

Phenotypic and tumor molecular characterization of colorectal cancer in relation to a susceptibility *SMAD7* variant associated with survival

Xabier Garcia-Albeniz^{1,2}, Hongmei Nan^{2,3}, Linda Valeri⁴,
Teppei Morikawa^{5,6}, Aya Kuchiba^{6,7}, Amanda I.Phipps⁸,
Carolyn M.Hutter⁸, Ulrike Peters⁸, Polly A.Newcomb⁸,
Charles S.Fuchs^{2,6}, Edward L.Giovannucci^{1,2,7},
Shuji Ogino^{1,5,6,9,†} and Andrew T.Chan^{2,10,*}

¹Department of Epidemiology, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA, ²Department of Medicine, Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA, ³Department of Epidemiology and Public Health, Division of Cancer Epidemiology, University of Maryland School of Medicine, Baltimore, MD 21201, USA, ⁴Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA, ⁵Center for Molecular Oncologic Pathology, Dana-Farber Cancer Institute, Boston, MA 02215, USA, ⁶Department of Medical Oncology, Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA 02215, USA, ⁷Department of Nutrition, Harvard School of Public Health, Boston, MA 02115, USA, ⁸Cancer Prevention Program, Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA, ⁹Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA and ¹⁰Division of Gastroenterology, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA

*To whom correspondence should be addressed. Tel: +1-617-726-3212; email: achan@partners.org

The minor allele (G) of rs4939827, a *SMAD7* (18q21) intronic variant, is associated with a lower risk of developing colorectal cancer (CRC) and poorer survival after diagnosis. Our objective was to evaluate the associations of this variant with different tumor phenotype and intratumoral molecular characteristics. We evaluated 1509 CRC cases and 2307 age-matched controls nested within the Nurses' Health Study and the Health Professionals Follow-up Study. We used the TaqMan assay to genotype rs4939827 and logistic regression to assess the association of rs4939827 with risk of CRC according to different phenotypic and molecular characteristics. We found that the minor allele (G) in rs4939827 (*SMAD7*, 18q21) was associated with a lower risk of developing tumor stage pT1 or pT2 CRC [multivariate odds ratio (OR), 0.73; 95% confidence interval (CI) 0.62–0.87] but not tumor stage pT3 or pT4 (multivariate OR, 1.07; 95% CI 0.93–1.23, *P* for heterogeneity = 1.2×10^{-4}). The association between rs4939827 and CRC also significantly differed by methylation of *RUNX3* (*P* for heterogeneity = 0.005). Among those with CRC, the minor allele (G) in rs4939827 was significantly associated with poorer overall survival (hazards ratio, 1.20; 95% CI, 1.02–1.42). We can conclude that the minor allele (G) of the germline intronic *SMAD7* variant rs4939827 is associated with a lower risk of CRC with earlier tumor stage and CRC without methylation of the tumor suppressor *RUNX3*. These findings suggest that individuals with this *SMAD7* variant that develop CRC are more probably to have tumors with greater invasiveness and methylation of *RUNX3*, which potentially contributes to their poorer observed survival.

Abbreviations: BH, Benjamini and Hochberg; BMI, body mass index; CI, confidence interval; CIMP, CpG island methylator phenotype; CRC, colorectal cancer; GWAS, genome-wide association study; HPFS, Health Professionals Follow-up Study; HR, hazards ratio; LINE-1, long interspersed nucleotide element-1; MPE, molecular pathological epidemiology; MSI, microsatellite instability; MSS, microsatellite stable; NHS, Nurses' Health Study; NSAIDs, non-steroidal anti-inflammatory drugs; OR, odds ratio; RERI, relative excess risk due to interaction.

[†]These authors contributed equally to this work.

Introduction

Genome-wide association studies (GWAS) have identified 20 colorectal cancer (CRC) susceptibility variants, corresponding to 16 loci (1–8). One of the earliest identified and most consistently validated variants, rs4939827 (18q21.1), resides in an intronic region of the gene *SMAD* family member 7 (*SMAD7*), which is associated with the TGFB1 (transforming growth factor- β) pathway. The minor allele (G) in rs4939827 is associated with a lower risk of CRC, with the most recent GWAS observing an odds ratio (OR) of 0.88 [95% confidence interval (CI) 0.85–0.93, *P*-value = 1.1×10^{-7}] (9). Recently, in a pooled analysis of five prospective cohorts, we observed that the G allele is also associated with poorer survival [hazards ratio (HR) = 1.16, *P*-value = 0.02] (10). Based on these results, we hypothesized that rs4939827 may be differentially associated with CRC subtypes associated with worsened outcomes or are related to the TGFB1 pathway, including methylation of *RUNX3*. Thus, we examined the association of rs4939827 with CRC risk according to specific clinical or molecular phenotypes within the Nurses' Health Study (NHS) and the Health Professionals Follow-up Study (HPFS), two of the prospective cohorts included in our prior analysis (10).

Materials and methods

Study population

The NHS was initiated in 1976, when 121 700 female USA registered nurses between the ages of 30 and 55 years returned a mailed health questionnaire. The HPFS was established in 1986 with a parallel cohort of 51 529 USA male dentists, pharmacists, optometrists, osteopath physicians, podiatrists and veterinarians between the ages of 40 and 75 years. In both cohorts, we have subsequently updated information biennially with greater than 90% follow-up. In 1989–90, we collected blood samples from 32 826 participants in the NHS cohort. In 1993–95, we collected blood samples from 18 159 participants in HPFS. The samples were collected in heparinized tubes and sent to us by overnight courier in chilled containers. In 2001–04, 29 684 women in NHS and 13 956 men in HPFS mailed in a 'swish-and-spit' sample of buccal cells. Participants who provided buccal cells had not previously provided a blood specimen. On receipt, blood and buccal cells were centrifuged, aliquoted and stored at -70°C .

In both cohorts, we requested permission to obtain medical records and pathology reports from participants who reported CRC on our biennial questionnaires. We identified fatal cases from the National Death Index and from next-of-kin (11). Study physicians blinded to exposure data reviewed all medical records to confirm cases of CRC. As described previously, we randomly selected between one and three controls within the same cohort from participants who were free of CRC at the same time the CRC was diagnosed in the cases (12). We matched controls with blood specimens (1013 in NHS and 673 in HPFS) to cases with blood specimens (483 in NHS and 372 in HPFS); similarly, controls with buccal specimens (423 in NHS and 198 in HPFS) were separately matched to cases with buccal specimens (439 in NHS and 215 in HPFS). Controls were matched to each case on ethnicity, year of birth and month/year of blood or buccal sampling (12). In total, 1509 (922 in the NHS and 587 in the HPFS) cases and 2307 (1436 in the NHS and 871 in the HPFS) controls were included in our analysis. The institutional review boards at Brigham and Women's Hospital and the Harvard School of Public Health approved this study.

Genotyping

The SNP rs4939827 was genotyped by the 5' nuclease assay (TaqMan®), using the ABI PRISM 7900HT Sequence Detection System (Applied Biosystems, Foster City, CA). TaqMan® primers and probes were designed using the Primer Express® Oligo Design software v2.0 (ABI PRISM). Laboratory personnel were blinded to case-control status, and 10% blinded quality control samples (duplicated samples) were inserted to validate genotyping procedures; concordance for the blinded quality control samples was 100%. Primers, probes and conditions for genotyping are available upon request. We successfully genotyped rs4939827 in 98% of samples in NHS and 99.6% of samples in HPFS. We confirmed that rs4939827 was in Hardy-Weinberg Equilibrium among the controls (χ^2 *P*-values = 0.93 for NHS and 0.85 for HPFS).

Pathological assessment of molecular characteristics

Beginning in 1997 in the HPFS and 2001 in the NHS, we began retrieving, from the pathology departments of treating hospitals, available pathological specimens from participants whom we confirmed had received a diagnosis of CRC up to 2006 (13). Among the 1509 cases with blood or buccal samples in this study, we were able to successfully obtain tissue suitable for molecular analyses in 658 cases. Real-time PCR (MethylLight) was used for quantitative DNA methylation to determine CpG island methylator phenotype (CIMP) status using DNA extracted from paraffin-embedded tissue. We quantified DNA methylation in eight CIMP-specific promoters (including *RUNX3*) as detailed elsewhere (14). CIMP-high was defined as >6 of 8 methylated markers using the eight-marker CIMP panel, and CIMP-low/0 was defined as 0 of 8 to 5 of 8 methylated markers, according to the previously established criteria (15). Microsatellite instability (MSI) analysis was performed using 10 microsatellite markers (D2S123, D5S346, D17S250, BAT25, BAT26, BAT40, D18S55, D18S56, D18S67 and D18S487). MSI-high was defined as the presence of instability in >30% of the markers, and MSI-low/microsatellite stable (MSS) as instability in 0–29% of markers (15). Long interspersed nucleotide element-1 (LINE-1) methylation level was assessed by pyrosequencing using the PyroMark kit and the PSQ HS 96 System (Qiagen, Valencia, CA). The average of the relative amounts of C in the four CpG sites evaluated was used as overall LINE-1 methylation level in a given sample (16). We performed PCR and pyrosequencing targeted for *KRAS* (codons 12 and 13), *BRAF* (codon 600) and *PIK3CA* (exons 9 and 20 (15,17,18)). TP53 expression was assessed by immunohistochemistry as detailed elsewhere (19).

Statistical analysis

We used logistic regression to estimate OR and corresponding 95% CIs for the association of variant rs4939827 with CRC among subgroups defined by tumor phenotype (T3-4 and T1-2; N1-2 and N0; M1 and M0; poorly differentiation and moderately well differentiation; rectum and colon location and age of diagnosis <60 and >60 years) and by molecular characteristics [*RUNX3* promoter methylation and absence of methylation; *KRAS* mutant and wild type (WT); *BRAF* mutant and wild type; *PIK3CA* mutant and wild type; LINE-1 methylation-low and methylation-high; MSI-H and MSI-L/MSS]. We obtained similar results using unconditional or conditional logistic regression adjusting for matching factors (data not shown). Thus, we present unconditional regression models adjusting for matching factors and other known or suspected risk factors. The Cochran's chi-square-based Q statistic test was used to assess the extent of heterogeneity across the two studies. Because there was little evidence for heterogeneity for the association of the SNP rs4939827 with CRC risk between women and men (P for heterogeneity = 0.32), we pooled data from the two studies.

We modeled each SNP using a log-additive approach, relating genotype dose (i.e. number of copies of the minor allele) to risk of CRC. We adjusted all analyses for age at sample collection, race, gender, regular aspirin use (yes

or no), regular non-steroidal anti-inflammatory drugs (NSAIDs) use (yes or no), body mass index (BMI; in tertiles), physical activity (in tertiles), history of CRC in a parent or sibling (yes or no), smoking status (never, former or current smoker), alcohol consumption (0–4.9, 5–9.9, 10–14.9 or ≥ 15.0 g per day), consumption of beef, pork or lamb as a main dish (0–3 times per month, once a week, 2–4 times per week or ≥ 5 times per week), energy-adjusted calcium and folate intake (in tertiles) and type of sample (blood versus cheek). To assess heterogeneity in the association between rs4939827 and tumors according to clinical phenotype or molecular characteristics, we used a case–case design using logistic regression model comparing tumor subtypes (20).

We performed mediation analyses as described in (21,22). In brief, this approach decomposes the total effect of the exposure (rs4939827) on the outcome (T3-T4 tumors) into a 'direct effect' plus an 'indirect effect.' The 'direct effect' consists of the effect of rs4939827 on T-stage at a fixed level of the mediator variable [i.e. the direct effect can be interpreted as the OR comparing the risk of T3-T4 tumor stage with the genetic variant present versus absent if the mediator (e.g. *RUNX3*) were what it would have been without the genetic variant]. The 'indirect effect' is the effect on the outcome (T-stage) of changes of the exposure (rs4939827) that operate through the methylation status of *RUNX3*. (The indirect effect can be interpreted as the OR for T3-T4 tumor stage for those with the genetic variant present comparing the risk if the mediator were what it would have been with versus without the genetic variant.) This approach allows for the adjustment of covariates and for the presence of interaction between the variables of interest (21). We computed additive interaction in the form of relative excess risk due to interaction (RERI) using the delta method (23) and multiplicative interaction by entering a product term in the model and assessing its significance by the Wald method.

For analyses of rs4929827 in relation to survival, we confined the cohort to the 646 incident cases of CRC diagnosed after collection of blood or buccal cells. We calculated the time from CRC diagnosis to death from CRC, death from any cause or the end of follow-up (1 June 2010), whichever came first. For CRC-specific survival, deaths from other causes were censored at the time of death. We used Cox proportional hazards regression to calculate HRs and 95% CI for the association between the SNP and survival, adjusting for age, race, sex (cohort), tumor stage, grade of differentiation, regular aspirin use, smoking status at diagnosis, alcohol consumption, consumption of beef, pork or lamb as a main dish and energy-adjusted calcium and folate intake as described above. The proportional hazards assumption was verified by plotting the cumulative martingale residuals and assessing for significance. We corrected for multiple comparisons with several tumor subtypes using the Benjamini and Hochberg (BH) false discovery rate adjustment (24). We corrected for the 14 assessed interactions; the values of the corrected P -values, when significant, are presented as footnotes in the tables. SAS V9.2 (SAS Institute, Cary, NC) was used for the analysis.

Table 1. Baseline characteristics of cases and controls^a

	NHS (women)		HPFS (men)		Total	
	Cases	Controls	Cases	Controls	Cases	Controls
	(n = 922)	(n = 1436)	(n = 587)	(n = 871)	(n = 1509)	(n = 2307)
Age at diagnosis [years, mean (SD)]	66.8 (9.2)		69.2 (9.2)		67.7 (9.3)	
Age at sample draw (years, mean)	59.5 (6.6)	59.6 (6.5)	64.8 (8.5)	64.9 (8.4)	62.5 (8.2)	62.1 (7.9)
Non-white (%)	13 (1.4)	6 (0.4)	49 (8.4)	59 (6.8)	62 (4.1)	65 (2.5)
Regular aspirin use (%) ^b	324 (35.1)	656 (45.7)	269 (45.8)	450 (51.7)	593 (39.3)	1106 (47.9)
Regular NSAID drug use (%) ^c	303 (32.9)	570 (39.7)	136 (23.2)	190 (21.8)	439 (29.1)	760 (32.9)
BMI [kg/m ² , mean (SD)] ^d	26.1 (5.1)	26.0 (5.0)	26.2 (3.3)	25.6 (3.3)	26.1 (4.5)	25.8 (4.5)
Physical activity [METs - h/week, mean (SD)] ^e	15.7 (14.3)	16.0 (14.0)	29.4 (24.2)	31.1 (25.2)	21.1 (20.0)	21.8 (20.5)
CRC in a parent or sibling (%)	212 (23.0)	239 (16.6)	114 (19.4)	132 (15.2)	326 (21.6)	371 (16.1)
Former or current smoker (%)	537 (59.0)	791 (55.0)	319 (60.0)	441 (55.0)	856 (59.0)	1232 (55.0)
Alcohol consumption [g/day, mean (SD)]	6.8 (10.2)	6.3 (9.4)	13.4 (15.7)	11.8 (13.2)	9.4 (13.1)	8.4 (11.3)
Beef, pork or lamb as a main dish [servings/week, mean (SD)]	2.8 (0.8)	2.8 (0.8)	2.7 (0.9)	2.6 (0.9)	2.7 (0.8)	2.7 (0.8)
Total calcium intake (mg/day, mean) ^f	952.4 (364.1)	1007.5 (386.9)	916.1 (399.4)	950.6 (382.5)	938.1 (378.7)	985.8 (386.2)
Total folate intake (μ g/day, mean) ^f	426.8 (179.1)	447.8 (194.9)	526.0 (226.2)	566.2 (231.6)	465.9 (204.7)	492.9 (217.3)

^aAt blood draw or cheek cell sampling.

^bRegular aspirin user was defined as the consumption of at least two 325 mg tablets per week in the NHS and at least two times per week in the HPFS. Non-regular user was defined otherwise.

^cRegular NSAID user was defined as the consumption of NSAIDs at least two times per week. Non-regular user was defined otherwise.

^dThe BMI is the weight in kilograms divided by the square of the height in meters.

^eMetabolic Equivalent of Task (MET) denotes metabolic equivalent. Met (h) = sum of the average time/week in each activity \times MET value of each activity. One MET, the energy spent sitting quietly, is equal to 3.5 ml of oxygen uptake per kilograms of body weight per minute for a 70 kg adult.

^fNutrient values (calcium and folate) represent the mean of energy-adjusted intakes.

Results

Among our total study population of 1509 CRC cases and 2307 controls, the minor allele (G) frequency for rs4939827 was 48.6%, consistent with the HapMap CEU population. The baseline characteristics of the cases and controls are presented in Table I. Compared with controls, CRC cases in both women and men were less probably to have regularly used aspirin or NSAIDs, have a family history of CRC, were more probably to smoke, consumed higher amounts of alcohol and had lower intakes of calcium and folate. Among men, cases had a higher BMI and were less physically active. Among women, cases were less probably to have used postmenopausal hormones. Although the minor allele (G) of rs4939827 was not significantly associated with CRC in our total population, the trend was in the direction previously reported (multivariate OR, 0.93; 95% CI 0.84–1.04, $P = 0.22$).

We confirmed our prior findings relating rs4939827 with survival among cases of CRC (10). Among the 646 cases of CRC diagnosed

after blood or buccal cell collection, we observed that the minor allele of rs4939827 was significantly associated with poorer overall survival (multivariate HR, 1.20; 95% CI 1.02–1.42, $P = 0.031$) and a trend toward poorer CRC-specific survival (HR = 1.20, 95% CI 0.98–1.47, $P = 0.08$).

To determine if the survival difference we observed according to the G allele was related to a differential association with tumor phenotypes, we examined the influence of rs4939827 on CRC susceptibility according to traditional clinicopathological features in our total study population. The association of the G allele of rs4929827 with CRC risk appeared to differ according to pT stage or the depth of invasion of the tumor (P for heterogeneity = 1.0×10^{-4}). The G allele of rs4929827 was inversely associated with a risk of tumors staged as pT1 or pT2 (multivariate OR, 0.73; 95% CI 0.62–0.87, P -value = 2.8×10^{-4}). In contrast, there was no association between rs4939827 and tumors staged as pT3 or pT4 (multivariate OR, 1.07; 95% CI 0.93–1.23, P -value = 0.38). Hence, the prevalence of tumors

Table II. Associations of rs4939827 with tumor phenotypes^a

	Controls	Cases	OR (95% CI)	Cases	OR (95% CI)	Case-only analysis
		pT1/pT2		pT3/pT4		pT3/pT4 versus pT1/pT2
TT	600	155	1	200	1	1
TG	1120	241	0.84 (0.65–1.10)	363	1.19 (0.93–1.53)	1.54 (1.10–2.15)
GG	538	95	0.51 (0.36–0.72)	182	1.14 (0.85–1.53)	2.31 (1.50–3.56)
P trend			2.8×10^{-4}		0.38	1.0×10^{-4b}
G allele ^c			0.73 (0.62–0.87)		1.07 (0.92–1.23)	1.52 (1.23–1.88)
		N0		N1-2		N1-2 versus N0
TT	600	236	1	114	1	1
TG	1120	407	1.10 (0.87–1.38)	191	0.88 (0.65–1.20)	0.79 (0.55–1.14)
GG	538	190	0.84 (0.63–1.11)	81	0.71 (0.49–1.04)	0.83 (0.53–1.31)
P trend			0.27		0.08	0.36
G allele			0.93 (0.81–1.06)		0.85 (0.70–1.02)	0.90 (0.72–1.13)
		M0		M1		M1 versus M0
TT	600	358	1	40	1	1
TG	1120	622	1.04 (0.85–1.27)	68	0.90 (0.56–1.44)	0.88 (0.54–1.44)
GG	538	287	0.84 (0.66–1.07)	26	0.70 (0.39–1.26)	0.78 (0.41–1.46)
P trend			0.20		0.24	0.43
G allele			0.93 (0.82–1.04)		0.84 (0.63–1.12)	0.88 (0.64–1.20)
		Moderately or well-differentiated tumors		Poorly differentiated tumors		Poorly versus moderately well differentiated tumors
TT	600	299	1	56	1	1
TG	1120	494	1.00 (0.81–1.24)	104	1.01 (0.66–1.54)	1.08 (0.68–1.69)
GG	538	235	0.85 (0.66–1.10)	50	0.95 (0.57–1.57)	1.28 (0.74–2.21)
P trend			0.23		0.84	0.39
G allele			0.93 (0.82–1.05)		0.98 (0.76–1.25)	1.13 (0.86–1.48)
		Colon cancer		Rectal cancer		Rectum versus colon
TT	600	118	1	36	1	1
TG	1120	233	1.14 (0.86–1.52)	66	1.03 (0.65–1.64)	0.90 (0.55–1.49)
GG	538	90	0.84 (0.59–1.18)	28	0.83 (0.48–1.45)	0.88 (0.48–1.62)
P trend			0.36		0.54	0.66
G allele			0.92 (0.78–1.09)		0.92 (0.70–1.20)	0.93 (0.69–1.27)
		Age of diagnosis >60 years		Age of diagnosis <60 years		<60 versus >60 years
TT	600	337	1	80	1	1
TG	1120	598	1.04 (0.85–1.28)	134	0.96 (0.65–1.40)	0.83 (0.52–1.33)
GG	538	263	0.83 (0.65–1.07)	67	0.95 (0.61–1.48)	1.19 (0.68–2.08)
P trend			0.18		0.81	0.63
G allele			0.92 (0.82–1.04)		0.97 (0.78–1.21)	1.07 (0.81–1.42)

^aORs and P -values are adjusted for age at sample collection, race, gender, regular aspirin use, regular NSAIDs use, BMI, physical activity, history of CRC in a parent or sibling, smoking status, alcohol consumption, consumption of beef, pork or lamb as a main dish, energy-adjusted calcium and folate intake and type of sample (more details in the text).

^bBH-adjusted P -value = 0.0014.

^cLog-additive model, representing the OR for each additional G allele as compared with TT.

with pT3–4 stage was 65.7% for cases of CRC with two variant alleles, 60.1% for cases of CRC with one variant allele, compared with 56.3% for cases of CRC with no variant alleles of rs4939827 (Mantel–Haenszel P -value = 0.018). The rs4939827 SNP was not significantly differentially associated with any of other evaluated clinical and pathological phenotypes, including nodal status, presence of distant metastases, tumor grade, anatomic location and age at diagnosis (Table II).

We also examined rs4939827 genotype according to molecular subtypes among 658 cases for whom we had available tumor tissue data. Because rs4939827 is located in an intronic region of *SMAD7*, which modulates TGF β 1 signaling, we evaluated rs4939827 according to promoter methylation of the *RUNX3* gene, which is related to the TGF β 1 pathway. The G allele of rs4939827 was associated with a lower CRC risk in tumors without methylation of *RUNX3* (multivariate OR, 0.84; 95% CI 0.71–0.99, P = 0.035). In contrast, there was a suggestive, although not statistically significant, increased risk of tumors with methylation of *RUNX3* associated with the minor allele of rs4939827 (multivariate OR, 1.27; 95% CI 0.94–1.74, P = 0.12) (Table III). The difference between rs4939827 and risk of tumors with methylation of *RUNX3* compared with tumors without methylation of *RUNX3* was statistically significant after adjustment for multiple comparisons (P for heterogeneity = 0.0053, BH-corrected P -value = 0.037). The risk of CRC associated with rs4939827 also appeared to be modified by *BRAF* mutational status. However, the P for heterogeneity was 0.021, non-significant after adjustment for multiple comparisons (BH-corrected P -value = 0.10). The association between the rs4939827 SNP and CRC did not significantly differ according to any of the other markers examined, including CIMP status, LINE-1 methylation, *KRAS* mutation, *PIK3CA* mutation, MSI status and TP53 mutation.

RUNX3 methylation was associated with a higher risk of pT3 and pT4 tumors (multivariate OR, 2.12; 95% CI 1.18–3.80, P = 0.012) even after adjusting for rs4939827. The multivariate ORs estimating the indirect effect of rs4939827 on risk of pT3 and pT4 tumors was 1.05 (95% CI 0.96–1.15, P = 0.32) for one variant allele and 1.50 (95% CI 0.44–5.10, P = 0.51) for two variant alleles (Table IV). Given the width of the CIs, we cannot exclude a possible indirect effect of rs4939827 through *RUNX3* methylation on pT stage. The multivariate ORs that estimate the direct effect of rs4939827 were 1.96 (95% CI 1.18–3.25, P -value = 0.009) for one variant allele and 4.46 (95% CI 1.19–16.6, P -value = 0.038) for two variant alleles. There was no evidence of interaction between the *SMAD7* variant rs4939827 and *RUNX3* methylation in either the additive scale (RERI = 0.48; 95% CI –1.42 to 2.37, P -value = 0.62) or multiplicative scale (OR, 0.93; 95% CI 0.39–2.19, P -value = 0.87).

Discussion

In this large study of 1509 cases of CRC and 2307 controls nested in two prospective cohorts, the CRC-susceptibility locus *SMAD7* intronic rs4939827 was associated with a lower risk of CRCs with pT1 and pT2 stage but not pT3 and pT4 stage. Hence, among individuals diagnosed with CRC, those with at least one G allele had 1.5-fold higher odds of having a pT3 or pT4 tumor compared with a pT1 or pT2 tumor. This finding may explain in part observed associations of the G allele of rs4939827 with lower overall risk of incident CRC (9), but worsened survival after diagnosis (10,25).

This study represents a strategy of investigations, which has been recently termed ‘molecular pathological epidemiology’ (MPE) (20,26). MPE is conceptually based on inherent heterogeneity of a disease, which is typically regarded as a single entity in traditional epidemiology studies, including GWAS (20,26). MPE can decipher how risk factors are associated with specific alterations in molecular pathways in cancer. Moreover, the MPE design can be used to help shed insight into the function of susceptibility variants identified by GWAS based upon their association with specific molecular subtypes of CRC (20).

The seemingly contrary associations of this SNP in *SMAD7* with decreased risk of incident disease but poorer outcomes in patients with established disease may be due to *SMAD7*’s known involvement in modulating the TGF β 1 pathway. In normal epithelium, TGF β 1 functions as a tumor suppressor through induction of cell arrest and inhibition of cell proliferation. However, once cells are resistant to TGF β 1-mediated proliferative inhibition (i.e. in established tumors), TGF β 1 appears to promote metastasis by enhancing angiogenesis and extracellular matrix disruption and inhibiting infiltrating tumor immune cells. A role for rs4939827 in the TGF β 1 pathway is further supported by our findings of significant association of rs4939827 with *RUNX3* methylation status in the tumors. *RUNX3* is a Runt domain transcription factor 3 involved in TGF β 1 signaling by interaction with SMAD transcription factors and is considered a suppressor of solid tumors. The *RUNX3* promoter is commonly aberrantly methylated in a CpG island in CRC (27–29), leading to gene inactivation.

Differential methylation of *RUNX3* according to rs498327 could plausibly explain the worsened survival among individuals with CRC who have a variant G allele (30). In addition to TGF β 1, *RUNX3* is involved in WNT regulation by forming a ternary complex with TCF4 and CTNNB1 (β -catenin) (31). Both TGF β 1 and WNT signaling are involved in the induction of epithelial-mesenchymal transition, a process that mediates invasion and metastasis in CRC. However, our mediation analysis does not suggest that the association between rs4939827 on risk of CRC according to pT stage is primarily related to *RUNX3*. Nonetheless, the mediation analysis should be interpreted in the context of some limitations (22), including the assumption of a lack of other, strong unmeasured confounders.

Previous studies of CRC-susceptibility variants and tumor phenotype have observed only a few significant differential associations between rs4939827 and CRC according to other clinical phenotype or molecular features. A case-only analysis of 1531 cases did not find any association between this variant and pT3–4 compared with pT1–2-staged tumors (P = 0.94) (32). In a study of 1096 patients from the Epicolon I cohort (33), the G allele of rs4939827 was significantly associated with well-differentiated tumors under a log-additive model (HR, 0.67; P -value = 0.027). However, this association was not replicated in a validation cohort, EPICOLON II (34). Neither of these two studies found any association between rs4939827 and other tumor features, including grade of differentiation, stage or tumor site. On the other hand, two studies have observed that the G allele of rs4939827 increased the risk of harboring a rectal tumor as compared with colon tumors (35,36).

We also examined several other molecular markers other than *RUNX3* and did not observe strong differences in the association with rs4939827. Our findings are consistent with Slattery *et al.*, who did not observe significant heterogeneity in associations of rs4939827 with CIMP status, MSI, *KRAS* or p53 mutational status (36). Nonetheless, in our study, it is notable that *BRAF*-status did appear to be differentially associated with rs4939827, despite the lack of a significant P for heterogeneity after correction for multiple comparisons. An interplay between TGF β 1 and *BRAF* mutations is biologically plausible (37), and *BRAF* mutations are strongly associated with CRC prognosis. Thus, given the low frequency of *BRAF* mutations in CRC, it is possible that statistically significant heterogeneity may be evident with a larger sample size.

Our study has several strengths. First, we used prospectively collected, biennially updated, detailed data on CRC risk factors over long-term follow-up. This permitted us to account for the main potential confounders of our associations. Second, our case-control study was nested within two large well-characterized cohorts within which matched controls were selected from the same cohort from which the case developed, minimizing the likelihood of population stratification or selection bias. Third, our findings were consistent between two independent cohorts. Moreover, we have previously demonstrated a consistent association between rs498327 and CRC and overall survival that has been validated by results in other populations (10). Last, among a large number of participants, we had germline DNA as well as available tumor tissue to examine the expression of molecular

Table III. Association of rs4939827 with tumor molecular characteristics^a

	Controls	Cases	OR (95% CI)	Cases	OR (95% CI)	Case-only analysis
		<i>RUNX3</i> promoter unmethylated		<i>RUNX3</i> promoter methylated		Methylation versus unmethylation
TT	600	142	1	25	1	1
TG	1120	254	1.07 (0.82–1.40)	63	1.40 (0.79–2.50)	1.43 (0.74–2.78)
GG	538	95	0.66 (0.47–0.94)	35	1.65 (0.87–3.12)	2.97 (1.39–6.35)
P trend			0.04		0.12	0.0053 ^b
G allele ^c			0.84 (0.71–0.99)		1.27 (0.94–1.74)	1.74 (1.18–2.56)
		WT <i>BRAF</i>		Mutant <i>BRAF</i>		Mutant versus WT <i>BRAF</i>
TT	600	162	1	19	1	1
TG	1120	286	1.05 (0.81–1.35)	41	1.09 (0.53–2.24)	1.86 (0.77–4.48)
GG	538	126	0.80 (0.58–1.09)	24	1.64 (0.75–3.58)	3.17 (1.19–8.47)
P trend			0.19		0.21	0.02
G allele			0.90 (0.78–1.05)		1.29 (0.86–1.93)	1.77 (1.09–2.88)
		LINE-1 methylation-high		LINE-1 methylation-low		Methylation-high versus low
TT	600	108	1	67	1	1
TG	1120	210	1.23 (0.91–1.66)	115	0.92 (0.63–1.33)	0.80 (0.51–1.27)
GG	538	101	1.13 (0.80–1.60)	43	0.57 (0.35–0.93)	0.55 (0.31–1.00)
P trend			0.48		0.03	0.05
G allele			1.06 (0.90–1.26)		0.78 (0.62–0.98)	0.75 (0.56–1.00)
		CIMP-low/negative		CIMP-high		CIMP-high versus low/negative
TT	600	146	1	21	1	1
TG	1120	267	1.07 (0.82–1.40)	50	1.53 (0.78–3.00)	1.65 (0.76–3.59)
GG	538	108	0.76 (0.55–1.06)	24	1.41 (0.65–3.05)	2.27 (0.93–5.56)
P trend			0.15		0.40	0.07
G allele			0.89 (0.76–1.04)		1.17 (0.81–1.68)	1.50 (0.97–2.34)
		MSI-L/MSS		MSI-H		MSI-H versus MSI-L/MSS
TT	600	159	1	21	1	1
TG	1120	268	1.00 (0.77–1.29)	54	1.63 (0.85–3.14)	1.93 (0.95–3.90)
GG	538	120	0.79 (0.58–1.08)	26	1.47 (0.70–3.09)	1.99 (0.87–4.55)
P trend			0.16		0.33	0.10
G allele			0.90 (0.77–1.04)		1.19 (0.84–1.68)	1.40 (0.94–2.07)
		WT <i>KRAS</i>		Mutant <i>KRAS</i>		Mutant versus WT <i>KRAS</i>
TT	600	113	1	69	1	1
TG	1120	215	1.24 (0.91–1.67)	113	0.86 (0.60–1.22)	0.69 (0.44–1.09)
GG	538	93	0.94 (0.65–1.36)	55	0.77 (0.51–1.18)	0.87 (0.50–1.51)
P trend			0.85		0.23	0.51
G allele			0.98 (0.82–1.17)		0.88 (0.71–1.09)	0.91 (0.69–1.20)
		WT <i>PIK3CA</i>		Mutant <i>PIK3CA</i>		Mutant versus WT <i>PIK3CA</i>
TT	600	139	1	31	1	1
TG	1120	251	1.07 (0.81–1.40)	57	0.94 (0.56–1.58)	0.85 (0.48–1.52)
GG	538	120	0.90 (0.65–1.25)	18	0.62 (0.31–1.23)	0.64 (0.30–1.37)
P trend			0.58		0.19	0.26
G allele			0.96 (0.82–1.12)		0.81 (0.58–1.11)	0.81 (0.56–1.17)
		TP53 negative		TP53 positive		TP53 positive versus negative
TT	600	76	1	67	1	1
TG	1120	151	1.14 (0.81–1.61)	120	1.15 (0.78–1.69)	0.83 (0.50–1.39)
GG	538	66	0.86 (0.56–1.31)	51	0.84 (0.52–1.36)	0.81 (0.43–1.55)
P trend			0.54		0.56	0.50
G allele			0.94 (0.77–1.15)		0.93 (0.74–1.17)	0.90 (0.65–1.24)

^aORs and *P*-values are adjusted for age at sample collection, race, gender, regular aspirin use, regular NSAIDs use, BMI, physical activity, history of CRC in a parent or sibling, smoking status, alcohol consumption, consumption of beef, pork or lamb as a main dish, energy-adjusted calcium and folate intake and type of sample (more details in the text).

^bBH-adjusted *P*-value = 0.0371.

^cLog-additive model, representing the OR for each additional G allele as compared with TT.

markers, including *RUNX3*, which is directly relevant to the TGFB1 pathway. This permitted us to examine rs4939827 in relation to tumor subtypes with greater mechanistic specificity.

We acknowledge the limitations to our study. First, we did not have tumor specimens available for analysis for all of our cases. However, the risk factors in cases with available tumor tissue did not appreciably differ

Table IV. Mediation and interaction analysis of *RUNX3* methylation in the association of rs4939817 with pT stage

Association with pT3/4 stage	OR (95% CI)		P-value	
	2.12 (1.18–3.80)		0.0120	
Mediation analysis	TG versus TT		GG versus TT	
	OR (95% CI)	P-value	OR (95% CI)	P-value
Direct effect	1.96 (1.18–3.25)	0.0088	4.46 (1.19–16.6)	0.038
Indirect effect	1.05 (0.96–1.15)	0.32	1.50 (0.44–5.10)	0.51
Marginal total effect	2.06 (1.22–3.46)	0.0068	6.70 (0.89–50.35)	0.064
Proportion mediated, %	9.3		39.2	
Interaction analysis	RERI ^a (95% CI)		P-value	
	0.48 (–1.42 to 2.37)		0.62	
Multiplicative	OR (95% CI)		P-value	
	0.93 (0.39–2.19)		0.87	

^aRERI (also known as ‘interaction contrast ratio’).

from those in cases without tumor tissue (38). Second, although we did correct for multiple comparisons in our primary analyses, we cannot rule out the possibility that the associations we observed with rs4939827 and tumor subtypes represent false-positive findings. However, the biological plausibility of our findings and the consistent association of this allele with CRC survival increase the likelihood that our observed associations are true. Third, we did include some cases of CRC in which participants provided DNA samples after diagnosis. However, for analyses of survival, we restricted our cohort to incident cases diagnosed after DNA was collected to minimize any effect of survival bias.

In summary, individuals with the rs4939827 CRC-susceptibility locus diagnosed with CRC tend to develop tumors with greater invasiveness (pT stage). Moreover, we also observed a differential association of rs4939827 with *RUNX3* methylation status, supporting an effect of rs4939827 or causal variants tagged by this SNP on carcinogenesis mediated through the TGFBI pathway. These findings support present understanding of the dual function of TGFBI/SMAD7 pathways in inhibiting early tumorigenesis yet facilitating metastasis. Taken together, these results could explain, at least in part, the lower risk of CRC associated with the G allele of rs4939827 yet poorer survival (10).

Funding

National Institutes of Health (NIH) / National Cancer Institute (NCI) (grant CA137178, CA059045, CA151993, CA055075, CA087969, CA127003, CA094880, CA154337); ‘La Caixa’ fellowship to X.G.A.; A.T.C. is a Damon Runyon Cancer Foundation Clinical Investigator.

Acknowledgements

We wish to acknowledge Patrice Soule and Hardeep Ranu for genotyping at the Dana-Farber Harvard Cancer Center High Throughput Polymorphism Core, under the supervision of Immaculata Devivo, as well as Carolyn Guo for programming assistance. We also thank the participants of the Health Professionals Follow-up Study and Nurses’ Health Study and the following state cancer registries: AL, AZ, AR, CA, CO, CT, DE, FL, GA, ID, IL, IN, IA, KY, LA, ME, MD, MA, MI, NE, NH, NJ, NY, NC, ND, OH, OK, OR, PA, RI, SC, TN, TX, VA, WA and WY.

Contribution

All the authors have contributed significantly to the submitted work and have read and approved the final version of this manuscript.

Conflict of Interest Statement: None declared.

References

- Tenesa, A. *et al.* (2010) Ten common genetic variants associated with colorectal cancer risk are not associated with survival after diagnosis. *Clin. Cancer Res.*, **16**, 3754–3759.
- Houlston, R.S. *et al.* (2010) Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. *Nat. Genet.*, **42**, 973–977.
- Houlston, R.S. *et al.* (2008) Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat. Genet.*, **40**, 1426–1435.
- Tomlinson, I.P. *et al.* (2008) A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nat. Genet.*, **40**, 623–630.
- Tomlinson, I.P. *et al.* (2011) Multiple common susceptibility variants near BMP pathway loci GREM1, BMP4, and BMP2 explain part of the missing heritability of colorectal cancer. *PLoS Genet.*, **7**, e1002105.
- Kocarnik, J.D. *et al.* (2010) Characterization of 9p24 risk locus and colorectal adenoma and cancer: gene-environment interaction and meta-analysis. *Cancer Epidemiol. Biomarkers Prev.*, **19**, 3131–3139.
- Hutter, C.M. *et al.* (2010) Characterization of the association between 8q24 and colon cancer: gene-environment exploration and meta-analysis. *BMC Cancer*, **10**, 670.
- Dunlop, M.G. *et al.* (2012) Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk. *Nat. Genet.*, **44**, 770–776.
- Peters, U. *et al.* (2012) Meta-analysis of new genome-wide association studies of colorectal cancer risk. *Hum. Genet.*, **131**, 217–234.
- Phipps, A.I. *et al.* (2012) Association between colorectal cancer susceptibility loci and survival time after diagnosis with colorectal cancer. *Gastroenterology*, **143**, 51–4.e4.
- Stampfer, M.J. *et al.* (1984) Test of the National Death Index. *Am. J. Epidemiol.*, **119**, 837–839.
- Pai, J.K. *et al.* (2004) Inflammatory markers and the risk of coronary heart disease in men and women. *N. Engl. J. Med.*, **351**, 2599–2610.
- Morikawa, T. *et al.* (2011) Association of CTNNB1 (beta-catenin) alterations, body mass index, and physical activity with survival in patients with colorectal cancer. *JAMA*, **305**, 1685–1694.
- Ogino, S. *et al.* (2007) Evaluation of markers for CpG island methylator phenotype (CIMP) in colorectal cancer by a large population-based sample. *J. Mol. Diagn.*, **9**, 305–314.
- Ogino, S. *et al.* (2009) CpG island methylator phenotype, microsatellite instability, BRAF mutation and clinical outcome in colon cancer. *Gut*, **58**, 90–96.
- Ogino, S. *et al.* (2008) A cohort study of tumoral LINE-1 hypomethylation and prognosis in colon cancer. *J. Natl. Cancer Inst.*, **100**, 1734–1738.
- Ogino, S. *et al.* (2005) Sensitive sequencing method for KRAS mutation detection by Pyrosequencing. *J. Mol. Diagn.*, **7**, 413–421.
- Ogino, S. *et al.* (2009) PIK3CA mutation is associated with poor prognosis among patients with curatively resected colon cancer. *J. Clin. Oncol.*, **27**, 1477–1484.
- Morikawa, T. *et al.* (2012) Tumor TP53 expression status, body mass index and prognosis in colorectal cancer. *Int. J. Cancer*, **131**, 1169–1178.
- Ogino, S. *et al.* (2011) Molecular pathological epidemiology of colorectal neoplasia: an emerging transdisciplinary and interdisciplinary field. *Gut*, **60**, 397–411.
- Vanderweele, T.J. *et al.* (2010) Odds ratios for mediation analysis for a dichotomous outcome. *Am. J. Epidemiol.*, **172**, 1339–1348.

22. Valeri, L. *et al.* Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. *Psychol. Methods*, in press.
23. Knol, M.J. *et al.* (2007) Estimating interaction on an additive scale between continuous determinants in a logistic regression model. *Int. J. Epidemiol.*, **36**, 1111–1118.
24. Benjamini, Y. *et al.* (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B. Stat. Methodol.*, **57**, 289–300.
25. Passarelli, M.N. *et al.* (2011) Common colorectal cancer risk variants in SMAD7 are associated with survival among prediagnostic nonsteroidal anti-inflammatory drug users: a population-based study of postmenopausal women. *Genes. Chromosomes Cancer*, **50**, 875–886.
26. Ogino, S. *et al.* (2010) Lifestyle factors and microsatellite instability in colorectal cancer: the evolving field of molecular pathological epidemiology. *J. Natl. Cancer Inst.*, **102**, 365–367.
27. Ahlquist, T. *et al.* (2008) Gene methylation profiles of normal mucosa, and benign and malignant colorectal tumors identify early onset markers. *Mol. Cancer*, **7**, 94.
28. Soong, R. *et al.* (2009) The expression of RUNX3 in colorectal cancer is associated with disease stage and patient outcome. *Br. J. Cancer*, **100**, 676–679.
29. Subramaniam, M.M. *et al.* (2009) RUNX3 inactivation in colorectal polyps arising through different pathways of colonic carcinogenesis. *Am. J. Gastroenterol.*, **104**, 426–436.
30. Slattery, M.L. *et al.* (2011) Genetic variation in the transforming growth factor- β signaling pathway and survival after diagnosis with colon and rectal cancer. *Cancer*, **117**, 4175–4183.
31. Ito, K. *et al.* (2008) RUNX3 attenuates beta-catenin/T cell factors in intestinal tumorigenesis. *Cancer Cell*, **14**, 226–237.
32. Ghazi, S. *et al.* (2010) Colorectal cancer susceptibility loci in a population-based study: associations with morphological parameters. *Am. J. Pathol.*, **177**, 2688–2693.
33. Piñol, V. *et al.* (2005) Accuracy of revised Bethesda guidelines, microsatellite instability, and immunohistochemistry for the identification of patients with hereditary nonpolyposis colorectal cancer. *JAMA*, **293**, 1986–1994.
34. Abulí, A. *et al.* (2010) Susceptibility genetic variants associated with colorectal cancer risk correlate with cancer phenotype. *Gastroenterology*, **139**, 788–96, 796.e1.
35. Lubbe, S.J. *et al.* (2012) Relationship between 16 susceptibility loci and colorectal cancer phenotype in 3146 patients. *Carcinogenesis*, **33**, 108–112.
36. Slattery, M.L. *et al.* (2010) Increased risk of colon cancer associated with a genetic polymorphism of SMAD7. *Cancer Res.*, **70**, 1479–1485.
37. Fleming, Y.M. *et al.* (2009) TGF-beta-mediated activation of RhoA signaling is required for efficient (V12)HaRas and (V600E)BRAF transformation. *Oncogene*, **28**, 983–993.
38. Chan, A.T. *et al.* (2007) Aspirin and the risk of colorectal cancer in relation to the expression of COX-2. *N. Engl. J. Med.*, **356**, 2131–2142.

Received August 23, 2012; revised October 11, 2012; accepted October 18, 2012