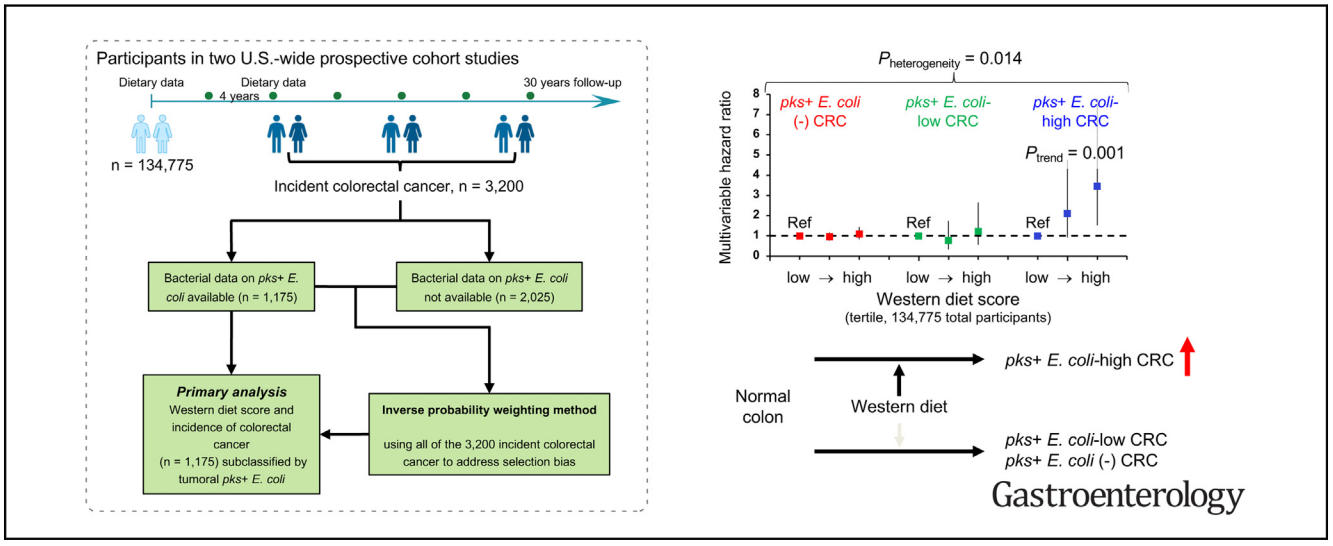
Gastroenterology 2022;163:862–874



# Western-Style Diet, *pks* Island-Carrying *Escherichia coli*, and Colorectal Cancer: Analyses From Two Large Prospective Cohort Studies

**Kota Arima**,[1,2,3,*] **Rong Zhong**,[1,4,5,*] **Tomotaka Ugai**,[1,4,*] **Melissa Zhao**,[1,*] Koichiro Haruki,[1,3] Naohiko Akimoto,[1] Mai Chan Lau,[1] Kazuo Okadome,[1] Raaj S. Mehta,[6,7] Juha P. Väyrynen,[1,3,8] Junko Kishikawa,[1] Tyler S. Twombly,[1] Shanshan Shi,[1] Kenji Fujiyoshi,[1] Keisuke Kosumi,[1] Yoko Ogata,[2] Hideo Baba,[2] Fenglei Wang,[9] Kana Wu,[9] Mingyang Song,[6,7,9] Xuehong Zhang,[9,10] Charles S. Fuchs,[11,12,13,14] Cynthia L. Sears,[15] Walter C. Willett,[9,10] Edward L. Giovannucci,[4,9,10] Jeffrey A. Meyerhardt,[3] Wendy S. Garrett,[3,16,17] Curtis Huttenhower,[18] Andrew T. Chan,[6,7,10,16,19,§] Jonathan A. Nowak,[1,§] Marios Giannakis,[3,17,19,§] and Shuji Ogino[1,4,17,20,§]

*[1]Program in MPE Molecular Pathological Epidemiology, Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts; [2]Department of Gastroenterological Surgery, Graduate School of Medical Sciences, Kumamoto University, Kumamoto, Japan; [3]Department of Medical Oncology, Dana-Farber Cancer Institute and Harvard Medical School, Boston, Massachusetts; [4]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts; [5]Department of Epidemiology and Biostatistics and Ministry of Education Key Lab of Environment and Health, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei, China; [6]Clinical and Translational Epidemiology Unit, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts; [7]Division of Gastroenterology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts; [8]Cancer and Translational Medicine Research Unit, Medical Research Center Oulu, Oulu University Hospital and University of Oulu, Oulu, Finland; [9]Department of Nutrition, Harvard T.H. Chan School of Public Health, Boston, Massachusetts; [10]Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts; [11]Yale Cancer Center, New Haven, Connecticut; [12]Department of Medicine, Yale School of Medicine, New Haven, Connecticut; [13]Smilow Cancer Hospital, New Haven, Connecticut; [14]Genentech, South San Francisco, California; [15]Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland; [16]Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Boston, Massachusetts; [17]Broad Institute of MIT and Harvard, Cambridge, Massachusetts; [18]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts; [19]Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts; and [20]Cancer Immunology and Cancer Epidemiology Programs, Dana-Farber Harvard Cancer Center, Boston, Massachusetts*




See editorial on page 812.

**BACKGROUND & AIMS:** Evidence supports a carcinogenic role of *Escherichia coli* carrying the *pks* island that encodes enzymes for colibactin biosynthesis. We hypothesized that the association of the Western-style diet (rich in red and processed meat) with colorectal cancer incidence might be stronger for tumors containing higher amounts of *pks*$^{+}$ *E coli*. **METHODS:** Western diet score was calculated using food frequency questionnaire

data obtained every 4 years during follow-up of 134,775 participants in 2 United States-wide prospective cohort studies. Using quantitative polymerase chain reaction, we measured *pks*$^+$ *E coli* DNA in 1175 tumors among 3200 incident colorectal cancer cases that had occurred during the follow-up. We used the 3200 cases and inverse probability weighting (to adjust for selection bias due to tissue availability), integrated in multivariable-adjusted duplication-method Cox proportional hazards regression analyses. **RESULTS:** The association of the Western diet score with colorectal cancer incidence was stronger for tumors containing higher levels of *pks*$^+$ *E coli* ($P_{heterogeneity}$ = .014). Multivariable-adjusted hazard ratios (with 95% confidence interval) for the highest (vs lowest) tertile of the Western diet score were 3.45 (1.53–7.78) ($P_{trend}$ = 0.001) for *pks*$^+$ *E coli*-high tumors, 1.22 (0.57–2.63) for *pks*$^+$ *E coli*-low tumors, and 1.10 (0.85–1.42) for *pks*$^+$ *E coli*-negative tumors. The *pks*$^+$ *E coli* level was associated with lower disease stage but not with tumor location, microsatellite instability, or *BRAF*, *KRAS*, or *PIK3CA* mutations. **CONCLUSIONS:** The Western-style diet is associated with a higher incidence of colorectal cancer containing abundant *pks*$^+$ *E coli*, supporting a potential link between diet, the intestinal microbiota, and colorectal carcinogenesis.

*Keywords:* Immunology; Microbiome; Molecular Pathological Epidemiology.

**WHAT YOU NEED TO KNOW**

BACKGROUND AND CONTEXT

The Western-style diet has been weakly associated with colorectal cancer risk; however, whether the association of a Western-style diet with colorectal cancer incidence varies by gut microbe remains unclear.

NEW FINDINGS

This analysis of 2 United States longitudinal prospective cohort studies showed that the association of a Western-style diet with colorectal cancer incidence was stronger for tumors containing higher amounts of *pks* island-carrying *Escherichia coli*.

LIMITATIONS

Our cohorts consisted predominantly of non-Hispanic Whites. Therefore, further studies using other populations are needed, as well as experimental confirmation to investigate the mechanisms.

IMPACT

These findings provide evidence supporting the role of the specific bacterium in mediating a pathogenic link between diet and colorectal cancer and the importance of diet for cancer prevention.



Accumulating evidence indicates that certain intestinal microorganisms influence colorectal tumor development through DNA damage, inflammation, and other mechanisms.[1–5] Among intestinal bacteria, *Escherichia coli* (*E coli*) strains of the B2 phylotype commonly harbor the 54-kilobase *pks* pathogenicity island that encodes enzymes for colibactin biosynthesis.[1,3,6] Experimental studies have shown that colibactin can alkylate DNA, induce DNA double-strand breaks, and cause a specific somatic mutational pattern in human cells.[7–9] However, human population studies are needed to better understand the role of *pks* island-carrying *E coli* (hereafter referred to as *pks*$^+$ *E coli*) in colorectal cancer.

Diet and nutrition are considered crucial factors for colorectal cancer development. A meta-analysis has shown a weak-to-modest relationship between Western dietary patterns and colorectal cancer risk.[10] An experimental study indicates that a Western-style diet (characterized by a high intake of red and processed meat, sugar, and refined grains and low intake of vegetables and legumes) can induce systemic and intestinal inflammation.[11] Considering the possible interplay between diet and pathogenic bacteria, studying the Western-style diet in relation to *pks*$^+$ *E coli* within colorectal tumor tissue is of particular interest. Such analyses may contribute to the development of cancer-prevention strategies targeting diet and microbiota.

In this study, we tested the hypothesis that the association of the Western-style diet with colorectal cancer incidence might be stronger for tumors containing higher amounts of *pks*$^+$ *E coli*. We used a molecular pathologic epidemiology database of 2 United States (US)-wide longitudinal prospective cohort studies with incident colorectal cancer cases. This comprehensive data set offered a unique opportunity to examine the long-term dietary patterns of individuals (who had not known whether they would develop cancers or not) in relation to colorectal cancer incidence subclassified by *pks*$^+$ *E coli* levels, while adjusting for potential confounders and selection bias due to tissue availability. In addition, we comprehensively assessed clinical, pathologic, molecular, and prognostic features according to the amount of *pks*$^+$ *E coli* in colorectal carcinoma tissue.

## Materials and Methods

### Study Population and Dietary Assessment

We used 2 prospective cohort studies in the US, namely, the Nurses' Health Study (NHS; 121,700 women aged 30–55 years at enrollment in 1976)[12,13] and the Health Professionals Follow-up Study (HPFS; 51,529 men aged 40–75 years at enrollment in 1986)[13,14] (Figure 1). Study participants had been monitored by use of questionnaires every 2 years on lifestyle and diagnoses of major diseases. The response rate has

*** Authors contributed equally as co-first authors; § Authors contributed equally as co-last authors.**

***Abbreviations used in this paper:*** **CI, confidence interval; Ct, cycle threshold; FFPE, formalin-fixed paraffin-embedded; HPFS, Health Professionals Follow-up Study; IPW, inverse probability weighting; MSI, microsatellite instability; NHS, Nurses' Health Study; PCR, polymerase chain reaction; SD, standard deviation; US, United States.**

**Most current article**


0016-5085/$36.00
https://doi.org/10.1053/j.gastro.2022.06.054

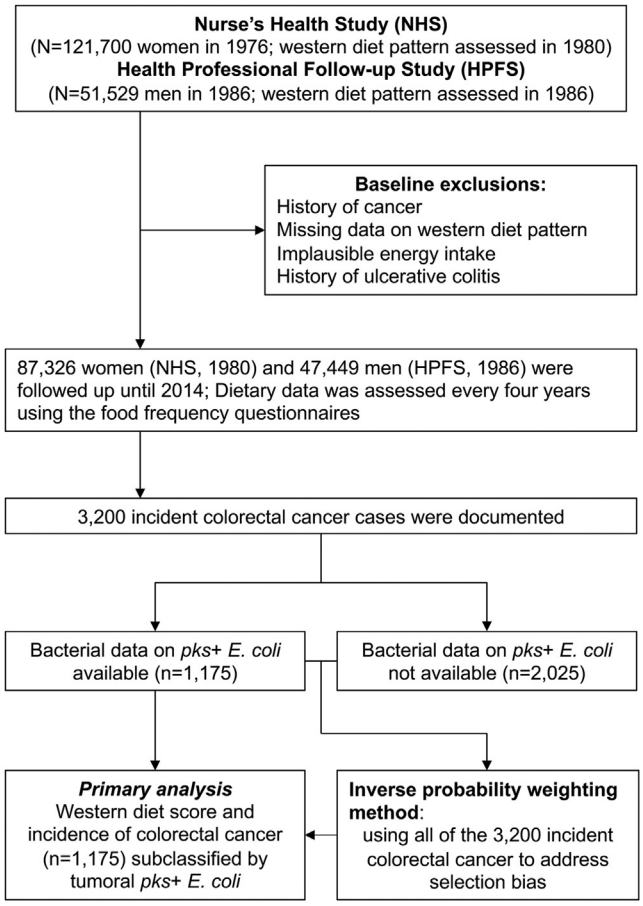



**Figure 1.** Flow diagram of the study population in the Nurses' Health Study and the Health Professionals Follow-up Study.

exceeded 90% for each follow-up questionnaire cycle in both cohorts. Dietary data were collected using self-administrated semiquantitative food frequency questionnaires in 1984, 1986, and every 4 years thereafter in the NHS, and every 4 years since 1986 in the HPFS. Validity of semiquantitative food frequency questionnaires in the assessment of dietary intake was extensively assessed and documented in studies by using diet records and plasma nutrients.[15–17] Total nutrient intakes were calculated by summing intakes from all foods and adjusted for total energy intake by the residual method. In this study, we used data from 134,775 participants who provided sufficient longitudinal dietary information.

The participants had been monitored since the baseline questionnaire return until colorectal cancer diagnosis, loss to follow-up, end of follow-up (June 1, 2014, for the NHS; January 1, 2014, for the HPFS), or death, whichever came first. Participants who had major illnesses, including colorectal cancer, reported those through questionnaires. Unreported lethal colorectal cancer cases were ascertained through use of the National Death Index. Clinical information, such as tumor location and disease stage based on the American Joint Committee on Cancer classification, was extracted from medical record by a study physician.[18] We included both colon and rectal carcinomas based on the colorectal continuum model.[19]

We gathered formalin-fixed paraffin-embedded (FFPE) tissue blocks from pathology files of hospitals throughout the US where the patients' tumors were resected. Histopathologic features, including tumor differentiation, extracellular mucin, and signet ring cells, were evaluated by the study pathologist (S.O.).[20] In this study, the inverse probability weighting (IPW) method using cases with available tissue bacterial data (n = 1175) and those without tissue bacterial data (n = 2025) was integrated into duplication method Cox proportional hazards regression analysis to adjust for selection bias due to tissue bacterial data availability (Figure 1). Characteristics of the cases with tissue bacterial data were similar to those without tissue bacterial data (Supplementary Table 1). In addition, during an assay validation step, we used tissues from 21 anonymized colorectal cancer patients who underwent surgical resections at the Brigham and Women's Hospital or Kumamoto University.

Informed consent was obtained from all participants at enrollment and consent for tissue specimen use was additionally obtained before tissue collection. This study protocol was approved by the Institutional Review Boards of the Brigham and Women's Hospital (Boston, MA) and Kumamoto University (Kumamoto, Japan), and those of participating registries as required.

## Tumor Tissue Analyses

Genomic DNA was extracted from archival FFPE tissue sections of colorectal carcinoma using the QIAamp DNA FFPE Tissue Kit and GeneRead DNA FFPE Kit (Qiagen, Hilden, Germany). We used custom TaqMan primer-probe sets (Applied Biosystems, Foster City, CA) for the *clbB* gene DNA sequence of *pks*$^+$ *E coli*[4] and for the reference human gene *SLCO2A1* that has been used in other bacterial assays on FFPE tissue-derived DNA[21] (the names used follow the recommendations for standardized nomenclature of genes and their products by an expert panel[22]). Genomic DNA concentration derived from samples was measured by Qubit Fluorometer (Thermo Fisher Scientific, Waltham, MA). Each reaction contained 20 ng of genomic DNA and was assayed in 20 $\mu$L reactions containing 1$\times$ final concentration TaqMan Environmental Master Mix 2.0 (Applied Biosystems) in a 96-well optical polymerase chain reaction (PCR) plate. Amplification and detection of DNA were performed with a QuantStudio 3 Real-time PCR System (Thermo Fisher Scientific) using the following reaction conditions: 10 minutes at 95°C and 45 cycles of 15 seconds at 95°C, 30 seconds at 57°C, and 30 seconds at 72°C. The primer and probe sequences for each TaqMan Gene Expression Assay were as follows: *pks*$^+$ *E coli* forward primer, 5′-GCAACATACTCGCCCAGACT-3′; *pks*$^+$ *E coli* reverse primer, 5′-TCTCAAGGCGTTGTTGTTTG-3′; *pks*$^+$ *E coli* FAM probe, 5′-CAAGGTGCGCGCTAGGCTGT-3′; *SLCO2A1* forward primer, 5′-ATCCCCAAAGCACCTGGTTT-3′; *SLCO2A1* reverse primer, 5′-AGAGGCCAAGATAGTCCTGGTAA-3′; and *SLCO2A1* VIC probe, 5′-CCATCCATGTCCTCATCTC-3′.

To validate our PCR assay, Sanger dideoxy sequencing was performed on the PCR product from 3 anonymized patients with colorectal carcinoma in which the PCR assay detected *pks*$^+$ *E coli* DNA. The PCR product (165 base pairs) using the forward and reverse primer sets was isolated by agarose gel electrophoresis. The isolated PCR product was amplified by subcloning and sequenced by Sanger dideoxy sequencing using the BigDye Terminator v3.1 Cycle Sequencing Kit (Thermo Fisher Scientific). We used Competent Quick DH5a (Toyobo, Osaka, Japan) and QIAprep Spin Miniprep Kit (Qiagen) in the transformation and extraction.

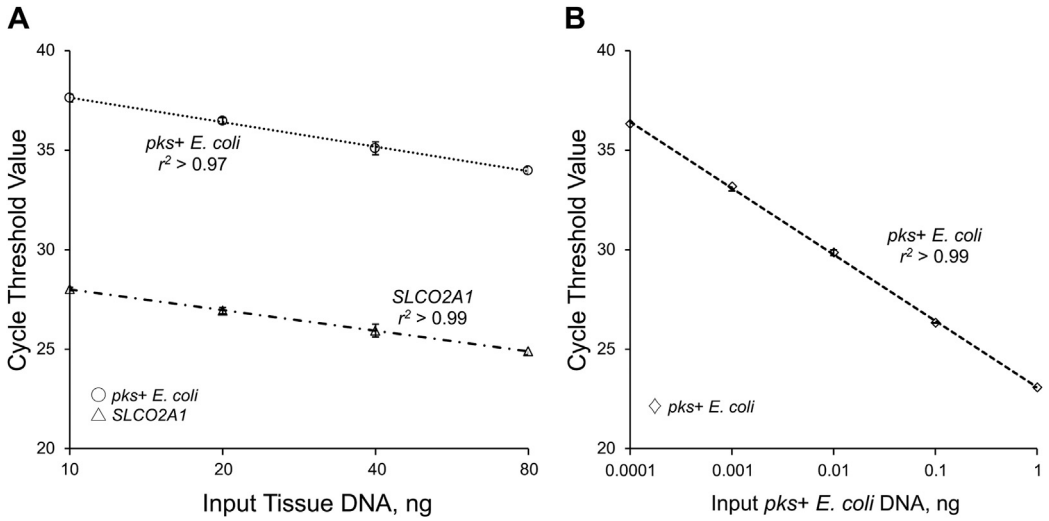



**Figure 2.** Assessment of linearity in quantitative real-time PCR assay. (*A*) Quantitative real-time PCR assay for *pks*$^+$ *E coli* DNA and the human reference gene *SLCO2A1* using 2-fold dilution series (10, 20, 40, and 80 ng) from the same DNA specimen from FFPE tissue. (*B*) Quantitative real-time PCR assay for *pks*$^+$ *E coli* DNA using 10-fold dilution series (0.0001, 0.001, 0.01, 0.1, and 1 ng) from DNA from cultured *pks*$^+$ *E coli*. Symbols indicate mean and the *error bars* show the standard deviation of CT values of quadruplicate runs. The coefficient of determination ($r^2$) in the assays for *pks*$^+$ *E coli* DNA and *SLCO2A1* is shown.

We confirmed that the PCR product had sequence of the *clbB* gene of *pks*$^+$ *E coli* in all 3 patients. In 2 patients with colorectal carcinoma with detectable *pks*$^+$ *E coli*, the cycle threshold (Ct) values for *pks*$^+$ *E coli* and *SLCO2A1* decreased linearly with the amount of input DNA (in a log scale) from the same specimens ($r^2 > 0.95$) (Figure 2*A*). We also confirmed that the Ct values (for *pks*$^+$ *E coli*) decreased linearly ($r^2 > 0.99$) with the amount of input DNA (in a log scale) from *pks*$^+$ *E coli* DNA (American Type Culture Collection, Manassas, VA) (Figure 2*B*) and that there was no amplification of DNA from *E coli* without the *pks* island (DH10B) (Thermo Fisher Scientific) as a negative control. These positive and negative controls were used in each PCR run on the HPFS and the NHS specimens. Furthermore, in 6 patients with colorectal carcinoma (3 positive and 3 negative for *pks*$^+$ *E coli* DNA), the interassay coefficient of variation of Ct values from each specimen was <1% for both targets in repeated assays of 5 different batches (Supplementary Table 2).

In the cohort cases, each specimen was analyzed in duplicate for each target in a single batch, and we used the mean of the 2 Ct values for each target. The amount of *pks*$^+$ *E coli* was calculated as a relative unit-less value normalized with *SLCO2A1* using the $2^{-\Delta Ct}$ method (where $\Delta Ct$ = the average Ct value of *pks*$^+$ *E coli* − the average Ct value of *SLCO2A1*), as previously described.[23] Cases with detectable *pks*$^+$ *E coli* were dichotomized into high level vs low level based on the median cutoff point. Microsatellite instability (MSI) status was determined based on PCR of 10 microsatellite markers (D2S123, D5S346, D17S250, BAT25, BAT26, BAT40, D18S55, D18S56, D18S67, and D18S487), as previously described.[18] CpG island methylator phenotype was determined using MethyLight assays[24] of the 8 promoter CpG islands (*CACNA1G*, *CDKN2A*, *CRABP1*, *IGF2*, *MLH1*, *NEUROG1*, *RUNX3*, and *SOCS1*).[25] Methylation level of long-interspersed nucleotide element 1 was measured using bisulfite PCR and pyrosequencing, as previously described.[26] PCR and pyrosequencing were performed for *KRAS* (codons 12, 13, 61, and 146),[27,28] *BRAF* (codon 600), and *PIK3CA* (exons 9 and 20), as previously described.[29]

## Statistical Analysis

Detailed statistical analysis methods are described in the Supplementary Methods. All statistical analyses were performed using SAS 9.4 software (SAS Institute, Inc, Cary, NC), and all *P* values were 2-sided. We adjusted the 2-sided $\alpha$ level to 0.012 (~0.05/4) for multiple hypothesis testing by Bonferroni correction, considering our use of 1 heterogeneity trend test (for levels of *pks*$^+$ *E coli*) and 3 stratum-specific (high, low, and negative *pks*$^+$ *E coli*) statistical trend tests.

The Western-style diet was derived by principal component analyses of the extensive diet data, as previously described and validated.[12,14] Each participant was assigned a factor score, determined by adding the reported frequencies of food item intakes weighted by the factor loadings (Table 1). To capture long-term habitual consumption, we calculated the cumulative mean of the Western diet scores from all data-available preceding food frequency questionnaires up to each questionnaire cycle. Table 1 reports the distribution of the Western diet score in each cohort.

To limit the number of primary hypotheses, our primary hypothesis testing was the assessment of heterogeneity of the association of the Western diet score with the incidence of colorectal cancer subclassified by tissue bacterial amount. We examined heterogeneity across the ordinal tumor subtypes (by the 1 degree-of-freedom statistical trend test for negative vs low vs high) in the multivariable-adjusted duplication-method Cox proportional hazards model using the meta-regression method with a subtype-specific random effect term.[30] For statistical trend tests, the diet score was used as a continuous variable with cohort-specific ceilings at the 10th and 90th percentiles to

**Table 1.** Distribution of Western Diet Scores and Factor Loading Matrix for Western-Style Diet in Each Cohort

| Variable | Health Professionals Follow-up Study | Nurses' Health Study |
|---|---|---|
| Distribution (percentile) | | |
| Minimum | −3.58 | −3.98 |
| 1% | −1.51 | −1.45 |
| 5% | −1.17 | −1.06 |
| 10% | −0.97 | −0.85 |
| 25% | −0.60 | −0.47 |
| 50% | −0.10 | 0.015 |
| 75% | 0.49 | 0.56 |
| 90% | 1.14 | 1.12 |
| 95% | 1.59 | 1.50 |
| 99% | 2.52 | 2.31 |
| Maximum | 10.0 | 9.66 |
| Food item[a] | | |
| Unprocessed red meat | 0.66 | 0.61 |
| Processed meat | 0.61 | 0.58 |
| High fat dairy food | 0.51 | 0.50 |
| French fries | 0.49 | 0.46 |
| Eggs | 0.47 | 0.41 |
| Desserts[b] | 0.43 | 0.45 |
| Condiments[c] | 0.39 | 0.36 |
| Refined grains | 0.38 | 0.38 |
| Butter | 0.38 | 0.50 |
| Mayonnaise | 0.36 | 0.34 |
| Margarine | 0.34 | 0.32 |
| Snacks[d] | 0.34 | |
| Pizza | 0.33 | 0.36 |
| Creamy soups | 0.31 | 0.32 |
| Sugar-sweetened beverages | 0.31 | 0.33 |
| Potatoes | | 0.34 |

[a]Only items with correlation coefficients >0.30 are presented. With the orthogonal rotation used, correlations are identical to factor loading matrix.
[b]Desserts includes chocolate, candy bars, cookies, brownies, cake, pie, and pastries.
[c]Condiments includes soy sauce, nondairy creamer, Worcestershire sauce, red chili sauce, and pepper.
[d]Snacks includes chips, popcorn, and crackers.

eliminate outlier effects. We also examined hazard ratios for each cancer subgroup by comparing dietary score tertiles as secondary analyses.

To control for selection bias due to tissue bacterial data availability in the 1175 cases, we used the 3200 incident colorectal cancer cases and the IPW method[31] combined with Cox proportional hazards regression models. Multivariable Cox regression models were stratified by age, sex (cohort), and questionnaire year and additionally adjusted for body mass index (continuous, with 35 kg/m$^2$ ceiling), pack-years smoked (continuous, with 50 pack-years ceiling), first-degree relative family history of colorectal cancer (yes vs no), previous colonoscopy/sigmoidoscopy (yes vs no), physical activity (continuous, with 50 metabolic equivalent task score h/wk ceiling), aspirin or nonsteroidal anti-inflammatory drug use (≥2 tablets/wk: yes vs no), multivitamin use (yes vs no), and alcohol consumption (continuous, with 30 g/d ceiling). In the NHS (female)-only analyses, we additionally adjusted for postmenopausal hormone use (yes vs no). For individuals with missing data in one questionnaire, data from preceding questionnaires were used.

In secondary analyses to assess clinical, pathologic, and molecular features according to *pks*$^+$ *E coli* status (negative, low, and high), we used the $\chi^2$ test for categorical variables, an analysis of variance for continuous variables, or Spearman correlation analysis for ordinal variables. In secondary analyses to assess patient survival, we used IPW-adjusted Kaplan-Meier analysis and multivariable Cox proportional hazards regression models (see details in the Supplementary Methods). In secondary analyses of the association of red meat variables (total, unprocessed, and processed red meat intake) with the incidence of colorectal cancer subclassified by *pks*$^+$ *E coli* status, we examined heterogeneity across the ordinal tumor subtypes (by the 1 degree-of-freedom statistical trend test for *pks*$^+$ *E coli* negative vs low vs high) in the multivariable-adjusted duplication-method Cox proportional hazards model using the meta-regression method with a subtype-specific random effect term.

## Results

We used data from 134,775 participants of the HPFS and the NHS (Table 2 and Figure 1). During 3,766,179 person-years of follow-up, we documented 3200 incident colorectal cancer cases. In multivariable analyses using each cohort, the Western diet score was weakly associated with colorectal cancer incidence (Supplementary Table 3). Because the results were similar in the 2 cohorts ($P_{\text{heterogeneity}} > .6$), we combined the 2 cohorts for further analyses to maximize statistical power while adjusting for cohort (ie, sex).

We developed and validated the assay to quantify *pks*$^+$ *E coli* DNA in tumor tissue. The assay, which was successfully conducted in duplicate in 1175 patients among the 3200 patients with colorectal cancer, detected *pks*$^+$ *E coli* in 111 patients, whereas 1064 patients were negative for this bacterium. Clinical, pathologic, and molecular features according to the amount of *pks*$^+$ *E coli* in colorectal carcinoma tissue are summarized in Table 3. The amount of *pks*$^+$ *E coli* DNA was inversely associated with American Joint Committee on Cancer stage ($P = .008$) but not with the other features examined.

We examined the association of the Western diet score with colorectal cancer incidence, using all 3200 incident cases, the 1175 cases with bacterial data, and the remaining 2025 cases without bacterial data (Supplementary Table 4). There was no substantial difference in the results from these 3 analyses. To adjust for selection bias due to bacterial data availability, we used the IPW method[31] on the 3200 patients for further analyses.

Our analysis showed that the association of Western diet scores with colorectal cancer incidence differed by tissue *pks*$^+$ *E coli* levels ($P_{\text{heterogeneity}} = .014$) (Table 4), and was stronger for tumors containing higher-level *pks*$^+$ *E coli*. Multivariable hazard ratios in individuals with scores in the highest (vs the lowest) tertile of Western diet scores were 3.45 (95% confidence interval [CI], 1.53–7.78; $P_{\text{trend}} = .001$

**Table 2.** Age-Standardized Characteristics According to Western Diet Score Tertiles in the Health Professionals Follow-up Study (Men, 1986–2014) and the Nurses' Health Study (Women, 1980–2014)

| | Health Professionals Follow-up Study | | | Nurses' Health Study | | |
|---|---|---|---|---|---|---|
| | Western diet score | | | Western diet score | | |
| Characteristic[a] | Tertile 1 | Tertile 2 | Tertile 3 | Tertile 1 | Tertile 2 | Tertile 3 |
| Participants, No. | 17,429 | 14,217 | 15,803 | 27,645 | 27,702 | 31,979 |
| Mean age, *y* | 65.1 | 64.7 | 63.3 | 62.8 | 61.3 | 60.0 |
| Mean body mass index, *kg/m*$^2$ | 25.3 | 25.9 | 26.3 | 24.9 | 25.3 | 25.9 |
| Mean physical activity, *METS h/wk*[b] | 29.6 | 27.2 | 26.2 | 12.4 | 10.5 | 9.3 |
| Mean pack-years smoked | 9.0 | 11.3 | 15.2 | 11.6 | 12.5 | 13.8 |
| Family history of colorectal cancer | 15.1 | 14.7 | 14.5 | 19.1 | 18.8 | 18.9 |
| Previous endoscopy | 25.1 | 24.9 | 22.6 | 28.6 | 27.6 | 26.3 |
| Current multivitamin use | 48.5 | 47.8 | 44.3 | 55.1 | 52.3 | 48.2 |
| Regular aspirin or NSAID use[c] | 43.7 | 48.4 | 49.4 | 57.7 | 60.6 | 60.8 |
| Postmenopausal | … | … | … | 74.8 | 74.0 | 72.7 |
| Current hormone use[d] | … | … | … | 47.1 | 44.8 | 42.4 |
| Dietary intake, mean | | | | | | |
| Total calorie intake, *kcal/d* | 1,610 | 1,900 | 2,420 | 1,430 | 1,620 | 1,990 |
| Unprocessed red meat, *servings/d* | 0.30 | 0.56 | 0.88 | 0.45 | 0.63 | 0.82 |
| Processed red meat, *servings/d* | 0.12 | 0.28 | 0.56 | 0.15 | 0.28 | 0.47 |
| Poultry, *servings/d* | 0.40 | 0.39 | 0.39 | 0.34 | 0.31 | 0.31 |
| Fruit, *servings/d* | 2.82 | 2.39 | 2.25 | 2.49 | 2.18 | 2.07 |
| Vegetable, *servings/d* | 3.31 | 3.15 | 3.26 | 2.88 | 2.62 | 2.63 |
| Alcohol, *g/d* | 8.0 | 11.2 | 13.9 | 6.5 | 6.1 | 5.5 |
| Folate, *μg/d* | 619 | 545 | 481 | 494 | 424 | 371 |
| Calcium, *mg/d* | 1,030 | 928 | 860 | 1,070 | 927 | 819 |
| Vitamin D, *IU/d* | 515 | 428 | 368 | 429 | 355 | 299 |
| Dietary fiber, *g/d* | 25.7 | 21.5 | 19.1 | 18.9 | 16.1 | 14.3 |

NOTE. Data are shown as percentages unless indicated otherwise.
METS, metabolic equivalent task score; NSAID, nonsteroidal anti-inflammatory drug.
[a]Updated information throughout follow-up was used to calculate the mean for continuous variables and the percentage for categorical variables. All variables are age-standardized except age.
[b]Physical activity is represented by the product sum of the METS of each specific recreational activity and hours spent on that activity per week.
[c]Regular users are defined as ≥2 standard (325-mg) tablets of aspirin or ≥2 tablets of NSAIDs per week.
[d]Proportion of current menopausal hormone use is calculated among postmenopausal women only.

across the tertiles) for colorectal cancer with high-level *pks*$^+$ *E coli*, 1.22 (95% CI, 0.57–2.63) for cancer with low-level *pks*$^+$ *E coli*, and 1.10 (95% CI, 0.85–1.42) for cancer without detectable *pks*$^+$ *E coli*. In a sensitivity analysis, we confirmed that the analysis without IPW yielded results (Supplementary Table 5) similar to the IPW-adjusted analysis.

In secondary subgroup analyses, we found similar differential associations by *pks*$^+$ *E coli* status in men and women (Table 5). In analyses using patients stratified by tumor MSI status (Supplementary Table 6), the differential association by *pks*$^+$ *E coli* status was apparent for the non–MSI-high subtype, whereas statistical power was limited for the MSI-high subtype.

In secondary analyses to assess the prognostic association of the amount of *pks*$^+$ *E coli*, we conducted survival analysis using the Kaplan-Meier method and Cox proportional hazard regression (Supplementary Figure 1 and Supplementary Table 7). In univariable analyses of colorectal cancer–specific survival, there was a statistically insignificant favorable prognostic association of the amount of *pks*$^+$ *E coli* ($P_{\text{trend}} = .028$, with the $\alpha$ level of 0.012), which did not persist in multivariable analyses ($P_{\text{trend}} > .16$).

We further examined whether red meat variables (total, unprocessed, and processed red meat intake amounts), which was the largest component of the Western diet score, might be differentially associated with colorectal cancer by *pks*$^+$ *E coli* status (Supplementary Table 8). We found that

**Table 3.** Clinical, Pathologic, and Molecular Characteristics of Colorectal Cancer Cases According to the Amount of *pks*$^+$ *Escherichia coli* DNA in Colorectal Cancer Tissue

| | | Amount of *pks*$^+$ *E coli* DNA in colorectal cancer tissue | | | |
|---|---|---|---|---|---|
| | All cases | Negative | Low | High | |
| Characteristic[a] | (n = 1175) | (n = 1064) | (n = 55) | (n = 56) | *P* value[b] |
| Sex | | | | | .28 |
| Female (NHS) | 656 (56) | 588 (55) | 31 (56) | 37 (66) | |
| Male (HPFS) | 519 (44) | 476 (45) | 24 (44) | 19 (34) | |
| Age, *y* | 69.0 ± 8.8 | 68.9 ± 8.8 | 69.6 ± 10.0 | 69.9 ± 8.1 | .61 |
| Year of diagnosis | | | | | .09 |
| ≤1995 | 399 (34) | 371 (35) | 12 (22) | 16 (29) | |
| 1996–2000 | 375 (32) | 333 (31) | 18 (33) | 24 (43) | |
| 2001–2008 | 401 (34) | 360 (34) | 25 (45) | 16 (29) | |
| Family history of colorectal cancer in first-degree relative(s) | | | | | .17 |
| Absent | 935 (80) | 854 (81) | 42 (76) | 39 (71) | |
| Present | 235 (20) | 206 (19) | 13 (24) | 16 (29) | |
| Tumor location | | | | | .70 |
| Cecum | 202 (17) | 182 (17) | 11 (20) | 9 (16) | |
| Ascending to transverse | 364 (31) | 334 (32) | 16 (29) | 14 (25) | |
| Descending to sigmoid | 355 (30) | 314 (30) | 19 (35) | 22 (39) | |
| Rectum | 249 (21) | 229 (22) | 9 (16) | 11 (20) | |
| AJCC disease stage | | | | | .008 |
| I | 262 (24) | 230 (23) | 12 (24) | 20 (42) | |
| II | 354 (33) | 319 (32) | 19 (37) | 16 (33) | |
| III | 311 (29) | 288 (29) | 14 (27) | 9 (19) | |
| IV | 157 (14) | 148 (15) | 6 (12) | 3 (6.3) | |
| Tumor size, *cm* | 4.4 ± 2.0 | 4.4 ± 2.0 | 4.7 ± 2.0 | 4.5 ± 2.1 | .39 |
| Tumor differentiation | | | | | .42 |
| Well to moderate | 1053 (90) | 954 (90) | 47 (85) | 52 (93) | |
| Poor | 118 (10) | 106 (10) | 8 (15) | 4 (7.1) | |
| MSI status | | | | | .49 |
| Non-MSI high | 947 (83) | 860 (83) | 40 (78) | 47 (87) | |
| MSI high | 188 (17) | 170 (17) | 11 (22) | 7 (13) | |
| CIMP status | | | | | .63 |
| Low/negative | 885 (82) | 803 (82) | 36 (82) | 46 (87) | |
| High | 197 (18) | 182 (18) | 8 (18) | 7 (13) | |
| LINE-1 methylation level | 63.0 ± 9.8 | 63.0 ± 9.8 | 63.4 ± 11.5 | 63.5 ± 7.2 | .88 |
| *KRAS* mutation | | | | | .50 |
| Wild-type | 645 (59) | 586 (59) | 30 (64) | 29 (53) | |
| Mutant | 443 (41) | 400 (41) | 17 (36) | 26 (47) | |
| *BRAF* mutation | | | | | .55 |
| Wild-type | 942 (84) | 852 (84) | 41 (82) | 49 (89) | |
| Mutant | 177 (16) | 162 (16) | 9 (18) | 6 (11) | |
| *PIK3CA* mutation | | | | | 0.51 |
| Wild-type | 878 (84) | 795 (84) | 39 (85) | 44 (90) | |
| Mutant | 168 (16) | 156 (16) | 7 (15) | 5 (10) | |

NOTE. Data are presented as number (%) or as the mean ± standard deviation.
AJCC, American Joint Committee on Cancer; CIMP, CpG island methylator phenotype; LINE-1, long-interspersed nucleotide element 1.
[a]Percentage indicates the proportion of patients with a specific clinical, pathologic, or molecular characteristic among all patients or in strata of the amount of *pks*$^+$ *E coli* DNA in colorectal cancer tissue.
[b]To assess associations between the categories (negative, low, and high) of *pks*$^+$ *E coli* DNA in colorectal cancer tissue and categorical data, the $\chi^2$ test was performed. To compare age, and LINE-1 methylation level, an analysis of variance was performed. To compare AJCC disease stage, Spearman analysis was performed.

**Table 4.** Incidence of Colorectal Cancer by *pks*+ *Escherichia coli* Status in Relation to Cumulative Average Western Diet Score in the Combined Cohorts of the Health Professionals Follow-up Study (1986–2014) and the Nurses' Health Study (1980–2014)

| Variable | Western diet score | | | *P* for trend[a] | *P* for heterogeneity[b] |
|---|---|---|---|---|---|
| | Tertile 1 | Tertile 2 | Tertile 3 | | |
| Person-years | 1,255,030 | 1,254,558 | 1,256,591 | | |
| Overall colorectal cancer | | | | | |
| Cases, No. (total n = 1175) | 392 | 391 | 392 | | |
| Age-adjusted HR (95% CI)[c] | 1 [Referent] | 1.04 (0.93–1.15) | 1.28 (1.14–1.44) | <.001 | |
| Multivariable HR (95% CI)[d] | 1 [Referent] | 0.98 (0.88–1.09) | 1.14 (1.01–1.29) | .010 | |
| *pks*+ *E coli* status | | | | | .014 |
| *pks*+ *E coli* negative | | | | | |
| Cases, No. (total n = 1064) | 364 | 354 | 346 | | |
| Age-adjusted HR (95% CI)[c] | 1 [Referent] | 1.00 (0.82–1.21) | 1.23 (0.96–1.58) | .068 | |
| Multivariable HR (95% CI)[d] | 1 [Referent] | 0.95 (0.78–1.15) | 1.10 (0.85–1.42) | .40 | |
| *pks*+ *E coli* low | | | | | |
| Cases, No. (total n = 55) | 18 | 15 | 22 | | |
| Age-adjusted HR (95% CI)[c] | 1 [Referent] | 0.82 (0.36–1.83) | 1.37 (0.64–2.97) | .51 | |
| Multivariable HR (95% CI)[d] | 1 [Referent] | 0.77 (0.35–1.73) | 1.22 (0.57–2.63) | .76 | |
| *pks*+ *E coli* high | | | | | |
| Cases, No. (total n = 56) | 10 | 22 | 24 | | |
| Age-adjusted HR (95% CI)[c] | 1 [Referent] | 2.24 (1.00–5.04) | 3.83 (1.69–8.66) | <.001 | |
| Multivariable HR (95% CI)[d] | 1 [Referent] | 2.11 (0.94–4.73) | 3.45 (1.53–7.78) | .001 | |

HR, hazard ratio.
[a]The trend test was performed using the Western diet score as a continuous variable with cohort-specific ceilings at 10th and 90th percentiles in the regression model. The 90th and 10th percentile values were used for scores >90th percentile and those <10th percentile, respectively, to eliminate outlier effects.
[b]The meta-regression method with a subtype-specific random effect term was used to test whether the association has a trend across the ordinal subtypes in the multivariable-adjusted model, where the Western diet score was used as a continuous variable with cohort-specific ceilings at 10th and 90th percentiles.
[c]Duplication-method Cox proportional hazards model weighted by inverse probabilities based on tissue bacterial data availability for competing risks data was used with total caloric intake adjusted and stratification by age (in months), sex (ie, cohort), and year of questionnaire return.
[d]Additionally adjusted for body mass index (continuous, with 35 kg/m$^2$ ceiling), cumulative pack-years smoked (continuous, with 50 pack-years ceiling), family history of colorectal cancer in any first-degree relative (yes vs no), previous lower gastrointestinal endoscopy (yes vs no), physical activity (continuous, with a ceiling at 50 metabolic equivalent task score hours/wk), regular use of aspirin or nonsteroidal anti-inflammatory drugs (>2 tablets/wk: yes vs no), multivitamin use (yes vs no), and alcohol consumption (continuous, with 30 g/d ceiling).

the differential association by the amount of *pks*+ *E coli* was not statistically significant for any of these red meat variables ($P_{\text{heterogeneity}} > .05$).

## Discussion

Colorectal cancer is a heterogeneous group of neoplastic diseases influenced by many factors, including diet, lifestyle, and intestinal microbiota.[32–37] Using 2 prospective cohort studies in the US with 3 decades of follow-up, we discovered a stronger association of the Western-style diet with the incidence of colorectal carcinoma containing higher amounts of *pks*+ *E coli*. Our findings provide evidence for a Western-style diet characterized by high intake of red and processed meat, sugar, and refined grains as a risk factor for colorectal cancer, especially its subtype containing a high amount of *pks*+ *E coli*. Our novel data can inform research efforts devoted to developing cancer prevention strategies that modify diet and the intestinal microbiome.

Previous metagenomic studies have cast light on the role of the intestinal microbiome in colorectal carcinogenesis.[38–40] Molecular pathologic analyses of colorectal cancer have also supported the role of specific intestinal microbes, such as colibactin-producing *pks*+ *E coli*, in tumor development.[41,42] A recent study has elucidated the structure of colibactin and enabled the synthesis of colibactin.[6] Experimental studies indicate that the genotoxic colibactin can alkylate DNA on adenine residues[7] and induce double-strand breaks, leading to a specific mutational signature.[8,9] Another study showed that organoids that recovered from short-term infection with *pks*+ *E coli* reveal characteristics of colorectal carcinoma cells, such as enhanced proliferation, WNT-independence, and impaired differentiation, at least in part through alterations in TP53-signaling.[43] In addition, evidence suggests that *pks*+ *E coli* suppresses the host immune response in the tumor microenvironment.[44] Taken together, although *pks*+ *E coli* likely plays a role in colorectal carcinogenesis, it currently remains uncertain when and how *pks*+ *E coli* exerts an effect on tumor development. Investigating

**Table 5.** Incidence of Colorectal Cancer by *pks*$^+$ *Escherichia coli* Status in Relation to Cumulative Average Western Diet Score in the Health Professionals Follow-up Study (1986–2014) and the Nurses' Health Study (1980–2014)

| Variable | Western diet score | | | *P* for trend[a] | *P* for heterogeneity[b] |
|---|---|---|---|---|---|
| | Tertile 1 | Tertile 2 | Tertile 3 | | |
| HPFS (men) | | | | | |
| Person-years | 365,506 | 365,602 | 365,680 | | |
| *pks*$^+$ *E coli* status | | | | | 0.71 |
| *pks*$^+$ *E coli* negative | | | | | |
| Cases, No. (total n = 476) | 156 | 158 | 162 | | |
| Age-adjusted HR (95% CI)[c] | 1 [Referent] | 1.19 (0.95–1.50) | 1.41 (1.09–1.82) | <.001 | |
| Multivariable HR (95% CI)[d] | 1 [Referent] | 1.12 (0.89–1.42) | 1.26 (0.96–1.65) | .015 | |
| *pks*$^+$ *E coli* positive | | | | | |
| Cases, No. (total n = 43) | 10 | 15 | 18 | | |
| Age-adjusted HR (95% CI)[c] | 1 [Referent] | 1.77 (0.80–3.95) | 2.33 (1.05–5.14) | .042 | |
| Multivariable HR (95% CI)[d] | 1 [Referent] | 1.62 (0.72–3.61) | 2.05 (0.93–4.50) | .10 | |
| NHS (women) | | | | | |
| Person-years | 889,524 | 888,957 | 890,911 | | |
| *pks*$^+$ *E coli* status | | | | | .018 |
| *pks*$^+$ *E coli* negative | | | | | |
| Cases, No. (total n = 588) | 208 | 196 | 184 | | |
| Age-adjusted HR (95% CI)[c] | 1 [Referent] | 0.95 (0.75–1.20) | 1.18 (0.87–1.61) | .28 | |
| Multivariable HR (95% CI)[d] | 1 [Referent] | 0.89 (0.70–1.13) | 1.03 (0.75–1.41) | .90 | |
| *pks*$^+$ *E coli* positive | | | | | |
| Cases, No. (total n = 68) | 18 | 22 | 28 | | |
| Age-adjusted HR (95% CI)[c] | 1 [Referent] | 1.17 (0.59–2.31) | 2.09 (1.07–4.09) | .019 | |
| Multivariable HR (95% CI)[d] | 1 [Referent] | 1.10 (0.55–2.18) | 1.81 (0.92–3.54) | .058 | |

HR, hazard ratio.
[a]The trend test was performed using the Western diet score as a continuous variable with cohort-specific ceilings at 10th and 90th percentiles in the regression model. The 90th and 10th percentile values were used for scores >90th percentile and those <10th percentile, respectively, to eliminate outlier effects.
[b]The meta-regression method with a subtype-specific random effect term was used to test whether the association has a trend across the ordinal subtypes (negative vs low vs high) in the multivariable-adjusted model, where the Western diet score was used as a continuous variable with cohort-specific ceilings at 10th and 90th percentiles.
[c]Duplication-method Cox proportional hazards model weighted by inverse probabilities based on tissue bacterial data availability for competing risks data was used with total caloric intake adjusted and stratification by age (in months) and year of questionnaire return.
[d]Additionally adjusted for body mass index (continuous, with 35 kg/m$^2$ ceiling), cumulative pack-years smoked (continuous, with 50 pack-years ceiling), family history of colorectal cancer in any first-degree relative (yes vs no), previous lower gastrointestinal endoscopy (yes vs no), physical activity (continuous, with a ceiling at 50 metabolic equivalent task score h/wk), regular use of aspirin or nonsteroidal anti-inflammatory drugs (>2 tablets/week: yes vs no), multivitamin use (yes vs. no), and alcohol consumption (continuous, with 30 g/day ceiling). We additionally adjusted for postmenopausal hormone use (yes vs no) for the NHS analysis.

the detailed mechanism and the associations of this bacterium with lifestyle and dietary risk factors is of particular interest.

Dietary influences on the microbiome in stool and colonic tissue have been investigated. Experimental studies have shown that daily microbiome variation is related to food group choices[45] and that a high-fat diet can alter intestinal bacterial composition[11] and lead to the development of systemic inflammation.[46,47] Observational studies have found relationships between a low-quality diet and an inflammatory diet with intestinal dysbiosis[48,49] as well as between a Western-style diet and a high level of plasma-soluble CD14, a biomarker of mucosal barrier dysfunction.[50] These lines of evidence suggest that dietary factors can influence intestinal microbial composition and inflammatory status.

The prior principal component analysis on diet data in the population revealed 2 dominant dietary patterns,; namely, the Western-style pattern and the prudent dietary pattern.[51] A meta-analysis indicates a weak-to-moderate association between the Western-style diet and colorectal cancer risk.[10] In contrast, the prudent dietary pattern, characterized by high intake of fruits, vegetables, fish, poultry, and whole grains, has been inversely associated with colorectal cancer risk.[10] Nonetheless, the strength of the association remains uncertain due to residual or unmeasured confounding by other healthy or unhealthy behaviors associated with the dietary patterns.

Using the molecular pathologic epidemiology approach,[33,52–54] we found a strong association between the Western diet and the colorectal cancer subgroup containing high levels of *pks*$^+$ *E coli*. This specific link between the

Western diet and *pks*$^+$ *E coli* suggests potentially interactive carcinogenic effects. In further analysis of the red meat variable, we did not observe a statistically significant association of any red meat variable with the incidence of colorectal cancer by *pks*$^+$ *E coli* status. Our data suggest that red meat intake by itself is unlikely the sole factor that contributed to the differential association of the Western diet with colorectal cancer by *pks*$^+$ *E coli* status. One possibility is that the Western diet may promote the proliferation and activity of *pks*$^+$ *E coli* or strengthen the carcinogenic effects of *pks*$^+$ *E coli*, or both, through alteration of the local tissue microenvironment. It is evident that the molecular pathologic epidemiology approach allows for the generation of intriguing hypotheses based on human population data. Although our analyses showed the correlation between the Western diet and the incidence of colorectal cancer containing high abundance of *pks*$^+$ *E coli*, a replication using additional independent cohorts and experimental research is necessary.

In addition, we found that the association of the Western diet with colorectal cancer incidence according to *pks*$^+$ *E coli* might be different by sex (Table 5). Although intriguing, those results were obtained by our secondary subgroup analyses, and as such, generalizability needs to be tested in independent data sets. If replicated, our findings may inform differential interactive influences of the Western diet and *pks*$^+$ *E coli* in men vs women. While the mechanisms underlying these sex-specific effects remain to be elucidated, differences in biological features of colorectal cancer between men and women have been demonstrated.[55–57] Additional studies are warranted to investigate how the Western diet and *pks*$^+$ *E coli* may exert interactive carcinogenic effects and which specific food items might contribute to the observed differential associations between the Western diet and colorectal cancer incidence according to *pks*$^+$ *E coli* status.

We acknowledge limitations in the current study. First, unmeasured or residual confounding, or both, might have substantially influenced our findings. We included most established risk factors in our analysis models, with little evidence for substantial confounding by the included variables.

Second, tissue bacterial data were unavailable for some incident cancer cases within the cohorts, which might have caused selection bias. However, by using all 3200 incident colorectal cancers and the IPW method,[31] we were able to adjust for selection bias with the available covariates. Analyses with and without the IPW adjustment yielded similar results.

Third, measurement errors were inherently present in the assessments of diet and tissue bacterial amounts, particularly with the use of FFPE tissue specimens. We used repeated assessments of diet every 4 years, which allowed us to estimate the effects of long-term dietary patterns. For bacterial analyses, we carefully optimized and validated our quantitative PCR assay for FFPE tissue specimens to ensure high analytical sensitivity and specificity. Our validation study also demonstrated a high linearity ($r^2 > 0.95$) and high precision (with $<1\%$ interassay coefficient of variation) of the assay.

Fourth, our cohort populations mainly consisted of non-Hispanic Whites, and thus, our findings need to be replicated in independent populations.

Fifth, we used information on microbial contents in tumor tissue, which was not prospectively collected, unlike dietary data. Therefore, establishing a cause-and-effect relationship between the microbial species and colorectal cancer requires additional studies.

Finally, our findings were based on the observational cohort studies, which had certain inherent limitations in data collections. Hence, additional epidemiologic studies and experimental confirmation are ultimately needed.

There exist notable strengths in the current study. First, our dietary data were prospectively and repeatedly collected for >30 years through validated food frequency questionnaires.[58]

Second, our prospective cohort design enabled the collection of diet and other lifestyle data without knowing who would develop colorectal cancer later, thereby eliminating differential recall bias between patients with cancer and individuals who were free of cancer.

Third, the prospective study design also enabled us to leverage all 3200 incident colorectal cancer cases with the IPW method to adjust for selection bias caused by tissue bacterial data availability.

Fourth, we used molecular pathologic epidemiology methods, which can provide novel etiologic insights into diet and bacterial species, thereby augmenting causal inference.

Fifth, the cancer patient group was assembled from hundreds of hospitals located throughout the US, which increases the generalizability of our findings in contrast to studies based on only 1 or a few hospitals. Nonetheless, our findings should be replicated in independent populations.

## Conclusion

We have found that the association of the Western diet with colorectal cancer incidence is stronger for tumors containing higher amounts of *pks*$^+$ *E coli*. Our findings provide evidence supporting the role of the gut microbiota in mediating the pathogenic link between diet and colorectal cancer. This study also underscores the importance of diet as a modifiable factor that may contribute to cancer prevention.

## Supplementary Material

Note: To access the supplementary material accompanying this article, visit the online version of *Gastroenterology* at www.gastrojournal.org, and at https://dx.doi.org/10.1053/j.gastro.2022.06.054.

## References

1. Nougayrede JP, Homburg S, Taieb F, et al. *Escherichia coli* induces DNA double-strand breaks in eukaryotic cells. Science 2006;313:848–851.
2. Wu S, Rhee KJ, Albesiano E, et al. A human colonic commensal promotes colon tumorigenesis via activation

of T helper type 17 T cell responses. Nat Med 2009; 15:1016–1022.
3. Arthur JC, Perez-Chanona E, Muhlbauer M, et al. Intestinal inflammation targets cancer-inducing activity of the microbiota. Science 2012;338:120–123.
4. Dejea CM, Fathi P, Craig JM, et al. Patients with familial adenomatous polyposis harbor colonic biofilms containing tumorigenic bacteria. Science 2018; 359:592–597.
5. Barrett M, Hand CK, Shanahan F, et al. Mutagenesis by microbe: the role of the microbiota in shaping the cancer genome. Trends Cancer 2020;6:277–287.
6. Xue M, Kim CS, Healy AR, et al. Structure elucidation of colibactin and its DNA cross-links. Science 2019; 365:365:eaax2685.
7. Wilson MR, Jiang Y, Villalta PW, et al. The human gut bacterial genotoxin colibactin alkylates DNA. Science 2019;363:eaar7785.
8. **Pleguezuelos-Manzano C, Puschhof J, Rosendahl Huber A**, et al. Mutational signature in colorectal cancer caused by genotoxic pks$^{+}$ *E. coli*. Nature 2020; 580:269–273.
9. Dziubanska-Kusibab PJ, Berger H, Battistini F, et al. Colibactin DNA-damage signature indicates mutational impact in colorectal cancer. Nat Med 2020; 26:1063–1069.
10. Magalhaes B, Peleteiro B, Lunet N. Dietary patterns and colorectal cancer: systematic review and meta-analysis. Eur J Cancer Prev 2012;21:15–23.
11. O'Keefe SJ, Li JV, Lahti L, et al. Fat, fibre and cancer risk in African Americans and rural Africans. Nat Commun 2015;6:6342.
12. Fung T, Hu FB, Fuchs C, et al. Major dietary patterns and the risk of colorectal cancer in women. Arch Intern Med 2003;163:309–314.
13. **Nishihara R, Wu K, Lochhead P**, et al. Long-term colorectal-cancer incidence and mortality after lower endoscopy. N Engl J Med 2013;369:1095–1105.
14. Hu FB, Rimm EB, Stampfer MJ, et al. Prospective study of major dietary patterns and risk of coronary heart disease in men. Am J Clin Nutr 2000;72:912–921.
15. Feskanich D, Rimm EB, Giovannucci EL, et al. Reproducibility and validity of food intake measurements from a semiquantitative food frequency questionnaire. J Am Diet Assoc 1993;93:790–796.
16. Yuan C, Spiegelman D, Rimm EB, et al. Relative validity of nutrient intakes assessed by questionnaire, 24-hour recalls, and diet records as compared with urinary recovery and plasma concentration biomarkers: findings for women. Am J Epidemiol 2018; 187:1051–1063.
17. Al-Shaar L, Yuan C, Rosner B, et al. Reproducibility and validity of a semiquantitative food frequency questionnaire in men assessed by multiple methods. Am J Epidemiol 2021;190:1122–1132.
18. **Yamauchi M, Morikawa T, Kuchiba A**, et al. Assessment of colorectal cancer molecular features along bowel subsites challenges the conception of distinct dichotomy of proximal versus distal colorectum. Gut 2012;61:847–854.
19. Yamauchi M, Lochhead P, Morikawa T, et al. Colorectal cancer: a tale of two sides or a continuum? Gut 2012; 61:794–797.
20. Inamura K, Yamauchi M, Nishihara R, et al. Prognostic significance and molecular features of signet-ring cell and mucinous components in colorectal carcinoma. Ann Surg Oncol 2015;22:1226–1235.
21. **Mima K, Sukawa Y, Nishihara R**, et al. *Fusobacterium nucleatum* and T cells in colorectal carcinoma. JAMA Oncol 2015;1:653–661.
22. **Fujiyoshi K, Bruford EA, Mroz P**, et al. Opinion: Standardizing gene product nomenclature—a call to action. Proc Natl Acad Sci U S A 2021;118:e2025207118.
23. Schmittgen TD, Livak KJ. Analyzing real-time PCR data by the comparative C(T) method. Nat Protoc 2008; 3:1101–1118.
24. Ogino S, Kawasaki T, Brahmandam M, et al. Precision and performance characteristics of bisulfite conversion and real-time PCR (MethyLight) for quantitative DNA methylation analysis. J Mol Diagn 2006;8:209–217.
25. **Nosho K, Irahara N, Shima K**, et al. Comprehensive biostatistical analysis of CpG island methylator phenotype in colorectal cancer using a large population-based sample. PLoS One 2008;3:e3698.
26. **Irahara N, Nosho K, Baba Y**, et al. Precision of pyrosequencing assay to measure LINE-1 methylation in colon cancer, normal colonic mucosa, and peripheral blood cells. J Mol Diagn 2010;12:177–183.
27. Ogino S, Kawasaki T, Brahmandam M, et al. Sensitive sequencing method for KRAS mutation detection by Pyrosequencing. J Mol Diagn 2005;7:413–421.
28. **Imamura Y, Lochhead P, Yamauchi M**, et al. Analyses of clinicopathological, molecular, and prognostic associations of KRAS codon 61 and codon 146 mutations in colorectal cancer: cohort study and literature review. Mol Cancer 2014;13:135.
29. **Liao X, Lochhead P, Nishihara R**, et al. Aspirin use, tumor PIK3CA mutation, and colorectal-cancer survival. N Engl J Med 2012;367:1596–1606.
30. Wang M, Spiegelman D, Kuchiba A, et al. Statistical methods for studying disease subtype heterogeneity. Stat Med 2016;35:782–800.
31. **Liu L, Nevo D, Nishihara R**, et al. Utility of inverse probability weighting in molecular pathological epidemiology. Eur J Epidemiol 2018;33:381–392.
32. **Hamada T, Nowak JA, Milner DA Jr**, et al. Integration of microbiology, molecular pathology, and epidemiology: a new paradigm to explore the pathogenesis of microbiome-driven neoplasms. J Pathol 2019;247:615–628.
33. **Akimoto N, Ugai T, Zhong R**, et al. Rising incidence of early-onset colorectal cancer—a call to action. Nat Rev Clin Oncol 2021;18:230–243.
34. Gunter MJ, Alhomoud S, Arnold M, et al. Meeting report from the joint IARC-NCI international cancer seminar series: a focus on colorectal cancer. Ann Oncol 2019; 30:510–519.
35. Tilg H, Adolph TE, Gerner RR, et al. The intestinal microbiota in colorectal cancer. Cancer Cell 2018;33:954–964.
36. Rajpoot M, Sharma AK, Sharma A, et al. Understanding the microbiome: emerging biomarkers for exploiting the

microbiota for personalized medicine against cancer. Semin Cancer Biol 2018;52:1–8.
37. Li J, Zhang AH, Wu FF, et al. Alterations in the gut microbiota and their metabolites in colorectal cancer: recent progress and future prospects. Front Oncol 2022; 12:841552.
38. Yachida S, Mizutani S, Shiroma H, et al. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. Nat Med 2019;25:968–976.
39. Thomas AM, Manghi P, Asnicar F, et al. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. Nat Med 2019;25:667–678.
40. Ternes D, Karta J, Tsenkova M, et al. Microbiome in colorectal cancer: how to get from meta-omics to mechanism? Trends Microbiol 2020;28:401–423.
41. Thakur BK, Malaise Y, Martin A. Unveiling the mutational mechanism of the bacterial genotoxin colibactin in colorectal cancer. Mol Cell 2019;74:227–229.
42. Arthur JC. Microbiota and colorectal cancer: colibactin makes its mark. Nat Rev Gastroenterol Hepatol 2020; 17:317–318.
43. Iftekhar A, Berger H, Bouznad N, et al. Genomic aberrations after short-term exposure to colibactin-producing *E. coli* transform primary colon epithelial cells. Nat Commun 2021;12:1003.
44. Lopes A, Billard E, Casse AH, et al. Colibactin-positive *Escherichia coli* induce a procarcinogenic immune environment leading to immunotherapy resistance in colorectal cancer. Int J Cancer 2020;146:3147–3159.
45. Johnson AJ, Vangay P, Al-Ghalith GA, et al. Daily sampling reveals personalized diet-microbiome associations in humans. Cell Host Microbe 2019;25:789–802 e5.
46. Wan Y, Wang F, Yuan J, et al. Effects of dietary fat on gut microbiota and faecal metabolites, and their relationship with cardiometabolic risk factors: a 6-month randomised controlled-feeding trial. Gut 2019; 68:1417–1429.
47. Groschel C, Prinz-Wohlgenannt M, Mesteri I, et al. Switching to a healthy diet prevents the detrimental effects of Western diet in a colitis-associated colorectal cancer model. Nutrients 2019;12:45.
48. Liu Y, Ajami NJ, El-Serag HB, et al. Dietary quality and the colonic mucosa-associated gut microbiome in humans. Am J Clin Nutr 2019;110:701–712.
49. Zheng J, Hoffman KL, Chen JS, et al. Dietary inflammatory potential in relation to the gut microbiome: results from a cross-sectional study. Br J Nutr 2020; 124:931–942.
50. Tabung FK, Birmann BM, Epstein MM, et al. Influence of dietary patterns on plasma soluble CD14, a surrogate marker of gut barrier dysfunction. Curr Dev Nutr 2017;1: e001396.
51. **Mehta RS, Nishihara R, Cao Y**, et al. Association of dietary patterns with risk of colorectal cancer subtypes classified by *Fusobacterium nucleatum* in tumor tissue. JAMA Oncol 2017;3:921–927.
52. Ogino S, Chan AT, Fuchs CS, et al. Molecular pathological epidemiology of colorectal neoplasia: an emerging transdisciplinary and interdisciplinary field. Gut 2011;60:397–411.
53. Ogino S, Nowak JA, Hamada T, et al. Insights into pathogenic interactions among environment, host, and tumor at the crossroads of molecular pathology and epidemiology. Annu Rev Pathol 2019;14:83–103.
54. **Mima K, Kosumi K, Baba Y**, et al. The microbiome, genetics, and gastrointestinal neoplasms: the evolving field of molecular pathological epidemiology to analyze the tumor-immune-microbiome interaction. Hum Genet 2021;140:725–746.
55. Lopes-Ramos CM, Kuijjer ML, Ogino S, et al. Gene regulatory network analysis identifies sex-linked differences in colon cancer drug metabolism. Cancer Res 2018;78:5538–5547.
56. Berntsson J, Eberhard J, Nodin B, et al. Pre-diagnostic anthropometry, sex, and risk of colorectal cancer according to tumor immune cell composition. Oncoimmunology 2019;8:e1664275.
57. Abancens M, Bustos V, Harvey H, et al. Sexual dimorphism in colon cancer. Front Oncol 2020;10:607909.
58. Rimm EB, Giovannucci EL, Stampfer MJ, et al. Reproducibility and validity of an expanded self-administered semiquantitative food frequency questionnaire among male health professionals. Am J Epidemiol 1992; 135:1114–1126; [discussion: 1127–1136].

**Author names in bold designate shared co-first authorship.**

**Received December 31, 2021. Accepted June 15, 2022.**

**Correspondence**
Address correspondence to: Shuji Ogino, MD, PhD, MS, Program in MPE Molecular Pathological Epidemiology, Department of Pathology, Brigham and Women's Hospital, 221 Longwood Avenue, Eugene Braunwald Research Center, Room 404A, Boston, Massachusetts 02115. e-mail: sogino@bwh.harvard.edu.

**Acknowledgments**
The authors would like to thank the participants and staff of the Health Professionals Follow-up Study and the Nurses' Health Study for their valuable contributions as well as the following state cancer registries for their help: Alabama, Arizona, Arkansas, California, Colorado, Connecticut, Delaware, Florida, Georgia, Idaho, Illinois, Indiana, Iowa, Kentucky, Louisiana, Maine, Maryland, Massachusetts, Michigan, Nebraska, New Hampshire, New Jersey, New York, North Carolina, North Dakota, Ohio, Oklahoma, Oregon, Pennsylvania, Rhode Island, South Carolina, Tennessee, Texas, Virginia, Washington, and Wyoming. The authors assume full responsibility for analyses and interpretation of these data.

**CRediT Authorship Contributions**
Kota Arima, MD, PhD (Data curation: Lead; Formal analysis: Supporting; Investigation: Lead; Methodology: Lead; Project administration: Supporting; Validation: Lead; Writing – original draft: Lead).

Rong Zhong, PhD (Data curation: Supporting; Formal analysis: Lead; Investigation: Supporting; Methodology: Supporting; Project administration: Supporting; Writing – original draft: Supporting).

Tomotaka Ugai, MD, PhD (Data curation: Supporting; Formal analysis: Supporting; Methodology: Supporting; Project administration: Supporting; Writing – original draft: Supporting).

Melissa Zhao, MD, MS (Data curation: Supporting; Writing – original draft: Supporting).

Koichiro Haruki, MD, PhD (Data curation: Supporting; Investigation: Supporting; Writing – original draft: Supporting).

Naohiko Akimoto, MD, PhD (Data curation: Supporting; Investigation: Supporting; Writing – original draft: Supporting).

Mai Chan Lau, PhD (Data curation: Supporting; Writing – original draft: Supporting).

Kazuo Okadome, MD, PhD (Validation: Supporting).

Raaj S. Mehta, MD (Data curation: Supporting; Methodology: Supporting).

Juha P. Väyrynen, MD, PhD (Data curation: Supporting; Investigation: Supporting; Methodology: Supporting; Writing – original draft: Supporting).

Junko Kishikawa, MD, PhD (Data curation: Supporting; Investigation: Supporting; Writing – original draft: Supporting).
Tyler S. Twombly, BS (Data curation: Supporting; Writing – review & editing: Supporting).
Shanshan Shi, PhD (Data curation: Supporting; Investigation: Supporting).
Kenji Fujiyoshi, MD, PhD (Data curation: Supporting).
Keisuke Kosumi, MD, PhD (Data curation: Supporting; Methodology: Supporting; Writing – original draft: Supporting).
Yoko Ogata, MS (Validation: Supporting).
Hideo Baba, MD, PhD (Validation: Supporting; Writing – review & editing: Supporting).
Fenglei Wang, PhD (Writing – original draft: Supporting).
Kana Wu, MD, PhD (Data curation: Supporting; Writing – review & editing: Supporting).
Mingyang Song, MD, ScD (Data curation: Supporting; Writing – review & editing: Supporting).
Xuehong Zhang, MD, ScD (Data curation: Supporting; Funding acquisition: Supporting; Writing – review & editing: Supporting).
Charles S. Fuchs, MD, MPH (Funding acquisition: Supporting; Supervision: Supporting; Writing – review & editing: Supporting).
Cynthia L. Sears, MD (Funding acquisition: Supporting; Investigation: Supporting; Methodology: Supporting; Supervision: Supporting; Writing – review & editing: Supporting).
Walter C. Willett, MD, DrPH (Funding acquisition: Supporting; Supervision: Supporting; Writing – review & editing: Supporting).
Edward L. Giovannucci, MD, ScD (Supervision: Supporting; Writing – review & editing: Supporting).
Jeffrey A. Meyerhardt, MD, MPH (Supervision: Supporting; Writing – review & editing: Supporting).
Wendy S. Garrett, MD, PhD (Funding acquisition: Supporting; Writing – review & editing: Supporting).
Curtis Huttenhower, PhD (Funding acquisition: Supporting; Writing – review & editing: Supporting).
Andrew T. Chan, MD, MPH (Funding acquisition: Supporting; Supervision: Supporting; Writing – review & editing: Supporting).
Jonathan A. Nowak, MD, PhD (Investigation: Supporting; Methodology: Supporting; Supervision: Supporting; Writing – review & editing: Supporting).
Marios Giannakis, MD, PhD (Funding acquisition: Supporting; Methodology: Supporting; Supervision: Supporting; Writing – review & editing: Supporting).
Shuji Ogino, MD, PhD, MS (Conceptualization: Lead; Data curation: Lead; Funding acquisition: Lead; Investigation: Supporting; Methodology: Supporting; Project administration: Lead; Resources: Lead; Supervision: Lead; Writing – original draft: Supporting; Writing – review & editing: Lead).

**Conflicts of interest**

These authors disclose the following: Andrew T. Chan previously served as a consultant for Bayer Healthcare and Pfizer Inc. Jeffrey A. Meyerhardt has served as an advisor/consultant to Ignyta, Array Pharmaceutical, and Cota Healthcare. Charles S. Fuchs is currently employed by Genentech, a subsidiary of Roche, previously served as a consultant for Agios, Bain Capital, Bayer, Celgene, Dicerna, Five Prime Therapeutics, Gilead Sciences, Eli Lilly, Entrinsic Health, Genentech, KEW, Merck, Merrimack Pharmaceuticals, Pfizer Inc, Sanofi, Taiho, and Unum Therapeutics, and also serves as a director for CytomX Therapeutics and owns unexercised stock options for CytomX and Entrinsic Health. Marios Giannakis receives research funding from Bristol-Myers Squibb, Merck, Servier, and Janssen. This study was not funded by any of these commercial entities. The remaining authors disclose no conflicts.

**Funding**

This work was supported by National Institutes of Health (NIH) grants (P01 CA87969 and UM1 CA186107; P01 CA55075 and UM1 CA167552 to Walter C. Willett; U01 CA167552 to Walter C. Willett and L.A. Mucci; P50 CA127003, R01 CA118553, and R01 CA169141 to Charles S. Fuchs; R01 CA137178 and K24 DK098311 to Andrew T. Chan; R35 CA197735 and R01 CA151993 to Shuji Ogino; and K07 CA188126 to Xuehong Zhang); by Cancer Research UK Grand Challenge Award (UK C10674/A27140 to Cynthia L. Sears, Wendy S. Garrett, Curtis Huttenhower, Marios Giannakis, and Shuji Ogino); by the Stand Up to Cancer (SU2C) Colorectal Cancer Dream Team Translational Research Grant (SU2C-AACR-DT22-17 to Charles S. Fuchs and Marios Giannakis), administered by the American Association for Cancer Research, a scientific partner of SU2C; and by grants from the Project P Fund, The Friends of the Dana-Farber Cancer Institute, Bennett Family Fund, and the Entertainment Industry Foundation through the National Colorectal Cancer Research Alliance. Kota Arima and Tomotaka Ugai were supported by a grant from Overseas Research Fellowship (201860083 to Kota Arima, 201960541 to Tomotaka Ugai) from Japan Society for the Promotion of Science. Rong Zhong was supported by a fellowship grant from Huazhong University of Science and Technology. Tomotaka Ugai, Koichiro Haruki, and Kenji Fujiyoshi were supported by fellowship grants from the Uehara Memorial Foundation. Tomotaka Ugai was supported by a grant from Yasuda Medical Foundation. Koichiro Haruki was supported by the Mitsukoshi Health and Welfare Foundation. Andrew T. Chan is a Stuart and Suzanne Steele MGH Research Scholar. Jeffrey A. Meyerhardt's research is supported by the Douglas Gray Woodruff Chair fund, the Guo Shu Shi Fund, Anonymous Family Fund for Innovations in Colorectal Cancer, Project P fund, and the George Stone Family Foundation. Marios Giannakis was supported by a Conquer Cancer Foundation of American Society of Clinical Oncology Career Development Award. The content is solely the responsibility of the authors and does not necessarily represent the official views of National Institutes of Health. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Use of standardized official symbols**

We use HUGO (Human Genome Organisation) Gene Nomenclature Committee-approved official symbols (or root symbols) for genes and gene products, including BRAF, CACNA1G, CD14, CDKN2A, CRABP1, IGF2, KRAS, MLH1, NEUROG1, PIK3CA, RUNX3, SLCO2A1, SOCS1, TP53, and WNT, all of which are described at www.genenames.org. Gene symbols are italicized, whereas symbols for gene products are not italicized.