Characterization of Gene–Environment Interactions for Colorectal Cancer Susceptibility Loci

Carolyn M. Hutter1,2, Jenny Chang-Claude4, Martha L. Slattery7, Bethann M. Pluymeisen1, Yi Lin1, David Duggan9, Hongmei Nan9,10, Mathieu Lemire14, Jagadish Rangrej14, Jane C. Figueiredo18, Shuo Jiao1, Tabitha A. Harrison1, Yan Liu19, Lin S. Chen20, Deanna L. Stelling1, Greg S. Warnick1, Michael Hoffmeister5, David J. Hunter10, Steven Gallinger16, Brent W. Zanke22, Hermann Brenner5, Jing Ma9, Richard B. Hayes26, Bette J. Caan27, John D. Potter1,2,28, Li Hsu1,3,5,10, Mathieu Lemire14, Jagadish Rangrej14, Jane C. Figueiredo18, Shuo Jiao1, Thomas J. Hudson15,17, and Ulrike Peters1,2

Abstract

Genome-wide association studies (GWAS) have identified dozens of loci associated with colorectal cancer (CRC) risk. Here, we examined potential effect-modification between single-nucleotide polymorphisms (SNP) at 10 of these loci and probable or established environmental risk factors for CRC in 7,016 CRC cases and 9,723 controls from nine cohort and case–control studies. We used meta-analysis of an efficient empirical-Bayes estimator to detect potential multiplicative interactions between each of the SNPs [rs16892766 at 8q23.3 (EIF3H/UTP23)], rs983267 at 8q24 (MYC), rs10795668 at 10p14 (FLJ3802842), rs3802842 at 11q23 (LOC120376), rs444235 at 1q22.2 (BMP4), rs4779584 at 15q13 (GREM1), rs9929218 at 16q22.1 (CDH1), rs4939827 at 18q12 (SMAD7), rs10411210 at 19q13.1 (RHPN2), and rs961253 at 20p12.3 (BMP2)] and select major CRC risk factors (sex, body mass index, height, smoking status, aspirin/nonsteroidal anti-inflammatory drug use, alcohol use, and dietary intake of calcium, folate, red meat, processed meat, vegetables, fruit, and fiber). The strongest statistical evidence for a gene–environment interaction across studies was for vegetable consumption and rs16892766, located on chromosome 8q23.3, near the EIF3H and UTP23 genes (nominal pInteraction = 1.3 × 10−4; adjusted P = 0.02). The magnitude of the main effect of the SNP increased with increasing levels of vegetable consumption. No other interactions were statistically significant after adjusting for multiple comparisons. Overall, the association of most CRC susceptibility loci identified in initial GWAS seems to be invariant to the other risk factors considered; however, our results suggest potential modification of the rs16892766 effect by vegetable consumption. Cancer Res; 72(8): 2036–44. ©2012 AACR.

Authors' Affiliations: 1Public Health Sciences Division, Fred Hutchinson Cancer Research Center; 2Department of Epidemiology, School of Public Health, University of Washington; 3Department of Biostatistics, University of Washington, Seattle; Divisions of 4Cancer Epidemiology, 5Clinical Epidemiology and Aging Research, and 6Preventive Oncology, German Cancer Research Center, Heidelberg, Germany; 7Department of Internal Medicine, University of Utah Health Sciences Center, Salt Lake City; 8Translational Genomics Research Institute, Phoenix; 9Channing Laboratory, Brigham and Women’s Hospital and Harvard Medical School; 10Program in Molecular and Genetic Epidemiology, Department of Epidemiology, and 11Departments of Epidemiology and Nutrition, Harvard School of Public Health; 12Department of Medical Oncology, Dana-Farber Cancer Institute; 13Division of Gastroenterology, Massachusetts General Hospital and Harvard Medical School, Boston; 14,15Ontario Institute for Cancer Research; 16Department of Surgery, University Health Network, Toronto General Hospital; 17Departments of Medical Biophysics and Molecular Genetics, University of Toronto, Toronto, Canada; 18Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles; 19Stephens and Associates, Carrollton, Texas; 20Department of Health Studies, University of Chicago, Chicago; 21Service de Génétique Médicale, CHU Nantes, Nantes, France; 22Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Canada; 23Division of Endocrinology, Diabetes and Metabolism, Ohio State University, Columbus; 24Department of Epidemiology, University of Pittsburgh Medical Center, Pittsburgh; 25Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Bethesda; 26Division of Epidemiology, Department of Environmental Medicine, New York University School of Medicine, New York City; 27Division of Research, Kaiser Permanente Medical Care Program, Oakland; and 28Centre for Public Health Research, Massey University, Wellington, New Zealand.

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

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

Corresponding Author: Carolyn M. Hutter, 1100 Fairview Ave. N., M4-B402, P.O. Box 19024, Seattle, WA 98109. Phone: 206-667-7960; Fax: 206-667-7850; E-mail: chutter@fhcrc.org

doi: 10.1158/0008-5472.CAN-11-4067
©2012 American Association for Cancer Research.

Cancer Res; 72(8) April 15, 2012
Introduction

Approximately one third of colorectal cancer (CRC), the second leading cancer in the United States, is attributable to inherited factors (1). Identification of associated genetic variants may elucidate mechanisms underlying this disease. First, results from genome-wide association studies (GWAS) have shown considerable success in identifying genetic variants associated with CRC (2–10). However, these variants currently explain only a small fraction of the genetic heritability (9). Recent work postulates that there may be up to 65 to 70 common loci underlying CRC susceptibility, requiring large sample sizes for detection (11); additional avenues of work are also needed to identify other factors underlying the "missing heritability" (12). Less common genetic variants, and gene–environment interactions (GxE) are postulated to explain an important component (12, 13). In addition, alternative models (e.g., recessive models) have generally not been tested. A full examination of the role of GxE underlying CRC will require genome-wide scans incorporating genetic and environmental factors and interaction terms across the genome. Nonetheless, a logical first step in exploring the GxE contribution is to characterize potential effect modification of genetic risk variants already identified as having marginal effects.

This article focuses on potential GxE interactions for the first 10 CRC GWAS loci identified: 8q24 (MYC), 15q13 (GREM1), 18q21 (SMAD7), 11q23 (EIF3H/UTP23), 10p14 (FIJ3982842), 14q22.2 (BMP4), 16q22.1 (CDH1), 19q13.1 (RHPN2), and 20p12.3 (BMP2). In the context of this article, we use the term "environmental risk factors" broadly to include nonsingle-nucleotide polymorphisms (SNP) risk factors, including sex, which is genetically determined, as well as factors like tobacco use and height, which themselves may be intermediate phenotypes with genetic and environmental determinants. Previous studies have examined gene–environment interactions with selected sets of these known variants for some environmental covariates. However, these studies have either focused only on single variants (14–16) or had relatively small sample sizes and results have been inconsistent (17, 18). Here, we carry out a more comprehensive examination of these loci and 12 probable or established CRC risk factors [sex, body mass index (BMI), height, smoking status, aspirin/nonsteroidal anti-inflammatory drug (NSAID) use, and intake of alcohol, dietary calcium, dietary folate, red meat, processed meat, vegetables, fruit, and fiber] in a combined analysis of 9 case–control and nested case–control studies comprising 7,016 CRC cases and 9,723 controls.

Patients and Methods

Study participants

The studies used are listed in Table 1 and have been described in detail previously (10). In brief, we used data from 5 nested case–control studies in prospective U.S. cohorts [Health Professionals Follow-up Study (HPFS); Nurses’ Health Study (NHS); Physician’s Health Study (PHS); Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (PLCO); and Woman’s Health Initiative (WHI)] and 4 case–control studies from the United States, Canada, and Europe [Assessment of Risk in Colorectal Tumors in Canada (ARCTIC); French Association Study Evaluating RISK for sporadic CRC (ASTERISK); Darmkrebs: Chancen der Verhütung durch Screening (DACHS); Diet, Activity and Lifestyle Survey (DALS)]. The ARCTIC study used subjects from the Ontario Colon Cancer Family Registry (19). All cases were defined as invasive colorectal adenocarcinoma (International Classification of Disease Code 153-154) and confirmed by medical record, pathology report, or death certificate. The studies used nested case–control or case–control designs with study-specific eligibility and matching criteria, except for PLCO. For PLCO, controls were drawn from the controls used in previous GWAS studies of prostate cancer and lung cancer available through dbGaP (20, 21). To account for the different eligibility and matching criteria used in those GWAS, sampling fraction weights, based on sex, smoking status, age at entry, and year of entry were used to weight the PLCO case and controls to be representative of eligible subjects in the full PLCO cohort. PHS subjects were matched on smoking status, so that study is excluded from the summary of main effects of smoking-related variables. Due to small numbers, we excluded samples reported as racial/ethnic groups other than “White”; European ancestry was confirmed in GWAS samples using principal components analysis (22).

Table 1. Overview of the studies included in this analysis

<table>
<thead>
<tr>
<th>Study</th>
<th>Case</th>
<th>Control</th>
<th>Total</th>
<th>% Female</th>
<th>% Colon</th>
<th>Median age, y (range)</th>
<th>Genotyping platform</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARCTIC</td>
<td>821</td>
<td>883</td>
<td>1,704</td>
<td>50.3</td>
<td>66.1</td>
<td>62 (27–77)</td>
<td>Affymetrix GWAS platforms</td>
</tr>
<tr>
<td>ASTERISK</td>
<td>954</td>
<td>1,060</td>
<td>2,014</td>
<td>41.6</td>
<td>70.8</td>
<td>67 (40–99)</td>
<td>BeadXpress</td>
</tr>
<tr>
<td>DACHS</td>
<td>1,731</td>
<td>1,742</td>
<td>3,473</td>
<td>40.3</td>
<td>60.7</td>
<td>69 (33–98)</td>
<td>Illumina GWAS platforms</td>
</tr>
<tr>
<td>DALS I</td>
<td>689</td>
<td>720</td>
<td>1,409</td>
<td>47.4</td>
<td>100</td>
<td>66 (30–79)</td>
<td>BeadXpress</td>
</tr>
<tr>
<td>DALS II</td>
<td>706</td>
<td>710</td>
<td>1,416</td>
<td>43.4</td>
<td>100</td>
<td>66 (30–79)</td>
<td>Illumina GWAS platforms</td>
</tr>
<tr>
<td>HPFS</td>
<td>344</td>
<td>635</td>
<td>979</td>
<td>0</td>
<td>76.2</td>
<td>69 (48–82)</td>
<td>TaqMan OpenArray</td>
</tr>
<tr>
<td>NHS</td>
<td>465</td>
<td>1,009</td>
<td>1,474</td>
<td>100</td>
<td>79.8</td>
<td>61 (44–69)</td>
<td>TaqMan OpenArray</td>
</tr>
<tr>
<td>PLCO</td>
<td>544</td>
<td>1,976</td>
<td>2,520</td>
<td>26.7</td>
<td>94.9</td>
<td>64 (55–74)</td>
<td>Illumina GWAS platforms</td>
</tr>
<tr>
<td>PHS</td>
<td>288</td>
<td>454</td>
<td>846</td>
<td>0</td>
<td>76.3</td>
<td>58 (40–84)</td>
<td>Illumina GWAS platforms</td>
</tr>
<tr>
<td>WHI</td>
<td>474</td>
<td>534</td>
<td>1,008</td>
<td>97</td>
<td>68 (50–79)</td>
<td>TaqMan OpenArray</td>
<td></td>
</tr>
</tbody>
</table>

*DALS I, initial GWAS of subjects in the DALS study; DALS II, follow-up replication for a subset of SNPs for subjects in the DALS study.
All participants gave informed consent, and studies were approved by the Institutional Review Board.

**Genotype data**

We examined 10 SNPs identified through published CRC GWAS before September, 2010 (Table 2). For WHI, PCLO, and DALS I, genotype data were generated with Illumina Human-Hap300k and 240k (PCLO), 550k (WHI, DALS), and 610k (DALS, PLCO) BeadChip Array Systems on the Infinium platform as previously described (10). ARCTIC samples were genotyped on Affymetrix platforms (3) and imputed with BEAGLE (23), using the phased HapMap release 22 as the reference sample (24). We used imputed SNPs, coded as the best call genotype, for all 10 SNPs in ARCTIC. The imputation quality was moderate for rs4939827 ($r^2 = 0.49$) and rs10411210 ($r^2 = 0.82$), and was high for all other SNPs with imputation $r^2$ ranging from 0.90 to 1.00 (see Supplementary Table S1). For DACHS, DALS II, and ASTERISK, samples were genotyped with TaqMan results for rs10505477 as a linkage disequilibrium (LD) substitute for rs6983267 (see Supplementary Table S1). For DACHS, the 8q24 SNP, rs6983267, was not successfully genotyped on BeadXpress and was replaced, for a subset of the samples (2,849 total), with previous TaqMan genotyping. For ASTERISK, we used TaqMan results for rs10505477 as a linkage disequilibrium (LD) substitute for rs6983267 (see Supplementary Table S1). The LD $r^2$ between rs10505477 and rs6983267 in the HapMap Utah residents from ancestry from northern and western Europe (CEU) population is 0.93. The NHS, HPFS, and PHS samples were genotyped using TaqMan OpenArray technology. All genotyping underwent standard quality control procedures, including concordance checks for blinded and unblinded duplicates, examination of sample and SNP call rates, and checking Hardy–Weinberg equilibrium (HWE) in controls. Call rate, HWE $P$ value, and minor allele frequency (MAF) for each SNP in each study are included in Supplementary Table S1. One SNP, rs4444235 at 14q22.2, was excluded from the study because of the HWE $P$ value in controls ($P = 3 \times 10^{-3}$).

**Harmonization of environmental data**

Information on basic demographics and environmental risk factors was collected by self-report using in-person interviews and/or structured questionnaires, as detailed previously (19, 26–34). We carried out a multistep data harmonization procedure, reconciling each study’s unique protocols and data collection instruments. First, we defined common data elements (CDE). We examined the questionnaires and data dictionaries for each study to identify study specific data elements that could be mapped to the CDEs. Through an iterative process, we communicated with each data contributor to obtain relevant data and coding information. The data elements were written to a common data platform, transformed via a SQL programming script, and combined into a single data set with common definitions, standardized permissible values, and standardized coding. The mapping and resulting data were reviewed for quality assurance, and range and logic checks were carried out to assess data and data distributions within and between studies. Outlying samples were truncated to the minimum or maximum value of

<table>
<thead>
<tr>
<th>Chromosomal location</th>
<th>Gene/locus</th>
<th>SNP</th>
<th>Minor allele</th>
<th>Other allele MAF</th>
<th>Current study, OR* (95% CI)</th>
<th>Current study, N</th>
<th>Published ref.</th>
<th>Published OR (95% CI)a</th>
<th>Current study, P value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>8q24</td>
<td>EIF3H/UTP23</td>
<td>rs16892766 A</td>
<td>0.22</td>
<td>0.78</td>
<td>1.23 (1.19–1.27)</td>
<td>16,775</td>
<td>68</td>
<td>1.17 (1.08–1.26)</td>
<td>0.05</td>
</tr>
<tr>
<td>8q24</td>
<td>LOC338591</td>
<td>rs10795668 G</td>
<td>0.33</td>
<td>0.67</td>
<td>0.93 (0.86–0.99)</td>
<td>16,775</td>
<td>68</td>
<td>0.96 (0.89–1.02)</td>
<td>0.17</td>
</tr>
<tr>
<td>8q24</td>
<td>MYC</td>
<td>rs6983267 A</td>
<td>0.50</td>
<td>0.50</td>
<td>0.83 (0.79–0.87)</td>
<td>16,775</td>
<td>68</td>
<td>0.89 (0.84–0.95)</td>
<td>0.03</td>
</tr>
<tr>
<td>8q24</td>
<td>LOC120376</td>
<td>rs3802842 A</td>
<td>0.41</td>
<td>0.59</td>
<td>1.12 (1.04–1.20)</td>
<td>16,769</td>
<td>101</td>
<td>1.12 (1.05–1.17)</td>
<td>0.03</td>
</tr>
<tr>
<td>8q24</td>
<td>14q22.2</td>
<td>rs4444235 A</td>
<td>0.60</td>
<td>0.40</td>
<td>1.14 (1.07–1.21)</td>
<td>16,762</td>
<td>96</td>
<td>1.06 (1.02–1.11)</td>
<td>0.07</td>
</tr>
<tr>
<td>8q24</td>
<td>BMP2</td>
<td>rs961253 A</td>
<td>0.36</td>
<td>0.64</td>
<td>0.96 (0.89–1.04)</td>
<td>16,783</td>
<td>60</td>
<td>0.92 (0.85–1.00)</td>
<td>0.30</td>
</tr>
<tr>
<td>11q23</td>
<td>ENPP3</td>
<td>rs10505477 A</td>
<td>0.49</td>
<td>0.51</td>
<td>1.17 (1.08–1.26)</td>
<td>16,652</td>
<td>74</td>
<td>1.17 (1.09–1.24)</td>
<td>0.05</td>
</tr>
<tr>
<td>11q23</td>
<td>LOC289511</td>
<td>rs6983267 G</td>
<td>0.48</td>
<td>0.52</td>
<td>0.94 (0.87–1.01)</td>
<td>16,652</td>
<td>74</td>
<td>0.87 (0.80–0.95)</td>
<td>0.02</td>
</tr>
<tr>
<td>11q23</td>
<td>LOC120376</td>
<td>rs3802842 A</td>
<td>0.33</td>
<td>0.67</td>
<td>1.12 (1.05–1.19)</td>
<td>16,652</td>
<td>74</td>
<td>0.94 (0.87–1.01)</td>
<td>0.01</td>
</tr>
<tr>
<td>11q23</td>
<td>LOC120376</td>
<td>rs6983267 G</td>
<td>0.48</td>
<td>0.52</td>
<td>0.94 (0.87–1.01)</td>
<td>16,652</td>
<td>74</td>
<td>0.94 (0.87–1.01)</td>
<td>0.01</td>
</tr>
<tr>
<td>11q23</td>
<td>BMP2</td>
<td>rs961253 A</td>
<td>0.36</td>
<td>0.64</td>
<td>0.96 (0.89–1.04)</td>
<td>16,783</td>
<td>60</td>
<td>0.92 (0.85–1.00)</td>
<td>0.30</td>
</tr>
<tr>
<td>11q23</td>
<td>BMP2</td>
<td>rs961253 A</td>
<td>0.36</td>
<td>0.64</td>
<td>0.96 (0.89–1.04)</td>
<td>16,783</td>
<td>60</td>
<td>0.92 (0.85–1.00)</td>
<td>0.30</td>
</tr>
</tbody>
</table>

Abbreviations: N, number; NA, not available. *OR and $P$ values for the additive model. †Presented ORs for current data are based on subjects that overlap with main effects previously published (10). ‡NA, number missing for each SNP. §8q24 has a higher proportion missing because it was not successfully genotyped using BeadXpress.
established range for each variable. The reference time for cohort studies was time of enrollment (WHI and PLCO) or blood draw (HPFS, NHS, and PHS). The data elements considered were analyzed as continuous variables (BMI and height); dichotomous variables [sex (male/female), smoking (ever/never at reference time), aspirin/NSAID use (yes/no for regular use at reference time; see exact definitions in Supplementary Table S2)]; ordered categorical variables [alcohol consumption (3 categories defined by g/d)], and study-specific quartiles for smoking pack years (using never smokers as reference, other quartiles coded 1–4); and sex- and study-specific quartiles, in which the quartile groups were coded with the median value of the quartile within each study and sex and scaled to a unit scale reflective of the distribution for that variable [dietary calcium (units of 500 mg/d), dietary folate (units of 500 mcg/d), red meat (units of servings per day), processed meat (units of servings per day), fruit (units of 5 servings per day), and dietary fiber (units of 10 g/d)]. We use scales such as 500 mg/d for calcium, to provide more meaningful and easier to interpret effect sizes. All quartile variables had 4 categories for each sex within each study. Because some studies collected dietary information in categories that could not be converted to study-specific quartiles, we also examined red meat, processed meat, vegetables, and fruit as dichotomous variables, cut at sex- and study-specific medians. We accounted for the multiple testing burden and potential correlation between these additional variables using permutation testing, as described in the Statistical methods section. For all variables, the lowest category of exposure (or no use) was used as the reference.

Statistical methods

Unless otherwise indicated, we adjusted all regression analyses described below for age, center, and sex, as appropriate. We used fixed effects meta-analysis methods to obtain summary ORs and 95% CIs across studies. The test values from the meta-analysis, unadjusted for multiple comparisons, are termed nominal test values. We report the test value for heterogeneity and examine forest plots for results showing evidence for heterogeneity. For PLCO, we used inverse sampling fractions as weights in all analyses to account for study design; for all other studies, we used equal weights.

Inadequate modeling of the marginal association can bias interaction testing (35). Therefore, for each SNP and environmental factor, we employed a screening method, based on logistic regression main-effect associations, to find a reasonable form to use for GxE testing. Nested models were compared using likelihood ratio tests, with a test value of less than 0.05 indicating significantly better performance. For SNPs, we considered assumptions of log-additive (SNPs coded 0/1/2, representing counts of the minor allele) and recessive (SNPs coded 0/1 in which 0 represents homozygous for common allele or heterozygous and 1 represents homozygous for the minor allele) modes of inheritance in comparison with an unrestricted model with indicator variables for heterozygote and homozygote minor alleles. We did not consider a dominant mode of inheritance, because the log-additive model usually does not lose power if the true model is dominant. If the unrestricted model did not significantly outperform the log-additive model, we used the log-additive model. If the unrestricted model performed significantly better than the log-additive model, but not the recessive model, we used the recessive model. If the unrestricted model performed significantly better than the log-additive and the recessive, we used the unrestricted model. Under this procedure, we selected the recessive model for rs6983267, and the log-additive model for the other 9 SNPs. Dichotomous environmental variables were coded 0/1 and did not require model selection. For the continuous variables, BMI and height, we compared main-effects models with and without a quadratic term. In both cases, the model with the quadratic term did not perform significantly better, so we modeled these variables using only a linear term. For the categorical variables (alcohol, pack years, and the quartile version of the dietary variables), we compared a model using a group-linear variable to a saturated model with indicator variables for each non-reference category. For alcohol, the saturated model performed significantly better, so we modeled alcohol with indicator variables. In contrast, for the other variables, the saturated model was not significantly better than a model with a single group-linear term. Thus, we modeled these variables with their sex- and study-specific medians, as described above in the section on data harmonization.

To test for interactions between SNPs and environmental risk factors, we used an efficient empirical-Bayes (EB) shrinkage method (36). This method creates a weighted average of the standard case-only and case–control estimators, which is weighted toward the unbiased case–control estimator when the assumption of gene–environment independence in the population is suspect and toward the more efficient case-only estimator when the assumption is supported by the data. We modeled both the main effect and interaction based on the model selected from the main effects, as described above. Subjects missing data for a particular SNP or environmental factor were dropped from the analysis for that SNP × factor interaction test.

Because we carried out 180 tests (10 SNPs × 18 versions of the environmental risk factors), with correlation among some tests, we used permutations to account for multiple testing. We ran the analysis 1,000 times using a permuted case–control status in each run. Then, we used the Westfall & Young step down procedure (37) to derive the adjusted test value for each GxE interaction based on the permuted test values. We term these the adjusted test values and used them to evaluate the statistical significance of a given interaction at the 0.05 level.

For situations in which the EB interaction-term adjusted test value of less than 0.05, we also examined the results from the traditional logistic regression case–control estimate and examined results adjusting for additional covariates (smoking history, BMI, alcohol consumption, and red meat consumption). As follow-up analysis, we examined the main effect for the SNP in strata defined by the environmental risk factor. We also pooled the data across studies and examined (i) the main effect of the environmental factor in strata defined by the SNP; and (ii) the combined effect in strata defined by both the SNP and the environmental factor. As a supplemental analysis, we
examined all 180 SNP × environmental factor GxE interactions in substratum analyses restricted to colon only and rectal only cases.

Data harmonization was carried out with SAS and T-SQL. All other analyses were conducted with the R programming language.

Results

Study characteristics are described in Table 1 and Supplementary Table S3. Table 2 shows the marginal results for each SNP. As we have previously reported using an overlapping set of subjects (10), 8 of the 10 loci show statistical evidence for association with CRC with nominal P values ranging from 0.03 to 4.1 × 10⁻⁷. One SNP (rs16892766) had a heterogeneity P value of 0.03; the heterogeneity P value for all other SNPs ranged from 0.18 to 0.96, indicating little evidence for heterogeneity in the main effects of SNPs across studies. The 2 established SNPs not showing statistical evidence for association are rs10795668 at 10p14 and rs10411210 at 19q13; how-ever, both showed a statistically nonsignificant OR in the same direction of association as previous reports (Table 2). Our model selection procedure indicated a recessive model for association with CRC, an OR for the AA genotype (homozygous for the minor allele) compared with the AC (Q) and medians (M), are presented for the model and units used in the main effects for quartiles of alcohol per day. The main effects for quartiles of alcohol per day, drinkers of 1 to 28 grams of alcohol per day, and Alc >28 g/day, drinkers of 1 to 28 grams of alcohol per day, each compared to drinkers of <1 gram per day. Veg, vegetable.

Table 3. Gene–environment interactions with Pinteraction < 0.01

<table>
<thead>
<tr>
<th>SNP/chromosomal location/environmental variable</th>
<th>Fixed effects meta-analysis</th>
<th>ORINT (95% CI) nom.p</th>
<th>adj.p</th>
<th>het.p</th>
<th>Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs16892766/8q23.3/Vegetable quartile medians</td>
<td>1.56 (1.06–2.30)</td>
<td>1.3 × 10⁻⁴</td>
<td>0.03</td>
<td>0.68</td>
<td>ARCTIC/DALS/PLCO/WHI/HPFS/NHS/PHS</td>
</tr>
<tr>
<td>rs16892766/8q23.3/Fiber quartile medians</td>
<td>1.33 (1.13–1.56)</td>
<td>6.0 × 10⁻⁴</td>
<td>0.09</td>
<td>0.87</td>
<td>DALS/PLCO/WHI/HPFS/NHS</td>
</tr>
<tr>
<td>rs49398277/18q21/Red meat above/below median</td>
<td>1.14 (1.05–1.24)</td>
<td>2.9 × 10⁻³</td>
<td>0.36</td>
<td>0.59</td>
<td>ARCTIC/DALS/PLCO/WHI/DACHS/ASTERISK/HPFS/NHS/PHS</td>
</tr>
<tr>
<td>rs16892766/8q23.3/Vegetables above/below median</td>
<td>1.29 (1.09–1.53)</td>
<td>3.5 × 10⁻³</td>
<td>0.40</td>
<td>0.31</td>
<td>ARCTIC/DALS/PLCO/WHI/DACHS/ASTERISK/HPFS/NHS/PHS</td>
</tr>
<tr>
<td>rs3082042/11q23/Folate quartile medians</td>
<td>1.34 (1.08–1.67)</td>
<td>8.2 × 10⁻³</td>
<td>0.71</td>
<td>0.65</td>
<td>DALS/PLCO/WHI/HPFS/NHS</td>
</tr>
<tr>
<td>rs3082042/11q23/Red meat quartile medians</td>
<td>1.17 (1.04–1.32)</td>
<td>8.6 × 10⁻³</td>
<td>0.73</td>
<td>0.58</td>
<td>ARCTIC/DALS/PLCO/WHI/HPFS/NHS/PHS</td>
</tr>
</tbody>
</table>

NOTE: Permutation-based significance threshold, 3.75 × 10⁻⁴; Bonferroni-based significance threshold, 2.78 × 10⁻⁴. Abbreviations: ORINT, multiplicative interaction OR from EB; 95% CI for ORinteraction nom.p, nominal P value; adj.p, adjusted P value based on permutations; het.p, Phetogeneity. Studies, studies included in estimating the interaction term.
The results for the 180 gene–environment interactions tested are presented in Supplementary Table S4. Six SNP/environmental factor interactions showed nominal \( P < 0.01 \) Table 3; forest plots for individual study results are in Supplementary Fig. S1). The lowest nominal \( P \) value was for rs16892766, with vegetables as quartiles \( \text{OR}_{\text{interaction}} = 1.88, 95\% \text{CI: 1.36–2.59}; \text{nominal } P_{\text{interaction}} = 1.3 \times 10^{-3} \). rs16892766 has a MAF of 0.1 in the CEU population and is located on chromosome 8q23.3. This was the only finding with an adjusted \( P < 0.05 \) (adjusted \( P = 0.02 \)). Because of potential correlations between the environmental factors tested, we used permutations methods to adjust for multiple comparisons. A Bonferroni correction assumes the tests are independent. For the permutations, the cutoff that corresponds to a family wise error rate of 0.05 can be calculated by taking the 5th percentile of the minimum of \( P \) values of all tests across all permutation runs. For our data, it was \( 3.75 \times 10^{-4} \), slightly less conservative than the Bonferroni cutoff 0.05/180 = \( 2.87 \times 10^{-4} \). The rs16892766/vegetable consumption interaction was statistically significant with either correction. This same SNP had a nominal \( P_{\text{interaction}} < 0.01 \) for processed fiber as quartiles \( (\text{nominal } P_{\text{interaction}} = 6.0 \times 10^{-4}; \text{adjusted } P = 0.09) \) and for vegetables dichotomized at sex- and study-specific medians \( (\text{nominal } P_{\text{interaction}} = 3.5 \times 10^{-3}; \text{adjusted } P = 0.40) \). The correlation between vegetable quartiles and fiber quartiles in this data set was 0.65. Table 4 shows the association with CRC risk in strata defined by quartiles of vegetable consumption. The magnitude of the main effect of the minor (C) allele for this SNP increased with increasing levels of vegetable consumption, ranging from no evidence for association \( (\text{OR} = 0.94; 95\% \text{CI: 0.77–1.15}; \text{nominal } P = 0.54) \) in the lowest quartile to a relatively strong association for a common genetic factor \( (\text{OR} = 1.40; 95\% \text{CI: 1.13–1.74}; \text{nominal } P = 0.002) \) in the highest quartile. Results of the pooled analysis showing associations for vegetables in strata defined by levels of the SNP, and the combined association in strata defined by rs16892766 genotype and vegetable consumption are shown in Supplementary Materials (Supplemental Tables S5 and S6).

The rs16892766/vegetable-consumption results were not altered when we adjusted for additional covariates; interaction OR (adjusted for ever-smoked, BMI, alcohol use, red meat, and processed meat consumption) = 1.90 (95% CI: 1.35–2.67; nominal \( P_{\text{interaction}} = 2.48 \times 10^{-4} \)). A similar magnitude of interaction was seen using traditional case–control logistic analysis \( (\text{OR}_{\text{interaction}} = 1.79; 95\% \text{CI: 1.23–2.59}; \text{nominal } P_{\text{interaction}} = 2.3 \times 10^{-3} \).

In supplementary analyses of all GxE interactions stratified by cancer site (colon vs. rectum; Supplementary Table S7), the strongest statistical evidence for gene–environment interaction among colon cancer cases were for the same rs16892766/vegetable-consumption \( (\text{OR}_{\text{interaction}} = 1.79; 95\% \text{CI: 1.28–2.51}; \text{nominal } P_{\text{interaction}} = 6.5 \times 10^{-4} \) and rs16892766/fiber \( (\text{OR}_{\text{interaction}} = 1.31; 95\% \text{CI: 1.12–1.53}; \text{nominal } P_{\text{interaction}} = 9.8 \times 10^{-4} \) interactions observed for the combined CRC. For rectal cases, with a smaller sample size, the rs16892766/vegetable-consumption interaction was not statistically significant \( (\text{interaction OR for rectal cancer } = 1.51; 95\% \text{CI: 0.57–4.03}; \text{interaction } P = 0.41) \), and the only interaction with nominal \( P < 0.01 \) was rs4779584 and dietary calcium \( (\text{nominal } P = 6.7 \times 10^{-3} \).

### Table 4. Main effect of rs16892766 overall and by quartiles of vegetable consumption

<table>
<thead>
<tr>
<th>Group</th>
<th>OR* (95% CI)</th>
<th>( P^a )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>1.17 (1.08–1.27)</td>
<td>1.6 \times 10^{-4}</td>
</tr>
<tr>
<td>By vegetable quartiles</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quartile 1</td>
<td>0.94 (0.77–1.15)</td>
<td>0.541</td>
</tr>
<tr>
<td>Quartile 2</td>
<td>1.19 (0.96–1.47)</td>
<td>0.114</td>
</tr>
<tr>
<td>Quartile 3</td>
<td>1.26 (1.02–1.55)</td>
<td>0.029</td>
</tr>
<tr>
<td>Quartile 4</td>
<td>1.40 (1.13–1.74)</td>
<td>2.2 \times 10^{-3}</td>
</tr>
</tbody>
</table>

*OR and \( P \) values for log-additive model; ORs represent each additional copy of minor (C) allele for rs16892766.

The rs16892766 SNP is in an LD region on chromosome 8q23.3. Two studies have fine-mapped this region in relation to CRC risk (38, 39). Both found the strongest signals for a cluster of 5 SNPs, including rs16892766, that are in high LD; Pittman and colleagues also identified a sixth SNP in the cluster that is not in the public databases (39). Neither study found evidence for secondary independent signals in this region. The eukaryotic translation initiation factor 3 subunit H (\( \text{EIF3H} \)) gene is the closest gene to this cluster, with the closest gene to this cluster, with the identified SNPs ~140kb downstream from the gene transcript. Initial functional studies indicated that the rs16892766 region interacts with the \( \text{EIF3H} \) promoter and represses gene expression (39); however, a subsequent examination of ENCODE data and eQTLs suggests that the variants in this region may be influencing expression levels of the neighboring UTP23, small subunit processome component, homolog (yeast; \( \text{UTP23} \)) gene, rather than \( \text{EIF3H} \) itself. The variants may also impact expression of both genes (38). Additional work is needed to elucidate the functional relationship between \( \text{EIF3H} \) or \( \text{UTP23} \), or both, and CRC etiology. As the functional role of this SNP and other variants in the region is unknown, we cannot currently make informed speculations on how it might relate to vegetable consumption.

Vegetable consumption has long been hypothesized to be protective against CRC (40), although epidemiologic studies are not fully consistent [see review in (41)]. A recent meta-analysis of 1,694,236 participants including 16,057 colorectal cases with data on vegetable consumption from prospective
cohort studies found a statistically significant nonlinear inverse association between both fruit and vegetable intake with CRC risk and the summary relative risk for the highest versus lowest intake for vegetables was 0.91 (95% CI: 0.86–0.96; ref. 42). The postulated mechanisms have primarily focused on vegetables as a source of fiber and micronutrients, including folate (43). We also observed some evidence for interaction between the rs16892766 SNP and quartiles of both fiber intake (ORinteraction = 1.33; 95% CI (1.13–1.56); Pinteraction = 6.0 × 10−4; adjusted P = 0.09), and dietary folate intake (ORinteraction = 1.34; 95% CI: 1.08–1.67; Pinteraction = 8.2 × 10−3; adjusted P = 0.71). As with vegetable consumption, the pattern was for an increased risk associated with the minor (C) allele at higher levels of consumption. Vegetable consumption shows a positive correlation with both fiber (correlation = 0.65) and folate (correlation = 0.49) in these studies and it is difficult to disentangle the different measures using reported dietary-intake measures. Future follow-up of this interaction could focus more specifically on biomarkers for different dietary components.

Although we did not observe statistically significant evidence for heterogeneity in the rs16892766/vegetable-consumption interaction, we did observe minor evidence for heterogeneity for the main effect of the rs16892766 SNP in the full sample (P heterogeneity = 0.030). We considered the possibility that the underlying GxE interaction may have been contributing to the observed heterogeneity. However, we observed similar evidence of heterogeneity for the main effect of rs16892766 in strata defined by levels of vegetable consumption (P heterogeneity = 0.02–0.20). These results indicate that the minor level of observed heterogeneity for this SNP did not result from the rs16892766/vegetable-consumption interaction. Additional avenues would need to be explored for the source of this potential heterogeneity in association.

Previous studies of CRC risk have reported potential interactions with the 10 known loci in relation to age, family history, and sex (2, 5, 7, 17, 44, 45); however, the results have been inconsistent. Additional studies have looked at GxE for a broader range of environmental factors (14–18), but ours is the first to report a statistically significant interaction between rs16892766 and vegetable consumption. Using the DALS study, Slattery and colleagues observed an interaction between rs4939827, on 18q21 near the SMAD7 gene, and recent aspirin/NSAID use (16). We observed evidence for that interaction in the DALS study alone (Pinteraction = 0.03). However, we did not observe evidence for this association across the other studies, including analysis restricted to colon cancer only. This may reflect differences in how aspirin/NSAID use was collected across studies (Supplementary Table S2): for example, the time frame was 2 years before diagnosis for DALS and the other case-control studies, whereas for the cohort studies, baseline data describe a variable number of years before diagnosis. It might also reflect other underlying differences among the studies, a false positive in the initial report, or a false negative in the present study. Using a discovery set not included in this report, Figueiredo and colleagues examined GxE interactions for these same 10 loci with more than 10 environmental factors in a sample of 1,191 and 999 unrelated population-based controls (18). They observed several suggestive gene–environment interactions, although none were replicated in an independent sample that overlaps with the ARCTIC sample used in this article. Furthermore, that analysis was restricted to MSS/MSI-L CRC cases, which have different environmental risk factors (46) and, therefore, perhaps different underlying gene–environment interactions than the more broadly defined CRC cases used in this study.

Strengths of this study include the large sample size and standardized harmonization. We adopted a flexible approach to retrospective harmonization, using methods similar to those proposed by other projects (47, 48). Not every study was included for some of the environmental factors considered, either because they did not collect that particular variable or because they did not collect information in a way that was considered inferentially equivalent. We limited our study to variables that could be combined across at least 50% of the studies and we used yes/no and study-specific quartiles as forms of variables. These forms are more easily comparable across studies. As in many epidemiologic studies, measurement error may be leading to false negatives. We may be missing interactions that would have been found through inclusion of other environmental factors, through different assessments of the environmental variables or through different models, including fully saturated models (35).

The lack of evidence for other GxE interactions for most loci identified through initial GWAS is similar to what has been observed in prostate and breast cancer (49–52). This is not surprising given that the loci were identified through large-scale discovery and replication. SNPs with a strong GxE might show more heterogeneity across studies and may be less likely to appear as the strongest marginal signals. A full examination of the role of gene–environment interactions in CRC will require large, well-powered, genome-wide investigations with well measured and harmonized environmental risk factor data.

Disclosure of Potential Conflicts of Interest

A.T. Chan is a consultant for Bayer Healthcare, Pfizer Inc., and Millennium Pharmaceuticals. A.Z. LaCroix is a scientific advisory board member for Amgen and the University of Massachusetts. No potential conflicts of interest were disclosed by the other authors.

Acknowledgments

The authors thank all participants from all studies included in this manuscript for making this work possible; from DACHS, cooperating clinicians, and Ute Handte-Daub, Mahabbet Celik, and Ursula Elber for excellent technical assistance; Patrice Soule and Hardeep Ram of the Dana Farber Harvard Cancer Center High-throughput Genotyping Core who assisted in the genotyping for NHS, HPFS, and PHS under the supervision of David J. Hunter; Carolyn Guo who assisted in programming for NHS and HPFS; Hanian Zhang who assisted in programming for the PHS; staff of the NHS and the HPFS, for their valuable contributions as well as the following state cancer registries for their help: AL, AZ, AR, CA, CO, CT, DE, FL, GA, ID, IL, IN, IA, KY, LA, ME, MD, MA, MI, NH, NJ, NY, NC, ND, OH, OK, OR, PA, RI, SC, TN, TX, VA, WA, WV. The authors also thank Drs. Christine Berg and Philip Prevec, Division of Cancer Prevention, at the National Cancer Institute, the screening center investigators and staff of the PLCO Cancer Screening Trial, Mr. Thomas Riley and staff at Information Management Services, Inc., and Ms. Barbara O'Brien and staff at Westat, Inc. for their contributions to the PLCO Cancer Screening Trial; WHI investigators and staff for their dedication. A full listing of WHI investigators can be found on the WHI website (53).

Published OnlineFirst February 24, 2012; DOI: 10.1158/0008-5472.CAN-11-4067
Grant Support
ARTIC was supported by a GL2 grant from the Ontario Research Fund, the Canadian Institutes of Health Research, and the Cancer Risk Evaluation (CaRe) Program grant from the Canadian Cancer Society Research Institute. T.J. Hudson and B.W. Zanke are recipients of Senior Investigator Awards from the Ontario Institute for Cancer Research, through generous support from the Ontario Ministry of Research.

ASTERISK was supported by a regional Hospital Clinical Research Program (PHRC) and supported by the Regional Council of Pays de la Loire, the Groupe-ment des Entreprises Francaises dans la Lutte contre le Cancer (GEFLUC), the Association Anne de Bretagne Génétique, and the Ligue Régionale Contre le Cancer (LRCRC).

CCFR was supported by the National Cancer Institute, NIH under RFA 09-024 and through cooperative agreements with members of the Colon Cancer Family Registry–Ontario (CCFR) Center, Ontario Registry for Studies of Familial CRC, contributed data to this manuscript and was supported by (U01 CA074785).

DACHS was supported by grants from the German Research Council (Deutsche Forschungsgemeinschaft, BR 1704/6-1, BR 1704/6-3, BR 1704/6-4, and CH 1171-1), and the German Federal Ministry of Education and Research (01K8B04 and 01EH0814).

DALS was supported by the National Cancer Institute, NIH, U.S. Department of Health and Human Services (R01 CA09898 to S.L. Slattery). Funding for the genome-wide scan of DALS, PLCO, and WHI was provided by the National Cancer Institute, National Institutes of Health, U.S. Department of Health and Human Services (R01 CA099045 to U. Peters). C.M. Hutter was supported by a training grant from the National Cancer Institute, National Institutes of Health, U.S. Department of Health and Human Services (U01 CA137088 to U. Peters). HPFS was supported by the NIH (P01 CA 055075 to C.S. Fuchs, R01 137178 to A.T. Chan, P50 CA 127003 to C.S. Fuchs, and P01 CA 087969 to E.L. Giovannucci) and NHS by the NIH (R01 CA059045 to U. Peters). C.M. Hutter was supported by the National Cancer Institute, NIH, U.S. Department of Health and Human Services (R25 CA094880). Additional funding for this work was provided by National Cancer Institute, National Institutes of Health, U.S. Department of Health and Human Services (U10 CA137088 to U. Peters).

The WHI program was supported by the National Heart, Lung, and Blood Institute, NIH, U.S. Department of Health and Human Services through contracts HHSN268201100001C-4C, HHSN268201100046C, and HHSN271201100004C, and NIMH421. 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 December 16, 2011; revised January 27, 2012; accepted February 13, 2012; published OnlineFirst February 24, 2012.

References
Characterization of Gene–Environment Interactions for Colorectal Cancer Susceptibility Loci

Carolyn M. Hutter, Jenny Chang-Claude, Martha L. Slattery, et al.


Updated version
Access the most recent version of this article at:
doi:10.1158/0008-5472.CAN-11-4067

Supplementary Material
Access the most recent supplemental material at:
http://cancerres.aacrjournals.org/content/suppl/2012/02/24/0008-5472.CAN-11-4067.DC1

Cited articles
This article cites 48 articles, 6 of which you can access for free at:
http://cancerres.aacrjournals.org/content/72/8/2036.full#ref-list-1

Citing articles
This article has been cited by 10 HighWire-hosted articles. Access the articles at:
http://cancerres.aacrjournals.org/content/72/8/2036.full#related-urls

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, use this link
http://cancerres.aacrjournals.org/content/72/8/2036.
Click on "Request Permissions" which will take you to the Copyright Clearance Center's (CCC) Rightslink site.