients have received chemo- or radiotherapy before surgery. Microbiome profiling identified 865 filtered amplicon sequence variants (ASVs – a marker for distinct taxonomic classification) among 66 saliva samples, with a median library size of 36,250 paired-end reads [IQR: 31,827–50,317]. Due to their lower microbial biomass, the

library size of gut mucosal microbiomes was smaller (median: 17,711 [IQR: 9037–30,135]), with 1073 filtered ASVs detected across 129 tissue samples (seven removed). Initial quality check showed that the salivary and gut mucosal microbiomes were well separated on Bray-Curtis ordination (Supplementary Fig. 1A), and the sequenced mock community's composition matched the manufacturer's description (Supplementary Fig. 1B). Assessment of the rarefaction curves showed that both sequenced salivary and gut tissue samples attained sufficient sampling depth to recover the respective microbiome diversity (Supplementary Fig. 2).

Ordination by principal coordinate analysis (PCoA), based on phylogenetic-assisted isometric log-ratio (PhIRL) transformed value, showed that the salivary microbiomes of CRC and controls completely overlapped (Fig. 1a). Only active smoking within the last two years, but not CRC status, was significantly associated with the salivary microbiome structure (RDA, p -value = 0.033). Likewise, only two ASVs belonging to the genera *Leptotrichia* and *Solobacterium* were consistently identified as significantly more abundant in the CRC's salivary microbiome (log₂ fold change of 2.25 and 1.82 respectively, adjusted p -values < 0.05) (Supplementary Fig. 3A; Supplementary Table 1). These point to the high structural similarity in the salivary microbiome between the two groups. By contrast, the gut mucosal microbiomes differ significantly based on CRC status (Fig. 1b). CRC and diabetes significantly contributed to the variance in the gut microbiome (RDA, p -value < 0.05). Gut mucosa collected within a participant (tumour and non-tumour for CRC, biopsy and polyp for control) shared more similarity in their microbiomes than those of the same sample type between participants (Fig. 1c), resembling findings from previous research⁸. We also conducted these analyses using the weighted Unifrac and Bray-Curtis distances, which produced similar interpretations (Supplementary Fig. 4). Additionally, we performed unsupervised clustering on gut mucosal microbiomes, which showed the presence of two robust community state types (CSTs) supported by a mean accuracy of 90.67% (assessed by 50 iterations of nested cross-validation) in a random forest classification. This algorithm also identified that several 'balances' contributed significantly in separating the two CSTs (Supplementary Fig. 5). CST1 was generally more enriched in Gammaproteobacteria (mostly *Escherichia*) while CST2 had higher abundance of Actinobacteria (mainly *Collinsella*) and Lachnospiraceae (Fig. 1d). The two CSTs were similar in library size (p -value = 0.15, t -test), but different in CRC status (p -value = 0.002, Fisher-exact test), with the majority of control samples (72%) belonging to CST1. Samples from the same patients mostly shared the same CST membership (90.3%, n = 56/62 patients with paired microbiomes). These findings suggest that while CRC status mainly explained the dissimilarity observed in the gut mucosal microbiomes, their overall configurations were determined by the dominant presence of Gammaproteobacteria (*Escherichia*), possibly driven by an unknown or stochastic factor.

Enrichment of oral bacteria in the tumour gut tissues

We applied differential abundance analysis to rigorously detect bacteria enriched in the CRC tumours, by comparing results from different approaches, including ANCOMBC, DESeq2 and corncob (see Methods)^{23–25}. Our analyses revealed that ASVs classified as bacteria of putative oral origin (*Gemella*, *Peptostreptococcus*, *F. nucleatum*, *Leptotrichia*, *Selenomonas sputigena*, and *Campylobacter rectus*) were overabundant in the tumour microbiomes, compared to control biopsies (Figs. 1d and 2b) (log₂ foldchange of 0.84 to 4.17, adjusted p -values < 0.05; Supplementary Table 1). Within the CRC patients, tumours also showed an elevated presence of the aforementioned oral bacteria, alongside *Hungatella*, *Lachnoclostridium*, and *Osillibacter*, when compared to adjacent non-tumour tissues, albeit with less pronounced fold change (log₂ foldchange of 0.87 to 1.95,

Table 1. Baseline characteristics of patients recruited in this study.

	CRC cases ($n = 42$)	Controls ($n = 21$)	p -value
Age	64 [54–69]	60 [53–66]	0.359
Male sex	62%	76%	0.395
BMI	22.9 [20.85–24.95]	22.2 [21.1–23.4]	0.387
Overweight/obesity ^a	47.60%	33%	0.409
Diabetes	19%	19%	1
High blood pressure	52%	47.60%	0.79
Active smoking in the last two years	21.40%	19%	1
Oral diseases ^b	33%	38%	0.782
Family history of cancer	19%	19%	1
Location of sampled mucosa			0.533
Descending colon	7	3	
Sigmoid colon	28	12	
Rectum	7	6	
Size of tumour/polyp (cm)	5 [4–5.75]	1 [0.7–1.2]	
TNM stage of cancer	II (18), III (20), IV (4)		
Polyp dysplasia grade		Low (4), none (17)	

^aOverweight/obesity was classified using WHO recommendation for Asian populations.

^bOral diseases include self-reported gingivitis, periodontitis or halitosis. The number in each cell refers to median with interquartile range in brackets, percentage or count number for each category.

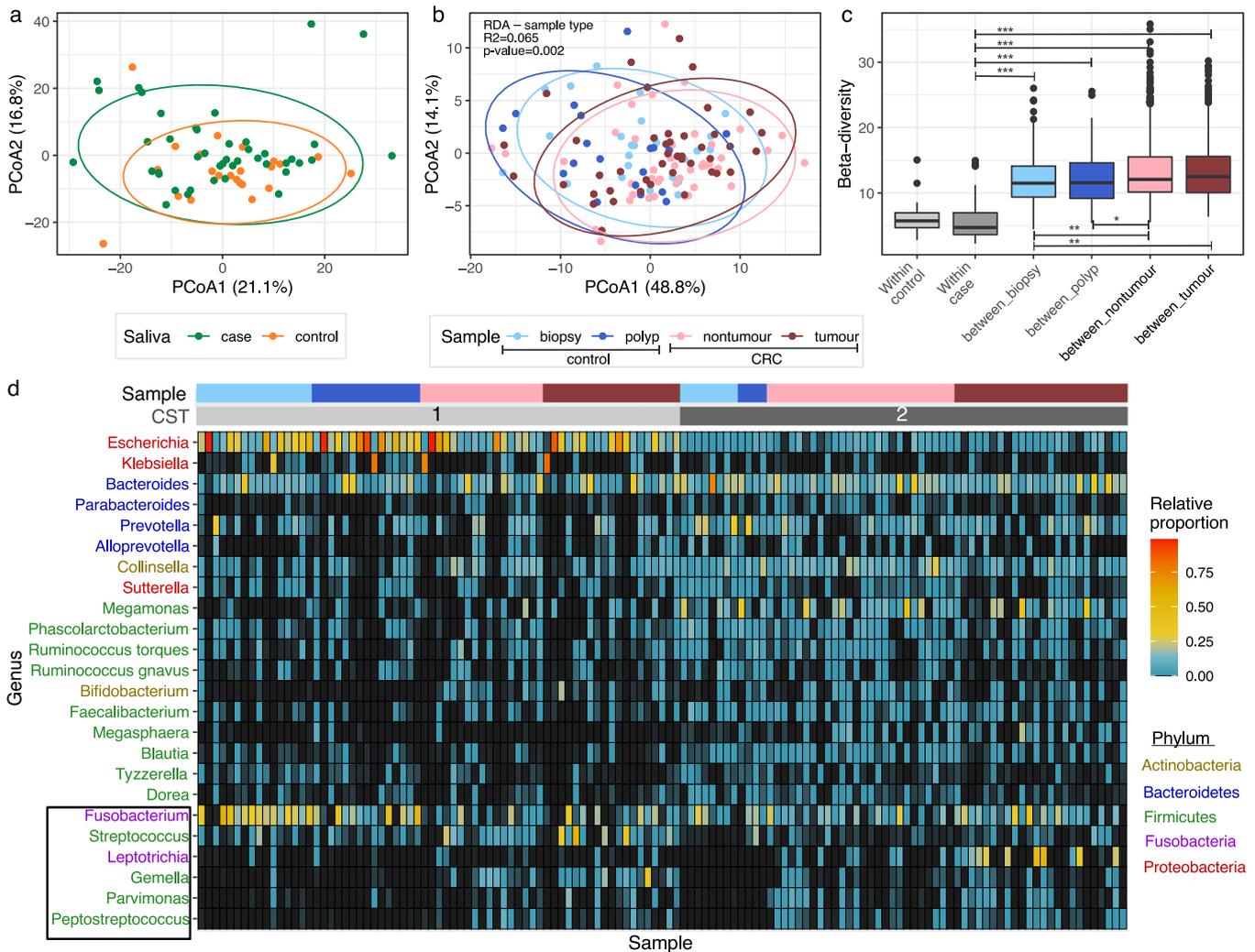


Fig. 1 The salivary and gut mucosal microbiomes of colorectal cancer patients. Principal coordinate analyses (PCoA), conducted on phylogenetic-assisted isometric log-ratio (PhILR) transformed data, of **(a)** 66 salivary microbiomes, and **(b)** 129 gut mucosal microbiomes, with different CRC groups and sample types denoted by different colours (see Keys; biopsies and polyps collected from controls, nontumours and tumours collected from cases). **c** Boxplot showing the distribution of pairwise beta-diversity, calculated on PhILR transformed values, observed in each gut microbiome category. Bold central lines denote the median, the upper whisker extends from the 75th percentile to the highest value within the 1.5*interquartile range (IQR) of the hinge, the lower whisker extends from the 25th percentile to the lowest value within 1.5*IQR of the hinge. Data points beyond the end of the whiskers are outliers. Asterisk markings represent statistically significant differences between groups, as calculated by posthoc Tukey test (p -values ranging from >0.01 to ≤ 0.05 (*); from $>1e-5$ to ≤ 0.01 (**); $\leq 1e-5$ (***)). **d** Heatmap displaying the proportional abundances of 24 most abundant genera (prevalence $\geq 15\%$, mean relative abundance $\geq 1\%$, and accounting for $\sim 85\%$ of the gut mucosal microbiome composition), with headers showing the samples' community state type (CST): CST1 (light gray), CST2 (dark gray), and the corresponding sample type: biopsy (light blue), polyp (dark blue), nontumour (pink), tumour (dark red). Genera were coloured according to their classifications at Phylum level (see Keys). Genera in black box represent ones with probable origin from the oral cavity. The contributions of 24 genera listed here were summarized in Supplementary Table 3. Source data are provided as a Source Data file.

adjusted p -values < 0.05 , Fig. 2a). These increases were coupled with the reduction in abundances of commensal anaerobes in the tumour tissues, such as *Blautia*, *Parabacteroides*, *Dorea*, and *Collinsella* (log₂ foldchange of -1.06 to -1.34 , adjusted p -values < 0.05). Collectively, CRC-associative taxa ($n = 11$, log₂ foldchange > 0 , Fig. 2a) accounted for a mean cumulative relative abundance of 12% across 43 tumour microbiomes, with prevalence exceeding 90% ($n = 39/43$). When comparing between different cancer stages, the increased abundance of one taxon (*Leptotrichia* ASV-13, log₂ fold-change 2.63, adjusted p -values < 0.05) was consistently associated with tumours of advanced stages (III-IV), compared to stage II (Fig. 2c). Results from DESeq2 alone additionally showed that *F. nucleatum* was also enriched in advanced CRC stages (adjusted p -value < 0.05). ASVs confidently

assigned as *F. nucleatum* ($n = 14$) and *Leptotrichia* spp. ($n = 16$) were present at mean relative abundance of 4.6% (prevalence = 26/43) and 6.3% (prevalence = 22/43) in tumour microbiomes, respectively (Supplementary Table 2). We performed similar analysis within the control group and showed that only one ASV (*Faecalibacterium*) was consistently depleted in polyps compared to paired biopsies (log₂ fold-change = -0.99 , adjusted p -value < 0.05). However, when compared to CRC samples, *Fusobacterium mortiferum*, *Tyzzerella*, and *Sutterella* were significantly enriched in the control gut microbiomes (Fig. 2b, log₂ foldchange of -2.6 to -5.27 , adjusted p -value < 0.05 ; Supplementary Table 1).

To investigate bacterial co-occurrence and their potential interactions, we next constructed a correlation network of gut microbiomes from CRC patients ($n = 86$) (Fig. 3)^{26,27}. Two oral

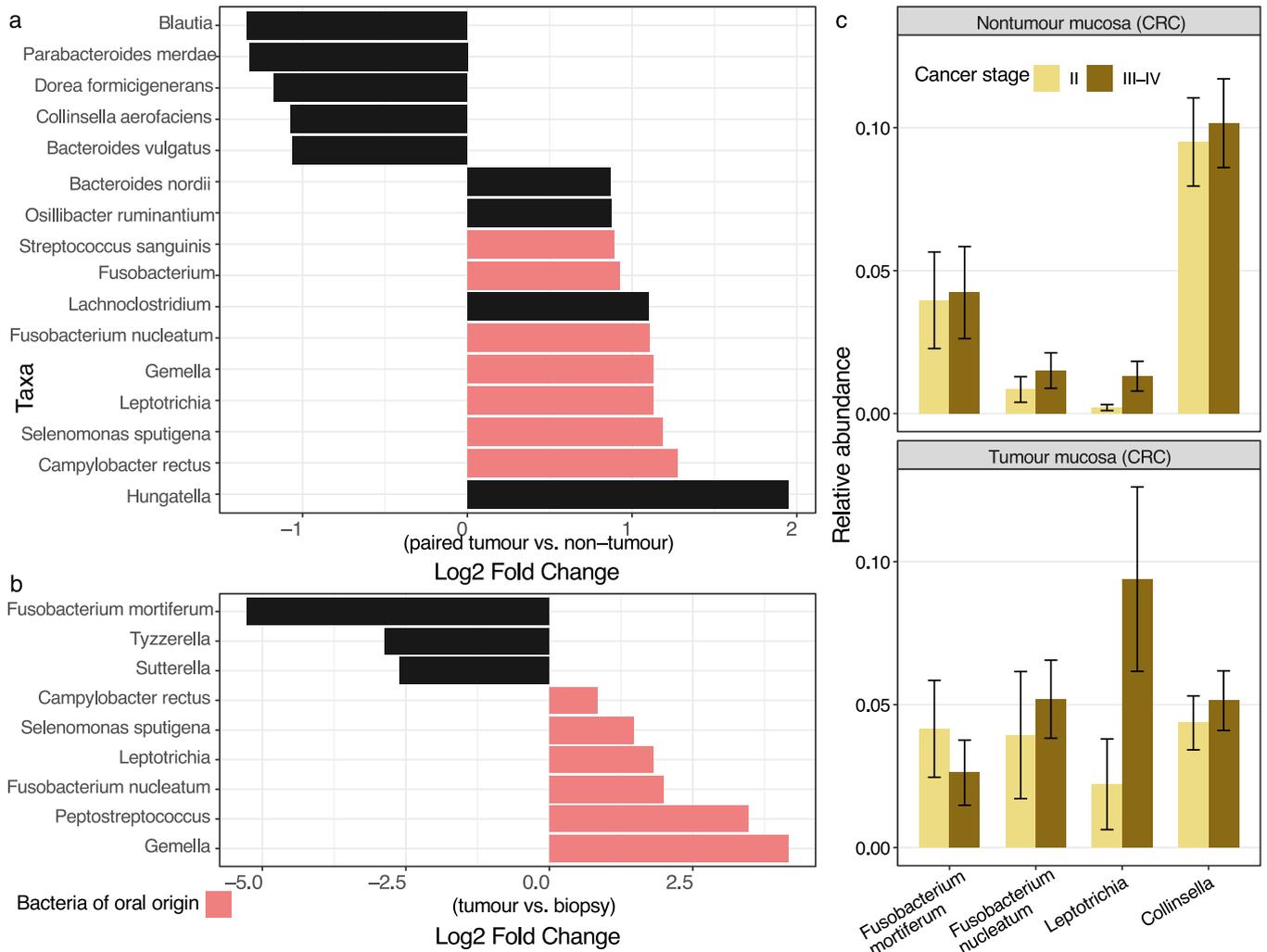


Fig. 2 Bacterial taxa significantly abundant among the examined classes. Taxa, or amplicon sequence variants (ASVs), were determined as significant and visualized in **a** and **b** if they were detected in at least two of the three tested approaches (ANCOMBC, DESeq2, corncob; adjusted p -value ≤ 0.05). **a** Log₂ fold change of ASVs that differ between paired tumour and non-tumour mucosal microbiomes from case participants, using the full model 'Patient + sample type' ($n = 86$). **b** Log₂ fold change of ASVs that differ between tumour and biopsy (control) mucosal microbiomes ($n = 67$). Log₂ fold change was derived from ANCOMBC test output, and taxa of oral origin were coloured in pink. **c** Relative abundance of ASVs assigned as *Fusobacterium mortiferum* ($n = 14$), *Fusobacterium nucleatum* ($n = 14$), *Leptotrichia* ($n = 16$), and *Collinsella* ($n = 14$) in the tumour and nontumour mucosal microbiomes, stratified by cancer stages (III-IV vs. II). The error bar represents standard error of the mean relative abundance in each bar plot. Source data are provided as a Source Data file.

bacteria clusters emerged from this network, one consisting of several *Streptococcus* and *Veillonella* taxa, and another composed mostly of aforementioned tumour-associated ASVs (*Leptotrichia*, *Selenomonas*, *F. nucleatum*, *Streptococcus*, *Granulicatella*, *Gemella*, *Peptostreptococcus*, and *Parvimonas*). The latter cluster exhibited positive correlation with *E. coli* ($cor = 0.488$, p -value = $9.2e-07$), and antagonism toward the gut commensal *Blautia* ($cor = -0.466$, p -value = $2.97e-06$). Besides, other tumour-associated ASVs such as *Hungatella*, *Lachnospirillum*, and *C. rectus* were clustered alongside *Negativibacillus* and *Eggerthella*, which showed strong negative correlations with anaerobic gut commensals *Dorea*, *Bacteroides*, and *Faecalibacterium*. These findings highlight the potential competition between tumour-associated taxa and common gut commensal anaerobes. Other *Fusobacterium* species, *F. mortiferum* and *F. varium* were not linked to the oral clusters, showing that they were mainly gut inhabitants. Comparison with the network constructed from salivary microbiomes revealed that the same tumour-associated ASVs (*F. nucleatum*, *Gemella*, *Selenomonas*)

formed similar clusters as observed in the CRC gut microbiomes (Supplementary Fig. 6).

Diverse *Fusobacterium* colonizes CRC patients

Since *F. nucleatum* was more enriched in the tumour microbiomes and previously demonstrated to promote tumourigenesis, we next studied the population structure of *Fusobacterium* recovered from CRC patients. Six patients with a *Fusobacterium* relative abundance at the tumour site exceeding 10% (except for patient 18) and covered different cancer stages were selected for *Fusobacterium* isolation. In total, we isolated 56 presumptive *Fusobacterium* organisms, as identified by the matrix-assisted laser desorption/ionization time of flight mass spectrometer (MALDI-TOF), from the oral, nontumour and tumour samples of these patients (Table 2). Fifty-three successfully sequenced genomes belong to *F. nucleatum* ($n = 38$) and *F. periodonticum* ($n = 15$) species complexes, of which phylogenetic reconstruction was performed separately. Across the two phylogenies, we identified 14 phylogenetic clusters (PCs; each with 2–6 isolates exhibiting negligible genetic

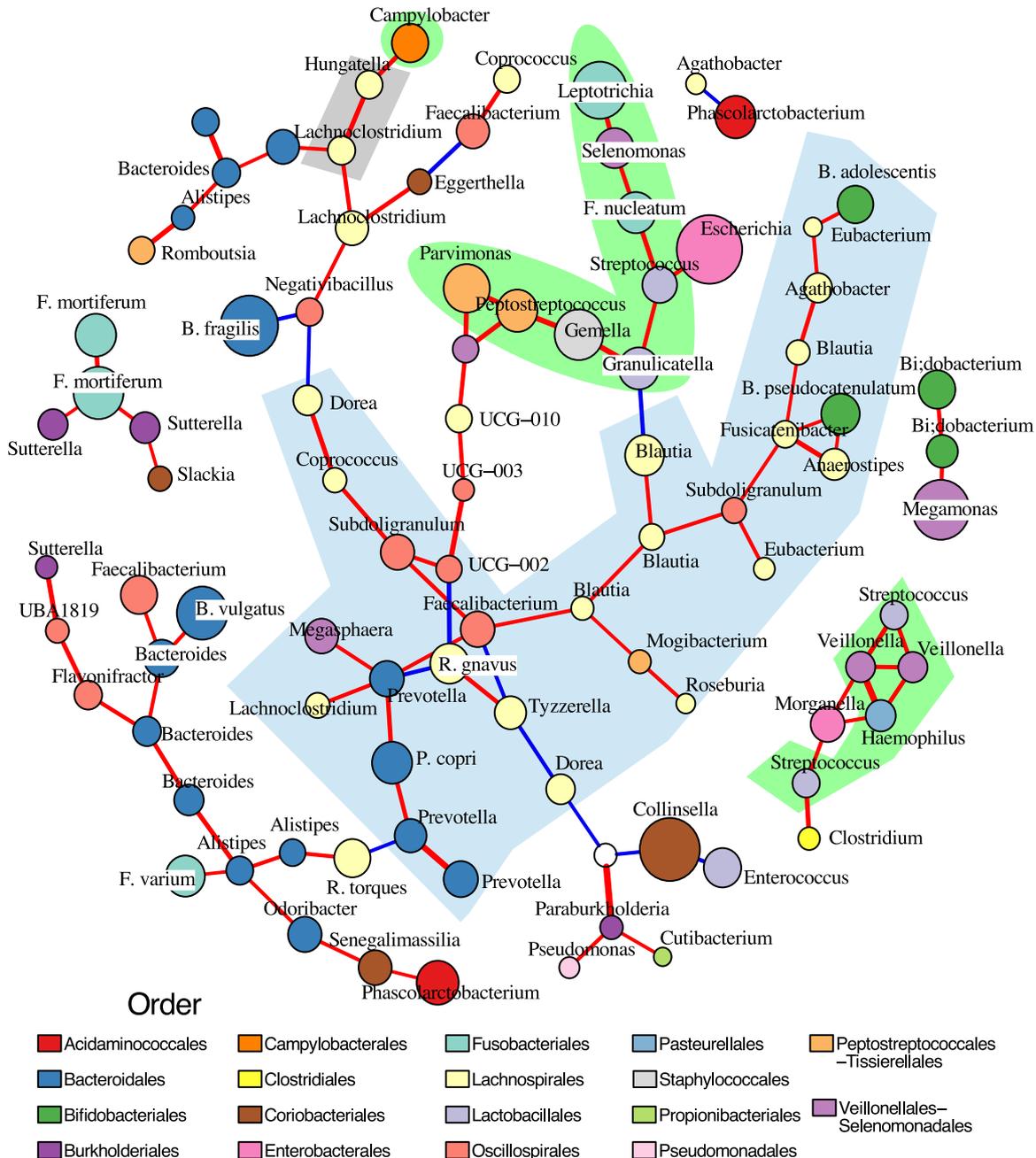


Fig. 3 Correlation network of colorectal cancer gut mucosal microbiomes. The network was constructed from 117 most representative ASVs sampled from 86 mucosal microbiomes, outlining significant interactions detected by both CCLasso (p value ≤ 0.01 and absolute correlation strength > 0.37) and SpiecEasi. Positive and negative interactions were coloured as red and blue lines respectively, with line weight proportional to correlation strength. The ASVs (nodes) were coloured based on taxonomic family (see Legend), with sizes proportional to their relative abundances. The light green shaded area entails ASVs identified as members of the human oral microbiome (comparison with expanded Human Oral Microbiome Database); the blue shaded area covers ASVs identified as gut anaerobic commensals (Lachnospirales, Bacteroidales, Bifidobacteriales, Oscillospirales); and gray shaded area covers other tumour-associated ASVs (as identified in Fig. 2a). Source data are provided as a Source Data file.

differences) and 12 singletons originating from this study's collection, which were collectively named as PCs herein (17 *F. nucleatum*, 7 *F. periodonticum*, and 2 *F. hwasookii*). Each PC represents a single *Fusobacterium* strain, with multiple colonies picked from the same patient (Table 2). Core-genome phylogeny of *F. nucleatum* showed that tumour-associated isolates were detected in all four subspecies (*animalis*, *vincentii*, *nucleatum*, *polymorphum*) (Fig. 4a). In the *F. periodonticum* phylogeny, tumour-associated isolates (2 PCs isolated from P18, P40) formed

a distinct cluster that is phylogenetically separated from the available references (Fig. 4b). These isolates all showed ~91% average nucleotide identity (ANI) to the closest *F. periodonticum* references, suggesting that they constitute a novel subspecies of this species complex, denoted herein as novel *F. periodonticum* (novelFperi). Likewise, one gut PC (H16_Fa) shared 93% ANI to the closest *F. nucleatum* references and were phylogenetically distant from the remaining *F. nucleatum* isolates, potentially indicative of a novel *F. nucleatum* subspecies.

Table 2. Summary of *Fusobacterium* isolates recovered from this study.

Phylogenetic cluster	Species	Subspecies	Isolation source	Number of isolates	intracranial SNV	Patient ID	Tumour location	Cancer stage	<i>Fusobacterium</i> relative abundance (%) ^b
T10_Fa1	<i>F. nucleatum</i>	<i>animalis</i>	Tumour	2	5	P10	Sigmoid	IIB	2.6/24
TH10_Fa2	<i>F. nucleatum</i>	<i>animalis</i>	Tumour-nontumour	5	2				
S16-12	<i>F. nucleatum</i>	<i>polymorphum</i>	Oral	1		P16	Sigmoid	IIB	10/34
S16-17	<i>F. periodonticum</i>	-	Oral	1					
T16_Fp	<i>F. nucleatum</i>	<i>polymorphum</i>	Tumour	2	3				
H16_Fa	<i>F. nucleatum</i>	<i>animalis</i>	Nontumour	2	0				
S16_Fa	<i>F. nucleatum</i>	<i>animalis</i>	Oral	2	10				
TH16_Fvi	<i>F. nucleatum</i>	<i>vincentii</i>	Tumour-nontumour	6	1				
H18-18	<i>F. nucleatum</i>	<i>animalis</i>	Nontumour	1		P18	Descending	IIA	0.3/0.8
S18-79	<i>F. nucleatum</i>	<i>polymorphum</i>	Oral	1					
S18-78	<i>F. hwasookii</i>	-	Oral	1					
S18-66	<i>F. periodonticum</i>	-	Oral	1					
S18_Fperi1	<i>F. periodonticum</i>	-	Oral	2	1				
S18_Fperi2	<i>F. periodonticum</i>	-	Oral	2	1				
S18_Fh	<i>F. hwasookii</i>	-	Oral	2	4				
TH18_novelFperi	Novel <i>F. periodonticum</i> ¹	-	Tumour-nontumour	6	2				
S28-2	<i>F. periodonticum</i>	-	Oral	1		P28	Sigmoid	IV	14/39
T28_Fn	<i>F. nucleatum</i>	<i>nucleatum</i>	Tumour	3	1				
S40-28	<i>F. nucleatum</i>	<i>animalis</i>	Oral	1		P40	Sigmoid	IIIB	27/10
TH40_novelFperi	Novel <i>F. periodonticum</i> ^a	-	Tumour-nontumour	2	2				
S40_Fp	<i>F. nucleatum</i>	<i>polymorphum</i>	Oral	2	2				
S46-13	<i>F. nucleatum</i>	<i>vincentii</i>	Oral	1					
S46-17	<i>F. nucleatum</i>	<i>vincentii</i>	Oral	1					
S46-7	<i>F. nucleatum</i>	<i>polymorphum</i>	Oral	1					
S46-16	<i>F. nucleatum</i>	<i>polymorphum</i>	Oral	1					
T46_Fa	<i>F. nucleatum</i>	<i>animalis</i>	Tumour	3	3	P46	Sigmoid	IIIB	3.8/20

SNV single nucleotide variant.

^aDenotes a potential new subspecies.^b*Fusobacterium* relative abundance, inferred from 16S rRNA gene profiling results, showed the values for non-tumour/tumour samples for each cancer patient.

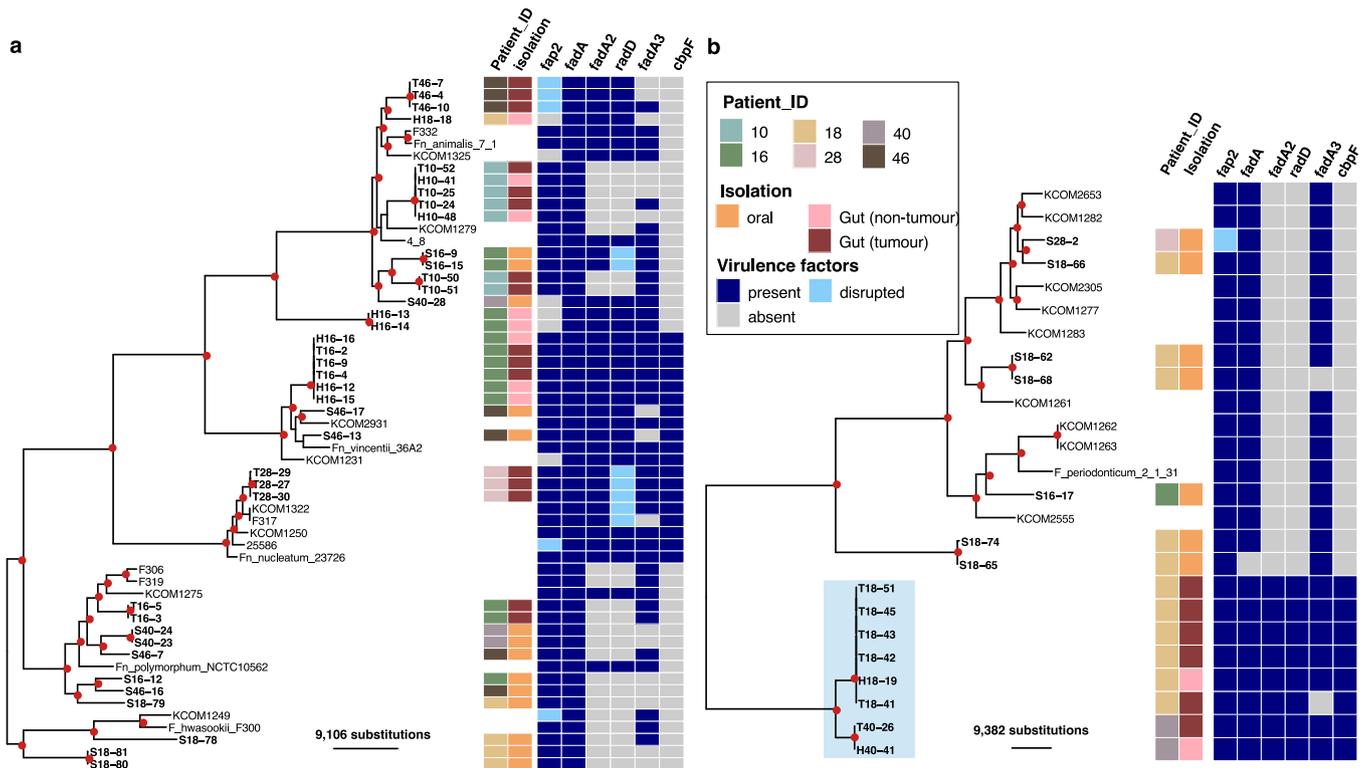


Fig. 4 Maximum likelihood phylogenies of *Fusobacterium* isolates from this study. **a** *F. nucleatum* species phylogeny constructed from the alignment of 516 core genes (89,900 variant sites; $N = 57$), using *F. hwasookii* clade as an outgroup. **b** *F. periodonticum* species phylogeny constructed from alignment of 863 core genes (106,738 variant sites, $N = 26$). Red circles at internal nodes denote bootstrap values ≥ 80 . The associated metadata on the right describe the patient ID and clinical origin of isolates (where reference genomes were left blank), and the genomic presence of several virulence factors (*fap2*, *fadA*, *fadA2*, *radD*, *fadA3*, *cbpF*). Light blue shaded area covers isolates identified as novel *F. periodonticum* subspecies. The scale bars denote the estimated number of substitutions.

The *Fusobacterium* population within each individual patient was diverse (2–7 PCs). Several *Fusobacterium* species/subspecies were detected in each patient's saliva, sometimes with more than one PCs of the same subspecies (P18, P46) (Table 2). Likewise, we observed similar diversity in gut-associated isolates, with more than one PCs detected in three patients (P10, P16, P18). Most patients did not share the same *Fusobacterium* subspecies recovered from both oral- and gut-associated isolates, except for P16 (*polymorphum*). However, phylogenetic evidence confirmed that the two niches harboured distinct populations, which were ~16,955 SNPs apart (Fig. 4a). Particularly, oral *Fusobacterium* isolates from P18 ($n = 9$) belonged to six different PCs (mostly *F. periodonticum* and *F. hwasookii*), while 6/7 gut isolates were of a single novelFperi clone. By contrast, *Fusobacterium* from tumour and nontumour sites were frequently clustered in the same PC ($n = 4$; in P10, P16, P18 and P40), indicating that the same bacterial clones have colonised and spread beyond the tumour microenvironment. We used the mapping approach to confidently inspect the intraclonal variations within these PCs, and showed that they shared minimal genetic differences in the core genome (1–2 SNVs). These values fall in range with the variation observed in five other gut PCs (with either tumour or nontumour isolates; 0–5 SNVs) and five other oral PCs (1–10 SNVs).

Variation in *Fusobacterium* virulence gene content

We next sought to examine the presence of several *Fusobacterium* virulence factors, of which pathogenicity has been proven in experimental studies, including genes encoding adhesin (*fadA*, *cbpF*), lectin (*fap2*), and bacterial co-aggregation factor (*radD*)^{12,13,28,29}. RadD is an autotransporter facilitating *Fusobacterium*'s interspecies interaction in polymicrobial biofilms²⁹, while

CbpF inhibits CD4⁺ T-cell response through CEACAM1 binding and activation³⁰. Genomic screening showed that *fap2* was present and intact in the majority of genomes from both species (49/53), with disruptive mutations occurring in some isolates, such as the tumour-associated *F. nucleatum animalis* in P46 (Fig. 4a). We also detected *fadA* in all isolates (except S18-65), with all *F. periodonticum* variants one amino acid shorter (codon A22) than the canonical FadA found in *F. nucleatum* (129 aa). The other elements showed variable presence among the examined genomes. For example, *cbpF* was present in all *F. nucleatum nucleatum*, *F. nucleatum vincentii*, and novelFperi, while *radD* was co-localised with *fadA2/radA* (a 122 aa *fadA* homolog) in 28 isolates. Another *fadA* homolog (*fadA3*) with unknown function was prevalent in both two *Fusobacterium* species. Phylogenies of FadA and CbpF showed that the two tree topologies were largely in agreement with those inferred from the core genomes, suggesting the absence of horizontal gene transfer (Supplementary Fig. 7). By contrast, the clustering pattern observed in the Fap2 phylogeny was concordant to subspecies classification for *F. nucleatum nucleatum*, *F. nucleatum vincentii*, and *F. periodonticum*, but was admixed for *F. nucleatum polymorphum*, *F. hwasookii* and *F. nucleatum animalis* (Supplementary Fig. 8A). *fap2* encodes a very large protein of variable length (median of 3938 aa [range: 3436–4669]), and the protein length showed some correlation with its phylogenetic clustering, with variants >4200 aa ($n = 6$) all belonging to a monophyly composed of *F. hwasookii* and *F. nucleatum polymorphum*. Similarly, the RadD phylogeny did not concur with those inferred from the core genomes, and its length variation (median 3526 aa [range: 3461–3602]) also showed association with the tree topology (Supplementary Fig. 8B). *radD* was ~800 bp downstream of *fadA2*, which is flanked by an IS150

transposase on the *F. nucleatum* 23726 reference genome. This could explain the mobilization mechanism of *radD-fadA2* across the *Fusobacterium* phylogeny. These data indicate that the autotransporter encoding genes *fap2* and *radD* may have undergone frequent horizontal gene transfer or recombination in the *F. nucleatum* species complex.

DISCUSSION

Our study revealed the composition of microbiome perturbations at the tumours of Vietnamese patients with CRC and non-cancerous colorectal polyps. Tumour-enriched taxa include mostly bacteria of putative oral origin, such as *F. nucleatum*, *Leptotrichia*, *Gemella*, *C. rectus*, and *Selenomonas*, which agrees with findings from previous studies profiling either gut mucosal or faecal microbiomes in different CRC populations^{8–10,31}. Compared to these studies, some CRC-indicative taxa (*Parvimonas*, *Solobacterium*, *Porphyromonas*) were not included in our findings because we applied a conservative approach in reporting differential abundance testing. Indeed, ASVs assigned to *Parvimonas* and *Porphyromonas* only showed significant enrichment in tumour microbiomes in either DESeq2 or ANCOMBC test. Thus, these slight differences likely stem from technical rather than biological reasons, highlighting that the proliferation of oral bacteria at the gut mucosa could be a universal signature of CRC microbiomes. We found that several of these oral taxa shared identical ASVs between the oral and gut niches, pointing to a probable oral origin of tumour-associated taxa. Our analysis found that these bacteria also display a co-occurrence pattern in the tumour microbiome, which agrees with the frequent presence of polymicrobial biofilms composed of oral taxa (*F. nucleatum*, *Peptostreptococcus*, *Gemella*) in colorectal tumours³². Among the oral bacteria, *F. nucleatum* stands out for its ability to form “bridging” interactions with other bacteria via the presence of several adhesins¹¹. *F. nucleatum* was recently reported to secrete FadA with amyloid properties, which confers acid tolerance and provides a scaffold for biofilm formation³³. In addition, our analyses pointed to the significant presence of *Leptotrichia* in tumour microbiomes, especially in advanced cancer. This association, however, has only been noted in few studies^{31,34}. This may be due to the differences in sampling location, as tumours excised from the distal colon (as performed for all cases in our study) were reported to harbour a higher abundance of *Leptotrichia*, compared to those originating from the proximal colon³⁴. Regarding the oral microbiome, we reported that only two taxa (*Leptotrichia* and *Solobacterium*) were significantly enriched in the saliva of CRC patients. Previous research has revealed that several bacteria (*Parvimonas*, *Haemophilus*, *Prevotella*, *Neisseria*) were significantly depleted in the CRC’s oral microbiome, compared to healthy controls³⁵. This discrepancy could stem from different oral sampling methods (saliva vs. cheek swab), study population (Asian vs. European), control populations (polyp vs. healthy), and analytical tools employed.

Asides from oral taxa, *Hungatella* overabundance was the most significant signature of CRC microbiome in our dataset. This falls in line with results from a recent metagenomic meta-analysis, showing that *Hungatella hathewayi*’s specific choline trimethylamine-lyase gene (*cutC*) was significantly enriched in the faecal microbiomes of CRC patients¹⁰. Moreover, colonic *H. hathewayi* could induce hypermethylation in prominent tumour suppressor genes, thus silencing their functions and promoting intestinal epithelial cell proliferation³⁶. On the other hand, we found that *F. mortiferum* was the most significantly enriched taxon in the polyp control group. *F. mortiferum* was known as a hallmark for dysbiosis in infectious diarrhoea³⁷, and recent studies have also reported the abundance of *F. mortiferum* in patients with colorectal polyps^{38,39}. Furthermore, this species was shown to be present in the gut microbiomes of ~60% of a cohort in

Southern China, albeit in very low abundance (~0.5%)⁴⁰. Unlike other *Fusobacterium* species, *F. mortiferum* was devoid of distinctive virulence factors such as adhesins FadA and Fap2⁴¹. The association between *F. mortiferum* and colorectal polyps will need to be further addressed in future studies.

Despite the increasing importance of *F. nucleatum* in the pathogenesis of CRC and other invasive diseases¹¹, genomic characterisation of these bacteria from patient populations is currently limited due to technical difficulties in *Fusobacterium* isolation. Here, we applied targeted culturomics approach, which combines anaerobic culturing, high-throughput identification by MALDI-TOF and WGS, to study the *Fusobacterium* population in high resolution and help uncover novel bacteria⁴². Indeed, we discovered novel subspecies of both *F. nucleatum* and *F. periodonticum* from culturing the gut tissues, showing that the microbiomes in non-Western settings offer untapped diversity. Using metagenomic assemblies from Chinese faecal microbiomes, Yeoh and colleagues have proposed several new *Fusobacterium* species (based on 95% ANI cutoff)⁴¹. Our WGS approach provided more accurate and complete realization of the bacterial genomes, which contributes to the global representation of *Fusobacterium* diversity (with 26 non-duplicate assemblies added). Furthermore, our approach allows for delineation of bacteria from tumour and non-tumour sites, which is inaccessible by faecal metagenomes. Nevertheless, targeted culturomics generally has low sensitivity, and bacterial recovery is subjected to factors such as storage time and condition. Therefore, our approach could not capture the high diversity of *Fusobacterium* in the oral niche⁴³, which likely explains the absence of close genetic relatedness between oral and gut *Fusobacterium* isolates. Previous research deploying WGS has demonstrated that oral and tumour-originated *F. nucleatum* shared little genetic divergence (0–183 SNVs), supporting the notion that oncogenic *Fusobacterium* arise from the patient’s oral microbiome⁴⁴. Similarly, using arbitrarily primed PCR, Komiya and colleagues showed that identical *F. nucleatum* strains were isolated from 6/14 paired gut tissue-saliva samples⁴⁵, but WGS was not conducted to verify the exact genetic differences. The populations of *Fusobacterium* colonising the oral cavity and gut were heterogeneous within some individuals, even at the subspecies level, which mirrors the diversity observed previously for gut commensals such as *Bifidobacterium*⁴⁶. Chronic infections with *Helicobacter pylori* at the stomach, which increases the risk of gastric cancer, usually result in extensive clonal propagations detected by WGS within each patient, though isolates were collected in a single timepoint⁴⁷. This prolonged colonization scenario contrasts with our observations in three CRC patients (P10, P16, P18), in which two to three *Fusobacterium* strains (with minimal intraclonal variation) were present at the tumour and non-tumour gut tissues. Given that CRC could take years to develop, we hypothesize that the *Fusobacterium* population at tumour sites might fluctuate in response to the frequent seedings from the highly diverse oral source. Additionally, identical *Fusobacterium* strains have been retrieved from the colonic tumours and liver metastasis of the same patient, suggesting the metastatic potential of tumour-borne *Fusobacterium*²⁰. Future longitudinal study design is necessary to investigate the *Fusobacterium* population dynamic within CRC patients, including those from different geographical regions.

The two well-described major virulence genes (*fadA* and *fap2*) were identified in the majority of *Fusobacterium* genomes, regardless of niche. This concurs with previous research reporting the high prevalence of *fadA* and *fap2* in *F. nucleatum* and *F. periodonticum* metagenomic assemblies from a cohort in China⁴¹. These suggest that *Fusobacterium* with high virulence potential are prevalent in the human population, and the genetic presence of *fadA* and *fap2* is not suitable for predicting the risk of

Fusobacterium-related CRC. All gut-derived novelFperi isolates harboured the examined virulence genes (*fadA*, *fap2*, *radD*, and *cbpF*), which was more similar to *F. nucleatum* compared to *F. periodonticum*. Moreover, *fap2* and *radD* showed variation in gene length and evidence of horizontal gene transfer, underlying the significance of dynamic evolutionary processes in shaping *Fusobacterium*'s virulence landscape. A recent study using *Fusobacterium* WGS has also reported that *fap2* could be either missing or highly divergent in tumour-derived *F. nucleatum*, suggesting the mobile nature of *fap2*⁴⁸. Since Fap2 orchestrates *F. nucleatum* invasion into CRC tumour cells via specific binding to Gal-GalNAc, this ligand-receptor interaction was recently proposed as a target for clinical intervention in *Fusobacterium*-enriched CRC⁴⁹. Interestingly, our genetic analysis predicted that *fap2* was either missing or truncated in some gut-associated *Fusobacterium* isolates, which may indicate the complex lifestyle of *Fusobacterium* once colonising the gut environment.

Some limitations were notable in our study design. Due to ethical concerns, patients with colorectal polyps were selected as the control group, instead of healthy age-matched individuals. Our interpretations do not extend to cancer in the proximal colon, though previous reports have noted that proximal CRC tumours had a higher *Fusobacterium* abundance⁵⁰. The sample size of cultured *Fusobacterium* isolates was moderate and did not include longitudinal sampling, so it was not possible to investigate the bacterial evolution in longer timeframe. Besides, our saliva sampling might not fully reflect the microbiome compositions at other oral sites, as well as to capture the whole diversity of *Fusobacterium*, which is more abundant in subgingival dental biofilms. Notwithstanding these shortcomings, our study reconfirmed the prominent role of oral anaerobic conglomerates in CRC microbiome in an understudied Asian population, and provided new insights into the genomic diversity of the oncobacterium *Fusobacterium*. The observed diversity in this organism should be taken into account when designing future diagnostic or therapeutic tools that target *Fusobacterium*.

METHODS

Study design and sample collection

This prospective case-control study enrolled adult Vietnamese patients (≥18 years old) admitted at Binh Dan Hospital, a large surgical hospital in Ho Chi Minh City Vietnam, from December 2018 to January 2020. This study received ethical approval from the Ethics Committee of Binh Dan Hospital (690/BVBD-QD), and the study was performed in compliance with the Declaration of Helsinki. Written informed consent was obtained from all study participants. Cases were defined as patients diagnosed with left-sided colorectal cancer (distal colon and rectum) of stage II onward, who received colectomy treatment and underwent non-antibiotic pre-operative bowel preparation. Controls were patients diagnosed with colorectal polyps (single/scattered non-cancerous polyps at distal colon or rectum), who received polypectomy at the hospital.

Demographic and clinical information were collected from study participants at recruitment. Cancer stage classification was based on the TNM Staging system⁵¹. A saliva sample (~3 mL) was collected within three hours pre-operation from each study participant (by spitting into a sterile container). For cases, the mucosa epithelia at the tumour and adjacent non-tumour (2–10 cm away from the tumour) sites were collected aseptically from the excised colon. For controls, we collected colorectal polyps and 2–3 biopsies of non-polyp mucosal epithelium (~50 mg) during colonoscopy. All clinical samples were stored on ice and transported back to the laboratory within 4 h, then were stored in -80 °C until further experiments.

16S rRNA gene sequencing

Microbiome profiling was performed on recruited 43 cases and 25 controls. Total DNA was extracted from whole biopsies and polyps (due to their small size), whole tumour (mucosa plus tumour tissue), and nontumour tissues ($n = 136$) using the FastDNA spin kit for soil (MP Biomedicals, USA), with bead-beating step on Precellys 24 homogenizer (Bertin Instruments, France). Though our approach targets the whole tissue and not just the mucosa, the bacterial biomass in the mucosa still comprise the majority of tissue-associated microbiome. Thus, the terms mucosal and tissue-associated microbiomes were used interchangeably. DNA from the saliva samples ($n = 67$, one missing) was extracted using the ReliaPrep Blood gDNA Miniprep (Promega, USA). For microbiome profiling, all samples underwent primary PCR amplification (30 cycles) using the conventional V4 primers (515F-806R) and KAPA HiFi Hot Start DNA polymerase (KAPA Biosystems, USA), and secondary PCR was performed to add dual-indexes (IDT, USA) to each sample, following procedures optimized in a published protocol^{52,53}. Additionally, we applied the same procedures to a positive control (Zymo mock community, Zymo Research, USA) and six negative controls (two for each DNA extraction kits, and two no-template PCR amplifications). 16S rRNA sequencing was performed for all samples on one run of the Illumina MiSeq platform, to generate 250 bp paired-end reads.

Microbiome data analysis

All data analyses were conducted in R (v4.1.1) and Rstudio using multiple packages, including 'dada2', 'phyloseq', 'DESeq2', 'ANCOMBC', 'corncob', 'philr', 'ggplot2', 'vegan', 'SpiecEasi' and others^{23–25,27,54–57}. Generated sequence reads were analysed under the amplicon sequence variant framework (ASV) using DADA2^{58,59}. Chimeric sequences were detected and removed independently for each sample. Taxonomic assignment (up to the species level) was performed using the RDP Naïve Bayesian Classifier implemented in 'dada2' package, on the SILVA v138 train dataset⁶⁰. Further filtering removed ASVs matching the following criteria (1) classified as 'Mitochondria' or 'Archaea', (2) unclassified at Kingdom or Phylum level, (3) identified as kitome or contamination from mock community (except *Escherichia* and *Enterococcus* ASVs), or (4) identified as low abundant singletons (abundance ≤ 10 counts and present in only one sample). This resulted in 2,461 ASVs detected across 203 samples (68 participants), totalling 5,250,754 sequences.

Saliva and gut mucosal microbiomes were then analysed separately. For saliva microbiomes, we removed singleton ASVs with abundances < 79 sequences (third quartile threshold) and one sample with low sequencing depth. The filtered ASVs ($n = 865$) were aligned using PASTA⁶¹, and a maximum likelihood phylogeny was constructed under the GTR + G model using IQ-Tree (with 1000 rapid bootstrap)⁶². The resulting phylogeny was used to transform the ASV count matrix into isometric log-ratio (ILR) 'balances' (weighted log-ratio between two ASVs), using the 'philr' package^{56,63}. Ordination was performed using principle coordinate analysis (PCoA) on a calculated Euclidean distance matrix. To identify covariates which explain the salivary microbiome structures, we performed redundancy analysis on the 'balance' value matrix of 62 samples with complete metadata. We repeat the same analytical procedures on the gut mucosal microbiome data. Low-abundance singleton ASVs (< 44 sequences – third quartile threshold) and seven samples with low sequencing depth (< 1300 sequences each, as assessed by rarefaction curve) were removed, retaining 1073 ASVs across 129 samples for downstream analyses. We tested the association between covariates and the gut mucosal microbiome structures using redundancy analysis, performed on the ILR-transformed 'balance' values of 120 samples with complete metadata. The ILR-transformed values were used to calculate the beta-diversity,

within and between participants. In addition, the gut mucosal microbiomes ($n = 129$) were clustered into community state types (CSTs) using the partition around medoid (pam) algorithm on the calculated ILR-transformed distance matrix, with the optimal number of CSTs ($k = 2$) determined by gap statistic and average silhouette width (asw)⁶⁴. The random forest classification algorithm (10,000 trees) was then used to identify 'balances' differentiating the two CSTs, using the package 'randomforestSRC'⁶⁵. We further assessed the performance of this model using 50 iterations of nested cross-validation (five-fold cross-validations for both the outer and inner loops), as implemented in Python's Sklearn library.

Evaluating differential abundances

In order to detect ASVs that showed significantly differential abundance between two examined groups, we utilized the compositional data analysis approach implemented in ANCOMBC²³. In addition, the same comparisons were performed using DESeq2 and corncob to check for consistent results, as recommended in recent benchmark studies^{66,67}. The comparisons include salivary microbiomes in cases ($n = 43$) and controls ($n = 23$); paired tumours ($n = 43$) against adjacent non-tumours ($n = 43$); paired polyps ($n = 16$) against non-polyp biopsies ($n = 16$); tumours ($n = 43$) against non-polyp biopsies ($n = 24$); tumours of cancer stage III-IV ($n = 24$) against stage II ($n = 18$). For paired comparison within cases and controls, the model design was set to "~Patient + sample_type" to increase statistical power⁶⁸. Multiple hypothesis testing was corrected using Holm or Benjamini-Hochberg method, setting false discovery rate as 0.05. ANCOMBC and corncob approaches were carried out using default parameters. For DESeq2, library size corrections were estimated using 'poscounts' method. All comparisons were performed using likelihood ratio test, and ASVs with adjusted p -value < 0.05 (and base mean > 20 for DESeq2) were considered significant hits. To minimize the number of false positives, ASVs which showed significant hits in at least two tested methods were considered differentially abundant and included in final interpretation. We performed BLAST for ASV sequences of interest against the expanded Human Oral Microbiome Database (HOMD; www.homd.org/), and species identification was assigned if the ASV showed >99% nucleotide similarity to that in the database.

Correlation network

We constructed a correlation network of gut mucosal microbiomes from colorectal cancer patients ($n = 86$), using 117 most representative ASVs, defined as ones with abundance of at least 10 sequences detected in at least 15 samples. This filtering resulted in a median sample retainment rate of 77% [70–85%]. The correlation network was constructed using CCLasso, with 250 bootstrap and three-fold cross validation²⁶. Interactions with adjusted p values < 0.01 and absolute correlation strength > 0.37 were considered significant hits. Additionally, a separate correlation network was inferred using SpiecEasi on the same dataset²⁷. Both these methods have been demonstrated to produce robust performance in a recent benchmark study⁶⁹. To avoid spurious hits, only significant interactions detected by both the CCLasso and SpiecEasi approaches were included in the final visualization. We applied the same procedures to construct correlation networks of microbiomes in saliva samples ($n = 66$, 115 ASVs) and controls' gut mucosa ($n = 43$, 90 ASVs).

Fusobacterium isolation and whole genome sequencing

Fusobacterium isolation was performed on six selected case patients (P10, P16, P18, P28, P40, P46), whose *Fusobacterium* relative abundance in the tumour microbiome exceeded 0.5% as

inferred by microbiome profiling. The respective samples (saliva, tumour, and non-tumour tissues) were subjected to anaerobic culturing in a Whitley A35 anaerobic workstation (Don Whitley Scientific, UK) supplied with 5% CO₂, 10% H₂, and 85% nitrogen gas, following an established *Fusobacterium* isolation procedures⁴⁵. Briefly, gut tissues were thawed on ice, and ~100 mg tissues were aseptically excised and anaerobically homogenized, using sterile surgical blades, in phosphate buffer supplemented with L-cysteine HCl (500 mg/L), Tween 80 (500 mg/L), and 0.1% resazurin. The suspension (100 µL) was then plated onto the selective media (EG agar supplemented with L-cysteine HCl, 50 ml/L of defibrinated sheep blood, 7 mg/L of crystal violet, 5 mg/L of vancomycin, 30 mg/L of neomycin, and 25 mg/L of nalidixic acid; Sigma-Aldrich, Germany). Thawed saliva samples were plated directly on the selective media. Plates were incubated at 37 °C for 48–72 h, and colonies (up to 10) resembling that of *Fusobacterium* were picked from each plate and sub-cultured on new EG media to confirm purity and select for single colonies. The isolate's taxonomic identities were queried using MALDI-TOF, and those characterised as *Fusobacterium* species were retained. A total of 56 *Fusobacterium* isolates were recovered and subjected to DNA extraction using the Wizard genomic extraction kit (Promega, USA). For each isolate, 1 ng DNA was used to prepare the sequencing library using the Nextera XT library preparation kit, following the manufacturer's instruction. Normalized libraries were pooled and sequenced on an Illumina MiSeq platform to generate 250 bp paired-end reads.

Pangenome analysis, phylogenetic reconstruction and screening for virulence genes

FASTQC was used to check the sequencing quality of each read pair⁷⁰, and Trimmomatic v0.36 was used to trim sequencing adapters and low-quality reads⁷¹. For each isolate, the trimmed read set was input into Unicycler v0.4.9 to construct the de novo assembly, using default parameters, and contigs of size over 500 bp were retained⁷². The assemblies were checked for traces of contamination using Checkm, and three assemblies were shown contaminated and discarded⁷³. The resulting assemblies were of adequate quality, with median size of 2,125,169 bp [IQR: 2,067,843–2,168,429], median number of contigs of 133 [IQR: 86–173] and the median N50 of 35,535 bp [IQR: 21,685–51,953]. Prokka v1.13 was used to annotate the assemblies, using the well-annotated *F. nucleatum* 23726 (accessed via FusoPortal) as reference⁷⁴. To provide preliminary taxonomic classification up to the subspecies level, FastANI was used to calculate the average nucleotide identity (ANI) between the individual assembly and a set of *Fusobacterium* references, with an ANI value ≥ 95% denoting a shared species/subspecies⁷⁵. The pangenomes of 57 *F. nucleatum/hwasookii* isolates (38 sequenced herein plus 19 references) and 25 *F. periodonticum* isolates (15 sequenced herein plus 10 references) were constructed separately using panX⁷⁶. The respective core genome from each species complex was aligned, with invariant sites removed, producing SNP alignments of 89,900 bp (*F. nucleatum/hwasookii* complex) and 106,738 bp (*F. periodonticum* complex). These were input into RAxML to construct maximum likelihood phylogenies, under the GTRGAMMA substitution model with 300 rapid bootstraps⁷⁷. Using the pangenome analysis output, we screened for the presence of several known *Fusobacterium* virulence genes (*fap2*, *fadA*, *radD*, *cbpF*). The intact presence or synteny of each genetic element was checked manually by gene alignment (Seaview) or genome visualization (Artemis) tools^{78,79}. Visualization of phylogenetic tree and associated metadata was performed using package 'ggtree'⁸⁰. Individual protein sets were aligned and inspected in Seaview, and phylogenies were constructed in RAxML, using the PROTGAM-MAGTR model and 200 rapid bootstraps.

Intra-clonal variation examination

To investigate intra-clonal variation with high confidence, we examined single nucleotide variants (SNV) among isolates belonging to the same phylogenetic cluster (Fig. 4 and Table 2), using the mapping approach recommended previously⁸¹. For each phylogenetic cluster, trimmed fastq files from the isolates were concatenated and input into Unicycler to construct a pan-assembly, with contigs less than 500 bp removed. This pan-assembly was ordered against an appropriate *Fusobacterium* reference using ABACAS, creating a pseudogenome reference⁸². Trimmed paired-end reads from each isolate were mapped against this reference using a custom wrapper script. Briefly, mapping was conducted using BWA MEM algorithm and samtools v1.8^{83,84}, with duplicate reads removed using PICARD, followed by indel realignment by GATK⁸⁵. SNVs were detected using the haplotype-based caller FreeBayes⁸⁶, and low quality SNVs were removed using bcftools if they met any of the following criteria: consensus quality <30, mapping quality <30, read depth <4, ratio of SNVs to reads at a position (AO/DP) < 85%, coverage on the forward or reverse strand <1. The bcftools 'consensus' command was used to generate a pseudosequence⁸⁷, integrating the filtered SNVs and invariant sites, and masking the low mapping region (depth <4) and low-quality SNVs with 'N'. The presence of high quality SNVs were validated by manual visualization of output bam files in Artemis, and SNV pertaining to recombination, transposons, plasmids, or repetitive elements were excluded from interpretation.

DATA AVAILABILITY

Raw sequence data are available in the NCBI Sequence Read Archive, including ones for 16S rRNA sequencing (BioProject PRJNA791834) and *Fusobacterium* whole genome sequencing (BioProject PRJNA791829). The source data underlying Figs. 1c, 1d, 2a-c, 3 and Supplementary Figures 5B and 6 are provided as a Source Data file.

CODE AVAILABILITY

Source data and R codes used for the microbiome analysis and visualization are deposited in Github (https://github.com/Hao-Chung/Vietnam_CRC_microbiome).

Received: 27 April 2022; Accepted: 14 October 2022;

Published online: 29 October 2022

REFERENCES

- Bray, F. et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **68**, 394–424 (2018).
- Arnold, M. et al. Global patterns and trends in colorectal cancer incidence and mortality. *Gut* **66**, 683–691 (2017).
- Keum, N. N. & Giovannucci, E. Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. *Nat. Rev. Gastroenterol. Hepatol.* **16**, 713–732 (2019).
- Schirmer, M. et al. Linking the human gut microbiome to inflammatory cytokine production capacity. *Cell* **167**, 1125–1136.e8 (2016).
- Song, M. & Chan, A. T. Environmental factors, gut microbiota, and colorectal cancer prevention. *Clin. Gastroenterol. Hepatol.* **17**, 275–289 (2019).
- Castellarin, M. et al. *Fusobacterium nucleatum* infection is prevalent in human colorectal carcinoma. *Genome Res.* **22**, 299–306 (2012).
- Kostic, A. D. et al. Genomic analysis identifies association of *Fusobacterium* with colorectal carcinoma. *Genome Res.* **22**, 292–298 (2012).
- Flemer, B. et al. Tumour-associated and non-tumour-associated microbiota in colorectal cancer. *Gut* **66**, 633–643 (2017).
- Wirbel, J. et al. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat. Med.* **25**, 679–689 (2019).
- Thomas, A. M. et al. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat. Med.* **25**, 667–678 (2019).

- Brennan, C. A. & Garrett, W. S. *Fusobacterium nucleatum* — symbiont, opportunist and oncobacterium. *Nat. Rev. Microbiol.* **17**, 156–166 (2019).
- Rubinstein, M. R. et al. *Fusobacterium nucleatum* promotes colorectal carcinogenesis by modulating E-Cadherin/ β -catenin signaling via its FadA adhesin. *Cell Host Microbe* **14**, 195–206 (2013).
- Abed, J. et al. Fap2 mediates *Fusobacterium nucleatum* colorectal adenocarcinoma enrichment by binding to tumor-expressed Gal-GalNAc. *Cell Host Microbe* **20**, 215–225 (2016).
- Casasanta, M. A. et al. *Fusobacterium nucleatum* host cell binding and invasion induces IL-8 and CXCL1 secretion that drives colorectal cancer cell migration. *Sci. Signal.* **13**, 1–13 (2020).
- Geng, F., Zhang, Y., Lu, Z., Zhang, S. & Pan, Y. *Fusobacterium nucleatum* caused DNA damage and promoted cell proliferation by the Ku70/p53 pathway in oral cancer cells. *DNA Cell Biol.* **39**, 144–151 (2020).
- Guo, P. et al. FadA promotes DNA damage and progression of *Fusobacterium nucleatum*-induced colorectal cancer through up-regulation of chk2. *J. Exp. Clin. Cancer Res.* **39**, 1–13 (2020).
- Mima, K. et al. *Fusobacterium nucleatum* in colorectal carcinoma tissue and patient prognosis. *Gut* **65**, 1973–1980 (2016).
- Serna, G. et al. *Fusobacterium nucleatum* persistence and risk of recurrence after preoperative treatment in locally advanced rectal cancer. *Ann. Oncol.* **31**, 1366–1375 (2020).
- Salvucci, M. et al. Patients with mesenchymal tumours and high Fusobacteriales prevalence have worse prognosis in colorectal cancer (CRC). *Gut* **71**, 1600–1612 (2022).
- Bullman, S. et al. Analysis of *Fusobacterium* persistence and antibiotic response in colorectal cancer. *Science* **358**, 1443–1448 (2017).
- Althoff, T. et al. Large-scale physical activity data reveal worldwide activity inequality. *Nature* **547**, 336–339 (2017).
- Nguyen, S. M. et al. Projecting cancer incidence for 2025 in the 2 largest populated cities in Vietnam. *Cancer Control* **26**, 1–13 (2019).
- Lin, H. & Peddada, S. D. Analysis of compositions of microbiomes with bias correction. *Nat. Commun.* **11**, 1–11 (2020).
- Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
- Martin, B. D., Witten, D. & Willis, A. D. Modeling microbial abundances and dysbiosis with beta-binomial regression. *Ann. Appl. Stat.* **14**, 94–115 (2020).
- Fang, H., Huang, C., Zhao, H. & Deng, M. CCLasso: correlation inference for compositional data through Lasso. *Bioinformatics* **31**, 3172–3180 (2015).
- Kurtz, Z. D. et al. Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.* **11**, e1004226 (2015).
- Brewer, M. L. et al. *Fusobacterium* spp. target human CEACAM1 via the trimeric autotransporter adhesin CbpF. *J. Oral Microbiol.* **11**, 1565043 (2019).
- Wu, C. et al. Genetic and molecular determinants of polymicrobial interactions in *Fusobacterium nucleatum*. *Proc. Natl Acad. Sci. USA* **118**, e2006482118 (2021).
- Galaski, J. et al. *Fusobacterium nucleatum* CbpF mediates inhibition of T cell function through CEACAM1 activation. *Front. Cell. Infect. Microbiol.* **11**, 1–8 (2021).
- Warren, R. L. et al. Co-occurrence of anaerobic bacteria in colorectal carcinomas. *Microbiome* **1**, 1–12 (2013).
- Drewes, J. L. et al. High-resolution bacterial 16S rRNA gene profile meta-analysis and biofilm status reveal common colorectal cancer consortia. *npj Biofilms Microbiomes* **3**, 34 (2017).
- Meng, Q. et al. *Fusobacterium nucleatum* secretes amyloid-like FadA to enhance pathogenicity. *EMBO Rep.* **22**, 1–19 (2021).
- Gao, Z., Guo, B., Gao, R., Zhu, Q. & Qin, H. Microbiota dysbiosis is associated with colorectal cancer. *Front. Microbiol.* **6**, 1–9 (2015).
- Flemer, B. et al. The oral microbiota in colorectal cancer is distinctive and predictive. *Gut* **67**, 1454–1463 (2018).
- Xia, X. et al. Bacteria pathogens drive host colonic epithelial cell promoter hypermethylation of tumor suppressor genes in colorectal cancer. *Microbiome* **8**, 1–13 (2020).
- Chung The, H. et al. Assessing gut microbiota perturbations during the early phase of infectious diarrhea in Vietnamese children. *Gut Microbes* **9**, 38–54 (2018).
- Liang, S. et al. Gut microbiome associated with APC gene mutation in patients with intestinal adenomatous polyps. *Int. J. Biol. Sci.* **16**, 135–146 (2020).
- Wei, P. L. et al. Classification of changes in the fecal microbiota associated with colonic adenomatous polyps using a long-read sequencing platform. *Genes (Basel)* **11**, 1–14 (2020).
- He, Y. et al. Non-nucleatum *Fusobacterium* species are dominant in the Southern Chinese population with distinctive correlations to host diseases compared with *F. nucleatum*. *Gut* **70**, 810–812 (2021).
- Yeoh, Y. K. et al. Southern Chinese populations harbour non-nucleatum *Fusobacterium* possessing homologues of the colorectal cancer-associated FadA virulence factor. *Gut* **69**, 1998–2007 (2020).

42. Lagier, J. C. et al. Culturing the human microbiota and culturomics. *Nat. Rev. Microbiol.* **16**, 540–550 (2018).
43. Richardson, M. et al. Analysis of 16S rRNA genes reveals reduced Fusobacterial community diversity when translocating from saliva to GI sites. *Gut Microbes* **12**, 1–13 (2020).
44. Abed, J. et al. Colon cancer-associated Fusobacterium nucleatum may originate from the oral cavity and reach colon tumors via the circulatory system. *Front. Cell. Infect. Microbiol.* **10**, 1–12 (2020).
45. Komiya, Y. et al. Patients with colorectal cancer have identical strains of Fusobacterium nucleatum in their colorectal cancer and oral cavity. *Gut* **68**, 1335–1337 (2019).
46. Chung The, H. et al. Exploring the genomic diversity and antimicrobial susceptibility of Bifidobacterium pseudocatenulatum in a Vietnamese population. *Microbiol. Spectr.* **9**, e0052621 (2021).
47. Ailloud, F. et al. Within-host evolution of Helicobacter pylori shaped by niche-specific adaptation, intragastric migrations and selective sweeps. *Nat. Commun.* **10**, 2273 (2019).
48. Queen, J. et al. Comparative analysis of colon cancer-derived fusobacterium nucleatum subspecies: inflammation and colon tumorigenesis in murine models. *MBio* **13**, e0299121 (2022).
49. Slade, D. J. New roles for Fusobacterium nucleatum in cancer: target the bacteria, host, or both? *Trends Cancer* **7**, 185–187 (2021).
50. Mima, K. et al. Fusobacterium nucleatum in colorectal carcinoma tissue according to tumor location. *Clin. Transl. Gastroenterol.* **7**, e200 (2016).
51. American Joint Committee on Cancer. Chapter 20 - colon and rectum. In: *AJCC Cancer Staging Manual 8th edition* (Springer, 2017).
52. Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K. & Schloss, P. D. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl. Environ. Microbiol.* **79**, 5112–5120 (2013).
53. Gohl, D. et al. An optimized protocol for high-throughput amplicon-based microbiome profiling. *Protoc. Exch.* 1–28 <https://doi.org/10.1038/protex.2016.030> (2016).
54. R Core Team. *R: a language and environment for statistical computing* (R Core Team, 2016).
55. McMurdie, P. J. & Holmes, S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* **8**, e61217 (2013).
56. Silverman, J. D., Washburne, A. D., Mukherjee, S. & David, L. A. A phylogenetic transform enhances analysis of compositional microbiota data. *Elife* **6**, 1–20 (2017).
57. Oksanen, J. et al. *vegan: Community ecology package*. R package version 2.3-5. (2016).
58. Callahan, B. J., Sankaran, K., Fukuyama, J. A., McMurdie, P. J. & Holmes, S. P. Bioconductor workflow for microbiome data analysis: from raw reads to community analyses. *F1000Research* **5**, 1492 (2016).
59. Callahan, B. J., McMurdie, P. J. & Holmes, S. P. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* **11**, 2639–2643 (2017).
60. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73**, 5261–5267 (2007).
61. Mirarab, S. et al. PASTA: ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. *J. Comput. Biol.* **22**, 377–386 (2015).
62. Nguyen, L. T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
63. Washburne, A. D. et al. Methods for phylogenetic analysis of microbiome data. *Nat. Microbiol.* **3**, 652–661 (2018).
64. Tibshirani, R., Walthers, G. & Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc.* **63**, 411–423 (2001).
65. Ishwaran, H. & Kogalur, U. B. *Random Forests for Survival, Regression and Classification (RF-SRC)*, R package version 2.2.0. (2016).
66. Calgano, M., Romualdi, C., Waldron, L., Rizzo, D. & Vitulo, N. Assessment of statistical methods from single cell, bulk RNA-seq, and metagenomics applied to microbiome data. *Genome Biol.* **21**, 1–31 (2020).
67. Nearing, J. T. et al. Microbiome differential abundance methods produce different results across 38 datasets. *Nat. Commun.* **13**, 1–16 (2022).
68. Stevens, J. R., Herrick, J. S., Wolff, R. K. & Slattery, M. L. Power in pairs: assessing the statistical value of paired samples in tests for differential expression. *BMC Genom.* **19**, 1–13 (2018).
69. Hirano, H. & Takemoto, K. Difficulty in inferring microbial community structure based on co-occurrence network approaches. *BMC Bioinforma.* **20**, 1–14 (2019).
70. Andrews, S. FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/> (2010).
71. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
72. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.* **13**, 1–22 (2017).
73. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
74. Todd, S. M., Settledge, R. E., Lahmers, K. K. & Slade, D. J. Fusobacterium genomics using MinION and illumina sequencing enables genome completion and correction. *mSphere* **3**, 1–9 (2018).
75. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 1–8 (2018).
76. Ding, W., Baumdicker, F. & Neher, R. A. panX: pan-genome analysis and exploration. *Nucleic Acids Res.* **46**, 1–12 (2018).
77. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
78. Gouy, M., Guindon, S. & Gascuel, O. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* **27**, 221–224 (2010).
79. Carver, T. et al. Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics* **24**, 2672–2676 (2008).
80. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T. Y. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2016).
81. Key, F. M. et al. On-person adaptive evolution of Staphylococcus aureus during atopic dermatitis increases disease severity. Preprint at <https://doi.org/10.1101/2021.03.24.436> (2021).
82. Assefa, S., Keane, T. M., Otto, T. D., Newbold, C. & Berriman, M. ABACAS: Algorithm-based automatic contiguation of assembled sequences. *Bioinformatics* **25**, 1968–1969 (2009).
83. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at <https://arxiv.org/abs/1303.3997> (2013).
84. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
85. McKenna, A. et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
86. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. Preprint at <https://doi.org/10.48550/arXiv.1207.39> (2012).
87. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, 1–4 (2021).

ACKNOWLEDGEMENTS

HCT is a Wellcome International Training Fellow (218726/Z/19/Z). L.J.H. is supported by Wellcome Trust Investigator Awards 100974/C/13/Z and 220876/Z/20/Z; the Biotechnology and Biological Sciences Research Council (BBSRC), Institute Strategic Programme Gut Microbes and Health BB/R012490/1, and its constituent projects BBS/E/F/000PR10353 and BBS/E/F/000PR10356. S.B. is a Wellcome Senior Research Fellow (215515/Z/19/Z). The authors wish to thank all participants and their caretakers for their participation in the study, and the pathology laboratory of Binh Dan Hospital for the tumour/biopsy pathology results.

AUTHOR CONTRIBUTIONS

H.N.H.T. and T.N.H.T. performed laboratory experiments, validated and curated data, performed data analysis and wrote the paper. H.N.H.T. and T.N.H.T. contributed equally to this work. P.H.N., C.N.V., N.T.N., V.N.D.T., K.A.V., T.D.H., B.L.D., P.M.N., V.C.H. coordinated and performed the patient recruitment, sample and data collection at the study site. C.N.N.M., T.N.T.N., T.T.V., and T.P.D. helped perform laboratory experiments. T.N.H.T., K.V.D., S.B. and H.C.T. conceived the study's design. D.T.P., G.E.T., S.B. and H.C.T. provided resources for the study. D.T.P., G.E.T., L.J.H., D.J.S., V.H.T. supervised the study. K.V.D., D.T.P., G.E.T., L.J.H., D.J.S., S.B., V.H.T. revised and edited the paper. H.C.T. revised data analysis plan, performed data analysis and wrote the paper.

COMPETING INTERESTS

The authors declare no competing interests.

CONSENT FOR PUBLICATION

This study does not publish any identifiable details related to the participant.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41522-022-00351-7>.

Correspondence and requests for materials should be addressed to Hao Chung The.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022