

Discovery and Features of an Alkylating Signature in Colorectal Cancer

Carino Gurjao^{1,2}, Rong Zhong^{3,4}, Koichiro Haruki³, Yvonne Y. Li^{1,2}, Liam F. Spurr^{1,2,5}, Henry Lee-Six⁶, Brendan Reardon^{1,2}, Tomotaka Ugai^{3,7}, Xuehong Zhang^{8,9}, Andrew D. Cherniack^{1,2}, Mingyang Song^{7,8,9,10,11}, Eliezer M. Van Allen^{1,2}, Jeffrey A. Meyerhardt¹, Jonathan A. Nowak¹², Edward L. Giovannucci^{7,8,9}, Charles S. Fuchs¹³, Kana Wu⁹, Shuji Ogino^{2,3,7,12}, and Marios Giannakis^{1,2}

Downloaded from AACRjournals.org/cancerdiscovery/article-pdf/11/10/2446/3082370/2446.pdf by guest on 29 July 2023

ABSTRACT

Several risk factors have been established for colorectal cancer, yet their direct mutagenic effects in patients' tumors remain to be elucidated. Here, we leveraged whole-exome sequencing data from 900 colorectal cancer cases that had occurred in three U.S.-wide prospective studies with extensive dietary and lifestyle information. We found an alkylating signature that was previously undescribed in colorectal cancer and then showed the existence of a similar mutational process in normal colonic crypts. This alkylating signature is associated with high intakes of processed and unprocessed red meat prior to diagnosis. In addition, this signature was more abundant in the distal colorectum, predicted to target cancer driver mutations KRAS p.G12D, KRAS p.G13D, and PIK3CA p.E545K, and associated with poor survival. Together, these results link for the first time a colorectal mutational signature to a component of diet and further implicate the role of red meat in colorectal cancer initiation and progression.

SIGNIFICANCE: Colorectal cancer has several lifestyle risk factors, but the underlying mutations for most have not been observed directly in tumors. Analysis of 900 colorectal cancers with whole-exome sequencing and epidemiologic annotations revealed an alkylating mutational signature that was associated with red meat consumption and distal tumor location, as well as predicted to target KRAS p.G12D/p.G13D.

¹Department of Medical Oncology, Dana-Farber Cancer Institute and Harvard Medical School, Boston, Massachusetts. ²Broad Institute of MIT and Harvard, Cambridge, Massachusetts. ³Program in MPE Molecular Pathological Epidemiology, Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts. ⁴Department 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. ⁵Pritzker School of Medicine, Biological Sciences Division, University of Chicago, Chicago, Illinois. ⁶Wellcome Sanger Institute, Hinxton, United Kingdom. ⁷Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts. ⁸Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts. ⁹Department of Nutrition, Harvard T.H. Chan School of Public Health, Boston, Massachusetts. ¹⁰Clinical and Translational Epidemiology Unit, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts. ¹¹Division of Gastroenterology, Massachusetts General Hospital, Boston, Massachusetts.

¹²Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts. ¹³Yale Cancer Center, Yale School of Medicine, Smilow Cancer Hospital, New Haven, Connecticut.

Note: Supplementary data for this article are available at Cancer Discovery Online (<http://cancerdiscovery.aacrjournals.org/>).

C. Gurjao, R. Zhong, and K. Haruki are the co-first authors of this article. K. Wu, S. Ogino, and M. Giannakis are the co-senior authors of this article.

Corresponding Author: Marios Giannakis, Dana-Farber Cancer Institute, 450 Brookline Avenue, Boston, MA 02215-5450. Phone: 617-582-7263; E-mail: Marios_Giannakis@dfci.harvard.edu

Cancer Discov 2021;11:2446–55

doi: 10.1158/2159-8290.CD-20-1656

This open access article is distributed under the Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 International (CC BY-NC-ND).

©2021 The Authors; Published by the American Association for Cancer Research

INTRODUCTION

Most tumor mutations are passengers that have little to no functional role in cancer. However, their positional context in the genome may reveal information about the underlying mutational processes (1). Snapshots of these processes, called mutational signatures, were originally deconvoluted using a nonnegative matrix factorization (NMF) approach (2) on a large collection of whole-genome sequencing and whole-exome sequencing (WES) data (3). Mutational signatures may elucidate the roles of mutagens in cancer and inform prevention and treatment efforts. Several studies have been conducted to associate mutational signatures with cellular processes or exposures. These include rare cancer predisposition syndromes (4), environmental agents (5), and microbiota (6). Such association studies have relied on either DNA-sequencing data sets or preclinical models, such as organoids. However, although many lifestyle-related factors have been linked to colorectal cancer (7), larger and more comprehensive data sets are needed to enable the discovery of the associated signatures. Consequently, past efforts have not been able to capture the cumulative effect of putative mutagens, such as dietary components, over decades. In particular, red meat consumption has been consistently linked to the incidence of colorectal cancer (8–10). The suggested mechanism is mutagenesis through alkylating damage induced by N-nitroso-compounds (NOC), which are metabolic products of blood heme iron or meat nitrates/nitrites (11). Nevertheless, this mutational damage is yet to be observed directly in patients' tumors.

RESULTS

Active Mutational Signatures in Colorectal Tumors and Normal Colonic Crypts

To address this gap, we leveraged a database of incident colorectal cancer cases that had occurred in three U.S.-wide prospective cohort studies, namely the Nurses' Health Studies (NHS) I and II and the Health Professionals Follow-up Study (HPFS; ref. 12). Study participants (more than 230,000 women and 50,000 men) repeatedly provided data on diet, lifestyle, and other factors without knowing their future colorectal cancer diagnosis, if any. We performed WES on matched primary untreated tumor-normal pairs in 900 patients with colorectal cancer with adequate tissue materials (Fig. 1A; Supplementary Table S1).

NMF signal separation revealed the existence of seven mutational processes (see Methods and Fig. 1B and C; Supplementary Fig. S1). We confirmed the robustness of the deconvolution by using another signature assignment program (SigProfiler; ref. 3); we again found seven mutational processes (Supplementary Fig. S2, left) that are highly similar to the ones obtained using the standard NMF approach (Supplementary Fig. S2, right).

To uncover the etiology of these colorectal signatures (that we name *c*-signatures), we first used a cosine similarity metric (cossim) to compare the deconvoluted signatures to reference COSMIC Single Base Substitution (SBS) signatures (3). The seven *de novo* signatures displayed the highest similarity with four known mutational processes (Supplementary Fig. S3), namely *POLE* deficiency (*c*-POLEa/SBS10a, cossim = 0.95 and *c*-POLEb/SBS10b, cossim = 0.86), aging (*c*-Age/SBS1,

cossim = 0.95), deficient mismatch repair (dMMR; *c*-dMMRa/SBS15, cossim = 0.90 and *c*-dMMRb/SBS26, cossim = 0.90), and exposure to alkylating agents (*c*-Alkylation/SBS11, cossim = 0.94). *c*-SBS40 matched the closest to SBS40 (cossim = 0.84), which is a featureless signature with unknown etiology and found in most cancers (3).

We substantiated the etiology of the four mutational processes by integrating clinical, pathology, and methylation data (Fig. 2A). Tumors harboring a *POLE* exonuclease domain mutation were significantly enriched in signatures *c*-POLEa and *c*-POLEb ($P = 2.3 \times 10^{-5}$ and $P = 1.8 \times 10^{-6}$, respectively, Mann-Whitney *U* test). Similarly, patients with orthogonally assessed microsatellite instability (MSI)-high status were significantly enriched in signatures *c*-dMMRa and *c*-dMMRb ($P < 2 \times 10^{-16}$ for both, Mann-Whitney *U* test). Signature *c*-Age also displayed a significant association with patients' age at diagnosis ($P = 1.7 \times 10^{-5}$, Mann-Whitney *U* test). Last, we support the etiology of the alkylating-like signature, not previously described in colorectal cancer, by assessing the *MGMT* (O-6-methylguanine-DNA methyltransferase) promoter methylation status in tumors from the NHS/HPFS cohorts. *MGMT* is a central gene in the repair of alkylating lesions. Among the sequenced specimens with available *MGMT* promoter methylation data, we observed that tumors with methylated *MGMT* promoters were enriched in the signature *c*-Alkylation ($P = 6.6 \times 10^{-3}$, Mann-Whitney *U* test; Fig. 2A), further supporting that this signature represents the biological consequence of increased alkylating damage. Of note, SBS18, which is associated with *MUTYH*-associated polyposis (3), is absent in the tumor samples we sequenced. We believe this is the case because of the low occurrence of *MUTYH* deficiency generally in colorectal cancer (less than 1%; ref. 13), as well as further undersampling of patients with germline predisposition mutations as only healthy individuals were enrolled prospectively in NHS/HPFS.

NMF signal separation in The Cancer Genome Atlas (TCGA) colorectal tumors ($n = 540$) revealed the existence of seven signatures (Supplementary Figs. S4 and S5 for SigProfiler results) similar to the ones found in NHS/HPFS (Supplementary Fig. S6), thus suggesting the existence of the same underlying mutational processes in all colorectal cancer cohorts. Analysis of the TCGA colorectal tumors (Fig. 2B) substantiated the same etiologies for the *POLE* signatures *c*-POLEa and *c*-POLEb ($P = 6.2 \times 10^{-7}$ and $P = 7.3 \times 10^{-7}$, respectively, Mann-Whitney *U* test), as well as dMMR signatures *c*-dMMRa and *c*-dMMRb ($P < 2 \times 10^{-16}$ for both, Mann-Whitney *U* test). We also observed that TCGA tumors with *MGMT* promoter methylation were enriched in signature *c*-Alkylation ($P = 9.7 \times 10^{-5}$, Mann-Whitney *U* test). Of note, in TCGA, signature *c*-Alkylation displayed the highest similarity with SBS30 (cossim = 0.81), followed by SBS11. Conversely, SBS30 was the second most similar signature to the *c*-Alkylation one in the NHS/HPFS cohorts (Fig. 2C; Supplementary Fig. S3). SBS30 resembles SBS11 (cossim of 0.76, Fig. 2C) and is attributed to base excision repair (BER) deficiency (3), which is also a central pathway in repairing damage from alkylated bases. We nevertheless found no association between germline polymorphisms in *NTHL1* and other genes of the BER pathway and the alkylating signature in the TCGA specimens (see Methods and Supplementary

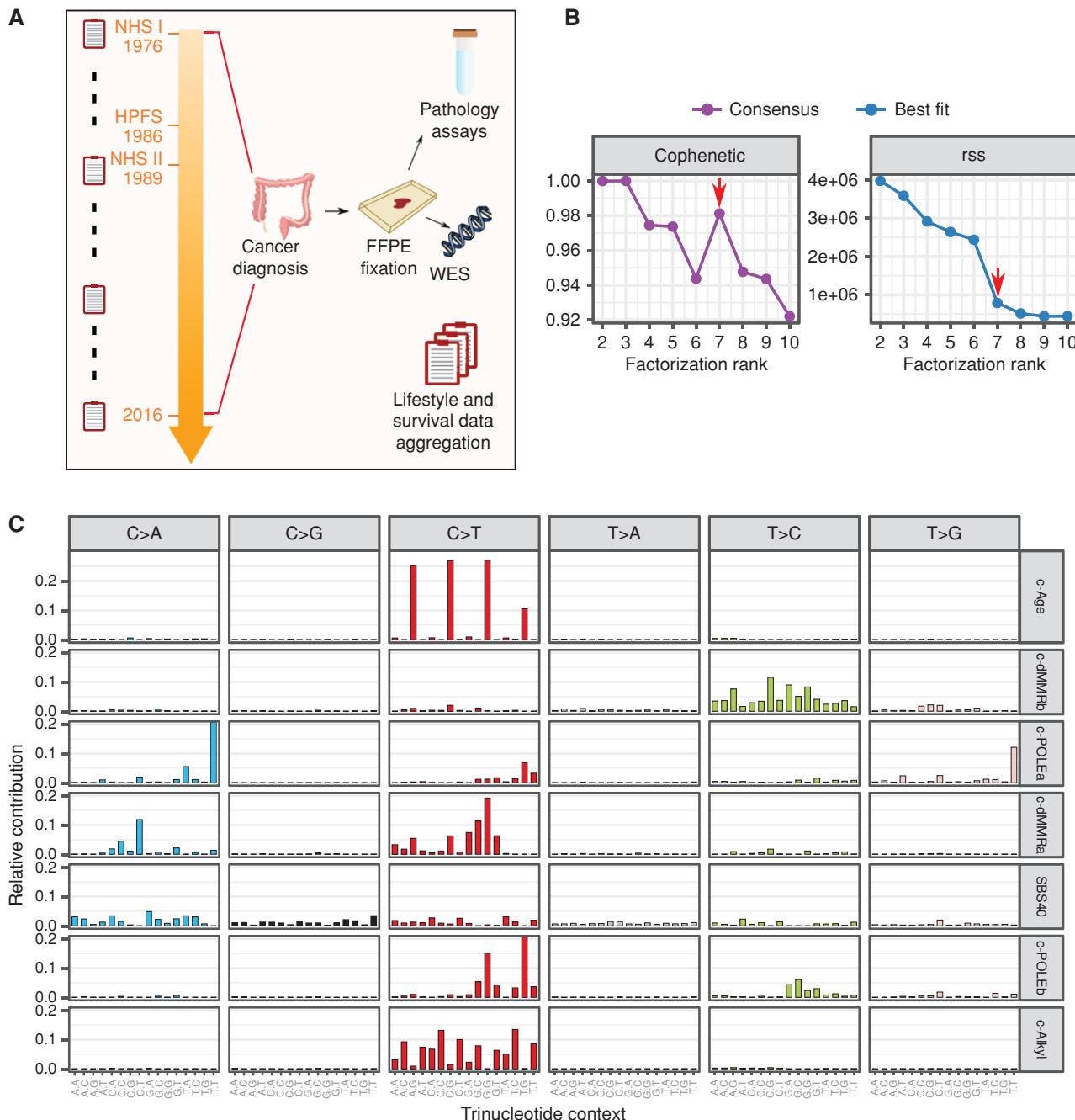


Figure 1. *De novo* signature deconvolution in NHS/HPFS colorectal cancers. **A**, Cohort and data overview. FFPE, formalin-fixed, paraffin-embedded. **B**, Quality measures for NMF in NHS/HPFS. Arrows indicate the estimated rank of mutational signatures. rss, residual sum of squares. **C**, The consensus seven signatures found by NMF in NHS/HPFS.

Fig. S7). The presence of SBS30 ahead of SBS11 in the TCGA colorectal cancer data set could instead be attributed to a smaller sample size of colorectal cancers in TCGA compared with NHS/HPFS (see “Undersampling Simulations” in Methods and Supplementary Fig. S8). The Fanconi anemia (FA) and translesion synthesis (TLS) DNA damage repair pathways also do not show an association with the alkylating signature (see Methods and Supplementary Fig. S9A and S9B).

We also estimated the effect size for the Mann–Whitney *U* tests by calculating the rank-biserial correlation r_{rb} for each mutational signature and the respective molecular or clinical phenotype shown in Fig. 2B. We observed that the effect sizes were similar for the alkylating signatures and the aging signature ($r_{rb} = 0.14$ and $r_{rb} = 0.16$, respectively) and smaller than the hypermutator dMMR and POLE signatures ($r_{rb} > 0.8$ for dMMR and POLE signatures in both TCGA and NHS/HPFS).

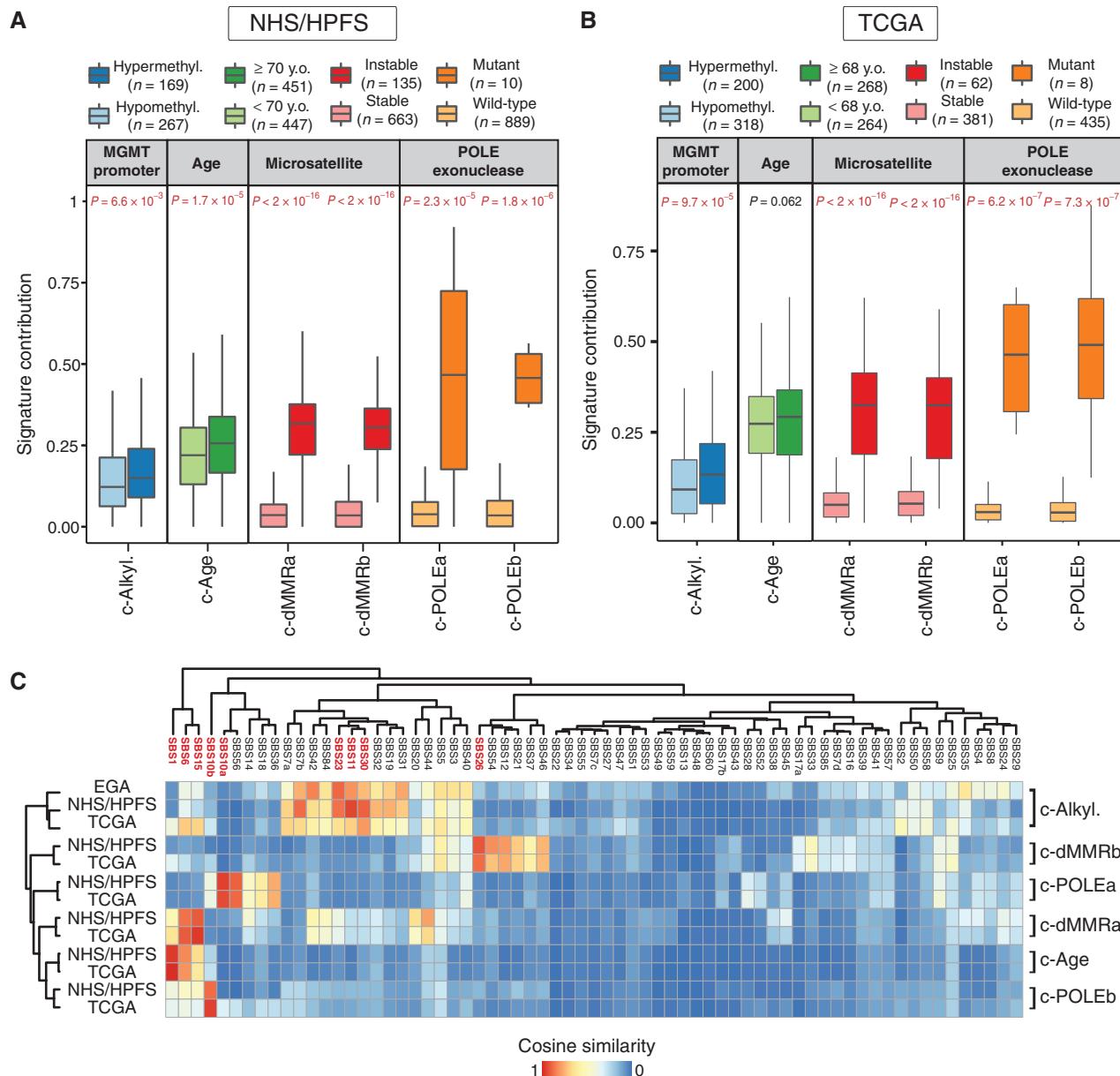


Figure 2. Active mutational signatures in colonic cells. Proportion of mutations assigned to *de novo* extracted signatures in colorectal cancers from NHS/HPFS (A) and TCGA (B), segregated by MGMT promoter methylation status, POLE exonuclease mutations, microsatellite instability, and age at diagnosis. Box plot outliers not shown. C, Heat map of the similarity scores between colorectal tumor (from TCGA and NHS/HPFS) signatures, clustered on the y-axis, and reference COSMIC signatures, clustered on the x-axis. COSMIC signatures found in either NHS/HPFS or TCGA are bolded. The alkylating normal colon (from EGA) signature is also shown. Clustering has been performed according to cosine similarity.

Downloaded from http://aacjournals.org/cancerdiscovery/article-pdf/11/10/2446/3082370/2446.pdf by guest on 29 July 2023

Interestingly, a previously published survey of mutational signatures in normal colorectal crypts (14) from the European Genome-phenome Archive (EGA) showed the existence of a signature (named SBSC) that we found to be similar to the alkylating one that we observed in NHS/HPFS colorectal cancers (*cossim* = 0.85). Of note, SBSC matched closely to SBS23, which, similar to SBS30, also resembles SBS11 (*cossim* of 0.77; Fig. 2C). The hierarchical clustering of SBSC with the seven signatures deconvoluted from NHS/HPFS and TCGA confirmed the similarity of EGA SBSC with the alkylating imprints (Fig. 2C).

Dietary Patterns of Alkylation Damage

To test whether dietary components contributed to the alkylating signature in colorectal cancer, we leveraged prospectively collected repeated measurements of meat, poultry, and fish consumption in grams per day in the NHS and HPFS cohorts. All available red meat variables showed significant positive associations between prediagnosis intakes and alkylating damage in colorectal cancers (Fig. 3A; overall red meat, $P = 0.017/r_{rb} = 0.14$; unprocessed red meat, $P = 7.8 \times 10^{-3}/r_{rb} = 0.16$; and processed red meat, $P = 7.3 \times 10^{-3}/r_{rb} = 0.16$, Mann-Whitney

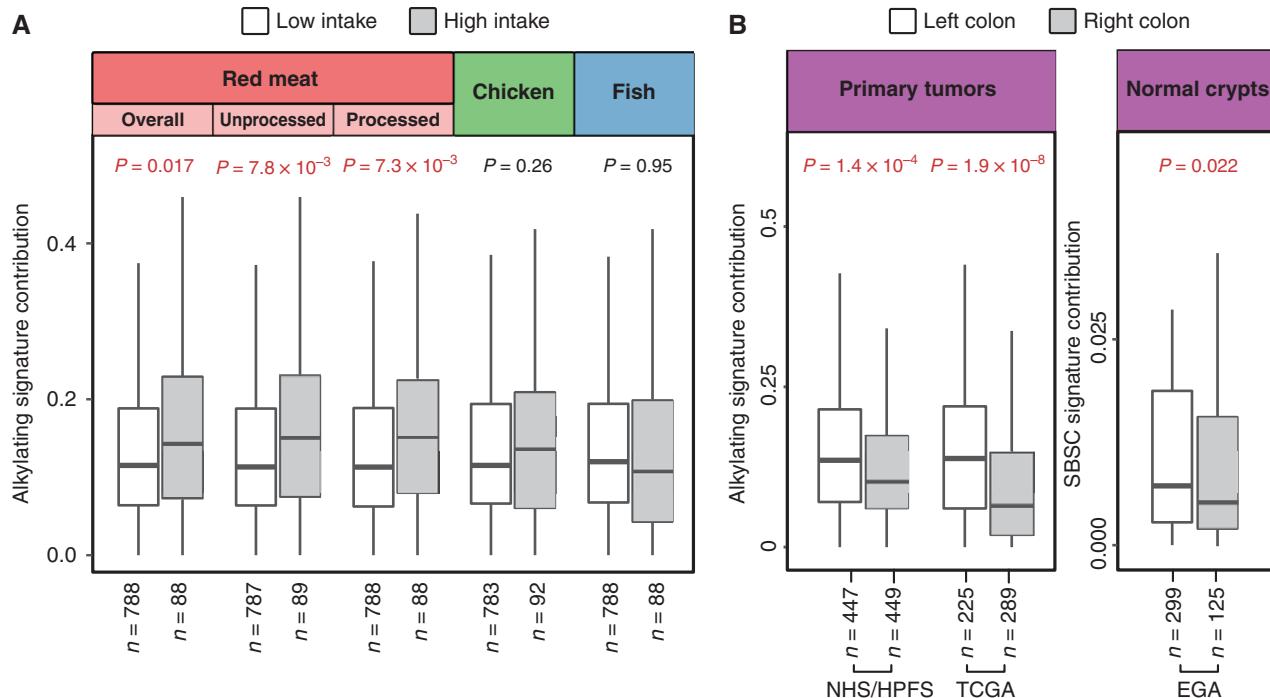


Figure 3. Epidemiology and distribution of alkylating damage. **A**, Proportion of mutations assigned to alkylating damage in NHS/HPFS, segregated by intake (top decile in grams per day vs. the rest) of overall, processed, and unprocessed red meat, as well as chicken and fish. **B**, Proportion of mutations assigned to alkylating damage in colorectal cancer and normal colon, segregated by tumor location. Box plot outliers not shown.

U test). Other dietary variables (fish and chicken intake, Fig. 3B) and lifestyle factors (body mass index, alcohol consumption, smoking, and physical activity in Supplementary Fig. S10) did not show any significant association with the alkylating signature. In addition, no other colorectal cancer mutational process showed a significant association with red meat intake (Supplementary Fig. S11). Of note, *MGMT* promoter methylation did not differ by red meat consumption (two-sided Mann–Whitney *U* test, *P* = 0.51; Supplementary Fig. S12). When adjusted for red meat intake, there was no difference in alkylating damage between male and female patients with colorectal cancer (two-sided Mann–Whitney *U* test, *P* = 0.27 for patients with high overall red meat consumption).

Previous studies (9, 10) showed a positive association between processed red meat and colorectal cancer incidence in the distal colon. Thus, we also investigated how the alkylating damage might differ by tumor location. We found that, compared with the proximal colon, the distal colorectal specimens exhibited higher alkylating damage in tumors (*P* = 1.4 × 10⁻⁴ in NHS/HPFS and *P* = 1.9 × 10⁻⁸ in TCGA, Mann–Whitney *U* test) and normal crypts (*P* = 0.022, Mann–Whitney *U* test; Fig. 3B).

Carcinogenicity of Alkylation Damage

Mutational processes increase the likelihood of specific driver mutations in certain trinucleotide contexts. To find such driver mutations that associate with the alkylating signature, we devised a simple model (Fig. 4A; see Methods) that predicts the relative likelihood of mutational processes to target colorectal cancer recurrent drivers in non-MSI-high, non-*POLE*-mutated tumors.

In particular, the alkylating signature appeared to be the dominant one that targets *KRAS* p.G13D (relative likelihood = 1) and *KRAS* p.G12D (relative likelihood = 0.91; Fig. 4A). This is due to p.G12D and p.G13D being in trinucleotide contexts (ACC>ATC and GCC>GTC, respectively) mainly targeted by the alkylating signature. *PIK3CA* p.E545K (TCA>TTA) is also predicted to be predominantly targeted by the alkylating signature (relative likelihood = 0.87). Supporting this, we showed that colorectal cancers having *KRAS* p.G12D, *KRAS* p.G13D, or *PIK3CA* p.E545K-mutant colorectal cancers were enriched with the alkylating signature compared with all other tumors (Fig. 4B, *P* = 0.013, Mann–Whitney *U* test).

Last, we examined patient survival across ordinal alkylating mutational signature quartiles and found that patients whose tumors have high alkylation damage (top quartile) had a worse colorectal cancer-specific survival (log-rank test *P*_{trend} = 0.036; Fig. 4C; Supplementary Tables S2 and S3). Furthermore, higher alkylating signature contribution was associated with worse colorectal cancer-specific survival in both univariable and multivariable Cox proportional hazards regression analyses (*P*_{trend} = 0.015 and *P*_{trend} = 0.036, respectively, Fig. 4D and Supplementary Table S3).

DISCUSSION

Our work demonstrated the presence of a novel alkylating mutational signature, which we deconvoluted directly from WES of colorectal tumors. Interestingly, this signature is highly similar to SBS11, which was originally discovered in patients with prior exposure to temozolomide (1).

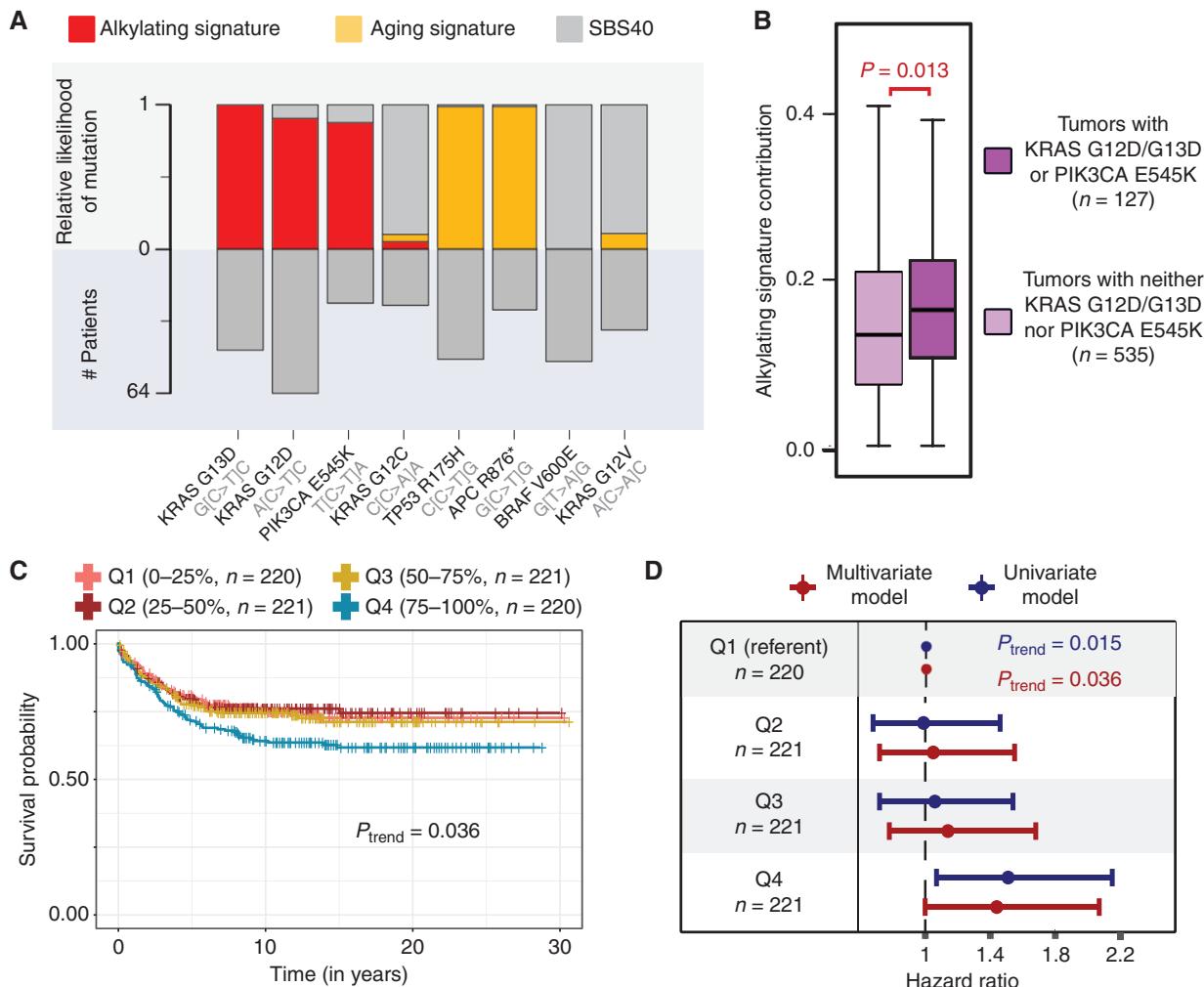


Figure 4. Carcinogenic potency of alkylating damage. **A**, Relative likelihood of mutational processes to target recurrent hotspots in nonhypermutated colorectal cancer. As hotspots, we considered all point mutations that were present in at least 25 patients with nonhypermutated (non-MSI-high, non-POLE-mutated) colorectal cancer. Each stacked bar represents the relative likelihood of a given signature to target a given hotspot. **B**, Proportion of mutations assigned to alkylating damage in NHS/HPFS, TCGA colorectal cancers, segregated by KRAS G12D/KRAS G13D/PIK3CA E545K mutation status. Box plot outliers not shown. **C**, Kaplan-Meier plot illustrating colorectal cancer-specific survival of the patients stratified into quartiles of alkylating signature contribution. **D**, Forest plot of the association between the colorectal cancer-specific survival and quartiles of alkylating signature contribution in univariable and multivariable Cox regression models.

Temozolamide is an alkylating agent used as a treatment of brain gliomas with *MGMT* promoter methylation (1) and induces the same lesions as dietary NOCs and in the same proportions (80% of N7-methylguanine and N3-methylguanine, as well as 10% of O6-methylguanine; refs. 15, 16).

Previous attempts have shown the existence of alkylating lesions in normal colorectal mucosa, notably caused by NOCs (17). The latter can be formed endogenously after nitrosylation of heme iron from blood (17, 18) but have also been associated with red meat intake in a small cohort of participants (19). However, these previous studies were based on limited data sets (small sample sizes and/or use of laboratory methylating agents) and lack comprehensive sequencing that would enable the discovery of the full mutational spectrum induced by red meat. Crucially, past efforts have focused on normal colorectal tissues and not examined colorectal cancer.

Our analysis reveals the existence of an alkylating signature in colorectal cancer, which is associated with high prediagnosis intake of processed and unprocessed red meat.

Earlier work also hypothesized that the distal colon has increased DNA damage from exposure to dietary carcinogens, as a result of feces storage and water resorption in this portion of the large intestine (20). This is believed to explain the association observed between distal cancer incidence and red meat consumption (9, 10, 20). Consistently, we found an enrichment in tumors and normal crypts in the distal colon and rectum.

In support of the International Agency for Research on Cancer (IARC) Monograph Working Group, which classified processed meat as carcinogenic (8), our results provide molecular evidence of this dietary factor's mutagenic impact. In addition, our analyses further implicate unprocessed meat intake and suggest *MGMT* as a factor of susceptibility to red

meat-induced damage. The existence of a similar alkylating signature in normal colorectal crypts also suggests that mutational changes due to such damage may start to occur early in the path of colorectal carcinogenesis.

Our analysis predicted *KRAS* p.G12D, *KRAS* p.G13D, and *PIK3CA* p.E545K to be mainly targeted by the alkylating signature in nonhypermutated colorectal cancers. We showed that there was indeed higher alkylating damage in tumors harboring these driver mutations. Independent epidemiologic analyses have also shown a positive association between high consumption of red meat products and *KRAS* p.G12D and *KRAS* p.G13D (21, 22). Although the number of mutations due to alkylation damage was lower than other mutational processes, we showed that alkylation might have considerable carcinogenic potential by targeting driver mutations in *KRAS* and *PIK3CA*. We also demonstrated a significantly worse survival for patients with high levels of the alkylation signature contribution.

Our study has leveraged a comprehensive data set with repeated dietary measures over years, without patients knowing their upcoming colorectal cancer diagnosis, and WES on a large collection of colorectal tumors. It provides unique evidence supporting the direct impact of dietary behaviors on colorectal carcinogenesis. Moreover, the presence of a similar alkylating signature in normal mucosa advocates for the utility of early dietary interventions and suggests potential precision prevention approaches in *MGMT*-methylated premalignant tissue. Similarly, the association of the signature with cancer driver mutations, such as *KRAS* and *PIK3CA* ones, may offer future potential therapeutic opportunities. More generally, our study exemplifies the potential role of large-scale molecular epidemiologic studies in elucidating cancer pathogenesis (23) and guiding prevention efforts through lifestyle modifications, such as dietary interventions.

METHODS

Study Population, Specimens, and Sequencing

We used data from three prospective cohort studies in the United States: the Nurses' Health Study I (NHS1, including 121,701 women ages 30 to 55 years at enrollment who had been followed since 1976), the Nurses' Health Study II (NHS2, including 116,429 women ages 25 to 42 years who had been followed since 1989), and the HPFS (including 51,529 men ages 40 to 75 years followed since 1986; ref. 12). The study participants had been sent questionnaires biennially to update information on lifestyle factors and newly diagnosed diseases, including colorectal cancer. The follow-up rate had been more than 90% for each follow-up questionnaire cycle in the three cohort studies. The patients were followed until death or end of follow-up (January 1, 2016, for HPFS; June 1, 2016, for NHS1; and June 1, 2015, for NHS2), whichever came first. Study physicians, who were blinded to exposure data, reviewed medical records of 4,855 incident colorectal cancer cases to confirm the disease diagnosis and to collect data on tumor size, tumor anatomic location, and disease stage. Archival formalin-fixed, paraffin-embedded (FFPE) tissue blocks of tumor and normal colon were collected in a subset of colorectal cancer. We previously showed that in our cohorts, demographic features of cases did not differ appreciably by tissue availability (24). The study protocol was approved by the institutional review boards of the Brigham and Women's Hospital and Harvard T.H. Chan School of Public Health (Boston, MA) and those of participating registries as required. Written informed consent was obtained from all patients with colorectal cancer.

We prioritized relatively more recent colorectal cancer cases for sequencing to mitigate the potential impact of FFPE artifacts. Given

the number of NHS versus HPFS participants (2:1 female/male ratio), we also sequenced relatively more specimens from male patients to obtain more balanced sequencing data. Supplementary Table S4 shows the clinical and pathologic characteristics of the 4,855 patients with colorectal cancer.

WES was carried as previously described (25). Briefly, using guide hematoxylin and eosin-stained slides, tumor areas were selected to extract tumor-enriched DNA from tissue sections of tumor FFPE blocks. Normal DNA was extracted from resection margins or other areas free from tumors. DNA specimens underwent hybrid capture with SureSelect v.2 Exome bait (Agilent Technologies), followed by sequencing on Illumina HiSeq 2000 instruments. The obtained average coverage was 85 \times in tumors and matched adjacent normal colon tissue (see Supplementary Table S5).

Dietary Variables

Ascertainment of diet was carried out as previously described (9). To assess dietary intake in each cohort, food frequency questionnaires (FFQ) were initially collected in 1980 for NHS and in 1986 for HPFS. For the NHS, a 61-item semiquantitative FFQ was used at baseline (26), which was expanded to approximately 130 food and beverage items in 1984, 1986, and every 4 years thereafter. For the HPFS cohorts, baseline dietary intake was assessed using a 131-item FFQ that was also used for updates generally every 4 years subsequently (27). In particular, unprocessed red meat consumption was evaluated based on forms on the intake of "beef or lamb as main dish," "pork as main dish," "hamburger," and "beef, pork, or lamb as a sandwich or mixed dish." Processed meat diets included "bacon"; "beef or pork hot dogs"; "salami, bologna, or other processed meat sandwiches"; and "other processed red meats such as sausage, kielbasa, etc." Consumption of red meat, chicken, poultry, and fish was evaluated in grams per day. For the remainder of our analysis, we considered the top decile of each variable to determine the "high-intake" patients and considered the rest as "low-intake" patients, because only the top-decile patients show a substantial difference in overall red meat intake (Supplementary Fig. S13A and S13B). Data were based on the most recent prediagnosis reported intake for each patient.

MGMT Promoter Methylation, MSI, and POLE Deficiency Status

MGMT promoter methylation analysis in the NHS/HPFS cohorts was carried out using bisulfite conversion and real-time PCR as previously described (28). MSI status was evaluated using 10 microsatellite markers (D2S123, D5S346, D17S250, BAT25, BAT26, BAT40, D18S55, D18S56, D18S67, and D18S487) as formerly detailed (12).

POLE deficiency was assessed by sequencing and manual Integrated Genome Viewer curation of *POLE* exonuclease domain mutations in hypermutated non-MSI-high tumors (>400 mutations).

Somatic Variant Calling

We have used the Cancer Genome Analysis (CGA) WES characterization pipeline (https://github.com/broadinstitute/CGA_Production_Analysis_Pipeline) developed at the Broad Institute of MIT and Harvard to call, filter, and annotate somatic mutations. All analyses were carried out on the human genome build hg19. The pipeline employs the following tools: MuTect (29), ContEst (30), Strelka (31), DeTiN (32), AllelicCapSeg (33), MAFPoNFilter (34), RealignmentFilter, GATK (35), and PicardTools. FFPE-specific artifacts are filtered similarly to previous publications (25, 36). Briefly, FFPE artifacts arise from formaldehyde deamination of cytosines resulting in C-to-T transition mutations, which presents itself as an "Orientation bias" (excess of C>T sites in F1R2 read pairs and an excess of G>A in F2R1 read pairs). In the pipeline we used, the "Orientation Bias Filter" tool (37) filters out FFPE-specific artifacts. To further filter spurious single-nucleotide variant calls, we used Burrows-Wheeler Aligner BWA-MEM (<http://bio-bwa.sourceforge.net/>) to realign sequenced reads associated with the mutations to a set

of sequences derived from the human reference assembly. The Panel of Normal was created using normal samples with less than 1% of cross-sample contamination (as evaluated by Contest; ref. 30) and less than 1% of tumor in normal (as outputted by DeTIN; ref. 32). We illustrate the variant calling pipeline in Supplementary Fig. S14.

TCGA Data Analysis

Clinical, methylation, and somatic mutation data from TCGA were downloaded from the Data Coordination Center (DCC) data portal at <https://dcc.icgc.org/releases/current/Projects/COAD-US> and <https://dcc.icgc.org/releases/current/Projects/READ-US> (as of March 2020). For consistency, only WES data sets were used. Altogether, we pooled 540 TCGA patients with somatic mutation data, among whom 523 patients also had methylation data.

We evaluated *MGMT* promoter methylation status using the MGMT-STP27 prediction model (38). In short, two probes (cg12434587 and cg12981137) were used to predict *MGMT* promoter methylation. An M value cutoff of 0.358, which empirically maximized the sum of sensitivity and specificity, was then used to discriminate *MGMT* promoter methylation status (Supplementary Fig. S15).

Nonnegative Matrix Factorization

Mutations were deconvoluted into separate signatures based on the number of mutations in each of 96 possible trinucleotide contexts. Deconvolution was carried out with a standard NMF method based on Kullback–Leibler divergence using the “NMF” R package (39). This method is particularly adapted for mutational signature analysis as recent studies demonstrated (40).

A critical parameter in NMF is the estimation of the rank (i.e., the number of expected mutational signatures). To determine this, we performed quality measures on a range of ranks ($n = 2$ to 10) for the 900 colorectal cancer exomes in the NHS/HPFS cohorts. This showed a sharp increase in the cophenetic (i.e., the stability of the NMF classes) and dispersion (i.e., the reproducibility of the class assignments) metrics after rank = 7. For this rank, we also observed that the residual sum of squares (RSS) reached a lower plateau (Supplementary Fig. S1). A similar rank survey on an independent cohort of 540 colorectal cancer exomes from the TCGA (Supplementary Fig. S4) revealed the same dispersion and cophenetic peaks at rank = 7 and a lower plateau RSS. For the rest of the analysis, we consequently used rank = 7. We confirmed the robustness of these seven signatures by running NMF with different variant allele frequency (VAF) cutoffs (Supplementary Fig. S16). This demonstrates that the signature discovery is not affected by low VAF mutations, which are more likely to represent sequencing artifacts, such as those due to FFPE preservation.

SigProfiler was run on NHS/HPFS and TCGA colorectal cancer exomes as previously described (3).

Undersampling Simulations

To show that the difference in sample size between TCGA ($n = 540$) and NHS/HPFS ($n = 900$) can explain the presence of SBS30 instead of SBS11 in the former cohort, we (i) randomly sampled 540 patients of the 900 from NHS/HPFS; (ii) extracted seven signatures from the 540 patients and found their closest fit among SBS1 (aging signature), SBS10a and SBS10b (POLE signatures), SBS15 and SBS26 (dMMR signatures), and SBS11 and SBS30; and (iii) repeated steps (i) and (ii) a hundred times.

Crypt Mutational Signature Analysis

Mutational signatures from normal colonic crypts (14) were used in our analysis. These signatures were extracted from WGS data from 571 crypts from 42 individuals from the EGA (14). Deconvolution was performed using a hierarchical Dirichlet process, which produces results similar to NMF (14).

Analysis of Recurrent Hotspot Mutations

To compute the relative likelihood of mutational processes to target a specific hotspot, we (i) localized the trinucleotide context of the hotspot, (ii) extracted the signatures contribution for the specific trinucleotide context, and (iii) normalized the contribution of each signature, such that the sum became 1. Recurrent hotspots were defined as specific point mutations occurring in at least 25 patients.

TCGA Germline Polymorphisms Analysis

TCGA genotyping data (Affymetrix SNP 6.0 array platform) were used to select germline variants from genes in the BER, FA, and TLS pathways extracted from the GSEA database (refs. 41, 42; <http://www.gsea-msigdb.org/gsea/msigdb/>). We imputed autosomal variants for TCGA samples using IMPUTE2 (43), with haplotypes of 1000 Genomes Phase 3 (44) as the reference panel. We used the following criteria to select SNPs with the plink software (45): (i) average imputation confidence score, also called INFO score, ≥ 0.4 ; (ii) minor allele frequency $\geq 5\%$; (iii) SNP missing rate $< 5\%$ for best-guessed genotypes at posterior probability ≥ 0.9 ; and (iv) Hardy–Weinberg equilibrium P value $> 1 \times 10^{-6}$. After imputation, 2,041 variants were included in our subsequent analysis. We tested for an additive effect (genotype 0,1,2 as a continuous variable) for each SNP and found no association with the alkylating signature [Supplementary Fig. S7 and Supplementary Fig. S9, FDR-adjusted P value (q value) less than 0.1 for all SNPs tested].

Statistical Analysis

We used R version 3.6.2 to perform statistical analyses. Significance for two-group comparisons was evaluated by a one-sided Mann–Whitney U test unless otherwise indicated. $P < 0.05$ was considered statistically significant. For the comparisons of the alkylating signature by age in the NHS/HPFS cohorts and TCGA colorectal cancer database, the patients' median age (70 and 67 years, respectively) was used as the cutoff.

Eight hundred eighty-two patients with available colorectal cancer survival data were subsequently used for survival analyses. Univariable- and multivariable-adjusted Cox proportional hazards regression analysis as used to calculate the HR of colorectal cancer-specific survival and overall survival according to ordinal alkylating mutational signature quartiles (Q1–Q4). The multivariable Cox regression model initially included sex (female vs. male), age at diagnosis (<60, 60–64, 65–69, and ≥ 70 years), year of diagnosis (1995 or before, 1996–2000, 2001–2005, and 2006–2014), family history of colorectal cancer (present vs. absent), current smoking status (never smoking, past smoking, 1–14 pack-years, 15–24 pack-years, ≥ 25 pack-years), alcohol consumption (women: 0–<0.15, 0.15–<2.0, 2.0–<7.5, and ≥ 7.5 g/day; men: 0 to <1, 1–<6, 6–<15, and ≥ 15 g/day), tumor location (proximal colon vs. distal colon vs. rectum), CpG island methylator phenotype (high vs. low/negative; ref. 46), KRAS mutation (mutant vs. wild-type; ref. 47), BRAF mutation (mutant vs. wild-type; ref. 47), tumor differentiation (well to moderate vs. poor), disease stage (I/II vs. III/IV), microsatellite instability status (MSI-high vs. non-MSI-high; ref. 46), and long-interspersed nucleotide element 1 (LINE-1) methylation level (continuous; ref. 48). A backward elimination with a threshold P of 0.05 was used to select variables for the final models. Cases with missing data were assigned to the majority category of a given categorical covariate to limit the degrees of freedom, except for cases with missing LINE-1 methylation, for which we assigned a separate indicator variable. We confirmed that excluding the cases with missing information in any of the covariates did not substantially alter results.

Data Availability

WES data have been deposited in dbGAP (accession number phs000722). WES quality metrics and a subset of clinical annotations are included in this article. Additional clinical and epidemiology data from the NHS1, NHS2, and HPFS can be requested through the NHS/HPFS consortia.

Code Availability Statement

All analysis scripts are available upon request.

Authors' Disclosures

Y.Y. Li reports other support from g.Root Biomedical Services outside the submitted work. A.D. Cherniack reports other support from Bayer outside the submitted work. E.M. Van Allen reports grants from Novartis and BMS; personal fees from Tango Therapeutics, Genome Medical, Invitae, Monte Rosa Therapeutics, Manifold Bio, Illumina, Enara Bio, and personal fees from Janssen outside the submitted work; in addition, E.M. Van Allen has a patent for institutional patents filed on chromatin mutations and immunotherapy response, and methods for clinical interpretation pending. J.A. Meyerhardt reports personal fees from COTA Healthcare and personal fees from Taiho Pharmaceutical outside the submitted work. C.S. Fuchs reports personal fees from Amylin Pharma, AstraZeneca, Bain Capital, CytonX Therapeutics, Daiichi-Sankyo, Eli Lilly, Entrinsec Health, Evolveimmune Therapeutics, Genentech, Merck, Taiho, and personal fees from Unum Therapeutics outside the submitted work; in addition, C.S. Fuchs serves as a director for CytonX Therapeutics and owns unexercised stock options for CytonX and Entrinsec Health; is a cofounder of Evolveimmune Therapeutics and has equity in this private company; has provided expert testimony for Amylin Pharmaceuticals and Eli Lilly. C.S. Fuchs is now an employee of Genentech and Roche. S. Ogino reports grants from NIH during the conduct of the study. M. Giannakis reports grants from CRUK, SU2C, and grants from NIH/NCI during the conduct of the study; grants from Bristol-Myers Squibb, Merck, Servier, Janssen, and grants from ASCO Conquer Cancer Foundation outside the submitted work. No disclosures were reported by the other authors.

Authors' Contributions

C. Gurjao: Data curation, formal analysis, investigation, visualization, methodology, writing-original draft, writing-review and editing. **R. Zhong:** Data curation, formal analysis, writing-review and editing. **K. Haruki:** Data curation, formal analysis, writing-review and editing. **Y.Y. Li:** Writing-review and editing. **L.F. Spurr:** Writing-review and editing. **H. Lee-Six:** Resources, writing-review and editing. **B. Reardon:** Writing-review and editing. **T. Ugai:** Writing-review and editing. **X. Zhang:** Writing-review and editing. **A.D. Cherniack:** Writing-review and editing. **M. Song:** Writing-review and editing. **E.M. Van Allen:** Writing-review and editing. **J.A. Meyerhardt:** Resources, writing-review and editing. **J.A. Nowak:** Resources, writing-review and editing. **E.L. Giovannucci:** Resources, data curation, writing-review and editing. **C.S. Fuchs:** Resources, funding acquisition, writing-review and editing. **K. Wu:** Data curation, funding acquisition, writing-review and editing. **S. Ogino:** Resources, data curation, funding acquisition, writing-review and editing. **M. Giannakis:** Conceptualization, resources, supervision, funding acquisition, investigation, methodology, writing-original draft, writing-review and editing.

Acknowledgments

We thank N. Abdennur and S. Abraham for technical feedback, as well as W.L. Chiu, J. Elhai, and L. Fossecave for useful comments. This work was supported by the NIH grants P01 CA87969 (to R.M. Tamini), UM1 CA186107 (to M.J. Stampfer), P01 CA55075 (to W.C. Willett), UM1 CA167552 (to W.C. Willett), U01 CA167552 (to W.C. Willett and L.A. Mucci), U01 CA176726 (to W.C. Willett), U54 HG003067 (to E.S. Lander and S.B. Gabriel), P50 CA127003 (to B. Wolpin), P30CA016359 (to C.S. Fuchs), R01 CA118553 (to C.S. Fuchs), R01 CA169141 (to C.S. Fuchs), R35 CA197735 (to S. Ogino), R01 CA151993 (to S. Ogino), K07 CA190673 (to R. Nishihara), K07 CA188126 (to X. Zhang), R21 CA238651 (to X. Zhang), R03 CA197879 (to K. Wu), R21 CA222940 (to K. Wu and M. Giannakis),

and R21 CA230873 (to K. Wu and S. Ogino); by Cancer Research UK Grand Challenge Award (UK C10674/A27140, to M. Giannakis and S. Ogino); by Nodal Award (2016-02) from the Dana-Farber Harvard Cancer Center (to S. Ogino); by the Stand Up To Cancer Colorectal Cancer Dream Team Translational Research Grant (SU2C-AACR-DT22-17, to C.S. Fuchs and M. 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 National Colorectal Cancer Research Alliance. Stand Up To Cancer is a division of the Entertainment Industry Foundation. K. Haruki was supported by fellowship grants from the Uehara Memorial Foundation and the Mitsukoshi Health and Welfare Foundation. X. Zhang was supported by American Cancer Society Research Scholar Grant (RSG NEC-130476). X. Zhang was supported by the Dana-Farber Harvard Cancer Center (DF/HCC) GI SPORE Developmental Research Project Award (P50CA127003), DF/HCC Nodal Award (Cancer Center Support Grant, P30CA006516-55), the Karin Grunbaum Cancer Research Foundation, and the Zhu Family PEER Award. J.A. Meyerhardt 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. M. Giannakis was supported by a Conquer Cancer Foundation of ASCO Career Development Award. T. Ugai was supported by a grant from Overseas Research Fellowship (201960541) from Japan Society for the Promotion of Science. R. Zhong was supported by a fellowship grant from Huazhong University of Science and Technology, Wuhan, Hubei, China. We thank the participants and staff of the Nurses' Health Studies and the Health Professionals Follow-up Study for their valuable contributions as well as the following state cancer registries for their help: AL, AZ, AR, CA, CO, CT, DE, FL, GA, ID, IL, IN, IA, KY, LA, ME, MD, MA, MI, NE, NH, NJ, NY, NC, ND, OH, OK, OR, PA, RI, SC, TN, TX, VA, WA, and WY. The authors assume full responsibility for analyses and interpretation of these data.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received November 17, 2020; revised March 3, 2021; accepted May 27, 2021; published first June 17, 2021.

REFERENCES

1. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. *Nature* 2013;500:415–21.
2. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. *Nature* 1999;401:788–91.
3. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. *Nature* 2020;578:94–101.
4. Grollman JE, de Voer RM, Elsayed FA, Nielsen M, Weren RDA, Palles C, et al. Mutational signature analysis reveals NTHL1 deficiency to cause a multi-tumor phenotype. *Cancer Cell* 2019;35:256–66.
5. Kucab JE, Zou X, Morganella S, Joel M, Nanda AS, Nagy E, et al. A compendium of mutational signatures of environmental agents. *Cell* 2019;177:821–36.
6. Pleguezuelos-Manzano C, Puschhof J, Huber AR, van Hoeck A, Wood HM, Nomburg J, et al. Mutational signature in colorectal cancer caused by genotoxic *pkcsE. coli*. *Nature* 2020;580:269–73.
7. Platz EA, Willett WC, Colditz GA, Rimm EB, Spiegelman D, Giovannucci E. Proportion of colon cancer risk that might be preventable in a cohort of middle-aged US men. *Cancer Causes Control* 2000;11:579–88.

8. Bouvard V, Loomis D, Guyton KZ, Grosse Y, Ghissassi FE, Benbrahim-Tallaa L, et al. Carcinogenicity of consumption of red and processed meat. *Lancet Oncol* 2015;16:1599–600.
9. Bernstein AM, Song M, Zhang X, Pan A, Wang M, Fuchs CS, et al. Processed and unprocessed red meat and risk of colorectal cancer: analysis by tumor location and modification by time. *PLoS One* 2015;10:e0135959.
10. Larsson SC, Rafter J, Holmberg L, Bergkvist L, Wolk A. Red meat consumption and risk of cancers of the proximal colon, distal colon and rectum: the Swedish Mammography Cohort. *Int J Cancer* 2005;113:829–34.
11. Bastide NM, Pierre FHF, Corpet DE. Heme iron from meat and risk of colorectal cancer: a meta-analysis and a review of the mechanisms involved. *Cancer Prev Res* 2011;4:177–84.
12. Liao X, Lochhead P, Nishihara R, Morikawa T, Kuchiba A, Yamauchi M, et al. Aspirin use, tumor PIK3CA mutation, and colorectal-cancer survival. *N Engl J Med* 2012;367:1596–606.
13. Lubbe SJ, Di Bernardo MC, Chandler IP, Houlston RS. Clinical implications of the colorectal cancer risk associated with MUTYH mutation. *J Clin Oncol* 2009;27:3975–80.
14. Lee-Six H, Olafsson S, Ellis P, Osborne RJ, Sanders MA, Moore L, et al. The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* 2019;574:532–7.
15. Zhang J, Stevens MFG, Bradshaw TD. Temozolomide: mechanisms of action, repair and resistance. *Curr Mol Pharmacol* 2012;5:102–14.
16. Fahrer J, Frisch J, Nagel G, Reißig S, Waisman A, Samson LD, et al. Dose-response of alkylation-induced colorectal carcinogenesis in MGMT-proficient and -deficient mice. *Toxicol Lett* 2013;221:S71.
17. Povey AC, Badawi AF, Cooper DP, Hall CN, Harrison KL, Jackson PE, et al. DNA alkylation and repair in the large bowel: animal and human studies. *J Nutr* 2002;132:3518S–21S.
18. Bingham SA, Day NE, Luben R, Ferrari P, Slimani N, Norat T, et al. Dietary fibre in food and protection against colorectal cancer in the European Prospective Investigation into Cancer and Nutrition (EPIC): an observational study. *Lancet* 2003;361:1496–501.
19. Billson HA, Harrison KL, Lees NP, Hall CN, Margison GP, Povey AC. Dietary variables associated with DNA N7-methylguanine levels and O6-alkylguanine DNA-alkyltransferase activity in human colorectal mucosa. *Carcinogenesis* 2009;30:615–20.
20. Chao A, Thun MJ, Connell CJ, McCullough ML, Jacobs EJ, Flanders WD, et al. Meat consumption and risk of colorectal cancer. *JAMA* 2005;293:172–82.
21. Brink M, Weijenberg MP, de Goeij AFPM, Roemen GMJM, Lentjes MHFM, de Bruïne AP, et al. Meat consumption and K-ras mutations in sporadic colon and rectal cancer in The Netherlands Cohort Study. *Br J Cancer* 2005;92:1310–20.
22. Gilsing AMJ, Fransen F, de Kok TM, Goldbohm AR, Schouten LJ, de Bruïne AP, et al. Dietary heme iron and the risk of colorectal cancer with specific mutations in KRAS and APC. *Carcinogenesis* 2013;34:2757–66.
23. Song M, Vogelstein B, Giovannucci EL, Willett WC, Tomasetti C. Cancer prevention: molecular and epidemiologic consensus. *Science* 2018;361:1317–8.
24. Nishihara R, Lochhead P, Kuchiba A, Jung S, Yamauchi M, Liao X, et al. Aspirin use and risk of colorectal cancer according to BRAF mutation status. *JAMA* 2013;309:2563–71.
25. Giannakis M, Mu XJ, Shukla SA, Qian ZR, Cohen O, Nishihara R, et al. Genomic correlates of immune-cell infiltrates in colorectal carcinoma. *Cell Rep* 2016;17:1206.
26. Willett WC, Sampson L, Stampfer MJ, Rosner B, Bain C, Witschi J, et al. Reproducibility and validity of a semiquantitative food frequency questionnaire. *Am J Epidemiol* 1985;122:51–65.
27. Rimm EB, Giovannucci EL, Stampfer MJ, Colditz GA, Litin LB, Willett WC. Reproducibility and validity of an expanded self-administered semiquantitative food frequency questionnaire among male health professionals. *Am J Epidemiol* 1992;135:1114–26.
28. Ogino S, Kawasaki T, Brahmmandam M, Cantor M, Kirkner GJ, Spiegelman D, et al. Precision and performance characteristics of bisulfite conversion and real-time PCR (MethylLight) for quantitative DNA methylation analysis. *J Mol Diagn* 2006;8:209–17.
29. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* 2013;31:213–9.
30. Cibulskis K, McKenna A, Fennell T, Banks E, DePristo M, Getz G. ContEst: estimating cross-contamination of human samples in next-generation sequencing data. *Bioinformatics* 2011;27:2601–2.
31. Saunders CT, Wong WSW, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* 2012;28:1811–7.
32. Taylor-Weiner A, Stewart C, Giordano T, Miller M, Rosenberg M, Macbeth A, et al. DeTiN: overcoming tumor-in-normal contamination. *Nat Methods* 2018;15:531–4.
33. Landau DA, Carter SL, Stojanov P, McKenna A, Stevenson K, Lawrence MS, et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* 2013;152:714–26.
34. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* 2014;505:495–501.
35. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kerntsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20:1297–303.
36. Van Allen EM, Wagle N, Stojanov P, Perrin DL, Cibulskis K, Marlow S, et al. Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nat Med* 2014;20:682–8.
37. Costello M, Pugh TJ, Fennell TJ, Stewart C, Lichtenstein L, Meldrim JC, et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res* 2013;41:e67.
38. Bady P, Sciuscio D, Diserens A-C, Bloch J, van den Bent MJ, Marosi C, et al. MGMT methylation analysis of glioblastoma on the Infinium methylation BeadChip identifies two distinct CpG regions associated with gene silencing and outcome, yielding a prediction model for comparisons across datasets, tumor grades, and CIMP-status. *Acta Neuropathol* 2012;124:547–60.
39. Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* 2010;11:367.
40. Degasperi A, Amarante TD, Czarnecki J, Shooter S, Zou X, Glodzik D, et al. A practical framework and online tool for mutational signature analyses show intertissue variation and driver dependencies. *Nat Cancer* 2020;1:249–63.
41. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005;102:15545–50.
42. Mootha VK, Lindgren CM, Eriksson K-F, Subramanian A, Sihag S, Lehar J, et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* 2003;34:267–73.
43. van Leeuwen EM, Kanterakis A, Deelen P, Kattenberg MV, Genome of the Netherlands Consortium, Slagboom PE, et al. Population-specific genotype imputations using minimac or IMPUTE2. *Nat Protoc* 2015;10:1285–96.
44. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. *Nature*. 2015;526:68–74.
45. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007;81:559–75.
46. Ogino S, Kawasaki T, Kirkner GJ, Kraft P, Loda M, Fuchs CS. Evaluation of markers for CpG island methylator phenotype (CIMP) in colorectal cancer by a large population-based sample. *J Mol Diagn* 2007;9:305–14.
47. Noshio K, Kawasaki T, Ohnishi M, Suemoto Y, Kirkner GJ, Zepf D, et al. PIK3CA mutation in colorectal cancer: relationship with genetic and epigenetic alterations. *Neoplasia* 2008;10:534–41.
48. Ogino S, Kawasaki T, Noshio K, Ohnishi M, Suemoto Y, Kirkner GJ, et al. LINE-1 hypomethylation is inversely associated with microsatellite instability and CpG island methylator phenotype in colorectal cancer. *Int J Cancer* 2008;122:2767–73.