

Post Genome-Wide Gene–Environment Interaction Study Using Random Survival Forest: Insulin Resistance, Lifestyle Factors, and Colorectal Cancer Risk

Su Yon Jung¹, Jeanette C. Papp², Eric M. Sobel², and Zuo-Feng Zhang³



Abstract

Molecular and genetic pathways of insulin resistance (IR) connecting colorectal cancer and obesity factors in postmenopausal women remain inconclusive. We examined the IR pathways on both genetic and phenotypic perspectives at the genome-wide level. We further constructed colorectal cancer risk profiles with the most predictive IR SNPs and lifestyle factors. In our earlier genome-wide association gene–environmental interaction study, we used data from a large cohort of postmenopausal women in the Women's Health Initiative Database for Genotypes and Phenotypes Study and identified 58 SNPs in relation to IR phenotypes. In this study, we evaluated the identified IR SNPs and selected 34 lifestyles for their association with colorectal cancer risk in a total of 11,078 women (including 736 women with colorectal cancer) using a 2-stage multimodal random survival forest analysis. In overall and sub-

group (defined via body mass index, exercise, and dietary-fat intake) analyses, we identified 2 SNPs (*LINC00460* rs1725459 and *MTRR* rs722025) and lifetime cumulative exposure to estrogen (oral contraceptive use) and cigarette smoking as the most common and strongest predictive markers for colorectal cancer risk across the analyses. The combinations of genetic and lifestyle factors had much greater impact on colorectal cancer risk than any individual risk factors, and a possible synergism existed to increase colorectal cancer risk in a gene-behavior dose-dependent manner. Our findings may inform research on the role of IR in the etiology of colorectal cancer and contribute to more accurate prediction of colorectal cancer risk, suggesting potential intervention strategies for women with specific genotypes and lifestyles to reduce their colorectal cancer risk.

Introduction

Colorectal cancer ranks third among women for both cancer incidence and mortality in the United States and other westernized countries (1, 2), and the majority (about 90%) of new cases and deaths occur in women ages 50 years and older (3). Of nonmodifiable and modifiable environmental factors that together account for more than

60% of colorectal cancer susceptibility (4), obesity (both overall and central obesity), and obesity-related behavioral factors such as physical inactivity and unbalanced diet have been considered risk factors (5–7).

Insulin is a potential mediator underlying the biologic mechanism by explaining 40% of the association between obesity and colorectal cancer (8). A recent *in vivo* study (9) reported that the elevated circulating levels of insulin and glucose, reflecting insulin resistance (IR), increased colorectal epithelial proliferation in a dose-dependent manner, suggesting the molecular IR pathways connecting to colorectal cancer. Furthermore, obesity and IR, influencing mutually, lead to hyperglycemia and compensatory hyperinsulinemia and have been associated with increased risk for postmenopausal colorectal cancer (10, 11).

Thus, the potential existence of pathways between IR, colorectal cancer, and obesity on molecular-genetic perspectives are convincingly presumed. Particularly, our previous study (12) revealed that genetic variants (SNPs) in relation to IR phenotypes were associated with greater increases in IR among obese, physically inactive, and high dietary-fat groups, indicating the role of obesity and

¹Translational Sciences Section, Jonsson Comprehensive Cancer Center, School of Nursing, University of California, Los Angeles, Los Angeles, California.

²Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, California. ³Department of Epidemiology, Fielding School of Public Health, University of California, Los Angeles, Los Angeles, California.

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

Corresponding Author: Su Yon Jung, University of California Los Angeles, 700 Tiverton Ave, 3-264 Factor Building, Los Angeles, CA 90095. Phone: 310-825-2840; Fax: 310-267-0413; E-mail: sjung@sonnet.ucla.edu

Cancer Prev Res 2019;12:877–90

doi: 10.1158/1940-6207.CAPR-19-0278

©2019 American Association for Cancer Research.

obesity-related lifestyle factors as an effect modifier in the pathway between IR SNPs and phenotypes (Supplementary Fig. S1). Furthermore, the effect of IR SNPs on colorectal cancer risk through an IR gene–phenotype pathway can be modified by obesity. Therefore, IR (genotype and phenotype) and obesity may conjunctionally influence the risk of colorectal cancer (Supplementary Fig. S1, yellow lines).

Understanding how obesity and related lifestyles interact with IR genes and phenotypes and modify the IR pathway, influencing the risk of colorectal cancer, is important to develop a gene–lifestyle preventive tool in primary cancer prevention. However, no published reports at the genome-wide level have examined the IR pathway connecting to colorectal cancer risk by incorporating obesity factors. In addition, published studies generating risk profiles for colorectal cancer with both genetic and lifestyle factors are scarce.

We hoped to address these gaps in this study. As the first step, by using data from postmenopausal women of the Women's Health Initiative Database for Genotypes and Phenotypes (WHI dbGaP) Study, we previously conducted a genome-wide association (GWA) gene–environmental (i.e., behavioral) interaction ($G \times E$) study. We identified SNPs associated with IR phenotypes [homeostatic model assessment–IR (HOMA-IR), hyperglycemia, and hyperinsulinemia] by testing for interactions with obesity and obesity-related lifestyles (13). By performing a stratification analysis, we identified 58 SNPs that had genome-wide significance in women stratified by obesity (4 SNPs), physical activity (36 SNPs), and dietary-fat intake (18 SNPs).

In the current study, as the second step, we evaluated whether those 58 SNPs were associated with the risk of colorectal cancer in the identically stratified subgroups (i.e., obese/exercise/dietary-fat subgroups) in which those SNPs had been found to be associated with IR in the earlier GWA $G \times E$ study (13). This approach would allow us to test our hypothetical pathways, in which IR genes and phenotypes identified through interactions with obesity pathways are associated with colorectal cancer risk (Supplementary Fig. S1), and thus improve our understanding of the etiology of colorectal cancer.

Moreover, in addition to obesity and related lifestyle factors, we further selected 31 nonmodifiable and modifiable lifestyle factors for use in this study in constructing risk profiles with the SNPs and lifestyle factors in relation to colorectal cancer by performing 2-stage random survival forest (RSF) analyses. The RSF is a machine-learning, nonparametric tree-based ensemble method, which can deal with the nonlinear effects of variables (that are not handled in a traditional regression model) and evaluate high-order interactions among variables; thus, the RSF may successfully yield accurate colorectal cancer–risk profile predictions (14, 15).

By using the most predictive SNPs and lifestyle factors identified via the two-stage RSF, we constructed prediction models for colorectal cancer risk. We further estimated the combined and joint effect of predictive variables on colorectal cancer risk by performing a regression analysis. By employing the two complementary statistical approaches, we finally tested whether the most predictive genetic and lifestyle factors in combination predict the risk for colorectal cancer in a dose–response fashion.

Materials and Methods

Study population

We used data from postmenopausal women in the WHI dbGaP, the Harmonized and Imputed GWA Studies, under dbGaP study accession phs000200.v11.p3, which came from a joint imputation and harmonization effort for the GWA study within the WHI 2 representative study arms Clinical Trials and Observational Studies. The detailed studies' design and rationale have been described elsewhere (16, 17). Briefly, the WHI is a long-term prospective cohort study that has focused on strategies for preventing chronic diseases, including breast cancer and colorectal cancer, in postmenopausal women. Healthy postmenopausal women had been enrolled in the WHI study between 1993 and 1998 from more than 40 clinical centers across the United States; the women were 50 to 79 years old, expected to live in close proximity to the clinical centers for at least 3 years after enrollment, and able to provide written consent. Women enrolled in the WHI study were eligible for the WHI dbGaP study if they had met eligibility requirements for submission to dbGaP and provided DNA samples. The Harmonized and Imputed GWA Studies consist of 6 sub-studies (MOPMAP[AS264]; GARNET; GECCO-CYTO; GECCO-INIT; HIPFX; and WHIMS); in them, we initially identified 16,088 women who reported their race or ethnicity as non-Hispanic white (Supplementary Fig. S2). We applied the exclusion criteria to our earlier GWA $G \times E$ study and excluded (i) women ($n = 2,714$) diagnosed with diabetes at or after enrollment and (ii) women ($n = 1,580$) whose genetic data were related to others (kinship estimate > 0.25) and/or outliers based on principal components. In the current study, we additionally excluded 716 women with less than 1 year follow-up period and/or a diagnosis of any types of cancer at enrollment. Thus, a total of 11,078 women (including 736 women with colorectal cancer), who had been followed up through August 29, 2014, with 16-year median follow-up period, were finally analyzed in this study. The Institutional Review Boards of each participating WHI clinical center and the University of California, Los Angeles (Los Angeles, CA), have approved this study.

Data collection and colorectal cancer outcome

The coordinating clinical centers had collected data from participants' self-administered questionnaires via a

uniform written protocol and performed data quality assurance. Through the questionnaires at enrollment, participants provided demographic, socioeconomic, and lifestyle factors as well as family, medical, and reproductive histories. In this study, we initially pulled out all available variables; on the basis of their association with IR and colorectal cancer through the literature review (3, 18) and the initial analysis process including univariate and stepwise multiple regression analyses and a multicollinearity test, we selected 34 variables to evaluate in this study. In detail, we evaluated demographic (age, education, and marital status) and socioeconomic factors (family income and employment), family histories of colorectal cancer and diabetes, and medical (depressive symptoms, hypertension, high cholesterol, and cardiovascular disease) and reproductive histories (ages at menarche and menopause, number of pregnancies, months of breastfeeding, hysterectomy, one or both ovaries removed, and durations of past oral contraceptive (OC), unopposed estrogen [exogenous estrogen (E) only], and opposed estrogen [E plus progestin (P)] use). We further examined lifestyle factors including physical activity, cigarettes smoked per day, and diet per day [dietary intake of alcohol, fiber, total sugars, fruits, and vegetables; and percentage of calories from protein, saturated fatty acids (SFA), monounsaturated fatty acids, and polyunsaturated fatty acids]. We also included anthropometric variables, including height, weight, and waist and hip circumferences, which had been assessed by trained staff.

The colorectal cancer outcomes were confirmed by a centralized review of medical charts. Cancer sites were coded corresponding to the NCI's Surveillance, Epidemiology, and End-Results guidelines (19). The colorectal cancer variables were defined as (i) cancer development (yes/no) and (ii) the time to develop the cancer, estimated as the time in days between enrollment and colorectal cancer development, censoring, or study endpoint, then, computed as years.

Genotyping and laboratory methods

We obtained the genotyped data from the WHI Harmonized and Imputed GWA Studies. The genotype calls were normalized to the reference panel GRCh37, and genotype imputation was performed via 1,000 Genomes Project reference panel. SNPs for harmonization were checked for pairwise concordance among all samples across the GWA studies (17). The detailed genetic data-quality cleaning (QC) process has been described previously (13). In the initial QC process, SNPs with a missing-call rate of <3% and a Hardy-Weinberg equilibrium of $P \geq 10^{-4}$ were included. We further performed the secondary QC and included SNPs with $\hat{R}^2 \geq 0.6$ imputation quality (20), but excluded women with a kinship estimate of $\hat{R}^2 \geq 0.25$ to minimize possible confounding effect from shared environment.

We also obtained information from fasting blood samples in the WHI dbGaP Study, which had been extracted by trained phlebotomists from each woman at enrollment. The glucose serum concentrations were analyzed by the hexokinase method on a Hitachi 747 instrument (Boehringer Mannheim Diagnostics) and the insulin levels by radioimmunoassay (Linco Research, Inc.), with average coefficients of variation of 1.28% and 10.93%, respectively. The HOMA-IR levels were estimated as glucose (unit: mg/dL) \times insulin (unit: μ IU/mL)/405 (21).

Statistical analysis

We examined the distributions of participants' characteristics by colorectal cancer status by using unpaired 2-sample *t* tests (for continuous variables) and χ^2 tests (for categorical variables). If continuous variables were skewed or had outliers, Wilcoxon rank-sum test was conducted. In our previous GWA study, we had tested for the gene-environment interaction in the strata by body mass index (BMI: cutoff, 30 kg/m²), metabolic equivalents (METs)·hours/week (cutoff, 10 METs), and percentage of calories from SFA (cutoff, 7%). The results (either $G \times E$ formal test and stratified analysis) from the sub-GWA studies were combined in a meta-analysis assuming a fixed-effect model. In the current study, we evaluated the SNPs identified in the particular behavioral setting of obesity/physical activity/dietary fat intake in relation to colorectal cancer risk in the identical behavioral setting.

In this study, we performed the RSF analysis. The RSF first generates bootstrap samples using approximately 63% of the original data and then grows a tree from each bootstrap sample via a splitting rule, with which a tree node maximizes survival differences across daughter nodes. This tree-building process is repeated numerous times (5,000 times in this study) to create ultimately a forest of trees (22, 23). Next, an ensemble cumulative hazard estimate was calculated from each tree and averaged over all trees for each individual; using this ensemble estimate, we estimated a predicted cumulative colorectal cancer incidence rate. Furthermore, by using this ensemble estimate and creating the out-of-bag (OOB) data (on average, the 37% of the original data not used for bootstrapping), the OOB ensemble cumulative hazard estimate was calculated to compute the prediction parameter (i.e., prediction error interpreted as a misclassification probability). Finally, the OOB concordance index (*c*-index) was estimated from the formula (*c*-index = 1 - prediction error), which is a measure of prediction performance conceptually similar to the area under the receiver operating characteristic (AUROC) curve (22, 24).

The rank of each variable according to its predictability of developing colorectal cancer was determined by 2 predictive parameters: (i) minimal depth (MD), in which variables having a small MD value split the tree close to the root are considered highly predictive and (ii) variable

importance (VIMP), estimated from the difference between the OOB c-indexes from the original OOB data and from the permuted OOB data, in which variables having greater VIMP values are the more predictive (14).

We performed a 2-stage RSF analysis. In the first stage, we evaluated SNPs using an RSF for their association with colorectal cancer risk by incorporating obesity (Supplementary Fig. S3B–S3F). We also examined the lifestyle factors separately in relation to colorectal cancer risk (Supplementary Fig. S3A). With only the SNPs and lifestyle factors that had significantly low MD and high VIMP values, we further conducted the second stage of RSF to generate risk profiles for colorectal cancer that account for both IR-genetic and lifestyle factors. In the stage 2, we took a multimodal approach. In detail, in the overall and stratified subgroups (defined by physical activity and SFA intake), we (i) estimated the two MD and VIMP parameters and compared the two measures in the plot (Fig. 1A; Supplementary Figs. S4A1.B1 and S5A1.B1); (ii) generated the OOB c-index from the nested RSF model (Fig. 1B; Supplementary Figs. S4.A2.B2 and S5.A2.B2); and (iii) estimated the incremental error rate of each variable in the nested sequence of RSF models, beginning with the top variable, and calculated a dropping error rate as the difference between the error rates from the nested sequence models. The 2-stage RSF and multimodal approaches (Supplementary Fig. S6) allowed us to remove the SNPs and lifestyle factors that were not significantly associated with colorectal cancer risk, leading to greater statistical power with the correct type I error rate than the power we obtained with the original RSF-based analysis (23).

To obtain the HRs and 95% confidence intervals for the single and combined effects of SNPs and behavioral factors on colorectal cancer risk, we performed multiple Cox proportional hazards regression while checking assumptions via a Schoenfeld residual plot and rho. The regression analyses were adjusted for potential confounding factors listed in Table 1. We considered a 2-tailed $P < 0.05$ statistically significant. A multiple-comparison adjustment by using the Benjamini–Hochberg method (25) was conducted. We used R version 3.5.1 with several packages, including survival, survivalROC, randomForestSRC, ggRandomForests, and gamlss.

Results

The allele frequencies of the 58 SNPs identified at genome-wide significance in our earlier study are presented in Supplementary Table S1. The distributions of participants' baseline characteristics by colorectal cancer status (Table 1) reflected that patients with colorectal cancer were relatively younger, more highly educated, taller, heavier smokers (≥ 15 cigarettes/day), and more depressed than patients without colorectal cancer. Women with colorectal cancer were also likely to have shorter

durations of breastfeeding and past OC use before menopause but to have higher frequencies and longer durations of E-only and E + P use after menopause.

Two-stage RSF and multimodal approach to determine the most predictive SNPs and behavioral factors for colorectal cancer risk

To identify the most influential variables with the highest predictability and lowest prediction errors for colorectal cancer risk, we conducted a 2-stage RSF with a multimodal approach using the 2 measures MD and VIMP. These 2 methods use different prediction algorithms; thus, having variables with somewhat different ranking is expected. In the first-stage RSF, we created a plot (Supplementary Fig. S3) to compare the 2 measures for each SNP and behavioral factor. Given that SNPs and behavioral variables in agreement with high ranks in both MD and VIMP are the strongest predictive markers for colorectal cancer risk, we selected 13 of the 34 behavioral factors (Supplementary Fig. S3A); 18 of the 58 SNPs in overall analysis (Supplementary Fig. S3B); 9 (Supplementary Fig. S3C) and 11 (Supplementary Fig. S3D) of the 36 SNPs in METs ≥ 10 and < 10 , respectively; and 2 (Supplementary Fig. S3E) and 5 (Supplementary Fig. S3F) of the 18 SNPs in calories from SFA $< 7.0\%$ and $\geq 7.0\%$, respectively.

With the 13 behavioral factors and selected SNPs together, in overall- and subgroups, we next performed the second-stage multimodal RSF to construct risk profiles with the most predictive factors. In the overall analysis of the total population, we initially computed the 2 measures MD and VIMP (Table 2) and plotted them for comparison (Fig. 1A); the dashed red line reflects where the 2 measures were in agreement. By selecting variables with high ranks in both measures, we determined that 2 SNPs (*LINC00460* rs17254590 and *MTRR* rs722025) and 1 behavioral factor (OC use) were the strongest predictive markers of colorectal cancer risk. Second, we generated the OOB c-index (conceptually similar to the AUROC) within the nested RSF model (Table 2). In the plot (Fig. 1B) where variables were arranged by MD (low to high values), we identified the same top 3 variables (2 SNPs and 1 behavioral factor) as those identified in Fig. 1A. These 3 variables improved the OOB c-index, whereas the others did not substantially improve the prediction accuracy, suggesting that the OOB c-index has complementary predictive value. Finally, we computed a dropping error rate for each variable in the nested sequence of RSF models (Table 2) and determined that once again the same 3 variables (those identified with the aforementioned 2 strategies) contributed the most to decreasing the error rate, thus improving the prediction accuracy.

For each subgroup analysis, we continuously applied the 3 approaches (agreement between MD and VIMP; OOB c-index; and contribution to dropping error rate) and determined the most predictive variables as follows:

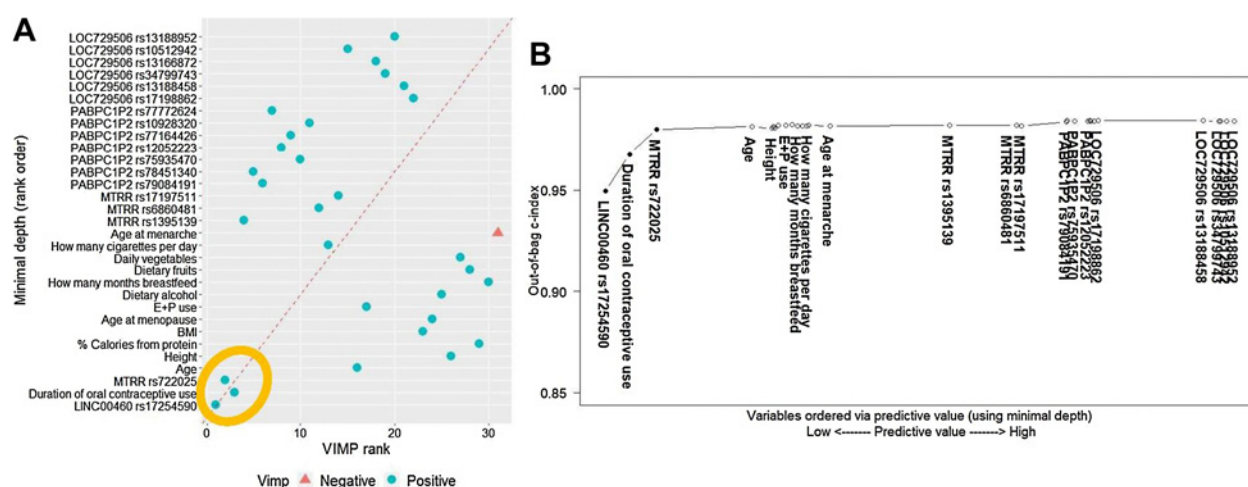


Figure 1.

Overall analysis: the second stage of RSF with 18 SNPs and 13 behavioral factors selected from the first stage of RSF analysis. **A**, Comparing minimal depth and VIMP rankings (E + P, exogenous estrogen + progestin; VIMP, variable of importance). Note: 3 variables within the orange ellipse were identified as the most influential predictors. **B**, Out-of-bag c-index. Improvement in out-of-bag c-index was observed when the top 3 variables [●] were added to the model, whereas other variables [○] did not further improve the accuracy of prediction.

(i) in the active group (≥ 10 METs; Supplementary Table S2A; Supplementary Fig. S4.A1.A2), 2 SNPs (*MTRR* rs722025 and *MKLN1* rs117911989) and 3 lifestyle factors (OC use, age, and cigarette smoking); (ii) in the inactive group (< 10 METs; Supplementary Table S2B; Supplementary Fig. S4.B1.B2), 1 SNP (*MTRR* rs722025) and 3 lifestyle factors (OC use, cigarette smoking, and E + P use); (iii) in the low fat-intake group ($< 7.0\%$ calories from SFA; Supplementary Table S3A; Supplementary Fig. S5.A1.A2), 1 SNP (*LINC00460* rs17254590) and 2 lifestyle factors (OC use and age); and 4) in the high fat-intake group ($\geq 7.0\%$ calories from SFA; Supplementary Table S3B; Supplementary Fig. S5.B1.B2), 2 SNPs (*LINC00460* rs17254590 and *PABPC1P2* rs10928320) and 1 lifestyle factor (OC use).

Combined and joint effects of the most predictive SNPs and behavioral factors on colorectal cancer risk

We estimated the cumulative colorectal cancer incidence rate for each predictive variable by accounting for its nonlinear effect using RSF (Fig. 2). The genotypes of SNPs were evaluated as continuous variables. On the basis of Fig. 2A–D, we considered *PABPC1P2* rs10928320 CC, *MTRR* rs722025 GA+AA, *MKLN1* rs117911989 GG, and *LINC00460* rs17254590 GG to be risk genotypes, targeted for further analysis as categorical variables. In addition, using a cut-off value bisecting variables (Fig. 2E–H), we defined the high-risk lifestyle groups as those with < 5 years of past OC use, a history of E + P use, smoking ≥ 15 cigarettes/day, and age older than 60 years and analyzed them as binary variables.

In the overall analysis, with the top 3 most influential variables, we developed a multivariate model predicting

colorectal cancer risk (Table 3), indicating that the individual SNPs had a stronger effect than the individual behavioral factors on colorectal cancer risk even after adjusting for confounding factors. A similar trend was observed in the physical activity- and SFA-subgroup multivariate analyses (Supplementary Table S4.A.B); in particular, single risk-genotypes had > 5.0 HRs while single risk-behavioral factors had ≤ 3 HRs.

However, the combinations of SNPs and lifestyle factors yielded different results (Tables 4 and 5; Supplementary Table S5). For example, in the active subgroup (Table 4), 2 SNPs (*MTRR* rs722025 and *MKLN1* rs117911989) were combined and stratified by cigarette smoking. Heavier smokers (≥ 15 /day) with the 2 risk genotypes had an almost 10 times higher risk of colorectal cancer than less heavy smokers (< 15 /day) with null-risk genotypes; and their (the heavy smokers with combined risk genotypes) risk was much greater than the risk of those with any single risk-genotypes (Supplementary Table S4A). Consistently, the high-risk lifestyle group (with 2 lifestyle factors, such as OC use and age) of heavier smokers had a 7-times higher risk than the null-risk lifestyle group of less heavy smokers; their (the heavy smokers with 2 risk-lifestyle factors) risk was also higher than that of those with any single risk-lifestyles (Supplementary Table S4A). When the 2 SNPs and the 3 lifestyle factors were combined, the high-risk group (with 2 risk genotypes and 3 risk behaviors) had 32 times the excess risk for colorectal cancer than the low-risk group (with ≤ 1 risk genotype and ≤ 2 risk behaviors), suggesting a cumulative effect of genetic and lifestyle factors in an additive interaction model. Multiple testing was corrected to control the false-discovery rate. When stratified by smoking, heavier smokers with high risk of

Table 1. Characteristics of participants

Characteristics	Participants without CRC	Participants with CRC
	(n = 10,342) n (%)	(n = 736) n (%)
Age in years, median (range)	67 (50–81)	66 (50–79) ^a
Education		
≤High school	3,708 (35.9)	222 (30.2) ^a
>High school	6,634 (64.1)	514 (69.8)
Family income		
<\$35,000	4,600 (44.5)	312 (42.4)
≥\$35,000	5,742 (55.5)	424 (57.6)
METs·hour·week ⁻¹ , median (range) ^b	7.25 (0.00–134.17)	7.79 (0.00–79.00)
METs·hour·week ⁻¹ ^b		
≥10.0	4,322 (41.8)	321 (43.6)
<10.0	6,020 (58.2)	415 (56.4)
How many cigarettes per day		
<15 ^c	8,034 (77.7)	548 (74.5) ^a
≥15	2,308 (22.3)	188 (25.5)
BMI in kg/m ² , median (range)	26.94 (15.42–58.49)	26.72 (17.33–55.62)
BMI ^d		
<30.0	7,305 (70.6)	538 (73.1)
≥30.0	3,037 (29.4)	198 (26.9)
Waist-to-hip ratio, median (range)	0.807 (0.444–1.282)	0.808 (0.604–1.393)
Height in cm, median (range)	161.8 (146.2–177.0)	162.4 (146.2–177.0) ^a
Depressive symptom ^e , median (range)		
<0.06	9,574 (92.6)	665 (90.4) ^a
≥0.06	768 (7.4)	71 (9.6)
Cardiovascular disease ever		
No	8,804 (85.1)	608 (82.6)
Yes	1,538 (14.9)	128 (17.4)
Hypertension ever		
No	7,167 (69.3)	492 (66.8)
Yes	3,175 (30.7)	244 (33.2)
Family history of diabetes		
No	7,467 (72.2)	528 (71.7)
Yes	2,875 (27.8)	208 (28.3)
Family history of colorectal cancer		
No	8,704 (84.2)	605 (82.2)
Yes	1,638 (15.8)	131 (17.8)
Dietary alcohol in g/day ^f		
<6.1	7,218 (69.8)	506 (68.8)
≥6.1	3,124 (30.2)	230 (31.2)
Daily fruits in med portion, median (range)	1.74 (0.00–8.72)	1.91 (0.03–7.00)
Daily vegetables in med portion, median (range)	2.03 (0.03–11.71)	2.08 (0.06–11.49)
% Calories from SFA, median (range)	11.30 (2.22–32.39)	11.38 (2.60–26.77)
% calories from SFA ^g		
<7.0	937 (9.1)	72 (9.8)
≥7.0	9,405 (90.9)	664 (90.2)
% Calories from MFA, median (range)	12.79 (2.16–27.64)	12.78 (2.82–23.04)
% Calories from PFA, median (range)	6.61 (1.19–21.77)	6.63 (1.63–19.30)
% Calories from protein, median (range)	16.59 (5.48–35.97)	16.42 (7.28–26.92)
Dietary total sugars in g, median (range)	93.15 (4.59–525.92)	95.99 (13.75–267.06)
Hysterectomy ever		
No	6,682 (64.6)	456 (62.0)
Yes	3,660 (35.4)	280 (38.0)
One or both ovaries removed		
No	7,877 (76.2)	549 (74.6)
Part of an ovary taken out	87 (0.8)	6 (0.8)
One taken out	728 (7.0)	57 (7.7)
Both taken out	1,559 (15.1)	118 (16.0)
Unknown number taken out	91 (0.9)	6 (0.8)
Age at menarche in years, median (range)	13 (≤9–≥17)	13 (≤9–≥17)
Age at menopause in years, median (range)	50 (20–63)	50 (22–60)
Total months of breastfeeding ^g		
≤12	8,621 (83.4)	643 (87.4) ^a
>12	1,721 (16.6)	93 (12.6)
Oral contraceptive duration in years, median (range)	7.6 (0.1–47.0)	5.0 (0.1–21.0) ^a

(Continued on the following page)

Table 1. Characteristics of participants (Cont'd)

Characteristics	Participants without CRC	Participants with CRC
	(n = 10,342) n (%)	(n = 736) n (%)
E-only in years		
Never	7,295 (70.5)	489 (66.4) ^a
<5	1,442 (13.9)	105 (14.3)
5-<10	506 (4.9)	54 (7.3)
≥10	1,099 (10.6)	88 (12.0)
E + P in years		
Never	8,533 (82.5)	572 (77.7) ^a
<5	992 (9.6)	85 (11.5)
5-<10	428 (4.1)	38 (5.2)
10-<15	243 (2.3)	23 (3.1)
≥15	146 (1.4)	18 (2.4)

Abbreviations: CRC, colorectal cancer; E, exogenous estrogen; E + P, E + progestin; MET, metabolic equivalent; MFA, monounsaturated fatty acids; PFA, polyunsaturated fatty acids.

^a $P < 0.05$, χ^2 or Wilcoxon rank-sum test.

^bPhysical activity was estimated from recreational physical activity combining walking and mild, moderate, and strenuous physical activity. Each activity was assigned a MET value corresponding to intensity; the total MET-hours-week⁻¹ was calculated by multiplying the MET level for the activity by the hours exercised per week and summing the values for all activities. The total MET was stratified using 10 METs as the cutoff according to current American College of Sports Medicine and American Heart Association recommendations (49).

^cThe women with smoking <15 cigarettes/day included nonsmokers.

^dBMI variable was categorized using 30 kg/m², where 30.0 or higher falls within the obese range (<https://www.cdc.gov/obesity/adult/defining.html>).

^eDepression scales were estimated using a short form of the Center for Epidemiologic Studies Depression Scale.

^fDietary alcohol per day and total months of breastfeeding were stratified using the mean values of 6.1 g/day and 12 months, respectively, as the cut-off points.

^g% Calories from SFA was stratified using 7% as the cut-off value, which adheres to the American Heart Association/American College of Cardiology dietary guidelines, which are aligned with the 2015–2020 Dietary Guidelines for Americans to help cardiovascular and metabolic diseases reductions (50).

both genotypes and behavioral factors had a 40-times higher risk than less heavy smokers with low risk of both genotypes and behavioral factors (Table 4). This suggests a gene–lifestyle dose–response relationship, and further, a potential joint effect of smoking with genetic and lifestyle factors on colorectal cancer risk in both additive and multiplicative models (effect size for $G \times E = 1.00$ and $p 0.993$). The results in Table 4, after being adjusted for the years of regular smoking (excluding the time the participants stayed off cigarettes), were consistent. The analyses of the inactive group yielded similar results but with a less strong impact of gene-lifestyle combinations on colorectal cancer risk.

Comparable results from the SFA-stratified analyses were observed (Table 5). Particularly, in the low-SFA group with 1 SNP (*LINC00460* rs17254590) and 2 lifestyle factors (OC use and age), the combination effects of the risk genotype and risk lifestyle factors on colorectal cancer risk were 15 times greater than null-risk or either of the risk genotype and risk lifestyle factors. This implies a combined gene–lifestyle effect in both additive and multiplicative interaction models (effect size for $G \times E = 9.31$ and $p 0.006$). The Benjamini–Hochberg correction for multiple comparisons was conducted. Furthermore, when stratified by the duration of past OC use, women with a history of shorter use (<5 years) and high risk of both genotype and lifestyle had a 29-times greater risk than women with a history of longer OC use (≥5 years) and null-risk or either of risk genotype and risk lifestyle. Thus the combined risk of the SNP and lifestyle factors was much greater than the risk of any single SNP and lifestyle

factors (Supplementary Table S4B). Furthermore, these findings may indicate a possible joint effect of past OC use with the risk factors on colorectal cancer risk in both additive and multiplicative models (effect size for $G \times E = 2.06$ and $p 0.140$). The high-SFA group analyses (Table 5) yielded similar results but with attenuated gene-lifestyle joint effect on colorectal cancer risk.

Using the nested RSF model with the strongest predictive markers (*MTRR* rs722025, *LINC00460* rs1725459, cigarette smoking, and OC use), we further constructed contour plots to visualize the cumulative colorectal cancer incidence rates of individual SNPs with different combinations of cigarette smoking and OC use, stratified by physical activity (Supplementary Fig. S7) and by SFA intake (Supplementary Fig. S8); the results were consistent with and illustrative of the aforementioned findings.

Discussion

Understanding how obesity and obesity-related lifestyle factors interact with IR pathways (genes and phenotypes), influencing colorectal cancer risk, and further generating colorectal cancer risk profiles that account for both genetic and lifestyle factors is important for the development of a gene–lifestyle combination tool for primary cancer prevention efforts. We performed a 2-stage multimodal RSF analysis to identify the most predictive genetic and lifestyle variables overall and in subgroups (stratified by well-established risk-effect modifiers including BMI, physical activity, and dietary-fat intake; refs. 3, 26). Two SNPs (*LINC00460* rs1725459 and *MTRR* rs722025) and 2

Table 2. Overall analysis: predictive values of variable from the second stage of RSF analysis

Variable ^a	Minimal depth ^b	VIMP	C-index	Error ^c	Drop error ^d
LINC00460 rs17254590	2.0578	0.1518	0.9498	0.0502	0.4498
Duration of oral contraceptive use	2.4190	0.0338	0.9679	0.0321	0.0181
MTRR rs722025	2.8214	0.0927	0.9800	0.0200	0.0122
Age	4.2452	0.0008	0.9813	0.0187	0.0013
Height	4.5468	0.0001	0.9806	0.0194	-0.0008
% Calories from protein	4.5594	0.0000	0.9814	0.0186	0.0008
BMI	4.5794	0.0004	0.9806	0.0194	-0.0007
Age at menopause	4.6370	0.0002	0.9819	0.0181	0.0013
E + P use	4.7536	0.0007	0.9820	0.0180	0.0001
Dietary alcohol	4.8406	0.0002	0.9823	0.0177	0.0003
Total months of breastfeeding	4.9228	0.0000	0.9817	0.0183	-0.0007
Daily intake of fruits	4.9926	0.0001	0.9819	0.0181	0.0002
Daily intake of vegetables	5.0544	0.0001	0.9818	0.0182	-0.0001
Cigarettes per day	5.0838	0.0022	0.9820	0.0180	0.0002
Age at first period	5.4070	-0.0001	0.9818	0.0182	-0.0002
MTRR rs1395139	7.1922	0.0051	0.9820	0.0180	0.0002
MTRR rs6860481	8.1890	0.0023	0.9821	0.0179	0.0001
MTRR rs17197511	8.2598	0.0015	0.9818	0.0182	-0.0003
PABPC1P2 rs79084191	8.9278	0.0046	0.9837	0.0163	0.0019
PABPC1P2 rs78451340	8.9404	0.0047	0.9844	0.0156	0.0007
PABPC1P2 rs75935470	9.0516	0.0040	0.9843	0.0157	-0.0002
PABPC1P2 rs12052223	9.2484	0.0044	0.9842	0.0158	-0.0001
PABPC1P2 rs77164426	9.2824	0.0044	0.9843	0.0157	0.0001
PABPC1P2 rs10928320	9.2862	0.0039	0.9840	0.0160	-0.0004
PABPC1P2 rs77772624	9.3394	0.0044	0.9841	0.0159	0.0001
LOC729506 rs17198862	9.3996	0.0005	0.9845	0.0155	0.0004
LOC729506 rs13188458	10.9650	0.0006	0.9843	0.0157	-0.0002
LOC729506 rs34799743	11.2040	0.0006	0.9842	0.0158	-0.0001
LOC729506 rs13166872	11.2262	0.0007	0.9841	0.0159	-0.0001
LOC729506 rs10512942	11.3146	0.0008	0.9840	0.0160	-0.0001
LOC729506 rs13188952	11.4290	0.0006	0.9842	0.0158	0.0002

NOTE: Variables in bold face were selected as the most predictive markers.

Abbreviation: E + P, exogenous estrogen + progestin.

^aVariables are ordered according to minimal depth.

^bPredictive value for each variable was assessed via minimal depth method in the nested random survival forest models. A lower value is likely to have a greater influence on prediction.

^cThe incremental error rate of each variable was estimated 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 estimated from the third nested model (including the first, second, and third variables).

^dThe drop error rate was estimated by the difference between the error rates from the nested models with a prior and corresponding variables. 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 to 0.5; thus, the drop error rate for the first variable was obtained by subtracting the error rate (0.0502) from 0.5.

lifestyle factors, including lifetime cumulative exposure to estrogen (past OC use) and cigarette smoking, were the most common and strongest predictive markers for colorectal cancer risk across the analyses. With those influential variables, we constructed risk profiles for colorectal cancer in the overall population and within subgroups. It is worthy of note that the combinations of genetic and lifestyle factors had a far greater impact on colorectal cancer risk than any individual risk factors and had a possible synergism to increase colorectal cancer risk.

LINC00460 rs1725459, in our earlier GWA study, was associated with IR phenotypes and in this study, by interacting with dietary fat intake, it is associated with increased risk of colorectal cancer. *LINC00460* is a long intergenic noncoding RNA (lncRNA) 460 (27). Many lncRNAs regulate oncogenes and tumor-suppressive genes' expression and thus have been shown to be involved in carcinogenesis (28). A recent *in vitro* study found that lncRNA *LINC00460* was associated with nasopharyngeal cancer

(NPC; ref. 27) and upregulated substantially in NPC tissues, suggesting its function as an oncogene. Furthermore, miR-149 represses tumor-suppressive miRNA, resulting in dysregulation of AKT1 cellular pathways (29); through the miR-149 pathway, *LINC00460* promotes cell proliferation, migration, and invasion (30). Thus, *LINC00460* may regulate insulin cell-signaling and be involved in tumorigenesis. To the best of our knowledge, our study is the first to show the association of the lncRNA with colorectal cancer development through IR pathways, which is supported for its biologic plausibility by the previous studies (27–30). In addition, a previous GWA study (31) found that *LINC00460* was associated with subcutaneous adipose tissue, supporting our finding that its associations with IR phenotypes and colorectal cancer are observed in fatty-acid strata.

One SNP in *MTRR*, in relation to IR phenotypes by interacting with physical activity, is associated with increased colorectal cancer risk in this study. This is

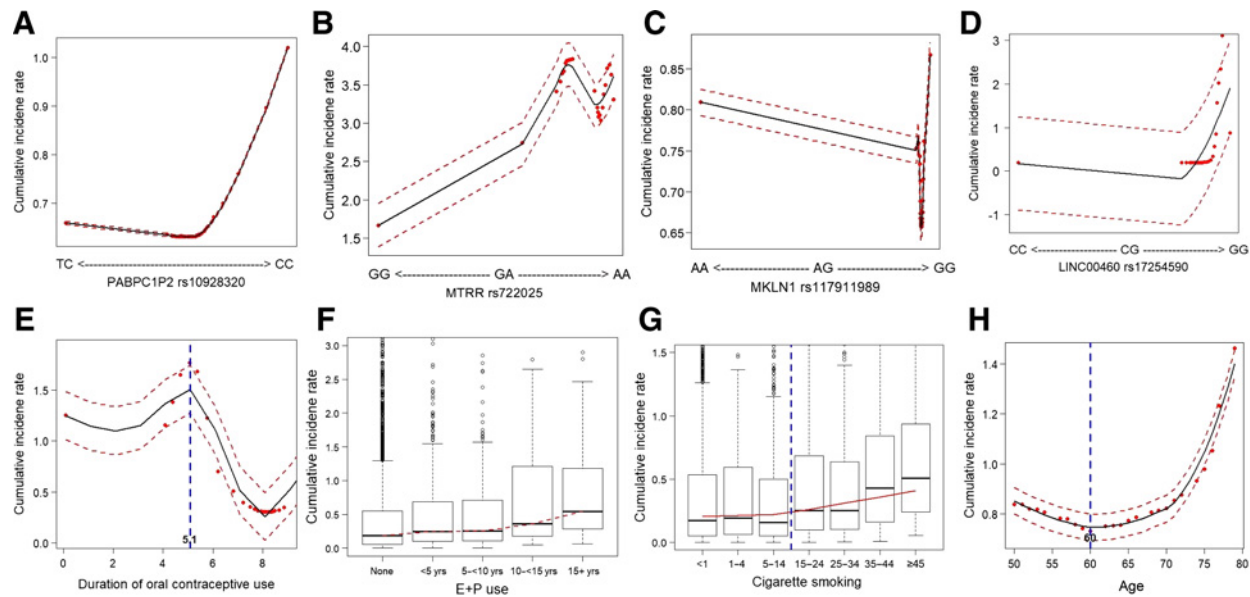


Figure 2.

Cumulative colorectal cancer incidence rate for the 8 most influential variables (4 SNPs and 4 behavioral factors) based on a random survival forest analysis (E + P, exogenous estrogen + progestin). Dashed red lines indicate 95% confidence intervals).

consistent with previous findings that *MTRR* SNPs were associated with type 2 diabetes in adipose tissue (32). The underlying mechanisms are uncertain, but mutations in *MTRR* cause hyperhomocysteinemia, leading to endoplasmic reticular stress and resulting in inhibited insulin signaling in adiposity. In addition, a couple of previous studies have reported a significant association between *MTRR* SNPs and cancers, particularly in lung and colorectal cancers (33, 34). Our finding of the association between the *MTRR* SNP and risk for colorectal cancer is consistent with those from the previous studies but calls attention to the interactions with obesity factors because the colorectal cancer risk of the SNP would be missed without incorporation of physical activity.

Lifetime cumulative exposure to estrogen may play a key role in colorectal carcinogenesis. Particularly, the past use

of exogenous estrogen (e.g., OC) has been considered a protective factor for postmenopausal colorectal cancer risk. Several *in vivo* and *in vitro* studies indicated that estrogen upregulates several cell-cycle regulators such as p53, leading to growth-inhibiting effects on colorectal cancer cells (35) and is involved in the epigenetic pathway of the CpG-island, resulting in a hypermethylation phenotype (36). However, epidemiologic evidence for the relationship between OC use and colorectal cancer is not conclusive: no associations (18, 37), reduced risk with increased duration of use (38), no clear risk reduction with the duration of use (39, 40), reduced risk with (39) or without (40) recency of use, and possible increased risk for colorectal cancer (41). These mixed findings may be in part explained by a lack of consideration of the duration of OC use by accounting for its nonlinear effect. Our RSF cumulative colorectal cancer incidence rate showed nonlinear associations with colorectal cancer; the risk increases up to 5 years of OC use, but drops thereafter.

Few studies have reported that OC use is associated with a reduced risk of colorectal cancer in the presence of specific molecular features [e.g., estrogen receptor- β (42) and microsatellite instability positivity (35)]. Because we had no data on the molecular features of the tumors, our findings should be revisited with independent samples and data on molecular subtypes. In addition, earlier OC formulations (in pre-1980) had high estrogen levels and the formulations of OCs have since been changed (38); thus, different OC preparations could result in different effects on cancer risk. Our data did not include the types of OC formulations, so future studies are warranted to

Table 3. Overall analysis: results from multivariate regression predicting CRC risk

Variable	HR ^a (95% CI)	P
SNP (Ref/Alt)		
<i>LINC00460</i> rs17254590 (CC + CG/GG)	12.48 (1.76–88.77)	0.012
<i>MTRR</i> rs722025 (GG/GA + AA)	7.54 (3.38–16.85)	<0.001
Behavioral factor		
Duration of oral contraceptive use^b	3.12 (2.69–3.63)	<0.001

NOTE: Numbers in bold face are statistically significant.

Abbreviations: Alt, alternative allele; CI, confidence interval; CRC, colorectal cancer; Ref, reference allele.

^aMultivariate regression was adjusted by age, smoking, height, BMI, dietary alcohol, daily fruits, daily vegetables, % calories from protein, age at menarche, age at menopause, breastfeeding, and exogenous estrogen plus progestin use.

^bOral contraceptive use was analyzed as a binary variable using 5.1 years as a cut-off value, at which colorectal cancer risk was diverged into high (<5.1 years) and low (\geq 5.1 years) in a cumulative graph for the colorectal cancer incidence rate.

Table 4. Physical activity–stratified analysis: combined effect of risk genotypes of *MTRR* rs722025 and *MKLN1* rs117911989 and behavioral factors on CRC risk

n	Total		n	Cigarette smoking < 15		n	Cigarette smoking ≥ 15	
	HR ^a (95% CI)	P ^b		HR ^a (95% CI)	P ^b		HR ^a (95% CI)	P ^b
< Active group (MET ≥ 10; n = 4,643) >								
Risk genotype (<i>MTRR</i> rs722025 GA + AA and <i>MKLN1</i> rs117911989 GG)								
0	Reference		328	Reference		92	2.10 (0.50–8.81)	0.310
1	0.82 (0.37–1.79)	0.616	1,519	0.93 (0.35–2.48)	0.892	426	1.30 (0.44–3.89)	0.637
2	7.00 (3.46–14.17)	<0.001	1,799	8.39 (3.45–20.40)	<0.001	479	9.79 (3.94–24.31)	<0.001
Behavioral factors (oral contraceptive use, cigarette smoking, and age at enrollment) ^c								
0	Reference		928	Reference		300	1.32 (0.78–2.23)	0.308
1	1.55 (1.13–2.14)	0.011	2,228	1.12 (0.79–1.59)	0.529	578	1.33 (0.85–2.07)	0.213
2	6.75 (4.18–10.90)	<0.001	490	4.92 (3.40–7.14)	<0.001	119	7.10 (4.39–11.49)	<0.001
Risk genotypes combined with behavioral factors ^d								
0	Reference		1,634	Reference		465	0.71 (0.24–2.11)	0.543
1	10.02 (6.87–14.62)	<0.001	1,735	8.09 (4.96–13.21)	<0.001	466	12.25 (7.21–20.83)	<0.001
2	32.26 (18.21–57.15)	<0.001	277	36.29 (21.62–60.92)	<0.001	66	39.96 (20.99–76.06)	<0.001
< Inactive group (MET < 10; n = 6,435) >								
Risk genotype (<i>MTRR</i> rs722025 GA + AA)								
0	Reference		1,885	Reference		590	1.30 (0.66–2.54)	0.452
1	6.01 (4.35–8.29)	<0.001	3,051	6.34 (4.33–9.29)	<0.001	909	6.73 (4.42–10.25)	<0.001
Behavioral factors (oral contraceptive use, cigarette smoking, and E + P use) ^e								
0	Reference		3,066	Reference		825	1.26 (0.91–1.74)	0.157
1	1.57 (1.28–1.93)	<0.001	1,702	1.68 (1.33–2.12)	<0.001	601	1.64 (1.17–2.30)	0.006
2	2.49 (1.25–4.95)	0.012	168	1.94 (1.16–3.24)	0.014	73	2.52 (1.27–5.01)	0.012
Risk genotypes combined with behavioral factors ^d								
0	Reference		1,820	Reference		570	1.19 (0.57–2.44)	0.645
1	6.25 (4.49–8.69)	<0.001	3,013	6.02 (4.08–8.89)	<0.001	876	7.19 (4.69–11.04)	<0.001
2	8.43 (3.74–19.02)	<0.001	103	10.26 (5.50–19.12)	<0.001	53	9.07 (3.92–20.99)	<0.001

NOTE: Numbers in bold face are statistically significant.

Abbreviations: CI, confidence interval; CRC, colorectal cancer; E + P, exogenous estrogen + progestin; MET, metabolic equivalent.

^aMultivariate regression for risk genotype analysis was adjusted by age, height, BMI, dietary alcohol, daily fruits, daily vegetables, % calories from protein, age at menarche, age at menopause, breastfeeding, duration of oral contraceptive use, E + P use, and smoking (in total analysis); in behavioral factor analysis, variables tested for risk factors and joint effect were not included as a covariate in the multivariate regression.

^bP values were adjusted to correct for multiple testing via the Benjamini–Hochberg approach.

^cThe number of behavioral factors was defined as 0 (low risk: null risk behaviors), 1 (moderate risk: 1 or 2 risk behaviors), and 2 (high risk: 3 risk behaviors).

^dThe combined number of risk genotypes and behavioral factors was based on risk genotypes defined as 0 [low risk: none or 1 risk allele (active group); none (inactive group)] and 1 [high risk: 2 risk alleles (active); 1 risk allele (inactive group)] and based on behavioral factors defined as 0 (low risk: ≤2 risk behaviors) and 1 (high risk: 3 risk behaviors). The ultimate number of risk genotypes combined with behavioral factors was defined as 0 (low risk for genotypes and behaviors), 1 (high risk for either genotypes or behaviors), and 2 (high risk for both genotypes and behaviors).

Table 5. SFA-stratified analysis: combined effect of risk genotypes of *LINC00460* rs17254590 and *PABPC1P2* rs10928320 and behavioral factors on CRC risk

n	Total		n	Oral contraceptive use ≥ 5 years		n	Oral contraceptive use < 5 years	
	HR ^a (95% CI)	P ^b		HR ^a (95% CI)	P ^b		HR ^a (95% CI)	P ^b
< % Calories from SFA < 7.0 % (n = 1,009) >								
Risk genotype (<i>LINC00460</i> rs17254590 GG)								
0	Reference		531	Reference		169	2.29 (0.85–6.17)	0.101
1	8.42 (4.84–14.66)	<0.001	168	5.51 (2.59–11.74)	<0.001	141	25.41 (12.80–50.45)	<0.001
Behavioral factors (oral contraceptive use and age at enrollment) ^c								
0	Reference		78	Reference		84	1.74 (0.48–6.37)	0.402
1	5.24 (3.22–8.51)	<0.001	621	1.05 (0.35–3.15)	0.928	226	5.83 (1.97–17.26)	0.002
Risk genotypes combined with behavioral factors ^d								
0	Reference		556	Reference		212	1.80 (0.74–4.39)	0.199
1	15.68 (9.59–25.62)	<0.001	143	5.47 (2.58–11.58)	<0.001	98	28.76 (14.86–55.66)	<0.001
< % Calories from SFA ≥ 7.0 % (n = 10,069) >								
Risk genotype (<i>LINC00460</i> rs17254590 GG and <i>PABPC1P2</i> rs10928320 CC) ^e								
0	Reference		1,208	Reference		628	1.10 (0.44–2.76)	0.839
1	8.68 (5.55–13.57)	<0.001	5,816	5.38 (3.08–9.40)	<0.001	2,417	16.82 (9.65–29.30)	<0.001

NOTE: Numbers in bold face are statistically significant.

Abbreviations: CI, confidence interval; CRC, colorectal cancer.

^aMultivariate regression for risk genotype analysis was adjusted by age, smoking, height, BMI, dietary alcohol, daily fruits, daily vegetables, % calories from protein, age at menarche, age at menopause, breastfeeding, exogenous estrogen plus progestin use, and duration of oral contraceptive use (in total analysis); in behavioral factor analysis, variables tested for risk factors and joint effect were not included as a covariate in the multivariate regression.

^bP values were adjusted to correct for multiple testing via the Benjamini–Hochberg approach.

^cThe number of behavioral factors was defined as 0 (low risk: null or 1 risk behavior) and 1 (high risk: 2 risk behaviors).

^dThe combined number of risk genotypes and behavioral factors was based on risk genotypes defined as 0 (low risk: none) and 1 (high risk: 1 risk allele) and based on behavioral factors defined as 0 (low risk: null or 1 risk behavior) and 1 (high risk: 2 risk behaviors). The ultimate number of risk genotypes combined with behavioral factors was defined as 0 (neither or either of high risk for genotypes and behaviors) and 1 (high risk for both genotypes and behaviors).

^eThe number of risk genotypes was defined as 0 (none or 1 risk allele) and 1 (high risk: 2 risk alleles).

Downloaded from http://aacrjournals.org/cancerpreventionresearch/article-pdf/12/12/877/224394/877.pdf by guest on 18 April 2024

examine the different effects on colorectal cancer risk according to OC preparation.

In this study, using 5 years of OC use as the cut-off point, we observed a possible joint effect of OC use with SNPs and lifestyle factors on colorectal cancer risk and this joint effect was attenuated in the high dietary-fat subgroup. This may suggest a potential trade-off pathway between female hormones and fatty acids, reflecting the minimized effect of estrogen in high fatty-acid levels.

Cigarette smoking may contribute to 15% to 20% of colorectal cancer cases (26, 43) with a dose-response relationship, including daily cigarette consumption, years of smoking (44), and the induction period (the time since the onset of smoking; ref. 26). Tobacco-derived carcinogens reach the colorectal mucosa through the digestive tract and the circulatory system, which may cause the potential carcinogenesis in this target organ (45). Our study population had, on average, a 15-year induction period, 50% of the 30-year period suggested in previous studies (46, 47) between smoking onset and cancer formation; however, the combined effect on colorectal cancer risk of daily consumption of 15 or more cigarettes with selected IR SNPs and lifestyle factors was tremendous in our physical activity strata (i.e., interactions with degree of exercise). This finding is supported by those from a previous report (48) of the interaction pathways of smoking, colorectal cancer, and obesity, and further suggests biologic mechanism studies such as IR-gene signature and cell signaling in relation to colorectal cancer cells of postmenopausal women with a history of smoking by different levels of obesity and/or exercise.

Our findings should not be extrapolated to other populations as our study population was limited to non-Hispanic white postmenopausal women. Despite several advantages of the 2-stage RSF multimodal approach, it could over-fit the model owing to noisy tasks, especially in relatively small subgroups. Our findings need to be replicated in an independent study with large samples.

Overall, in this study, the IR SNPs identified through the GWA study have a potential synergistic effect on colorectal cancer risk with lifestyle factors including lifetime exposure to exogenous estrogen and cigarette smoking. Our findings may inform future research on the role of IR in the etiology of colorectal cancer and contribute to greater accuracy in predicting colorectal cancer risk, suggesting the potential for the development of intervention strategies for women who carry the risk genotypes, which may reduce their risk for colorectal cancer.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

References

1. American Cancer Society. Cancer Fact and Figures, 2019. Atlanta: American Cancer Society, Inc.: <https://www.cancer.org/content/>

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

Brigham and Women's Hospital, Harvard Medical School, Boston, MA: JoAnn E. Manson; MedStar Health Research Institute/Howard University, Washington, DC: Barbara V. Howard; Stanford Prevention Research Center, Stanford, CA: Marcia L. Stefanick; The Ohio State University, Columbus, OH: Rebecca Jackson; University of Arizona, Tucson/Phoenix, AZ: Cynthia A. Thomson; University at Buffalo, Buffalo, NY: Jean Wactawski-Wende; University of Florida, Gainesville/Jacksonville, FL: Marian Limacher; University of Iowa, Iowa City/Davenport, IA: Robert Wallace; University of Pittsburgh, Pittsburgh, PA: Lewis Kuller; Wake Forest University School of Medicine, Winston-Salem, NC: Sally Shumaker.

Authors' Contributions

Conception and design: S.Y. Jung, J.C. Papp

Development of methodology: S.Y. Jung, J.C. Papp, E.M. Sobel

Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): S.Y. Jung, J.C. Papp, E.M. Sobel

Writing, review, and/or revision of the manuscript: S.Y. Jung, J.C. Papp, E.M. Sobel, Z.-F. Zhang

Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): S.Y. Jung

Study supervision: J.C. Papp

Acknowledgments

This study was supported by the National Institute of Nursing Research of the NIH under Award Number K01NR017852 and a University of California Cancer Research Coordinating Committee grant (CRN-18-522722). 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 NIH, and the U.S. Department of Health and Human Services through contracts HHSN268201100046C, HHSN268201100001C, HHSN268201100002C, HHSN268201100003C, HHSN-26820-1100004C, and HHSN271201100004C. The datasets 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).

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 June 2, 2019; revised August 28, 2019; accepted September 20, 2019; published first September 25, 2019.

dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2019/cancer-facts-and-figures-2019.pdf.

2. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2018;68:394–424.
3. American Cancer Society. Colorectal Cancer Facts & Figures 2017–2019. Atlanta: American Cancer Society, Inc. 2017: <https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/colorectal-cancer-facts-and-figures/colorectal-cancer-facts-and-figures-2017-2019.pdf>.
4. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, et al. Environmental and heritable factors in the causation of cancer—analyses of cohorts of twins from Sweden, Denmark, and Finland. *N Engl J Med* 2000;343:78–85.
5. Ma Y, Yang Y, Wang F, Zhang P, Shi C, Zou Y, et al. Obesity and risk of colorectal cancer: a systematic review of prospective studies. *PLoS One* 2013;8:e53916.
6. Boyle T, Keegel T, Bull F, Heyworth J, Fritschi L. Physical activity and risks of proximal and distal colon cancers: a systematic review and meta-analysis. *J Natl Cancer Inst* 2012;104:1548–61.
7. Aune D, Lau R, Chan DS, Vieira R, Greenwood DC, Kampman E, et al. Dairy products and colorectal cancer risk: a systematic review and meta-analysis of cohort studies. *Ann Oncol* 2012;23:37–45.
8. Ho GY, Wang T, Gunter MJ, Strickler HD, Cushman M, Kaplan RC, et al. Adipokines linking obesity with colorectal cancer risk in postmenopausal women. *Cancer Res* 2012;72:3029–37.
9. Tran TT, Naigamwalla D, Oprescu AI, Lam L, McKeown-Eyssen G, Bruce WR, et al. Hyperinsulinemia, but not other factors associated with insulin resistance, acutely enhances colorectal epithelial proliferation in vivo. *Endocrinology* 2006;147:1830–7.
10. Kabat GC, Kim MY, Peters U, Stefanick M, Hou L, Wactawski-Wende J, et al. A longitudinal study of the metabolic syndrome and risk of colorectal cancer in postmenopausal women. *Eur J Cancer Prev* 2012;21:326–32.
11. Gunter MJ, Hoover DR, Yu H, Wassertheil-Smoller S, Rohan TE, Manson JE, et al. Insulin, insulin-like growth factor-I, endogenous estradiol, and risk of colorectal cancer in postmenopausal women. *Cancer Res* 2008;68:329–37.
12. Jung SY, Sobel EM, Papp JC, Crandall CJ, Fu AN, Zhang ZF. Obesity and associated lifestyles modify the effect of glucose metabolism-related genetic variants on impaired glucose homeostasis among postmenopausal women. *Genet Epidemiol* 2016;40:520–30.
13. Jung SY, Mancuso N, Yu H, Papp J, Sobel E, Zhang ZF. Genome-wide meta-analysis of gene-environment interaction for insulin resistance phenotypes and breast cancer risk in postmenopausal women. *Cancer Prev Res* 2019;12:31–42.
14. Mogensen UB, Ishwaran H, Gerds TA. Evaluating random forests for survival analysis using prediction error curves. *J Stat Softw* 2012;50:1–23.
15. Hamidi O, Poorolajal J, Farhadian M, Tapak L. Identifying important risk factors for survival in kidney graft failure patients using random survival forests. *Iran J Public Health* 2016;45:27–33.
16. Design of the Women's Health Initiative clinical trial and observational study. The Women's Health Initiative Study Group. *Control Clin Trials* 1998;19:61–109.
17. NCBI: WHI Harmonized and Imputed GWAS Data. A sub-study of Women's Health Initiative. Available from: http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000746.v1.p3.
18. Nichols HB, Trentham-Dietz A, Hampton JM, Newcomb PA. Oral contraceptive use, reproductive factors, and colorectal cancer risk: findings from Wisconsin. *Cancer Epidemiol Biomarkers Prev* 2005;14:1212–8.
19. National Cancer Institute. SEER Program: Comparative Staging Guide For Cancer. 1993. Available from: https://seer.cancer.gov/archive/manuals/historic/comp_stage1.1.pdf.
20. Schumacher FR, Al Olama AA, Berndt SI, Benlloch S, Ahmed M, Saunders EJ, et al. Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat Genet* 2018;50:928–36.
21. Matthews DR, Hosker JP, Rudenski AS, Naylor BA, Treacher DF, Turner RC. Homeostasis model assessment: insulin resistance and beta-cell function from fasting plasma glucose and insulin concentrations in man. *Diabetologia* 1985;28:412–9.
22. Ishwaran H, Kogalur UB. Random Survival Forests for R. 2007. Available from: <https://pdfs.semanticscholar.org/951a/84f0176076fb6786fdf43320e8b27094dcfa.pdf>.
23. Chung RH, Chen YE. A two-stage random forest-based pathway analysis method. *PLoS One* 2012;7:e36662.
24. Harrell FE Jr, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA* 1982;247:2543–6.
25. Wiens BL, Dmitrienko A, Marchenko O. Selection of hypothesis weights and ordering when testing multiple hypotheses in clinical trials. *J Biopharm Stat* 2013;23:1403–19.
26. Chan AT, Giovannucci EL. Primary prevention of colorectal cancer. *Gastroenterology* 2010;138:2029–43.
27. Kong YG, Cui M, Chen SM, Xu Y, Tao ZZ. LncRNA-LINC00460 facilitates nasopharyngeal carcinoma tumorigenesis through sponging miR-149-5p to up-regulate IL6. *Gene* 2018;639:77–84.
28. Fang J, Sun CC, Gong C. Long noncoding RNA XIST acts as an oncogene in non-small cell lung cancer by epigenetically repressing KLF2 expression. *Biochem Biophys Res Commun* 2016;478:811–7.
29. Ghasemi A, Fallah S, Ansari M. MicroRNA-149 is epigenetically silenced tumor-suppressive microRNA, involved in cell proliferation and downregulation of AKT1 and cyclin D1 in human glioblastoma multiforme. *Biochem Cell Biol* 2016;94:569–76.
30. Zhang M, Gong W, Zuo B, Chu B, Tang Z, Zhang Y, et al. The microRNA miR-33a suppresses IL-6-induced tumor progression by binding Twist in gallbladder cancer. *Oncotarget* 2016;7:78640–52.
31. Sung YJ, Perusse L, Sarzynski MA, Fornage M, Sidney S, Sternfeld B, et al. Genome-wide association studies suggest sex-specific loci associated with abdominal and visceral fat. *Int J Obes* 2016;40:662–74.
32. Zhi X, Yang B, Fan S, Li Y, He M, Wang D, et al. Additive interaction of MTHFR C677T and MTRR A66G polymorphisms with being overweight/obesity on the risk of type 2 diabetes. *Int J Environ Res Public Health* 2016;13:1243.
33. Wu PP, Tang RN, An L. A meta-analysis of MTRR A66G polymorphism and colorectal cancer susceptibility. *J BUON* 2015;20:918–922.
34. Aksoy-Sagirlı P, Erdenay A, Kaytan-Saglam E, Kizir A. Association of three single nucleotide polymorphisms in MTR and MTRR genes with lung cancer in a Turkish population. *Genet Test Mol Biomarkers* 2017;21:428–32.
35. Slattery ML, Potter JD, Curtin K, Edwards S, Ma KN, Anderson K, et al. Estrogens reduce and withdrawal of estrogens increase risk of microsatellite instability-positive colon cancer. *Cancer Res* 2001;61:126–30.
36. Issa JP. Colon cancer: it's CIN or CIMP. *Clin Cancer Res* 2008;14:5939–40.
37. Brandstedt J, Wangefjord S, Nodin B, Eberhard J, Jirstrom K, Manjer J. Associations of hormone replacement therapy and oral contraceptives with risk of colorectal cancer defined by clinicopathological factors, beta-catenin alterations, expression of cyclin

- D1, p53, and microsatellite-instability. *BMC Cancer* 2014;14:371.
38. Martinez ME, Grodstein F, Giovannucci E, Colditz GA, Speizer FE, Hennekens C, et al. A prospective study of reproductive factors, oral contraceptive use, and risk of colorectal cancer. *Cancer Epidemiol Biomarkers Prev* 1997;6:1–5.
 39. Bosetti C, Bravi F, Negri E, La Vecchia C. Oral contraceptives and colorectal cancer risk: a systematic review and meta-analysis. *Hum Reprod Update* 2009;15:489–98.
 40. Levi F, Pasche C, Lucchini F, La Vecchia C. Oral contraceptives and colorectal cancer. *Dig Liver Dis* 2003;35:85–7.
 41. Kune GA, Kune S, Watson LF. Oral contraceptive use does not protect against large bowel cancer. *Contraception* 1990;41:19–25.
 42. Rudolph A, Toth C, Hoffmeister M, Roth W, Herpel E, Schirrmacher P, et al. Colorectal cancer risk associated with hormone use varies by expression of estrogen receptor-beta. *Cancer Res* 2013;73:3306–15.
 43. Giovannucci E, Martinez ME. Tobacco, colorectal cancer, and adenomas: a review of the evidence. *J Natl Cancer Inst* 1996;88:1717–30.
 44. Hartz A, He T, Ross JJ. Risk factors for colon cancer in 150,912 postmenopausal women. *Cancer Causes Control* 2012;23:1599–605.
 45. 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 U S A* 1977;74:3555–9.
 46. Giovannucci E. An updated review of the epidemiological evidence that cigarette smoking increases risk of colorectal cancer. *Cancer Epidemiol Biomarkers Prev* 2001;10:725–31.
 47. Giovannucci E, Colditz GA, Stampfer MJ, Hunter D, Rosner BA, Willett WC, et al. A prospective study of cigarette smoking and risk of colorectal adenoma and colorectal cancer in U.S. women. *J Natl Cancer Inst* 1994;86:192–9.
 48. Arnold M, Freisling H, Stolzenberg-Solomon R, Kee F, O'Doherty MG, Ordonez-Mena JM, et al. Overweight duration in older adults and cancer risk: a study of cohorts in Europe and the United States. *Eur J Epidemiol* 2016;31:893–904.
 49. Haskell WL, Lee IM, Pate RR, Powell KE, Blair SN, Franklin BA, et al. Physical activity and public health: updated recommendation for adults from the American College of Sports Medicine and the American Heart Association. *Med Sci Sports Exerc* 2007;39:1423–34.
 50. Van Horn L, Carson JA, Appel LJ, Burke LE, Economos C, Karmally W, et al. Recommended Dietary Pattern to Achieve Adherence to the American Heart Association/American College of Cardiology (AHA/ACC) Guidelines: A Scientific Statement From the American Heart Association. *Circulation* 2016;134:e505–29.

