npj | biofilms and microbiomes

Article

Published in partnership with Nanyang Technological University

https://doi.org/10.1038/s41522-025-00826-3

Fusobacterium lineage profiling facilitates the clarification of the associations between non-*nucleatum Fusobacterium* and colorectal cancer

Yuli Wu[1,5], Guo Ji[1,5], Dongyan Han[1,5], Youhua Zhang[1], Xingchen Zhu[1], Hao Li[2], Man Li[1], Yaohui Gao[1], Ruting Xie[1], Min Xu[3], Ling Lu[1], Zixin Deng[3,4], Qing Wei[1] ✉, Huanlong Qin[2] ✉ & Dexi Bi[1] ✉

Non-*nucleatum Fusobacterium* may play a nonnegligible role in colorectal cancer (CRC) and certain *Fusobacterium* lineages (namely, L1 and L5) have shown specific associations with CRC. We aim to clarify the complex connections between *Fusobacterium* and CRC. We found that the widely adopted quantitative PCR (qPCR) method could overestimate *F. nucleatum* abundance and, in fact, reflect L1 levels in clinical samples. A lineage-specific qPCR assay targeting L1/L5 was developed and validated using mock and clinical samples. Its application in independent cohorts confirmed that L1 was overabundant in CRC, whereas L5 correlated with lymphovascular invasion. Importantly, faecal L1 abundance was more predictive of CRC than *F. nucleatum*, supported also by cross-population metagenomic data. CRC-associated virulence and colonisation genes were found in various L1 species other than *F. nucleatum*. Our results highlight the clinical importance of L1/L5 in CRC with high-diversity *Fusobacterium* contexts and suggest that non-*nucleatum Fusobacterium* may also contribute to CRC.

The bacterial genus *Fusobacterium* in the human gut microbiota is strongly associated with the development and progression of colorectal cancer (CRC)[1–5]. Among its species, *Fusobacterium nucleatum* is the most extensively studied and is enriched in CRC and contributes to pathogenesis via multiple mechanisms[6–8]. A recent study revealed that *F. nucleatum* subsp. *animalis* clade 2 (*Fna* C2) dominates the CRC niche[9]. However, increasing evidence also supports a nonnegligible role of non-*nucleatum Fusobacterium* in CRC, especially in populations with expanded *Fusobacterium* community diversity in the gut[10–15].

Fusobacterium includes diverse species[14,16,17]. *Fusobacterium* community diversity in the gut varies across populations. A metagenomic study revealed that distinct non-*nucleatum Fusobacterium* species bearing FadA virulence factor homologues exist in the gut microbiota of the southern Chinese population but are rare in other populations[10]. A 16S rRNA gene sequencing-based study roughly classified *Fusobacterium* into four lineages on the basis of V4 polymorphisms and confirmed the predominance of non-*nucleatum Fusobacterium* species in the gut of the same population, although this approach did not provide species-level resolution[11]. Interestingly, that study revealed that *Fusobacterium* lineages exhibit distinct correlations with host diseases[11]. Consistently, genomic analysis has also categorised *Fusobacterium* species into certain lineages with different sets of virulence-associated genetic determinants[17]. We recently developed a *rpoB*-based sequencing technique (FrpoB-seq) that enables species-resolved, precise quantification of *Fusobacterium* and revealed surprisingly diverse compositional patterns at the species level in the gut microbiota of the Chinese population[14]. We have further established that *Fusobacterium* has nine genetic lineages (L1--L9), four of which (L1, L4, L5, and L9) are frequently detected in the gut microbiota, but only two (L1 and L5) displayed distinct associations with CRC[14]. In particular, L1, consisting of *F. nucleatum*, *Fusobacterium periodonticum*, *Fusobacterium pseudoperiodonticum*, *Fusobacterium hwasookii* and many closely related species, was enriched in CRC, whereas L5, represented by *Fusobacterium varium* and *Fusobacterium*

[1]Department of Pathology, Shanghai Tenth People's Hospital, Tongji University School of Medicine, Shanghai, China. [2]Department of Gastrointestinal Surgery, Shanghai Tenth People's Hospital, Tongji University School of Medicine, Shanghai, China. [3]Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin, China. [4]State Key Laboratory of Microbial Metabolism, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China. [5]These authors contributed equally: Yuli Wu, Guo Ji, Dongyan Han. ✉e-mail: weiqing1971@tongji.edu.cn; qinhuanlong@tongji.edu.cn; bidexi@tongji.edu.cn

ulcerans, was associated with lymphovascular invasion (LVI)[14]. Therefore, in a microbiota background containing diverse *Fusobacterium*, it is very likely that multiple *Fusobacterium* species are involved in CRC by exerting pathogenicity in a lineage-dependent manner.

Moreover, molecular genetic studies have confirmed that the four subspecies of *F. nucleatum* are separate species[14,17–19], which has been endorsed by the NCBI Taxonomy Database (https://www.ncbi.nlm.nih.gov/taxonomy/). The former subsp. *animalis*, *polymorphum*, *vincentii* and *nucleatum* are now renamed *F. animalis*, *F. polymorphum*, *F. vincentii* and *F. nucleatum*, respectively[18,19]. Thus, the formerly called '*F. nucleatum*' is in fact a subgroup of closely related species, again highlighting that it is not just one single species engaging in CRC. However, given that the updated taxonomy has not yet been generally recognised, we continue using the term '*F. nucleatum*' to refer to the collection of its former members in this study. It should be noted that the updated taxonomy does not invalidate previous '*F. nucleatum*'-centred findings but highlights the need to better interpret those important results within the updated framework. In addition, to date, targeted quantification of '*F. nucleatum*' rely mostly on a primer/probe set (Fn-F/R) targeting the *nusG* gene[20,21]. However, we previously revealed that the target regions of the primers and the cognate probe were not specific, appearing as well in *F. periodonticum*, *F. pseudoperiodonticum* and *F. hwasookii*[15], which are also L1 members frequently observed in the gut[14]. Hence, there is also a need to revisit the association between '*F. nucleatum*' and CRC.

Given that the clinical detection of CRC-associated *Fusobacterium* lineages may be more meaningful than just one species, this study designed a PCR-based method for accurate detection of L1 and L5 and validated their clinicopathological significance in CRC, which allows a revisit of the associations between '*F. nucleatum*' and CRC in comparison with L1. This study helps clarify the intricate associations between *Fusobacterium* and CRC and also provides a novel approach to investigate the role of non-*nucleatum* *Fusobacterium* in CRC.

Results

The widely used primers did not accurately detect the abundance of '*F. nucleatum*' in the gut microbiota with expanded *Fusobacterium* community diversity

To examine whether the widely used primers Fn-F/R detected the actual level of '*F. nucleatum*', we performed qPCR on clinical samples (Supplementary Data 1) that had previously undergone FrpoB-seq[14], including tumour and matched normal tissues from CRC patients as well as faeces from CRC patients and healthy controls (Fig. 1A, B). Interestingly, in both types of samples, the qPCR results revealed only moderate correlations (Spearman's *rho* = 0.617 and =0.591 in the faecal and tissue samples, respectively) with the relative abundance of '*F. nucleatum*' obtained by FrpoB-seq, but in contrast, exhibited strong correlations (Spearman's *rho* = 0.843 and =0.863 in faecal and tissue samples, respectively) with L1 abundance, which is consistent with the fact that the nonspecific targets of Fn-F/R in L1 members frequently occur in the human gut microbiota. To substantiate the cross reactivity of Fn-F/R with other L1 species, we constructed synthetic L1-community samples containing different ratios of *F. nucleatum* ATCC 25586 and other L1 strains *F. periodonticum* THCT14E2 and *F. hwasookii* THCT18E1, previously isolated from CRC patients. Strikingly, qPCR with Fn-F/R failed to distinguish the *F. nucleatum* percentage gradience in mixed communities but instead correlated with total L1 bacterial loads of each sample, compared with *F. nucleatum*-alone controls (Fig. 1C). The quantitative performance with Fn-F/R on samples containing the non-*nucleatum* strains (THCT14E2 and THCT14E2) was near-identical to that on *F. nucleatum* alone, except

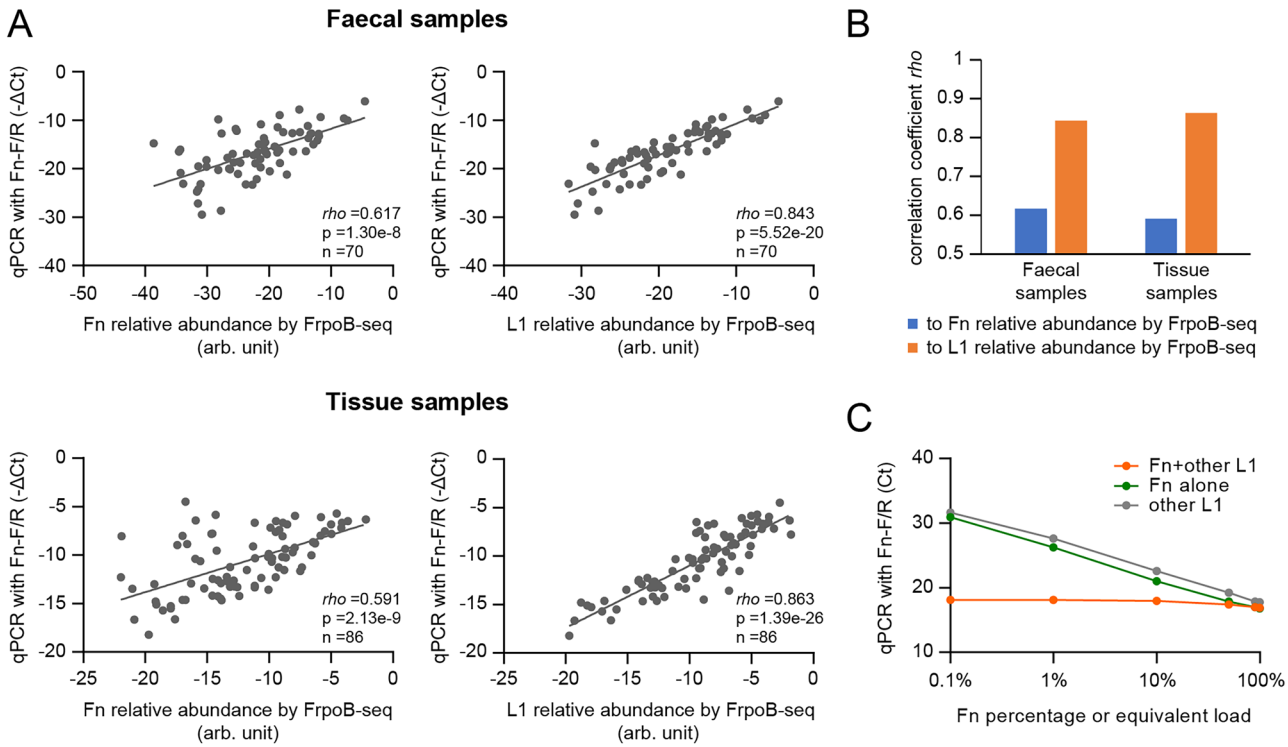


Fig. 1 | The widely used primer set Fn-F/R did not accurately detect the abundance of '*F. nucleatum*'. A Correlations between the qPCR results with primer set Fn-F/R and FrpoB-seq results in clinical samples. Samples with '*F. nucleatum*' (Fn) detected by FrpoB-seq were subjected to qPCR with Fn-F/R. The faecal samples included 46 from CRC patients and 24 from healthy controls. The tissue samples included 45 tumour samples and 41 normal samples. The Spearman correlation test was applied. **B** The Spearman correlation coefficients in **A** summarised in a histogram graph. **C** The qPCR results with primer set Fn-F/R in L1-community samples containing different percentages (0.1%, 1%, 10%, 50%, 90%, 100%) of *F. nucleatum* (Fn; ATCC 25586) and other L1 species *F. periodonticum* (THCT14E2) and *F. hwasookii* (THCT18E1). Note that the L1-community samples were designed to have an equal total CFU (10^6). The qPCR was meanwhile conducted with '*F. nucleatum*' alone or other L1, loaded equivalently to the '*F. nucleatum*' amounts in the L1-mixture samples.

Table 1 | The identified L1-, L5- and '*F. nucleatum*'-specific markers and corresponding primers designed

Taxon	Marker	Reference gene	Annotation	Primer	sequence (5' to 3')	Amplicon size (bp)
L1	L1_125	*C7Y58_06055*	hypothetical protein	L1_125-F1	RGTRTTGGAGTTGTTCAATATGG	93
				L1_125-R1	TTCMGATAARGAATTAACTTCATTTTC	
				L1_125-F2	ATGAARYTAACATTACAACAAGCTATATT	103
				L1_125-R2	TTTAAAGGYACTACATAAYTRTCACGAAT	
				L1_125-F3	ATTCGTGAYARTTATGTAGTRCCTTTAAA	293
				L1_125-R3	CCATATTGAACAACTCCAAYACY	
	L1_746	*C7Y58_07255*	DUF4037 domain-containing protein	L1_746-F	YTAAATGARTTRGGWGGAACTTTAAA	163
				L1_746-R	TTTGTRCAAATCTRTATRCYTCAGTTA	
	L1_820	*C7Y58_01415*	hypothetical protein	L1_820-F	GTGCAACAACWATHTCWTAYTGGTATGG	139
				L1_820-R	TAYTTYTCWACATARGCYTTTCCCAT	
L5	L5_211	*C4N18_13790*	hypothetical protein	L5_211-F	TTTRAATAATAAAATTGGAGTGAGA	285
				L5_211-R	TCTTCTGCTAGTGTTCTAGGAAGAA	
	L5_962	*C4N18_14645*	PaaI family thioesterase	L5_962-F	TTATYACTYTAGCTGATACTACATGTGG	153
				L5_962-R	CAAGATGCTGAGCAGTRGCAGTACA	
	L5_1277	*C4N18_13335*	hypothetical protein	L5_1277-F	GAGGTATGGGATAGAGAACAGCCTG	113
				L5_1277-R	TCATTAATRTTYTTTACATCTTCTGGTGTT	
	L5_1430	*C4N18_12565*	hypothetical protein	L5_1430-F	TTTCAYTACCTTGAAGAAGGAAA	103
				L5_1430-R	TTTCTTTATCTTTTCCTGTAAGCATAA	
	L5_1621	*C4N18_10820*	hypothetical protein	L5_1621-F	GAGAATCAATTCAGAGASTACAMATTCTAT	143
				L5_1621-R	AATAAAGRTTRACAGTATCCATCACTGTA	
'*F. nucleatum*'	Fn_SR	*C7Y58_05970*	peptidase	Fn_SR-F	YTATCCTCAATCAAGAGTTACAA	348
				Fn_SR-Ra	AGRATTTTAYCRGTATTAGCATCRAT	
				Fn_SR-Rb	AARATTTTWCCASTATTAGCATCAAT	
				Fn_SR-Rc	ATAACTTTCCCACTATTAGCATCAAT	

the marginally elevated Ct values attributable to imperfect primer-template complementarity in those species (Fig. 1C). Thus, the results of qPCR with Fn-F/R reflected the total abundance of major L1 members and therefore might overestimate the abundance of '*F. nucleatum*,' particularly in samples containing other L1 members. These results highlight the necessity of detecting *Fusobacterium* lineages that are actually associated with CRC.

Identification of L1- and L5-specific markers via comparative genomic analysis

We sought to develop a PCR-based method that can be easily used to detect L1 and L5 in clinical samples. Comparative genomic analysis of 157 unique, publicly available *Fusobacterium* genomes with lineages defined previously (Supplementary Data 2)[14] was conducted to screen for L1- and L5-specific markers. There were 94 genomes of ten species that belong to L1; eleven genomes of four species belonging to L5; 50 genomes of eight species belonging to L2, L3, L4, L8 or L9; and two *F. naviforme* genomes. *F. naviforme* has no lineage designation, as its classification as a *Fusobacterium* species remains questionable and has been proposed for reclassification into a new genus[14,16,22]. Although L6 and L7 have no genome available and are unable to be included, these two lineages rarely occur in the human gut[14]. Genes that were highly conserved within one lineage but completely absent in other lineages and non-*Fusobacterium* organisms were selected as lineage-specific markers. Three and five markers exhibiting strict specificity were identified for L1 and L5, respectively (Table 1). These marker genes were shared by all members of L1 or L5 with 80%–100% nucleotide identities to reference sequences but were not present in any of the genomes of the other included lineages (Fig. 2). Homologous sequences were also not found in other organisms via examination of the NCBI nucleotide database.

Development of a PCR-based assay for *Fusobacterium* lineages L1 and L5 detection

Universal primers were then designed for the above markers (Table 1). Putative nonspecific amplification was not found via in silico analysis in the NCBI nucleotide database. Primer performance was then experimentally assessed in the available strains (Supplementary Table 1)[14] of eight *Fusobacterium* species belonging to L1, L4 or L5. The primer sets L1_746-F/R and L5_1621-F/R presented the best specificity for L1 and L5, respectively, with amplification results that were completely consistent with the lineage classification (Fig. 3A). An optimised PCR condition was also set up by assessing different primer concentrations, template inputs, and thermal cycling parameters (see 'Materials and methods' and Supplementary Fig. 1). The two primer sets were subsequently tested in human host (represented by whole-blood DNA) and real-world *Fusobacterium*-free gut microbiota backgrounds, and nonspecific amplification was not detected (Fig. 3B). They were further evaluated with mock samples for feasibility of qPCR. For samples containing defined colony-forming units (CFUs) of an L1 or L5 reference strain, the corresponding quantification results displayed good correlation with the bacterial loads, as linearity was observed between Ct values and log10-transformed CFU values (Fig. 3C). The correlation was stably maintained for a given culture sample of the reference strain and its tenfold serial dilutions and when they were mixed with DNA of the *Fusobacterium*-free gut microbiota and reference strains of other lineages (Fig. 3D). The L1_746-F/R also performed well with the L1-community samples mentioned above (Supplementary Fig. 2). Finally, we examined the quantification accuracy of the two primer sets in the clinical samples mentioned above (Supplementary Data 1)[14]. Strikingly, the L1 and L5 levels determined by qPCR were highly consistent with the FrpoB-seq results (Fig. 3E, F and Supplementary Fig. 3). Meanwhile, qPCR melt curve analysis further demonstrated the primer specificity (Supplementary Fig. 4). With such

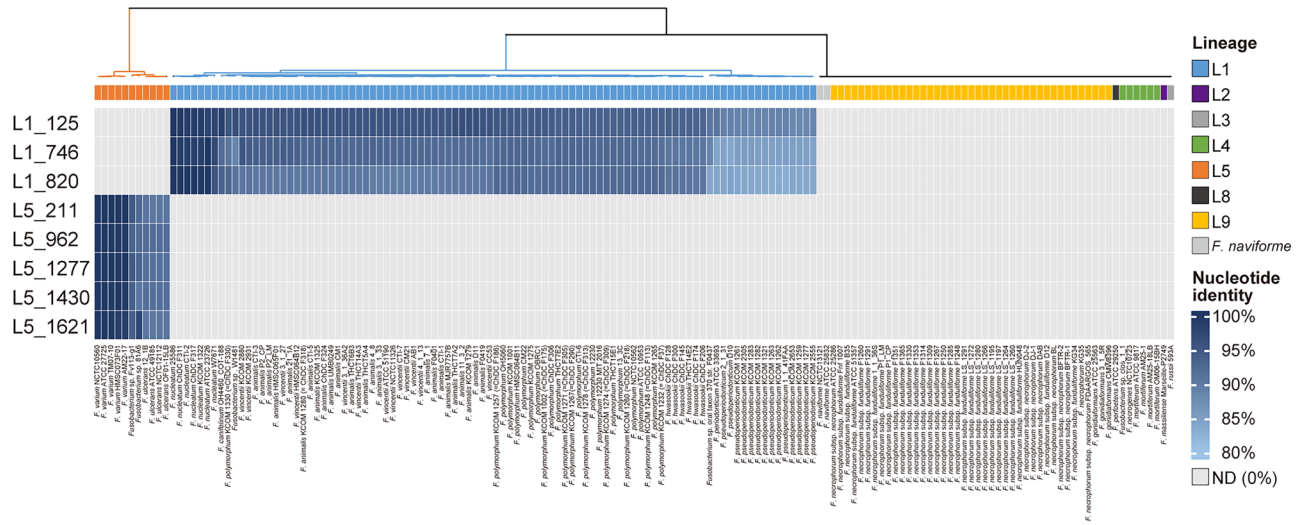


Fig. 2 | Distribution of the identified L1- or L5-specific markers in *Fusobacterium* genomes. The presence and nucleotide identities to reference sequences of the markers are visualised via a heatmap. The marker identifiers are listed on the left of the heatmap. The strain names are listed under the heatmap. See Table 1 for information of the reference sequences. ND not detected.

efforts, we developed a PCR-based quantitative assay termed FL-typing (*Fusobacterium* lineage typing) for L1 and L5 detection. Notably, the qPCR results for Fn-F/R were strongly correlated (Spearman's *rho* = 0.950 and =0.932 in the faecal and tissue samples, respectively) with the L1 levels detected via qPCR (Supplementary Figs. 3 and 5), indicating that Fn-F/R indeed detected L1 members rather than '*F. nucleatum*' alone. The minimal CFUs required for reliable quantification were currently determined to be 50 CFUs per reaction (Supplementary Table 2) for FL-typing. The sensitivity and specificity of FL-typing in qualitative detention were 98.2% and 100%, respectively for L1; and 96.4% and 96.6%, respectively for L5 (Supplementary Table 3).

Validation of L1- and L5-specific correlations in CRC with FL-typing

We intended to validate the previously reported L1- and L5-specific correlations in CRC with FL-typing in newly collected faecal and tissue samples (Fig. 4). We collected faecal samples from 60 CRC patients and 53 healthy controls (Supplementary Data 3). We also collected tumour and adjacent normal tissue samples from 100 CRC patients (Supplementary Data 4). We found that the faecal L1 levels in CRC patients were significantly greater than those in healthy controls (Fig. 4A). Similarly, the L1 levels in CRC tumours were also higher than those in adjacent normal tissues (Fig. 4B). Moreover, the L5 levels in CRC tumours with LVI were significantly greater than those in those without LVI (Fig. 4C). Thus, with this method, we confirmed that L1 is enriched in CRC, whereas high intratumour L5 levels correlate with LVI.

The abundance of L1 was more predictive of CRC than that of '*F. nucleatum*'

We also compared the predative values for the CRC between L1 and '*F. nucleatum*'. We developed a new method to detect '*F. nucleatum*' similarly as above. Briefly, a genetic marker (Fn_SR) highly specific to '*F. nucleatum*' on the basis of comparative analysis was identified with a new primer set designed for this purpose (Supplementary Fig. 6A and Table 1). PCR revealed that the primer set had no nonspecific amplification in strains of other *Fusobacterium* species (Supplementary Fig. 6B). The primer set could precisely distinguish *F. nucleatum* percentage gradience in the L1-community samples (Supplementary Fig. 2), and the qPCR results in clinical samples strongly correlated with the FrpoB-seq-detected '*F. nucleatum*' abundance (Supplementary Figs. 3 and 6C). Upon qPCR detection, receiver operating characteristic (ROC) curves and precision-recall (PR) curves of L1, '*F. nucleatum*' and Fn-F/R results for CRC prediction were generated for both previously and newly collected faecal samples (cohorts F1 and F2, respectively) from CRC patients and healthy controls (Fig. 4D, E and Supplementary Table 4). In both cohorts, the area under curves (AUCs) of ROC or PR of L1 were greater than that of '*F. nucleatum*'. It was also numerically larger than that with Fn-F/R, but the difference was not statistically significant, likely because of the small sample size. The findings were further validated in publicly available faecal metagenomic datasets from eight independent geographically distinct cohorts, including CRC patients and healthy controls across diverse populations in China (Shanghai[23] and Hong Kong[24]), Japan[25], Germany[26], France[27], the United States (USA)[28], Austria[29], and Italy[30] (Fig. 4F; Supplementary Fig. 7 and Supplementary Table 4). Remarkably, the L1 consistently demonstrated better predictive performance than '*F. nucleatum*' in the Shanghai (China), Hong Kong (China), Japan, Germany, France and USA cohorts. But variations in AUCs of ROCs or PRs across cohorts were observed, suggesting that the predictive performance could be population-specific. These discrepancies may be attributable to interpopulation variations in *Fusobacterium* prevalence and the compositional complexity of L1 subcommunity. In cohorts where L1 and '*F. nucleatum*' showed comparable predictive power (AUCs not significantly different), this phenomenon likely reflects simplified L1 community structures with '*F. nucleatum*' dominance within those populations.

Virulence- and colonisation-associated genetic determinants are shared by L1 species other than '*F. nucleatum*'

To explore potential conserved genetic mechanisms underlying the lineage-specific CRC association of L1, the distribution of virulence- and colonisation-associated genetic determinants originally characterised in '*F. nucleatum*' were examined in L1 species. Certain virulence factors and the *eut* and *pdu* operons have been suggested to contribute to CRC niche colonisation and pathogenicity of *F. nucleatum* subsp. *animalis* clade 2 (*Fna* C2)[9]. In particular, the two operons were found to be specific to *Fna* C2 but absent in *Fna* C1[9]. Here, the presence of those genetic determinants was examined in a wider taxonomic context, including L1 species other than '*F. nucleatum*'. Notably, we found that these genes were also shared by diverse L1 species (Fig. 5A). The canonical virulence genes *fadA*, *fadA3*, *fap2*, *fplA*, *cmpA* and *fusolisin* were found to be widespread across L1 species, whereas *fadA2*, *radD* and *aim1* were present in certain strains of non-*nucleatum* L1 species. It has been suggested that *fap2*, *cmpA* and *fusolisin* are responsible for promoting gastrointestinal niche adaptation, whereas *fadA2*, *radD* and *aim1* are related to adaptation in the oral niche[9]. The *eut* and *pdu*

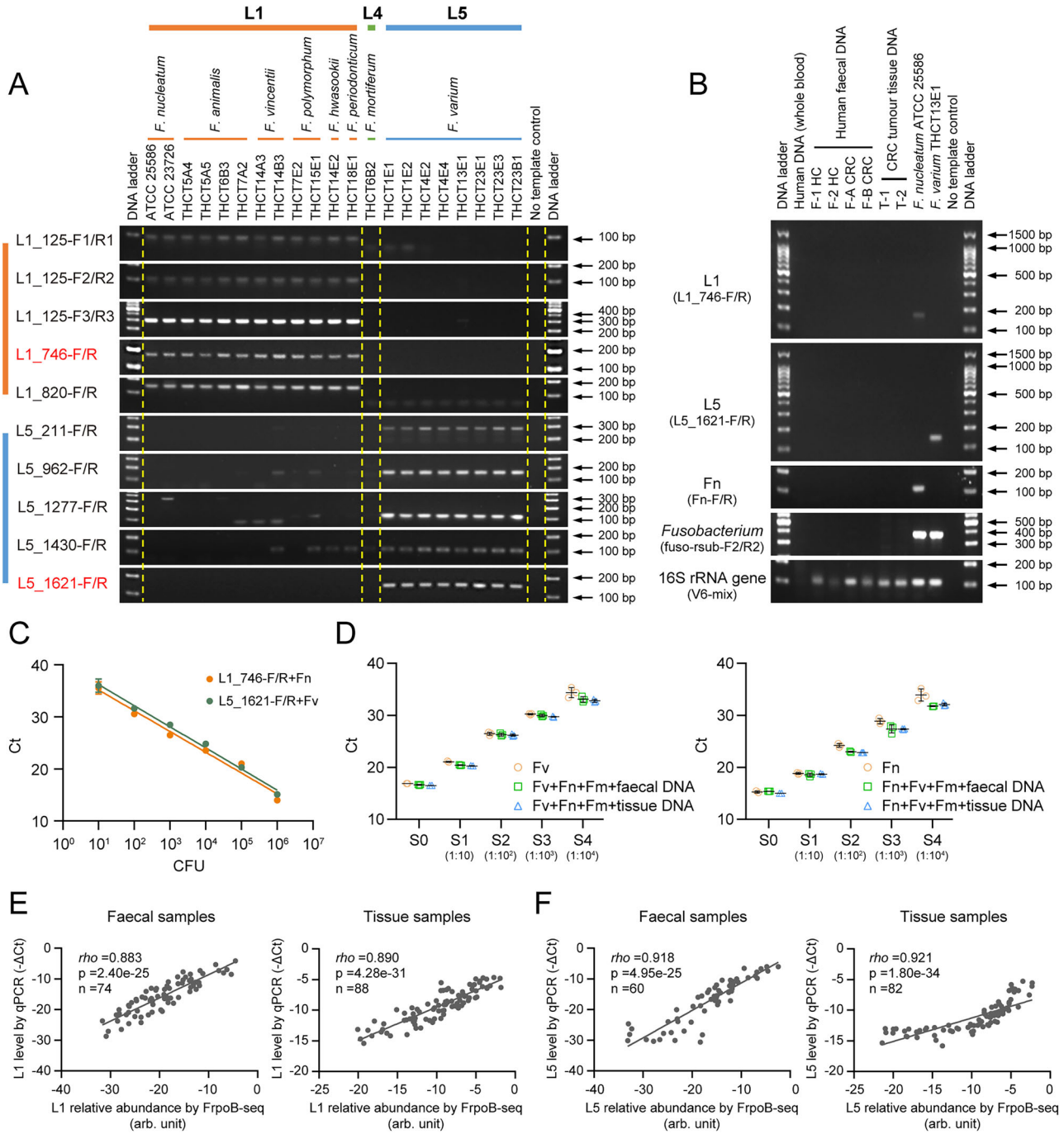
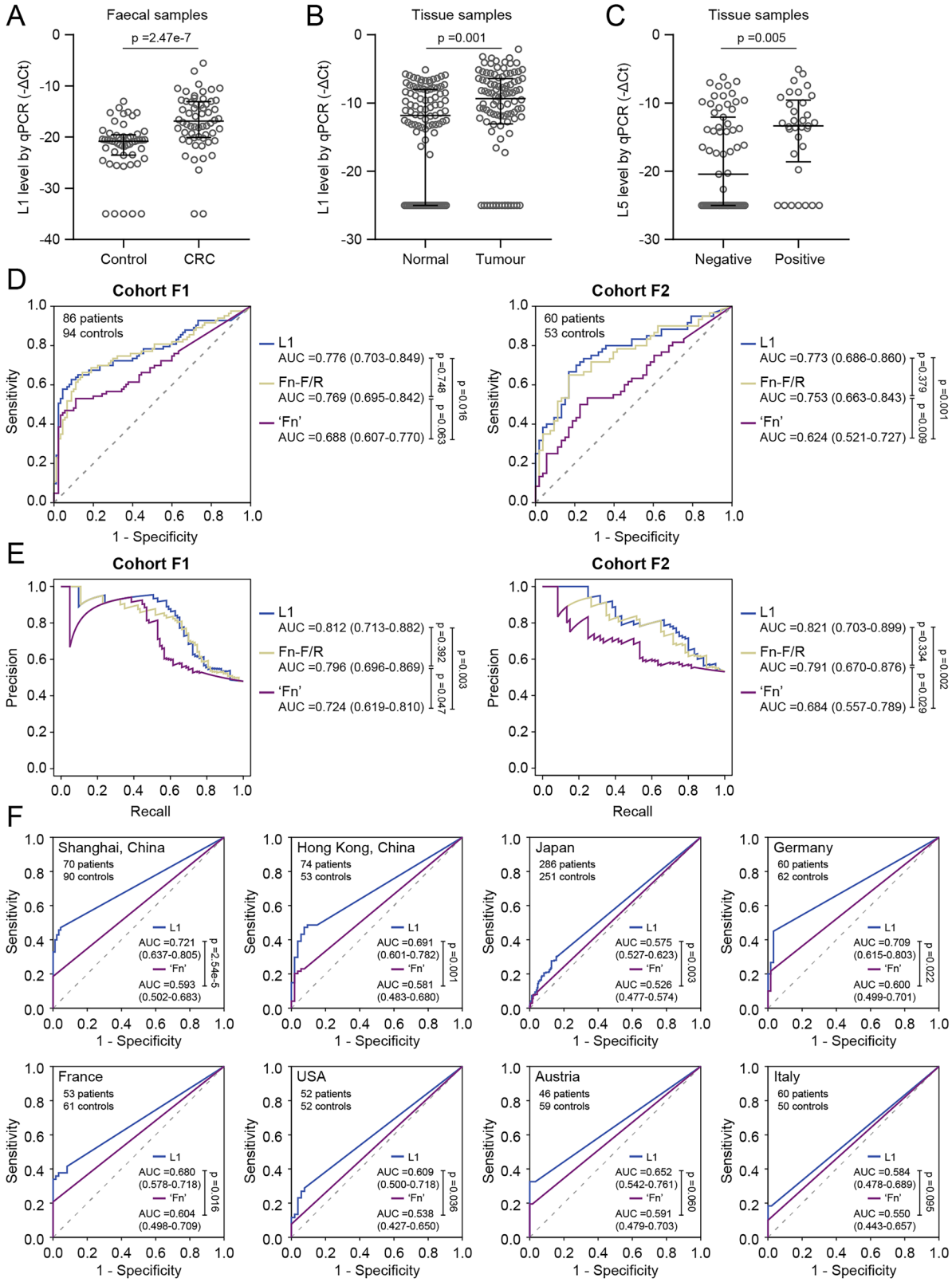


Fig. 3 | Development of a PCR-based assay for *Fusobacterium* lineages L1 and L5 detection. A Amplification performance of the designed primers in *Fusobacterium* strains of different lineages. The primer sets selected for further analyses are marked in red. **B** The selected primer sets did not amplify DNA from human whole blood or *Fusobacterium*-free faecal or tissue samples. The presence of bacteria but the absence of *Fusobacterium* was confirmed by amplification with the primer sets V6-mix and fuso-rsub-F2/R2, respectively. *F. nucleatum* ATCC 25586 and *F. varium* THCT13E1 were used as positive controls. HC healthy control, CRC colorectal cancer. **C** qPCR with the selected primer sets in the corresponding L1 (*F. nucleatum* ATCC 25586, Fn) and L5 (*F. varium* THCT13E1, Fv) reference strains loaded with defined CFUs (colony-forming units). **D** qPCR with the selected primer sets in a given culture of Fn or Fv and its tenfold serial dilutions, with or without DNA from a *Fusobacterium*-free faeces/tissue sample and strains of other lineages. Fm, *F. mortiferum* THCT6B2. NC, negative control. For **C**, **D** the experiments were performed in triplicate, and the data are expressed as the means and standard deviations. Spearman correlation between qPCR- and FrpoB-seq-obtained L1 (**E**) or L5 (**F**) levels in clinical samples. Samples with L1 or L5 detected by FrpoB-seq were subjected to qPCR with L1_746-F/R or L5_1621-F/R, respectively. For (**E**), the faecal samples were from 50 CRC patients and 24 healthy controls, while the tissue samples included 45 tumour samples and 43 normal samples. For (**F**), the faecal samples were from 32 CRC patients and 28 healthy controls, while the tissue samples included 43 tumour samples and 39 normal samples.

operons, which promote gastrointestinal niche adaptation by enhancing metabolic capabilities[9], were also found in various L1 species in addition to *F. animalis*. The *eut* operon was well conserved across most L1 species, showing high synteny to that of *Fna* C2 (Fig. 5B). Syntenic *pdu* operons were also found in *F. polymorphum*, *F. canifelinum*, *F. hwasookii* and a strain (KCOM 2305) of *F. pseudoperiodonticum* but were absent in the other studied L1 species (Fig. 5C). Therefore, CRC niche adaptation-associated genetic determinants are also shared by diverse L1 species.



Distinct co-occurrence patterns of L1 and L5 with other bacteria

Given the key role of *Fusobacterium*, particularly '*F. nucleatum*', in organising biofilms[31], which are also observed in CRC[32], we investigated the microbial co-occurrence network associated with L1 and L5 using available metagenomic sequencing data from cohort F1. Distinct co-occurrence patterns of L1 and L5 with other bacteria were revealed (Fig. 6 and Supplementary Data 5). L1 exhibited positive correlations with 49 species and negative correlations with eight species, whereas L5 showed positive correlations with 40 species and negative correlations with 59 species. Only two species (*Streptococcus pyogenes* and *Gemella sanguinis*) correlated with both L1 and L5 to varying degrees within the network. Interestingly, L1 correlated preferentially with biofilm-associated taxa[31,33], including

Fig. 4 | Validation of L1- and L5-specific correlations in newly collected clinical CRC samples with FL-typing. A L1 levels in faecal samples compared between CRC patients ($n = 60$) and healthy controls ($n = 53$). Mann–Whitney test. **B** L1 levels were compared between tumour and matched normal tissues from 100 CRC patients. The p value of the Mann–Whitney test is shown, whereas that of the Wilcoxon matched-pairs signed-rank test is 1.86e-11. **C** L5 levels in CRC tumours compared between lymphovascular invasion (LVI)-positive ($n = 33$) and LVI-negative tumours ($n = 67$). Mann–Whitney test. Medians and interquartile ranges are shown. If not detected, −ΔCt values of -35 and -25 were manually defined for the faecal and tissue samples, respectively. **D** ROC curves of L1, '*F. nucleatum*' and Fn-F/R for the previously and newly collected faecal samples (cohorts F1 and F2, respectively). ROC curves were compared with the DeLong method, Benjamini–Hochberg correction was additionally applied, and adjusted p values are shown. **E** PR curves of L1, '*F. nucleatum*' and Fn-F/R for cohorts F1 and F2. PR curves were compared with the bootstrap method followed by Benjamini–Hochberg correction, and adjusted p values are shown. **F** ROC curves of L1 and '*F. nucleatum*' for eight published faecal metagenomic datasets across diverse populations. L1 and '*F. nucleatum*' abundance was represented by the FPKM values of their specific makers L1_746 and Fn_SR, respectively. The ROC curves were compared with the DeLong method. The corresponding PR curves are provided in Supplementary Fig. 7. For (**D**, **E**), ROC comparison with the bootstrap method was also conducted and the p or adjusted p values are listed in Supplementary Table 4. All AUCs are reported with their 95% confidence intervals in the parentheses.

Streptococcus spp., *Campylobacter* spp., *Gemella* spp. and certain non-L1/L5 *Fusobacterium* species. Notably, *Streptococcus* and *Gemella* are core members of the oral microbiota and the coaggregation between *Streptococcus* and '*F. nucleatum*' represents a critical step in oral biofilm formation[31]. In contrast, L5 correlated with gut commensals such as *Lachnoclostridium* spp., *Bacteroides* spp., *Clostridium* spp. and *Ruminococcus* spp. While most L5-associated species spanned broad taxonomic groups, several belonged specifically to the *Bacteroides* genus.

Discussion

In this study, we uncovered that the widely used primers did not accurately detect '*F. nucleatum*' but rather indicated the abundance of L1, and further provided the FL-typing method that could precisely detect L1 and L5. We confirmed that L1 and L5 had specific correlations with CRC and that L1 exhibited better predictive value than '*F. nucleatum*' for CRC.

This study highlights the clinical importance of *Fusobacterium* lineages L1 and L5 in CRC. We previously revealed L1- and L5-specific correlations in CRC via FrpoB-seq, but the method could not be implemented in all samples, as it required successful sequencing library construction[14]. We herein developed a PCR-based L1/L5 detection method (FL-typing) that possesses general applicability. Its application in clinical samples confirmed that L1 was enriched in the gut microbiota of CRC patients, whereas the L5 abundance in CRC tumours was positively associated with LVI. The major L1 member '*F. nucleatum*' has been previously recognised to play an important role in promoting CRC[6–8]. However, we showed that qPCR with the primer set (Fn-F/R) adopted by most studies in fact estimated the abundance of L1 rather than '*F. nucleatum*,' as the results strongly correlated with L1 levels, and the corresponding AUCs of ROCs or PRs for CRC prediction were comparable to those of L1, which is in line with the fact that the target region was not confined to '*F. nucleatum*' but is also conserved in other species belonging to L1[15]. Thus, the previously recognised clinicopathological relevance of '*F. nucleatum*' in CRC could be overestimated and was very likely attributed to L1. Indeed, we repeatedly observed that the actual levels of '*F. nucleatum*' were less predictive for CRC than L1[14]. Therefore, our study strongly suggests the need to revisit the association between '*F. nucleatum*' and CRC and the less characterised species should be evaluated for pathogenicity. However, these results do not contradict those of CRC-related experimental studies that focused on '*F. nucleatum*'. As a major member of L1, it remains an important model organism, and the uncovered pathogenicity may represent features of this lineage. Notably, the current AUCs of ROCs or PRs of L1 seemed relatively modest, which reflect the complex multifactorial nature of CRC pathogenesis. These results align with our understanding that CRC development involves intricate host-microbe interactions, where L1 represents only one component of the microbial landscape. Future diagnostic approaches should consider strategies that combine L1 detection with complementary biomarkers. Targeting experimentally verified virulence genes and coupling with additional biomarkers such as co-occurring microbes and serum or faecal metabolite signatures[23] could be implemented to improve diagnostic precision. The L5, however, is not a considered as a marker for CRC prediction, but rather a factor promoting the disease course, as it is not overabundant in the gut microbiota of CRC patients[14]. The other *Fusobacterium* lineages were not covered in this study, as their clinical significance in CRC has not been observed[14].

A recent study revealed that *Fna* C2 is predominant among '*F. nucleatum*' in CRC niches[9]. Our results additionally indicate that the less prevalent *Fusobacterium* members should not be neglected. With shared genetic determinants responsible for virulence and niche adaptation, closely related *Fusobacterium* members rather than a single species may collectively contribute to pathogenesis. Notably, this scenario might be population-specific and common in populations with expanded gut *Fusobacterium* community diversity. Indeed, the studied subjects in this study all belonged to the Chinese population, which has an increased prevalence of non-*nucleatum Fusobacterium* and diverse *Fusobacterium* community compositions in the gut microbiota[10,11,14]. Validation using metagenomic datasets further demonstrated that this phenomenon did not occur in all examined populations. Therefore, population differences may be considered for better probing of the gut microbes contributing to diseases as diagnostic targets.

The pathogenic mechanisms underlying L1 and L5 in CRC remain incompletely understood. However, evidence suggests that their distinct associations with CRC pathogenesis may stem from their divergent repertoires of lineage-specific virulence genes. Comparative genomic study analyses have revealed that the distribution of virulence genes across *Fusobacterium* species is diverse but exhibit a lineage-specific pattern[17]. Intriguingly, the L1 and L5 correspond to evolutionarily distinct 'active invader' subgroups with unique genetic signatures[17], implying that their differential pathogenicity in CRC may arise from mechanistically distinct virulence pathways. Furthermore, our metagenomic co-occurrence analysis of faecal samples revealed distinct ecological networks associated with L1 and L5. Critically, these patterns align with previous observations in CRC tumour microbiomes[12], wherein '*F. nucleatum*' (the major L1 member) resided within a co-occurrence cluster featuring oral bacteria like *Streptococcus* and *Gemella*, while *F. varium* (the major L5 member) associates with a distinct cluster containing multiple *Bacteroides* taxa. These findings also provide mechanistic insights into their differential pathogenicity in CRC. The L1-centric network supports its capacity to form invasive biofilms in CRC alongside disease-associated microbes[32], especially the oral bacteria. The correlations between L5 and *Bacteroides* species imply an alternative ecological strategy, likely involving mucin degradation facilitated by *Bacteroides*[34,35], that shapes its unique pathogenic niche[36]. To fully elucidate these mechanisms, systematic investigations could be conducted, including comprehensive comparative genomic analyses followed by functional validation through targeted mutagenesis and phenotypic characterisation of candidate virulence determinants. The lineage-specific markers identified in this study mostly encode proteins with uncharacterised functions. It is worthy to explore whether they play a role in CRC pathogenesis. It should also be noted that there are other lineage-specific, but less conserved genes in L1 and L5, including several putative virulence genes. While these genes did not satisfy the stringent criteria for marker selection, they may nevertheless encode critical pathogenic functions. Their exclusion from marker panels should not preclude further investigation into their potential contributions to CRC progression.

This study employed a stringent strategy to develop and maximise the accuracy of the FL-typing method. We leveraged exhaustive comparative

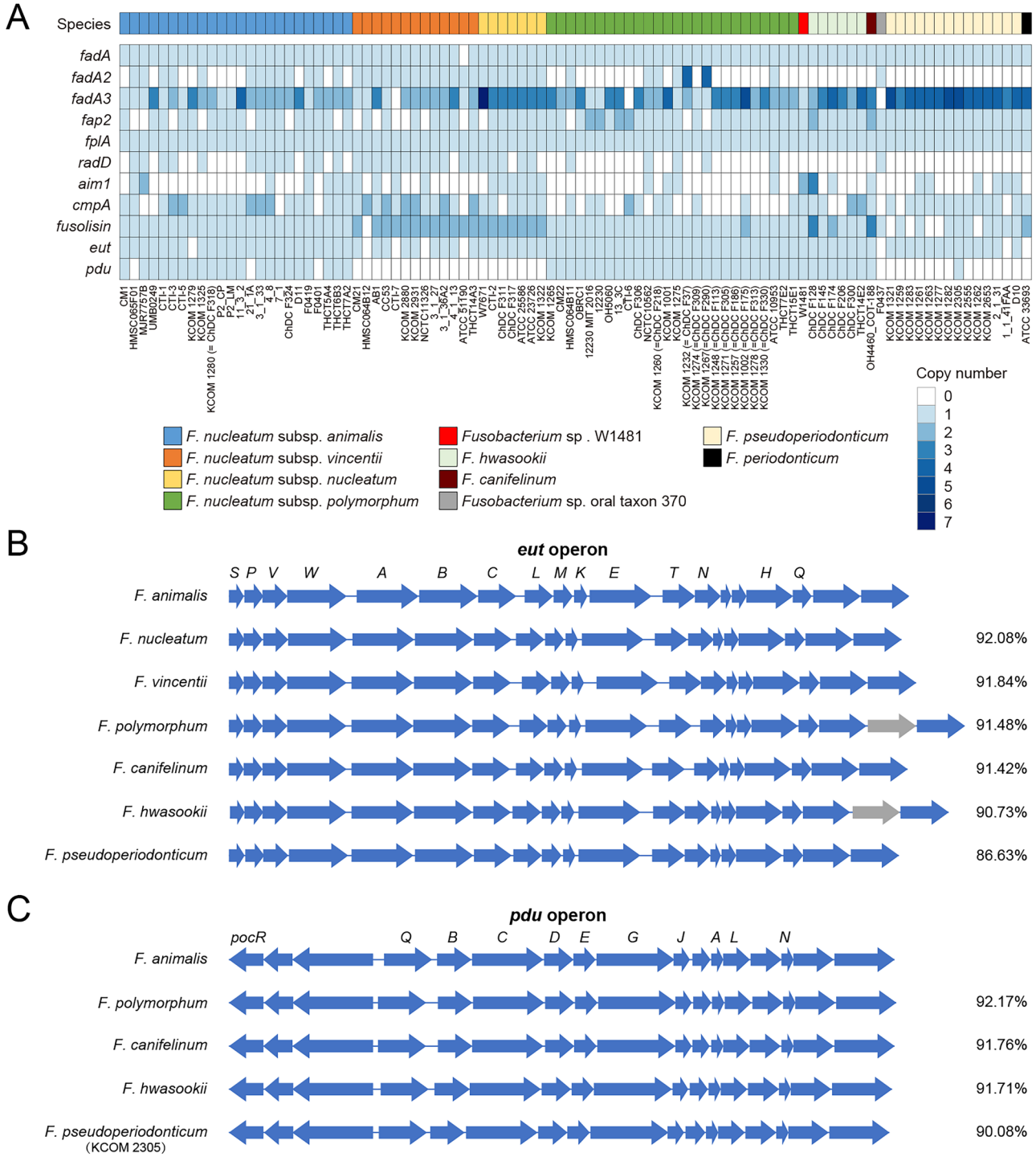


Fig. 5 | Virulence gene distribution across L1 species. A A heatmap illustrating the distribution of known '*F. nucleatum*' virulence factors across L1 species. Gene/operon copy numbers are visualised by colour gradience. **B** The *eut* operon is conserved in L1 species. Representative gene clusters are drawn for the strains SB010 for *F. animalis* (reference), KCOM 1261 for *F. pseudoperiodonticum*, ChDC F174 for *F. hwasookii*, OH4460_COT-188 for *F. canifelinum*, ATCC 25586 for *F. nucleatum*, THCT14A3 for *F. vincentii* and THCT15E1 for *F. polymorphum*. Genes with no counterparts in the reference operon are marked in grey. **C** The *eut* operons exist in *various* L1 species. Gene clusters are drawn based on strains SB010 for *F. animalis* (reference), OH4460_COT-188 for *F. canifelinum*, THCT15E1 for *F. polymorphum*, and ChDC F174 for *F. hwasookii*. Nucleotide identities to the reference operon are given.

genomic analysis to obtain highly specific lineage markers and designed universal primers. Given that the number of publicly available genomes is inevitably biased among lineages, multiple evaluations were conducted to assess the specificity and accuracy of the primers, including qualitative performance in strains of different lineages and complex host and/or microbiota backgrounds, as well as quantitative performance in mock and clinical faecal and tissue samples. Indeed, FL-typing showed good accuracy when compared to FrpoB-seq results. Notably, although the primer set Fn-F/R also informed the abundance of L1, its target *nusG* was not identified as an L1 marker, suggesting that Fn-F/R was not sufficiently stringent for L1 detection. Indeed, its qPCR outputs displayed lower correlation coefficients to L1 abundance than the results with L1_746-F/R. Although the ROC

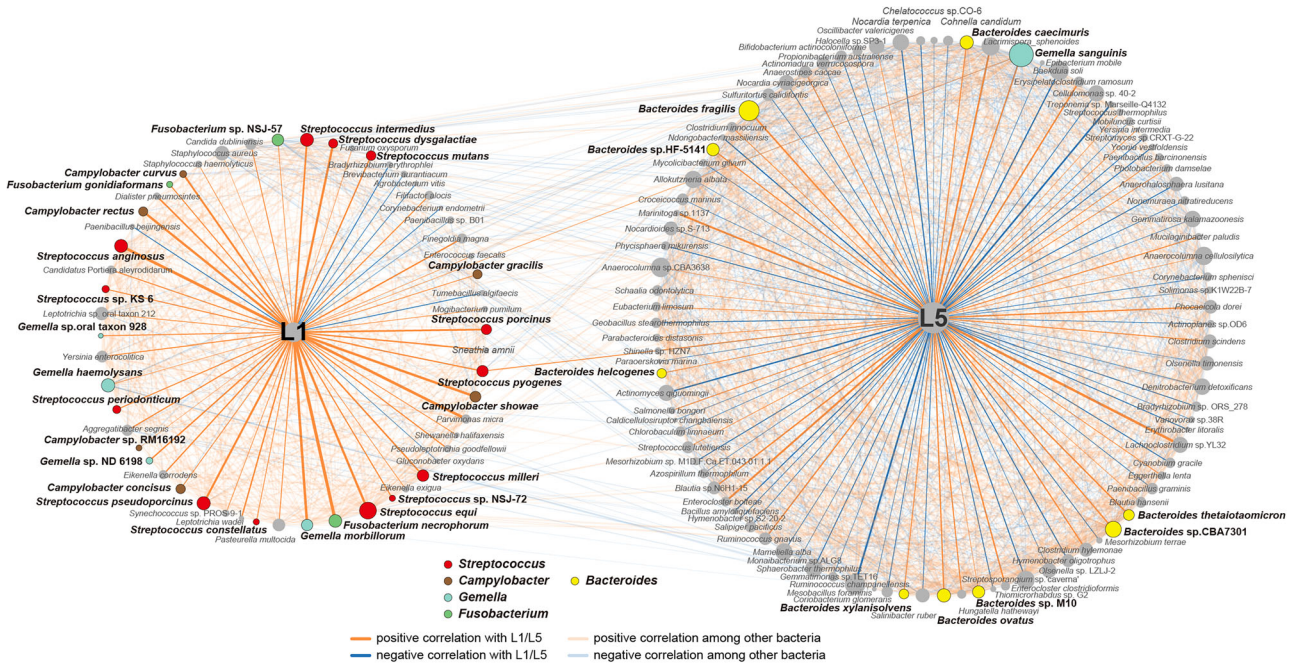


Fig. 6 | Distinct co-occurrence patterns of L1 and L5 with other bacteria. The microbial co-occurrence network, constructed from 67 faecal microbiomes, depicts significant correlations ($p < 0.05$ and absolute correlation coefficient >0.4) among L1, L5 and bacterial species correlating with either lineage. Nodes represent L1, L5 or microbial species, scaled by node degree. Edges denote significant associations, with thickness referring to correlation strength. Species of interest are highlighted as indicated. Detailed correlation data are provided in Supplementary Data 5.

curves for CRC prediction with Fn-F/R and L1_746-F/R did not significantly differ across cohorts, likely due to the small sample sizes, the AUCs of L1_746-F/R were consistently greater than those of Fn-F/R in both analysed cohorts. In light of the importance of precise taxonomic profiling to unravel real disease-associated microbes, we envision that such a systemic strategy can serve as a paradigm of stringent primer design for gut microbe detection. In addition, while we mapped the lineage information to Genome Taxonomy Database (GTDB)[37] representatives and MetaPhlAn species-level genome bins (SGBs)[38] for the studied genomes, future efforts could integrate this method, particularly its lineage markers, into established bioinformatics databases and/or tools. This would extend the method's utility, facilitating deeper investigation into lineage-associated clinical relevance.

This study has several limitations. This was a single-centre study with a relatively small sample size. The results of this study need further confirmation with larger sample sizes in different populations, especially through multi-centre prospective studies, and with in-depth mechanistic interrogation. More *Fusobacterium* isolates and genomes with wide lineage spans are needed to validate and refine the developed method for L1/L5 detection. Significant *Fusobacterium* genome expansion has recently been documented, particularly among strains isolated from CRC tumours as described by Tran et al.[12] and Zepeda-Rivera et al.[9]. Future efforts will focus on incorporating rapidly expanding genomic resources to iteratively enhance the method, following comprehensive taxonomic revalidation and lineage classification. The clinical relevance of other less common lineages could also be explored in future studies, as they were not covered here. Meanwhile, the qPCR-based approach, while widely adopted, is semi-quantitative and limited with low-abundance targets. The ddPCR represents a promising alternative for detecting low-abundance targets and enables absolute quantification. Our method could be further adapted to the ddPCR platform with necessary optimisation and validation, to enhance its utility in future investigations. Furthermore, the specific pathogenic mechanisms of L1 and L5 in CRC development and/or progression remain incompletely elucidated and warrant systemic investigations in future research.

In conclusion, this study retaliated the distinct correlations of *Fusobacterium* lineages L1 and L5 in CRC by developing and applying a highly specific PCR-based quantification method and further showed that L1 is more predictive of CRC than *F. nucleatum* alone in the Chinese population. Our findings suggest that CRC is likely driven by the collective pathogenicity of L1/L5 rather than a single species, providing an important reference for future research. Lineage-specific virulence determinants represent promising targets for subsequent functional studies to elucidate the precise etiological relationship between *Fusobacterium* and CRC. Our results also highlight the clinical diagnostic value of L1 detection, particularly for CRC screening and diagnosis. The FL-typing method is readily integrable into existing clinical and research laboratory platforms.

Methods

Public bacterial genomic data

A total of 157 unique *Fusobacterium* genomes with curated species designation and defined lineage designation[14] as previously described, including 40 complete and 110 draft ones, were retrieved from the NCBI database in February 2022 and included for analysis (Supplementary Data 2). Those genomes were checked for update in February 2023. The genomes of *F. nucleatum* ATCC 25586 and *F. varium* ATCC 27725 were used as references for L1 and L5 for bioinformatic analyses, respectively. Nucleotide sequence search and comparison were performed with BLASTN or Mega BLAST (BLAST+ version 2.9.0) for virulence genes and operons. The corresponding GTDB[37] representatives and matched MetaPhlAn SGBs[38] are also indicated in Supplementary Data 2. The GTDB representatives were identified by direct accession number mapping. The SGBs were identified by aligning the genomes against the MetaPhlAn database (vJan25_202503)[38] via Mega BLAST (BLAST+ version 2.9.0). Note that SGBs, defined as species-level genome bins, are usually specific to defined species.

Bacterial strains

The bacterial strains used in this study are listed in Supplementary Table 1 and were described previously[14,15]. They were cultivated on Columbia agar supplemented with 5% defibrinated sheep blood (Comagal Microbial) at 37 °C in a jar containing the AnaeroPack System (MITSUBISHI Gas Chemical). The bacterial cultures were harvested, resuspended, tenfold serially diluted and plated to determine colony-forming units (CFUs), on

the basis of which the suspensions were adjusted to the desired concentration, followed by reexamination.

Clinical specimens

Clinical samples were retrieved from the biobank of the Shanghai Tenth People's Hospital. To assess the accuracy of the canonical '*F. nucleatum*' detection method and newly developed methods here, DNA from previously collected fresh frozen tissue and faecal samples[14] that had undergone FrpoB-seq were used, including tumour and matched adjacent normal tissue samples from 45 CRC patients and faecal samples from 52 CRC patients and 37 healthy controls (Supplementary Data 1). To validate the clinical significance of L1 and L5, fresh-frozen surgically resected tumour and adjacent normal tissue samples from 100 patients with primary CRC, as well as faecal samples (cohort F2) from 60 CRC patients and 53 healthy volunteers, were newly collected (Supplementary Data 3 and 4). The inclusion and exclusion criteria were the same as those previously described[14]. Briefly, patients (aged >18 years old, either sex) with pathologically diagnosed primary CRC were included. Healthy volunteers (aged >18 years, either sex) were recruited from individuals undergoing routine health examination who had no personal and family history of cancer, no clinical signs of tumour, and a negative faecal immunochemical test. Participants who had taken antibiotics or probiotics, or received faecal microbiota transplantation within one month before sample collection were excluded. All tissue and faecal samples were freshly collected upon admission, immediately snap-frozen in liquid nitrogen and then stored at −80 °C. Clinical and pathological information was obtained from medical records. For ROC curve analysis, previously collected faecal samples (Supplementary Data 1)[14] without FrpoB-seq results from 34 CRC patients and 57 healthy controls were also used together with the abovementioned samples with FrpoB-seq results (collectively, cohort F1). The relative abundance in arbitrary unit obtained by FrpoB-seq was from our previous study[14] and reanalysed here. It was calculated as follows. The compositional percentages in the *Fusobacterium* communities were calculated based on the FrpoB-seq data and then normalised to the relative abundance of total *Fusobacterium* determined by qPCR[14]. The raw FrpoB-seq data are available in the NCBI SRA database under accession number PRJNA715828.

Identification of L1- and L5-specific markers

The identification of L1- and L5-specific markers was performed as previously described[15]. The 157 *Fusobacterium* genomes spanning seven lineages were used to identify genetic markers specific to L1 or L5. For either lineage, pangenome analysis was conducted via PGAP (version 1.2.1)[39] with the GeneFamily method to identify conserved genes. Genes shared by all the genomes of the lineage were then compared at the nucleotide and protein levels against the genomes of other lineages to screen for lineage-specific genes among the conserved ones, which were further searched in the NCBI nucleotide database to exclude those with homologous sequences in non-*Fusobacterium* organisms, yielding a set of genes as candidate lineage markers. Considering that the alleles of each candidate gene had varied sequences, especially across species, we sought to screen for genes that might contain consensus regions sufficient for primer design. The alleles of each candidate gene were aligned against the reference sequence (from the reference genome, see Table 1) via Mega BLAST (BLAST+ version 2.9.0), as this algorithm uses a longer sequence seed to initiate an alignment than BLASTn does. Candidate genes with two or more alleles whose alignment failed were excluded. For L5, as the above method yields more than 50 candidate genes, we selected only those with all alleles having >90% nucleotide identities to the references. The obtained genes were ultimately defined as L1- or L5-specific markers. For each marker, multiple sequence alignment of its alleles was performed via MUSCLE (version 3.8.31)[40], and the corresponding universal primers were designed on consensus or near-consensus regions as previously described[15]. Putative nonspecific amplification of the primers was assessed with the NCBI nucleotide database prior to experimental examination. The genetic marker and corresponding primer set of '*F. nucleatum*' were identified similarly.

DNA preparation

The genomic DNA of the bacterial strains was prepared with the TIANamp Bacteria DNA Kit (TIANGEN) according to the manufacturer's instructions. Total DNA from the faecal and tissue samples was prepared via the cetyltrimethylammonium bromide-based method as previously described[14,15] and examined with Qubit fluorometry. The DNA of tissue or faecal samples from a same cohort were prepared in a same batch.

Polymerase chain reaction (PCR)

PCR was conducted on a Mastercycler nexus gradient thermal cycler (Eppendorf) with Premix Taq™ (Ex Taq™ Version 2.0 plus dye, Takara) according to the manufacturer's instructions. In each reaction, 10 ng of bacterial or 100 ng of faecal or tissue DNA was used, and 30 cycles or 35 cycles for the latter of 98 °C, 30 seconds at 50°C, and 20 seconds at 72 °C were applied. The final concentration of each primer was 0.4 μM. The PCR conditions were established based on assessing different primer concentrations (0.1–1 μM), template inputs (1–1000 ng), and thermal cycling conditions (45–50 °C). The PCR products were electrophoresed on 0.8% agarose gels premixed with SYBR Green I dye (Yeasen Biotech) and visualised via a Tanon 3500 UV gel imaging system (Tanon Science & Technology). The sequences of the widely used primer set to detect '*F. nucleatum*' were Fn-F: 5′-CAACCATTACTTTAACTCTACCATGTTCA-3′ and Fn-R: 5′-GTTGACTTTACAGAAGGAGATTATGTAAAAATC-3′[20,21]. *Fusobacterium* was detected with the primers fuso-rsub-F2 (5′-GCCTCATTTTYTDGARTTYCAATT-3′) and fuso-rsub-R2 (5′-ACDACTCTTTCHGCHCCATTKAT-3′)[14]. Bactria were detected with a primer mixture (V6-mix) targeting the 16S rRNA gene V6 region (5′-CNACGCGAAGAACCTTANC-3′, 5′-ATACGCGARGAACCTTACC-3′, 5′-CTAACCGANGAACCTYACC-3′, 5′-CAACGCGMARAACCTTACC-3′, 5′-CGACRRCCATGCANCACCT-3′)[41]. The primers designed in this study are listed in Table 1. The total DNA of *F. nucleatum* ATCC 25586 and *F. varium* THCT13E1 were used as positive controls and representatives of L1 and L5, respectively. Whole-blood DNA from healthy volunteers was used to represent the host DNA background. Tissue and faecal samples without *Fusobacterium* detected were used to represent real-world *Fusobacterium*-free microbiota background.

Quantitative PCR

Quantitative PCR (qPCR) was performed on a 7500 Real-Time PCR system (Applied Biosystems) with a TB Green Premix Ex Taq II kit (Takara) according to the manufacturer's instructions. For clinical samples, 90 ng of DNA was used in each reaction. The conditions were 50 °C for 2 min; 95 °C for 30 s; and 40 cycles of 95 °C for 5 s, 50 °C for 30 s and 70 °C for 30 s. The L1, L5, and actual '*F. nucleatum*' levels were quantified by the newly designed L1_746-F/R, L5_1621-F/R, and Fn_SR-F/Ra/Rb/Rc, respectively (Table 1). qPCR with Fn-F/R was also performed when needed. The reference was the bacterial 16S rRNA gene quantified by V6-mix. A -ΔCt value was calculated for each sample. For outlier values, repeated detection was performed to exclude experimental errors.

Mock sample creation

The *F. nucleatum* ATCC 25586, *F. varium* THCT13E1, *F. mortiferum* THCT6B2 were used as reference strains of L1, L5 and L4, respectively. To prepare standardised single-strain DNA samples with quantified CFUs, genomic DNA was extracted from 1×10^9 bacterial cells, then normalised to 100 μl, yielding a 1×10^7 CFU/μl standardised suspension. Subsequently, a series of tenfold dilutions was performed, producing concentrations from 10^6 to 10^1 CFU/μl. Real-world single-strain DNA samples were prepared from 1.5 ml bacterial culture, followed by tenfold serial dilutions. Those DNA samples were meanwhile spiked into complex biological matrices containing DNA from *Fusobacterium*-free tissue or faecal samples (70 ng per reaction) alone with DNA from the reference strains of non-target lineages (2.5 ng per strain per reaction). Synthetic L1-community samples containing different ratios (0.1%, 1%, 10%, 50%, 90%, and 100%) of *F. nucleatum* ATCC 25586 and other L1 strains *F. periodonticum* THCT14E2

and *F. hwasookii* THCT18E1 were constructed. The L1-community samples were fixed to have an equal total CFU (10^6 per reaction). The *F. pseudoperiodonticum* and *F. hwasookii* were mixed at a ratio of 1:1. Corresponding control samples containing *F. nucleatum* alone or non-*nucleatum* mixture, equivalently to the *F. nucleatum* portions in the L1-mixture samples were also constructed. The strains THCT13E1, THCT6B2, THCT14E2 and THCT18E1 were originally isolated from CRC tumour tissues (Supplementary Table 1)[14,15].

Metagenomic and 16S rRNA gene amplicon sequencing

Metagenomic and 16S rRNA gene V3-V4 amplicon sequencing was conducted at BGI, Shenzhen, following standard procedures. High-throughput sequencing was performed on the BGI DNBSEQ platform. For metagenomic data, clean reads were assembled with MEGAHIT (v1.2.9)[42] and taxonomic profiling was conducted with Kraken 2 (v2.1.2)[43] against the Human Gastrointestinal Genome database[44]. For 16S rRNA gene sequencing data, operational taxonomic units (OTUs) were generated via USEARCH (usearch11.0.667_i86; https://drive5.com/usearch/) and taxonomic profiling was performed using the RDP classifier (v2.2)[45] against the Silva database[46]. Default parameters were applied in these procedures.

Sensitivity and specificity assessment of the lineage-specific qPCR assay

Qualitative detection sensitivity and specificity of the developed lineage-specific qPCR assay were assessed. A subgroup of faecal samples ($n = 56$) from cohort F1, which contained both L1 and L5 as confirmed by FrpoB-seq[14] and metagenomic sequencing, was used as actual L1/L5-positive samples. Human faecal samples absent from the L1 and L5 commensals were extremely rare. Given that *Fusobacterium* hardly colonise murine gut[47], 29 mouse faecal samples, absent of *Fusobacterium*, as confirmed by 16S rRNA gene amplicon sequencing, were used as actual L1/L5-negative samples. They were collected from seventeen four-week-old male and twelve six-week-old female C57BL/6 mice housed under standard specific pathogen-free conditions with a 12-h light-dark cycle and standard chow and water *ad libitum*. Mice were purchased from Shanghai Slark Experimental Animal Co., Ltd. and acclimatised for at least one week prior to sample collection. They were not anesthetised/unconscious or euthanised during the sample collection. No experimental intervention was conducted on the mice. The qPCR with L1_746-F/R and L5_1621-F/R was conducted, and positive or negative detection were recorded. The sensitivity was calculated as True Positives/(True Positives + False Negatives); while specificity = True Negatives/(True Negatives + False Positives).

L1 and '*F. nucleatum*' abundance estimation in public metagenomic sequencing data

Publicly available faecal metagenomic datasets from eight independent cohorts were collected, encompassing CRC patients and healthy controls across diverse populations in China (Shanghai[23] and Hong Kong[24]), Japan[25], Germany[26], France[27], the United States (USA)[28], Austria[29], and Italy[30]. Raw paired-end sequencing reads in FASTQ format were retrieved from the NCBI Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra) under accession numbers PRJNA706060, PRJNA514108, PRJEB10878, PRJDB4176, PRJEB27928, PRJEB6070, PRJEB7774, PRJEB12449, and SRP136711. Subject metadata were systematically extracted from the NCBI SRA repositories, corresponding original publications and an integrated summary provided by a previous study[48] re-analysing these datasets. Only subjects explicitly classified as CRC or healthy controls were retained for analysis. The Fragments Per Kilobase Million (FPKM) values of the established markers L1_746 and Fn_SR were computed as previously described in ref. 49 to represent L1 and '*F. nucleatum*' abundance, respectively. Specifically, a customised database containing all available L1_746 and Fn_SR sequences was built, and Bowtie 2 (version 2.3.4.3)[50] configured with default parameters, were used for read alignment. FPKM were calculated based on the Bowtie 2 results.

Microbial co-occurrence network analysis of metagenomic sequencing data

Microbial co-occurrence network was inferred with FastSpar (version 1.0.0)[51], a fast, parallelisable C++ implementation of the SparCC algorithm[52], based on the metagenomic sequencing data available for 67 faecal samples from cohort F1 (Supplementary Data 1). Default parameters and 200 bootstraps were applied. Correlations with p values < 0.05 and absolute correlation coefficients >0.4 were retained for further analysis. Here, Kraken 2-derived taxonomic profiles (generated as described above) were used as input. L1 and L5 abundances were calculated as the summed abundance of their constituent species. The resulting network was visualised using Cytoscape (version 3.10.3)[53].

Heatmap

Heatmaps were generated with the ComplexHeatmap (version 2.6.2) package[54] in R (version 4.03). An accompanying dendrogram was drawn with default parameters.

Statistics

Data were analysed with GraphPad Prism (version 7) or R (version 4.03). Exploratory data analysis including descriptive statistics and distribution analysis was conducted to examine the data. The Spearman rank correlation test was used for correlation analysis between qPCR and FrpoB-seq results. The Mann–Whitney test was used for comparisons between two groups (e.g., CRC patients vs. healthy controls; tumour vs. normal tissues; LVI-positive vs. LVI-negative tumour tissues). The Wilcoxon signed-rank test was additionally used to analyse paired data between the tumour and normal tissue groups. Receiver operating characteristic (ROC) curves and Precision-recall (PR) curves were generated to assess and compare the predictive performance for CRC (among L1, '*F. nucleatum*' and Fn-F/R results for the in-house cohorts, or between L1 and '*F. nucleatum*' for the public metagenomic cohorts). The ROC curves were compared with the DeLong and bootstrap (stratified; 1000 replicates) methods through the R package pROC v1.18.0[55], while the PR curves were compared with the bootstrap (stratified; 1000 replicates) method through the R package usefun v0.5.2 (https://cran.r-project.org/web/packages/usefun/). The Benjamini–Hochberg correction was applied for multiple comparison. The ROC and PR curve analyses were conducted in the two in-house faecal sample cohorts F1 (86 CRC patients and 94 healthy controls) and F2 (60 patients and 53 controls) mentioned above, as well as in the published metagenomic cohorts. A two-tailed p value < 0.05 was considered statistically significant.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Ethics

This study was approved by the Ethics Committee of Shanghai Tenth People's Hospital (No. SHSY-IEC-4.1/20-85/01) and was registered with ChiCTR (No. ChiCTR2000037022). Written consent was obtained from each participant. The study was conducted in accordance with the guidelines of the Declaration of Helsinki and the national regulations on ethics reviews of life sciences and medical research of China. The mouse experiment was approved by the Animal Ethics Committee of Shanghai Tenth People's Hospital (No. SHSY-2025-3887) and conducted in accordance with institutional guidelines and national laboratory animal regulations of China for animal experimentation and care.

Data availability

The data generated in this study are available upon request from the corresponding authors. The metagenomic sequencing data, 16S rRNA gene amplicon sequencing data and previously published FrpoB-seq data (reanalysed in this work) are available under the accession number PRJNA715828. The bacterial genomic data were obtained from the NCBI with full accession numbers listed in Supplementary Data 2. The publicly

available metagenomic datasets were retrieved from the NCBI Sequence Read Archive under accession numbers PRJNA706060, PRJNA514108, PRJEB10878, PRJDB4176, PRJEB27928, PRJEB6070, PRJEB7774, PRJEB 12449 and SRP136711.

Received: 23 August 2024; Accepted: 1 September 2025;
Published online: 09 October 2025

References

1. Bullman, S. et al. Analysis of *Fusobacterium* persistence and antibiotic response in colorectal cancer. *Science* **358**, 1443–1448 (2017).
2. Zhou, Z., Chen, J., Yao, H. & Hu, H. *Fusobacterium* and colorectal cancer. *Front. Oncol.* **8**, 371 (2018).
3. Tahara, T. et al. *Fusobacterium* in colonic flora and molecular features of colorectal carcinoma. *Cancer Res.* **74**, 1311–1318 (2014).
4. Kostic, A. D. et al. Genomic analysis identifies association of *Fusobacterium* with colorectal carcinoma. *Genome Res.* **22**, 292–298 (2012).
5. Wong, S. H. & Yu, J. Gut microbiota in colorectal cancer: mechanisms of action and clinical applications. *Nat. Rev. Gastroenterol. Hepatol.* **16**, 690–704 (2019).
6. Wang, N. & Fang, J. Y. *Fusobacterium nucleatum*, a key pathogenic factor and microbial biomarker for colorectal cancer. *Trends Microbiol.* **31**, 159–172 (2023).
7. Brennan, C. A. & Garrett, W. S. *Fusobacterium nucleatum* - symbiont, opportunist and oncobacterium. *Nat. Rev. Microbiol.* **17**, 156–166 (2019).
8. Clay, S. L., Fonseca-Pereira, D. & Garrett, W. S. Colorectal cancer: the facts in the case of the microbiota. *J. Clin. Investig.* **132**, e155101 (2022).
9. Zepeda-Rivera, M. et al. A distinct *Fusobacterium nucleatum* clade dominates the colorectal cancer niche. *Nature* **628**, 424–432 (2024).
10. Yeoh, Y. K. et al. Southern Chinese populations harbour non-nucleatum Fusobacteria possessing homologues of the colorectal cancer-associated FadA virulence factor. *Gut* **69**, 1998–2007 (2020).
11. 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).
12. Tran, H. N. H. et al. Tumour microbiomes and *Fusobacterium* genomics in Vietnamese colorectal cancer patients. *npj Biofilms Microbiom.* **8**, 87 (2022).
13. Fukuoka, H. et al. Elucidating colorectal cancer-associated bacteria through profiling of minimally perturbed tissue-associated microbiota. *Front Cell Infect. Microbiol.* **13**, 1216024 (2023).
14. Bi, D. et al. Profiling *Fusobacterium* infection at high taxonomic resolution reveals lineage-specific correlations in colorectal cancer. *Nat. Commun.* **13**, 3336 (2022).
15. Bi, D. et al. A newly developed PCR-based method revealed distinct *Fusobacterium nucleatum* subspecies infection patterns in colorectal cancer. *Micro. Biotechnol.* **14**, 2176–2186 (2021).
16. Fatahi-Bafghi, M. Genomic and phylogenomic analysis of Fusobacteriaceae family and proposal to reclassify *Fusobacterium naviforme* Jungano 1909 into a novel genus as *Zandiella naviformis* gen. nov., comb. nov. and reclassification of *Fusobacterium necrophorum* subsp. *funduliforme* as later heterotypic synonym of *Fusobacterium necrophorum* subsp. *necrophorum* and *Fusobacterium equinum* as later heterotypic synonym of *Fusobacterium gonidiaformans*. *Antonie Leeuwenhoek* **117**, 34 (2024).
17. Manson McGuire, A. et al. Evolution of invasion in a diverse set of *Fusobacterium* species. *mBio* **5**, e01864 (2014).
18. Kook, J. K. et al. Genome-based reclassification of *Fusobacterium nucleatum* subspecies at the species level. *Curr. Microbiol.* **74**, 1137–1147 (2017).
19. Munson, E., Carella, A. & Carroll, K. C. Valid and accepted novel bacterial taxa derived from human clinical specimens and taxonomic revisions published in 2022. *J. Clin. Microbiol.* **61**, e0083823 (2023).
20. Castellarin, M. et al. *Fusobacterium nucleatum* infection is prevalent in human colorectal carcinoma. *Genome Res.* **22**, 299–306 (2012).
21. Mima, K. et al. *Fusobacterium nucleatum* and T cells in colorectal carcinoma. *JAMA Oncol.* **1**, 653–661 (2015).
22. Gharbia, S. E. & Shah, H. N. Biochemical properties of *Fusobacterium naviforme* and phenotypically similar isolates. *Lett. Appl. Microbiol.* **12**, 177–179 (1991).
23. Gao, R. et al. Integrated analysis of colorectal cancer reveals cross-cohort gut microbial signatures and associated serum metabolites. *Gastroenterology* **163**, 1024–1037.e1029 (2022).
24. Yu, J. et al. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut* **66**, 70–78 (2017).
25. Yachida, S. et al. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nat. Med.* **25**, 968–976 (2019).
26. 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).
27. Zeller, G. et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* **10**, 766 (2014).
28. Vogtmann, E. et al. Colorectal cancer and the human gut microbiome: reproducibility with whole-genome shotgun sequencing. *PLoS ONE* **11**, e0155362 (2016).
29. Feng, Q. et al. Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nat. Commun.* **6**, 6528 (2015).
30. 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).
31. Kolenbrander, P. E., Palmer, R. J. Jr., Periasamy, S. & Jakubovics, N. S. Oral multispecies biofilm development and the key role of cell-cell distance. *Nat. Rev. Microbiol.* **8**, 471–480 (2010).
32. Queen, J. et al. *Fusobacterium nucleatum* is enriched in invasive biofilms in colorectal cancer. *npj Biofilms Microbiom.* **11**, 81 (2025).
33. Ma, L., Feng, J., Zhang, J. & Lu, X. *Campylobacter* biofilms. *Microbiol. Res.* **264**, 127149 (2022).
34. Macfarlane, S., Woodmansey, E. J. & Macfarlane, G. T. Colonization of mucin by human intestinal bacteria and establishment of biofilm communities in a two-stage continuous culture system. *Appl Environ. Microbiol.* **71**, 7483–7492 (2005).
35. Berkhout, M. D. et al. Mucin-driven ecological interactions in an in vitro synthetic community of human gut microbes. *Glycobiology* **34**, cwae085 (2024).
36. Kvich, L. et al. Biofilms and core pathogens shape the tumor microenvironment and immune phenotype in colorectal cancer. *Gut Microbes* **16**, 2350156 (2024).
37. Parks, D. H. et al. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* **50**, D785–d794 (2022).
38. Blanco-Míguez, A. et al. Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat. Biotechnol.* **41**, 1633–1644 (2023).
39. Zhao, Y. et al. PGAP: pan-genomes analysis pipeline. *Bioinformatics* **28**, 416–418 (2012).
40. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
41. Geller, L. T. et al. Potential role of intratumor bacteria in mediating tumor resistance to the chemotherapeutic drug gemcitabine. *Science* **357**, 1156–1160 (2017).
42. Li, D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T. W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).

43. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
44. Almeida, A. et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* **39**, 105–114 (2021).
45. Wang, Q. & Cole, J. R. Updated RDP taxonomy and RDP Classifier for more accurate taxonomic classification. *Microbiol. Resour. Announc.* **13**, e0106323 (2024).
46. Yilmaz, P. et al. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Res.* **42**, D643–D648 (2014).
47. Queen, J. et al. Comparative analysis of colon cancer-derived *Fusobacterium nucleatum* subspecies: inflammation and colon tumorigenesis in murine models. *mBio* **13**, e0299121 (2021).
48. Liu, N. N. et al. Multi-kingdom microbiota analyses identify bacterial-fungal interactions and biomarkers of colorectal cancer across cohorts. *Nat. Microbiol.* **7**, 238–250 (2022).
49. Danko, D. et al. A global metagenomic map of urban microbiomes and antimicrobial resistance. *Cell* **184**, 3376–3393.e3317 (2021).
50. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
51. Watts, S. C., Ritchie, S. C., Inouye, M. & Holt, K. E. FastSpar: rapid and scalable correlation estimation for compositional data. *Bioinformatics* **35**, 1064–1066 (2019).
52. Friedman, J. & Alm, E. J. Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* **8**, e1002687 (2012).
53. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
54. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
55. Robin, X. et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinforma.* **12**, 77 (2011).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (82072236 to D.B. and 82072634 to Q.W.), the Clinical Research Plan of SHDC (SHDC2020CR2069B to Q.W.) and the Special Development Funds of Zhangjiang National Innovation Demonstration Zone (ZJ2022-ZD-005 to H.Q.).

Author contributions

D.B. designed the study. D.B., Q.W. and H.Q. supervised the study and acquired funding. D.B., Y.W., G.J., D.H., X.Z., H.L., M.L., Y.Z., Y.G., L.L. and R.X. acquired the data. Y.W., G.J., D.H., X.Z. and Y.G. performed data curation. D.B., Y.W., G.J., D.H., Y.G., Y.Z., Z.D. and M.X. performed the data analysis and interpretation. D.B., Y.W., G.J., D.H., M.X., Z.D., Q.W. and H.Q. drafted and revised the manuscript. All the authors have approved the submitted version.

Competing interests

D.B., Q.W. and H.Q. are the inventors on patent applications filed by the Shanghai Tenth People's Hospital related to the lineage-specific detection method and improved detection method for the formerly called 'F. nucleatum' (nos. 2023108938211 and 2023118427636, respectively) described herein. The other authors declare that they have no potential conflicts of interest.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41522-025-00826-3.

Correspondence and requests for materials should be addressed to Qing Wei, Huanlong Qin or Dexi Bi.