Supplementary Materials and Methods for "Patients with mesenchymal tumours and high *Fusobacteriales* prevalence have worse prognosis in colorectal cancer (CRC)"

Manuela Salvucci[1], Nyree Crawford[2], Katie Stott[2], Susan Bullman[3,4], Daniel B. Longley[2], and Jochen H.M. Prehn[1]*

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

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

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

[4]Fred 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.

Contents

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 McCoys 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.

Salvucci M, *et al. Gut* 2021;0:1–13. doi: 10.1136/gutjnl-2021-325193

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**: CAGCCTCTTCTCCTTCCTGAT;
- **TNFα R**: GCCAGAGGGCTGATTAGAGA;
- **β-tubulin F**: CGCAGAAGAGGAGGAGGATT;
- **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

Salvucci M, *et al. Gut* 2021;0:1–13. doi: 10.1136/gutjnl-2021-325193

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

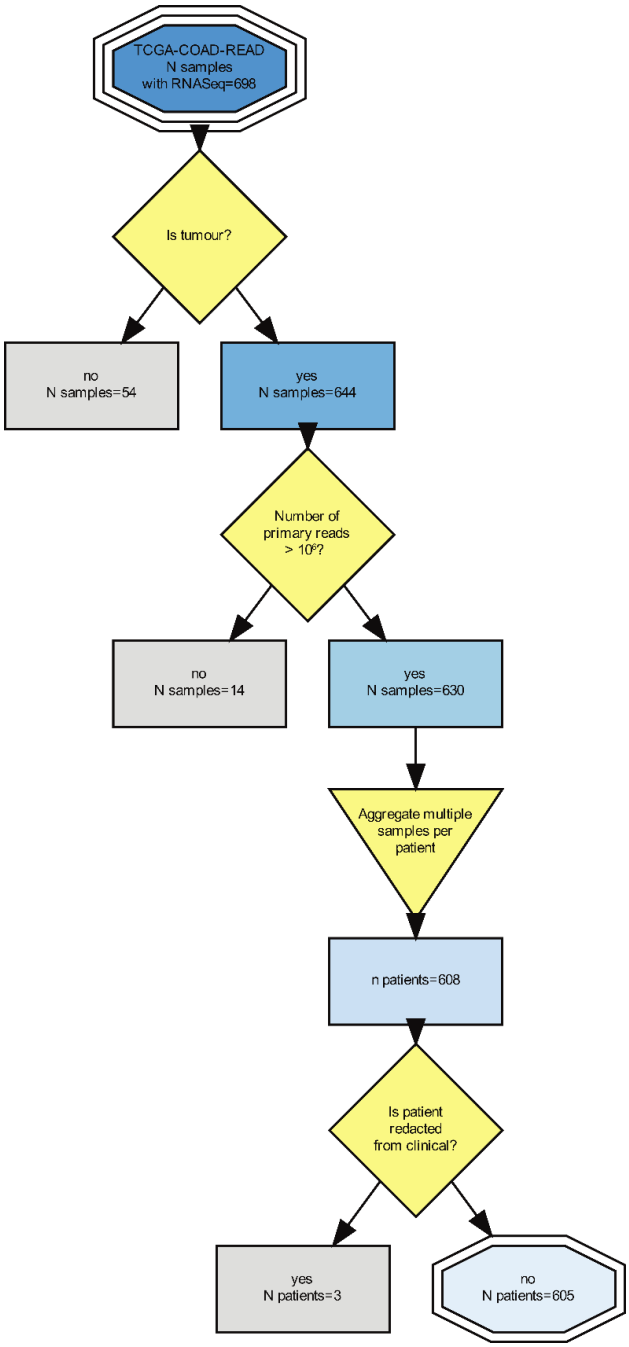
Stage I to IV patients with cancer of the colon (COAD) or rectum (READ) accrued by The Cancer Genome Atlas (TCGA) network with available fresh frozen tumour resections of sufficient quality and quantity for sequencing analysis (https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/studied-cancers) 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.

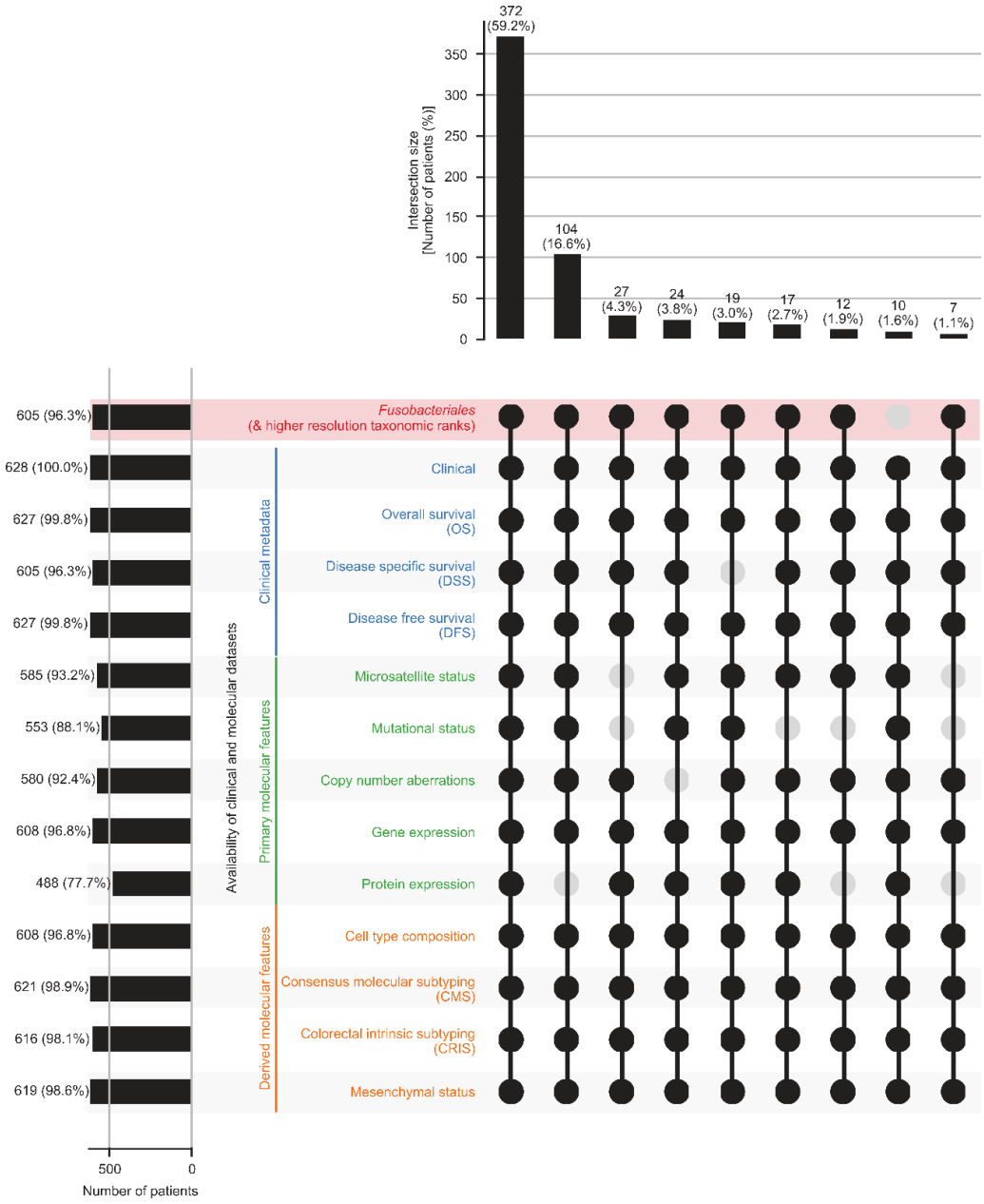
Salvucci M, *et al. Gut* 2021;0:1–13. doi: 10.1136/gutjnl-2021-325193

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/Fusobacteriales* impact.



Supplementary Materials and Methods Figure 2. (Cross-)availability of *Fusobacteriales* 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 $2^{-\Delta Ct}$ 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

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 *Fusobacteriales* 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/Fusobacteriales* impact.

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

Sub-species/strain	Candidate parent species	Per. identity	Remapped parent species
Cetobacterium_sp._ZOR0034	Cetobacterium_somerae	100%	Cetobacterium_somerae
Cetobacterium_sp._ZWU0022	Cetobacterium_somerae	99.78%	Cetobacterium_somerae
Fusobacterium_sp._CM1	Fusobacterium_nucleatum	99.86%	Fusobacterium_nucleatum
Fusobacterium_sp._CM21	Fusobacterium_nucleatum	99.86%	Fusobacterium_nucleatum
Fusobacterium_sp._CM22	Fusobacterium_nucleatum	99.70%	Fusobacterium_nucleatum
Fusobacterium_sp._HMSC064B11	Fusobacterium_nucleatum	99.93%	Fusobacterium_nucleatum
Fusobacterium_sp._HMSC064B12	Fusobacterium_nucleatum	99.82%	Fusobacterium_nucleatum
Fusobacterium_sp._HMSC065F01	Fusobacterium_nucleatum	99.87%	Fusobacterium_nucleatum
Fusobacterium_sp._OBRC1	Fusobacterium_nucleatum	100%	Fusobacterium_nucleatum
Fusobacterium_sp._HMSC073F01	Fusobacterium_varium	100%	Fusobacterium_varium
Leptotrichia_sp._Marseille-P3007	Leptotrichia_buccalis	98.37%	Leptotrichia_buccalis
Leptotrichia_sp._oral_taxon_225	Leptotrichia_trevisanii	99.52%	Leptotrichia_trevisanii
Leptotrichia_sp._oral_taxon_879	Leptotrichia_hongkongensis?	96.86%	un-mapped
Leptotrichia_sp._oral_taxon_212	Leptotrichia_hongkongensis?	92.67%	un-mapped
Leptotrichia_sp._oral_taxon_847	Leptotrichia_massiliensis?	92.04%	un-mapped
Leptotrichia_sp._oral_taxon_215	No candidate parent species found	un-mapped	
Fusobacterium_sp._oral_taxon_370	Fusobacterium_nucleatum or Fusobacterium_periodonticum?	95.33% for both	un-mapped

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).

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, CAMK2D, 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, NKD1, NKD2, NLK, PLCB1, PLCB2, PLCB3, PLCB4, 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,

ROCK1, ROCK2, RUVBL1, SENP2, SFRP1, SFRP2, SFRP4, SFRP5, SIAH1, SKP1, SMAD2, SMAD3, SMAD4, SOX17, TBL1X, TBL1XR1, TBL1Y, TCF7, TCF7L1, TCF7L2, TP53, VANGL1, VANGL2, WIF1, 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) 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

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), 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) 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 75th percentile as cut-off) and mutational status (number of aberrations) was assessed with χ^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 χ^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*) were retrieved from the TCGA PanCanAtlas data-freeze release (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 (75^{th} 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/Fusobacteriales* impact.

- **model 2**: logistic regression model with an interaction term for mesenchymal status (*Fusobacteriales ~ gene/signature * mesenchymal status*).

Patients were grouped into *Fusobacteriales*-low vs. high using the 75th percentile of *Fusobacteriales* 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 *Fusobacteriales* (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 χ^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

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 75^{th} 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,

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

across the whole unselected patient population and withing 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

1 Crawford N, Stasik I, Holohan C *et al.* SAHA overcomes flip-mediated inhibition of smac mimetic-induced apoptosis in mesothelioma. *Cell Death & Disease* 2013;**4**:e733–3. doi:10.1038/cddis.2013.258

2 Buckley NE, Haddock P, Simoes RDM *et al.* A brca1 deficient, nfkappab driven immune signal predicts good outcome in triple negative breast cancer. *Oncotarget* 2016;**7**:19884–96. doi:10.18632/oncotarget.7865

3 Allen WL, Dunne PD, McDade S *et al.* Transcriptional Subtyping and CD8 Immunohistochemistry Identifies Patients With Stage II and III Colorectal Cancer With Poor Prognosis Who Benefit From Adjuvant Chemotherapy. *JCO Precision Oncology* 2018;**44**:1–15. doi:10.1200/po.17.00241

4 McCorry AM, Loughrey MB, Longley DB *et al.* Epithelial-to-mesenchymal transition signature assessment in colorectal cancer quantifies tumour stromal content rather than true transition. *Journal of Pathology* 2018;**246**:422–6. doi:10.1002/path.5155

5 Pollard TJ, Johnson AEW, Raffa JD *et al.* Tableone: An open source python package for producing summary statistics for research papers. *JAMIA Open* 2018;**1**:26–31. doi:10.1093/jamiaopen/ooy012

6 Liu J, Lichtenberg T, Hoadley KA *et al.* An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell* 2018;**173**:400–416.e11. doi:10.1016/j.cell.2018.02.052

7 Biosciences M. Nbt.1868.Pdf. *Nature Publishing Group* 2011;**29**. doi:10.1038/nbt0511-393

Transcriptomic-dependent *Fn/Fusobacteriales* impact.

8 Walker MA, Pedamallu CS, Ojesina AI *et al.* GATK PathSeq: a customizable computational tool for the discovery and identification of microbial sequences in libraries from eukaryotic hosts. *Bioinformatics (Oxford, England)* 2018;**34**:4287–9. doi:10.1093/bioinformatics/bty501

9 McKenna A, Hanna M, Banks E *et al.* The genome analysis toolkit: A mapreduce framework for analyzing next-generation dna sequencing data. *Genome Research* 2010;**20**:1297–303. doi:10.1101/gr.107524.110

10 Nielsen TO, Parker JS, Leung S *et al.* A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptorPositive breast cancer. *Clinical Cancer Research* 2010;**16**:5222–32. doi:10.1158/1078-0432.ccr-10-1282

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

12 Ramaswamy S, Ross KN, Lander ES *et al.* A molecular signature of metastasis in primary solid tumors. *Nature Genetics* 2002;**33**:49–54. doi:10.1038/ng1060

13 Chang WH, Lai AG. Transcriptional landscape of DNA repair genes underpins a pan-cancer prognostic signature associated with cell cycle dysregulation and tumor hypoxia. *DNA Repair* 2019;**78**:142–53. doi:10.1016/j.dnarep.2019.04.008

14 Damotte D, Warren S, Arrondeau J *et al.* The tumor inflammation signature (TIS) is associated with anti-PD-1 treatment benefit in the CERTIM pan-cancer cohort. *Journal of Translational Medicine* 2019;**17**. doi:10.1186/s12967-019-2100-3

Transcriptomic-dependent *Fn/Fusobacteriales* impact.

15 Rooney MS, Shukla SA, Wu CJ *et al.* Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* 2015;**160**:48–61. doi:10.1016/j.cell.2014.12.033

16 Higgs BW, Morehouse CA, Streicher K *et al.* Interferon gamma messenger RNA signature in tumor biopsies predicts outcomes in patients with nonSmall cell lung carcinoma or urothelial cancer treated with durvalumab. *Clinical Cancer Research* 2018;**24**:3857–66. doi:10.1158/1078-0432.ccr-17-3451

17 Wang S, Song R, Wang Z *et al.* S100A8/a9 in inflammation. *Frontiers in Immunology* 2018;**9**. doi:10.3389/fimmu.2018.01298

18 Kim N, Kim HK, Lee K *et al.* Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nature Communications* 2020;**11**. doi:10.1038/s41467-020-16164-1

19 Becht E, Giraldo NA, Lacroix L *et al.* Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biology* 2016;**17**:218. doi:10.1186/s13059-016-1070-5

20 Finotello F, Mayer C, Plattner C *et al.* Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. *Genome Medicine* 2019;**11**. doi:10.1186/s13073-019-0638-6

21 Guinney J, Dienstmann R, Wang X *et al.* The consensus molecular subtypes of colorectal cancer. *Nature Medicine* 2015;**21**:1350–6. doi:10.1038/nm.3967

Transcriptomic-dependent *Fn/Fusobacteriales* 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

23 Thorsson V, Gibbs DL, Brown SD *et al.* The Immune Landscape of Cancer. *Immunity* 2018;**48**:812–830.e14. doi:10.1016/j.immuni.2018.03.023

24 Bonneville R, Krook MA, Kautto EA *et al.* Landscape of microsatellite instability across 39 cancer types. *JCO Precision Oncology* 2017;1–15. doi:10.1200/po.17.00073

25 Mayakonda A, Lin D-C, Assenov Y *et al.* Maftools: Efficient and comprehensive analysis of somatic variants in cancer. *Genome Research* 2018;**28**:1747–56. doi:10.1101/gr.239244.118

26 Mermel CH, Schumacher SE, Hill B *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology* 2011;**12**. doi:10.1186/gb-2011-12-4-r41

27 Nilsen G, Liestøl K, Loo PV *et al.* Copynumber: Efficient algorithms for single- and multi-track copy number segmentation. *BMC Genomics* 2012;**13**:591. doi:10.1186/1471-2164-13-591

28 Vallat R. Pingouin: Statistics in python. *Journal of Open Source Software* 2018;**3**:1026. doi:10.21105/joss.01026

29 Fang Z. GSEApy: Gene set enrichment analysis in python. 2020. doi:10.5281/ZENODO.3748085

30 Chen EY, Tan CM, Kou Y *et al.* Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* 2013;**14**:128. doi:10.1186/1471-2105-14-128

Transcriptomic-dependent *Fn/Fusobacteriales* impact.

31 Kuleshov MV, Jones MR, Rouillard AD *et al.* Enrichr: A comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research* 2016;**44**:W90–7. doi:10.1093/nar/gkw377

32 Hunter JD. Matplotlib: A 2D graphics environment. *Computing in science & engineering* 2007;**9**:90–5.

33 Waskom M. Seaborn: Statistical data visualization. *Journal of Open Source Software* 2021;**6**:3021. doi:10.21105/joss.03021

34 Davidson-Pilon C. Lifelines: Survival analysis in python. *Journal of Open Source Software* 2019;**4**:1317. doi:10.21105/joss.01317

35 R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria:: R Foundation for Statistical Computing 2020. https://www.R-project.org/

36 Van Rossum G, Drake FL. *Python 3 reference manual*. Scotts Valley, CA:: CreateSpace 2009.

37 McKinney W, others. Data structures for statistical computing in python. In: *Proceedings of the 9th python in science conference*. Austin, TX 2010. 51–6.

38 Oliphant TE. *A guide to numpy*. Trelgol Publishing USA 2006.

39 Pedregosa F, Varoquaux G, Gramfort A *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 2011;**12**:2825–30.

40 Ellson J, Gansner E, Koutsofios L *et al.* Graphviz — open source graph drawing tools. In: *Lecture notes in computer science*. Springer-Verlag 2001. 483–4.

Transcriptomic-dependent *Fn/Fusobacteriales* impact.

41 Alexander Lex HS Nils Gehlenborg. UpSet: Visualization of intersecting sets. *IEEE transactions on visualization and computer graphics* 2014;**20**:1983–92. doi:10.1109/TVCG.2014.2346248

42 Seabold S, Perktold J. Statsmodels: Econometric and statistical modeling with python. In: *9th python in science conference*. 2010.