SUPPLEMENTARY MATERIALS AND METHODS

1. Description of participating studies

Colon Cancer Family Registry (CCFR) (1): The CCFR is an NCI-supported consortium consisting of six centers dedicated to the establishment of a comprehensive collaborative infrastructure for interdisciplinary studies in the genetic epidemiology of colorectal cancer. The CCFR includes data from approximately 42500 total subjects in 15000 families (10500 probands, and 26770 unaffected and affected relatives and 4276 unrelated controls and 923 spouse controls). Cases and controls, age 20 to 74 years, were recruited at the six participating centers beginning in 1998. The CCFR implemented a standardized questionnaire that was administered to all participants, and included established and suspected risk factors for colorectal cancer, including questions on medical history and medication use, reproductive history (for female participants), family history, physical activity, demographics, alcohol and women who also enrolled tobacco use, and dietary factors. This study selected tumor samples for this molecular subtype study from two of the CCFR sites, Ontario (OFCCR) and Seattle (SCCFR).

Cancer Prevention Study-II (CPS-II) (2,3): The CPS-II Nutrition Survey cohort is a prospective study of cancer incidence and mortality in the United States, established in 1992 and described in detail elsewhere. At enrollment, participants completed a mailed self-administered questionnaire including information on demographic, medical, diet, and lifestyle factors. Follow-up questionnaires to update exposure information and to ascertain newly diagnosed cancers were sent biennially starting in 1997. Reported cancers were verified through medical records, state cancer registry linkage, or death certificates. The Emory University Institutional Review Board approves all aspects of the CPS II Nutrition Cohort.

Colorectal Cancer Study of Austria (CORSA) (4): In the ongoing CORSA study, more than 16,000 Caucasian participants have been recruited within the province-wide screening project “Burgenland Prevention Trial of Colorectal Disease with Immunological Testing” (B-PREDICT) since 2003. All inhabitants of the Austrian province Burgenland aged between 40 and 80 years are annually invited to participate in fecal immunochemical testing and haemoccult positive screening participants are invited for colonoscopy. CORSA participants have been recruited in the four KRAGES hospitals in Burgenland, Austria, and additionally, at the Medical University of Vienna (Department of Surgery), the Viennese hospitals “Rudolfstiftung” and the “Sozialmedizinisches Zentrum Süd”, and at the Medical University of Graz (Department of Internal Medicine).

Tumors analyzed for the presence of *F nucleatum* and somatic mutations across the GECCO studies, and the availability of tumor characteristics and survival data are described in the Table below.

Participant studies, number of tumors, and availability of data.

Study	**Tumors (n)**	**Tumor characteristics (n)**	**Tumor characteristics with survival data (n)**
OFCCR	730	674	462
SCCFR	540	523	396
CPS-II	576	536	462
CORSA	148	140	0
Total (n)	**1994**	**1873**	**1320**

2. Methods

Statistical analysis in R

Logistic and Cox regression analyses were performed in R.3.6.0 (https://cran.r-project.org/). P-values obtained from univariable and multivariable logistic regressions for *APC*, *TP53*, *KRAS*, *ERBB2*, *ERBB3, POLE*, *PIK3CA*, *SMAD4*, and *BRAF* genes were adjusted for multiple testing with the 'p.adjust' function using the Benjamini-Hochberg method (aka 'fdr' procedure).

Cox proportional hazards regression was performed using 'coxph'. We checked the proportional hazards assumption with the 'cox.zph' function and used stratification for the Cox model to allow for non-proportionality as necessary. To control for confounders, the Cox proportional hazards regression model was adjusted for sex, age at diagnosis, tumor site, tumor stage, hypermutation status, tumor burden, *POLE*, *TP53* and *ERBB3* mutation status, and MSI status.

We performed a series of sensitivity analyses in which different propensity score based methods were used for modeling the association of *F nucleatum*/subspecies with the survival outcome. Specifically, we considered the following three models based on propensity score (Model 1 and 2 for binary *F nucleatum* status, and Model 3 for *F nucleatum* subspecies that has three levels).

1. Model 1: Generate the propensity score for *F nucleatum*. Categorize propensity score into 5 groups. Then run the Cox model including only *F nucleatum*, stratified by the categorized propensity score.
2. Model 2: Generate the propensity score for *F nucleatum*. Categorize propensity score into 5 groups. Then run the Cox model with two predictors: *F nucleatum* and the categorized propensity score in terms of indicators.
3. Model 3: Generate the propensity score for *F nucleatum* subspecies. Then run the Cox model that includes only *F nucleatum* subspecies, imposed with inverse probability weighting (IPW) based on the propensity score.

We generated propensity scores using the following three approaches respectively for each model.

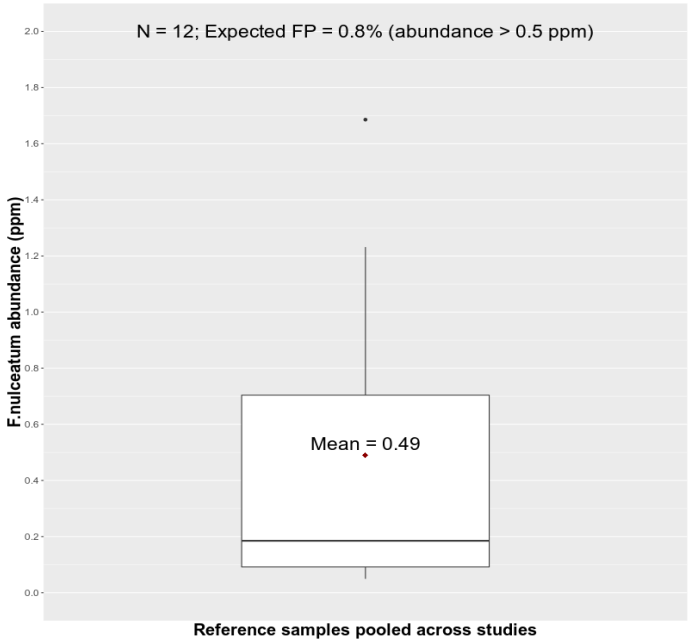
1) Logistic regression for *F nucleatum* (multinomial logistic regression for *F nucleatum* subspecies) to predict the exposure of *F nucleatum*/subspecies status with the predictors of sex, age at diagnosis, tumor site, hypermutation status, tumor burden, mutations in *POLE*, *TP53*, and *ERBB3*, MSI status, and with/without tumor stage. We used the "glm" and "multinom" functions to implement the analysis.
2) Covariate balancing propensity score (CBPS) which is estimated such that it maximizes the resulting covariate balance as well as the prediction of treatment assignment. We used the "CBPS" function in the "CBPS" package to implement the analysis.
3) Generalized boosted model (GBM) which is a nonparametric, piecewise constant model for predicting the treatment. We used the "gbm" function in the "gbm" package to implement the analysis.

For inference, the confidence intervals are based on 1000 bootstrapping samples. Results of these sensitivity analyses are shown in Supplementary Tables 3, 4, 5, and 6. Supplementary Table 3 shows results for the association between *F nucleatum* and CRC-specific survival by Model 1 and 2 for all three approaches (Logistic regression, CBPS, GBM) respectively. These results should be compared with results from the multivariate Cox models shown in Table 2. In the same Supplementary Table 3 we also show the results for the association between *F nucleatum* subspecies and CRC-specific survival by Model 3 for all three approaches (Logistic regression, CBPS, GBM) respectively. These results should be compared with results from the multivariate Cox models shown in Table 3. Supplementary Tables 4 and 5 show the results for the association between *F nucleatum* and CRC-specific survival by Model 1 and 2 for all three approaches (Logistic regression, CBPS, GBM) stratified by MSI status and chemotherapy, respectively. These results should be compared to results described in the main text under results, sub header "Impact of *F nucleatum* on survival".

For causal effects between *ERBB3* mutation and *F nucleatum*, we calculated the average treatment effect (ATE), which is a causal measure used to compare treatments (or interventions) in randomized trials. In our analysis, the target ATE is the proportion difference of *F nucleatum* positive between *ERBB3* mutation and non-mutation groups after marginalizing over covariate distribution. We calculated it based on multivariate logistic regression. Confidence intervals are based on 1000 bootstrapping samples. The results are reported in Supplementary Table 6.

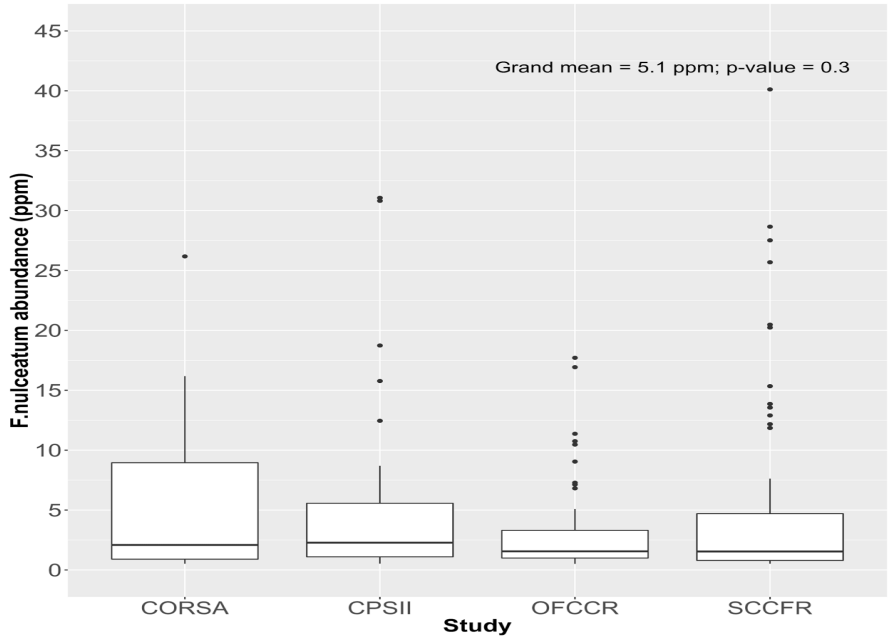
Controlling for possible sequencing and alignment artifacts

To better control for possible sequencing and analysis artifacts, our approach relied on the background model (or baseline) estimated using patients matched blood samples. The stringent cutoff for *F nucleatum* abundance ≥ 0.5 ppm was determined based on the average abundance of *F nucleatum* detected in 0.8% matched blood normal samples, as shown below in Supplementary Figure S1.



Supplementary Figure S1. Box plot representing the mean abundance of *F nucleatum* detected in 12/1568 matched blood samples.

Using blood as a background signal and by setting the abundance threshold to ≥0.5 ppm, we estimated the false positive rate of detection to be less than 0.8%. The *F nucleatum* abundance levels detected were consistent across all four studies (average abundance = 5.1 ppm, p-value = 0.3, Supplementary Figure S2).

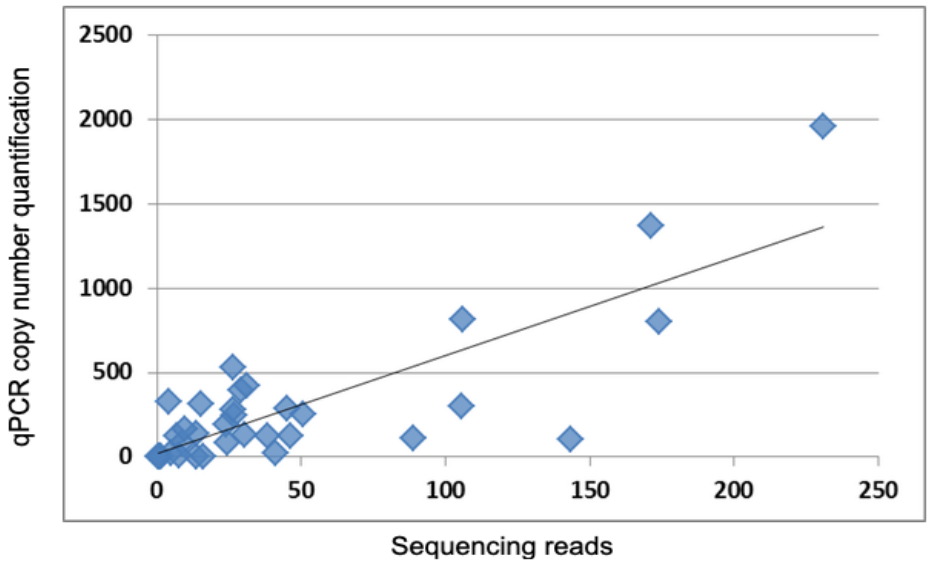


Supplementary Figure S2. *F nucleatum* abundance across studies.

Validation of the next-generation sequencing method to quantify F nucleatum with quantitative PCR

Quantitative PCR assay was developed to amplify the *nusG* of *F nucleatum*. The TaqMan probe contained the 6-FAM attached to the 5' end, and the MGB attached to the 3' end. The human *TERT* reference assay (Applied Biosystems) was added in duplex to our custom *nusG* primers and probe for normalization of *nusG* quantification. The *TERT* probe was labeled with VIC dye and TAMRA quencher.

For validation, 40 representative DNA samples isolated from FFPE colorectal tumor tissues were used (Supplementary Figure S3. Five of these tumor DNA samples were negative for *F nucleatum* sequence reads. The remaining 35 tumor DNA samples were selected to represent a relatively even distribution of total *F nucleatum* reads. DNA samples were prepared in duplicates with negative controls containing CEPH DNA. TaqMan Universal Mastermix II, with UNG (Applied Biosystems), was used as our qPCR reagent solution.



Supplementary Figure S3. The *nusG* qPCR copy number quantification vs. *F nucleatum* sequencing reads. *F nucleatum* reads were normalized to the total number of reads sequenced for that sample. There is a strong correlation between qPCR copy number and sequencing reads for the *F nucleatum* ($R = 0.81$, $p<0.00001$).

4. References

1. Newcomb PA, Baron J, Cotterchio M, Gallinger S, Grove J, Haile R, et al. Colon Cancer Family Registry: an international resource for studies of the genetic epidemiology of colon cancer. Cancer Epidemiol Biomarkers Prev. 2007;16:2331–43.

2. Calle EE, Rodriguez C, Jacobs EJ, Almon ML, Chao A, McCullough ML, et al. The

American Cancer Society Cancer Prevention Study II Nutrition Cohort: rationale, study design, and baseline characteristics. Cancer. 2002;94:2490–501.

3. Campbell PT, Deka A, Briggs P, Cicek M, Farris AB, Gaudet MM, et al. Establishment of the cancer prevention study II nutrition cohort colorectal tissue repository. Cancer Epidemiol Biomarkers Prev. 2014;23:2694–702.

4. Hofer P, Baierl A, Feik E, Führlinger G, Leeb G, Mach K, et al. MNS16A tandem repeats minisatellite of human telomerase gene: a risk factor for colorectal cancer. Carcinogenesis. 2011;32:866–71.