Published OnlineFirst February 25, 2011; DOI: 10.1158/1055-9965.EPI-10-0675

Research Article

Cancer Epidemiology, Biomarkers & Prevention

# Genotype–Environment Interactions in Microsatellite Stable/Microsatellite Instability-Low Colorectal Cancer: Results from a Genome-Wide Association Study

Jane C. Figueiredo[1], Juan Pablo Lewinger[1], Chi Song[1], Peter T. Campbell[3], David V. Conti[1], Christopher K. Edlund[2], Dave J. Duggan[4], Jagadish Rangrej[5], Mathieu Lemire[5], Thomas Hudson[5], Brent Zanke[8], Michelle Cotterchio[6], Steven Gallinger[7], Mark Jenkins[9], John Hopper[9], Robert Haile[1], Polly Newcomb[10], John Potter[10], John A. Baron[11], Loic Le Marchand[12], and Graham Casey[1]

## Abstract

**Background:** Genome-wide association studies (GWAS) have led to the identification of a number of common susceptibility loci for colorectal cancer (CRC); however, none of these GWAS have considered gene–environment (G × E) interactions. Therefore, it is unclear whether current hits are modified by environmental exposures or whether there are additional hits whose effects are dependent on environmental exposures.

**Methods:** We conducted a systematic search for G × E interactions using genome wide data from the Colon Cancer Family Registry that included 1,191 cases of microsatellite stable (MSS) or microsatellite instability–low (MSI-L) CRC and 999 controls genotyped using either the Illumina Human1M or Human1M-Duo BeadChip. We tested for interactions between genotypes and 14 environmental factors using 3 methods: a traditional case–control test, a case-only test, and the recently proposed 2-step method by Murcray and colleagues. All potentially significant findings were replicated in the ARCTIC Study.

**Results:** No G × E interactions were identified that reached genome-wide significance by any of the 3 methods. When analyzing previously reported susceptibility loci, 7 significant G × E interactions were found at a 5% significance level. We investigated these 7 interactions in an independent sample and none of the interactions were replicated.

**Conclusions:** Identifying G × E interactions will present challenges in a GWAS setting. Our power calculations illustrate the need for larger sample sizes; however, as CRC is a heterogeneous disease, a tradeoff between increasing sample size and heterogeneity needs to be considered.

**Impact:** The results from this first genome-wide analysis of G × E in CRC identify several challenges, which may be addressed by large consortium efforts. *Cancer Epidemiol Biomarkers Prev; 20(5); 758–66. ©2011 AACR.*

## Introduction

Recently, several genome-wide association studies (GWAS) have led to the identification and replication of a number of susceptibility loci for CRC (1–6). Incorporating environmental exposures into GWAS data may aid in the identification of additional susceptibility alleles that would be otherwise masked by heterogeneity in subgroups, and would also clarify whether certain environmental exposures may modulate risk in susceptible individuals. However, there are limited data on the interaction between other susceptibility alleles and environmental risk factors for CRC. To date, no studies have examined the interaction between a wide range of environmental factors and genome-wide genotype data with respect to cancer risk.

Detecting gene–environment (G × E) interactions using a standard case–control test is challenging in a genome-wide context because of the stringent significance level required to adjust for multiple testing and because only weak G × E interactions are expected. The case-only test is known to be more powerful than the case–control test but in the presence of population level G × E association it can yield a severely inflated type I error (7). Recently, new methods to test for G × E interactions in GWAS have

**Authors' Affiliations:** [1]Department of Preventive Medicine, Keck School of Medicine, University of Southern California; [2]USC Epigenome Center, University of Southern California, Los Angeles, California; [3]Department of Epidemiology, American Cancer Society, Atlanta, Georgia; [4]Translational Genomics Research Institute, Phoenix, Arizona; [5]Ontario Institute for Cancer Research, The MaRS Center; [6]Cancer Care Ontario; [7]Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto; [8]The University of Ottawa Faculty of Medicine, Ottawa, Ontario, Canada; [9]School of Population Health, University of Melbourne, Victoria, Australia; [10]Fred Hutchinson Cancer Research Center, Seattle, Washington; [11]Departments of Medicine and Community and Family Medicine, Dartmouth Medical School, Lebanon, New Hampshire; and [12]Cancer Research Center of Hawaii, University of Hawaii, Honolulu, Hawaii

**Corresponding Author:** Jane C. Figueiredo, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, 1450 Biggy Street, room 1509J, Los Angeles, CA 90033. Phone: 3234427752; Fax: 3234427787; E-mail: janefigu@usc.edu.

**doi:** 10.1158/1055-9965.EPI-10-0675

Published OnlineFirst February 25, 2011; DOI: 10.1158/1055-9965.EPI-10-0675



been proposed. Murcray and colleagues introduced an efficient 2-step approach that is carried out independently of any initial scans for main effects (8). The method expands on the traditional test for G × E interaction in a case–control study by incorporating a preliminary screening step constructed to efficiently use all available information. This method has been shown to be more powerful for a wide range of environmental exposures, minor allele frequencies, and genetic effects compared to the traditional 1-step test (8). In this study, we take advantage of these methodologies to systematically search for G × E interactions within a GWAS of MSS (microsatellite stable)/MSI-L (microsatellite instability–low) CRC from the Colon Cancer Family Registry considering lifestyle and environmental exposures known to be involved in the etiology of CRC.

## Methods

### Study design and sample

Participants included in this analysis were recruited from 3 population-based registries based at the Fred Hutchinson Cancer Research Center (FHCRC, Seattle, WA), Cancer Care Ontario [Ontario Familial Colorectal Cancer Family Registry (OFCCR), Toronto, Canada], and the University of Melbourne (Victoria, Australia), which recruited families from both Australia and New Zealand as part of the Colon Cancer Family Registry (Colon CFR; ref. 9).

Cases from these registries met the following eligibility criteria: invasive CRC; self-identified as non-Hispanic White; no identified germline mutations in mismatch repair (MMR) genes; MSS or MSI-L CRC, and/or MMR protein immunohistochemistry positive determined using standard methods (10). All cases meeting these criteria and under age 50 or who had an affected first-degree relative with CRC were included, together with a 20% random sample of those over age 50 with no affected first-degree relative.

Population-based controls were randomly sampled from these same catchment areas as the 3 registries, frequency matched on age, as described recently (9). All controls were self-identified as non-Hispanic White and reported no personal or family history of CRC.

Written, informed consent was obtained from all participants. The study was approved by the Institutional Review Board at each of the institutions.

### Data collection and variable definitions

All participants completed mailed questionnaires (Cancer Care Ontario) or a telephone-based or face-to-face interview (FHCRC, University of Melbourne) at study enrollment information. Questions focused on exposures 2-years before the date of diagnosis for cases and 2-years before the date of recruitment for controls. Data were collected on personal and family histories of colorectal (cancer and polyps) and other cancers and colon polyps, and lifestyle risk factors, including, medication use, reproductive history, physical activity, body height and weight, demographics, alcohol intake, tobacco use, diet, and supplement use.

Ever-use (yes, no) of selected supplements (multivitamins, folic acid, and calcium) and medications (nonsteroidal anti-inflammatory drugs, NSAIDs) were defined as use at least 2 times per week for more than a month during a participant's lifetime. NSAIDs included ibuprofen and aspirin. Because folic acid is contained in nearly all multivitamins, the derived variable for folic acid included use of folic acid supplements and multivitamins. Alcohol use was defined as the consumption of any alcoholic beverage (beer, wine, hard cider, sake, liquor, spirits, mixed drinks, or cocktails) at least once a week for 6 months or longer during the most recent decade of life at enrollment. Being an ever smoker was defined as ever smoking at least 1 cigarette per day for 3 months or longer. Pack-years of smoking was calculated based on the number of cigarettes smoked per day and the number of years smoked. A person was considered to be physically active if they reported more than 20 metabolic equivalent (MET) hours per week of physical activity during the most recent decade of life at enrollment. The number of servings per week of fruits, vegetables, and red meat were also calculated. Body mass index (BMI) was calculated as the person's weight (kg) 2 years prior to study recruitment divided by adult height (m) squared.

### Genotyping

All participants provided a blood sample at the time of recruitment. DNA samples were genotyped with the Illumina Human1M ($n$ individuals = 1,973; $m$ = 1,072,820 SNPs) or Human1M-Duo ($n$ individuals = 374; $m$ = 1,199,187 SNPs) BeadChip platforms. Samples with GenCall scores less than 0.15 at any locus were considered "no calls." Each 96-well plate included 1 inter-plate positive quality control sample (NA06990—Coriell Cell Repositories). In addition, 27 blinded and 22 unblinded quality control replicates from the study sample were genotyped. SNP data obtained from both the Coriell and study sample replicates showed a very high concordance rate of called genotypes: 99.95% and more than 99.94%, respectively (for samples with call rates >90%). The Human1M and Human1M-Duo contain 415 and 436 SNPs, respectively, that were genotyped as part of a candidate gene study on the Illumina GoldenGate platform on a subset of the individuals genotyped in this study ($N$ = 444). A high concordance rate (>98%) was observed for more than 99% of the samples with a call rate more than 90%.

Individuals were excluded with (Fig. 1)

(1) missing phenotype data ($n$ = 2);
(2) self-identified as Caucasian ($n$ = 29);
(3) poor concordance (<98%) with genotypes on selected candidate genes genotyped on GoldenGate platform ($n$ = 3);

Published OnlineFirst February 25, 2011; DOI: 10.1158/1055-9965.EPI-10-0675



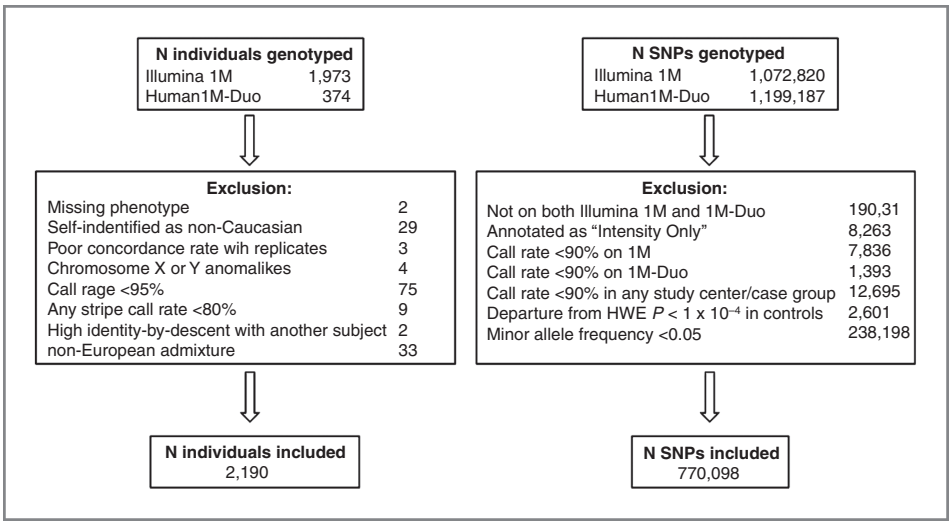


Figure 1. Flow diagram of the number of individuals and SNPs genotyped and excluded in this study.

(4) chromosome X or Y anomalies ($n = 3$ gender misclassifications, $n = 1$ low male chromosome Y intensity);
(5) a call rate less than 95% ($n = 75$);
(6) any stripe call rate less than 80% ($n = 9$);
(7) a high identity-by-descent with another study individual ($n = 2$); or
(8) non-European admixture as estimated by STRUCTURE (11; $n = 33$).

SNPs were excluded from analysis if

(1) they were not on both the Human1M and Human1M-Duo ($m = 190{,}301$);
(2) they were annotated as "Intensity Only" on either the Human1M or Human1M-Duo ($m = 8{,}263$);
(3) had a call rate less than 90% on the Human1M or Human1M-Duo ($m = 7{,}836$ and 1,393, respectively);
(4) had a call rate less than 90% in any study center/case status group ($m = 12{,}695$);
(5) they departed from Hardy–Weinberg equilibrium ($P \leq 1 \times 10^{-4}$) in controls ($m = 2{,}601$); or
(6) had a minor allele frequency (MAF) less than 0.05 ($m = 238{,}198$).

A total of 2,190 individuals and 770,098 SNPs were used in the final analysis.

All SNPs with borderline significant associations underwent additional quality control checks including:

(1) visual inspection of genotype clusters;
(2) verification of genotype concordance between HapMap samples genotyped by Illumina on the Human1M and Human1M-Duo and with HapMap phase II release 24;
(3) verification of no major MAF difference in controls genotyped on the Human1M and Human1M-Duo.

## Replication sample

We replicated significant interactions using data from the ARCTIC Study. Details of the ARCTIC Study are provided elsewhere (5). The selected SNPs were extracted from a larger set of SNPs genotyped in 2,433 unique samples on a custom 10,640-bead iSelect array from Illumina. All eligible cases of CRC were included irrespective of MSI status. After excluding samples from the Colon CFR that were included in this study and samples not self-identified as white, we were left with a total of 872 cases and 810 controls. The selected SNPs were extracted from a larger set of SNPs genotyped in 2,433 unique samples on a custom 10,640-bead iSelect array from Illumina designed for 7,703 SNPs. The call rate for this panel was 99.96% after excluding 3 failed DNAs (call rate < 69%) and 332 failed SNPs (4.3%). There were no discordant genotypes in 23 pairs of duplicates. Data collection on environmental risk factors and variable definitions were carried out in the same manner as described above. All subjects provided written informed consent. This study was approved by the ethics review boards of the Toronto Academic Health Sciences Council.

## Statistical analysis

We considered 3 approaches for genome-wide G × E testing. In the first 2 approaches, we exhaustively tested every SNP in the GWAS panel for G × E interaction with each of the environmental exposures using either a case–control test or a case-only test. The case–control test is based on the logistic regression model:

$$\log\left(\frac{\Pr(Y=1|G,\,E,\,Z)}{\Pr(Y=0|G,\,E,\,Z)}\right) = \alpha+\beta_e E+\beta_g G+\beta_{gxe} G\times E+\eta Z \quad \text{(A)}$$

where $Y$ indicates disease status, with $Y = 1$ for cases and $Y = 0$ for controls, $E$ is an environmental exposure, $G$ is the genotype at a particular SNP, and $Z$ represents any additional covariates to be adjusted for such as sex, age (continuous) and center.

Published OnlineFirst February 25, 2011; DOI: 10.1158/1055-9965.EPI-10-0675



For a binary exposure, the case-only test is based on logistic regression model:

$$\log\left(\frac{\Pr(E=1|G,Z)}{\Pr(E=0|G,Z)}\right)=\alpha+\beta G+\eta Z \quad \text{(B)}$$

For a quantitative exposure, the case-only test is based on logistic regression model:

$$E\,[E]=\alpha+\beta G+\eta Z \quad \text{(C)}$$

We used an additive coding for the genotypes for both the case–control and the case-only tests, i.e., $G$ indicates the number of copies of the reference allele ($G = 0, 1, 2$). The additive model is known to have good power for a very wide range of true modes of action (recessive, dominant, multiplicative; ref. 12). For the case–control test, the hypothesis of no SNP–environment (SNP × E) interaction corresponds to the $H_0$: $\beta_{ge} = 0$ in model (A). For the case-only test the hypothesis of no SNP × E interaction corresponds to $H_0$: $\beta = 0$ in models (B) or (C) depending on whether the exposure is binary or quantitative. We tested every SNP that passed quality control and was polymorphic in our sample using the case–control and case-only tests. We refer to these scans for SNP × E interactions as exhaustive case–control or exhaustive case-only respectively, in contrast with the 2-step scan described below. To control for multiple testing we used a simple Bonferroni correction for the number of SNPs that were actually tested ($0.05/770{,}098 = 6.5 \times 10^{-8}$). Because each exposure was considered an independent *a priori* hypothesis we corrected across SNPs for each exposure, but not across exposures. For continuous exposures, we also report the stratified estimates of risk by dichotomizing at the median unless otherwise reported.

The third method for genome-wide G×E testing we considered was the approach of Murcray and colleagues (8). This 2-step method consists of a screening first step followed by a formal test of interaction. Specifically, in the first step a test of association between the exposure $E$ and SNP $G$ is carried out on the *combined sample of cases and controls* based on the logistic regression model:

$$\log\left(\frac{\Pr(E=1|G,Z)}{\Pr(E=0|G,Z)}\right)=\alpha+\gamma G+\eta Z \quad \text{(B)}$$

for binary exposures, or the linear regression model:

$$\mathrm{E[E|G]}=\alpha+\gamma G+\eta Z \quad \text{(C)}$$

for continuous exposures.

The hypothesis $H_0$: $\gamma = 0$ is tested for each SNP at significance level $a_1 = 0.001$ using a $\chi^2$ 1df Wald test. The $m$ SNPs achieving $\alpha_1$ significance (i.e., with $P$-value $<\alpha_1$) pass the screening step and are tested for G×E interaction using model (1). To preserve a genome-wide type I error of $\alpha = 0.05$ of the overall 2-step procedure, Murcray and colleagues (8) showed that it suffices to correct in step 2 by the $m$ SNPs that pass the screening, i.e., by testing at significance level $\alpha/m$. For details of the rationale behind the screening step and the validity of the 2-step approach see Murcray and colleagues (8). All the genome-wide G×E analyses were carried out with the software PLINK (13).

In addition to the genome-wide G × E analyses using the exhaustive case–control, exhaustive case-only, and 2-step methods described above, we carried out focused testing of previously reported and replicated genetic variants associated with CRC from 5 published GWAS (1–5) and a meta-analysis (6) for G×E interaction with each of the exposures of interest: 8q23.3 (rs16892766, *EIF3H*; ref. 2); 8q24 (rs6983267, rs7014346; refs. 1, 4, 5, 14–16), 10p14 (rs10795668 ; ref. 2), 11q23 (rs3802824; ref. 1), 14q22.2 (rs4444235, *BMP4*; ref. 6); 15q13 (rs4779584; ref. 17); 16q22.1 (rs9929218, *CDH1*)(6); and 18q21 (rs4939827, *SMAD7*; refs. 1, 3), 19q13.1 (rs10411210, *RHPN2*; ref. 6), and 20p12.3 (rs961253; ref. 6). 8q24-rs1050477 and 9p24-rs719725 (5, 14) were not available on the Illumina Human1M or Human1M-Duo. We considered 9p24-rs7025295 and 9p24-rs7857628 as surrogates for the missing 9p24-rs719725 ($r^2 = 0.965$, $r^2 = 0.966$ using HapMap2_r24 CEU, respectively). We tested each variant using the case–control based on model (A) and case-only test based on model (B) at 5% significance level. In addition to testing individual SNPs, we tested a score that combines information from all the 13 SNPs into a single variable for interaction with the exposures of interest. For each subject, the score was constructed by counting the number of CRC risk-increasing variants across the 13 SNPs (i.e., the score ranges from 0 to 26). We tested the interaction of the score and the exposures using the standard logistic regression model (A), with $G$ representing now the quantitative score.

Only interactions that were identified to be significant from this GWA study were tested in the ARCTIC Study using the 1-step test on model (A) at 5% significance level. All models were adjusted for age, center, and sex unless otherwise specified.

## Results

This study included 1,191 population-based cases of MSS/MSI-L CRC and 999 unrelated population-based controls. Table 1 shows the distribution of selected characteristics for the study population. After adjustment for age, sex, and study center, we found BMI, smoking, and red meat intake were positively associated with risk of CRC. Ever use of folic acid and multivitamins were associated with an increased risk of disease in unadjusted models only. Alcohol use, NSAID use and calcium use were associated with statistically significant decreased risk of CRC. Among women, ever use of postmenopausal hormones or oral contraceptives were associated with statistically significant decreased disease risks. Servings of fruits and vegetables, physical activity and height were not associated with risk of CRC.

Using the exhaustive case–control and case-only tests, we observed no statistically significant interactions with

Published OnlineFirst February 25, 2011; DOI: 10.1158/1055-9965.EPI-10-0675



**Table 1.** Characteristics of the study sample

| | Cases ($n$ = 1,191) | Controls ($n$ = 999) | Unadjusted OR (95% CI) | Adjusted OR (95% CI)[a] |
|---|---|---|---|---|
| Mean age ± SD, y | 52.6 ± 11.3 | 60.4 ± 11.1 | – | – |
| Sex, *n* (%) | | | | |
| Male | 617 (51.8) | 478 (47.8) | – | – |
| Female | 574 (48.2) | 521 (52.2) | | |
| Center, *n* (%) | | | | |
| Ontario, Canada | 417 (35.0) | 501 (50.2) | – | – |
| Melbourne, Australia | 330 (27.7) | 189 (18.9) | | |
| Seattle | 444 (37.3) | 309 (30.9) | | |
| Mean height ± SD, m | 1.71 ± 0.10 | 1.70 ± 0.10 | 2.23 (0.96 -5.19) | 0.25 (0.06 -1.04) |
| Unknown/missing | 1 | 0 | | |
| BMI (kg/m$^2$), *n* (%) | | | | |
| 18.5–25.0 (normal) | 379 (32.8) | 408 (41.5) | 1.00 | 1.00 |
| 25.1–30 (overweight) | 490 (42.4) | 382 (38.8) | 1.38 (1.14–1.68) | 1.51 (1.20–1.90) |
| 30+ (obese) | 287 (24.8) | 194 (19.7) | 1.59 (1.23–2.00) | 1.57 (1.20–2.01) |
| Mean ± SD, kg/m$^2$ | 27.40 ± 5.43 | 26.63 ± 5.30 | 1.03 (1.01–1.04) | 1.03 (1.02 -1.05) |
| Unknown/missing | 29 | 14 | | |
| Alcohol use, *n* (%) | | | | |
| <1 per wk | 428 (36.1) | 336 (33.7) | 1.00 | 1.00 |
| ≥1 per wk | 757 (63.9) | 661 (66.3) | 0.90 (0.75–1.07) | 0.77 (0.62–0.95) |
| Unknown/missing | 6 | 2 | | |
| Smoking, *n* (%) | | | | |
| Never | 508 (43.2) | 437 (44.5) | 1.00 | 1.00 |
| Ever | 668 (56.8) | 544 (55.5) | 1.06 (0.89–1.25) | 1.27 (1.04–1.55) |
| Mean pack-years ± SD | 13.14 ± 20.36 | 12.85 ± 19.58 | 1.001 (0.996–1.005) | 1.008 (1.003–1.013) |
| Unknown/missing | 15 | 18 | | |
| NSAIDs, *n* (%) | | | | |
| Never | 765 (64.3) | 534 (53.5) | 1.00 | 1.00 |
| Ever | 425 (35.7) | 465 (46.5) | 0.64 (0.54–0.76) | 0.69 (0.56–0.85) |
| Unknown/missing | 1 | 0 | | |
| Folic acid,[b] *n* (%) | | | | |
| Never | 558 (47.0) | 494 (49.4) | 1.00 | 1.00 |
| Ever | 630 (53.0) | 505 (50.6) | 1.10 (0.93–1.31) | 0.98 (0.80–1.21) |
| Unknown/missing | 3 | 0 | | |
| Multivitamins, *n* (%) | | | | |
| Never | 608 (51.4) | 519 (52.2) | 1.00 | 1.00 |
| Ever | 576 (48.6) | 475 (47.8) | 1.04 (0.87–1.23) | 0.98 (0.79–1.20) |
| Unknown/missing | 7 | 5 | | |
| Calcium, *n* (%) | | | | |
| Never | 919 (77.6) | 645 (64.8) | 1.00 | 1.00 |
| Ever | 265 (22.4) | 350 (35.2) | 0.53 (0.44–0.64) | 0.65 (0.51–0.82) |
| Unknown/missing | 7 | 4 | | |
| Postmenopausal hormones,[c] *n* (%) | | | | |
| Never | 407 (71.2) | 255 (49.1) | 1.00 | 1.00 |
| Ever | 165 (28.8) | 264 (50.9) | 0.39 (0.31–0.50) | 0.60 (0.44–0.81) |
| Unknown/missing | 2 | 2 | | |
| Oral contraceptives,[c] *n* (%) | | | | |
| Never | 206 (36.3) | 206 (39.7) | 1.00 | 1.00 |
| Ever | 361 (63.7) | 313 (60.3) | 1.15 (0.90–1.47) | 0.59 (0.43–0.80) |
| Unknown/missing | 7 | 2 | | |
| Physically active, *n* (%) | | | | |
| No | 469 (58.7) | 393 (57.5) | 1.00 | 1.00 |
| Yes | 330 (41.3) | 291 (42.5) | 0.95 (0.77 -1.17) | 1.07 (0.84–1.37) |
| Unknown/missing | 392 | 315 | | |
| Mean number of servings per wk of fruit ± SD | 10.45 ± 8.53 | 11.42 ± 8.2 | 0.986 (0.976–0.996) | 0.991 (0.979–1.003) |
| Unknown/missing | 33 | 15 | | |
| Mean number of servings per wk vegetables ± SD | 14.26 ± 11.23 | 14.59 ± 10.37 | 0.997 (0.99 -1.005) | 0.994 (0.985–1.004) |
| Unknown/missing | 6 | 5 | | |
| Mean number of servings per wk of red meat ± SD | 4.75 ± 4.34 | 3.96 ± 3.76 | 1.053 (1.029–1.079) | 1.053 (1.025–1.082) |
| Unknown/missing | 41 | 48 | | |

[a]Adjusted for age, sex, and center.

[b]Includes multivitamin users.

[c]Females only.

Published OnlineFirst February 25, 2011; DOI: 10.1158/1055-9965.EPI-10-0675



any SNP at a genome-wide significance level of $6.5 \times 10^{-8}$ with any environmental exposure. The lowest interaction *P* values were between: oral contraceptive use and rs17329226 ($P = 7.0E-07$); and ever smoker and rs2486540 ($P = 3.1E-07$), rs2486538 ($P = 3.7E-07$), and rs538835 ($P = 5.3E-07$).

Using the 2-step method, between 662 and 1,004 SNPs (depending on the exposure variable) passed the significance level in the screening step and were carried on to the second step. Therefore, the appropriate number of corrections for multiple testing varied by exposure, dependent on the total number of SNPs in the second step, from $5.0 \times 10^{-5}$ to $7.6 \times 10^{-5}$. We identified no significant G × E interactions with *P* value less than $10^{-4}$.

Table 2 lists the known hits for CRC indentified through published GWA studies (1–6). We tested whether any of these SNPs showed a significant interaction with the selected environmental exposures. We identified the following interactions (using a case–control test): rs3802842 and postmenopausal hormones ($P = 0.01$); rs10795668 and oral contraceptive-use ($P = 0.04$), rs961253 and oral contraceptive-use ($P = 0.01$); rs9929218 and height ($P = 0.02$); rs9929218 and alcohol-use ($P = 0.04$); rs4939827 and servings of vegetable intake ($P = 0.01$); and rs9929218 and calcium-use ($P = 0.045$). In our replication sample, none of the interactions with the individual SNPs were significant at the 5% level. When we tested the interaction of the environmental covariates and the score that combines the previously reported GWAS hits, we only found 2 marginal significant interactions with red meat consumption ($P = 0.01$) and calcium ($P = 0.05$). We did not attempt replication of these interactions.

## Discussion

In this GWAS of early-onset MSS/MSI-L CRC, we identified no selected personal or lifestyle characteristic that significantly modified the effect of genetic variants on the risk of CRC at a strict genome-wide level of less than $6.5 \times 10^{-8}$ using an exhaustive case–control or case-only test or the appropriate significance levels for 2-step method of Murcray and colleagues (8). We identified 7 significant interactions with previously identified hits from published GWAS in CRC. Interestingly, one of the interactions was between rs3802842 and postmenopausal hormone use; rs3802842 has been previously reported to be associated with an increased risk of CRC among females with Lynch syndrome (18). However, none of these 7 interactions were statistically significant at the 5% level in an independent replication sample.

Little of the genetic variation in CRC has been explained and it is likely that many more variants remain to be identified. One potential way to identify additional susceptibility alleles is to search for G×E interactions, and thereby identify genetic variants that may have an effect only in a given subgroup of individuals, identified by a common environmental risk factor or molecular profile. We applied an efficient 2-step approach described by Murcray and colleagues for detecting loci involved in G × E interactions. It is carried out independently of any initial scans for main effects and that incorporates a preliminary screening step constructed to efficiently use all available information (8). Other methods have been proposed, such as a 2-df test for assessing genetic main effects and interactions jointly (19) and approaches designed to combine the case–control and case-only analyses (20, 21), but there has been no formal comparison of these methods.

Achieving sufficient statistical power is challenging in a genome-wide context, even with these recently described methodologies. Our power calculations highlight this point, especially where the expected gene, exposure and interaction effects are modest. Figure 2 shows the sample size required to attain 80% power with the 2-step approach for various combinations of minor allele frequencies, exposure prevalences, and interaction odds ratios. In this context, it was assumed that there were no SNP main effects, corresponding to the scenario where a G×E scan could detect a SNP that a standard GWAS based on SNP main effects would not. We found that a using data from a typical GWAS of 1,000 cases and 1,000 controls would detect interaction odds ratios of 2 or higher, with highly prevalent exposures and allele frequencies. There are likely to be many G × E interactions, but our study is underpowered to detect them. International consortia gathering GWAS data in CRC may aid in this effort if environmental covariates are available and there is potential for harmonization of variable definitions. However, even this increased sample size will not suffice to detect interaction odds ratios below 1.4, especially for less frequent exposures and lower allele frequencies.

We also investigated whether any of the recently reported and robustly replicated susceptibility loci identified through GWA studies of CRC were modulated by selected environmental factors. We considered only replicated susceptibility variants from published GWA studies of independent CRC cases and unaffected controls (1–6). We identified a few significant interactions at the less than 5% level, but none of these were significant in an independent case–control study of CRC that had collected epidemiologic data using the same questionnaires from individuals in one of the same geographic regions. One potential reason for our failure to replicate could be that we were unable to restrict our replication sample to only cases with early-onset MSS or MSI-L cancers. Common environmental exposures, such as alcohol intake, cigarette smoking, and obesity, have been reported to differ by MSI strata (22, 23). Furthermore, for 4 known susceptibility alleles we found no association with CRC in the Colon CFR and in the absence of a main effect the prospects of identifying a G × E interaction may be lower.

There are some limitations to this study. The main concern is the limited statistical power to investigate G

Published OnlineFirst February 25, 2011; DOI: 10.1158/1055-9965.EPI-10-0675

**Table 2.** Significant interactions at 5% level between known CRC-associated variants and environmental factors

| SNP | Chr | Gene | MAF[a] | Main SNP effect OR (95% CI) | Environment factor | Category | Stratified OR (95% CI) | *P*-int[b] | Replication sample: ARCTIC Study | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | Stratified OR (95% CI) | *P*-int[b] |
| rs16892766 | 8q23.3 | EIF3H (flanking 3′ end) | 0.08 | 1.28 (1.01–1.62) | – | | | | | |
| rs6983267 | 8q24.21 | POU5F1B (flanking 5′ end) | 0.48 | 0.81 (0.70–0.93) | – | | | | | |
| rs7014346 | 8q24.21 | POU5F1B (flanking 5′ end) | 0.38 | 1.17 (1.02–1.34) | – | | | | | |
| rs7025295 | 9p24.1 | UHRF2 (flanking 5′ end) | 0.40 | 0.83 (0.73–0.96) | – | | | | | |
| rs7857628 | 9p24.1 | UHRF2 (flanking 5′ end) | 0.41 | 0.82 (0.71–0.94) | – | | | | | |
| rs10795668[c] | 10p14 | Gene desert | 0.33 | 0.84 (0.73–0.98) | OC use | No | 0.97 (0.70–1.34) | 0.04 | 0.97 (0.71–1.32) | 0.30 |
| | | | | | | Yes | 0.61 (0.46–0.81) | | 1.23 (0.92–1.65) | |
| rs3802842 | 11q23.1 | C11orf93 (intron) | 0.28 | 1.22 (1.06–1.41) | PMH use | No | 1.48 (1.11–1.98) | 0.01 | 1.08 (0.80–1.47) | 0.53 |
| | | | | | | Yes | 0.79 (0.56–1.11) | | 1.24 (0.91–1.68) | |
| rs4444235 | 14q22.2 | BMP4 (flanking 3′ end) | 0.48 | 1.11 (0.97–1.26) | – | | | | | |
| rs4779584 | 15q13.3 | SCG5 (flanking 3′ end) | 0.19 | 1.22 (1.03–1.44) | – | | | | | |
| rs9929218 | 16q22.1 | CHD1 (intron) | 0.29 | 0.93 (0.80–1.08) | Height | <1.7 m | 1.13 (0.91–1.41) | 0.02 | 0.79 (0.62–0.99) | 0.58 |
| | | | | | | ≥1.7 m | 0.79 (0.65–0.97) | | 0.72 (0.57–0.90) | |
| | | | | | Calcium use | No | 0.85 (0.72–1.02) | 0.05 | 0.78 (0.64–0.96) | 0.55 |
| | | | | | | Yes | 1.21 (0.89–1.64) | | 0.70 (0.53–0.93) | |
| | | | | | Alcohol use | No | 1.04 (0.87–1.25) | 0.04 | 0.66 (0.49–0.88) | 0.30 |
| | | | | | | Yes | 0.74 (0.56–0.96) | | 0.79 (0.65–0.97) | |
| rs4939827 | 18q21.1 | SMAD7 (intron) | 0.48 | 0.80 (0.70–0.92) | Vegetable intake | <14 p/wk | 0.84 (0.68–1.04) | 0.01 | 0.79 (0.62–0.99) | 0.33 |
| | | | | | | ≥14 p/wk | 0.77 (0.65–0.93) | | 0.91 (0.75–1.10) | |
| rs10411210 | 19q13.11 | RHPN2 (intron) | 0.09 | 1.02 (0.81–1.28) | – | | | | | |
| rs961253 | 20p12.3 | Gene desert | 0.35 | 1.02 (0.89–1.18) | OC use | No | 1.48 (1.07–2.06) | 0.01 | 1.16 (0.86–1.57) | 0.22 |
| | | | | | | Yes | 0.83 (0.64–1.09) | | 0.90 (0.67–1.20) | |

Abbreviations: PMH, postmenopausal hormones; OC, oral contraceptives; –, no G × E interaction detected at significance level less than 5% for the selected environmental factors.

[a]MAF determine in unrelated controls.

[b]*P*-value for interaction between G and E.

Published OnlineFirst February 25, 2011; DOI: 10.1158/1055-9965.EPI-10-0675



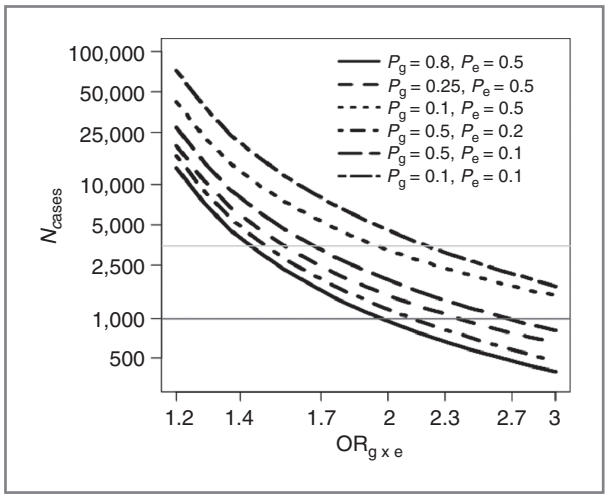


Figure 2. Sample size required to attain 80% power with the 2-step SNP × E approach for various combinations of minor allele frequencies (pg) and exposure prevalence (pe) for a binary environmental factor E. The marginal OR of the SNP was set to ORg = 1 (i.e., no marginal effect) and the marginal OR of the environmental factor was set to ORe = 1.5. An equal number of cases and control was assumed. The dark gray horizontal line represents the number of cases for a typical CRC GWAS. The light gray horizontal line represents the approximate number of cases achievable by pooling all existing CRC GWAS.

× E interactions for less common exposures and less frequent alleles. Collaborative consortia offer important advantages of increasing sample size; however, they also have important limitations, including the potential introduction of heterogeneity due to combining different study designs, measures of exposures, and cancer outcome. Consortia with central quality control procedures and careful standardization and harmonization of definitions and measurements may be helpful. However, large sample size alone does not guarantee quality and reliable results (24). In this study, we had uniform data collection protocols and all cases were defined in a standard manner as MSS or MSI-L. Another potential limitation is our relatively crude definitions of the environmental factors. Furthermore, because of the study design, we were unable to investigate the potential effects of ethnicity, family history of CRC, or other phenotypes of CRC (i.e., MSI high). Lastly, there is no consensus about the correct statistical method to model G × E interactions and more research is required.

In summary, we identified no genome-wide significant G × E interactions in this GWAS of early-onset MSS/MSI-L CRC. Much of the evidence from descriptive epidemiology, migrant studies, and changes in CRC rates in countries undergoing rapid economic development (most obviously Japan in the second half of the twentieth century; Japan now has the highest rates of CRC in the world) points to environmental risk factors as the major determinants of the international variation in CRC. It is crucial therefore that we gain a better understanding of susceptibility to these environmental factors. This, in turn, underscores the need to detect G × E interactions, which will require large collaborations of GWA studies with adequate data collection on exposures.

## Disclosure of Potential Conflicts of Interest

The content of this article does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the CFRs, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government or the CFR.

## Grant Support

This work was supported by the National Cancer Institute, NIH, under RFA no. CA-95-011 and through cooperative agreements with the Australasian Colorectal Cancer Family Registry (U01 CA097735), the Ontario Registry for Studies of Familial Colorectal Cancer (U01 CA074783), and the Seattle Colorectal Cancer Family Registry (U01 CA074794), as well as NIH/NCI U01CA122839 GWAS (G. Casey) and the Canadian Cancer Society Research Institute, the Ontario Institute for Cancer Research, and the Ontario Ministry of Research and Innovation.

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 24, 2010; revised October 25, 2010; accepted January 26, 2011; published OnlineFirst February 25, 2011.

## References

1. Tenesa A, Farrington SM, Prendergast JG, Porteous ME, Walker M, Haq N, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. Nat Genet 2008;40:631–7.
2. Tomlinson IP, Webb E, Carvajal-Carmona L, Broderick P, Howarth K, Pittman AM, et al. A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. Nat Genet 2008;40:623–30.
3. Broderick P, Carvajal-Carmona L, Pittman AM, Webb E, Howarth K, Rowan A, et al. A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. Nat Genet 2007;39:1315–7.
4. Tomlinson I, Webb E, Carvajal-Carmona L, Broderick P, Kemp Z, Spain S, et al. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. Nat Genet 2007;39:984–8.
5. Zanke BW, Greenwood CM, Rangrej J, Kustra R, Tenesa A, Farrington SM, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. Nat Genet 2007;39:989–94.
6. Houlston RS, Webb E, Broderick P, Pittman AM, Di Bernardo MC, Lubbe S, et al. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. Nat Genet 2008;40:1426–35.
7. Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. Stat Med 1994;13:153–62.
8. Murcray CE, Lewinger JP, Gauderman WJ. Gene-environment interaction in genome-wide association studies. Am J Epidemiol 2009;169:219–26.
9. Newcomb PA, Baron J, Cotterchio M, Gallinger S, Grove J, Haile R, et al. Colon cancer family registry: an international resource for studies

Published OnlineFirst February 25, 2011; DOI: 10.1158/1055-9965.EPI-10-0675



of the genetic epidemiology of colon cancer. Cancer Epidemiol Biomarkers Prev 2007;16:2331–43.
10. Lindor NM, Rabe K, Petersen GM, Haile R, Casey G, Baron J, et al. Lower cancer incidence in Amsterdam-I criteria families without mismatch repair deficiency: familial colorectal cancer type X. JAMA 2005;293:1979–85.
11. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics 2000;155:945–59.
12. Thomas DC. Statistical Methods in Genetic Epidemiology. Oxford: Oxford University Press; 2004.
13. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559–75.
14. Poynter JN, Figueiredo JC, Conti DV, Kennedy K, Gallinger S, Siegmund KD, et al. Variants on 9p24 and 8q24 are associated with risk of colorectal cancer: results from the Colon Cancer Family Registry. Cancer Res 2007;67:11128–32.
15. Haiman CA, Patterson N, Freedman ML, Myers SR, Pike MC, Waliszewska A, et al. Multiple regions within 8q24 independently affect risk for prostate cancer. Nat Genet 2007;39:638–44.
16. Gruber SB, Moreno V, Rozek LS, Rennerts HS, Lejbkowicz F, Bonner JD, et al. Genetic variation in 8q24 associated with risk of colorectal cancer. Cancer Biol Ther 2007;6:1143–7.
17. Jaeger E, Webb E, Howarth K, Carvajal-Carmona L, Rowan A, Broderick P, et al. Common genetic variants at the CRAC1 (HMPS) locus on chromosome 15q13.3 influence colorectal cancer risk. Nat Genet 2008;40:26–8.
18. Wijnen JT, Brohet RM, van Eijk R, Jagmohan-Changur S, Middeldorp A, Tops CM, et al. Chromosome 8q23.3 and 11q23.1 variants modify colorectal cancer risk in Lynch syndrome. Gastroenterology 2009;136:131–7.
19. Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ. Exploiting gene-environment interaction to detect genetic associations. Hum Hered 2007;63:111–9.
20. Mukherjee B, Ahn J, Gruber SB, Rennert G, Moreno V, Chatterjee N. Tests for gene-environment interaction from case-control data: a novel study of type I error, power and designs. Genet Epidemiol 2008;32:615–26.
21. Li D, Conti DV. Detecting gene-environment interactions using a combined case-only and case-control approach. Am J Epidemiol 2009;169:497–504.
22. Poynter JN, Haile RW, Siegmund KD, Campbell PT, Figueiredo JC, Limburg P, et al. Associations between smoking, alcohol consumption, and colorectal cancer, overall and by tumor microsatellite instability status. Cancer Epidemiol Biomarkers Prev 2009.
23. Campbell PT, Jacobs ET, Ulrich CM, et al. Associations between obesity/overweight and colorectal cancer risk: Overall and by microsatellite instability status. J Natl Cancer Inst 2010;102:391–400.
24. Ioannidis JP, Trikalinos TA, Ntzani EE, Contopoulos-Ioannidis DG. Genetic associations in large versus small studies: an empirical assessment. Lancet 2003;361:567–71.

Cancer Epidemiology, Biomarkers & Prevention

AACR American Association for Cancer Research

# Genotype–Environment Interactions in Microsatellite Stable/Microsatellite Instability-Low Colorectal Cancer: Results from a Genome-Wide Association Study

Jane C. Figueiredo, Juan Pablo Lewinger, Chi Song, et al.

*Cancer Epidemiol Biomarkers Prev* 2011;20:758-766. Published OnlineFirst February 25, 2011.