Genetic Signatures of Glucose Homeostasis: Synergistic Interplay With Long-Term Exposure to Cigarette Smoking in Development of Primary Colorectal Cancer Among African American Women

Su Yon Jung, PhD[1,2]

INTRODUCTION: **Insulin resistance (IR)/glucose intolerance is a critical biologic mechanism for the development of colorectal cancer (CRC) in postmenopausal women. Whereas IR and excessive adiposity are more prevalent in African American (AA) women than in White women, AA women are underrepresented in genome-wide studies for systemic regulation of IR and the association with CRC risk.**

METHODS: **With 780 genome-wide IR single-nucleotide polymorphisms (SNPs) among 4,692 AA women, we tested for a causal inference between genetically elevated IR and CRC risk. Furthermore, by incorporating CRC-associated lifestyle factors, we established a prediction model on the basis of gene–environment interactions to generate risk profiles for CRC with the most influential genetic and lifestyle factors.**

RESUTLS: **In the pooled Mendelian randomization analysis, the genetically elevated IR was associated with 9 times increased risk of CRC, but with lack of analytic power. By addressing the variation of individual SNPs in CRC in the prediction model, we detected 4 fasting glucose–specific SNPs in *GCK*, *PCSK1*, and *MTNR1B* and 4 lifestyles, including smoking, aging, prolonged lifetime exposure to endogenous estrogen, and high fat intake, as the most predictive markers of CRC risk. Our joint test for those risk genotypes and lifestyles with smoking revealed the synergistically increased CRC risk, more substantially in women with longer-term exposure to cigarette smoking.**

DISCUSSION: **Our findings may improve CRC prediction ability among medically underrepresented AA women and highlight genetically informed preventive interventions (e.g., smoking cessation; CRC screening to longer-term smokers) for those women at high risk with risk genotypes and behavioral patterns.**

SUPPLEMENTARY MATERIAL accompanies this paper at http://links.lww.com/CTG/A695, http://links.lww.com/CTG/A696, http://links.lww.com/CTG/A697, http://links.lww.com/CTG/A698, http://links.lww.com/CTG/A699, http://links.lww.com/CTG/A700

Clinical and Translational Gastroenterology 2021;12:e00412. https://doi.org/10.14309/ctg.0000000000000412

INTRODUCTION

Colorectal cancer (CRC) is the leading cause of cancer diagnosis and death in women in the United States and other westernized countries (1), and approximately 90% of new cases and deaths occur in women aged 50 years and older (2). African American (AA) women have the highest CRC incidence and mortality rates among all races/ethnic female groups. Although new cases and deaths due to CRC have decreased throughout all racial/ethnic groups since the mid-2000s (3), AA women still rank first, with incidence and mortality rates of 20% and 35%, respectively, higher than those in Whites during 2012–2016 (2,4). In addition, CRC is the third most common cancer diagnosis and cause of cancer deaths in AA women (4).

Excessive adiposity accounts for up to 60% of CRC susceptibility (5,6), and insulin resistance (IR) or glucose intolerance has been believed to be the major biologic mechanism of colorectal carcinogenesis owing to obesity (7) by explaining more than 40% of the association between obesity and CRC (8). In particular, elevated insulin concentrations promoted the growth of CRC in

[1]Translational Sciences Section, School of Nursing, University of California, Los Angeles, Los Angeles, California, USA; and [2]Jonsson Comprehensive Cancer Center, University of California, Los Angeles, Los Angeles, California, USA. **Correspondence:** Su Yon Jung, PhD. E-mail: sjung@sonnet.ucla.edu

Received June 25, 2021; accepted August 22, 2021; published online October 5, 2021

cell lines (9), and in an animal study (10), increased levels of glucose and insulin, which reflect IR, stimulated colorectal epithelial proliferation (11). In addition, several population-based epidemiologic studies reported that colorectal carcinogenesis is associated with IR or impaired glucose tolerance (12–15). IR promotes mitosis *via* insulin receptors and insulin-like growth factor 1 receptors by dysregulating downstream cellular signaling cascades, leading to enhancement of cellular anabolic status and increased antiapoptosis and cell proliferation (16,17). IR may, thus, initiate and promote CRC cell growth. Obesity and IR disproportionately affect AA women (18,19), suggesting that they tend to be more metabolically unhealthy than White women. A recent DNA methylation study (20) for AAs (mainly female individuals) with CRC detected aberrant methylations of CpG islands in the genes that are involved in an insulin network, supporting the critical role of IR in AA women's colorectal tumorigenesis.

The systemic development of IR can be influenced by not only environmental (21–23) but also genetic and epigenetic factors (24). To detect genetic variations of IR, extensive genomic studies have been performed, but mostly focusing on Whites. AAs, one of the racial/ethnic minorities, are underrepresented in the genome-wide genomic study of IR. Detecting IR-specific genetic characteristics within AAs can contribute to advanced understanding of molecular biology related to IR in the AA population and further, as potential risk biomarkers, improve prediction accuracy for CRC development. Thus, this may highlight the promotion of IR-specific, genetically informed, personalized interventions for CRC preventive and therapeutic efforts. In addition, IR phenotypes themselves explained a small to moderate proportion of CRC variation (13,15,25), implying a potential role of IR genetic signatures in validating the causal pathways of colorectal carcinogenesis.

Furthermore, adherence to the World Cancer Research Fund/American Institute for Cancer Research recommendations, such as those for diet, physical activity, and body weight control (26), did not substantially prevent CRC development in AA women (27). This finding suggests the need for alternative strategies that are more predictive of CRC risk, such as the combination of genetic and environmental factors (e.g., lifestyles) that synergistically interact, ultimately leading to CRC initiation and progression.

For those reasons, we performed a genomic gene–environment (G×E) interaction study by focusing on IR and relevant lifestyle factors among AA postmenopausal women. First, we examined 780 IR single-nucleotide polymorphisms (SNPs) that were detected as top signals from independent genome-wide association (GWA) studies (28–33). After validating GWA IR-SNPs in our data, we tested for a causal inference between genetically elevated IR and CRC development. Next, we identified CRC-related lifestyle factors from the literature, which disproportionately affect AA women. By incorporating those lifestyle factors with the validated IR genetic markers, we established a risk prediction model and computed risk prediction of variables for CRC. With the most influential genetic and lifestyle factors, we eventually generated CRC risk profiles and estimated their combined and joint effects on CRC risk. We surmised that our multimodal approach could resolve the inconclusive findings from previous studies of IR and lifestyle factors in association with CRC and, thus, improve the predictive power for CRC risk in medically and scientifically underrepresented AA women.

MATERIALS AND METHODS

Study population

We examined AA postmenopausal women who had been enrolled in the SNP Health Association Resource (SHARe), a prospective cohort of the AA and Hispanic minorities, which is part of Women's Health Initiative Database for Genotypes and Phenotypes (WHI dbGaP) Harmonized and Imputed GWA Studies, with an effort to detect genes/genetic variants associated with quantitative traits with enhanced statistical power in those racial/ethnic minorities. Detailed descriptions of the study design and rationale have been published (34–36). In brief, healthy women in the WHI study were recruited between 1993 and 1998 at 40 WHI-designated clinical centers across the United States if they were aged 50–79 years, postmenopausal, and able to provide written informed consent. Further, they were enrolled in the WHI dbGaP study if they had met eligibility for data submission to dbGaP and provided their DNA samples. For 7,470 women who reported their race or ethnicity as AA, we applied the following exclusion criteria: quality control (QC) of genomic data, diabetes history, less than 1-year follow-up, and diagnosis of any cancer type at screening. Our final study sample included 4,692 AA women. They were followed up through August 2014, with a 15-year median follow-up end point. By their last follow-up, 73 women (1.6%) in this group had developed primary CRC. Our study was approved by the institutional review boards of each WHI participating clinical center and the University of California, Los Angeles.

IR genetic variants selection

IR genetic variants were selected on the basis of the publicly available data resource on glycemic traits, the Meta-Analyses of Glucose and Insulin-related traits Consortium (www.magicinvestigators.org) (28–31). The MAGIC analyzed fasting glucose (FG) and fasting insulin (FI) levels as continuous variables. Two other GWA data resources for an AA cohort were used: 1 (32) found SNPs in a 500-kb linkage disequilibrium (LD) block associated with FG, and the other (33) detected functional SNPs for glucose intolerance. From a total of 1,344 FG-SNPs and 313 FI-SNPs identified in those GWA studies, 689 FG-SNPs and 91 FI-SNPs are available in our AA SHARe genomic data set; among those SNPs, 94 FG-SNPs (34 index in LD < 0.3) and 8 FI-SNPs (4 index in LD < 0.3) were validated with a relevant phenotype.

Genotyping and phenotyping

Genotyping data for AA women were extracted from the WHI dbGaP SHARe for our study. Detailed information on the genotyping has been reported (34,36). DNA samples were derived from participant blood samples at baseline and genotyped *via* Affymetrix 6.0 (Affymetrix, Inc., Santa Clara, CA) at the Fred Hutchinson Cancer Research Center in Seattle, WA. Data were normalized to Genome Reference Consortium Human Build 37, imputed with the 1,000 genomes reference panels, and harmonized *via* pairwise concordance among samples across WHI GWA studies. We performed genomic data QC by filtering out SNPs with a missing call rate of ≥2%, a Hardy-Weinberg equilibrium of $P < 1E–04$, and $R^2 < 0.6$ imputation quality (37). We further excluded individuals with unexpected duplicates, first-degree and second-degree relatives, and outliers on the basis of genetic principal components (PCs).

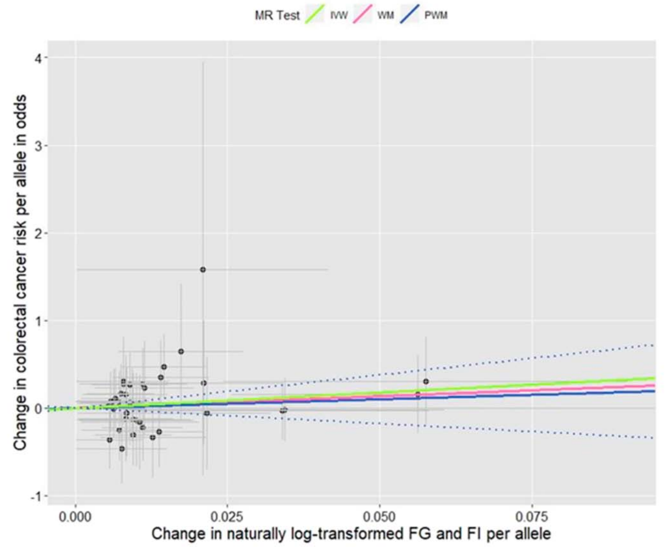

Figure 1. Scatter plot for the effects of 38 individual FG- and FI-genetic instrumental variables on colorectal cancer risk. Each black dot reflects a genome-wide FG/FI–raising genetic variant. The green and pink lines indicate IVW and WM estimates, respectively. The blue line indicates PWM estimates and 95% CIs. (PWM HR = 9.24, 95% CI: 0.03–2.95E+03; MR-Egger intercept *P* value = 0.701). CI, confidence interval; FG, fasting glucose; FI, fasting insulin; HR, hazard ratio; IVW, inverse variance–weighted; MR, Mendelian randomization; PC, principal components; PWM, penalized weighted median; WM, weighted median. Note: All MR estimates were based on the phenotype association and cancer association with genetic instruments, each of which was adjusted for age and 10 genetic PCs.

Lifestyle factors and CRC outcome

We performed a literature review (2,4,6,7,27,38–42) on lifestyle factors that are relevant to CRC in AA women and extracted lifestyle variables from the SHARe data: age at enrollment; family history of CRC (genetic inheritance), anthropometric measures (body mass index and abdominal adiposity, including waist circumference and waist-to-hip ratio), physical activity, alcohol intake (dietary alcohol per day and alcohol intake history), smoking (years as a regular smoker and number of cigarettes per day), and nutrition (dietary fiber, daily fruits and vegetables, percentage calories from protein, percentage calories from saturated and monounsaturated and polyunsaturated fatty acids [PFAs], dietary calcium, vitamin K, and total sugars). Furthermore, we added to our data analysis the following: demographic and socioeconomic variables (education, marital status, and employment), comorbid conditions (depressive symptoms, lipid metabolic profiles, cardiovascular disease ever, and hypertension ever), and reproductive histories (ages at menarche and menopause, oophorectomy and hysterectomy, duration of oral contraceptive use, number of pregnancies, duration of breast feeding, and exogenous estrogen [E] use [unopposed and opposed (E plus progestin)]). All the variables had been recorded at baseline from participants *via* self-administered questionnaires (except height, weight, and waist/hip circumferences, which were measured by trained staff). The coordinating clinical centers monitored the data collection process as part of data QC. With a total of 35 variables, we conducted preliminary univariate and stepwise multiple regressions for CRC risk and checked multicollinearity among variables.

Primary CRC diagnosis of the study participants was confirmed *via* a centralized review of medical records and pathology and cytology reports by the WHI committee of physicians, who followed the National Cancer Institute's Surveillance, Epidemiology, and End Results guidelines (43). The time between enrollment and CRC diagnosis, censoring, or study end point was estimated, first in days, and then converted into years.

Statistical analysis

Linear and Cox proportional hazards regressions, respectively, were used to estimate the associations of GWA IR-SNPs with naturally log-transformed FG (mg/dL)/FI (μIU/mL) and with CRC risk, both of which were adjusted for age and 10 genetic PCs. The assumptions for each regression were met. A 2-tailed $P <$ 0.05 for validation tests with FG and FI was considered nominally significant, and after the Bonferroni correction for multiple comparisons, $P <$ 7E-05 for FG and $P <$ 5E-04 for FI were considered statistically significant.

To test for the causal pathway between FG/FI and CRC risk, we performed Mendelian randomization (MR) analysis using GWA SNPs as genetic instruments. First, we checked the assumptions to make a valid inference: (i) F statistics (44) of 8.0 for

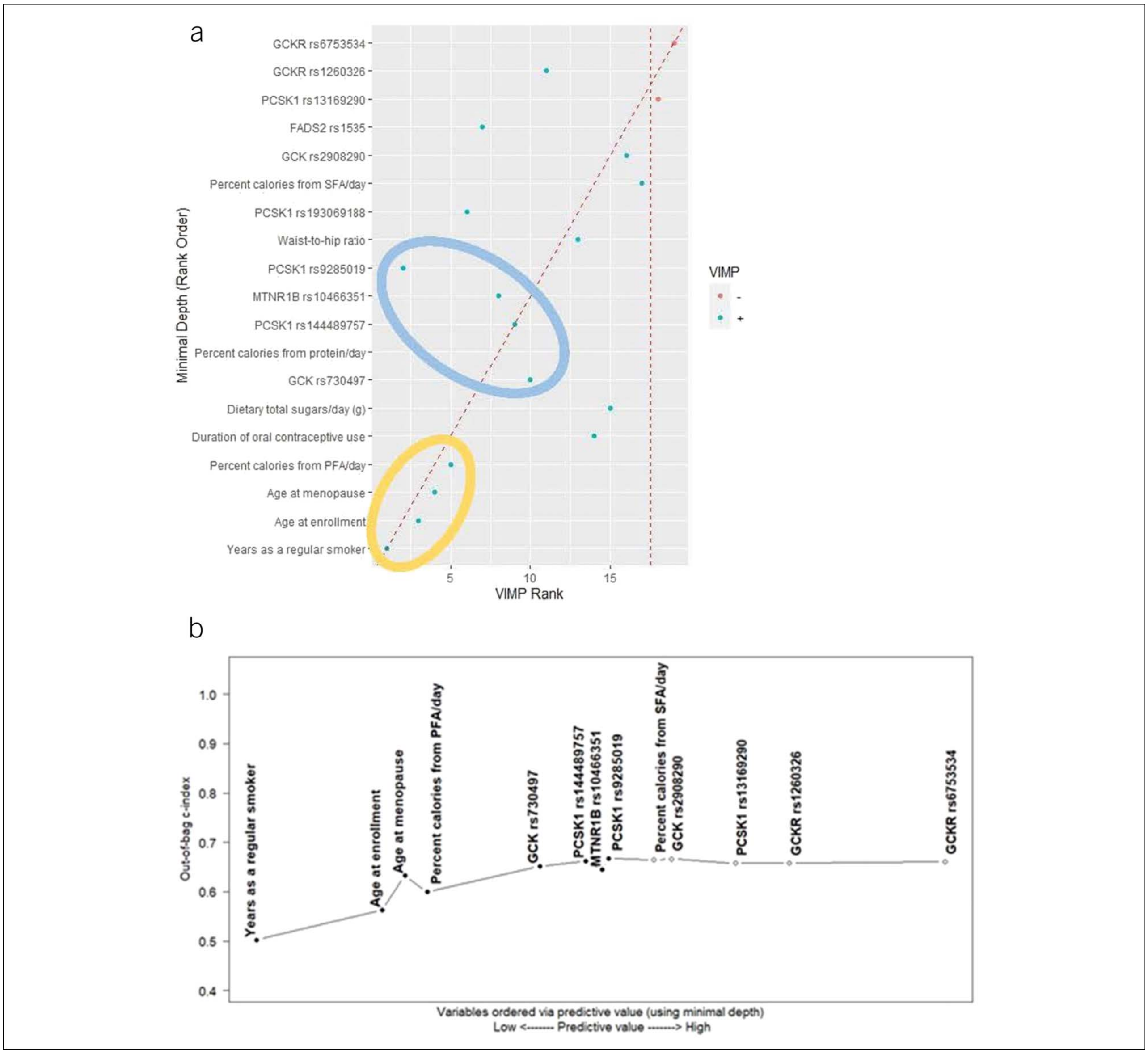

Figure 2. The second stage of RSF analysis using 10 single-nucleotide polymorphisms and 9 behavioral factors selected from the first stage of RSF analysis. (**a**) Comparing rankings between minimal depth and VIMP. PFA, polyunsaturated fatty acid; RSF, random survival forest; SFA, saturated fatty acid; VIMP, variable of importance. Note: The 4 behavioral variables within the gold ellipse and the 4 genetic markers within the blue ellipse were identified as the topmost influential predictors. (**b**) Out-of-bag concordance index (c-index). (Improvement in c-index was observed when the top 8 variables [●] were added to the model, whereas other variables [○] did not further improve the accuracy of prediction.)

FG and 10.2 for FI index-SNPs reflect enough strength (45); (ii) for horizontal pleiotropy, pleiotropic SNPs associated with BMI were identified (see Supplementary Table 1, http://links.lww.com/CTG/A696) (46,47), and no SNPs overlapped with our modeled SNPs; and (iii) for vertical directional pleiotropy, the MR-Egger regression analysis (48) was performed, and no evidence of pleiotropy was observed. Having confirmed that our genetic instruments met the MR conditions, we assumed additive effects of SNPs on phenotype and computed an inverse variance–weighed (IVW) estimate and standard error to test for the causal pathway between genetically determined FG/FI and CRC risk as follows (49,50):

Here X_k is the observed mean change in the phenotype per additional allele at genetic variant $_K$ ($_K = 1\ldots K$), and σ_{Xk} is the associated standard error; Y_k is the observed log odds change in the BC outcome per allele at genetic variant $_K$, and σ_{Yk} is the associated standard error.

In addition to the IVW estimate, we used weighted median and penalized weighted median (PWM) estimates that allow up to 50% of genetic variants' invalidity. These alternative methods can provide a more consistent estimate of the causal effect by assigning a weight to the ordered estimates, establishing linearity between neighboring estimates, and by down-weighting outlying genetic variants with heterogeneous estimates (51,52). In addition, the PWM is believed to be a better estimate if there is directional pleiotropy. The results from MR analysis were reported as risk ratios (hazard ratios [HRs]) and 95% confidence intervals (CIs) for the change in CRC risk per unit increase in naturally log-

Table 1. RSF second-stage analysis: predictive values of variables for colorectal cancer risk

Variable[a]	Minimal depth[b]	VIMP	C-index	Incremental error[c]	Drop error[d]
Years as a regular smoker[e]	2.9426	0.0202	0.5026	0.4974	0.0026
Age at enrollment[e]	3.4192	0.0072	0.5635	0.4366	0.0609
Age at menopause[e]	3.5056	0.0065	0.6325	0.3675	0.0691
Percentage calories from PFA/day[e]	3.5902	0.0047	0.5993	0.4007	−0.0332
Duration of oral contraceptive use	3.6041	0.0014	0.6217	0.3783	0.0224
Dietary total sugars/day (g)	3.8086	0.0011	0.6434	0.3567	0.0217
GCK rs730497[f]	4.0185	0.0034	0.6513	0.3487	0.0080
Percentage calories from protein/day	4.0939	0.0025	0.6595	0.3405	0.0082
PCSK1 rs144489757[f]	4.1934	0.0038	0.6622	0.3378	0.0027
MTNR1B rs10466351[f]	4.2558	0.0040	0.6446	0.3554	−0.0176
PCSK1 rs9285019[f]	4.2797	0.0072	0.6671	0.3329	0.0225
Waist-to-hip ratio	4.2984	0.0019	0.6626	0.3374	−0.0045
PCSK1 rs193069188	4.4524	0.0046	0.6666	0.3334	0.0040
Percentage calories from SFA/day	4.4530	0.0002	0.6643	0.3357	−0.0023
GCK rs2908290	4.5190	0.0009	0.6666	0.3334	0.0023
FADS2 rs1535	4.6099	0.0045	0.6747	0.3253	0.0082
PCSK1 rs13169290	4.7623	−0.0003	0.6576	0.3424	−0.0172
GCKR rs1260326	4.9678	0.0028	0.6578	0.3422	0.0002
GCKR rs6753534	5.5591	−0.0011	0.6603	0.3397	0.0025

C-index, concordance index; PFA, polyunsaturated fatty acid; RSF, random survival forest; SFA, saturated fatty acid; VIMP, variable of importance.

[a] Variables ordered by minimal depth.

[b] Minimal depth is the predictive value of the variable estimated from the nested RSF models with a lower value being likely to have a greater impact on prediction.

[c] The incremental error rate was calculated in the nested sequence of models starting with the top variable, followed by the model with the top 2 variables, then the model with the top 3 variables, and so on. For example, the third error rate was computed from the third nested model, including the first, second, and third variables.

[d] The drop error rate of the variable was calculated by the difference between the error rates of a previous and the corresponding variable from the nested models. For example, the drop error rate of the second variable was estimated by the difference between the error rates from the first and second nested models. The error rate for the null model is set at 0.5; thus, the drop error rate for the first variable was obtained by subtracting the error rate (0.4974) from 0.5.

[e] Variables were selected as the most predictive behavioral markers on the basis of multimodal predictive values.

[f] Variables were selected as the most predictive genetic markers on the basis of multimodal predictive values.

transformed FG/FI. As another measure of pleiotropy, the heterogeneity of the MR estimates across genetic instruments was evaluated *via* Cochran Q test.

Next, we performed the Random Survival Forest (RSF) analysis with index and individual SNPs and lifestyle variables. RSF is a nonparametric tree-based ensemble machine-learning method that accounts for the nonlinear effects and high-order interactions among variables (53); it has been shown to outperform traditional prediction models, thus successfully yielding accurate predictions (54–58). A tree from each bootstrapped sample was generated to maximize risk differences across daughter nodes, and this process was repeated numerous times (n = 5,000 trees in this study) to create a forest of trees (54,59). By using the out-of-bag (OOB) data, the prediction error (i.e., misclassification probability) was estimated to calculate the OOB concordance index (c-index = 1 − prediction error). The OOB c-index is a quantitative measure of prediction performance, conceptually similar to the area under the receiver operating characteristic curve (59,60). The predictive power of each variable was determined *via* 2 values: (i) minimal depth (MD), in which variables with a small MD are highly predictive, and (ii) variable importance (VIMP), calculated from the permutation of OOB datasets, in which variables with a larger VIMP are more predictive (53,61).

We applied a multimodal 2-stage RSF approach. The first RSF (see Supplementary Figure 1, http://links.lww.com/CTG/A695) comprised separate analysis of SNPs and lifestyle variables, and then only SNPs/lifestyles with significantly low MD and high VIMP estimates were carried over to the second RSF. This strategy excluded variables without sufficient effects on CRC risk, giving more statistical power with the correct type I error in the second stage. In the second RSF, we used a multimodal approach to detect the most predictive genetic and lifestyle factors: (i) comparison between the MD and VIMP estimates in the plot, (ii) estimation of OOB c-index with the nested RSF model, and (iii) computation of the incremental error rate of each variable within the nested sequenced RSF models. With the topmost influential variables identified, we eventually estimated the combined and joint effects on CRC risk using multiple Cox regression adjusted for covariates. A 2-tailed *P* value was corrected for multiple comparisons *via* the Benjamini-Hochberg method, and a 5% false discovery rate was statistically significant. Multiple R packages

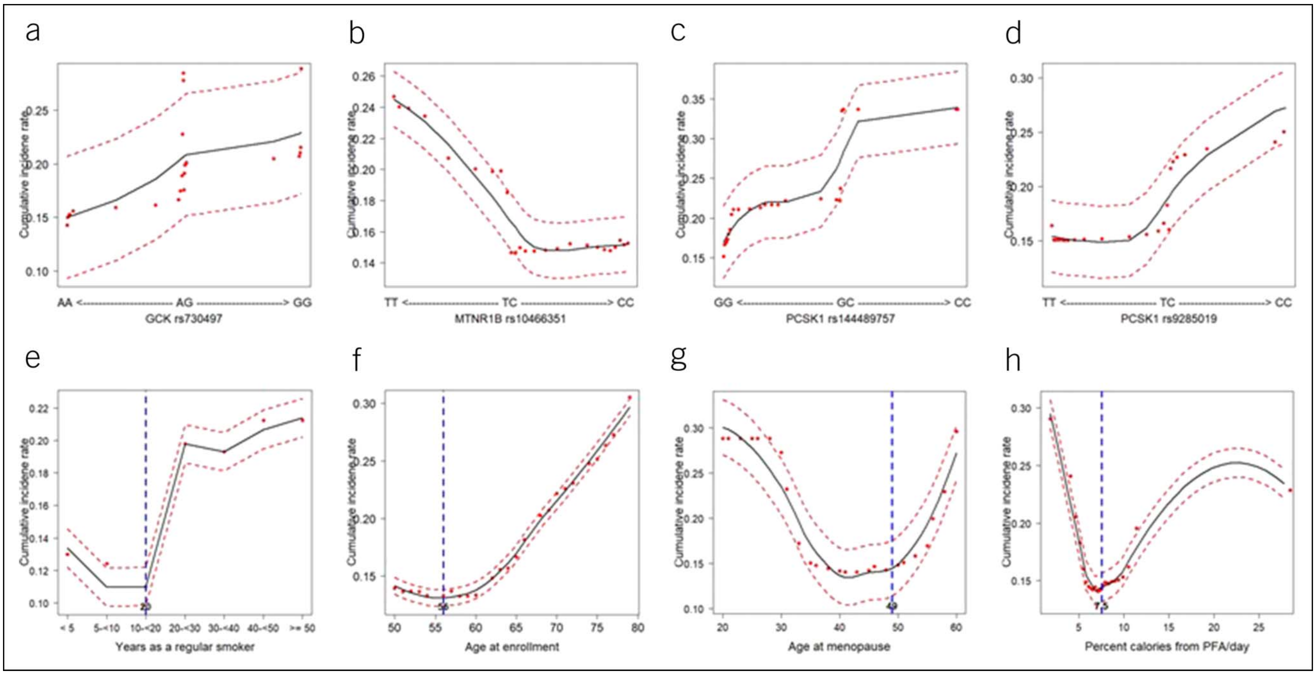

Figure 3. Cumulative incidence rates of colorectal cancer for the 8 topmost predictive variables: 4 single-nucleotide polymorphisms and 4 behavioral factors selected on the basis of a random survival forest analysis. AA, African American; PFA, polyunsaturated fatty acid. Note: Dashed red lines indicate 95% confidence intervals.

were used (R v4.0.4, survival, metaphor, forest plot, survival ROC, random Forest SRC, ggRandomForests, ggplot2, ggthemes, and gamlss).

Ethical approval and consent to participate: Our study was approved by the institutional review boards of each WHI participating clinical center and the University of California, Los Angeles.

RESULTS

MR analysis

Among the 94 FG-SNPs (see Supplementary Table 2A, http://links.lww.com/CTG/A697) and 8 FI-SNPs (see Supplementary Table 2B, http://links.lww.com/CTG/A697) that were validated at nominal significance and after multiple comparison corrections, 34 index FG-SNPs and 4 index FI-SNPs in LD < 0.3 were carried over to our MR analysis (see Supplementary Table 3A, http://links.lww.com/CTG/A698). In the pooled analysis of 38 FG/FI SNPs, genetically elevated IR was associated with 9 times increased risk of CRC, but that finding lacked statistical power (Figure 1; see Supplementary Table 4A, http://links.lww.com/CTG/A699 [PWM HR = 9.24, 95% confidence interval: 0.03–2.95E+03]). The subset analysis of 3 index SNPs (2 FG and 1 FI) at significance after correction for multiple comparisons revealed a greater effect of genetically elevated IR on CRC risk but, again, not reaching statistical significance (see Supplementary Table 4A, http://links.lww.com/CTG/A699). The phenotype-specific analyses for genetically determined FG and FI (see Supplementary Tables 4B and 4C, http://links.lww.com/CTG/A699) showed similar patterns. For example, a 1-unit increase in the log-transformed genetically elevated FI was associated with 7 times higher risk of CRC, but without sufficient statistical power.

Multimodal 2-stage RSF analysis

We analyzed all phenotype-specific individual SNPs in addition to index SNPs in the first RSF prediction model to address the possibility that a combined analysis of only index SNPs may mask individual SNP variation in the risk of CRC development. By using 2 prediction estimates, MD and VIMP, we generated plots for lifestyle factors and SNPs separately to compare the 2 measures in the plot (see Supplementary Figure 1, http://links.lww.com/CTG/A695) and detected the most influential lifestyle and genetic factors that were in agreement with high ranks: 9 of 35 lifestyle factors, 8 of 34 FG index SNPs, 6 (4 of which overlap with the selected FG indexes) of 94 FG individual SNPs, and 1 index (overlapping with the selected FG indexes) of 4 FI-index and 8 FI-individual SNPs each.

Next, we performed the second stage of RSF with a multimodal approach. Using the selected 9 lifestyle factors and 10 FG/FI SNPs, we first plotted the 2 estimates, MD and VIMP (Figure 2a). Both estimates with high ranks detected 4 lifestyle factors (years as a regular smoker, age at enrollment, age at menopause, and percentage daily calories from PFA) and 4 FG-specific SNPs (*GCK* rs730497, *PCSK1* rs144489757, *MTNR1B* rs10466351, and *PCSK1* rs9285019) as the most predictive variables for CRC development. Of note, 2 of the selected SNPs (*GCK* rs730497 and *MTNR1B* rs10466351) are indexes. Next, we computed the c-index (i.e., area under the receiver operating characteristic) from the nested RSF models and plotted with variables ordered by MD rank (Figure 2b), revealing the same set of the topmost 4 lifestyle and 4 genetic variables that showed distinction to improve prediction ability compared with the rest of the variables. This implies the usefulness of the c-index in complementary analysis to determine variables' prediction ability. Finally, we estimated a drop error in each variable ranked by MD in the nested sequence of RSF models (Table 1), detecting once again the same top 8

Table 2. **Results of combined and joint tests for smoking with risk genotypes predicting colorectal cancer risk**

SNP[a]	Total		Never smokers			Regular smoker for <20 yr		
No. of risk	HR[b] (95% CI)	*P*	n	HR[b] (95% CI)	*P*	N	HR[b] (95% CI)	*P*
0	Reference		810	Reference		733	2.87 (0.731–11.287)	0.1307
1	**2.18 (1.317–3.597)**	**0.0024**[c]	233	1.14 (0.119–10.992)	0.9086	221	**5.78 (1.275–26.164)**	**0.0229**
2	**2.08 (1.007–4.294)**	**0.0479**	90	3.32 (0.342–32.159)	0.3010	76	**7.49 (1.238–45.370)**	**0.0284**
P_{trend}		0.0500	P_{trend}					0.1000

	Never smokers			Regular smoker for ≥20 yr		
	n	HR[b] (95% CI)	*P*	N	HR[b] (95% CI)	*P*
	810	Reference		1,799	**5.01 (1.515–16.569)**	**0.0083**[c]
	233	1.14 (0.119–10.981)	0.9084	532	**11.70 (3.460–39.531)**	**0.0001**[c]
	90	2.90 (0.301–27.873)	0.3570	198	**9.55 (2.376–38.374)**	**0.0015**[c]
	P_{trend}					3.00E-04

Numbers in bold face are statistically significant.

CI, confidence interval; FDR, false discovery rate; HR, hazard ratio; SNP, single-nucleotide polymorphism.

[a] The number of risk genotypes (*GCK* rs730497 AG+GG; *MTNR1B* rs10466351 TT; *PCSK1* rs144489757GC+CC; and *PCSK1 rs9285019* TC+CC) was defined as follows: 0 (none or 1 risk allele) vs 1 (2 risk alleles) vs 2 (3 or more risk alleles).

[b] Multivariate regression for risk genotypes was adjusted by waist-to-hip ratio, duration of oral contraceptive use, dietary total sugars/day (g), percentage calories from protein/day, and percentage calories from saturated fatty acid/day.

[c] *P* value with FDR <0.05 was presented after multiple comparison corrections *via* the Benjamini-Hochberg method.

variables as the most influential lifestyle and genetic factors that contribute to reducing the prediction error rate.

The selected topmost IR SNPs and lifestyles: combined and joint effects on CRC risk

With the topmost influential IR-SNPs and lifestyle factors, we computed the predictive cumulative CRC incidence rate, implementing the RSF machine-learning process that accounts for the confounding variables and potential nonlinearity effect of each variable on CRC outcomes (Figure 3). The risk genotypes of each SNP were accordingly categorized for further analysis (Figure 3a–d): *GCK* rs730497 AG+GG; *MTNR1B* rs10466351 TT; *PCSK1* rs144489757GC + CC; and *PCSK1* rs9285019 TC+CC. In addition, corresponding to the cutoff values in Figure 3e–h, risk lifestyles were defined as ≥20 years of regular smoking; age older than56 years at enrollment; age older than 49 years at menopause; and ≥7.5% of daily calories from PFA. Having categorized those genetic and lifestyle variables, we first computed their individual risk of CRC (by adjusting or not adjusting for each other), with HRs of SNPs ranging from 1.64 to 1.81 (see Supplementary Tables 5A and 5B, http://links.lww.com/CTG/A700) and HRs of lifestyles ranging from 1.59 to 2.92 (see Supplementary Tables 5C and 5D, http://links.lww.com/CTG/A700), confirming their single effects on CRC risk.

However, when those genetic and lifestyle factors were combined and tested for their joint effects with smoking, much stronger risks for CRC development were observed (Tables 2 and 3). First, we combined the selected 4 SNPs and evaluated for the risk of CRC (Table 2), revealing that the presence of ≥3 risk genotypes was associated with 2 times greater risk of CRC than null or 1 risk genotype. Next, to test for the combined genetic effects on CRC risk jointly with smoking, we categorized smokers as never smokers and shorter-term (<20 years) and longer-term (≥20 years) regular smokers and compared never smokers with (i) shorter-term regular smokers and (ii) longer-term regular smokers. Overall, the joint effect of SNPs with smoking was apparent in both comparisons (Table 2). In detail, compared with the never smokers who carried null or 1 risk genotype, the shorter-term regular smokers who carried ≥3 risk genotypes had an almost 7 times higher risk of CRC. The joint effect of smoking was much greater when never smokers were compared with longer-term regular smokers: an almost 10 times higher CRC risk was detected in the longer-term smokers with ≥3 risk genotypes than in the never smokers with null or 1 risk genotype.

Similar patterns were observed when we examined the selected 4 lifestyles for their combined and joint effects with smoking (Table 3). For example, women who had 4 risk lifestyles had a 3.5 times higher risk of CRC than those who had ≤2 risk lifestyles. Furthermore, when women were stratified by smoking status, no joint effect of smoking was detected in the comparison between never smokers and shorter-term regular smokers. However, the joint effect of smoking was distinct between never smokers and longer-term regular smokers. That is, compared with the never smokers who had <2 risk lifestyles, the longer-term regular smokers who had 3 risk lifestyles were associated with 6.6 times higher risk of CRC, suggesting an apparent effect of prolonged exposure to cigarette smoking on lifestyles, which led to increased CRC risk.

Furthermore, we combined the IR SNPs and lifestyles and tested for CRC risk (Table 4), detecting a 5 times higher risk of CRC in women with combined SNPs and lifestyles than in those without both factors; this risk is substantially higher than those for the separate combinations of SNPs and lifestyles (e.g., 2 times and 3 times higher risk, respectively). In addition, a joint effect of the combined SNPs and lifestyles with smoking was detected. In detail, compared with the never smokers who had null genotypes and lifestyles, shorter-term regular smokers who had both risk genotypes and lifestyles had a 7 times greater CRC risk.

Table 3. Results of combined and joint tests for smoking with risk behaviors predicting colorectal cancer risk

Behavior[a]	Total			Never smokers			Regular smoker for <20 yr	
No. of risk	HR[b] (95% CI)	*P*	n	HR[b] (95% CI)	*P*	n	HR[b] (95% CI)	*P*
0	Reference		563	Reference		531	2.84 (0.721–11.149)	0.1356
1	**2.82 (1.691–4.703)**	**0.0001**[c]	427	0.48 (0.050–4.620)	0.5239	365	3.14 (0.726–13.541)	0.1256
2	**3.51 (1.767–6.979)**	**0.0003**[c]	143	1.76 (0.178–17.450)	0.6281	134	1.90 (0.192–18.888)	0.5828
P_{trend}		0.0010	P_{trend}					0.2000
				Never smokers			Regular smoker for ≥20 yrs	
			n	HR[b] (95% CI)	*P*	n	HR[b] (95% CI)	*P*
			563	Reference		1,072	2.50 (0.701–8.913)	0.1580
			427	0.46 (0.048–4.410)	0.4995	1,059	**6.80 (2.052–22.530)**	**0.0017**[c]
			143	1.30 (0.134–12.489)	0.8225	398	**6.61 (1.843–23.740)**	**0.0038**[c]
			P_{trend}					2.00E-04

Numbers in bold face are statistically significant.

CI, confidence interval; FDR, false discovery rate; HR, hazard ratio.

[a] The number of behavioral factors (years as a regular smoker <20 yr vs ≥20 yr [in overall analysis only]; age at enrollment ≤56 vs >56 yr; age at menopause ≤49 vs >49 yr; and percent calories from polyunsaturated fatty acid/day <7.5% vs ≥7.5%) was defined as follows: 0 (none, 1, or 2 risk behaviors) vs 1 (3 risk behaviors) vs 2 (4 risk behaviors).

[b] Multivariate regression for risk genotypes was adjusted by waist-to-hip ratio, duration of oral contraceptive use, dietary total sugars/day (g), percentage calories from protein/day, and percentage calories from saturated fatty acid/day.

[c] *P* value with FDR <0.05 was presented after multiple comparison corrections *via* the Benjamini-Hochberg method.

Furthermore, the difference in risk of CRC owing to smoking effect was greater in the comparison of never smokers with longer-term regular smokers (8 times higher risk) than in the comparison of never smokers with shorter-term regular smokers (7 times higher risk). Notably, across the joint tests for shorter-term/longer-term smoking effect, the CRC risk magnitude from the combined SNPs and lifestyles was not greater than the sum of both risks from the separate combinations, suggesting that genetic and lifestyle factors, when combined, may override smoking's modification on CRC risk to some degree.

DISCUSSION

Our genetic G × E study for AA postmenopausal women evaluated an extensive set of GWA-based IR SNPs for their potential causal pathways of CRC development in an MR framework and further, by incorporating lifestyle factors, tested for interactions in the CRC risk prediction model. Having considered the variations of individual SNPs in CRC risk, we detected 4 FG-specific SNPs (including 2 indexes) and 4 lifestyle factors as the most predictive variables for CRC risk. The joint analysis of those risk genotypes and lifestyles with smoking revealed a gene–lifestyle dose-dependent relationship with a synergistic increase of CRC risk, indicating that CRC risk profiles that combined genetic and behavioral factors substantially improved the CRC risk prediction; the results, thus, lead to the potential promotion of genetically informed interventions for cancer prevention/therapeutic effort.

All 4 selected FG SNPs are located in the intronic and intergenic regions of genes that play well-known roles in regulating glucose metabolism and insulin production/sensitivity, suggesting that their genetic variations may affect glucose homeostasis. In particular, the *GCK* gene encodes a member of the hexokinase family that phosphorylates glucose to produce glucose-6-phosphate (P), the first step in glucose metabolism pathways (62). *GCK* acts as a glucose sensor in the pancreatic beta cells by playing a crucial role in modulating insulin secretion and in the liver by facilitating glucose uptake and conversion to glycogen (62,63). Thus, this gene's mutation has been associated with multiple types of diabetes in both Whites and AAs (64–66). In addition, *GCK* showed potential in tumor development through the glycolysis process from glucose-6P to fructose-6P, guided by phosphoglucose isomerase, which promotes progression in the tumor cells (67,68). However, few population-level genomic studies (69) have been so far conducted for the association of this gene variation with cancer. Our study is the first, to our knowledge, to reveal the *GCK* variant associated with CRC risk, particularly among AA women.

The *PCSK1* gene is one of the genes linked to early-onset obesity, encoding prohormone convertase 1/3, which is involved in the biosynthetic process of prohormones in endocrine tissues, thus regulating food ingestion and glucose homeostasis (70,71). In detail, the convertase one-third mediates the cleavage of proinsulin in the process of insulin biosynthesis; thus, the loss-of-function mutation of the gene leads to increased circulating proinsulin and defects in insulin production, resulting in impaired glucose tolerance, diabetes, and obesity (70,72–74). In relation to carcinogenesis and cancer progress, the mutation of this gene is associated with liver metastasis but is shown partially in the primary CRC cells (75), indicating the selective process involving the convertases during metastasis to the liver. Our finding of the 2 genetic variants in the *PCSK1* associated with primary CRC is the first report and warrants further replication studies with larger independent data sets.

In addition, *MTNR1B* encodes melatonin receptor 1B, which plays a well-established role in insulin production and glucose metabolism in pancreatic islets (76,77). Its genetic variation may disturb circadian rhythm, resulting in glucose intolerance (78). In addition, melatonin receptors have been reported to be involved

Table 4. Results of combined and joint tests for smoking with risk genotypes and behaviors predicting colorectal cancer risk

SNPs combined with behaviors[a] No. of risk[b]	Total HR[c] (95% CI)	 *P*	Never smokers n	 HR[c] (95% CI)	 *P*	Regular smoker for <20 yr n	 HR[c] (95% CI)	 *P*
0	Reference		1,043	Reference		954	**3.41 (1.069–10.878)**	**0.0382**[d]
1	**1.79 (1.034–3.088)**	**0.0375**[d]	90	3.21 (0.355–28.935)	0.2992	76	**7.25 (1.313–40.055)**	**0.0231**[d]
2	**5.42 (1.316–22.347)**	**0.0193**[d]	P_{trend}					0.0900
P_{trend}		0.0800						

	Never smokers n	 HR[c] (95% CI)	 *P*	Regular smoker for ≥20 yr n	 HR[c] (95% CI)	 *P*
	1,043	Reference		1964	**6.09 (2.161–17.166)**	**0.0006**[d]
	90	2.82 (0.315–25.214)	0.3545	565	**8.41 (2.788–25.379)**	**0.0002**[d]
	P_{trend}					0.0020

Numbers in bold face are statistically significant.

CI, confidence interval; FDR, false discovery rate; HR, hazard ratio.

[a]The risk genotypes are *GCK* rs730497 AG+GG; *MTNR1B* rs10466351 TT; *PCSK1* rs144489757GC + CC; and *PCSK1 rs9285019* TC+CC. The behavioral factors are years as a regular smoker <20 vs ≥20 yr [in overall analysis only]; age at enrollment ≤56 vs >56 yr; age at menopause ≤49 vs >49 yr; and percentage calories from polyunsaturated fatty acid/day <7.5% vs ≥7.5%.

[b]The combined number of risk genotypes and risk behaviors was based on risk genotypes defined as 0 (none, 1, or 2 risk alleles) and 1 (3 or more risk alleles) and based on risk behaviors defined as 0 (none, 1, 2, or 3 risk behaviors) and 1 (4 risk behaviors). The ultimate number of risk genotypes combined with risk behaviors was defined as 0 (none of risk genotypes and risk behaviors), 1 (either risk genotypes or risk behaviors), and 2 (both risk genotypes and risk behaviors) in total analysis; in smoker-specific analyses, 0 (none) and 1 (either risk genotypes or risk behaviors or both).

[c]Multivariate regression for risk genotypes was adjusted by waist-to-hip ratio, duration of oral contraceptive use, dietary total sugars/day (g), percentage calories from protein/day, and percentage calories from saturated fatty acid/day.

[d]*P* value with FDR <0.05 was presented after multiple comparison corrections *via* the Benjamini-Hochberg method.

in the mechanism of melatonin-induced inhibitory proliferation in cancer cells of the breast, prostate, and colorectum (79,80). In particular, mRNA expression of those melatonin receptors decreased in colorectal tumor cells (80), suggesting its proactive effect on cancer development. Our genomic study, consistent with a previous genomic study (81), detected the decreased risk of CRC, particularly in those with a CC minor allele (vs TT) despite insufficient analytic power. All the aforementioned particular genotypes of SNPs detected in our study showed in combination a synergistic effect on CRC risk, requiring further validation and functional studies.

In addition to aging (7,42,82), older age at menopause is a critical risk factor of CRC development in postmenopausal women (83–85), indicating that the greater exposure to endogenous estrogen may increase CRC risk. Our analysis of these lifestyle factors combined showed greater risk of CRC than that from analysis of the individual factors. Furthermore, the joint effect of smoking with the genetic and those lifestyle factors was apparent in our study by presenting synergistically increased risk of CRC. Cigarette smoking can account for more than 20% of CRC risk with a dose-response relationship to the number of years of smoking (82). Carcinogens contained in tobacco reach the colorectal mucosa through the digestive system and bloodstream, causing tumorigenesis in the colorectum (86). Among AAs, compared with other races/ethnicities, higher total nicotine equivalents have been found after controlling for the number of cigarettes smoked per day (87). But, little study has been done on AAs in association with CRC. Our results from AA women showed that longer-term (≥20 years) regular smokers with the combination of genetic and lifestyle factors were associated with an 8 times higher risk of CRC than never smokers without both risk factors, although our induction period (median, 15 years) was somewhat shorter than the typical period (88). An analysis of CRC screening among AAs revealed lower screening prevalence in active smokers than in never smokers (89); thus, the importance of screening in those high-risk (i.e., active/longer-term regular smokers) subjects cannot be overemphasized. Furthermore, our finding suggests that smoking and glucose metabolism are interrelated with particular lifestyles in promoting CRC carcinogenesis, warranting future studies for biologic mechanisms of IR phenotypes/genotypes in smokers with different levels of lifestyles for CRC initiation and/or progression.

Our data on smoking were self-reported, thus subject to a possible misclassification bias, but a previous validation study (90) confirmed the high reliability of self-reported measures for active smoking assessment. In addition, our genetic instruments had relatively weak power in the causal inference testing, leading to variations of individual genetic markers on CRC risk. A 2-stage RSF may overfit the model with multiple tasks, requiring a replication study with an independent data set. Finally, we examined AA postmenopausal women, so the generalizability of our findings to other populations is limited.

Our study indicates that IR SNPs in combination with aging, prolonged lifetime exposure to endogenous estrogen, and a higher-fat diet, jointly with smoking, synergistically increased the risk of CRC, more substantially in women who had longer-term exposure to cigarette smoking. Our findings may improve CRC prediction ability among the medically underrepresented AA women and highlight the development of genetically focused preventive interventions (e.g., smoking cessation; encouraging CRC screening for longer-term smokers) for those women who are at high risk with particular risk genotypes and behavioral patterns.

CONFLICTS OF INTEREST

Guarantor of the article: Su Yon Jung, PhD.
Specific author contributions: S.Y.J.: designed the study, performed the genomic data QC. S.Y.J.: performed the statistical analysis and interpreted the data. S.Y.J.: secured funding for this project. All participated in the paper writing and editing. All authors have read and approved the submission of the manuscript.
Financial support: This study was supported by the National Institute of Nursing Research of the National Institutes of Health under Award Number K01NR017852.
Potential competing interests: None to report.
Consent for publication: This manuscript has not been published and is not being considered for publication elsewhere, in whole or in part, in any language.
Availability of data and materials: The data that support the findings of this study are available in accordance with policies developed by the NHLBI and WHI to protect sensitive participant information and approved by the Fred Hutchinson Cancer Research Center, which currently serves as the IRB of record for the WHI. Data requests may be made by emailing helpdesk@WHI.org.

Study Highlights

WHAT IS KNOWN

✓ Colorectal cancer (CRC) is the leading cause of cancer diagnosis and death in women in the United States and other westernized countries. African American (AA) women have the highest CRC incidence and mortality rates among all races/ethnic female groups. IR or glucose intolerance has been believed to be the major biologic mechanism of colorectal carcinogenesis owing to obesity. AA women are underrepresented in genome-wide studies for systemic regulation of IR and the association with CRC risk.

WHAT IS NEW HERE

✓ We established a prediction model on the basis of gene–environment interactions to generate risk profiles for CRC with the most influential genetic and lifestyle factors. We detected fasting glucose–specific single-nucleotide polymorphismss and lifestyles, including smoking, aging, prolonged lifetime exposure to endogenous estrogen, and high fat intake, as the most predictive markers for CRC risk among AA women. Our joint test for those risk genotypes and lifestyles with smoking revealed the synergistically increased CRC risk, more substantially in women with longer-term exposure to cigarette smoking.

ACKNOWLEDGEMENTS

Part of the data for this project was provided by the WHI program, which is funded by the National Heart, Lung, and Blood Institute, the National Institutes of Health, and the US Department of Health and Human Services through contracts HHSN268201100046C, HHSN268201100001C, HHSN268201100002C, HHSN268201100003C, HHSN268201100004C, and HHSN271201100004C. The data sets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap through dbGaP accession (phs000200.v11.p3). Program Office: National Heart, Lung, and Blood Institute, Bethesda, MD: Jacques Rossouw, Shari Ludlam, Dale Burwen, Joan McGowan, Leslie Ford, and Nancy Geller. Clinical Coordinating Center: Fred Hutchinson Cancer Research Center, Seattle, WA: Garnet Anderson, Ross Prentice, Andrea LaCroix, and Charles Kooperberg. Investigators and Academic Centers: JoAnn E. Manson, Brigham and Women's Hospital, Harvard Medical School, Boston, MA; Barbara V. Howard, MedStar Health Research Institute/Howard University, Washington, DC; Marcia L. Stefanick, Stanford Prevention Research Center, Stanford, CA; Rebecca Jackson, The Ohio State University, Columbus, OH; Cynthia A. Thomson, University of Arizona, Tucson/Phoenix, AZ; Jean Wactawski-Wende, University at Buffalo, Buffalo, NY; Marian Limacher, University of Florida, Gainesville/Jacksonville, FL; Robert Wallace, University of Iowa, Iowa City/Davenport, IA; Lewis Kuller, University of Pittsburgh, Pittsburgh, PA; and Sally Shumaker, Wake Forest University School of Medicine, Winston-Salem, NC.

REFERENCES

1. American Cancer Society. Cancer Fact and Figures 2021. American Cancer Society, Inc: Atlanta, 2021. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2021/cancer-facts-and-figures-2021.pdf.
2. American Cancer Society. Colorectal Cancer Facts & Figures 2020-2021. American Cancer Society, Inc: Atlanta, 2020. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/colorectal-cancer-facts-and-figures/colorectal-cancer-facts-and-figures-2020-2022.pdf.
3. Steele CB, Thomas CC, Henley SJ, et al. Vital signs: Trends in incidence of cancers associated with overweight and obesity - United States, 2005-2014. MMWR Morb Mortal Wkly Rep 2017;66(39):1052–8.
4. American Cancer Society. Cancer Fact and Figures for African Americans 2019-2021. American Cancer Society, Inc: Atlanta, 2021. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/cancer-facts-and-figures-for-african-americans/cancer-facts-and-figures-for-african-americans-2019-2021.pdf.
5. Ma Y, Yang Y, Wang F, et al. Obesity and risk of colorectal cancer: A systematic review of prospective studies. Plos One 2013;8(1):e53916.
6. Shokrani B, Brim H, Hydari T, et al. Analysis of beta-catenin association with obesity in African Americans with premalignant and malignant colorectal lesions. BMC Gastroenterol 2020;20(1):274.
7. Abdelsatir AA, Husain NE, Hassan AT, et al. Potential benefit of metformin as treatment for colon cancer: The evidence so far. Asian Pac J Cancer Prev 2015;16(18):8053–8.
8. Ho GY, Wang T, Gunter MJ, et al. Adipokines linking obesity with colorectal cancer risk in postmenopausal women. Cancer Res 2012;72(12):3029–37.
9. Bjork J, Nilsson J, Hultcrantz R, et al. Growth-regulatory effects of sensory neuropeptides, epidermal growth factor, insulin, and somatostatin on the non-transformed intestinal epithelial cell line IEC-6 and the colon cancer cell line HT 29. Scand J Gastroenterol 1993;28(10):879–84.
10. Tran TT, Medline A, Bruce WR. Insulin promotion of colon tumors in rats. Cancer Epidemiol Biomarkers Prev 1996;5(12):1013–5.
11. Tran TT, Naigamwalla D, Oprescu AI, et al. Hyperinsulinemia, but not other factors associated with insulin resistance, acutely enhances colorectal epithelial proliferation in vivo. Endocrinology 2006;147(4):1830–7.
12. Trevisan M, Liu J, Muti P, et al. Markers of insulin resistance and colorectal cancer mortality. Cancer Epidemiol Biomarkers Prev 2001;10(9):937–41.
13. Shin HY, Jung KJ, Linton JA, et al. Association between fasting serum glucose levels and incidence of colorectal cancer in Korean men: The Korean cancer prevention study-II. Metabolism 2014;63(10):1250–6.
14. Khaw KT, Wareham N, Bingham S, et al. Preliminary communication: Glycated hemoglobin, diabetes, and incident colorectal cancer in men and women: A prospective analysis from the European prospective

investigation into cancer-norfolk study. Am Soc Prev Oncol 2004;13(6): 915–9.

15. Hsu YC, Chiu HM, Liou JM, et al. Glycated hemoglobin A1c is superior to fasting plasma glucose as an independent risk factor for colorectal neoplasia. Cancer Causes Control 2012;23(2):321–8.
16. Sandhu MS, Dunger DB, Giovannucci EL. Insulin, insulin-like growth factor-I (IGF-I), IGF binding proteins, their biologic interactions, and colorectal cancer. J Natl Cancer Inst 2002;94(13):972–80.
17. Mulholland HG, Murray LJ, Cardwell CR, et al. Glycemic index, glycemic load, and risk of digestive tract neoplasms: A systematic review and meta-analysis. Am J Clin Nutr 2009;89(2):568–76.
18. Gallagher EJ, LeRoith D, Franco R, et al. Metabolic syndrome and pre-diabetes contribute to racial disparities in breast cancer outcomes: Hypothesis and proposed pathways. Diabetes Metab Res Rev 2016;32(7): 745–53.
19. Samson ME, Adams SA, Orekoya O, et al. Understanding the association of type 2 diabetes mellitus in breast cancer among African American and European American populations in South Carolina. J Racial Ethn Health Disparities 2016;3(3):546–54.
20. Ashktorab H, Daremipouran M, Goel A, et al. DNA methylome profiling identifies novel methylated genes in African American patients with colorectal neoplasia. Epigenetics 2014;9(4):503–12.
21. Weichhaus M, Broom J, Wahle K, et al. A novel role for insulin resistance in the connection between obesity and postmenopausal breast cancer. Int J Oncol 2012;41(2):745–52.
22. Liu J, Carnero-Montoro E, van Dongen J, et al. An integrative cross-omics analysis of DNA methylation sites of glucose and insulin homeostasis. Nat Commun 2019;10(1):2581.
23. Franks PW, Mesa JL, Harding AH, et al. Gene-lifestyle interaction on risk of type 2 diabetes. Nutr Metab Cardiovasc Dis 2007;17(2):104–24.
24. Arner P, Sahlqvist AS, Sinha I, et al. The epigenetic signature of systemic insulin resistance in obese women. Diabetologia 2016;59(11):2393–405.
25. Xu J, Ye Y, Wu H, et al. Association between markers of glucose metabolism and risk of colorectal cancer. BMJ Open 2016;6(6):e011430.
26. World Cancer Research Fund/American Association for Cancer Research. Recommendations for Cancer Prevention, 2015. http://www.aicr.org/reduce-your-cancer-risk/recommendations-for-cancer-prevention/. Accessed June 2021.
27. Nomura SJ, Dash C, Rosenberg L, et al. Is adherence to diet, physical activity, and body weight cancer prevention recommendations associated with colorectal cancer incidence in African American women? Cancer Causes Control 2016;27(7):869–79.
28. Dupuis J, Langenberg C, Prokopenko I, et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat Genet 2010;42(2):105–16.
29. Scott RA, Lagou V, Welch RP, et al. Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. Nat Genet 2012;44(9):991–1005.
30. Manning AK, Hivert MF, Scott RA, et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. Nat Genet 2012;44(6): 659–69.
31. Lagou V, Magi R, Hottenga JJ, et al. Sex-dimorphic genetic effects and novel loci for fasting glucose and insulin variability. Nat Commun 2021; 12(1):24.
32. Ramos E, Chen G, Shriner D, et al. Replication of genome-wide association studies (GWAS) loci for fasting plasma glucose in African-Americans. Diabetologia 2011;54(4):783–8.
33. Mondal AK, Sharma NK, Elbein SC, et al. Allelic expression imbalance screening of genes in chromosome 1q21-24 region to identify functional variants for Type 2 diabetes susceptibility. Physiol Genomics 2013;45(13): 509–20.
34. NCBI. WHI Harmonized and Imputed GWAS Data. A Sub-study of Women's Health Initiative, 2019. https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000746.v3.p3 Accessed January 1, 2021.
35. The Women's Health Initiative Study Group. Design of the women's health initiative clinical trial and observational study. Control Clin Trials. 1998;19(1):61–109.
36. NCBI. Women's Health Initiative - SNP Health Association Resource. A Sub-study of Women's Health Initiative, 2021. https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000386.v8.p3 Accessed January 1, 2021.
37. Schumacher FR, Al Olama AA, Berndt SI, et al. Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. Nat Genet 2018;50(7):928–36.
38. Akinyemiju T, Wiener H, Pisu M. Cancer-related risk factors and incidence of major cancers by race, gender and region; analysis of the NIH-AARP diet and health study. BMC Cancer 2017;17(1):597.
39. Dash C, Yu J, Nomura S, et al. Obesity is an initiator of colon adenomas but not a promoter of colorectal cancer in the Black Women's health study. Cancer Causes Control 2020;31(4):291–302.
40. Lai SM, Zhang KB, Uhler RJ, et al. Geographic variation in the incidence of colorectal cancer in the United States, 1998-2001. Cancer 2006;107(5 Suppl):1172–80.
41. Glenn BA, Hamilton AS, Nonzee NJ, et al. Obesity, physical activity, and dietary behaviors in an ethnically-diverse sample of cancer survivors with early onset disease. J Psychosoc Oncol 2018;36(4):418–36.
42. Ashktorab H, Paydar M, Yazdi S, et al. BMI and the risk of colorectal adenoma in African-Americans. Obesity (Silver Spring) 2014;22(5): 1387–91.
43. National Cancer Institute. SEER Program: Comparative Staging Guide for Cancer. 1993. https://seer.cancer.gov/archive/manuals/historic/comp_stage1.1.pdf Accessed January 1, 2021.
44. Burgess S, Thompson SG. Avoiding bias from weak instruments in Mendelian randomization studies. Int J Epidemiol 2011;40(3):755–64.
45. Pierce BL, Ahsan H, Vanderweele TJ. Power and instrument strength requirements for Mendelian randomization studies using multiple genetic variants. Int J Epidemiol 2011;40(3):740–52.
46. Shu X, Wu L, Khankari NK, et al. Associations of obesity and circulating insulin and glucose with breast cancer risk: A mendelian randomization analysis. Int J Epidemiol 2018;48:795–806.
47. Carreras-Torres R, Johansson M, Gaborieau V, et al. The role of obesity, type 2 diabetes, and metabolic factors in pancreatic cancer: A mendelian randomization study. J Natl Cancer Inst 2017;109(9):djx012.
48. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression. Int J Epidemiol 2015;44(2):512–25.
49. Bowden J, Del Greco MF, Minelli C, et al. A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. Stat Med 2017;36(11):1783–802.
50. Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. Genet Epidemiol 2013;37(7):658–65.
51. Qian F, Wang S, Mitchell J, et al. Height and body mass index as modifiers of breast cancer risk in BRCA1/2 mutation carriers: A mendelian randomization study. J Natl Cancer Inst 2019;111(4):350–64.
52. Bowden J, Davey Smith G, Haycock PC, et al. Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. Genet Epidemiol 2016;40(4):304–14.
53. Mogensen UB, Ishwaran H, Gerds TA. Evaluating random forests for survival analysis using prediction error curves. J Stat Softw 2012;50(11): 1–23.
54. Chung RH, Chen YE. A two-stage random forest-based pathway analysis methode36662. Plos One 2012;7(5):e36662.
55. Montazeri M, Beigzadeh A. Machine learning models in breast cancer survival prediction. Technology Health Care 2016;24(1):31–42.
56. Pang H, Lin A, Holford M, et al. Pathway analysis using random forests classification and regression. Bioinformatics 2006;22(16):2028–36.
57. Chang JS, Yeh RF, Wiencke JK, et al. Pathway analysis of single-nucleotide polymorphisms potentially associated with glioblastoma multiforme susceptibility using random forests. Am Soc Prev Oncol 2008; 17(6):1368–73.
58. Tong X, Feng Y, Li JJ. Neyman-Pearson classification algorithms and NP receiver operating characteristicseaao1659. Sci Adv 2018;4(2):eaao1659.
59. Ishwaran H, Kogalur UB. Random Survival Forests for R, 2007. https://pdfs.semanticscholar.org/951a/84f0176076fb6786fdf43320e8b27094dcfa.pdf.
60. Ishwaran H, Kogalur UB, Blackstone EH, et al. Random survival forests. Ann. Appl. Stat. 2008;2(3):841–60.
61. Inuzuka R, Diller GP, Borgia F, et al. Comprehensive use of cardiopulmonary exercise testing identifies adults with congenital heart disease at increased mortality risk in the medium term. Circulation 2012; 125(2):250–9.
62. GCK Glucokinase [Homo Sapiens (Human)]—Gene—NCBI. https://www.ncbi.nlm.nih.gov/gene/2645. 2021 Accessed January 1, 2021.

63. Gene Card: Human Gene Database: GCK Gene (Protein Coding). https://www.genecards.org/cgi-bin/carddisp.pl?gene=GCK. 2021 Accessed January 1, 2021.
64. An P, Miljkovic I, Thyagarajan B, et al. Genome-wide association study identifies common loci influencing circulating glycated hemoglobin (HbA1c) levels in non-diabetic subjects: The long life family study (LLFS). Metabolism 2014;63(4):461–8.
65. Pare G, Chasman DI, Parker AN, et al. Novel association of HK1 with glycated hemoglobin in a non-diabetic population: A genome-wide evaluation of 14,618 participants in the Women's genome health study. Plos Genet 2008;4(12):e1000312.
66. Leak TS, Langefeld CD, Keene KL, et al. Chromosome 7p linkage and association study for diabetes related traits and type 2 diabetes in an African-American population enriched for nephropathy. BMC Med Genet 2010;11:22.
67. Tong X, Zhao F, Thompson CB. The molecular determinants of de novo nucleotide biosynthesis in cancer cells. Curr Opin Genet Dev 2009;19(1):32–7.
68. Niizeki H, Kobayashi M, Horiuchi I, et al. Hypoxia enhances the expression of autocrine motility factor and the motility of human pancreatic cancer cells. Br J Cancer 2002;86(12):1914–9.
69. Dong X, Tang H, Hess KR, et al. Glucose metabolism gene polymorphisms and clinical outcome in pancreatic cancer. Cancer 2011;117(3):480–91.
70. Ramos-Molina B, Martin MG, Lindberg I. PCSK1 variants and human obesity. Prog Mol Biol Transl Sci 2016;140:47–74 Accessed January 1, 2021.
71. PCSK1 Proprotein Convertase Subtilisin/Kexin Type 1 [Homo Sapiens (Human)]—Gene—NCBI. https://www.ncbi.nlm.nih.gov/gene/5122. Accessed January 1, 2021.
72. Kaufmann JE, Irminger JC, Mungall J, et al. Proinsulin conversion in GH3 cells after coexpression of human proinsulin with the endoproteases PC2 and/or PC3. Diabetes 1997;46(6):978–82.
73. Bailyes EM, Shennan KI, Seal AJ, et al. A member of the eukaryotic subtilisin family (PC3) has the enzymic properties of the type 1 proinsulin-converting endopeptidase. Biochem J 1992;285(Pt 2):391–4.
74. Stijnen P, Ramos-Molina B, O'Rahilly S, et al. PCSK1 mutations and human Endocrinopathies: From obesity to gastrointestinal disorders. Endocr Rev 2016;37(4):347–71.
75. Tzimas GN, Chevet E, Jenna S, et al. Abnormal expression and processing of the proprotein convertases PC1 and PC2 in human colorectal liver metastases. BMC Cancer 2005;5:149.
76. McMullan CJ, Schernhammer ES, Rimm EB, et al. Melatonin secretion and the incidence of type 2 diabetes. JAMA 2013;309(13):1388–96.
77. Stumpf I, Muhlbauer E, Peschke E. Involvement of the cGMP pathway in mediating the insulin-inhibitory effect of melatonin in pancreatic beta-cells. J Pineal Res 2008;45(3):318–27.
78. Hu C, Jia W. Linking MTNR1B variants to diabetes: The role of circadian rhythms. Diabetes 2016;65(6):1490–2.
79. Pandi-Perumal SR, Trakht I, Srinivasan V, et al. Physiological effects of melatonin: Role of melatonin receptors and signal transduction pathways. Prog Neurobiol 2008;85(3):335–53.
80. Leon J, Casado J, Carazo A, et al. Gender-related invasion differences associated with mRNA expression levels of melatonin membrane receptors in colorectal cancer. Mol Carcinog 2012;51(8):608–18.
81. Deming SL, Lu W, Beeghly-Fadiel A, et al. Melatonin pathway genes and breast cancer risk among Chinese women. Breast Cancer Res Treat 2012;132(2):693–9.
82. Hartz A, He T, Ross JJ. Risk factors for colon cancer in 150,912 postmenopausal women. Cancer Causes Control 2012;23(10):1599–605.
83. Zervoudakis A, Strickler HD, Park Y, et al. Reproductive history and risk of colorectal cancer in postmenopausal women. J Natl Cancer Inst 2011;103(10):826–34.
84. Talamini R, Franceschi S, Dal Maso L, et al. The influence of reproductive and hormonal factors on the risk of colon and rectal cancer in women. Eur J Cancer 1998;34(7):1070–6.
85. Yoo KY, Tajima K, Inoue M, et al. Reproductive factors related to the risk of colorectal cancer by subsite: A case-control analysis. Br J Cancer 1999;79(11-12):1901–6.
86. Yamasaki E, Ames BN. Concentration of mutagens from urine by absorption with the nonpolar resin XAD-2: Cigarette smokers have mutagenic urine. Proc Natl Acad Sci USA 1977;74(8):3555–9.
87. Park SL, Carmella SG, Ming X, et al. Variation in levels of the lung carcinogen NNAL and its glucuronides in the urine of cigarette smokers from five ethnic groups with differing risks for lung cancer. Cancer Epidemiol Biomarkers Prev 2015;24(3):561–9.
88. Terry PD, Miller AB, Rohan TE. Cigarette smoking and breast cancer risk: A long latency period?. Int J Cancer 2002;100(6):723–8.
89. Oluyemi AO, Welch AR, Yoo LJ, et al. Colorectal cancer screening in high-risk groups is increasing, although current smokers fall behind. Cancer 2014;120(14):2106–13.
90. Soulakova JN, Hartman AM, Liu B, et al. Reliability of adult self-reported smoking history: Data from the tobacco use supplement to the current population survey 2002-2003 cohort. Nicotine Tob Res 2012;14(8):952–60.