Patients with mesenchymal tumours and high *Fusobacterales* prevalence have worse prognosis in colorectal cancer (CRC)

Manuela Salvucci, Nyree Crawford, Katie Stott, Susan Bullman, Daniel B Longley, Jochen H M Prehn

**ABSTRACT**

**Objectives** Transcriptomic-based subtyping, consensus molecular subtyping (CMS) and colorectal cancer intrinsic subtyping (CRIS) identify a patient subpopulation with mesenchymal traits (CMS4/CRIS-B) and poorer outcome. Here, we investigated the relationship between prevalence of *Fusobacterium nucleatum* (Fn) and *Fusobacterales*, CMS/CRIS subtyping, cell type composition, immune infiltrates and host contexture to refine patient stratification and to identify druggable context-specific vulnerabilities.

**Design** We coupled cell culture experiments with characterisation of Fn/Fusobacterales prevalence and host biology/microenvironment in tumours from two independent colorectal cancer patient cohorts (Taxonomy: n = 140, colon and rectal cases of The Cancer Genome Atlas (TCGA-COAD-READ) cohort: n = 605).

**Results** In vitro, Fn infection induced inflammation via nuclear factor kappa-light-chain-enhancer of activated B cells/tumour necrosis factor alpha in HCT116 and HT29 cancer cell lines. In patients, high Fn/Fusobacterales were found in CMS1, microsatellite unstable (M) tumours, with infiltration of M1 macrophages, reduced M2 macrophages, and high interleukin (IL)-6/IL-8 signalling. Analysis of the Taxonomy cohort suggested that Fn was prognostic for CMS4/CRIS-B patients, despite having lower Fn load than CMS1 patients. In the TCGA-COAD-READ cohort, we likewise identified a differential association between Fusobacterales relative abundance and outcome when stratifying patients in mesenchymal (either CMS4 or CRIS-B) versus non-mesenchymal (neither CMS4 nor CRIS-B). Patients with mesenchymal tumours and high Fusobacterales had approximately twofold higher risk of worse outcome. These associations were null in non-mesenchymal patients.

Modelling the three-way association between Fusobacterales, molecular subtyping and host contexture with logistic models with an interaction term disentangled the pathogen-host signalling relationship and identified aberrations (including NOTCH, CSF1-3 and IL-6/IL-8) as candidate targets.

**Conclusion** This study identifies CMS4/CRIS-B patients with high Fn/Fusobacterales prevalence as a high-risk subpopulation that may benefit from therapeutics targeting mesenchymal biology.

**Significance of this study**

What is already known on this subject?

- *Fusobacterium nucleatum* (Fn), a commensal Gram-negative anaerobe from the *Fusobacterales* order, is an oncobacterium in colorectal cancer (CRC), and a causal relationship between Fn prevalence and CRC pathogenesis, progression and treatment response has been reported in vivo.
- Broad-spectrum antibiotics have proven moderately successful in reducing tumour growth promoted by Fn in preclinical models. However, the use of antibiotics to treat bacteria-positive cases in the clinic is not a viable option as it may further alter the already dysbiotic gut microbiome of patients with CRC and may also have limited efficacy against Fn, which penetrates and embeds deeply within the tumour.
- The highly heterogeneous population of patients with CRC can be classified into distinct molecular subtypes (consensus molecular subtyping (CMS) and colorectal cancer intrinsic subtyping (CRIS)) based on gene expression profiles mirroring the underlying transcriptional programmes. Patients classified as CMS4 and CRIS-B exhibit a mesenchymal phenotype and have poorer outcome.

INTRODUCTION

Colorectal cancer (CRC) has one of the highest morbidities and mortality rates among solid cancers, and its incidence is steadily on the rise, accounting for circa 10% of newly diagnosed cancer cases worldwide. Patients with CRC with similar macroscopic clinicopathological characteristics exhibit a high degree of heterogeneity at the molecular level, which translates into heterogeneous and often suboptimal response to treatment. Thus, research has focused on molecular subtyping strategies based on single or multimics data from the host to categorise patients into subgroups to aid in risk stratification and disease management. Subtyping strategies such as the consensus molecular subtyping (CMS) and the colorectal cancer intrinsic subtyping (CRIS) classify patients into subgroups with more homogeneous signalling features based on key transcriptomic programmes. Among the four subtypes identified by the CMS classifier, CMS4 patients have high stroma infiltration along with upregulated angiogenesis and transforming growth.
What are the new findings?

- F. nucleatum prevalence is associated with immune involvement (decrease in antitumour M1 macrophages and increase in protumour M2 macrophages) and activation of specific signalling programmes (inflammation, DNA damage, WNT, metastasis, proliferation and cell cycle) in the host–tumours.
- The prevalence of bacteria from the Fusobacteriales order, largely driven by F. species, plays an active or opportunistic role, depending on the underlying host–tumour biology and microenvironment.
- F. and other species of the Fusobacteriales order are enriched in CMS1 (immune-high, microsatellite unstable tumours) patients compared with CMS2–4 cases.
- F. Fusobacteriales prevalence is associated with worse clinical outcome in patients with mesenchymal-rich CMS4/CRIS-B tumours but not in patients with other molecular subtypes.

How might it impact on clinical practice in the foreseeable future?

- F. Fusobacteriales screening and transcriptomic-based molecular subtyping should be considered to identify patients with mesenchymal-rich tumours and high bacterium prevalence to inform disease management.
- F. Fusobacteriales prevalence may need to be addressed exclusively in patients with mesenchymal-rich high-stromal infiltrating tumours rather than a blanket approach to treat all pathogen-positive patients.
- Clinical management of the disease for this subpopulation of high-risk patients with unfavourable clinical outcome could be attained by administering agents currently in clinical trials that target aberrations in the host signalling pathways (NOTCH, WNT and epithelial-mesenchymal transition) and tumour microenvironment (inflammamosome, activated T cells, complement system, and macrophage chemotaxis and activation).

Characterisation of cell signalling and tumour microenvironment in n = 745 patients to investigate the interaction between the dysregulation induced by Fusobacteriales prevalence (including F.) on the human host and, conversely, the characteristics of the host microenvironment that allow pathogens to thrive. Here, we provide evidence that the prognostic value of F. Fusobacteriales strongly relates to the molecular subtype of the host–tumour and is confined to subtypes showing mesenchymal involvement.

MATERIALS AND METHODS

Detailed methods for the in vitro experiments and the patients’ study (design, cohorts’ description and analysis steps) are provided in the online supplemental materials and methods.

RESULTS

F. infection induces inflammation mediated by tumour necrosis factor alpha (TNF-α) and NFκB in CRC cellular cultures

Due to the presence of F. in CRC tumour tissue, a causative role for this bacterium in exacerbating tumorigenesis has been put forward. Infection of colon cells with F. has previously been shown to induce inflammation, activate NFκB signalling and increase expression of the proinflammatory cytokine TNF-α (figure 1A). Hence, we infected HCT116 and HT29 colon cancer cell line cultures for 6 hours to assess epithelial cell response to increasing amounts of F. (multiplicity of infection, (MOI), bacteria-to-cancer-cells: 10, 100 and 1000). We found that NFκB signalling was activated on infection with F. in CRC cell lines, as evidenced by the degradation of IκBα (alpha nuclear factor of kappa light polypeptide gene enhancer in B-cell inhibitor) (figure 1B), an increase in NFκB transcriptional activity (figure 1C) and a marked increase in mRNA expression of the NFκB target gene, TNF-α (figure 1D). Taken together, these results confirm that F. and NFκB expression with human colon cancer epithelial cells promotes a proinflammatory response.

Prevalence of Fusobacteriales in tumour resections

We sought to investigate the relationship between inflammation in the human host and prevalence of F. Fusobacteriales in tumour resections of patients with CRC. We selected an in-house multicentre stage II–III cohort (Taxonomy, n = 140) and the colon and rectal cases of The Cancer Genome Atlas (TCGA-COAD-READ cohort, n = 605 patients; figure 2A) to encompass the heterogeneity of the CRC clinicopathological characteristics observed in the clinic. Demographic, clinicopathological characteristics for the Taxonomy and TCGA-COAD-READ cohorts are summarised in online supplemental table 1. We determined F. load by a targeted quantitative real-time PCR in tumour resections of the Taxonomy cohort where we detected F. in n = 101 of 140 (72%) patients (figure 2B and online supplemental table 2). The distribution of F. positivity levels (relative to the human PGT genome) was heterogeneous, and we categorised patients as F.-high or F.-low using the 75th percentile as cut-off (figure 2B). We estimated Fusobacteriales relative abundance (RA) in the TCGA-COAD-READ cohort from RNA sequencing data by mapping non-human reads to microbial reference databases and retaining only high-quality matches (see the Materials and methods section) with a PathSeq analysis (figure 2A and online supplemental table 2). For downstream analyses, we reported the RA at the family, genus and species taxonomic rank, and expressed it as percentage of the
total bacterial abundance. We detected *Fusobacteriales* (defined as RA over zero, at the order level) in n=558 of 605 (92%) of the TCGA-COAD-READ patients (figure 2D). *Fn* was the most abundant species and was detected in 82% of the TCGA-COAD-READ patients (compared with 72% in the Taxonomy cohort), accounting on average for approximately 45% of total *Fusobacteriales* RA and accounting for over 75% of total *Fusobacteriales* RA in 16% of cases (figure 2C). Analogous to the Taxonomy cohort, we categorised patients as *Fusobacteriales*-high or *Fusobacteriales*-low using the 75th percentile as cut-off.

**Higher Fn/Fusobacteriales prevalence correlates with inflammation and immune involvement**

We examined the association between host gene expression profiles of key inflammatory markers and either *Fn* load or *Fusobacteriales* RA in the Taxonomy and TCGA-COAD-READ cohorts, respectively. In line with the in vitro experiments (figure 1), we detected an increase in NFκB1 and a trend in TNF-α gene expression, recapitulated by transcriptomic-based signatures for an overall inflammation status mediated by the NFκB–TNF-α axis and IFN-γ with cytolytic involvement. *Fusobacteriales*-high patients overexpressed proinflammatory ILs (IL-6, IL-8, IL-10 and IL-1β), cytokines/chemokines (CCL8 and ICAM1), metalloproteinases (MMP1 and MMP3), NOS2 and inflammasome markers (NLRP3) (figure 2F).

As inflammation is strongly tied to immune cell migration and activity, we investigated whether there was a link between immune cell composition and either *Fn* load (taxonomy) or *Fusobacteriales* RA (TCGA-COAD-READ). Cell composition was computationally deconvoluted from gene expression profiles with quanTIseq<sup>16</sup> and microenvironment cell populations (MCP)-counter<sup>17</sup> (figure 2G,H). Despite observing high interpatient heterogeneity in cell composition within the Taxonomy and TCGA-COAD-READ cohorts, we detected higher immune cell activation and polarisation when comparing patients with high versus low *Fn* load (Taxonomy) or *Fusobacteriales* RA (TCGA-COAD-READ). Patients with high *Fn* load (Taxonomy) or *Fusobacteriales* (TCGA-COAD-READ) showed higher predicted abundance of regulatory T cells coupled with an increase in M1 macrophages and a decrease in M2 macrophages (figure 2L,J). MCP-counter identified a strong positive association between neutrophil infiltration and either *Fn* load (Taxonomy) or *Fusobacteriales* RA (TCGA-COAD-READ). However, no difference in predicted neutrophils abundance was detected by quanTIseq.
Figure 2  
Fusobacteria prevalence is associated with inflammation and immunosuppression in patients with CRC of the Taxonomy and TCGA-COAD-READ CRC cohorts. (A) Schematic representation of the cohorts included in the study and methods to estimate Fn load and Fusobacteria (order) RA in the Taxonomy and TCGA-COAD-READ cohorts, respectively. (B–D) Per-patient (waterfall plot, 1, left) and distribution (violin plot with overlaid data-points, 2, right) of bacterium prevalence in tumour resections of the Taxonomy (n=140, B) and TCGA-COAD-READ (n=605, D). In B,D 1, patients are sorted in ascending order of either Fn load (Taxonomy cohort, B) or Fusobacteria RA at the order taxonomic rank (TCGA-COAD-READ cohort, D). Cut-off of 75th percentile used for patients’ stratification in downstream analysis is also indicated (black dotted line). (C) Corresponding per-patient fraction of Fn species to total Fusobacteria order RA detected for the TCGA-COAD-READ cohort. (E,F) Violin plots depicting the expression distribution of key genes or signatures involved in inflammation and immunosuppression grouped by patients with low (in green) or high (in orange) either Fn load (Taxonomy cohort, E) or Fusobacteria RA at the order taxonomic rank (TCGA-COAD-READ cohort, F). Median and lower (25th) and upper (75th) percentiles are indicated by white solid or dashed lines, respectively. Statistical significance was evaluated using Kruskal-Wallis tests and p values are reported. (G,H) Stacked bar plots indicating cell type composition per patient estimated from gene expression by quanTIseq in tumours with low versus high either Fn load (Taxonomy cohort, G) or Fusobacteria RA at the order taxonomic rank (TCGA-COAD-READ cohort, H). Cell type composition is shown sorted in ascending order of tumour and stromal content (1 and 3) and aggregated (by mean, 2 across the low and high subgroups). (I,J) Distribution of specific tumour/stroma and immune cell types determined as indicated by either quantTIseq or MCP-counter grouped by either Fn load (Taxonomy cohort, I) or Fusobacteria RA at the order taxonomic rank (TCGA-COAD-READ cohort, J). Median and lower (25th) and upper (75th) percentiles are indicated by white solid or dashed lines, respectively. Statistical significance was evaluated using Kruskal-Wallis tests and p values are reported. CRC, colorectal cancer; Fn, Fusobacterium nucleatum; NK, natural killer cells; RA, relative abundance; TCGA-COAD-READ, colon and rectal cases of The Cancer Genome Atlas; Treg, regulatory T cell.
Importantly, no difference in fibroblasts and endothelial cells was observed by Fn/Fusobacteriales in either cohort by either method (figure 2LJ).

Multiomic characterisation of the association between Fusobacteriales RA and human host–tumour microenvironment in the TCGA-COAD-READ cohort

We leveraged the rich molecular characterisation of the TCGA-COAD-READ cohort to perform a systematic and unbiased characterisation of the association between Fusobacteriales RA and patient clinical and molecular features to identify human host vulnerabilities that may be conducive for tumour development (figure 3).

We observed higher Fusobacteriales in patients of older age, diagnosed with more advanced disease stage and tumours located in the colon, particularly in proximal site (figure 3A) cohorts. In contrast, we found no statistically significant differences in Fusobacteriales RA by sex, body mass index and either lymphovascular or perineural invasion (online supplemental figure 2). We observed similar patterns and a slightly higher prevalence in women (Taxonomy cohort, p=0.049), when assessing Fn in both the TCGA-COAD-READ and Taxonomy cohorts (online supplemental figure 3A), corroborating previous studies.18

Patients harbouring higher Fusobacteriales showed lower genomic intratumour heterogeneity, had higher silent and non-silent mutational burden and were enriched in microsatellite unstable cases (figure 3B and online supplemental figure 3B). Fusobacteriales-high patients had an increase in transversions, defined as the exchange of two-ring purines (A↔G) or of a one-ring pyrimidines (C↔T), coupled with a decrease in transversions, a substitution of purine for pyrimidine bases (online supplemental figure 4A) as evidenced by a decrease in conversion changes of C>G and T>A (online supplemental figure 4B). We found no difference in prevalence of common mutations in CRC by Fusobacteriales (low vs high) except for BRAF (figure 3C). BRAF mutations tended to be more common among Fusobacteriales-high and Fn-high patients, as previously reported when assessing Fn18 (figure 3C and online supplemental figure 3B). A comprehensive screen revealed that mutations in cell cycle (ATM), Hedgehog signalling (MEGF8), DNA damage/repair (TRIP12 and PRKDC), mitotic spindle (ASPM) and migration/adhesion (TRIO, GPR98) were more prevalent in Fusobacteriales-high patients (figure 3D) (online supplemental table 3).

We set out to investigate the relationship between copy number alterations (CNAs) and Fusobacteriales presence in the TCGA-COAD-READ cohort (figure 3E–G). We determined recurrent CNA amplifications and deletions across the whole cohort by applying the genomic identification of significant targets in cancer (GISTIC) algorithm19 (online supplemental figures 5 and 6 and online supplemental table 4). Fusobacteriales-high cases showed lower chromosomal instability with a lower fraction of the genome affected by recurrent CNAs, in line with the increased incidence of MSI. We identified CNA amplifications or deletions, the frequency of occurrence of which differed between Fusobacteriales-high versus Fusobacteriales-low patients and, thus, may be specifically associated with the bacterium presence (figure 3F). CNAs more frequently (>15%) observed in Fusobacteriales-high versus Fusobacteriales-low cases included deletions in 8p23.2 (tumour suppressor CSMD1 and LOC100287015), 18q21.1 (MIR4743 and RNA binding by CTHF) and 18q23, which impact the regulation of IL-6 and chemokine secretion, cell–cell adhesion and host response of viral transcription (figure 3G).

We then focused on the transcriptional level and combined enrichment analyses with pathway-activity signatures to compare the impact of Fusobacteriales RA on cellular processes (figure 3H–J). Transcriptional profiles that differed included mTORC1 and cMYC signalling, cell cycle (G2-M checkpoint), mitotic spindle, epithelial-to-mesenchymal transition, TGF-β and IL-1 regulation of extracellular matrix, matrix remodeling including focal adhesion, cytoskeleton and contractile actin filament bundle, mitochondrial translational elongation/termination, and protein complex assembly and stromal estimates (figure 3HIJ, online supplemental figure 7 and supplemental table 5). We corroborated these findings by comparing the activation of signalling pathways estimated by gene set signatures identified in the literature (see the Materials and methods section) in Fusobacteriales-low versus Fusobacteriales-high patients. Fusobacteriales RA was inversely linked to WNT signalling and positively associated with proliferation, metastasis (figure 3J) and DNA damage.

We sought to investigate whether the findings at the genomic and transcriptional levels were also observed in protein profiles determined by reverse phase protein array. We found that Fusobacteriales RA correlated with differential expression of proteins involved in microenvironment composition (Claudin7), cell cycle (Cyclin1), CMYC, apoptosis (cleaved Caspase7), proliferation (DLV3), Hippo pathway (Yap), DNA damage (Chk1 and ATM), receptor and mitogen-activated protein (MAP) kinases and PI3K signalling (figure 3K–M, online supplemental figure 8 and supplemental table 6).

Fn/Fusobacteriales prevalence differs by transcriptomic-based molecular subtype

The aforementioned systematic screen pinpointed host aberrations associated with Fusobacteriales hallmarking by transcriptomics-based molecular subtypes. Hence, we classified patients in the study by CMS2 and CRIS3 subtyping. We observed higher Fn load (Taxonomy, figure 4A) and Fusobacteriales RA (TCGA-COAD-READ, figure 4C) in immune-high CMS1 tumours, corroborating the link between pathogen prevalence and host immunity. We observed higher Fn load in CRIS-B tumours (figure 4B) and Fusobacteriales RA in CRIS-A cases (figure 4D) of the Taxonomy and TCGA-COAD-READ cohorts, respectively. At the family rank, Fusobacteriaceae were more abundant than Leptotrichiaceae, accounting for 77% and 23% of total Fusobacteriales RA and ~2% and ~<1% of the total bacteria RA, respectively. In line with the findings at the order level, we observed an increase in Fn, the most abundant Fusobacterium species, in CMS1 and CRIS-A cases (figure 4E,F). In line with the findings at the order level, we observed an approximately threefold increase when comparing patients classified as CMS1 versus the rest (figure 4E). Fn, the most abundant Fusobacterium species, was enriched in CMS1 and CRIS-A cases (figure 4E,F). We examined whether the positive association between inflammation and immune involvement by Fn/Fusobacteriales presence could be ascribed to the host CMS1 milieu or whether there was an additional pathogen-induced component. When restricting the analysis to CMS1 cases, we observed higher expression of proinflammatory markers in Fusobacteriales-high patients of the TCGA-COAD-READ cohort. We detected no association between pathogen prevalence and expression of anti-inflammatory markers or inflammation signatures in either CRC cohort (figure 4G,H). Taken together, these results suggest that
Figure 3  Multomic characterisation of the association between Fusobacteriales RA and human host–tumour microenvironment in the TCGA-COAD-READ cohort. (A,B). Association between Fusobacteriales at the order taxonomic rank binned into low versus high (cut-off 75th percentile) and clinicopathological (A) and mutational (B) characteristics of the human host. (C,D) Comparison of frequency of occurrence of mutations selected a priori (C) or identified by an unbiased scan (D) in Fusobacteriales-low versus Fusobacteriales-high patients. Colour bar indicates number of detected aberrations among frame shift deletions and insertions, in frame deletions and insertions, missense and nonsense mutations, and splice sites. P values were computed with χ² independence tests and adjusted for multiple comparisons (Benjamini-Hochberg false discovery rate). (E–G) Heatmap (E) displaying copy number alterations grouped by Fusobacteriales-low (in green) and Fusobacteriales-high (in orange) RA. Waterfall plot (F) displaying differences in recurrent copy number aberrations detected in patients with low Fusobacteriales versus high Fusobacteriales. Top panel (F) reports percentage of patients affected by recurrent copy number aberrations. Distribution of top 3 deletions, the frequency of occurrence of which differs between Fusobacteriales-low and Fusobacteriales-high patients (G). Red and blue shadings indicate amplification and deletions, respectively. (H–J) Heatmap (H) displaying expression of genes differentially expressed when comparing Fusobacteriales-low versus Fusobacteriales-high patients and corresponding pathway enrichment analysis (I). Expression distribution grouped by Fusobacteriales RA (low, in green, vs high, in orange) for selected gene expression signatures (J). (K–M) Heatmap (K) displaying expression of proteins differentially expressed when comparing Fusobacteriales-low versus Fusobacteriales-high patients and corresponding pathway enrichment analysis (L). Expression distribution grouped by Fusobacteriales RA (low, in green, vs high, in orange) for key proteins (M). In violin plots, the median and lower (25th) and upper (75th) percentiles are indicated by white solid or dashed lines, respectively. Green and orange annotation bars denote patients with low versus high Fusobacteriales RA (75th percentile cut-off). (Unadjusted) P values (J,M) were determined by Kruskal-Wallis tests. MSS, patients with microsatellite stable tumours; RA, relative abundance; TCGA-COAD-READ, colon and rectal cases of The Cancer Genome Atlas.
**Figure 4** Prevalence of *Fn/Fusobacteriales* by transcriptomic-based molecular subtypes of the host. (A–D) Boxplot with overlaid dot plots displaying the dependency by CMS (A,C) and CRIS (B,D) molecular subtyping by either *Fn* load (Taxonomy cohort; A,B) or *Fusobacteriales* RA at the order taxonomic rank (TCGA-COAD-READ cohort; C,D). (E,F) RA (to total bacterial kingdom) of *Fusobacteriales* reported at increasing resolution of taxonomic rank (family, genus and species) by CMS (E) and CRIS (F) subtypes (aggregated by mean). Genuses/species with an average RA lower than 0.05 were aggregated as ‘other’. (G,H) Distribution of key (pro-)/(anti-)inflammatory genes in CMS1 patients classified as ‘low’ (in green) or ‘high’ (in orange) using the 75th percentile as cut-off. Patients’ stratification was based on either *Fn* load (Taxonomy cohort) or *Fusobacteriales* RA at the order taxonomic rank (TCGA-COAD-READ cohort). Median and lower (25th) and upper (75th) percentiles are indicated by white solid or dashed lines, respectively. (Unadjusted) P values were determined by Kruskal-Wallis tests. CMS, consensus molecular subtyping; CRIS, colorectal cancer intrinsic subtyping; *Fn*, *Fusobacterium nucleatum*; RA, relative abundance; TCGA-COAD-READ, colon and rectal cases of The Cancer Genome Atlas.
*Fusobacteriales* may play an active role in mediating inflammation in the host.

**Patients with high *Fusobacteriales* have worse outcome in CMS4/CRIS-B**

We sought to investigate whether bacterium presence correlated with patient clinical outcome assessed by overall survival (OS), disease-specific survival (DSS) and disease-free survival (DFS) endpoints (figure 5 and online supplemental figures 9 and 10).

We found no statistically significant differences in either cohort when comparing survival curves from patients grouped by either *Fn* load or *Fusobacteriales* RA (figure 5A,E,1 and online supplemental figure 9-10). We hypothesised that *Fusobacteriales* may result in poorer outcome in a subtype-dependent context (ie, mesenchymal status; figure 5B,E,]). Indeed, we identified a differential association between *Fusobacteriales* RA and clinical outcome of the TCGA-COAD-READ cohort in mesenchymal (either CMS4 and/or CRIS-B) versus non-mesenchymal (neither CMS4 nor CRIS-B) tumours (figure 5G,H,K,L and online supplemental figure 10). *Fusobacteriales*-high mesenchymal patients had approximately twofold higher risk of worse outcome, whereas these associations were null in non-mesenchymal patients (figure 5G,H,K,L and online supplemental figure 10). Importantly, these findings held true when accounting for key (adjusted model 1) and more extensive (adjusted model 2) clinical–pathological characteristics that may represent confounders or disease modifiers (online supplemental table 7). We fitted two additional Cox regression models where, in addition to the interaction term between *Fusobacteriales* and mesenchymal status, we included adjustment covariates. In adjusted model 1, we included age, stage, tumour location and sex as key clinicopathological and demographic covariates. In adjusted model 2, we expanded on adjusted model 1 by also including history of colon polyps and history of other malignancy as comorbidities. We found that the risk of unfavourable outcome (HRs) and statistical significance were minimally impacted by accounting for potential disease modifiers in adjusted models 1 and 2, confirming the robustness of our findings (online supplemental table 7).

Although numbers in the Taxonomy cohort are more limited, when restricting the analysis to CMS4 and/or CRIS-B cases, we observed a trend in which *Fn*-high patients had shorter OS than those with low *Fn* load. Again, no difference in survival according to *Fn* load was observed in non-mesenchymal Taxonomy patients (figure 5C and online supplemental figure 9).

Exploratory analyses examining the association between clinical outcome and pathogen prevalence at taxonomic ranks of increasing resolution (order, family, genus and species) in the TCGA-COAD-READ cohort by fitting Cox regression models on the whole unsellected population and in mesenchymal versus non-mesenchymal settings revealed that the prognostic impact stems primarily from, but is not limited to, species, including *Fn*, from the *Fusobacterium* genus from the *Fusobacteriaceae* family (figure 5M and online supplemental figure 11).

**Putative mechanisms underlying selective *Fusobacteriales* virulence in mesenchymal tumours**

Having identified a patient subgroup that has an unfavourable clinical outcome when their tumours exhibit mesenchymal traits and are highly positive with *Fn/Fusobacteriales*, we reasoned that intervening by either clearing *Fn/Fusobacteriales* with broad-spectrum antibiotics or targeting the host–tumour biology could ameliorate clinical outcome for this subgroup of patients. Given that broad-spectrum antibiotics may not represent a viable avenue in the clinic and narrow-spectrum antibiotics currently do not exist, we set out to identify clinically actionable host-specific vulnerabilities that could be exploited. We examined the host signalling pathways and microenvironment to identify alterations that may be mediated by and/or exacerbated by *Fusobacteriales* (ie, interact) and, thus, may promote virulence and, ultimately, result in an unfavourable clinical outcome. To disentangle the three-way association between *Fusobacteriales* RA, gene/signature and molecular subtyping, we fitted two distinct logistic regression models for each feature of interest in the TCGA-COAD-READ cohort. The selection of features was hypothesis-driven and included key host signalling pathways and immunomodulators (figure 6A).

Figure 6A reports p values from the two models capturing the association between *Fusobacteriales* RA (high vs low) and either each gene/signature (model 1: *Fusobacteriales*–gene/signature, x-axis) or the interaction between each gene/signature with the molecular subtype (model 2: *Fusobacteriales*–gene/signature×molecular subtype, y-axis). The top half quadrant (darker grey shaded area) identifies a set of genes/signatures whose expression patterns differ by molecular subtype (statistically significant interaction p value in model 2) and thus may be mediating the signalling impact of *Fusobacteriales* and were prioritised for downstream analyses (figure 6B).

We tested whether the gene/signature we identified as candidate targets are indeed related to clinical outcome in patients of the TCGA-COAD-READ cohort with mesenchymal tumours and high *Fusobacteriales*. We restricted our analysis to patients with mesenchymal tumours, and for each clinical endpoint of interest, namely, OS, DSS or DFS, we fitted Cox regression models with an interaction term for *Fusobacteriales* RA (low vs high) and each of the gene/signature (low vs high) identified as statistically significant in the analysis presented in figure 6A. We reasoned that a gene/signature could be considered a candidate target with both specific and translatable impact on clinical outcome for patients with mesenchymal tumours if its association with unfavourable clinical outcome differed by *Fusobacteriales*. This analysis identified CSF1-3, IL-1β, IFN-γ, IL-8, IL-6, CD163, NOTCH2, ZEB2 and TFF2 as potential targets for patients with mesenchymal tumours and high *Fusobacteriales* (figure 6C and online supplemental figures 12–14).

**DISCUSSION**

*Fusobacteriales*, predominantly *Fn*, have been associated with pathogenesis, progression and treatment response in CRC. We coupled mechanistic studies in cell cultures with hypothesis-driven and unbiased screening in clinically relevant and omics-rich CRC cohorts to examine the cross-talk between pathogen–host and pathogen–tumour microenvironment. We demonstrate relationships between *Fn/Fusobacteriales* prevalence and host immunity, signalling and transcriptomic-based molecular subtypes. Our findings suggest that host–pathogen interactions can define patient subpopulations where *Fn/Fusobacteriales* play an active or opportunistic role, depending on the underlying host–tumour biology and microenvironment and identify putative druggable and clinically actionable vulnerabilities.

We observed higher *Fn/Fusobacteriales* prevalence in CMS1 patients, corroborating findings by Purcell et al. Interestingly, we found that overall, higher pathogen prevalence did not correlate with poorer disease outcome. In contrast, high *Fn/Fusobacteriales* levels were associated with poor prognosis in the CMS4/CRIS-B patient subset, suggesting that the presence of...
Figure 5  High Fn/Fusobacteriales prevalence is associated with negative clinical outcome in patients with mesenchymal-like tumours. (A–L) Kaplan-Meier estimates comparing survival curves in patients of the Taxonomy (OS, A–D) and TCGA-COAD-READ (DSS and DFS cohorts, E–L). Patients across the whole cohort were grouped by bacterium subgroup (low, in green, vs high, in orange; A,E,I) or mesenchymal status (CMS4 and/or CRIS-B, in light blue, vs remaining cases, in dark blue; B,F,J). Patients were grouped by bacterium group and further stratified by mesenchymal status (C,D,G,H,K,L). Patients were binned into a bacterium group (low vs high) using the 75th percentile as cut-off and based on either Fn load (Taxonomy cohort; A,C–D) or Fusobacteriales RA at the order level (TCGA-COAD-READ cohort; E,G–I,K,L). (M) Cox regression models fitted on bacterium RA reported at the order, family, genus and species taxonomic ranks. for each taxonomic rank, patients were classified as low or high subgroup using the corresponding 75th percentile RA abundance as cut-off. Univariate Cox regression models were fitted when evaluating the association between bacterium subgroup (high vs low, reference low) at each taxonomic rank and either DSS or DFS in the whole unselected patient population (left panel). Cox regression models with an interaction term between bacterium subgroup (high vs low, reference low) and mesenchymal status (mesenchymal, ie, either CMS4 and/or CRIS-B, vs non-mesenchymal, ie, neither CMS4 nor CRIS-B) at each taxonomic rank and either DSS or DFS was fitted to evaluate differential impact of bacterium on clinical outcome by tumour biology (right panels). CMS, consensus molecular subtyping; CRIS, colorectal cancer intrinsic subtyping; DFS, disease-free survival; DSS, disease-specific survival; Fn, Fusobacterium nucleatum; OS, overall survival; TCGA-COAD-READ, colon and rectal cases of The Cancer Genome Atlas.
Figure 6  Exploration of mechanism underlying differential impact of \( \text{Fn} / \text{Fusobacteriales} \) in mesenchymal versus non-mesenchymal tumours. (A) Scatterplot depicting p values derived by assessing with logistic regression models the relationship between genes/signatures associated with \( \text{Fusobacteriales} \) RA in univariate analysis (model 1, x-axis) or the interaction with mesenchymal status (model 2, y-axis). Gene/signature with statistically significant p values from model 2 are highlighted by a grey shaded area. (B) Breakdown of association including direction and effect size, in the unselected patients’ population and within mesenchymal versus non-mesenchymal cases. Only gene/signatures with significant interaction between \( \text{Fusobacteriales} \) RA and the gene/signature interaction with the molecular subtype (model 2, top quadrant, grey-shaded area) in the TCGA-COAD-READ cohort are included. Associations for both the TCGA-COAD-READ (\( \text{Fusobacteriales} \) RA) and Taxonomy (\( \text{Fn} \) load) cohorts are shown. Statistically significant associations are represented with circle markers, whereas non-significant associations are indicated by squared markers. (C) Association between gene/signature identified as candidate targets, A), and clinical outcome in patients of the TCGA-COAD-READ cohort with mesenchymal tumours. HRs and p values are derived from Cox regression models with an interaction term for \( \text{Fusobacteriales} \) relative abundance (low vs high) and each of the gene/signature (low vs high) being evaluated. CMS, consensus molecular subtyping; CRIS, colorectal cancer intrinsic subtyping; DFS, disease-free survival; DSS, disease-specific survival; \( \text{Fn} \), \( \text{Fusobacterium nucleatum} \); OS, overall survival; TCGA-COAD-READ, colon and rectal cases of The Cancer Genome Atlas.
Fn/Fusobacteriales has a specific clinical impact in mesenchymal-rich, high-stromal infiltrated tumours; this argues against a blanket approach for treating patients with Fn/Fusobacteriales-high tumours. Treatment with wide spectrum antibiotics reduces the growth of Fn-positive tumours in vivo. However, the use of antibiotics to treat Fn-positive CRC tumours may be limited as Fn penetrate deeply within tumour, immune and endothelial cells where they internalise with endosomes and lysosomes. In addition, long-term use of antibiotics can cause gut dysbiosis, which may impact disease progression and outcome.

Given that ‘it takes two to tango’, namely, a high pathogen prevalence and a conducive host milieu, we further examined this interdependence to identify druggable aberrations in the host signalling pathways and microenvironment. We identified putative targets related to (pro-)inflammation, inflammasome, activated T cells, complement system, metalloproteins and macrophage chemotaxis and activation. Fusobacteriales induce a constitutively activated NFκB-TNF-α-IL-6 state which results in activation of metalloproteins and inflammatory cytokines (CSF1-3) which mediate macrophage differentiation, inhibit cytotoxic immune cells and promote proliferation of myeloid-derived-suppressor (MDSC) cells. We observed an increase in inflammation and M1 macrophages and a decrease in M2 macrophages in patients with higher Fn/Fusobacteriales prevalence. We envisage that therapeutic options, such as NLRP3/AIM2 inflammasome suppression, IL-1β blockade, TNF-α or IL-6 inhibition, which have been approved for treatment of chronic inflammation and cytokine storm syndrome in multiple cancers, rheumatoid arthritis and COVID-19 may ameliorate the immunosuppressive microenvironment induced by Fn/Fusobacteriales. Importantly, these targets are involved in not only promoting an immunosuppressive microenvironment by recruiting tissue-associated macrophages (TAMs) and MDSCs, but also in orchestrating inflammation, angiogenesis, epithelial-to-mesenchymal transition and, ultimately, metastasis. The prometastatic impact of Fn/Fusobacteriales is further corroborated by findings in the literature linking higher pathogen prevalence in more advanced disease stage and metastasis in clinical specimens and higher metastatic burden in mice inoculated with Fn.

Cancer cells with an EMT phenotype secrete cytokines such as IL-10 and TGF-β that can further promote an immunosuppressive microenvironment. Additionally, secretion of IL-6 and IL-8 from stroma cells can further foster an EMT-associated fibroblast (CAFs) which, in turn, may promote angiogenesis and invasion. Taken together, these aberrations may result in a self-reinforcing mechanism that confers on cancer cells the ability to migrate, invade the extracellular matrix, extravasate and seed metastasis. When comparing the transcriptomic profiles by Fusobacteriales RA in the TCGA-COAD-READ cohort, we identified dysregulation affecting cell architecture involving apical surface dynamics and Aurora A kinase signalling, which regulate cMYC, DNA repair, cell motility/migration and induce EMT transition via β-catenin and TGF-β, leading to metastasis and resistance to treatment in multiple cancer types. Small molecule inhibitors against Aurora A have shown encouraging results in preclinical studies and clinical trials in CRC and other cancers. Cytoskeleton shape, filopodium protrusions and alterations in cell adhesion and structure are hallmarks of extracellular matrix invasion. EMT key effectors, SNAIL and ZEB1, alter apical surface dynamics by inhibiting scaffolding proteins and by inducing expression of matrix metalloproteins (MMP3 and MMP9), resulting in loosened tight junctions, altered cell polarity and increased plasticity which, in turn, enable cell invasion. Dysregulations in MMP expression may aid cancer cells that have reached the bloodstream to extravasate to distant tissues by priming the vascular endothelium via upregulation of VEGF-A and by increasing permeability via COX2 upregulation. Our analyses in the TCGA-COAD-READ cohort identified higher expression of vascular endothelial growth factor (VEGF) as well as an angio genesis signature in patients with higher Fusobacteriales RA. Indeed, a new generation of selective and highly penetrative MMP inhibitors is being trialled in GI cancers, and Mehta et al reported lower Fusobacteriales RA in subjects treated with aspirin, a COX2 inhibitor. Green et al demonstrated that MAPK7 is a master regulator of MMP9 and promotes the formation of metastasis. We observed a dysregulation in MAPK signalling at the protein level when comparing Fusobacteriales-high versus Fusobacteriales-low patients of the TCGA-COAD-READ cohort. MAPK7 induces EMT transition, cell migration and regulates TAM polarisation in a metalloprotein-dependent manner, rendering it an appealing upstream therapeutic target. IL-6 orchestrates MAPK-STAT3 signalling, which in turn regulates the dynamic transition between two CAFs subpopulations, EMT-CAFs and proliferation-CAFs, rendering the IL-6-TGF-β-EMT-CAFs cross-talk potentially a further therapeutic target. While directly targeting EMT via NOTCH or WNT has shown limited success in the clinic, microenvironment remodelling to reverse immunosuppression by inhibiting CXCL12 or promoting T-cell infiltration, or function via engineered oncolytic adenovirus, has shown promising results in reducing metastasis formation. Additionally, we observed a positive correlation between gene expression of IL-8, CXCL8, CXCR1 and CXCL10 and Fn/Fusobacteriales prevalence, corroborating findings from Casasanta et al assessing Fn in HCT116 CRC cells. In conclusion, our analyses have identified a patient subpopulation that has an unfavourable clinical outcome when their tumours exhibit mesenchymal traits and are highly positive with Fn/Fusobacteriales and pinpointed clinically actionable host-specific vulnerabilities that suggest new treatments for these patients that extend beyond broad-spectrum antibiotics.

Acknowledgements We thank the patients who kindly donated their samples and made this study possible.

Contributors MS, DL and JHMP conceptualised and designed the study. MS, NC, KS, SB, DL and JHMP were involved in collection, preparation, interpretation, validation and critical review of the data. NC and KS performed the cell culture experiments and quantified bacterium load in tumour samples of the patients of the Taxonomy cohort. MS performed formal analysis including bioinformatics analyses in the TCGA-AD-READ cohort and critical review of the data. NC and KS performed the cell culture experiments and quantified bacterium load in tumour samples of the patients of the Taxonomy cohort. MS performed formal analysis including bioinformatics and statistical analyses. MS created the manuscript figures and supplementary materials. MS and JHMP drafted the manuscript. All authors edited, reviewed, revised and approved the manuscript text. DL and JHMP acquired funding for the study.

Funding This study was supported by a grant from the Health Research Board and Science Foundation Ireland to JHMP (16/US3301), a studentship to KS sponsored by The Northern Ireland Department for the Economy (NI DFE), and by funding from NI DFE (SFI-16/IA/12582, STIL/5715/15). The results included here are in part based on data generated by the TCGA Research Network (https://wwwcancergov/tcga). We wish to acknowledge the information technology department at the Royal College of Surgeons in Ireland and the Irish Centre for High-End Computing (ICHEC) for the provision of computational facilities and support.

Competing interests None declared.

Patient consent for publication Not required.
Gut microbiota

Ethics approval The Taxonomy cohort collection was approved by the Medicine, Dentistry, and Biomedical Sciences School Ethics Committee (ref: 12/12v4), as previously described (Allen et al, JCO Precision Oncology, 2018; PMID: 30088816). Approval for the cohort of patients with colon (COAD) and rectal (READ) cases denoted as TCGA-COAD-READ in this manuscript was acquired by the original investigators of The Cancer Genome Atlas consortium.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository. Processing and analysis code along with bacterium estimates with corresponding clinical and molecular datasets for the Taxonomy and TCGA-COAD-READ cohorts included in this study are publicly available and archived at Zenodo (https://10.5281/zenodo.4019142). Bacterium estimates include Fusobacterium nucleatum load (Taxonomy cohort) and Fusobacteriales relative abundance, along with higher resolution estimates at genus, family and species taxonomic ranks, (TCGA-COAD-READ cohort).

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not be responsible for any error and/or omissions arising from translation and adaptation or otherwise.

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 license is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

ORCID Id Manuela Salvucci http://orcid.org/0000-0001-9941-4307

REFERENCES


Supplementary Figures for

“Patients with mesenchymal tumours and high *Fusobacteriales* prevalence have worse prognosis in colorectal cancer (CRC)”

Manuela Salvucci¹, Nyree Crawford², Katie Stott², Susan Bullman³⁴, Daniel B. Longley², and Jochen H.M. Prehn¹*

¹Centre for Systems Medicine, Department of Physiology and Medical Physics, Royal College of Surgeons in Ireland, Dublin, Ireland;

²Patrick G. Johnston Centre for Cancer Research, School of Medicine, Dentistry and Biomedical Science, Queen’s University Belfast, Northern Ireland, UK;

³Dana-Farber Cancer Institute, Harvard Medical School, Boston, USA;

⁴Fred Hutchinson Cancer Research Center, Human Biology Division, Seattle, USA.

**Corresponding author:** Prof. Jochen H. M. Prehn, Department of Physiology and Medical Physics, Royal College of Surgeons in Ireland, 123 St. Stephen’s Green, Dublin 2, Ireland. Tel.: +353-1-402-2255; Fax: +353-1-402-2447; E-mail: jprehn@rcsi.ie.

**Data and code availability:** Datasets and source code will be publicly available and archived upon publication at Zenodo (https://10.5281/zenodo.4019142).
Transcriptomic-dependent *Fn/Fusobacteriales* impact.

**Supplementary Figure 1.**

![Graph showing distribution of key player genes grouped by *Fn* load for patients of the Taxonomy cohort.]

**Association between Fn load and inflammation signalling in the human host.**

Distribution of key player genes grouped by *Fn* (high vs. low, using the 75th percentile as cut-off) for patients of the Taxonomy cohort.

Median and lower (25th) and upper (75th) percentiles are indicated by white solid or dashed lines, respectively. Statistical significance was evaluated Kruskal-Wallis tests and P-values are reported.
Transcriptomic-dependent *Fn/Fusobacteriales* impact.

**Supplementary Figure 2.**

**Association between *Fusobacteriales* relative abundance (RA) and human host clinico-pathological features in the TCGA-COAD-READ cohort.**

Mosaic plots depicting the relationship between categorical clinico-pathological characteristics of the human host and *Fusobacteriales* RA. Patients were classified as *Fusobacteriales*-low or -high using the 75th percentile as cut-off and indicated in green and orange, respectively. Statistical significance was evaluated with $\chi^2$ independence tests and the $\chi^2$ test statistic and the mod-log-likelihood P-values are reported.

**Abbreviations.** BMI: body mass index.

Salvucci et al.
Transcriptomic-dependent *Fn/Fusobacteriales* impact.

**Supplementary Figure 3.**

![Image of the figure showing the relationship between *Fn* species and other factors with statistical significance](image)

Salvucci et al.
Transcriptomic-dependent *Fn/Fusobacteriales* impact.

**Association between human host clinico-pathological (A) and mutational (B) features in *Fn*-low vs. high patients of the TCGA-COAD-READ (1) and Taxonomy (2) cohorts.**

*Fn* is expressed as relative abundance (RA) for patients of the TCGA-COAD-READ cohort or load for patients of the in-house Taxonomy cohort. Patients were categorised in low vs. high subgroups using the 75th percentile as cut-off and indicated in green and orange, respectively. Association between *Fn* and continuous variables is depicted with violin plots, median and lower (25th) and upper (75th) percentiles are indicated by white solid or dashed lines, respectively. Statistical significance was evaluated Kruskal-Wallis tests and P-values are reported. Association between *Fn* and categorical clinico-pathological characteristics is depicted with mosaic plots and statistical significance was evaluated with $\chi^2$ independence tests and the $\chi^2$ test statistic and the mod-loglikelihood P-values are reported.
Transcriptomic-dependent *Fn/Fusobacteriales* impact.

Supplementary Figure 4.

**TCGA-COAD-READ cohort**

**A-B.** Distribution of transitions (Ti) and transversions (Tv) (**A**) and conversion changes (**B**) in patients of the TCGA-COAD-READ cohort classified as *Fusobacteriales*-low (in green) or -high (in orange) based on a 75th percentile cut-off. Median and lower (25th) and upper (75th) percentiles are indicated by white solid or dashed lines, respectively. Statistical significance was evaluated using Kruskal-Wallis tests and P-values are reported.

**Association between Fusobacteriales relative abundance (RA) and DNA substitution mutations in the patients of TCGA-COAD-READ cohort.**

Salvucci et al.
Transcriptomic-dependent \textit{Fn/Fusobacteriales} impact.

Supplementary Figure 5.

\textbf{Recurrent copy number alterations in patients of the TCGA PanCancer cohort.}

Amplifications (in red, left hand-side) and deletions (in blue, right hand-side) computed by GISTIC2 analysis to detect recurrent copy number alterations in the TCGA PanCancer cohort (n=9142).

Chromosome bands are indicated (y axis) and cytobands that reached statistical significance (as indicated by q-values) are shown.

Salvucci et al.
Transcriptomic-dependent *Fn/Fusobacteriales* impact.

**Supplementary Figure 6.**

**Frequency of copy number alterations in patients of the TCGA-COAD-READ cohort.**

Frequency of occurrence of copy number amplifications (in red) or deletions (in blue) by chromosome in the whole unselected cohort (left panel) and in subgroups restricted to cases with low- (middle panel) or high- (right panel) *Fusobacteriales* relative abundance for patients of the TCGA-COAD-READ cohort.

Salvucci et al.
Transcriptomic-dependent Fn/Fusobacterales impact.

Supplementary Figure 7.

Pathway enrichment analysis for genes differentially expressed by Fusobacterales relative abundance (RA) in patients of the TCGA-COAD-READ cohort.

Enrichment analysis on genes identified as differentially expressed by Fusobacterales RA in patients of the TCGA-COAD READ cohort. Analysis was performed with EnrichR querying the BioCarta (version 2016) and NCI-Nature (version 2016) pathway databases. The number of identified altered genes for each pathway is encoded by the marker size and the magnitude of the associated P-values is color-coded, as indicated in the legend.
Transcriptomic-dependent *Fn/Fusobacteriales* impact.

**Supplementary Figure 8.**

Pathway enrichment analysis for proteins differentially expressed by *Fusobacteriales* relative abundance (RA) in patients of the TCGA-COAD-READ cohort.

Enrichment analysis on proteins identified as differentially expressed by *Fusobacteriales* RA in patients of the TCGA-COAD READ cohort. Analysis was performed with *EnrichR* querying the *BioCarta* (version 2016) and *NCI-Nature* (version 2016) pathway databases. The number of identified altered genes for each pathway is encoded by the marker size and the magnitude of the associated P-values is color-coded, as indicated in the legend.
Transcriptomic-dependent *Fn/Fusobacteriales* impact.

**Supplementary Figure 9.**

Kaplan-Meier plots comparing disease-free-survival (DFS) in patients of the Taxonomy cohort grouped by *Fn* load (A), mesenchymal status (B) and by *Fn* load within the non-mesenchymal and mesenchymal patients’ subpopulations (C-D). Patients were categorised in *Fn*-low or -high subgroups using the 75th percentile as cut-off. *Consensus Molecular Subtype* (CMS) and *Cancer Intrinsic Subtype* (CRIS) assignments were used to categorise patients in non-mesenchymal (“Neither CMS4 nor CRIS-B”) or mesenchymal, respectively (“Either CMS4 and/or CRIS-B”).
Transcriptomic-dependent \textit{F}n/\textit{Fusobacteriales} impact.

**Supplementary Figure 10.**

Kaplan-Meier plots comparing overall survival (OS) in patients of the TCGA-COAD-READ cohort grouped by \textit{Fusobacteriales} RA (A), mesenchymal status (B) and by \textit{Fusobacteriales} RA within the non-mesenchymal and mesenchymal patients’ subpopulations (C-D). Patients were categorised in \textit{Fusobacteriales}-low or -high subgroups using the 75\textsuperscript{th} percentile as cut-off. Consensus Molecular Subtype (CMS) and Cancer Intrinsic Subtype (CRIS) assignments were used to categorise patients in non-mesenchymal (“Neither CMS4 nor CRIS-B”) or mesenchymal, respectively (“Either CMS4 and/or CRIS-B”).

Salvucci et al.
Transcriptomic-dependent \textit{Fn/Fusobacterales} impact.

**Supplementary Figure 11.**

Cox regression models fitted on bacterium relative abundance reported at the order, family, genus, and species taxonomic ranks. For each taxonomic rank, patients were binned into -low or -high subgroups using the corresponding 75\textsuperscript{th} percentile RA as cut-off. Univariate Cox regression models were fitted when evaluating association between bacterium subgroup (high vs. low; reference low) at each taxonomic rank and OS in the whole unselected patient population (left panel). Cox regression models with an interaction term between bacterium subgroup (high vs. low; reference low) and mesenchymal status (mesenchymal, i.e either CMS4 and/or CRIS-B, vs. non-mesenchymal, i.e. neither CMS4 nor CRIS-B) at each taxonomic rank and OS were fitted to evaluate differential impact of bacterium on clinical outcome by tumour biology (right panels).
Transcriptomic-dependent *Fn/Fusobacteriales* impact.

**Supplementary Figure 12.**

Cox regression models fitted on patients of the TCGA-COAD-READ cohort with mesenchymal tumours (either CMS4 and/or CRIS-B) for each gene/signature identified from analysis presented in Fig. 6A. Patients were classified as *Fusobacteriales*-low or high using the corresponding 75th percentile relative abundance (RA) as cut-off. Univariate Cox regression models were fitted when evaluating association between *Fusobacteriales* (high vs. low; reference low) and OS in the whole unselected patient population (left panel). Cox regression models with an interaction term between *Fusobacteriales* (high vs. low; reference low) and gene/signature (high vs. low, reference low) and OS were fitted to evaluate differential impact of gene/signature on clinical outcome by

Salvucci et al.
Transcriptomic-dependent *Fn/Fusobacteriales* impact.

*Fusobacteriales* (right panels). * and ** denote interaction P-values lower than 0.05 and lower than or equal to 0.1, respectively.
Transcriptomic-dependent *Fn/Fusobacteriales* impact.

**Supplementary Figure 13.**

Cox regression models fitted on patients of the TCGA-COAD-READ cohort with mesenchymal tumours (either CMS4 and/or CRIS-B) for each gene/signature identified from analysis presented in Fig. 6A. Patients were classified as *Fusobacteriales*-low or high using the corresponding 75th percentile relative abundance (RA) as cut-off. Univariate Cox regression models were fitted when evaluating association between *Fusobacteriales* (high vs. low; reference low) and DSS in the whole unselected patient population (left panel). Cox regression models with an interaction term between *Fusobacteriales* (high vs. low; reference low) and gene/signature (high vs. low, reference low) and DSS were fitted to evaluate differential impact of gene/signature on clinical outcome by

Salvucci et al.
Transcriptomic-dependent *Fn/Fusobacteriales* impact.

*Fusobacteriales* (right panels). * and ** denote interaction P-values lower than 0.05 and lower than or equal to 0.1, respectively.
Transcriptomic-dependent *Fn/Fusobacteriales* impact.

**Supplementary Figure 14.**

Cox regression models fitted on patients of the TCGA-COAD-READ cohort with mesenchymal tumours (either CMS4 and/or CRIS-B) for each gene/signature identified from analysis presented in Fig. 6A. Patients were classified as *Fusobacteriales*-low or high using the corresponding 75\textsuperscript{th} percentile relative abundance (RA) as cut-off. Univariate Cox regression models were fitted when evaluating association between *Fusobacteriales* (high vs. low; reference low) and DFS in the whole unselected patient population (left panel). Cox regression models with an interaction term between *Fusobacteriales* (high vs. low; reference low) and gene/signature (high vs. low, reference low) and DFS were fitted to evaluate differential impact of gene/signature on clinical outcome by

Salvucci et al.
Transcriptomic-dependent *Fn/Fusobacteriales* impact.

*Fusobacteriales* (right panels). * and ** denote interaction P-values lower than 0.05 and lower than or equal to 0.1, respectively.
Captions for Supplementary Tables for

“Patients with mesenchymal tumours and high *Fusobacteriales* prevalence have worse prognosis in colorectal cancer (CRC)”

Manuela Salvucci¹, Nyree Crawford², Katie Stott², Susan Bullman³⁴, Daniel B. Longley², and Jochen H.M. Prehn¹

¹Centre for Systems Medicine, Department of Physiology and Medical Physics, Royal College of Surgeons in Ireland, Dublin, Ireland;
²Patrick G. Johnston Centre for Cancer Research, School of Medicine, Dentistry and Biomedical Science, Queen’s University Belfast, Northern Ireland, UK;
³Dana-Farber Cancer Institute, Harvard Medical School, Boston, USA;
⁴Fred Hutchinson Cancer Research Center, Human Biology Division, Seattle, USA.

**Corresponding author:** Prof. Jochen H. M. Prehn, Department of Physiology and Medical Physics, Royal College of Surgeons in Ireland, 123 St. Stephen’s Green, Dublin 2, Ireland. Tel.: +353-1-402-2255; Fax: +353-1-402-2447; E-mail: jprehn@rcsi.ie.

**Data and code availability:** Datasets and source code will be publicly available and archived upon publication at Zenodo ([https://10.5281/zenodo.4019142](https://10.5281/zenodo.4019142)).
Transcriptomic-dependent \textit{Fn/Fusobacteriales} impact.

**Supplementary Table 1.**

Clinico-pathological and demographic characteristics of the CRC patients included in this study (“Overall”) and grouped by cohort, namely “in house Taxonomy” and “TCGA-COAD-READ”. For continuous variables, median, interquartile range, and statistical significance (P-value) determined by Kruskal-Wallis tests are reported. For categorical values, number, and percentage of cases by level and statistical significance (P-value) determined by $\chi^2$ tests are reported.

**Supplementary Table 2.**

Patient-level bacterium data for cases of the Taxonomy and TCGA-COAD-READ cohort. \textit{Fn} load measured by qPCR for patients of the Taxonomy cohort is available in the sheet “Taxonomy cohort (n=140)”. Relative abundance of \textit{Fusobacteriales} and higher resolution taxonomic ranks (family, genus and species) including the \textit{Fn} species for the patients in the TCGA-COAD-READ cohort is available in the sheet “TCGA-COAD-READ cohort (n=605)”. For the TCGA-COAD-READ cohort, genera/species with an average relative abundance lower than 0.05 were aggregated as “Other”.

**Supplementary Table 3.**

Association between mutational status and \textit{Fusobacteriales} relative abundance in the TCGA-COAD-READ patients. Statistical significance was assessed by $\chi^2$ independence tests and $\chi^2$ statistics, unadjusted- and FDR-corrected mod-likelihood P-values are reported for each mutation that was either selected \textit{a priori} or was found to be statistically significant altered when comparing \textit{Fusobacteriales}-low vs. -high patients (75$^{\text{th}}$ percentile cut-off) of the TCGA-COAD-READ cohort.

Salvucci et al.
Transcriptomic-dependent *Fn/Fusobacteriales* impact.

**Supplementary Table 4.**

Association between recurrent copy number aberrations identified by GISTIC analysis when comparing *Fusobacteriales*-low vs. -high patients (75th percentile cut-off) of the TCGA-COAD-READ cohort.

**Supplementary Table 5.**

Association between gene expression profiles and *Fusobacteriales* relative abundance in the TCGA-COAD-READ patients was assessed by Spearman correlation. Correlation coefficient R, corresponding 95% confidence intervals, unadjusted- and FDR-corrected P-values are reported for each protein that was found to be statistically significant altered in the TCGA-COAD-READ cohort.

**Supplementary Table 6.**

Association between protein expression profiles and *Fusobacteriales* relative abundance in the TCGA-COAD-READ patients was assessed by Spearman correlation. Correlation coefficient R, corresponding 95% confidence intervals, unadjusted- and FDR-corrected P-values are reported for each protein that was found to be statistically significant altered in the TCGA-COAD-READ cohort.

**Supplementary Table 7.**

Un-adjusted and adjusted Cox regression models for patients of the TCGA-COAD-READ cohort. Cox regression models were fitted with an interaction term between *Fusobacteriales* (high vs. low, using the 75th percentile relative abundance as cut-off) and mesenchymal status.
Transcriptomic-dependent *Fn/Fusobacteriales* impact.

(mesenchymal vs. non-mesenchymal). Adjusted model 1 and 2 were fitted including precision variables. Model 1 used as adjustment covariates key clinical-pathological characteristics, namely age (continuous), stage (categorical, I to IV), tumour location (categorical, colon vs. rectum) and sex (categorical, male vs. female). Model 2 used as adjustment covariates a more extensive set (i.e. super-set) of clinico-pathological characteristics additionally including history of colon polyps (categorical, yes vs. no) and history of other malignancy as comorbidities.

**Supplementary Table 8.**

Detailed statistical output (coefficients and P-values) of logistic models 1 (*Fusobacteriales*–gene/signature) and 2 (*Fusobacteriales*–gene/signature:molecular subtype) fitted for a set of hypothesis-driven gene/signature profiles in patients of the TCGA-COAD-READ cohort presented in Fig. 6A. Table include all genes/signatures tested (regardless of statistical significance) reported in ascending order of interaction P-values from model 2 determined in the TCGA-COAD-READ cohort (discovery cohort). For completeness, detailed statistical output is also reported for the Taxonomy cohort.

Salvucci et al.
Supplementary Materials and Methods for “Patients with mesenchymal tumours and high Fusobacteriales prevalence have worse prognosis in colorectal cancer (CRC)”

Manuela Salvucci¹, Nyree Crawford², Katie Stott², Susan Bullman³, Daniel B. Longley², and Jochen H.M. Prehn¹*

¹Centre for Systems Medicine, Department of Physiology and Medical Physics, Royal College of Surgeons in Ireland, Dublin, Ireland;
²Patrick G. Johnston Centre for Cancer Research, School of Medicine, Dentistry and Biomedical Science, Queen’s University Belfast, Northern Ireland, UK;
³Dana-Farber Cancer Institute, Harvard Medical School, Boston, USA;
⁴Fred Hutchinson Cancer Research Center, Human Biology Division, Seattle, USA.

Corresponding author: Prof. Jochen H. M. Prehn, Department of Physiology and Medical Physics, Royal College of Surgeons in Ireland, 123 St. Stephen’s Green, Dublin 2, Ireland. Tel.: +353-1-402-2255; Fax: +353-1-402-2447; E-mail: jprehn@rcsi.ie.

Data and code availability: Datasets and source code will be publicly available and archived upon publication at Zenodo (https://10.5281/zenodo.4019142).
Transcriptomic-dependent \textit{Fn/Fusobacteriales} impact.

**Contents**

\textit{In vitro} experiments .................................................................................................................. 3

Cell culture.............................................................................................................................................. 3

\textit{Fn} culturing conditions .................................................................................................................... 3

Co-culture experiments ............................................................................................................................. 3

Western Blotting....................................................................................................................................... 3

NFκB activity assay.................................................................................................................................... 4

Quantitative polymerase chain reaction (qPCR)....................................................................................... 4

Association between \textit{Fusobacteriales} and \textit{Fn} prevalence in tumour resections with host
characteristics in CRC ............................................................................................................................... 4

Clinical cohorts......................................................................................................................................... 4

Taxonomy cohort....................................................................................................................................... 5

TCGA COAD-READ cohorts ..................................................................................................................... 5

Determination of \textit{Fn} load and \textit{Fusobacteriales} relative abundance in tumour resections of
CRC patients .............................................................................................................................................. 9

Taxonomy cohort....................................................................................................................................... 9

TCGA-COAD-READ cohort ..................................................................................................................... 9

Gene expression analysis .......................................................................................................................... 11

Transcriptomic-based signatures ............................................................................................................. 11

Markers for pro- and anti-inflammatory processes .............................................................................. 13

Characterization of the tumour microenvironment ............................................................................... 14

Patients’ classification into transcriptomic-based molecular subtypes................................................... 15

Unbiased and systematic analysis of human host associations with \textit{Fusobacteriales} in the
TCGA-COAD-READ cohort .................................................................................................................... 16

Mutational status ...................................................................................................................................... 16

Copy number alterations (CNAs)............................................................................................................. 17

Aberrations in transcriptional and protein profiles.................................................................................. 18

Exploration of putative mechanisms underlying differential impact of \textit{Fn/Fusobacteriales}
prevalence by tumour biology ............................................................................................................... 18

Statistical analysis .................................................................................................................................... 19

Comparative analyses ............................................................................................................................. 19

Outcome analysis ..................................................................................................................................... 20

Software and libraries ............................................................................................................................. 22

References .................................................................................................................................................. 23
Transcriptomic-dependent *Fn/Fusobacteriales* impact.

**In vitro experiments**

**Cell culture**

HCT116 and HT29 cells were purchased as authenticated stocks from ATCC (Teddington, UK). HT29 cells were cultured in DMEM medium (ThermoFisher Scientific Inc.) supplemented with 10% fetal bovine serum (Invitrogen, Paisley, UK). HCT116 cells were cultured in McCoy’s 5A medium (ThermoFisher Scientific Inc.) supplemented with 10% fetal bovine serum (Invitrogen, Paisley, UK). Cell lines were screened for the presence of mycoplasma utilising MycoAlert Mycoplasma Detection Kit (Lonza) monthly and cultured for no more than 20 passages.

**Fn culturing conditions**

*Fusobacterium nucleatum* subsp. *nucleatum* strain 25586 was purchased from American Type Culture Collection (ATCC, Middlesex, UK). *Fn* was cultured at 37ºC under anaerobic conditions (DG250, Don Whitley Scientific, West Yorkshire, UK) in Fastidious Anaerobic Broth (Neogen, formerly Lab M, Scotland, UK).

**Co-culture experiments**

HT29 and HCT116 cells were co-cultured with *Fn* at a Multiplicity of Infection (MOI) of 10:1, 100:1 and 1000:1 under normal culturing conditions for the CRC cell lines.

**Western Blotting**

Western blotting analysis was carried out as previously described [1]. IκBα antibody (#9242) was supplied by Cell Signaling Technology (Danvers, MA) and β-actin (#A5316) was supplied by Sigma.
Transcriptomic-dependent *Fn/Fusobacteriales* impact.

**NFκB activity assay**

Cells were co-transfected with NFκB luciferase reporter and Renilla constructs using X-tremeGENE HP (Promega, Madison, WI), as previously described [2]. Cells were lysed with Passive Lysis Buffer (Promega, Madison, WI) and Luciferase and Renilla activity assessed by luminescence using D-Luciferin and Colenterazine as substrates.

**Quantitative polymerase chain reaction (qPCR)**

RNA was extracted, according to manufacturer’s instructions using the High Pure RNA Isolation kit (Roche, Burgess Hill, UK). The Transcriptor First Strand cDNA synthesis kit (Roche, Burgess Hill, UK) was utilized to synthesize cDNA, according to manufacturer’s instructions. qPCR was performed on the LC480 light cycler, using Syber green, according to manufacturer’s instructions. Primer sequences:

- **TNFα F**: CAGCCTCTTCTCCTTCTGAT;
- **TNFα R**: GCCAGAGGGCTGATTAGAGA;
- **β-tubulin F**: CGCAGAAGAGGAGGAGATT;
- **β-tubulin R**: GAGGAAAGGGGCAGTTGAGT.

**Association between *Fusobacteriales* and *Fn* prevalence in tumour resections with host characteristics in CRC**

**Clinical cohorts**

In this study, we profiled *Fusobacteriales* and/or *Fn* in primary tumour tissue resections from n=645 CRC patients from an in-house (Taxonomy, [3-4]) and a public protected dataset (The
Transcriptomic-dependent \( Fn/\text{Fusobacterales} \) impact.

Cancer Genome Atlas, TCGA-COAD-READ). Demographic and clinical and pathological characteristics of the two cohorts are compared and contrasted in Suppl. Table 1, which was generated with the python package TableOne [5].

**Taxonomy cohort**

Stage II and III colorectal patients (n=156) from a multi-centre study (St Vincent’s Hospital, Dublin, IE; University Hospital Vall d’Hebron, Barcelona, ES; University of Aberdeen, UK; University of Florence, IT) were accrued, as previously described (Taxonomy cohort, [3]). The cohort collection was approved by the Medicine, Dentistry, and Biomedical Sciences School Ethics Committee (ref: 12/12v4), as previously described [3]. In downstream analyses, we included patients with available gene expression profiling (Almac Xcel array, Almac Diagnostics, Craigavon, UK, GSE103479, [3-4]) and estimation of \( Fn \) load from resected tumour tissue (at least 50% tumour content) by qPCR (n=140). The primary outcome for the Taxonomy cohort was overall survival (OS), but disease-free survival (DFS) records were also available.

**TCGA COAD-READ cohorts**

Stage I to IV patients with cancer of the colon (COAD) or rectum (READ) accrued by The Cancer Genome Atlas (TCGA) network with available fresh frozen tumour resections of sufficient quality and quantity for sequencing analysis (https://www.cancer.gov/about-neci/organization/ccg/research/structural-genomics/tcga/studied-cancers) were considered for inclusion in the study (n=629). In downstream analyses, we included all patients (n=605) that i) where not listed as “Redacted” in the clinical metadata retrieved from Liu et al. [6]; and ii) had at least a high quality RNASeq experiment from primary tumour from which bacterial relative abundance could be estimated (Supplementary Materials and Methods Figure 1).
Transcriptomic-dependent *Fn/Fusobacteriales* impact.

Throughout this study we investigated the relationship between the relative abundance of *Fusobacteriales* and higher resolution taxonomic ranks, including the *Fn* species, and characteristics of the host using several signatures and -omic views, namely mutations, copy number aberrations, gene and protein expression, (described in detail in the following sections).

**Supplementary Materials and Methods** Fig. 2 depicts data (cross-)availability and highlights what set of patients was included in each analysis.
Transcriptomic-dependent *Fn/Fusobacteriales* impact.

**Supplementary Materials and Methods Figure 1.** Flowchart depicting inclusion criteria with corresponding number of samples/patients available in the TCGA-COAD-READ cohort at each step of the analysis.
Transcriptomic-dependent *Fn/Fusobacterales* impact.

**Supplementary Materials and Methods Figure 2.** (Cross-)availability of *Fusobacterales* estimates (and higher resolution taxonomic ranks, including the *Fn* species), clinical and primary and derived -omic data for the TCGA-COAD-READ patients included in this study.
Transcriptomic-dependent *Fn/Fusobacteriales* impact.

**Determination of *Fn* load and *Fusobacteriales* relative abundance in tumour resections of CRC patients**

**Taxonomy cohort**

*Fn* abundance was quantified through qPCR analysis from tumour DNA, performed on the Roche Light Cycler 480 Real Time PCR Instrument (Roche, Burgess Hill, UK), using Syber green, according to manufacturer’s instructions. Each reaction contained 80 ng of genomic DNA which was assessed in duplicate, in 25 μl reactions. The abundance of *Fn* DNA in each tumour sample was normalised to the human reference gene Prostaglandin transporter (PGT) using the $2^{-ΔCt}$ method, where $ΔCt = Ct$ value for *Fn* – Ct value for PGT. Primer sequences:

- **Fn F**: CAACCATTACTTTAACTCTACCATGTTCA;
- **Fn R**: GTTGACTTTACAGAAGGAGATTATGTAAAAATC;
- **PGT F**: ATCCCCAAAGCACCTGGTTT;
- **PGT R**: AGAGGCCAAGATAGTCCTGGTAA.

**TCGA-COAD-READ cohort**

*Fusobacteriales* relative abundance in primary tumour specimens was estimated from RNASeq using a subtractive method implemented by the *PathSeq* pipeline (version 2, *PathSeqPipelineSpark* routine, [7-8]), powered by the Genome Analysis Toolkit engine (GATK, https://gatk.broadinstitute.org/, [9]) and the Apache Spark framework. Level 1 protected BAM sequencing files from RNASeq experiments for all TCGA-COAD-READ patients were accessed via the GDC Data Portal (https://portal.gdc.cancer.gov/) and served as input to the pipeline. Briefly, host reads (i.e. human) were filtered out and the remaining unmapped reads were aligned...
Transcriptomic-dependent *Fn/Fusobacterales* impact.

to microbial reads based on reference taxonomies for bacteria, fungi and viruses using a (default) min-clipped-read-length of 31. Host and microbe references files were retrieved from the GATK Resource Bundle (ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/pathseq/). We ran the *PathSeq* pipeline on n=698 patient samples of which n=644 were from tumour tissue. We restricted the analysis to samples which exceeded 10 million primary reads, resulting in n=630 high quality tumour samples for downstream analysis. Next, we collapsed microbial relative abundance from multiple samples and multiple tissue types (primary, recurrent and metastatic) of the same patient by mean. In downstream analyses, we included only patients with samples resected from primary tumours (n=605). We reported relative abundance for *Fusobacterales* at the order, family, genus and species taxonomic rank as normalized score expressed as percentage of the total relative abundance of the bacterial kingdom. Some of the species, denoted by the suffix "_sp," such as *Fusobacterium_sp._CM1*, reported by *PathSeq* are sub-species/strains. This may lead to under-reporting the relative abundance of e. g. *Fusobacterium nucleatum* as it does not include the abundances of its sub-species/strains. To avoid this issue, we manually re-mapped sub-species/strains to their parent species by blasting their sequence in NCBI (https://www.ncbi.nlm.nih.gov/nuccore/). We performed the re-mapping only when the percentage of identity between the sub-species/strain and its parent species exceeded 97%, as indicated in **Supplementary Materials and Methods** Table 1. The majority of the sub-species/strains mapped to *Fn.*
Transcriptomic-dependent *Fn/Fusobacterales* impact.

**Supplementary Materials and Methods Table 1.** Sub-species/strain mapping to parent species.

<table>
<thead>
<tr>
<th>Sub-species/strain</th>
<th>Candidate parent species</th>
<th>Per. identity</th>
<th>Remapped parent species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cetobacterium_sp._ZOR0034</td>
<td>Cetobacterium_somerae</td>
<td>100%</td>
<td>Cetobacterium_somerae</td>
</tr>
<tr>
<td>Cetobacterium_sp._ZWU0022</td>
<td>Cetobacterium_somerae</td>
<td>99.78%</td>
<td>Cetobacterium_somerae</td>
</tr>
<tr>
<td>Fusobacterium_sp._CM1</td>
<td>Fusobacterium_nucleatum</td>
<td>99.86%</td>
<td>Fusobacterium_nucleatum</td>
</tr>
<tr>
<td>Fusobacterium_sp._CM21</td>
<td>Fusobacterium_nucleatum</td>
<td>99.86%</td>
<td>Fusobacterium_nucleatum</td>
</tr>
<tr>
<td>Fusobacterium_sp._CM22</td>
<td>Fusobacterium_nucleatum</td>
<td>99.70%</td>
<td>Fusobacterium_nucleatum</td>
</tr>
<tr>
<td>Fusobacterium_sp._HM5064B11</td>
<td>Fusobacterium_nucleatum</td>
<td>99.93%</td>
<td>Fusobacterium_nucleatum</td>
</tr>
<tr>
<td>Fusobacterium_sp._HM64B12</td>
<td>Fusobacterium_nucleatum</td>
<td>99.82%</td>
<td>Fusobacterium_nucleatum</td>
</tr>
<tr>
<td>Fusobacterium_sp._HM65F01</td>
<td>Fusobacterium_nucleatum</td>
<td>99.87%</td>
<td>Fusobacterium_nucleatum</td>
</tr>
<tr>
<td>Fusobacterium_sp._OBRC1</td>
<td>Fusobacterium_nucleatum</td>
<td>100%</td>
<td>Fusobacterium_nucleatum</td>
</tr>
<tr>
<td>Fusobacterium_sp._HM073F01</td>
<td>Fusobacterium_varium</td>
<td>100%</td>
<td>Fusobacterium_varium</td>
</tr>
<tr>
<td>Leptotrichia_sp._Marseille-P3007</td>
<td>Leptotrichia_buccalis</td>
<td>98.37%</td>
<td>Leptotrichia_buccalis</td>
</tr>
<tr>
<td>Leptotrichia_sp._oral_taxon_225</td>
<td>Leptotrichia_trevisanii</td>
<td>99.52%</td>
<td>Leptotrichia_trevisanii</td>
</tr>
<tr>
<td>Leptotrichia_sp._oral_taxon_879</td>
<td>Leptotrichia_longgongensis?</td>
<td>96.86%</td>
<td>un-mapped</td>
</tr>
<tr>
<td>Leptotrichia_sp._oral_taxon_212</td>
<td>Leptotrichia_longgongensis?</td>
<td>92.67%</td>
<td>un-mapped</td>
</tr>
<tr>
<td>Leptotrichia_sp._oral_taxon_847</td>
<td>Leptotrichia_massiliensis?</td>
<td>92.04%</td>
<td>un-mapped</td>
</tr>
<tr>
<td>Leptotrichia_sp._oral_taxon_215</td>
<td>No candidate parent species</td>
<td>un-mapped</td>
<td>un-mapped</td>
</tr>
<tr>
<td>Fusobacterium_sp._oral_taxon_370</td>
<td>Fusobacterium_nucleatum or</td>
<td>95.33%</td>
<td>un-mapped</td>
</tr>
<tr>
<td></td>
<td>Fusobacterium_periodonticium?</td>
<td>both</td>
<td></td>
</tr>
</tbody>
</table>

**Gene expression analysis**

For the Taxonomy cohort, transcriptomics data (Almac Xcel array, Almac Diagnostics, Craigavon, UK; GSE103479) were processed as previously described [3-4]. For the TCGA-COAD-READ cohort, level 4 batch-corrected and normalised gene expression profiles by RNASeq were retrieved from the TCGA PanCanAtlas data-freeze release (*EBPlusPlusAdjustPANCAN_IlluminaHiSeq_RNASeqV2.geneExp.tsv*) from [https://gdc.cancer.gov/about-data/publications/pancanatlas](https://gdc.cancer.gov/about-data/publications/pancanatlas).

**Transcriptomic-based signatures**

We reviewed the literature and selected signatures encoding signalling pathways of interest including:
Transcriptomic-dependent *Fn/Fusobacteriales* impact.

- **proliferation**: mean gene expression of BIRC5, CCNB1, CDC20, NUF2, CEP55, NDC80, MKI67, PTTG1, RRM2, TYMS, and UBE2C ([10]).

- **epithelial-to-mesenchymal transition (EMT)**: difference in gene expression of epithelial (CDH1, DSP, OCLN) and mesenchymal (VIM, CDH2, FOXC2, SNAI1, SNAI2, TWIST1, FN1, ITGB6, MMP2, MMP3, MMP9, SOX10, GCS) genes ([11]).

- **metastasis**: difference in gene expression of markers promoting (SNRPF, EIF4EL3, HNRPAB, DHPS, PTTG1, COL1A1, COL1A2, and LMNB1) and inhibiting (ACTG2, MYLK, MYH11, CNN1, HLA-DPB1, RUNX1, MT3, NR4A1, and RBM5) metastasis ([12]).

- **DNA damage**: mean gene expression of PRKDC, NEIL3, FANCD2, BRCA2, EXO1, XRCC2, RFC4, USP1, UBE2T, and FAAP24 ([13]).

- **WNT signalling**: mean gene expression of AC023512.1, APC, APC2, AXIN1, AXIN2, BTRC, CACYBP, CAMK2A, CAMK2B, CAMK2G, CCND1, CCND2, CCND3, CER1, CHD8, CHP1, CHP2, CREBBP, CSNK1A1, CSNK1A1L, CSNK1E, CSNK2A1, CSNK2A2, CSNK2B, CTBP1, CTBP2, CTNNB1, CTNNBIP1, CUL1, CXXC4, DAAM1, DAAM2, DKK1, DKK2, DKK4, DVL1, DVL2, DVL3, EP300, FBXW11, FOSL1, FRAT1, FRAT2, FZD1, FZD10, FZD2, FZD3, FZD4, FZD5, FZD6, FZD7, FZD8, FZD9, GSK3B, JUN, LEF1, LRP5, LRP6, MAP3K7, MAPK10, MAPK8, MAPK9, MMP7, MYC, NFAT5, NFATC1, NFATC2, NFATC3, NFATC4, NFKB1, NFKB2, NLK, PLCB1, PLCB2, PLCP3, PLCP4, PORCN, PPARD, PPP2CA, PPP2CB, PPP2R1A, PPP2R1B, PPP2R5A, PPP2R5B, PPP2R5C, PPP2R5D, PPP2R5E, PPP3CA, PPP3CB, PPP3CC, PPP3R1, PPP3R2, PRICKLE1, PRICKLE2, PRKACA, PRKACB, PRKACG, PRKCA, PRKCB, PRKCG, PRKX, PSEN1, RAC1, RAC2, RAC3, RBX1, RHOA,
Transcriptomic-dependent *Fn/Fusobacteriales* impact.

ROCK1, ROCK2, RUVBL1, SENP2, SFRP1, SFRP2, SFRP4, SFRP5, SIAH1, SKP1, SMAD2, SMAD3, SMAD4, SOX17, TBL1X, TBL1XR1, TBL1Y, TCF7, TCF7L1, TCF7L2, TP53, VANGL1, VANGL2, WIFI1, WNT1, WNT10A, WNT10B, WNT11, WNT16, WNT2, WNT2B, WNT3, WNT3A, WNT4, WNT5A, WNT5B, WNT6, WNT7A, WNT7B, WNT8A (https://www.gsea-msigdb.org/gsea/msigdb/cards/KEGG_WNT_SIGNALING_PATHWAY).

- **Tumour Inflammation Signature (TIS):** mean gene expression of CD276, HLA-DQA1, CD274, IDO1, HLA-DRB1, HLA-E, CMKLR1, PDCD1LG2, PSMB10, LAG3, CXCL9, STAT1, CD8A, CCL5, NKG7, TIGIT, CD27, and CXCR6 ([14]).

- **Cytolytic activity:** mean gene expression of GZMA, and PRF1 ([15]).

- **Interferon gamma (IFNγ):** mean expression of IFNG, LAG3, CXCL9, and CD274 ([16]).

For both cohorts, we applied a robust scaling transformation (*sklearn.preprocessing.RobustScaler*) prior to computing the signatures. For the TCGA-COAD-READ cohort, gene expression profiles were quantile transformed (*sklearn.preprocessing.QuantileTransformer*) with the *output_distribution* flag set to *normal* prior to robust scaling.

**Markers for pro- and anti-inflammatory processes**

We selected NFKB1, TNF, IL6 and IL8 as key inflammatory markers to include in the analysis presented in Fig. 4G-H. Additionally, we performed a literature search and identified markers specific for pro- [17] and anti-inflammation [18] processes to further include in our analysis (Fig. 4G-H).
Transcriptomic-dependent *Fn/Fusobacteriales* impact.

**Characterization of the tumour microenvironment**

Cell type composition was computationally deconvoluted from bulk tumour gene expression data using 2 methods: *Microenvironment Cell Populations-counter* (MCP-counter, [19]); and *quantification of the Tumor Immune contexture from human RNA-seq data* (quanTIseq, [20]). MCP-counter, implemented as *R* package, uses marker genes to estimate the abundance (in arbitrary units) of endothelial cells, fibroblasts and 8 immune cell types including T cells, CD8+ T cells, cytotoxic lymphocytes, B lineage, natural killer (NK) cells, monocytic lineage, myeloid dendritic cells and neutrophils. For the Taxonomy cohort, we computed MCP-counter estimates as previously reported [4] and we normalized the resulting scores using a robust scaler ([*sklearn.preprocessing.RobustScaler*]). For the TCGA-COAD-READ cohort, we applied a quantile-transform ([*sklearn.preprocessing.QuantileTransformer*] with optimal distribution set to normal) followed by robust scaling ([*sklearn.preprocessing.RobustScaler*]) prior to applying the MCP-counter algorithm. Cell type composition was further characterized by applying the quanTIseq pipeline (step 3 in *quanTIseq_pipeline.sh* from [https://icbi.i-med.ac.at/software/quantiseq/doc/downloads/quanTIseq_pipeline.sh](https://icbi.i-med.ac.at/software/quantiseq/doc/downloads/quanTIseq_pipeline.sh)) to gene expression profiles of the Taxonomy ([4], flag set to account for the microarray nature of the data) or TCGA-COAD-READ cohort (*EBPlusPlusAdjustPANCAN_IlluminaHiSeq_RNASEqV2.geneExp.tsv*) without any additional pre-processing transformation. The quanTIseq algorithm uses a signature matrix to determine the fraction of tumour and stromal cells along with 10 immune cell types including non-regulatory CD4+ T cells, CD8+ T cells, regulatory T cells, dendritic cells, B cells, NK cells, neutrophils, monocytes, and classically- (M1) and alternatively- (M2) activated macrophages.
Transcriptomic-dependent *Fn/Fusobacteriales* impact.

**Patients’ classification into transcriptomic-based molecular subtypes**

Patients’ tumour samples were classified according to the *Consensus Molecular Subtype* (CMS, [21]) and *Cancer Intrinsic Subtype* (CRIS, [22]).

Circa 20% of primary tumour samples cannot be classified as CMS1 to CMS4 and they are marked as “no label” (NOLBL, [21]). In order to maximize the number of patients with CMS assignments, patients were classified in CMS groups using the nearest prediction from the random forest (RF) classifier (*R* package `CMSclassifier`, https://github.com/Sage-Bionetworks/CMSclassifier, [21]). For the Taxonomy cohort, we used the labels previously reported by McCorry *et al.* [4]. Similarly, for the TCGA-COAD-READ cohort, we retrieved the RF nearest prediction labels provided by Guinney *et al.* ([21], `cms_labels_public_all.txt` from synapse #: syn4978511). Additionally, we computed nearest prediction RF labels for the whole TCGA-COAD-READ cohort *de novo* to classify patients. We additionally included the CMS assignments for those patients that had not been subtyped as part of the Guinney *et al.* study. For both cohorts, subtype assignments mapping to multiple CMS classes were classified as indetermined and, thus, set to NOLBL.

Patients were subjected to CRIS subtyping and labelled as CRIS-A to CRIS-E or NOLBL (if Benjamini-Hochberg–corrected false discovery rate (BH.FDR) exceeded 0.2), as described in Isella *et al.* [22]. For the Taxonomy cohort, CRIS subtyping was performed using the nearest template prediction (NTP) classifier, available from GenePattern (https://genepattern.broadinstitute.org/gp/pages/login.jsf) as reported by McCorry *et al.* [4]. For the TCGA-COAD-READ cohort, we apply the CRIS subtyping to the whole TCGA-COAD-READ cohort. For the final CRIS assignments, we included either the labels provided from the
Transcriptomic-dependent *Fn/Fusobacteriales* impact.

Isella *et al.* publication [22] or the labels we computed *de novo* for patients that had not been subtyped as part of the original study.

**Unbiased and systematic analysis of human host associations with *Fusobacteriales* in the TCGA-COAD-READ cohort**

**Mutational status.**

Genomic intra-tumour heterogeneity and mutational burden expressed as number of silent and non-silent mutations per Mb was retrieved from the supplementary materials of Thorsson *et al.* [23] and corresponding data-freeze (*mutation-load_updated.txt* from [https://gdc.cancer.gov/about-data/publications/panimmune](https://gdc.cancer.gov/about-data/publications/panimmune)), respectively. Patients were classified as microsatellite stable (MSS) or unstable (MSI) using a cut-off of 0.4 applied to the MANTIS score retrieved from the supplementary materials of Bonneville *et al.* [24].

Somatic mutation data in Mutation Annotation Format (MAF, *mc3.v0.2.8.PUBLIC.maf.gz*) were retrieved from the TCGA PanCanAtlas data-freeze release ([https://gdc.cancer.gov/about-data/publications/pancanatlas](https://gdc.cancer.gov/about-data/publications/pancanatlas)) and restricted to the subset of patients diagnosed with COAD-READ cancers. We used the *maftools* *R* package (version 2.2.10, [25]) to compute conversion changes (*C>A*, *C>G*, *C>T*, *T>C*, *T>A*, *T>G*) and the percentage of transitions (*Ti*) and transversions (*Tv*) from the MAF file.

For each patient and each gene, we extracted from the MAF file the number of detected mutational aberrations. As aberrations, we included frame shift deletions and insertions, in frame deletions and insertions, missense and nonsense mutations and splice sites and we excluded the following variants: 3’ flank, 3’ UTR, 5’ flank, 5’ UTR, Intron, RNA, silent and non-stop mutations.
Transcriptomic-dependent *Fn/Fusobacteriales* impact.

Association between *Fusobacteriales* relative abundance (low vs. high using 75\textsuperscript{th} percentile as cut-off) and mutational status (number of aberrations) was assessed with $\chi^2$ independence tests. We restricted the analysis to genes with aberrations in at least 5\% of patients (n=818 genes out of 21332, ~4\%). We reported mod-log-likelihood P-values, adjusted for multiple comparisons with Benjamini-Hochberg FDR correction (Fig. 3C-D and Suppl. Table 3). Similarly, association between *Fn* and mutational status was assessed with $\chi^2$ independence tests in the TCGA-COAD-READ and Taxonomy cohorts (Suppl. Fig. 3). *Fn* refers to either relative abundance or load for the TCGA-COAD-READ and Taxonomy cohorts, respectively. Patients of the TCGA-COAD-READ cohort were considered wild-type for the gene of interest if the number of considered aberrations was null, mutant otherwise. Assessment of mutational status in the Taxonomy cohort has been previously described [3].

**Copy number alterations (CNAs)**

Copy number alterations ([broad.mit.edu_PANCAN_Genome_Wide_SNP_6_whitelisted.seg](https://broad.mit.edu_PANCAN_Genome_Wide_SNP_6_whitelisted.seg)) were retrieved from the TCGA PanCanAtlas data-freeze release ([https://gdc.cancer.gov/about-data/publications/pancanatlas](https://gdc.cancer.gov/about-data/publications/pancanatlas)). Recurrent CNAs were identified in the TCGA PanCancer collection via The Genomic Identification of Significant Targets In Cancer (GISTIC, version 2, [26]) using a cut-off q-value of 0.25 and confidence threshold of 0.90 for peak boundaries (Suppl. Fig. 5). A region was classified as amplification or deletion if the LogR was above or below the 0.1 threshold. Downstream analyses were restricted to patients from the TCGA-COAD-READ cohort with *Fusobacteriales* estimates (n=563). Copy number aberrations were visualised as a heatmap using the *python* package CNVkit (version 0.9.7, function *do_heatmap*), (Fig. 3E). Percentage of patients with aberrations at a given genomic position were visualised with the *R* package *copynumber* (version 1.26.0, function *plotFreq*, [27]), (Sup. Fig. 6).
Transcriptomic-dependent *Fn/Fusobacteriales* impact.

Differences in copy number aberrations at the cytoband level were computed by computing the difference in mean lesion frequency between patients with high vs. low *Fusobacteriales* relative abundance (75th percentile cut-off), (**Fig. 3F**). Top 3 differential copy number aberrations at the cytoband level were visualised in **Fig. 3G**.

**Aberrations in transcriptional and protein profiles**

A systematic screen was carried out to identify aberrations in transcriptional and protein profiles by *Fusobacteriales* relative abundance in patients of the TCGA-COAD-READ cohort. Association between *Fusobacteriales* relative abundance and either gene or protein expression was assessed by Spearman correlation (function `pairwise_corr`) from the `python` package `pingouin` (version 0.3.11, [28]). P-values were adjusted for multiple comparisons for False Discovery Rate with Benjamini-Hochberg (function `pingouin.multicomp` from the python package `pingouin`). For transcriptional profiles, we restricted the analysis to the 5000 most variant genes. All available proteins were tested (n=189 proteins). Genes and proteins whose expression differed by *Fusobacteriales* relative abundance were put forward for pathway enrichment analyses carried out with the `gseapy` package (version 0.10.2, [29]) which provides a wrapper (function `gseapy.enrichr`) for `EnrichR` [30-31], (**Fig. 3 H-I, K-L and Sup. Fig.7-8**).

**Exploration of putative mechanisms underlying differential impact of *Fn/Fusobacteriales* prevalence by tumour biology**

We fitted 2 logistic regression models to identify putative mechanisms underlying the differential impact of *Fn/Fusobacteriales* prevalence in mesenchymal vs. non-mesenchymal tumours. Specifically, we fitted:

- **model 1**: univariate logistic regression model (*Fusobacteriales ~ gene/signature*);
Transcriptomic-dependent *Fn/Fusobacteriaales* impact.

- **model 2**: logistic regression model with an interaction term for mesenchymal status

  \[(\text{Fusobacteriaales} \sim \text{gene/signature} \times \text{mesenchymal status})\].

Patients were grouped into *Fusobacteriaales*-low vs. high using the 75\(^{th}\) percentile of *Fusobacteriaales* relative abundance as cut-off. Selection of gene expression or signatures to include in model evaluation was hypothesis driven and this analysis was considered exploratory in nature. Thus, no P-value adjustment for multiple comparisons was performed. Tumour mesenchymal status was treated as binary (yes, no). Tumour were classed as mesenchymal if they were classified as CMS4 and/or CRIS-B based on transcriptomic assignments from the CMS [21] and/or CRIS [22] subtyping strategies. Logistic regression models were fitted using the function `statsmodels.formula.api.logit` from the python package `statsmodels` (version 0.11.1, [32]).

**Statistical analysis.**

Statistical significance was set at P<0.05, unless otherwise specified.

**Comparative analyses**

For hypothesis-driven investigations, we visualized the association between either *Fn* or *Fusobacteriaales* (order) relative abundance (high vs. low) with either split violin or mosaic plots drawn with the python packages `matplotlib` (version 3.3.1, [33]), `seaborn` (version 0.11.0, [34]), for continuous and categorical clinical or molecular features, respectively. For hypothesis-driven analysis, we evaluated statistical significance by either non-parametric Kruskal-Wallis or \(\chi^2\) independence tests for continuous or categorical variables, respectively. Given the hypothesis-driven and exploratory nature of these analyses, the P-values were not adjusted for multiple
Transcriptomic-dependent *Fn/Fusobacteriales* impact. Comparisons. In contrast, in unbiased and systematic analyses (Fig. 3) or when specified, P-values were adjusted for False Discovery Rate with Benjamini-Hochberg FDR correction (FDR-BH).

**Outcome analysis.**

As outcome endpoints, we evaluated disease-free (DFS), disease-specific (DSS) and overall (OS) survival where we consider relapse, cancer-related death or death by any cause as event, respectively. For the Taxonomy cohort where the cause of death was not annotated, we assessed exclusively DFS and OS. We used Kaplan-Meier estimators and we fit univariate and interaction Cox proportional hazards regression models to evaluate survival by covariates with the *python* package *lifelines* (version 0.25.5, [35]). We assessed statistical significance with log-rank and likelihood ratio tests, respectively. Interaction Cox regression models were fitted to evaluate the cross-talk between bacterium prevalence (high vs. low using the 75th percentile as cut-off) and mesenchymal phenotypes (*mesenchymal*: either CMS4 and/or CRIS-B; *vs.* *non-mesenchymal*: neither CMS4 nor CRIS-B). For the Taxonomy cohort, we used *Fn* load as pathogen prevalence (Fig. 5A, C-D and Sup. Fig. 9). For the TCGA-COAD-READ cohort, we used *Fusobacteriales* relative abundance as pathogen prevalence (Fig. 5E, G-I, K-L and Sup. Fig. 10).

In additional analysis we evaluated whether our findings were robust when accounting for covariates that may represent confounders or disease modifiers (Suppl. Table 7). For each clinical endpoint of interest, namely OS, DSS, DFS, for the patients of the TCGA-COAD-READ cohort, we fitted 2 additional Cox regression models where in addition to the interaction term between *Fusobacteriales* and mesenchymal status we included adjustment covariates. In adjusted model 1, we included age (continuous), stage (categorical, I to IV), tumour location (categorical,
Transcriptomic-dependent *Fn/Fusobacteriales* impact.

colon vs. rectum) and sex (categorical, male vs. female) as key clinical, pathological and
demographic covariates. We considered including resection margins (categorical, R0 vs. R1-R2)
and presence of lymphovascular invasion (categorical, yes vs. no) as disease modifiers, but
decided against as these covariates were missing for a high proportion of the patients. In adjusted
model 2, we expand upon adjusted model 1 by also including history of colon polyps
(categorical, yes vs. no) and history of other malignancy as comorbidities. However, the
covariate information was not available for all the patients included in the analysis in the
manuscript. Thus, for this additional analysis, we selected only patients with available covariates
(~85% of those included in Fig. 5 of the manuscript). Also, we re-fitted the unadjusted Cox
regression models reported in the manuscript to aid in the interpretation of the results (Suppl.
Table 7).

In exploratory analysis, we additionally assessed the association between clinical outcome and
pathogen relative abundance at higher taxonomic resolution (family, genus and species) for
patients of the TCGA-COAD-READ cohort (Fig. 5M and Sup. Fig. 11).

We evaluated whether the gene/signature identified by the analysis presented in Fig. 6A as
candidate targets are indeed related to clinical outcome in patients of the TCGA-COAD-READ
cohort with mesenchymal tumours and high *Fusobacteriales* (Suppl. Figs. 12-14). To this end,
we restricted our analysis to patients with mesenchymal tumours and for each clinical endpoint
of interest, namely OS, DSS, DFS, we fitted Cox regression models with an interaction term for
*Fusobacteriales* relative abundance (low vs. high) and each of the gene/signature (low vs. high)
identified as statistically significant in the analysis presented in Fig. 6A. Suppl. Figs. 12-14
visualise the association between clinical outcome (OS, DSS, DFS) and each gene/signature

21
Transcriptomic-dependent *Ft/Fusobacteriales* impact.

across the whole unselected patient population and within the low- and high-Fusobacteriales subgroups.

**Software and libraries**

Data processing and analyses were performed in *R* (version 3.6.3, [36]) and *python* (version 3.8.10, [37]). Key libraries used in this study include *pandas* (version 1.1.2, [38]), *numpy* (version 1.19.1, [39]), *sklearn* (version 0.23.1, [40]), *matplotlib* (version 3.3.1, [33]), *seaborn* (version 0.11.0, [34]), *graphviz* (version 0.14.1, [41]), *UpSetPlot* (version 0.5.0, [42]), *tableone* (version 0.7.6, [5]), *statsmodels* (version 0.11.1, [32]), *pingouin* (version 0.3.11, [28]), *gseapy* (version 0.10.2, [29]), *lifelines* (version 0.25.5, [35]). The full list of packages and their versions along with the data and code will be publicly available and archived upon publication at Zenodo (https://10.5281/zenodo.4019142).
Transcriptomic-dependent *Fn/Fusobacteriales* impact.

**References**


Transcriptomic-dependent *Fn/Fusobacteriales* impact.


11 Chae YK, Chang S, Ko T *et al*. Epithelial-mesenchymal transition (EMT) signature is inversely associated with t-cell infiltration in non-small cell lung cancer (NSCLC). *Scientific Reports* 2018;8. doi:10.1038/s41598-018-21061-1


Transcriptomic-dependent Fn/Fusobacteriales impact.


Transcriptomic-dependent *Fn/Fusobacteria*les impact.

22 Isella C, Brundu F, Bellomo SE et al. Selective analysis of cancer-cell intrinsic transcriptional traits defines novel clinically relevant subtypes of colorectal cancer. *Nature Communications* 2017;8:15107. doi: 10.1038/ncomms15107


Transcriptomic-dependent *Fn/Fusobacteriales* impact.


Transcriptomic-dependent *Fn/Fusobacteriales* impact.
