Southern Chinese populations harbour non-nucleatum Fusobacteria possessing homologues of the colorectal cancer-associated FadA virulence factor

Yun Kit Yeoh 1,2, Zigui Chen 1,2, Martin C S Wong 1,3, Mamie Hui 1,2, Jun Yu 1,4, Siew C Ng 1,4, Joseph J Y Sung 1, Francis K L Chan 1,4, Paul K S Chan 1,2

ABSTRACT

Objective Fusobacteria are not common nor relatively abundant in non-colorectal cancer (CRC) populations, however, we identified multiple *Fusobacterium* taxa nearly absent in western and rural populations to be comparatively more prevalent and relatively abundant in southern Chinese populations. We investigated whether these represented known or novel lineages in the *Fusobacterium* genus, and assessed their genomes for features implicated in development of cancer.

Methods Prevalence and relative abundances of fusobacterial species were calculated from 3157 CRC and non-CRC gut metagenomes representing 16 various biogeographies. Microbial genomes were assembled and compared with existing reference genomes to assess novel fusobacterial diversity. Phylogenetic distribution of virulence genes implicated in CRC was investigated.

Results Irrespective of CRC disease status, southern Chinese populations harboured increased prevalence (maximum 39% vs 7%) and relative abundances (average 0.4% vs 0.04% of gut community) of multiple recognised and novel fusobacterial taxa phylogenetically distinct from *Fusobacterium nucleatum*. Genomes assembled from southern Chinese gut metagenomes increased existing fusobacterial diversity by 14.3%. Homologues of the FadA adhesin linked to CRC were consistently detected in several monophyletic lineages sister to and inclusive of *F. varium* and *F. ulcerans*, but not *F. mortiferum*. We also detected increased prevalence and relative abundances of *F. varium* in CRC compared with non-CRC cohorts, which together with distribution of FadA homologues supports a possible association with gut disease.

Conclusion The proportion of fusobacteria in guts of southern Chinese populations are higher compared with several western and rural populations in line with the notion of environment/biogeography driving human gut microbiome composition. Several non-nucleatum taxa possess FadA homologues and were enriched in CRC cohorts; whether this imposes a risk in developing CRC and other gut diseases deserves further investigation.

INTRODUCTION

*Fusobacterium nucleatum* is a bacterial pathogen most well-known for its association with colorectal cancer (CRC) in humans. Irrespective of biogeography, multiple studies have consistently reported enrichment of *F. nucleatum* in the guts of CRC and tumour tissue of CRC subjects compared with non-CRC cohorts. Furthermore, the association between *F. nucleatum* and CRC has been demonstrated through cell model studies, implicating two proteins FadA and Fap2 in facilitating adherence, invasion and induction of oncogenic and inflammatory responses in CRC cells by *F. nucleatum*. We also detected increased prevalence of *F. varium*, sister to and inclusive of *F. ulcerans*, but not *F. mortiferum*. We also detected increased prevalence and relative abundances of *F. varium* in CRC compared with non-CRC cohorts, which together with distribution of FadA homologues supports a possible association with gut disease.
In contrast, relatively less is known about the biology of fusobacterial species other than *F. nucleatum* and their roles in human health, if any. According to the List of Prokaryotic names with Standing in Nomenclature (LPSN), there are 21 recognised species in the *Fusobacterium* genus at the time of writing. Apart from *F. nucleatum*, a few other species such as *F. necrophorum*, *F. gondiiiformans*, *F. periodonticum*, *F. mortiferum*, *F. ulcera*ns and *F. varium* have been reported in human-associated samples. For example, *F. necrophorum* are often associated with thrombophlebitis of the internal jugular vein (termed Lemiere’s syndrome), *F. gondiiiformans* found in urogenital and intestinal tracts, *F. periodonticum* in oral cavities associated with squamous cell carcinoma, *F. ulcera*ns in skin ulcers and *F. varium* in human guts associated with ulcerative colitis (UC). Apart from cases of disease, their prevalence and relative abundances in guts of healthy individuals are relatively low, often below detection thresholds consistent with the notion that the presence of *Fusobacterium* in human guts is specifically associated with CRC.

We initially produced shotgun metagenomes using stools collected from 556 self-reported healthy individuals recruited for establishing a gut microbiota databank in Hong Kong (HKGutMicMap project). These data were analysed together with publicly-available stool metagenomes of other healthy subjects from Hong Kong, Austria, China, Denmark, France, Germany, Israel, Spain, Sweden and the USA as well as several rural populations from El Salvador, Peru, Fiji, Mongolia and Tanzania to assess variation in gut microbial community composition across different biogeographies. We serendipitously observed a consistent increase in prevalence and relative abundance of multiple fusobacterial species in the Hong Kong, Chinese and Spanish but not American, European and other rural cohorts, concordant with the idea of variation in human gut microbiomes primarily driven by environment/geography. Here, we reconstructed fusobacterial genomes from the Hong Kong gut metagenomes and showed that *F. varium*, *F. ulcera*ns, *F. mortiferum* and other as yet uncharacterised fusobacterial taxa are prevalent in this population. We then investigated whether these genomes contained characteristics that could indicate potential associations with cancer and/or disease. Findings reported here suggest that the fusobacterial lineages prevalent in the Chinese gut possess genomic potential to facilitate development of CRC and possibly other diseases.

### MATERIALS AND METHODS

**HKGutMicMap cohort sample collection and DNA sequencing**

Subjects were recruited from the Hong Kong public as part of the HKGutMicMap study to generate gut microbiome profiles representative of the local, non-disease population. A research associate measured parameters such as body weight, waist circumference, body height and blood pressure and subjects were provided stool collection kits for self-collection. They were asked to deliver fresh stools to the laboratory within 2 hours of defecation. Faecal specimens were stored at −80°C until further processing. DNA was extracted from 0.1 g homogenised fractions of stool using the DNeasy PowerSoil Kit (QIAGEN, Hilden, Germany) following manufacturer’s instructions. The concentration of extracted DNA was determined using the Qubit dsDNA BR Assay Kit (Thermo Fisher Scientific, Waltham, Massachusetts) and normalised to 20 ng/μl with 10 mM Tris-Cl. Normalised DNA samples were sent to a sequencing service provider (Novogene HK Company Limited, Wan Chai, Hong Kong) for library preparation and paired-end shotgun metagenomic sequencing (Illumina NovaSeq 6000). A mock community sequencing control (ZymoBIOMICS Microbial Community DNA Standard, catalogue number D6305, Zymo Research, Irvine, California) was included.

**Prevalence and relative abundances of fusobacterial species based on gut metagenomes**

To examine prevalence and relative abundances of fusobacterial species in guts of human populations from different geographical backgrounds, we included non-CRC gut metagenome sequence data generated by previous studies in Hong Kong, China, USA, Austria, Denmark, France, Germany, Spain, Israel, Sweden, El Salvador, Peru, Fiji, Mongolia and Tanzania (online supplementary table S1). The Hong Kong, Austria, France, Germany and one USA cohort were comprised of CRC and non-CRC subjects. These gut metagenome data sets were chosen because they have been binned for microbial genomes. Together with data generated from the HKGutMicMap cohort (present study), raw sequences were quality-filtered using Trimmomatic V.0.38 to remove adapter and low quality regions. Next, microbial community compositional profiles were inferred from quality-filtered sequences (forward reads) using MetaPhlAn2 V.2.6 with the v20 database. For each fusobacterial species identified by MetaPhlAn2, their prevalence rates were calculated based on the number of samples each species was detected in (ie, relative abundance >0%) divided by the total number of samples in the respective cohorts.

**Binning fusobacterial population genomes from metagenomes**

To explore genomic diversity of fusobacterial species in the Hong Kong population, we assembled metagenomes from the Hong Kong cohorts and binned population genomes (termed metagenome-assembled genomes (MAGs)) from each de novo assembly. Overlapping sequence pairs in the quality-filtered data were first merged to produce longer sequences, and then assembled together with unmerged pairs using MEGAHIT V.0.1.1. Sequence coverage profiles were then obtained by mapping quality-filtered reads to their respective assemblies using BWA-MEM V.0.7.17. With these coverage information, MAGs were binned from each of the metagenomes using MetaBAT V.2.10.2, MetaBAT V.2.12.1 and MaxBin V.2.2.5. A non-redundant set of MAGs were calculated by merging output from the three sets of bins using DASTool V.1.1.0. The resulting non-redundant MAGs were quality-checked using the lineage workflow in CheckM V.1.0.13. MAGs with >90% completeness and <5% contamination were retained, and their taxonomy was inferred using Genome Taxonomy Database (GTDB) toolkit (GTDB-Tk) V.0.2.2 database release 86.

**Construction of fusobacterial phylogenetic trees**

We downloaded fusobacterial reference genomes from the National Centre for Biotechnology Information (NCBI) RefSeq database (release 89), and MAGs from recent publications that have assembled microbial genomes from the human metagenome data sets used above. These genomes were checked for completeness and contamination using CheckM, and only those with >90% completeness and <5% contamination were retained. We constructed two phylogenetic trees of the *Fusobacterium* genus—one using a dereplicated set of fusobacterial reference genomes and MAGs to highlight existing genomic diversity, and the other using all genomes to explore...
distribution of putative virulence protein homologues in this genus (described in section below). We included Cetobacterium, a member of the Fusobacteriaceae family as outgroup taxa in both trees. For the first tree, genomes were dereplicated using dRep V.1.4.3 based on genome distances and average nucleotide identities (ANI), and a concatenated amino acid alignment of 120 phylogenetically informative single-copy bacterial marker genes was generated using GTDB-Tk. A maximum-likelihood tree was constructed based on this alignment using RAxML. Both trees were midpoint-rooted using GenomeTreeTk V.0.0.53 (https://github.com/dparks1134/GenomeTreeTk).

Annotating genes in fusobacterial genomes
First, protein-coding sequences in all MAGs and reference genomes were translated into amino acid sequences using Prodigal. Amino acid sequences were aligned against the UniRef100 database (March 2018) using DIAMOND V.0.9.24 (≥30% sequence identity and ≥70% alignment length between query and reference) to identify gene families, and aligned counts were collated according to Kyoto Encyclopaedia of Genes and Genomes orthology to infer presence/absence of gene families. ANI comparisons were performed using FastANI V.1.1.

To explore presence/absence of two known CRC-associated fusobacterial genes FadA and Fap2, we annotated fusobacterial genomes using eggNOG-mapper V.2.0.1 with reference to eggNOG database V.5.0. The presence/absence of FadA and Fap2 homologues was visualized on a phylogenetic tree consisting of all 663 fusobacterial genomes. Amino acid gene trees were constructed for both FadA and Fap2 by aligning the putative homologues using MAFFT V.7.407 and inferring a maximum likelihood tree using RAxML. Both trees were midpoint-rooted using GenomeTreeTk V.0.0.53 (https://github.com/dparks1134/GenomeTreeTk).

Isolation of Fusobacterium from stools
Frozen stools were thawed and diluted in brain heart infusion culture medium. Dilutions were inoculated onto blood agar plates and anaerobically cultured at 37°C for 2 days. Colonies were identified using a MALDI Biotyper (Bruker, Billerica, Massachusetts). Colonies identified as Fusobacterium were subcultured onto fresh blood agar plates. Genomic DNA was extracted from pure cultures using a Gentra Puregene Yeast/Bact. DNA isolation Kit (Qiagen, Hilden, Germany), and sent to Novogene HK for library preparation and paired-end shotgun metagenomic sequencing (Illumina NovaSeq 6000). Trimomatic quality-filtered reads were assembled using MEGAHIT V.1.1.1 and annotated using eggNOG-mapper V.2.0.1 with reference to eggNOG database V.5.0.

Data availability
Raw sequence data generated for this study are available in the Sequence Read Archive under BioProject accession PRJNA557323.

RESULTS
The HKGutMicMap cohort
At the time of analysis, the HKGutMicMap cohort representing the general population in Hong Kong consisted of 556 subjects with shotgun metagenome data. These subjects were self-reported as healthy with no chronic disease. There were 294 females to 262 males, and their median age at time of sample collection was 51 years (SD 16.3 years). The median body mass index was 22.7 kg m⁻² (SD 3.4). These and other parameters such as body weight, blood pressure and waist circumference are listed in online supplementary table S2.

F. mortiferum, F. ulcerans and F. varium are prevalent in Chinese populations irrespective of CRC disease status
In total, 3157 stool metagenomes comprising non-CRC and CRC subjects were included in this study. These metagenomes represent populations from China (Hong Kong, Shenzhen and Zhejiang), USA, Austria, Denmark, France, Germany, Spain, Sweden, Israel, El Salvador, Peru, Fiji, Mongolia and Tanzania. To assess the distribution of fusobacterial species across biogeography of these populations, quality-filtered sequences from each metagenome were mapped to lineage-specific marker genes using MetaPhlAn2 to produce prevalence and relative abundance estimates.

In non-CRC subjects (n=2515), overall gut microbial community composition significantly differed among cohorts (p<0.05, permutational multivariate analysis of variance; figure 1A, online supplementary figure S1). At the phylum level, the three Chinese and USA cohorts had higher relative abundances of Bacteroidetes compared with Firmicutes (63% vs 30%), whereas Firmicutes were more relatively abundant in the other cohorts compared with Bacteroidetes (55% vs 29%). Peru was the exception with Actinobacteria being the most dominant phylum (60%) (online supplementary table S3). In addition, Spirochaetes were detected at >1% relative abundance only in El Salvador, Fiji and Tanzania cohorts, compared with an average 0.004% in the Western and Chinese cohorts. Fusobacterium were also more relatively abundant in the Chinese and Spanish (average 0.47%) compared with other cohorts (0.01%). We were interested in the comparatively higher relative abundances of Fusobacterium since F. nucleatum has been widely implicated in CRC. Within the fusobacterial genus, F. mortiferum, F. nucleatum, F. ulcerans and F. varium were more prevalent and relatively abundant in the Chinese cohorts relative to others including Spain (p<0.001, Kruskal-Wallis test adjusted for false discovery rate) (table 1, figure 1B).

In CRC subjects (n=642), average fusobacterial relative abundances in subjects from Hong Kong were higher than the USA, German and Austrian, but not the French cohort (online supplementary figure S2). F. nucleatum was detected in all six cohorts as expected for CRC. F. varium, F. ulcerans and F. mortiferum were more prevalent in Hong Kong (online supplementary table S4). In addition, the French were enriched in F. goniodiformans and F. necrophorum relative to the others. F. ulcerans was also present in the Austrian cohort, however, its prevalence was still six fold higher in Hong Kong. These findings indicate that F. mortiferum, F. ulcerans and F. varium are typically more common and detected at higher relative abundances in the guts of Hong Kong populations compared with some North Americans and Europeans, irrespective of CRC disease status.

Several fusobacterial species other than F. nucleatum are enriched in CRC
F. nucleatum and more broadly the Fusobacterium genus have been shown to be enriched in the guts of CRC patients compared with non-CRC controls, although associations between CRC and other fusobacterial species have not been specifically mentioned. Since F. mortiferum, F. ulcerans and F. varium were more prevalent and relatively abundant in the guts of Chinese populations, we wanted to know whether their distributions and abundances were changed in association with CRC akin to...
sequences from El Salvador, Peru,36 Fiji,37 Mongolia38 and Tanzania.39 40 Relative abundances were calculated using MetaPhlAn2 on quality-

subjects in studies with case-

Table S7). In addition, recent efforts in characterising genomic diversity of these less characterised fusobacterial lineages. Using metagenomes from the Hong Kong cohort (inclusive of non-CRC and CRC subjects), we binned 171 high quality fusobacterial MAGs (>90% complete, <5% contamination based on CheckM’s lineage workflow) (online supplementary table S7). Another four high quality fusobacterial MAGs we previously binned from gut metagenomes of patients from a clinic were also included in this study (annotated in online supplementary table S7). Together with 152 high quality fusobacterial genomes from RefSeq R89, we first dereplicated these 663 genomes and inferred a genome tree to establish their phylogenetic relationships to one another. Dereplication was performed to highlight existing genome diversity in the Fusobacterium genus, resulting in a phylogenetic tree comprising 218 unique fusobacterial genomes. Taxonomic information for all MAGs and reference genomes inferred according to the GTDB release 86 v2 was then appended onto the phylogenetic tree (figure 2).

Four major monophyletic lineages (termed clades) were resolved within the Fusobacterium genus congruent with taxonomic inferences produced by GTDB (denoted with suffixes _A, _B and _C) (figure 1). The clade denoted as Fusobacterium_A was comprised of F. nucleatum including its traditional subspecies animalis, vincentii, nucleatum and polymorphum, F. hwasookii, F. periodonticum, F. massiliense and F. russii. Fusobacterium_B was comprised of F. ulcerans, F. varium, F. mortiferum and various unclassified fusobacterial genomes; Fusobacterium_C comprised of F. perfoetens and other unclassified genomes; Fusobacterium_A and Fusobacterium_B clades were highly represented by genomes derived from Hong Kong and Chinese metagenomes (48 of 67 genomes; many genomes from RefSeq do not have accompanying geographic information and were assumed to be of non-Chinese origin) (figure 2). In contrast, the Fusobacterium_A and Fusobacterium_C clades were more represented by genomes from other regions (only 7 of 151 genomes were from Chinese sources). MAGs from Chinese populations collectively increased phylogenetic diversity of the overall tree by 14.3% based on branch lengths, indicating that the Chinese gut harbours novel fusobacterial diversity not yet represented by reference genomes.

To demonstrate that these novel fusobacteria were indeed more prevalent in Chinese populations, we mapped sequences from all non-CRC samples to the dereplicated set of 218 fusobacterial genomes and counted the proportion of aligned sequences in each cohort. The Chinese had 10–100-fold higher proportions of
| Country              | Fusobacterium gonidiaformans | F. mortiferum | F. necrophorum | F. nucleatum | F. periodonticum | F. oraltaxon 370 | F. ulcerans | F. varium |
|----------------------|------------------------------|---------------|----------------|--------------|------------------|-----------------|-------------|
| Hong Kong (China)    | 0.384                        | 15.493        | 0.128          | 4.609        | 0.640            | 0.128           | 9.475       | 13.828     |
| Prevalence (%)       |                              |               |                |              |                  |                 |             |
| Relative abundance (%)| 0.000                        | 0.236         | 0.000          | 0.002        | 0.000            | 0.000           | 0.025       | 0.044      |
| Shenzhen (China)     | 0.000                        | 10.526        | 0.000          | 13.158       | 7.895            | 0.000           | 23.684      | 39.474     |
| Prevalence (%)       |                              |               |                |              |                  |                 |             |
| Relative abundance (%)| 0.000                        | 0.299         | 0.000          | 0.001        | 0.000            | 0.000           | 0.077       | 0.167      |
| Zhejiang (China)     | 0.000                        | 30.345        | 0.000          | 0.690        | 0.690            | 0.000           | 10.345      | 6.207      |
| Prevalence (%)       |                              |               |                |              |                  |                 |             |
| Relative abundance (%)| 0.000                        | 0.525         | 0.000          | 0.001        | 0.000            | 0.000           | 0.030       | 0.003      |
| Israel               | 2.667                        | 0.667         | 0.000          | 3.333        | 0.667            | 0.000           | 2.000       | 0.000      |
| Prevalence (%)       |                              |               |                |              |                  |                 |             |
| Relative abundance (%)| 0.007                        | 0.013         | 0.000          | 0.006        | 0.001            | 0.000           | 0.009       | 0.000      |
| Austria              | 0.000                        | 0.000         | 0.000          | 0.000        | 0.000            | 0.000           | 0.000       | 0.000      |
| Prevalence (%)       |                              |               |                |              |                  |                 |             |
| Relative abundance (%)| 0.000                        | 0.000         | 0.000          | 0.000        | 0.000            | 0.000           | 0.000       | 0.000      |
| Denmark              | 0.000                        | 0.000         | 0.000          | 0.000        | 0.000            | 0.000           | 0.000       | 0.000      |
| Prevalence (%)       |                              |               |                |              |                  |                 |             |
| Relative abundance (%)| 0.000                        | 0.000         | 0.000          | 0.000        | 0.000            | 0.000           | 0.000       | 0.000      |
| France               | 1.639                        | 1.639         | 0.000          | 1.639        | 1.639            | 0.000           | 1.639       | 0.000      |
| Prevalence (%)       |                              |               |                |              |                  |                 |             |
| Relative abundance (%)| 0.000                        | 0.000         | 0.000          | 0.000        | 0.000            | 0.000           | 0.024       | 0.000      |
| Germany              | 0.000                        | 0.000         | 0.000          | 0.000        | 0.000            | 0.000           | 0.000       | 0.000      |
| Prevalence (%)       |                              |               |                |              |                  |                 |             |
| Relative abundance (%)| 0.000                        | 0.000         | 0.000          | 0.000        | 0.000            | 0.000           | 0.000       | 0.000      |
| Spain                | 4.110                        | 1.826         | 1.370          | 6.849        | 3.196            | 0.000           | 1.370       | 0.000      |
| Prevalence (%)       |                              |               |                |              |                  |                 |             |
| Relative abundance (%)| 0.085                        | 0.130         | 0.000          | 0.085        | 0.000            | 0.000           | 0.032       | 0.000      |
| Sweden               | 2.098                        | 0.000         | 0.000          | 1.399        | 0.000            | 0.000           | 0.699       | 0.000      |
| Prevalence (%)       |                              |               |                |              |                  |                 |             |
| Relative abundance (%)| 0.003                        | 0.000         | 0.000          | 0.000        | 0.000            | 0.000           | 0.000       | 0.000      |
| USA                  | 0.980                        | 1.471         | 0.000          | 2.451        | 2.941            | 0.000           | 2.941       | 0.490      |
| Prevalence (%)       |                              |               |                |              |                  |                 |             |
| Relative abundance (%)| 0.000                        | 0.010         | 0.000          | 0.000        | 0.000            | 0.000           | 0.003       | 0.000      |
| El Salvador          | 0.000                        | 0.000         | 0.000          | 0.000        | 0.000            | 0.000           | 0.000       | 0.000      |
| Prevalence (%)       |                              |               |                |              |                  |                 |             |
| Relative abundance (%)| 0.000                        | 0.000         | 0.000          | 0.000        | 0.000            | 0.000           | 0.000       | 0.000      |
| Fiji                 | 4.545                        | 6.818         | 0.000          | 1.705        | 2.273            | 0.000           | 0.000       | 0.000      |
| Prevalence (%)       |                              |               |                |              |                  |                 |             |
| Relative abundance (%)| 0.003                        | 0.039         | 0.000          | 0.000        | 0.000            | 0.000           | 0.000       | 0.000      |
| Mongolia             | 0.000                        | 2.727         | 0.000          | 0.909        | 0.000            | 0.000           | 0.000       | 0.000      |
| Prevalence (%)       |                              |               |                |              |                  |                 |             |
| Relative abundance (%)| 0.000                        | 0.030         | 0.000          | 0.000        | 0.000            | 0.000           | 0.000       | 0.000      |
| Peru                 | 1.299                        | 0.000         | 0.000          | 0.000        | 0.000            | 0.000           | 0.000       | 0.000      |
| Prevalence (%)       |                              |               |                |              |                  |                 |             |
| Relative abundance (%)| 0.001                        | 0.000         | 0.000          | 0.000        | 0.000            | 0.000           | 0.000       | 0.000      |
| Tanzania             | 1.493                        | 0.000         | 0.000          | 1.493        | 0.000            | 0.000           | 0.000       | 0.000      |
| Prevalence (%)       |                              |               |                |              |                  |                 |             |
| Relative abundance (%)| 0.011                        | 0.000         | 0.000          | 0.002        | 0.000            | 0.000           | 0.000       | 0.000      |

CRC, colorectal cancer.
Figure 2  Phylogenetic tree showing evolutionary relationships among 218 fusobacterial genomes. Seven *Cetobacterium* genomes serve as outgroup to root the tree. Genomes in this figure are from a dereplicated set of 676 fusobacterial and *Cetobacterium* genomes assembled from gut metagenomes from Hong Kong (HKGutMicMap, Yu et al.2 and Coker et al.29) and other regions,42–44 and reference genomes downloaded from RefSeq (release 89). Reference genomes obtained from RefSeq are labelled with their corresponding accession numbers, while metagenome-assembled genomes have branch labels showing their country of origin (those from Hong Kong are in red text). All 676 genomes were >90% complete and had <5% contamination based on the lineage workflow in CheckM,51 and were dereplicated using dRep53 to highlight existing genome diversity of the *Fusobacterium* genus in this figure. A concatenated amino acid alignment was produced to infer taxonomy of the genomes according to the genome taxonomy database (GTDB),52 and subsequently used to construct maximum likelihood trees using RAxML.54 Four major monophyletic clades in the *Fusobacterium* genus are shaded and denoted with suffixes according to the GTDB. Branch colours are intended to delineate species boundaries (indicated by labels) and do not represent any taxa in particular; genomes without species designations have black branches. Black circles at nodes represent 100% bootstrap support unless otherwise indicated (no less than 90% bootstrap). Scale bar indicates number of amino acid substitutions per site.

with reference to intraspecies and interspecies cutoffs derived from published studies (intraspecies >95% ANI; interspecies 78%–95%).59 60 With these cutoffs, we identified (i) six putative species in the *Fusobacterium_B* clade not including *F. perfoetens*, (ii) three species basal to *F. mortiferum*, (iii) one species sister to *F. ulcerans*, (iv) one species sister to the *F. ulcerans* and *F. varium* lineage, (v) one species basal to the lineage containing *F. polymorphum*, *nucleatum*, *vencentii* and *animalis* and (vi) two species basal/sister to *F. animalis* (online supplementary figure S5, table S10). These genomes share <95% ANI to any circumscribed fusobacterial taxa, and could represent novel or one of the 18 recognised species in the LPSN yet to have genome representation. In addition to drawing species boundaries, we could infer the degree of intraspecies genome similarity by comparing the initial number of 663 genomes to the resulting number of dereplicated genome clusters. For example, we observed that *F. mortiferum* was highly clonal despite its high prevalence in Chinese populations, whereas *F. periodonticum* genomes were...
Colon

Figure 3  Distribution of FadA and Fap2 homologues in the *Fusobacterium* genus. Red and blue ticks next to branch tips indicate detection of FadA and Fap2 homologues, respectively, in the corresponding genomes. Homologues were identified using the eggNOG-mapper with reference to the eggNOG database V.5.0. This phylogenetic tree consists of 663 fusobacterial and 13 *Cetobacterium* genomes assembled from gut metagenomes from Hong Kong (HKGutMicMap cohort from this study, Yu et al. and Coker et al. cohorts) and other regions, and reference genomes downloaded from RefSeq (release 89). Reference genomes obtained from RefSeq are labelled with their corresponding accession numbers, while metagenome-assembled genomes are labelled with bin IDs. Genomes from Hong Kong have labels in red. All genomes were >90% complete and had <5% contamination based on the lineage workflow in CheckM. A concatenated amino acid alignment was produced to infer taxonomy of the genomes according to the genome taxonomy database (GTDB), and subsequently used to construct maximum likelihood trees using RAxML. Four major monophyletic clades in the *Fusobacterium* genus are shaded and denoted with suffixes according to the GTDB. Branch colours are intended to delineate species boundaries (indicated by labels) and do not represent any taxa in particular; genomes without species designations have black branches. Scale bar indicates number of amino acid substitutions per site.

more variable and formed more clusters of unique genomes compared with *F. mortiferum* (figures 2 and 3, online supplementary table S11).

**Fusobacterial genome features possibly associated with disease**

Previous functional analyses of CRC gut metagenomes have revealed features such as shifts towards amino acid degradation and trimethylamine (TMA) production via choline metabolism. We annotated the fusobacterial MAGs and observed that while they did not contain key genes involved in TMA production (TMA-lyase (*cutC*, K20038), and L-carnitine/gamma-butyrobetaine antiporter (*catT*, K05245)), several orthologues such as proline iminopeptidase (K01259), glutamate formiminotransferase (K00603) and tryptophanase (K01667) were prevalent in genomes from the *Fusobacterium* clade (online supplementary table S12). Moreover, *Fusobacterium* genomes possess genes that may be involved in the catabolism of amino acids and production of glucose (phosphoenolpyruvate carboxykinase K01610, fructose-bisphosphate aldolase K01623, oxaloacetate decarboxylase K01571), as well as several other features that could be linked to cancer such as iron scavenging (K07230, K07243, K11707, K11708, K11709, K11710), ceramide glucosyltransferase (K00720) involved in production of glycosylated sphingolipids and para-aminobenzoate synthase (K01664, K01665) in the production of folate. Likewise, urease (K01428–K01430) was detected in *Fusobacterium_A* and *Fusobacterium_B* clades but not *Fusobacterium*. Some of these features are consistent with those identified in CRC gut microbiota metagenomes, though it is important to point out...
that these findings do not imply that fusobacteria contribute wholly to the altered microbial functional signature in CRC guts as they are typically <1% relative abundance. In addition, the distribution of features by clades suggest that disease associations, if any, likely vary among the fusobacterial lineages.

**Homologues of colorectal cancer-associated FadA and Fap2 are present in several fusobacterial species**

Previous cell model studies have identified two proteins in *F. nucleatum* that allow the bacterium to potentiate CRC, namely the FadA adhesin and Fap2 lectin. To identify whether Fusobacterium species other than *F. nucleatum* also possess similar genes that may allow them to interact with CRC cells, we annotated all 663 fusobacterial genomes with reference to the eggNOG database. A phylogenetic tree incorporating all 663 genomes was constructed to visualise distribution of these homologues within the *Fusobacterium* genus. For FadA, a total of 999 homologues (online supplementary table S13) were identified in 311 genomes including all lineages belonging to the *Fusobacterium* clade, a monophyletic subset of *F. necrophorum*, and in *F. varium*, *F. ulcerans* and several uncharacterised monophyletic taxa in the *Fusobacterium_A* clade (figure 3). These FadA homologues possibly comprise three or more protein families as shown by a protein tree constructed from amino acid alignments. Sequences from the *Fusobacterium_A* clade were distinct compared with those from the *Fusobacterium* clade, while homologues from *F. necrophorum* were placed together with the *Fusobacterium* homologues (figure 4). These observations indicate that FadA homologues from *F. varium*, *F. ulcerans* and uncharacterised *Fusobacterium_A* lineages could have distinct roles compared with homologues found in the *Fusobacterium* clade. For Fap2, we identified 754 putative homologues in 288 genomes (online supplementary table S14). Lineages in which Fap2 homologues were identified largely overlapped with FadA, encompassing members of the *Fusobacterium* clade, a monophyletic subset of *F. necrophorum* genotypes, several *F. varium* and *F. ulcerans*, and in a subset of *Fusobacterium_B* clade genomes (figure 3, online supplementary figure S6). Overall distribution of the FadA and Fap2 homologues in the *Fusobacterium* genus suggests that the potential association with CRC may be present in several distinct fusobacterial lineages. Since some of these fusobacterial species were increased in relative abundance and prevalence in CRC compared with non-CRC subjects, the detection of FadA and Fap2 homologues suggest that species such as *F. varium* may have the ability to potentiate disease akin to *F. nucleatum*.

To check that the fusobacterial MAGs recovered from metagenome data are representative of actual genomes, we isolated and sequenced genomes of eight fusobacteria obtained from five stool samples. These genomes were classified as *Fusobacterium_A* (seven genomes) and *F. ulcerans* (one), and scored >99%ANI to MAGs recovered from Hong Kong gut metagenomes (online supplementary table S15). Moreover, they contained FadA and Fap2 homologues consistent with their phylogeny, providing confidence that MAGs recovered here indeed represent real microbial genomes. Nevertheless, we recognise that MAG validation with only eight isolates is inadequate, and more work is needed to verify MAGs representing other fusobacterial lineages.

**DISCUSSION**

While it has been established in human populations from various geographical backgrounds that *F. nucleatum* is associated with CRC, less is known about the distribution of other fusobacterial species in human guts. Here, we showed that fusobacterial lineages such as *F. ulcerans*, *F. varium*, *F. mortiferum* and multiple uncharacterised taxa are more prevalent in the guts of non-CRC Chinese and Spanish cohorts compared with counterparts from several geographical regions. While these non-nucleatum taxa may simply reflect biogeographical differences in human gut microbiome composition, we saw two lines of evidence suggesting that they could possess oncogenic and/or disease-causing potential: (i) increased prevalence and relative abundances in CRC compared with non-CRC cohorts (online supplementary table S5) and (ii) detection of virulence gene homologues in multiple monophyletic lineages (figure 3). Taken together, these evidence suggest that *F. periodonticum*, a subset of *F. necrophorum*, *F. varium* and *F. ulcerans* together with their uncharacterised sister lineages, *F. hwasookii*, *F. massiliense* and *F. russii* might have a role in the development of CRC. These implicated lineages are consistent with a set of ‘active versus passive invader’ species proposed by Manson McGuire and colleagues based on genomic features such as genome size, presence of FadA-related proteins, expanded number of membrane protein-encoding genes and MORN2 protein domains. In addition, their link to disease is supported by independent microbial community data and cell model studies. For example, a recent microbial community composition survey indicated that *F. periodonticum* in the oral cavity is associated with oral squamous cell carcinoma. Another example is of *F. necrophorum*, in which blood culture-based surveys of fusobacterial infections have indicated that this species was the second most common isolate after *F. nucleatum*. As for *F. varium* and *F. ulcerans*, less is known about their distribution and association with cancers or disease. Gut microbiota surveys in Japanese cohorts
have indicated that *F. varium* is associated with UC, and a genome sequencing study of *F. varium* strain Fv113-g1 isolated from a UC patient reported expression of FaDA homologues in monocultures simulating in vivo conditions within the human gut. Our data on FaDA homologues show that sequences from the *Fusobacterium* A clade (*F. varium, F. ulcerans* and other uncharacterised sister taxa) are not identical to sequences from the *Fusobacterium* clade (in which the CRC-associated *F. nucleatum* is located) (figure 4), thereby suggesting distinct functions or targets among these homologues. While the presence/absence of these gene homologues do not directly translate to invasiveness, we postulate that *Fusobacterium* A taxa and their copies of the FaDA homologue can be risk factors for diseases other than CRC.

In light of the inference that *Fusobacterium* A taxa are prevalent in Chinese populations and could be potential risk factors of disease in humans, a limitation of this study is the lack of published data or cultured isolates to validate our observations. Findings reported here imply that disease-association is possible in this clade, from which we have isolated eight *Fusobacterium* A members with genomes that match MAGs recovered from metagenomes. The immediate next step is to test these isolates in cell and animal model experiments to determine whether they have the potential to facilitate CRC or other diseases akin to *F. nucleatum*. Specifically, the role of virulence gene homologues such as FaDA and Fap2 can be studied via knockout/knockdown experiments to assess their impact on disease outcomes. Following this, further isolation and testing of other uncharacterised fusobacterial lineages will provide a more comprehensive understanding of the biology and disease associations outside the *F. nucleatum* complex.

In conclusion, while fusobacterial species other than *F. nucleatum* have not been identified as risk factors likely owing to their almost non-existence in western populations and ubiquity in non-CRC southern Chinese populations, our findings suggest that there is potential in some of these prevalent but overlooked fusobacterial lineages to facilitate CRC. If any positive associations are confirmed, individuals carrying the corresponding taxa in their guts should be assessed for predispositions to disease. Findings reported here underscore the variability in gut microbiota composition across populations, and support ongoing efforts to characterise microbial diversity of the human microbiome.

Acknowledgements We thank staff and students involved in the HiKGuMMap project for coordinating collection, processing and maintaining inventory of samples, and Jin Yan Lim and Geicho Nakatsu for downloading and organising metagenome data and metadata.

Contributors YKY designed the study, analysed data and wrote the manuscript; ZC performed laboratory work; MCSC recruited subjects and edited the manuscript; MH revised the manuscript; JY, SCN and JJYS recruited subjects and acquired data; FKLC initiated the subject recruitment drive and provided funding; PSDK obtained funding, designed recruitment plan, recruited subjects, supervised the study and edited the manuscript.

Funding This study was supported by a seed fund for gut microbiota research provided by the Faculty of Medicine, The Chinese University of Hong Kong.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not required.

Ethics approval This study has been approved by the Joint Chinese University of Hong Kong-New Territories East Cluster Clinical Research Ethics Committee (reference number 2016.707). Written informed consent was obtained from all participants prior to collecting stool samples.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA557323. Gut metagenome and fusobacterial isolate genome sequence data are available in the Sequence Read Archive (SRA) under BioProject accession PRJNA557323.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

ORCID iDs
Yun Kit Yeoh http://orcid.org/0000-0003-0241-6117
Zigui Chen http://orcid.org/0000-0002-8577-1298
Martin C S Wong http://orcid.org/0000-0001-7030-9853
Jun Yu http://orcid.org/0000-0001-9239-2416
Siew C Ng http://orcid.org/0000-0002-6850-4454
Kian L Chan http://orcid.org/0000-0001-7388-2436

REFERENCES


2006


on 12 February 2020. Downloaded from http://gut.bmj.com/ by guest. Protected by copyright.


60 Parks DH, Chuvochina M, Chaumeil PA, et al. Selection of representative genomes for 24,706 bacterial and archaeal species clusters provide a complete genome-based taxonomy. bioRxiv 2019.


Southern Chinese populations harbour non-nucleatum Fusobacteria possessing homologues of the colorectal cancer-associated FadA virulence factor

Figure S1: Principal component analysis of gut microbiota composition in non-colorectal cancer (CRC) individuals from Hong Kong,[2,29] China,[30,31] USA,[3,27] Austria,[4] Denmark,[32] France, Germany,[5] Spain,[32], Israel,[33] Sweden,[34,35] El Salvador, Peru,[36] Fiji,[37] Mongolia,[38] and Tanzania [37,40]. Read mapping counts generated by MetaPhlAn2 were centered log ratio-transformed. Each circle represents community composition of one sample; the closer two circles are the more compositionally similar their microbial communities. The spokes connect circles to their respective centroids according to cohort (centroids indicated by position of cohort label).
Southern Chinese populations harbour non-nucleatum Fusobacteria possessing homologues of the colorectal cancer-associated FadA virulence factor

Figure S2: Average relative abundances of fusobacterial species detected in stool metagenomes of previously described CRC patients from various geographical backgrounds. Average relative abundances were calculated using MetaPhlAn2 on quality-filtered metagenome sequences, and values shown here for fusobacterial species are percentages of the total community. Each of the six stacked bars represent a study cohort: (i) Yu (Hong Kong), (ii) Coker (Hong Kong), (iii) Vogtmann (USA), (iv) Feng (Salzburg, Austria), (v) Zeller-f (France), and (vi) Zeller-g (Germany). Colours represent fusobacterial species detected in all metagenomes. Number of CRC subjects in each cohort are indicated in the cohort labels.
Southern Chinese populations harbour non-nucleatum Fusobacteria possessing homologues of the colorectal cancer-associated FadA virulence factor

Figure S3: Proportion of sequence reads in non-colorectal cancer gut metagenomes mapped to dereplicated fusobacterial genomes. Increasing proportions of mapped reads are represented by lighter colours. Phylogenetic tree on left of figure shows inferred evolutionary relationships between genomes. Branch colours delineate species boundaries and do not represent any taxa in particular.
Southern Chinese populations harbour non-nucleatum Fusobacteria possessing homologues of the colorectal cancer-associated FadA virulence factor

**Figure S4**: Proportions of sequence reads in colorectal cancer gut metagenomes mapped to dereplicated fusobacterial genomes. For each study location, the first and second columns indicate proportions from non-CRC controls and CRC cases, respectively. Increasing proportions of mapped reads are represented by lighter colours. Phylogenetic tree on left of figure shows inferred evolutionary relationships between genomes. Branch colours delineate species boundaries and do not represent any taxa in particular.
Southern Chinese populations harbour non-nucleatum Fusobacteria possessing homologues of the colorectal cancer-associated FadA virulence factor.

Figure S5: Pairwise average nucleotide identity (ANI) comparisons among dereplicated fusobacterial genomes. See Table S10 for ANI values and species boundaries of putatively novel genomes identified in this study. Increasing percentage ANI is represented by lighter colours. Phylogenetic tree on left and bottom of figure shows inferred evolutionary relationships between genomes. Branch colours delineate species boundaries and do not represent any taxa in particular. ANI comparisons were performed using FastANI.53
Southern Chinese populations harbour non-nucleatum Fusobacteria possessing homologues of the colorectal cancer-associated FadA virulence factor

Figure S6: Phylogenetic relationships Fap2 protein homologues identified in fusobacterial genomes. Figure shows a maximum-likelihood tree of aligned amino acid sequences of Fap2 homologues rooted at the midpoint. Homologues were identified using eggNOG-mapper. Each tip represents a homologue and is coloured according to species of the corresponding genomes they were found in. Background shading is according to the four major monophyletic clades identified in the genome-based phylogenetic tree in figure 2. Scale bars indicate amino acid substitutions per site. See Table S14 for output from eggNOG-mapper.