Title

Breast Cancer Risk and Insulin Resistance: Post Genome-Wide Gene–Environment Interaction Study Using a Random Survival Forest

Permalink
https://escholarship.org/uc/item/3dn6w4c0

Journal
Cancer Research, 79(10)

ISSN
0008-5472

Authors
Jung, Su Yon
Papp, Jeanette C
Sobel, Eric M
et al.

Publication Date
2019-05-15

DOI
10.1158/0008-5472.can-18-3688

Peer reviewed

Published OnlineFirst April 1, 2019; DOI: 10.1158/0008-5472.CAN-18-3688

Population and Prevention Science

Cancer Research

Breast Cancer Risk and Insulin Resistance: Post Genome-Wide Gene–Environment Interaction Study Using a Random Survival Forest

Su Yon Jung[1], Jeanette C. Papp[2], Eric M. Sobel[2], Herbert Yu[3], and Zuo-Feng Zhang[4]

Abstract

Obesity–insulin connections have been considered potential risk factors for postmenopausal breast cancer, and the association between insulin resistance (IR) genotypes and phenotypes can be modified by obesity-lifestyle factors, affecting breast cancer risk. In this study, we explored the role of IR in those pathways at the genome-wide level. We identified IR-genetic factors and selected lifestyles to generate risk profiles for postmenopausal breast cancer. Using large-scale cohort data from postmenopausal women in the Women's Health Initiative Database for Genotypes and Phenotypes Study, our previous genome-wide association gene–behavior interaction study identified 58 loci for associations with IR phenotypes (homeostatic model assessment–IR, hyperglycemia, and hyperinsulinemia). We evaluated those single-nucleotide polymorphisms (SNP) and additional 31 lifestyles in relation to breast cancer risk by conducting a two-stage multimodal random survival forest analysis. We identified the most predictive genetic and lifestyle variables in overall and subgroup analyses [stratified by body mass index (BMI), exercise, and dietary fat intake]. Two SNPs (*LINC00460* rs17254590 and *MKLN1* rs117911989), exogenous factors related to lifetime cumulative exposure to estrogen, BMI, and dietary alcohol consumption were the most common influential factors across the analyses. Individual SNPs did not have significant associations with breast cancer, but SNPs and lifestyles combined synergistically increased the risk of breast cancer in a gene–behavior, dose-dependent manner. These findings may contribute to more accurate predictions of breast cancer and suggest potential intervention strategies for women with specific genetic and lifestyle factors to reduce their breast cancer risk.

Significance: These findings identify insulin resistance SNPs in combination with lifestyle as synergistic factors for breast cancer risk, suggesting lifestyle changes can prevent breast cancer in women who carry the risk genotypes.

Introduction

Insulin resistance (IR), leading to glucose intolerance, such as high blood level of homeostatic model assessment–IR (HOMA-IR), hyperglycemia, and compensatory hyperinsulinemia, is thought to be central to the development of many obesity-relevant cancers such as postmenopausal breast cancer (1–3). For postmenopausal women, HOMA-IR, reflecting high blood levels of both insulin and glucose, and hyperglycemia contributes to 1.5 times higher risk for breast cancer (4). Hyperinsulinemia has been associated with a doubled risk for postmenopausal breast cancer (5, 6). Given the relationships between IR and breast cancer risk, IR-related genetic variants can potentially affect the risk of breast cancer.

Obesity is a well-established risk factor for postmenopausal breast cancer (3), and obesity–insulin connections might be crucial in the development of breast cancer (1). In particular, obesity status, physical inactivity, and high dietary-fat intake interact with the IR-related phenotypes, increasing breast cancer susceptibility (7–10). Furthermore, recent *in vitro* studies have shown IR-related gene signature and aberrantly amplified insulin signaling in breast cancer cells of obese postmenopausal women, implying the existence of molecular–genetic pathways between obesity, IR, and postmenopausal breast cancer (1, 11). In addition, our previous population-based epidemiology study (12) revealed that IR-relevant single-nucleotide polymorphisms (SNP) have greater increases in IR phenotypes among obese, inactive, and high-fat diet groups, suggesting that obesity modifies the associations between IR-genetic variants and their phenotypes, and thus jointly influences cancer susceptibility. Therefore, the association between IR (genotype and phenotype) and cancer risk can be modified by obesity status and obesity-related lifestyle factors (Supplementary Fig. S1).

For gene–phenotype association with behavioral interactions, no study at the genome-wide level in the published literature has explored the interacting role of obesity status and related lifestyle factors in the pathways among IR-relevant genetic variants, phenotypes, and postmenopausal breast cancer risk. Understanding how those lifestyle factors modify and interact with genes and phenotypes is important for developing a tool for use in primary

[1]Translational Sciences Section, Jonsson Comprehensive Cancer Center, School of Nursing, University of California, Los Angeles, Los Angeles, California. [2]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, California. [3]Cancer Epidemiology Program, University of Hawaii Cancer Center, Honolulu, Hawaii. [4]Department 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 Research Online (http://cancerres.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

doi: 10.1158/0008-5472.CAN-18-3688

Published OnlineFirst April 1, 2019; DOI: 10.1158/0008-5472.CAN-18-3688

breast cancer prevention efforts. Furthermore, few studies have incorporated genetic and lifestyle factors to generate risk profiles for breast cancer and to construct breast cancer risk models with risk profiles (13). Risk models including both factors will have greater accuracy in predicting breast cancer risk.

To address these critical gaps, by using large-scale cohort data from postmenopausal women in the Women's Health Initiative Database for Genotypes and Phenotypes (WHI dbGaP) Study, we have evaluated 58 loci (Supplementary Table S1) identified for their associations with IR phenotypes (HOMA-IR, hyperglycemia, and hyperinsulinemia) in our previous genome-wide association (GWA) gene–environmental (i.e., behavioral) interaction (G × E) study (14). Briefly, the 58 genome-wide significant loci were associated with IR phenotypes in women stratified by obesity (4 SNPs), physical activity (36 SNPs), and dietary-fat intake (18 SNPs).

In this study, we examined the association of these SNPs with breast cancer risk in obesity lifestyle-stratified subgroups in which the SNPs were associated with IR at genome-wide significance. It is to evaluate whether those SNPs that interact with obesity-related lifestyle factors in a particular behavioral setting (e.g., in obese/physical activity/dietary-fat intake groups) are associated with breast cancer risk in the identical behavioral setting. This may elaborate an empirical pathway where a significant proportion of the susceptibility of SNPs identified in the GWA study, through interactions with specific lifestyles, influence breast cancer risk (Supplementary Fig. S1).

In addition, we selected 31 lifestyle factors for this study. We evaluated the SNPs and lifestyle factors by conducting a two-stage random survival forest (RSF) analysis and ranked them according to their predictive value and accuracy for breast cancer. The RSF, a machine learning method, is a nonparametric tree-based ensemble method and accounts for the nonlinear effects of variables that may not be handled in a traditional regression model (15, 16). RSF also allows for high-order interactions among variables and has been successfully used to yield accurate predictions (15). With the most influential SNPs and lifestyle factors identified through the two-stage RSF, we fit predictive models for breast cancer risk. We further examined the combined effect of those identified variables on breast cancer risk and evaluated a gene–behavior dose–response relationship. By applying the two complementary strategies (RSF and regression), we ultimately tested the hypothesis that the most influential genetic and behavioral factors identified through the RSF analysis interact jointly to predict breast cancer risk.

Patients and Methods

Study population

Our study population is postmenopausal women who were enrolled in the WHI Harmonized and Imputed GWA Studies, a joint imputation and harmonization effort for GWA study within the WHI 2 study arms, including Clinical Trials and Observational Studies. The studies' detailed rationale and design have been described elsewhere (17, 18), but briefly, the WHI study included postmenopausal women enrolled between 1993 and 1998 at 40 clinical centers across the United States. Eligible women were 50–79 years old, postmenopausal, expected to live near the clinical centers for at least three years after enrollment, and able to provide written informed consent. Participants enrolled in the WHI study were eligible for the dbGaP study if they had met eligibility requirements for submission to dbGaP and provided DNA samples. The Harmonization and Imputation GWA Studies under the dbGaP study accession (phs000200.v11.p3) involved six GWA Studies [MOPMAP (AS264); GARNET; GECCO-CYTO; GECCO-INIT; HIPFX; and WHIMS]. With those six GWA studies, we initially included 16,088 women who reported their race or ethnicity as non-Hispanic white (Supplementary Fig. S2). In our previous GWA G × E study for the association with IR phenotypes, by applying exclusion criteria, we excluded (i) 2,714 who had diabetes at or after enrollment; (ii) 1,271 whose genetic data were duplicated and/or related to others; and (iii) 309 outliers based on principal components, resulting in 11,794 women. In this study, we excluded an additional 685 women who had been followed up for less than one year and/or had been diagnosed with any type of cancer at enrollment, leaving a total of 11,109 women (589 of them had developed breast cancer). Participants in this study had been followed up until August 29, 2014, with a median follow-up period of 16 years. This study has been approved by the Institutional Review Boards of each participating clinical center of the WHI and the University of California, Los Angeles.

Data collection and breast cancer outcome

Participants completed self-administered questionnaires at screening, providing demographic and socioeconomic information, medical and reproductive histories, and lifestyle behaviors. For this study, we evaluated information on demographic factors (age, education, marital status, family income, and family history of breast cancer), lifestyles (depressive symptoms, physical activity, cigarettes per day, and daily diet [dietary intake of alcohol, fiber, and total sugars, fruits, and vegetables; % calories from protein, carbohydrates, saturated fatty acids (SFA), monounsaturated fatty acids (MFA), and polyunsaturated fatty acids (PFA)], and medical (hypertension, high cholesterol, and cardiovascular disease) and reproductive histories [hysterectomy, age at menarche and menopause, number of pregnancies, months of breastfeeding, and durations of previous oral contraceptive and hormone replacement therapy of unopposed (exogenous estrogen only) and opposed estrogen use (exogenous estrogen plus progestin)]. We also used anthropometric measurements, including height, weight, and waist and hip circumferences that were measured by trained staff. These 31 variables were identified by literature review for their association with IR phenotypes and breast cancer (19), and after multicollinearity testing and univariate and stepwise regression analyses, were selected for inclusion in this study.

Participants' breast cancer outcomes were verified via a centralized review of medical charts, and cancer sites were coded according to the National Cancer Institute's Surveillance, Epidemiology, and End Results guidelines (20). The breast cancer outcome variables were (i) cancer development (yes/no) and (ii) the time to develop the cancer, estimated as the time in days between enrollment and breast cancer development, censoring, or study endpoint, and then converted into years.

Genotyping and laboratory methods

Details of the data-cleaning process applied to the genotyped data obtained from the WHI Harmonized and Imputed studies have been described previously (14). Briefly, the genotyped data were normalized via the reference panel GRCh37, and imputation was conducted via the 1000 Genomes Project reference panel (18); SNPs for harmonization were checked for pairwise concordance

Table 1. Characteristics of participants, stratified by breast cancer

Characteristics	Breast cancer cases (n = 589) n (%)	Controls (n = 10,520) n (%)
Age in years, median (range)	67 (50–79)	67 (50–81)
Education		
≤ High school	179 (30.4)	3,761 (35.8)[a]
> High school	410 (69.6)	6,759 (64.2)
Family income		
< $35,000	217 (37.5)	4,674 (45.4)[a]
≥ $35,000	361 (62.5)	5,630 (54.6)
Family history of breast cancer		
No	454 (77.1)	8,534 (81.1)[a]
Yes	135 (22.9)	1,986 (18.9)
Depressive symptom[b], median (range)	0.002 (0.001–0.880)	0.002 (0.000–0.937)
Dietary alcohol per day in g, median (range)	1.88 (0.00–127.15)	1.06 (0.00–183.76)[a]
Dietary alcohol per day[c]		
< 1.07	258 (43.8)	5,296 (50.3)[a]
≥ 1.07	331 (56.2)	5,224 (49.7)
% calories from SFA, median (range)	11.49 (3.73–21.50)	11.29 (2.22–32.39)
% calories from SFA[d]		
< 7.0	50 (8.5)	960 (9.1)
≥ 7.0	539 (91.5)	9,560 (90.9)
% calories from carbohydrates, median (range)	47.50 (18.98–80.77)	48.90 (1.51–85.84)[a]
% calories from MFA, median (range)	12.92 (4.08–24.51)	12.78 (2.16–27.64)
% calories from PFA, median (range)	6.55 (2.58–20.25)	6.61 (1.19–21.77)
METs·hour·week^{-1}[e]	7.00 (0.00–81.67)	7.50 (0.00–134.17)
METs·hour·week^{-1}[e]		
≥ 10.0	243 (41.3)	4,415 (42.0)
< 10.0	346 (58.7)	6,105 (58.0)
How many cigarettes per day		
≤ 15	278 (47.2)	5,960 (56.7)[a]
> 15	311 (52.8)	4,560 (43.3)
BMI in kg/m^2, median (range)	28.00 (17.55–49.31)	26.85 (15.42–58.49)[a]
BMI[f]		
< 30.0	357 (60.6)	7,505 (71.3)[a]
≥ 30.0	232 (39.4)	3,015 (28.7)
Waist-to-hip ratio, median (range)	0.810 (0.640–1.263)	0.807 (0.444–1.393)[a]
Age at menarche in years, median (range)	12 (≤ 9–≥ 17)	13 (≤ 9–≥ 17)[a]
Hysterectomy ever		
No	414 (70.3)	6,739 (64.1)[a]
Yes	175 (29.7)	3,781 (35.9)
Age at menopause in years, median (range)	50 (21–63)	50 (20–60)[a]
Age at menopause[c]		
< 47	152 (25.8)	3,207 (30.5)[a]
≥ 47	437 (74.2)	7,313 (69.5)
Oral contraceptive duration in years, median (range)	5.2 (0.1–21.0)	5.7 (0.1–47.0)[a]
Oral contraceptive duration[c]		
< 5.1	266 (45.2)	3,616 (34.4)[a]
≥ 5.1	323 (54.8)	6,904 (65.6)
Exogenous estrogen use (E-only) in years		
Never	451 (76.6)	7,360 (70.0)[a]
< 5	58 (9.8)	1,481 (14.1)
5 to < 10	18 (3.1)	546 (5.2)
≥ 10	62 (10.5)	1,133 (10.8)
Exogenous estrogen use (E+P) in years		
Never	454 (77.1)	8,681 (82.5)[a]
< 5	73 (12.4)	1,010 (9.6)
5 to < 10	30 (5.1)	434 (4.1)
10 to < 15	21 (3.6)	244 (2.3)
≥ 15	11 (1.9)	151 (1.4)

(Continued on the following column)

Abbreviations: MFA, monounsaturated fatty acids; PFA, polyunsaturated fatty acids.

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

[b] Depression scales were estimated using a short form of the Center for Epidemiologic Studies Depression Scale.

[c] Dietary alcohol per day, age at menopause, and oral contraceptive duration were stratified using the median values of 1.07 g/day, 47 years, and 5.1 years, respectively, as the cut-off points.

[d] % calories from SFA was stratified using 7% as the cutoff value, adherent 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).

[e] Physical 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 into two groups, with 10 METs as the cutoff according to current American College of Sports Medicine and American Heart Association recommendations (49).

[f] BMI was categorized using 30 kg/m^2, where 30.0 or higher falls within the obese range (https://www.cdc.gov/obesity/adult/defining.html).

among all samples. The initial data quality control process included SNPs with a missing-call rate of < 3% and a Hardy–Weinberg Equilibrium of $P \geq 10^{-4}$. The second quality control process included SNPs with $\hat{R}^2 \geq 0.6$ imputation quality (21) and excluded women with a kinship estimate with $\hat{R}^2 \geq 0.25$.

At baseline, fasting blood samples from each participant were collected by trained phlebotomists. Serum levels of glucose and insulin were measured by the hexokinase method on a Hitachi 747 instrument (Boehringer Mannheim Diagnostics) and by radioimmunoassay (Linco Research, Inc.), respectively, with average coefficients of variation of 1.28% and 10.93%, respectively. HOMA-IR was estimated as glucose (unit: mg/dL) × insulin (unit: μIU/mL)/405 (22).

Statistical analysis

Participants' baseline variables and allele frequencies stratified by breast cancer were examined via unpaired two-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 used. Multiple Cox proportional hazards regression, with an assumption test via a Schoenfeld residual plot and ρ evaluation, was conducted to obtain HRs and 95% confidence intervals (CI) for the single and combined effects of the most influential SNPs and lifestyle factors on breast cancer with adjustment for covariates (Table 1). For the gene–environment interaction, our previous GWA analysis was performed in strata defined by body mass index (BMI), metabolic equivalents (MET)·hours/week, and % calories from SFA, with respective cut-off values of 30 kg/m^2, 10 MET, and 7%. In this study, we evaluated the associations of the SNPs identified in the particular behavioral setting of obesity/physical inactivity/high-fat diet with breast cancer risk in the identical behavioral setting.

The RSF approach generates bootstrap samples from the original data and grows a tree from each bootstrapped sample, using a splitting rule applied to a tree node to maximize survival

Published OnlineFirst April 1, 2019; DOI: 10.1158/0008-5472.CAN-18-3688

differences across daughter nodes. This process is repeated numerous times (n = 5,000 trees in this study) to create a forest of trees (23, 24). Using an ensemble cumulative hazard estimate calculated from each tree and then averaged over all trees for each individual, we estimated a predicted cumulative incidence rate of breast cancer. The prediction parameter (i.e., prediction error interpreted as a misclassification probability) was created by using the out-of-bag (OOB) data (on average, 37% of the original data not used for bootstrapping) to calculate the OOB concordance index (c-index = 1 – prediction error), which is a measure of prediction performance (i.e., the probability of correctly classifying two cases) conceptually similar to the area under the receiver operating characteristic curve (23, 25). The importance of each variable was decided by two predicted values: (i) minimal depth (MD), where variables with a small MD split the tree close to the root and are considered highly predictive, and (ii) variable importance (VIMP), calculated as the difference between the OOB c-indexes from the original OOB data and from the permuted OOB data, where variables with larger VIMP are more predictive (15, 26).

A two-stage RSF analysis was conducted. In the first stage, we conducted an RSF on SNPs and lifestyle factors separately (Supplementary Fig. S3). Only those SNPs and lifestyle factors with significantly low MD and high VIMP values were selected for the second stage. During stage II, we performed another RSF with the selected SNPs and lifestyle factors from stage I. We took a multimodal approach: in overall, physical activity–stratified, and SFA-stratified subgroups, (i) estimating the values of MD and VIMP and comparing the two measures in the plot (Fig. 1A; Supplementary Fig. S4, S6, S8, and S10); (ii) generating the OOB c-index for the nested RSF model; and (iii) estimating the incremental error rate of each variable in the nested sequence of RSF models starting with the top variable and calculating a dropping error rate by the difference between the error rates from the nested sequence models. These approaches allow us to exclude the SNPs and lifestyle factors that may not have significant effects on breast cancer, resulting in more statistical power with the correct type I error rate in the stage II than the original RSF-based analysis (24). A two-tailed P value < 0.05 was considered statistically significant. R version 3.5.1 with survival, survivalROC, randomForestSRC, ggRandomForests, and gamlss packages was used.

Results

Participants' baseline characteristics and 58 SNPs that were previously identified in our GWA G × E study, stratified by breast cancer, are displayed (Table 1; Supplementary Table S1). Women with breast cancer were more likely to have higher education, higher family income, and family history of breast cancer; to consume more dietary alcohol/day and less % of calories from carbohydrates; to smoke more cigarettes/day; to be overall and abdominally obese; and to have experienced early menarche and late menopause. Patients with breast cancer also had shorter durations of oral contraceptive use (< 5 years) and exogenous estrogen (E)- only use, but a higher rate and longer duration of E + progestin (P) use.

Two-stage RSF to identify the most influential SNPs and behavioral variables in relation to breast cancer risk

With the 58 SNPs and 31 behavioral factors, we performed the two-stage RSF analysis to identify the most dominant variables with the highest predictive value and lowest prediction error for breast cancer risk. We used two predicted values, MD and VIMP measures. They use different prediction algorithms, so we expected the variable ranking to be somewhat different. In the first stage (Supplementary Fig. S3), we compared the two measures in a plot for each SNP and lifestyle and selected the strong predictive variables for cancer risk that were in agreement with high ranks: 12 of the 31 behavioral factors; 10 of the 58 SNPs in overall analysis; 7 and 10 of the 36 SNPs in MET ≥ 10 and < 10, respectively; and 2 and 5 of the 18 SNPs in calories from SFA < 7.0% and ≥ 7.0%, respectively.

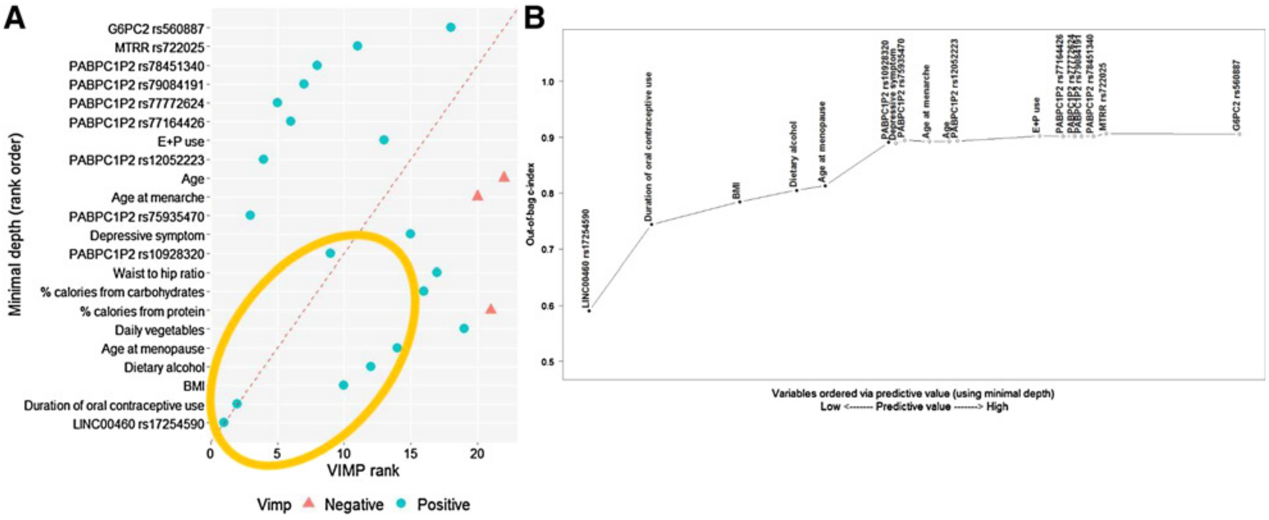

Figure 1.

Overall analysis: the second stage of RSF with 10 SNPs and 12 behavioral factors selected from the first stage of RSF analysis. **A,** Comparing minimal depth and VIMP rankings. **B,** Out-of-bag concordance index (c-index). (Improvement in out-of-bag c-index was observed when the top 6 variables [•] were added to the model, whereas other variables [○] did not further improve the accuracy of prediction).

Published OnlineFirst April 1, 2019; DOI: 10.1158/0008-5472.CAN-18-3688

Next, we performed the second RSF with the selected SNPs and 12 behavioral factors together in overall and subgroups to generate risk profiles with the most influential factors. Using a multimodal approach, we first estimated the values of MD and VIMP and plotted the two measures for comparison. Particularly, in the overall analysis plot (Table 2; Fig. 1A), the red dashed line indicates where the two measures were in agreement: both MD and VIMP indicated the following two SNPs and four behavioral factors as strong predictive markers of breast cancer risk: *LINC00460* rs17254590, *PABPC1P2* rs10928320, OC use, BMI, dietary alcohol, and age at menopause.

Second, we generated the OOB c-index using the nested RSF model. It ranks variables according to their predictive value estimated via MD. Results of the overall analysis (Fig. 1B) suggest that the top six variables improved the OOB c-index and thus had complementary predictive value, whereas the other variables did not significantly improve prediction accuracy.

We further calculated a dropping error rate of each variable in the nested sequence of RSF models (Table 2). By applying this complementary analysis with the aforementioned two approaches, we determined in the overall analysis the six variables that contributed the most to decreasing the error rate, and thus improving the prediction accuracy.

Table 2. Predictive values of variable in overall analysis from the second stage of random survival forest analysis

Variable[a]	Minimal depth[b]	VIMP	C-index	Error[c]	Drop error[d]
LINC00460 rs17254590	2.5218	0.0573	0.5907	0.4093	0.0907
Duration of oral contraceptive use	2.9940	0.0275	0.7437	0.2563	0.1531
BMI	3.6584	0.0079	0.7847	0.2153	0.0409
Dietary alcohol	4.0886	0.0067	0.8052	0.1948	0.0206
Age at menopause	4.3044	0.0025	0.8135	0.1865	0.0083
Daily vegetable	4.3096	0.0000	0.8178	0.1822	0.0043
% calories from protein	4.3910	-0.0001	0.8136	0.1864	-0.0042
% calories from carbohydrates	4.4212	0.0007	0.8169	0.1831	0.0033
Waist to hip ratio	4.4474	0.0005	0.8210	0.1790	0.0041
PABPC1P2 rs10928320	4.7834	0.0157	0.8910	0.1090	0.0700
Depressive symptom	4.8368	0.0009	0.8896	0.1104	-0.0014
PABPC1P2 rs75935470	4.9046	0.0271	0.8943	0.1057	0.0047
Age at menarche	5.0908	-0.0001	0.8927	0.1073	-0.0017
Age	5.2418	-0.0003	0.8925	0.1075	-0.0002
PABPC1P2 rs12052223	5.3036	0.0230	0.8934	0.1066	0.0009
E+P use	5.9216	0.0043	0.9024	0.0976	0.0090
PABPC1P2 rs77164426	6.1010	0.0174	0.9020	0.0980	-0.0005
PABPC1P2 rs77772624	6.1854	0.0178	0.9019	0.0981	0.0000
PABPC1P2 rs79084191	6.2392	0.0171	0.9014	0.0986	-0.0005
PABPC1P2 rs78451340	6.3298	0.0163	0.9017	0.0983	0.0003
MTRR rs722025	6.4268	0.0077	0.9066	0.0934	0.0048
G6PC2 rs560887	7.4328	0.0004	0.9059	0.0941	-0.0007

Abbreviations: C-index, concordance index; E+P, exogenous estrogen + progestin; VIMP, variable of importance.

[a] Variables are ordered by minimal depth.

[b] Predictive value of 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.

[c] The 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 1st, 2nd, and 3rd variables).

[d] The 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 (0.4093) from 0.5.

Consistently, in subgroup analyses, we applied the three approaches (agreement between MD and VIMP; OOB c-index; and contribution to dropping error rate) and determined the following SNPs and behavioral factors as the most predictive markers: (i) in the active group (MET ≥ 10; Supplementary Table S2; Supplementary Figs. S4 and S5), one SNP and seven lifestyles (*MKLN1* rs117911989; oral contraceptive use, estrogen + progestin (E+P) use, age at menopause, BMI, waist-to-hip ratio, dietary alcohol, and % calories from carbohydrates); (ii) in the inactive group (MET < 10; Supplementary Table S3; Supplementary Figs. S6 and S7), one SNP and six lifestyles (*MKLN1* rs117911989; oral contraceptive use, E+P use, age at menarche, BMI, waist-to-hip ratio, and dietary alcohol); (iii) in the low SFA intake group (calories from SFA < 7.0%; Supplementary Table S4; Supplementary Figs. S8 and S9), one SNP and two lifestyles (*LINC00460* rs17254590; oral contraceptive use and % calories from carbohydrates); and (iv) in the high SFA intake group (calories from SFA ≥ 7.0%; Supplementary Table S5; Supplementary Figs. S10 and S11), five SNPs, and three lifestyles (*LINC00460* rs17254590, and *PABPC1P2* rs75935470, rs12052223, rs78451340, and rs77164426; oral contraceptive use, age at menopause, and BMI).

Multivariate predictive model and combined effects of the most influential variables

Using the RSF model, the nonlinear effect of each predictive variable was accounted to estimate the cumulative incidence rate for breast cancer (Fig. 2A–H). The genotypes of SNPs were analyzed as a continuous variable. As shown in Fig. 2A–C, *LINC00460* rs17254590 GG, *PABPC1P2* rs10928320 TT, and *MKLN1* rs117911989 GG were considered risk genotypes and further analyzed as categorical variables. According to Fig. 2D, F, and G, using a cut-off value diverging each variable (similar to the median of each variable), the high-risk group was defined as < 5 years of oral contraceptive use, ≥ 47 years at menopause, or BMI ≥ 28 and analyzed as a binary variable. With the six most predictive variables in overall analysis, we developed a multivariate model predicting breast cancer risk (Supplementary Table S6), suggesting that the single effect of three lifestyles was significant even after adjusting for other covariates, while the single effect of two SNPs was not significant. We further estimated the single effect of other influential SNPs identified in the subgroup analyses (Supplementary Table S7–S9); no significant results were found.

However, the combined or joint effects of SNPs with lifestyles yielded different results (Supplementary Tables S3, S4, and S10). For example, in the active group [Table 3; one SNP (*MKLN1* rs117911989) and seven lifestyles], when stratified by E+P use, E+P ever-users with one risk genotype had a doubled risk for breast cancer than E+P never-users with null-risk genotype. Consistently, the high-risk lifestyle group (≥ 4 risk behaviors) of E+P ever-users had double the risk than the low-risk group (< 4 risk behaviors) of E+P never-users. When one SNP and seven lifestyles were combined, the high-risk group (≥ 4 risk behaviors and 1 risk genotype) had 90% excess risk for cancer than the low-risk group (< 4 risk behaviors and null-risk genotype), suggesting a cumulative effect of genetic and lifestyle factors in both additive and multiplicative interaction models (effect size and P for genes × lifestyles interaction test = 2.06 and 0.273, respectively). When

Published OnlineFirst April 1, 2019; DOI: 10.1158/0008-5472.CAN-18-3688

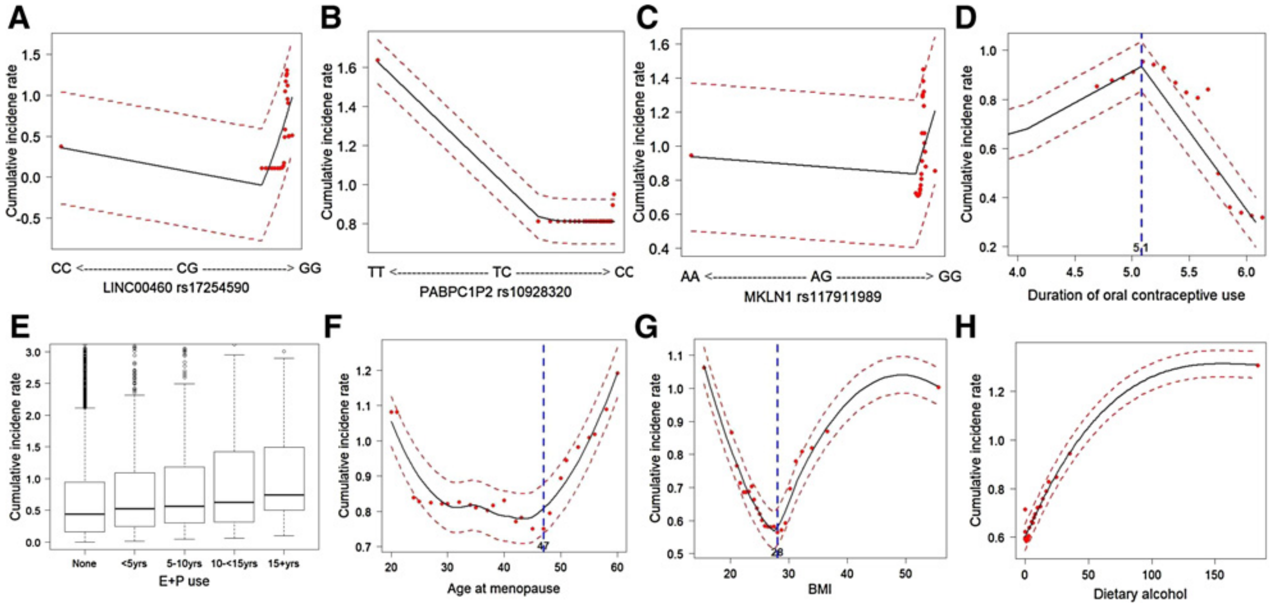

Figure 2.

Cumulative breast cancer incidence rate for the 8 most influential variables (3 SNPs and 5 behavioral factors) based on a random survival forest analysis. Dashed red lines indicate 95% confidence intervals.

Table 3. Physical activity–stratified analysis: combined effect of risk genotype of *MKLN1* rs117911989 GG and 7 behavioral factors (oral contraceptive use, E+P use, age at menopause, BMI, waist-to-hip ratio, dietary alcohol, and in active group only, % calories from carbohydrates) on breast cancer risk

Total			Never use of E + P			E + P ever use		
n[a]	HR[b] (95% CI)	*P*	*n*	HR[b] (95% CI)	*P*	*n*	HR[b] (95% CI)	*P*
Active Group (MET ≥ 10) (*n* = 4,658)								
Risk genotype								
0	reference		311	reference		62	2.00 (0.64–6.31)	0.235
1	1.33 (0.79–2.25)	0.284	3,409	1.37 (0.74–2.53)	0.312	876	**2.31 (1.21–4.39)**	**0.011**
Behavioral factors[c]								
0	reference		2,974	reference		777	**1.83 (1.33–2.53)**	**<0.001**
1	**1.62 (1.24–2.12)**	**<0.001**	746	**1.67 (1.19–2.33)**	**0.003**	161	**2.25 (1.29–3.94)**	**0.004**
Risk genotypes combined with behavioral factors[d]								
0	reference		69	reference		234	1.97 (0.62–6.27)	0.254
1	1.09 (0.61–1.97)	0.763	812	1.03 (0.54–1.97)	0.925	2,706	1.89 (0.96–3.72)	0.067
2	**1.86 (1.01–3.41)**	**0.047**	554	1.87 (0.94–3.70)	0.073	283	**2.49 (1.10–5.64)**	**0.029**
Inactive Group (MET < 10) (*n* = 6,451)								
Risk genotype								
0	reference		426	reference		78	1.00 (0.34–2.89)	0.997
1	0.96 (0.65–1.41)	0.824	4,989	0.93 (0.61–1.42)	0.747	958	1.12 (0.69–1.81)	0.647
Behavioral factors[c]								
0	reference		4,319	reference		868	1.04 (0.75–1.45)	0.810
1	**1.68 (1.35–2.10)**	**<0.001**	1,096	**1.49 (1.14–1.93)**	**0.003**	168	**2.13 (1.32–3.43)**	**0.002**
Risk genotypes combined with behavioral factors[d]								
0	reference		337	reference		71	1.16 (0.39–3.48)	0.789
1	1.02 (0.63–1.66)	0.925	4071	1.04 (0.63–1.73)	0.878	804	1.04 (0.58–1.87)	0.891
2	**1.67 (1.01–2.75)**	**0.044**	1,007	1.48 (0.86–2.54)	0.159	161	**2.25 (1.15–4.41)**	**0.018**

NOTE: Numbers in boldface are statistically significant.

Abbreviations: E+P, exogenous estrogen + progestin.

[a] The *n* indicates the cumulative number of risk genotypes or behavioral factors.

[b] Multivariate regression for behavioral factors was adjusted by age, depressive symptom, age at menarche, daily vegetables, % calories from protein, and % calories from carbohydrates (in inactive group); and in risk genotype analysis, 6 additional behavioral factors [oral contraceptive use, age at menopause, BMI, waist-to-hip ratio, dietary alcohol, and E+P use (in total analysis)] were added as covariates.

[c] The number of behavioral factors defined as 0 (low risk: ≤ 3 risk behaviors) and 1 (high risk: ≥ 4 risk behaviors).

[d] The combined number of risk genotypes and behavioral factors were estimated based on risk genotypes defined as 0 (low risk: none) and 1 (high risk: 1 risk allele) and behavioral factors defined as 0 (low risk: ≤ 3 risk behaviors) and 1 (high risk: ≥ 4 risk behaviors). The ultimate number of risk genotypes combined with behavioral factors defined as 0 (low risk of genotypes and behaviors), 1 (high risk of either genotypes or behaviors), and 2 (high risk of both genotypes and behaviors).

Published OnlineFirst April 1, 2019; DOI: 10.1158/0008-5472.CAN-18-3688

Table 4. SFA-stratified analysis: combined effect of risk genotypes of *LINC00460* rs17254590 GG, *PABPC1P2* rs75935470 TT, *PABPC1P2* rs12052223 GG, *PABPC1P2* rs78451340 GG, and *PABPC1P2* rs77164426 AA and 3 behavioral factors (oral contraceptive use; in low SFA intake group, % calories from carbohydrates; and in high SFA intake group, age at menopause and BMI) on breast cancer risk

Total			Oral contraceptive use ≥ 5 years			Oral contraceptive use < 5 years		
n[a]	HR[b] (95% CI)	*P*	*n*	HR[b] (95% CI)	*P*	*n*	HR[b] (95% CI)	*P*
% calories from SFA < 7.0 % (*n* = 1,010)								
Risk genotype								
0	reference		440	reference		338	**5.36 (2.44–11.78)**	**<0.001**
1	1.03 (0.52–2.01)	0.940	112	2.41 (0.79–7.32)	0.122	120	**3.60 (1.25–10.32)**	**0.017**
Behavioral factors[c]								
0	reference		196	reference		169	3.04 (0.91–10.17)	0.070
1	**3.47 (1.91–6.30)**	**<0.001**	356	1.38 (0.41–4.59)	0.602	289	5.75 (1.90–17.41)	0.002
Risk genotypes combined with behavioral factors[c]								
0	reference		475	reference		389	**4.48 (2.10–9.57)**	**<0.001**
1	**2.31 (1.00–5.33)**	**0.049**	77	2.64 (0.81–8.61)	0.107	69	**6.29 (2.27–17.45)**	**<0.001**
% calories from SFA ≥ 7.0 % (*n* = 10,099)								
Risk genotype[d]								
0	reference		75	reference		39	0.49 (0.06–4.43)	0.528
1	1.44 (0.59–3.48)	0.421	5,413	0.97 (0.36–2.61)	0.953	2,739	1.65 (0.61–4.46)	0.322
2	1.42 (0.57–3.50)	0.448	1,054	0.99 (0.36–2.77)	0.992	779	1.53 (0.55–4.26)	0.417
Behavioral factors[e]								
0	reference		825	reference		340	**2.72 (1.46–5.06)**	**0.002**
1	**2.13 (1.36–3.35)**	**0.001**	3,086	**1.60 (1.00–2.57)**	**0.051**	1,812	**2.97 (1.84–4.79)**	**<0.001**
2	**3.08 (1.89–5.01)**	**<0.001**	2,631	**2.28 (1.43–3.66)**	**<0.001**	1,405	**3.22 (1.97–5.24)**	**<0.001**
Risk genotypes combined with behavioral factors[e]								
0	reference		3,280	reference		1,682	**2.00 (1.54–2.59)**	**<0.001**
1	**1.24 (1.02–1.50)**	**0.029**	2,839	**1.43 (1.12–1.83)**	**0.004**	1,566	**2.13 (1.63–2.79)**	**<0.001**
2	1.37 (0.87–2.16)	0.179	423	1.54 (0.98–2.42)	0.060	309	**2.00 (1.23–3.24)**	**0.005**

NOTE: Numbers in boldface are statistically significant.

Abbreviation: SFA, saturated fatty acids.

[a]The *n* indicates the cumulative number of risk genotypes or behavioral factors.

[b]Multivariate regression for behavioral factors was adjusted by age, depressive symptom, age at menarche, daily vegetables, % calories from protein, waist-to-hip ratio, dietary alcohol, exogenous estrogen + progestin, age at menopause and BMI (in low SFA-intake group), and % calories from carbohydrates (in high SFA-intake group); and in risk-genotype analysis, 1 additional behavioral factor (oral contraceptive use in total analysis) was added as covariates.

[c]Low SFA intake group: the number of behavioral factors defined as 0 (low risk: 0 or 1 risk behavior) and 1 (high risk: 2 risk behaviors). The combined number of risk genotypes and behavioral factors were estimated based on risk genotypes defined as 0 (low risk: none) and 1 (high risk: 1 risk allele) and behavioral factors defined as 0 (low risk: 0 or 1 risk behavior) and 1 (high risk: 2 risk behaviors). The ultimate number of risk genotypes combined with behavioral factors defined as 0 (none of risk or high risk of either genotypes or behaviors) and 1 (high risk of both genotypes and behaviors).

[d]The number of risk genotypes defined as 0 (none), 1 (moderate risk: ≤ 4 risk alleles), and 2 (high risk: 5 risk alleles).

[e]High SFA intake group: the number of behavioral factors defined as 0 (low risk: none), 1 (moderate risk: ≤ 2 risk behaviors), and 2 (high risk: 3 risk behaviors). On the basis of risk genotypes (0: low/moderate risk with ≤ 4 risk alleles and 1: high risk with 5 risk alleles) and behavioral factors (0: low/moderate risk with ≤ 2 risk behaviors and 1: high risk with 3 risk behaviors), the ultimate number of risk genotypes combined with behavioral factors defined as 0 (low/moderate risk of both genotypes and behaviors), 1 (high risk of either genotypes or behaviors), and 2 (high risks of both genotypes and behaviors).

stratified by E+P use, E+P ever-users with high risk by both lifestyles and genotype had 2.5 times greater risk, compared with E+P never-users with low risk by both lifestyles and genotype. This indicates a gene and lifestyle dose–response relationship and a significant joint effect of E+P use with SNP and lifestyles on cancer risk. The inactive group analysis (Table 3) produced similar results.

Interestingly, the SFA-stratified analyses yielded a stronger combined effect of SNPs and lifestyles (Table 4). Specifically, in the low-SFA group [one SNP (*LINC00460* rs17254590) and two lifestyles], when stratified by the duration of oral contraceptive use, women who used < 5 years with one risk genotype had 3.6 times higher risk for breast cancer than women who used ≥ 5 years with null-risk genotype. Similarly, women using oral contraceptives for a shorter period with one risk lifestyle had 5.8 times greater risk for cancer than those using for a longer period with null-risk lifestyle. When one SNP and one lifestyle were combined and stratified by oral contraceptive use, shorter oral contraceptive users with both one risk genotype and one lifestyle had 6.3 times higher risk than longer oral contraceptive users with either a risk genotype or risk lifestyle. This suggests the combined effect of SNP and lifestyles in both additive and multiplicative interaction models (effect size and *P* for genes × lifestyles interaction = 1.10 and 0.887, respectively) and the joint effect of these factors with oral contraceptive use on breast cancer risk. The high-SFA group (Table 4) provided similar but attenuated combined results.

On the basis of a RSF model using the strongest variables (*MKLN1* rs117911989, *LINC00460* rs17254590, oral contraceptive use, and E+P use), we further constructed contour plots to provide the cumulative incidence rate for different combinations of SNP and hormone use by physical activity (Fig. 3A and B) and SFA intake (Supplementary Fig. S12), yielding consistent results.

Discussion

Understanding how lifestyle factors modify and interact with genes and phenotypes, affecting breast cancer risk, and further incorporating both genetic and lifestyle factors to generate risk profiles for breast cancer is important for developing a gene–behavior tool to use in primary breast cancer prevention efforts. Our two-stage multimodal RSF approach identified the most predictive genetic and lifestyle variables in overall and subgroup analyses (stratified by a well-established effect modifier such as

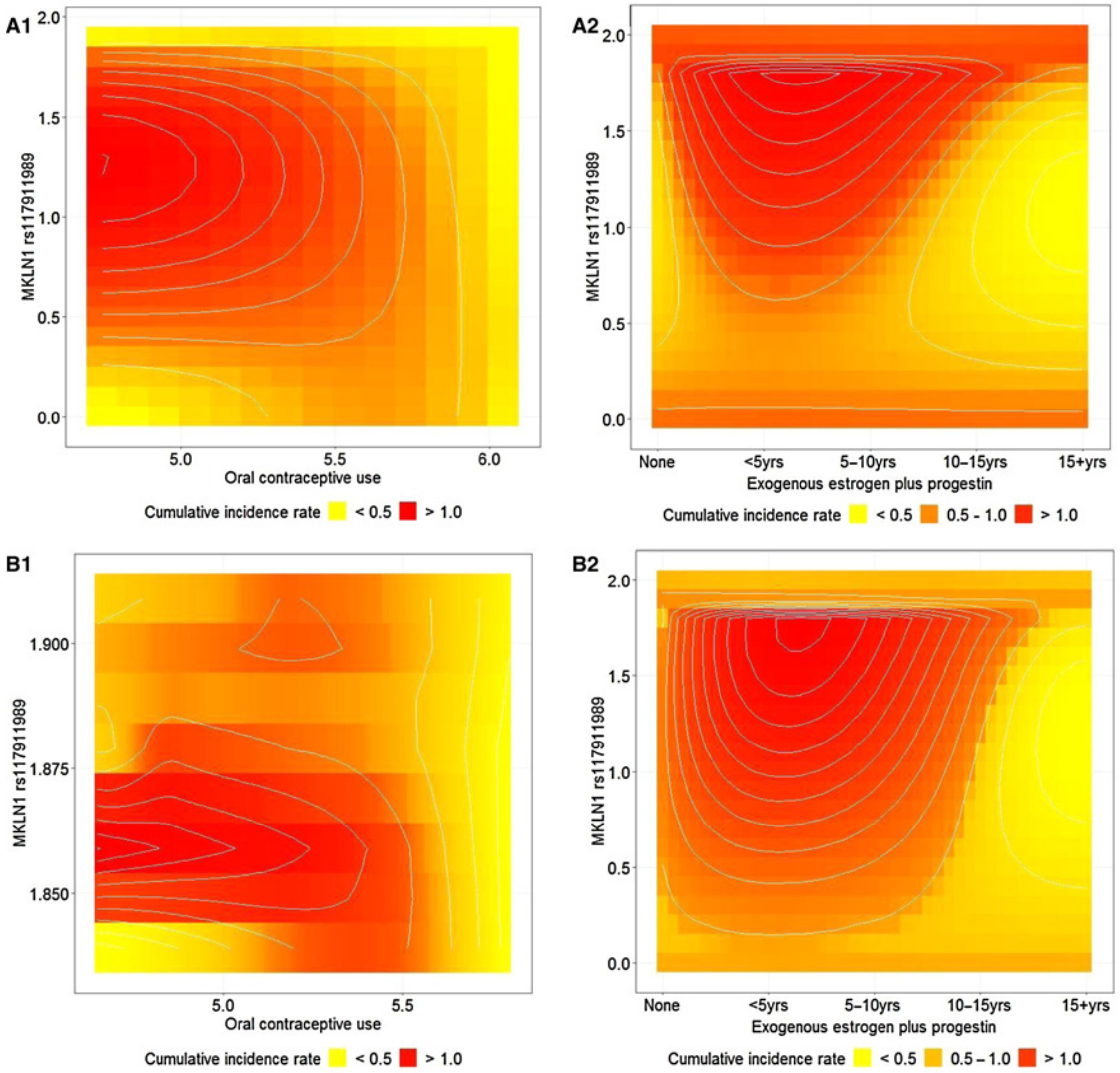

Figure 3.

Contour plot of cumulative breast cancer incidence rate for the combination of SNP (*MKLN1* rs117911989) and oral contraceptive use or E+P use, stratified by physical activity. Cumulative incidence rate estimated from the random survival forest model was adjusted by age, BMI, waist-to-hip ratio, depressive symptom, age at menarche, age at menopause, dietary alcohol, daily vegetables, % calories from protein, and % calories from carbohydrates. **A1,** Active Group (MET ≥ 10); **A2,** Active Group (MET ≥ 10); **B1,** Inactive Group (MET < 10); **B2,** Inactive Group (MET < 10).

BMI, exercise, and dietary fat intake; refs. 19, 27–29). Two SNPs (*LINC00460* rs17254590 and *MKLN1* rs117911989), lifestyle factors related to lifetime cumulative exposure to estrogen (oral contraceptive use, E+P use, and older age at menopause), BMI, and dietary alcohol consumption were the most common influential factors across the analyses. With those strongest variables, we constructed overall and within-subgroup risk profiles for breast cancer. In the individual SNP analyses, no significant associations were observed, but the combination of the SNPs and lifestyle factors synergistically increased the risk of breast cancer.

One SNP in *LINC00460*, in relation to IR phenotypes, by interacting with SFA intake, is associated with increased risk for breast cancer. *LINC00460* is long intergenic noncoding RNA (lncRNA) 460 (30). Several lncRNAs are involved in tumorigenesis via regulating oncogenes or tumor-suppressive genes' expression (31). Recently, cancer-related lncRNA *LINC00460* has been found in association with nasopharyngeal cancer (NPC; ref. 30). It was significantly upregulated in NPC tissues, and silencing of *LINC00460* repressed NPC cell proliferations, suggesting its function as an oncogene. miR-149 repressed tumor-suppressive miRNA, dysregulating AKT1 and cyclin D1 in cellular pathways (32). *LINC00460* produces its effect through sponging the miR-149-5p and then activating the *IL6* gene, which promotes cell proliferation, migration, and invasion (33). Our study is the first to report that this lncRNA is associated with breast cancer risk. In addition, *LINC00460* was associated with subcutaneous adipose tissue in a previous GWA study (34), supporting our finding of its associations with IR phenotypes and breast cancer observed in fatty acids strata.

Published OnlineFirst April 1, 2019; DOI: 10.1158/0008-5472.CAN-18-3688

MKLN1 is an intracellular protein that mediates cell responses to the extracellular matrix, influencing cell adhesion and cytoskeleton organization (35, 36). It has been associated with pancreatic (36) and lung cancer (37) and is a novel marker for cardiovascular risk (38). It has also been associated with type II diabetes (39). Our findings of its association with IR phenotypes are consistent with previous results, but our study newly reports the association of *MKLN1* with breast cancer risk. This association would have been missed without the incorporation of the physical activity factor, which will require further biologic functional study.

Because lifetime cumulative exposure to estrogen is a key factor for breast cancer risk, it is not surprising to find that oral contraceptive use was an important predictor in this study. Previous results for past oral contraceptive use in postmenopausal women in relation to breast cancer risk are conflicting: no associations (40–42) and slightly or modestly increased risk with longer duration of oral contraceptive use (29, 43). An *in vivo* study reported the use of oral contraceptive (especially combined E+P) increased the proliferation of human breast epithelial cells (44). The previous mixed findings may partly result from a lack of consideration of the duration of oral contraceptive use by accounting for its nonlinear effect. Our RSF analysis showed that the cumulative breast cancer incidence rate increases with up to 5 years of oral contraceptive use, but drops thereafter. According to previous studies reporting a higher risk for breast cancer only in active and recent oral contraceptive users (40, 42, 44), our findings may be confounded by the recency of use. In addition, because progestin formulations in oral contraceptive have changed, earlier preparation could have a different effect on cancer risk than those currently used. However, we had no data on the recency and the type of oral contraceptive preparation that our participants had taken, thus warranting future studies that examine the potential different effect on cancer risk according to time lags since last use and specific oral contraceptive configuration.

Using the cut-off value of oral contraceptive use (5 years), we further examined the combined effect of SNPs and lifestyles within the strata, suggesting the joint effect of the genetic and lifestyle factors with the duration of oral contraceptive use on breast cancer. Moreover, the joint effect was attenuated in high-SFA intake group, which may support potential trade-off pathways between sex hormones and fatty acids (i.e., the effect of estrogen levels minimized with high fatty acid levels).

Another strong exogenous factor we found that contributes to the women's lifetime exposure to estrogen is the use of E+P, a well-known risk factor for breast cancer (44–46). Synthetic progestin differs structurally from natural progesterone, resulting in different actions at the cellular level, such as cell proliferation and antiapoptosis by having an affinity for androgen, glucocorticoid, and mineralocorticoid receptors (44, 47). Furthermore, the joint effect of gene and lifestyles with E+P use on breast cancer was attenuated in an inactive group, implicating obesity–sex hormone pathways (48); that is, in obese women who have relatively higher levels of estrogen, the effect of E+P use can be reduced.

Our study population was confined to non-Hispanic white postmenopausal women, so the generalizability of our findings to other populations is limited. Also, owing to insufficient statistical power, we did not examine any breast cancer molecular subtypes. A two-stage RSF provides greater statistical power to identify the most predictive variables for breast cancer risk. Despite that benefit, it can over-fit the model due to noisy tasks with a relatively small sample size, so our results need to be replicated in independent studies with a large sample size.

This study suggests that IR SNPs identified at the GWA level interact with lifestyle factors, including exogenous lifetime exposures to estrogen, obesity, and dietary alcohol, to influence risk for breast cancer. The identified SNPs in combination with those lifestyles have a possible synergistic effect on breast cancer risk, which calls for further biologic mechanism studies such as gene regulation and aberrant cell signaling in relation to breast cancer cells of obese women with a history of estrogen use and alcohol intake. Our findings may contribute to greater accuracy in predicting breast cancer and suggest intervention strategies for the women who carry the risk genotypes, such as a shorter duration of exogenous estrogen use and reduced body weight and alcohol intake, which may lead to reduced potential impact of such risk factors on the epigenome and thus reduce their risk for breast cancer.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors' Contributions

Conception and design: S.Y. Jung, J.C. Papp, Z.-F. Zhang
Development of methodology: S.Y. Jung
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): S.Y. Jung
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): S.Y. Jung, E.M. Sobel, H. Yu, Z.-F. Zhang
Writing, review, and/or revision of the manuscript: S.Y. Jung, J.C. Papp, E.M. Sobel, H. Yu, Z.-F. Zhang
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): S.Y. Jung

Acknowledgments

S.Y. Jung was supported for this study by 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, HHSN268201100004C, 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).

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.

Received November 26, 2018; revised February 13, 2019; accepted March 25, 2019; published first April 1, 2019.

Published OnlineFirst April 1, 2019; DOI: 10.1158/0008-5472.CAN-18-3688

References

1. Weichhaus M, Broom J, Wahle K, Bermano G. A novel role for insulin resistance in the connection between obesity and postmenopausal breast cancer. Int J Oncol 2012;41:745–52.
2. Clayton PE, Banerjee I, Murray PG, Renehan AG. Growth hormone, the insulin-like growth factor axis, insulin and cancer risk. Nat Rev Endocrinol 2011;7:11–24.
3. Calle EE, Kaaks R. Overweight, obesity and cancer: epidemiological evidence and proposed mechanisms. Nat Rev Cancer 2004;4:579–91.
4. Sieri S, Muti P, Claudia A, Berrino F, Pala V, Grioni S, et al. Prospective study on the role of glucose metabolism in breast cancer occurrence. Int J Cancer 2012;130:921–9.
5. Kabat GC, Kim M, Caan BJ, Chlebowski RT, Gunter MJ, Ho GY, et al. Repeated measures of serum glucose and insulin in relation to postmenopausal breast cancer. Int J Cancer 2009;125:2704–10.
6. Gunter MJ, Hoover DR, Yu H, Wassertheil-Smoller S, Rohan TE, Manson JE, et al. Insulin, insulin-like growth factor-I, and risk of breast cancer in postmenopausal women. J Natl Cancer Inst 2009;101:48–60.
7. Rose DP, Vona-Davis L. The cellular and molecular mechanisms by which insulin influences breast cancer risk and progression. Endocr Relat Cancer 2012;19:R225–241.
8. Booth FW, Roberts CK, Laye MJ. Lack of exercise is a major cause of chronic diseases. Compr Physiol 2012;2:1143–211.
9. Wasserman L, Flatt SW, Natarajan L, Laughlin G, Matusalem M, Faerber S, et al. Correlates of obesity in postmenopausal women with breast cancer: comparison of genetic, demographic, disease-related, life history and dietary factors. Int J Obesity Related Metab Disord 2004;28:49–56.
10. Franks PW, Mesa JL, Harding AH, Wareham NJ. Gene-lifestyle interaction on risk of type 2 diabetes. Nutr Metab Cardiovasc Dis 2007;17:104–24.
11. Creighton CJ, Sada YH, Zhang Y, Tsimelzon A, Wong H, Dave B, et al. A gene transcription signature of obesity in breast cancer. Breast Cancer Res Treat 2012;132:993–1000.
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. Scannell Bryan M, Argos M, Andrulis IL, Hopper JL, Chang-Claude J, Malone KE, et al. Germline variation and breast cancer incidence: a gene-based association study and whole-genome prediction of early-onset breast cancer. Cancer Epidemiol Biomarkers Prev 2018;27:1057–64.
14. Jung SY, Mancuso N, Yu H, Papp J, Sobel EM, Zhang ZF. Genome-wide meta-analysis of gene-environmental interaction for insulin-resistance phenotypes and breast cancer risk in postmenopausal women. Cancer Prev Res 2019;12:31–42.
15. Mogensen UB, Ishwaran H, Gerds TA. Evaluating random forests for survival analysis using prediction error curves. J Statist Software 2012; 50:1–23.
16. Hamidi O, Poorolajal J, Farhadian M, Tapak L. Identifying important risk factors for survival in kidney graft failure patients using random survival forests. Iranian J Public Health 2016;45:27–33.
17. 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.
18. 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.
19. Pfeiffer RM, Webb-Vargas Y, Wheeler W, Gail MH. Proportion of U.S. trends in breast cancer incidence attributable to long-term changes in risk factor distributions. Cancer Epidemiol Biomarkers Prev 2018;27: 1214–22.
20. National Cancer Institute. SEER Program: Comparative Staging Guide For Cancer, 1993. https://seer.cancer.gov/archive/manuals/historic/comp_stage1.1.pdf.
21. 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.
22. 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.
23. Ishwaran H, Kogalur UB. Random Survival Forests for R; 2007. Available from: https://pdfs.semanticscholar.org/951a/84f0176076fb6786fdf43320e8b27094dcfa.pdf.
24. Chung RH, Chen YE. A two-stage random forest-based pathway analysis method. PLoS One 2012;7:e36662.
25. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. The Annals of Applied Statistics 2008;2:841–60. doi: 10.1214/08-AOAS169.
26. Inuzuka R, Diller GP, Borgia F, Benson L, Tay EL, Alonso-Gonzalez R, et al. Comprehensive use of cardiopulmonary exercise testing identifies adults with congenital heart disease at increased mortality risk in the medium term. Circulation 2012;125:250–9.
27. Rice MS, Eliassen AH, Hankinson SE, Lenart EB, Willett WC, Tamimi RM. Breast cancer research in the Nurses' Health Studies: exposures across the life course. Am J Public Health 2016;106:1592–8.
28. Memon ZA, Qurrat ul A, Khan R, Raza N, Noor T. Clinical presentation and frequency of risk factors in patients with breast carcinoma in Pakistan. Asian Pacific J Cancer Prev 2015;16:7467–72.
29. Rieck G, Fiander A. The effect of lifestyle factors on gynaecological cancer. Best practice & research. Clin Obstet Gynaecol 2006;20:227–51.
30. 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 2018;639:77–84.
31. 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.
32. 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.
33. 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.
34. 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.
35. Adams JC, Seed B, Lawler J. Muskelin, a novel intracellular mediator of cell adhesive and cytoskeletal responses to thrombospondin-1. EMBO J 1998; 17:4964–74.
36. Wolpin BM, Rizzato C, Kraft P, Kooperberg C, Petersen GM, Wang Z, et al. Genome-wide association study identifies multiple susceptibility loci for pancreatic cancer. Nat Genet 2014;46:994–1000.
37. McKay JD, Hung RJ, Han Y, Zong X, Carreras-Torres R, Christiani DC, et al. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. Nat Genet 2017;49:1126–32.
38. Warren HR, Evangelou E, Cabrera CP, Gao H, Ren M, Mifsud B, et al. Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk. Nat Genet 2017;49: 403–15.
39. Rouillard AD, Gundersen GW, Fernandez NF, Wang Z, Monteiro CD, McDermott MG, et al. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. Database 2016;2016:pii:baw100.
40. Dumeaux V, Fournier A, Lund E, Clavel-Chapelon F. Previous oral contraceptive use and breast cancer risk according to hormone replacement therapy use among postmenopausal women. Cancer Causes Control 2005; 16:537–44.
41. Marchbanks PA, McDonald JA, Wilson HG, Folger SG, Mandel MG, Daling JR, et al. Oral contraceptives and the risk of breast cancer. N Engl J Med 2002;346:2025–32.
42. Bhupathiraju SN, Grodstein F, Stampfer MJ, Willett WC, Hu FB, Manson JE. Exogenous hormone use: oral contraceptives, postmenopausal hormone therapy, and health outcomes in the Nurses' Health Study. Am J Public Health 2016;106:1631–7.
43. Van Hoften C, Burger H, Peeters PH, Grobbee DE, Van Noord PA, Leufkens HG. Long-term oral contraceptive use increases breast cancer risk in women over 55 years of age: the DOM cohort. Int J Cancer 2000; 87:591–4.

Published OnlineFirst April 1, 2019; DOI: 10.1158/0008-5472.CAN-18-3688

44. Cogliano V, Grosse Y, Baan R, Straif K, Secretan B, El Ghissassi F. Carcinogenicity of combined oestrogen-progestagen contraceptives and menopausal treatment. Lancet Oncol 2005;6:552–3.
45. Gartlehner G, Patel SV, Feltner C, Weber RP, Long R, Mullican K, et al. Hormone therapy for the primary prevention of chronic conditions in postmenopausal women: evidence report and systematic review for the US Preventive Services Task Force. JAMA 2017; 318:2234–49.
46. Rossouw JE, Anderson GL, Prentice RL, LaCroix AZ, Kooperberg C, Stefanick ML, et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results From the Women's Health Initiative randomized controlled trial. JAMA 2002;288: 321–33.
47. Asi N, Mohammed K, Haydour Q, Gionfriddo MR, Vargas OLM, Prokop LJ, et al. Progesterone vs. synthetic progestins and the risk of breast cancer: a systematic review and meta-analysis. Syst Rev 2016; 5:121.
48. Munsell MF, Sprague BL, Berry DA, Chisholm G, Trentham-Dietz A. Body mass index and breast cancer risk according to postmenopausal estrogen-progestin use and hormone receptor status. Epidemiol Rev 2014;36: 114–36.
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–e529.

Published OnlineFirst April 1, 2019; DOI: 10.1158/0008-5472.CAN-18-3688

Cancer Research

The Journal of Cancer Research (1916–1930) | The American Journal of Cancer (1931–1940)

AACR American Association for Cancer Research

Breast Cancer Risk and Insulin Resistance: Post Genome-Wide Gene–Environment Interaction Study Using a Random Survival Forest

Su Yon Jung, Jeanette C. Papp, Eric M. Sobel, et al.

Cancer Res 2019;79:2784-2794. Published OnlineFirst April 1, 2019.