Genetic variants within ultraconserved elements and susceptibility to right- and left-sided colorectal adenocarcinoma

Moubin Lin¹, Cathy Eng², Ernest T.Hawk³,
Maosheng Huang¹, Anthony J.Greisinger⁴, Jian Gu¹,
Lee M.Ellis^{5,6}, Xifeng Wu^{1,*} and Jie Lin¹

¹Department of Epidemiology, ²Department of Gastrointestinal Medical Oncology and ³Division of Cancer Prevention and Population Science, The University of Texas M. D. Anderson Cancer Center, Houston, TX 77030, USA, ⁴Kelsey Research Foundation, Houston, TX 77005, USA and ⁵Department of Surgical Oncology and ⁶Department of Cancer Biology, The University of Texas M. D. Anderson Cancer Center, Houston, TX 77030, USA

*To whom correspondence should be addressed. Department of Epidemiology, Unit 1340, The University of Texas M. D. Anderson Cancer Center, 1155 Pressler Blvd., Houston, TX 77030, USA. Tel: +1 713 745 2485; Fax: +1 713 792 4657; Email: xwu@mdanderson.org

We investigated whether single nucleotide polymorphisms within ultraconserved elements (UCEs) are associated with susceptibility to overall colorectal cancer (CRC) and susceptibility to tumor site-specific CRC. The study included 787 CRC patients and 551 healthy controls. The study comprised of a training set (520 cases and 341 controls) and a replication set (267 cases and 210 controls). We observed associations in rs7849 and rs1399685 with CRC risk. For example, a dose-dependent trend (per-allele odds ratio (OR), 0.78; 95% confidence interval (CI), 0.63–1.00; *P* for trend = 0.05) associated with the variant allele of rs7849 in the training set. The significant trend toward a decrease in CRC risk was confirmed in the replication set (per-allele OR, 0.72; 95% CI, 0.52–0.99; *P* for trend = 0.044). When stratified by tumor location, for left-sided CRC (LCRC) risk, significant association was observed for the variant-containing genotypes of rs1399685 (OR, 1.77; 95% CI, 1.02–3.06) and the risk was replicated in the replication population (OR, 2.04; 95% CI, 1.02–4.07). The variant genotypes of rs9784100 and rs7849 conferred decreased risk but the associations were not replicated. Three right-sided CRC (RCRC) susceptibility loci were identified in rs6124509, rs4243289 and rs12218935 but none of the loci was replicated. Joint effects and potential higher order gene–gene interactions among significant variants further categorized patients into different risk groups. Our results strongly suggest that several genetic variants in the UCEs may contribute to CRC susceptibility, individually and jointly, and that different genetic etiology may be involved in RCRC and LCRC.

Introduction

The carcinogenesis of colorectal cancer (CRC) is not well understood, but ~35% of all CRCs involve an inherited susceptibility (1). However, CRCs are heterogeneous and therefore need to be studied in discrete subsets that evolve through different genetic pathways (2). Accumulating evidence suggests that right-sided CRC (RCRC) and left-sided CRC (LCRC) relative to the splenic flexure can be viewed as two tumor entities with different epidemiological, pathological, molecular and clinical characteristics (3–5). These differences have been attributed to the two entities' differences in embryologic origin, anatomy, physiological function and genetic mechanisms, with microsatellite instability associated with RCRC and chromosomal instability associated with LCRC (6–8). However, site difference has seldom been considered in previous investigations of genetic risk factors for CRC.

Abbreviations: CART, classification and regression tree; CI, confidence interval; CRC, colorectal cancer; LCRC, left-sided CRC; OR, odds ratio; RCRC, right-sided CRC; SNP, single nucleotide polymorphism; UCE, ultraconserved element.

To date, studies of genetic predisposition to CRC have largely focused on single nucleotide polymorphisms (SNPs) in genes associated with metabolizing enzymes, DNA methylation, DNA repair, estrogen/androgen, inflammation, antioxidants and insulin (9). Bejerano *et al.* (10) discovered ultraconserved elements (UCEs), a unique category of potentially functional sequences (spanning at least 200 DNA base pairs) showing 100% identity among orthologous regions in human, mouse and rat species. The existence of UCEs for over 300 million years strongly suggests their functional role in the human genome (11). Additional studies have confirmed that UCEs function as long-range enhancers of flanking genes (12), splicing regulators (13), epigenetic modification (14) and transcriptional coactivators (15). UCEs have also been linked to cancer. Calin *et al.* (16) first discovered that transcribed UCEs had distinct expression signatures in human leukemias and carcinomas. Moreover, UCEs are frequently located at fragile sites and genomic regions involved in cancers (16). Given that variants within UCEs have been subjected to extreme purifying selection and have been conserved in humans over long evolutionary history, it is conceivable that few common SNPs within UCEs may harbor critical biological functions (17). To date, only six validated SNPs have been identified in 481 examined UCEs (10) and two (rs2056116 and rs9572903) of the six SNPs have been significantly associated with familial breast cancer risk (18).

Given that CRCs have distinct expression signatures of transcribed UCEs compared with normal tissue (16), we hypothesized that genetic variants within UCEs may modulate the risk of CRC. To test this hypothesis, we conducted a case–control study and systematically evaluated 48 SNPs, individually and jointly, within 481 UCEs for associations with CRC risk. Moreover, we further assessed whether the genetic polymorphisms conferred different risks for RCRC and LCRC.

Materials and methods

Study population and epidemiological data

We included 787 patients with histologically confirmed CRC and 551 healthy controls in this case–control study. CRC patients were enrolled from The University of Texas M. D. Anderson Cancer Center between January 1990 and June 2008. There were no restrictions on age, gender, ethnicity or cancer stage on participation. Of the 787 patients, 520 patients were incident cases of CRC diagnosed within 1 year of enrollment and the remaining 267 patients had CRCs that were diagnosed for >1 year before the patients were referred to M. D. Anderson Cancer Center. The control subjects were identified from an existing pool of control subjects enrolled in ongoing case–control studies of cancer. Briefly, the control subjects were recruited from Kelsey-Seybold Clinic, the largest private multispecialty group practice in the Houston metropolitan area, with 18 clinics and >325 physicians and over 400 000 patients. The majority of the control subjects were healthy individuals seen at the clinic for their annual physical exams. The rationale of recruiting controls from Kelsey-Seybold Clinic has been discussed previously (19). Kelsey-Seybold Clinic staff distributed a brief questionnaire about the study to potential control subjects when they arrived for annual physical exams. The questionnaire was used to elicit the patients' willingness to be contacted by staff at M. D. Anderson and to collect preliminary demographic data for matching. The potential control subjects were contacted by telephone at a later date to confirm their willingness to participate and to schedule an in-person interview at a Kelsey-Seybold Clinic convenient to the participant. On the day of the interview, the controls visited the clinic specifically for this study and not for any treatment purposes. The population covered by Kelsey-Seybold Clinic well represents the base population from which the majority of cases arise. Control subjects had no prior history of cancer (except for non-melanoma skin cancer). For incident cases (CRC diagnosed within 1 year of enrollment) and controls, epidemiological data were collected and recorded in a structured questionnaire. For prevalent cases (CRCs that were diagnosed for >1 year), demographic data were obtained from patients' medical records. All study participants donated a small blood sample for molecular analysis. We obtained

written informed consent from all patients and control subjects prior to their participation and donation of blood samples. The institutional review boards at M. D. Anderson and Kelsey-Seybold Clinic approved this study.

SNP selection and genotyping

We obtained information on the sequence position for each UCE from Supplementary data from a study by Bejerano *et al.* (<http://www.cse.ucsc.edu/~jill/ultra.html>) (10) and the UCbase and miRfunc database (<http://microrna.osu.edu/UCbase4>) (20). Then, we used the International HapMap Project (<http://www.hapmap.org>) to identify 141 tagging SNPs spaced across the 481 UCEs. Using the LDSelect program (version 1.0) (<http://droog.gs.washington.edu/ldSelect.html>) and the UCSC Golden Path Gene Sorter program (<http://genome.ucsc.edu>), we further identified 54 SNPs based on the selection criteria of pairwise linkage disequilibrium with an r^2 threshold of 0.8 and a minor allele frequency cutoff of 0.05 in Caucasian. The complete SNP set was then submitted to Illumina (San Diego, CA) technical support for further exclusion of SNPs with low Illumina quality design scores (<0.6). Finally, we selected 48 SNPs for genotyping in this study (Table I). Genomic DNA was isolated from peripheral blood using the QIAamp DNA Blood Maxi Kit (Qiagen, Germantown, MD). Genotyping was performed with the Golden Gate assay of the VeraCode technology using the Illumina BeadXpress Reader System according to the

manufacturer's protocol. Raw data were analyzed using Illumina Genome Studio software. Ten samples were genotyped in duplicate and showed 100% concordance. The mean call rate for the SNP array was 99.9%. One SNP, rs4412572, failed in genotyping for all samples due to low signal strength and was therefore not included in further analyses.

Statistical analysis

Statistical analyses were performed using Stata software (version 10; StataCorp LP, College Station, TX). Pearson's χ^2 test or Fisher's exact test was used to compare the difference in the distribution of categorical variables. Differences in continuous variables were tested using the Student's *t*-test. The Hardy-Weinberg equilibrium of alleles at individual locus was assessed using the χ^2 goodness-of-fit test. CRC risk associated with individual SNPs was estimated as odds ratios (ORs) and 95% confidence intervals (CIs) using multivariate logistic regression adjusted for potential confounding variables where appropriate. For each SNP, we tested three genetic models (specifically, the dominant, recessive and additive models). The model with the most significant *P* value was considered as the best fitting model. For polymorphisms with a frequency of homozygous minor allele $<5\%$, we only considered the dominant model that had the highest statistical power. We divided the study population into a training data set to guide data analysis and a validation set to

Table I. SNPs in UCEs selected for this study

Gene name (gene symbol)	SNP ID	Chromosome	UCE element	Position	Major/minor allele	Minor allele frequency (%)
<i>HS2ST1</i>	rs10747335	1	uc.29	Intron	C/A	0.23
<i>PKN2</i>	rs10493807	1	uc.31	5'-FR	G/A	0.08
<i>C1orf110</i>	rs4412572	1	uc.38	5'-FR		
<i>RGS4</i>	rs11580679	1	uc.39	5'-FR	A/G	0.38
<i>RGS4</i>	rs16847292	1	uc.39	5'-FR	A/G	0.07
<i>NUF2</i>	rs10917804	1	uc.40	3'-FR	A/G	0.08
<i>C1orf75</i>	rs7533689	1	uc.41	5'-FR	G/A	0.19
<i>LOC730134</i>	rs10211390	2	uc.54	3'-FR	C/G	0.36
<i>BCL11A</i>	rs12473113	2	uc.57	3'-FR	G/A	0.09
<i>BCL11A</i>	rs9784100	2	uc.60	3'-FR	G/C	0.39
<i>FLJ16124</i>	rs2954963	2	uc.65	3'-FR	G/A	0.27
<i>SFXN5</i>	rs2421099	2	uc.66	Intron	A/T	0.14
<i>LOC730124</i>	rs786255	2	uc.79	5'-FR	A/G	0.39
<i>LOC728773</i>	rs1399685	2	uc.81	3'-FR	T/A	0.06
<i>LOC728304</i>	rs12619842	2	uc.92	5'-FR	C/G	0.17
<i>PDK1</i>	rs6710129	2	uc.99	Intron	A/G	0.20
<i>ZBTB20</i>	rs16822925	3	uc.119	Intron	C/A	0.14
	rs9838168	3	uc.126		G/A	0.44
<i>RSRC1</i>	rs11713363	3	uc.131	Intron	A/G	0.26
<i>C5orf36</i>	rs13154972	5	uc.172	3'-FR	A/G	0.17
<i>EBF1</i>	rs4921445	5	uc.175	Intron	A/G	0.13
<i>C7orf30</i>	rs199657	7	uc.208	5'-FR	G/C	0.40
<i>LOC442660</i>	rs774265	7	uc.213	Intron	G/A	0.15
<i>SHFM1</i>	rs6953983	7	uc.220	5'-FR	G/A	0.36
<i>EBF2</i>	rs9942838	8	uc.235	Intron	G/A	0.50
<i>LOC347119</i>	rs7033100	9	uc.264	3'-FR	G/A	0.32
<i>LHX6</i>	rs1467737	9	uc.278	Intron	A/G	0.47
<i>ZNF503</i>	rs12782308	10	uc.287	5'-FR	G/C	0.14
<i>PKD2L1</i>	rs2305386	10	uc.294	Intron	G/A	0.05
<i>SCD</i>	rs7849	10	uc.298	3' UTR	A/G	0.17
<i>TECTB</i>	rs11195893	10	uc.310	3'-FR	G/A	0.10
<i>RAB11FIP2</i>	rs12218935	10	uc.311	3'-FR	A/G	0.46
<i>MGMT</i>	rs1711662	10	uc.318	Intron	G/A	0.28
<i>DLG2</i>	rs3815988	11	uc.331	Intron	A/G	0.35
<i>HNT</i>	rs6590611	11	uc.334	Intron	C/A	0.33
<i>LOC647277</i>	rs1395351	13	uc.351	3'-FR	A/G	0.29
<i>ARHGEF7</i>	rs9560010	13	uc.357	Intron	G/A	0.17
<i>AKAP6</i>	rs8007042	14	uc.367	5'-FR	C/A	0.12
<i>AKAP6</i>	rs1956211	14	uc.369	Intron	A/G	0.22
<i>MAP2K5</i>	rs8037887	15	uc.391	Intron	A/C	0.15
<i>IKZF3</i>	rs13313561	17	uc.410	Intron	G/C	0.05
<i>TCF4</i>	rs12455881	18	uc.436	Intron	G/A	0.16
<i>ZNF407</i>	rs4243289	18	uc.437	Intron	G/C	0.44
<i>RBL1</i>	rs6124509	20	uc.455	3'-FR	A/G	0.14
<i>FAM48B1</i>	rs16983007	X	uc.465	5'-FR	G/A	0.08
<i>LOC729188</i>	rs1029496	X	uc.466	3'-FR	A/G	0.14
<i>PDK3</i>	rs10482283	X	uc.469	5'-FR	G/A	0.19
<i>GRIA3</i>	rs1293524	X	uc.481	5'-FR	A/T	0.11

replicate findings from the training set. In particular, the training set comprised the incident cases (CRC patients diagnosed within 1 year of recruitment) and controls and the replication set comprised of CRC patients who had been diagnosed >1 year and the controls. In single-SNP analysis, genotypes associated with increased CRC risk in the training set and replicated in the replication set were defined as unfavorable genotypes. Significant SNPs in single-SNP analysis were included in the analysis of assessing joint effects of unfavorable genotypes and this analysis was performed in replication set only. Gene–gene interactions were evaluated using classification and regression tree (CART) analysis implemented in the HelixTree Genetics Analysis software (version 4.1.0; Golden Helix, Bozeman, MT). All *P* values were two-sided with a significance level of 0.05.

Results

Characteristics of the study population were shown in Supplementary Table 1, available at *Carcinogenesis* Online. In the training set, the mean ages of cases and controls were 59.69 (± 10.75) and 58.48 (± 13.120), respectively. Among the 517 cases, 303 (88.86%) were non-Hispanic White, 21 (6.16%) were Hispanic and 17 were Black (4.99%). One hundred and sixty-one (47.35%) patients were never smokers, 131 (38.53%) were former smokers and 48 (14.12%) were current smokers. By tumor site, 366 (70.79%) had LCRC and 151 (29.21%) had RCRC. In the replication set, 73 (27.3%) were RCRC patients and 185 (69.3%) were LCRC patients. In the replication set, similar percent distributions were observed in gender, ethnicity, smoking status and tumor location. In both training and validation sets, there was no significant differences in ethnicity between cases and controls. We performed all analyses in Caucasians only and in all study subjects. Since the results were similar, we only reported results on all subjects.

Significant deviations from the Hardy–Weinberg equilibrium were found for rs1029496, rs10482283, rs16983007 and rs1293524 in controls (*P* < 0.0001), and these SNPs were therefore excluded from further analysis. In individual SNP analysis, we found that four SNPs (rs7849, rs9784100, rs1399685 and rs6124509) showed significant or borderline associations with CRC risk (Table II). The most significant association was rs7849 in a recessive model, which was associated

with reduced risk among subjects carrying the homozygous variant genotype (OR, 0.37; 95% CI, 0.18–0.75; *P* = 0.006). The variant-containing genotype of both rs9784100 and rs6124509 conferred reduced CRC risks with borderline significance, with ORs of 0.76 (95% CI, 0.57–1.02; *P* = 0.067) and 0.75 (95% CI, 0.54–1.02; *P* = 0.069), respectively. In addition, the variant genotype of rs1399685 conferred a borderline increased risk (OR, 1.66; 95% CI, 0.99–2.80; *P* = 0.056). Also shown in Table II, the decreased CRC risk conferred by the variant allele of rs7849 showed a dose-dependent trend (per-allele OR, 0.78; 95% CI, 0.63–1.00; *P* for trend = 0.05) in the training set. The significant trend toward a decrease in CRC risk was confirmed in the replication set (per-allele OR, 0.72, 95% CI, 0.52–0.99; *P* for trend = 0.044) and in the pooled analysis (per-allele OR, 0.77; 95% CI, 0.63–0.93; *P* for trend = 0.008). The borderline association conferred by rs1399685 was replicated with an OR of 2.08 (95% CI, 1.10–3.94; *P* = 0.025). The OR was 1.74 (95% CI, 1.16–2.60; *P* = 0.007) in the pooled analysis. However, the associations with the other SNPs were not replicated (Table II).

We then stratified the risk by tumor location, i.e. RCRC and LCRC. For RCRC, a significant decrease in CRC risk was observed for the variant-containing genotype of rs6124509 compared with the wild-type genotype (OR, 0.62; 95% CI, 0.39–1.00; *P* = 0.049). Conversely, an increase in CRC risk was seen for the homozygous variant genotypes of rs4243289 and rs12218935 with ORs of 1.69 (95% CI, 1.05–2.72; *P* = 0.03) and 1.78 (95% CI, 1.10–2.89; *P* = 0.019), respectively. None of the associations were replicated in the replication set or in pooled analysis (Table III). For LCRC, significantly increased risk was observed for variant genotype of rs1399685 (OR, 1.77; 95% CI, 1.02–3.06; *P* = 0.041) and this increased risk was validated in the replication population (OR, 2.04; 95% CI, 1.02–4.07; *P* = 0.043). In the pooled analysis, the OR was 1.76 (95% CI, 1.15–2.70; *P* = 0.009). Significant reduced risk was observed in the variant-containing genotype of rs9784100 (OR, 0.66; 95% CI, 0.48–0.90; *P* = 0.009) as well as per-variant allele risk (OR, 0.78; 95% CI, 0.63–0.97; *P* for trend = 0.027) (Table IV). However, the significant associations were not replicated. In pooled analysis, both ORs for variant-containing genotype and per-variant allele reached statistical

Table II. Associations between SNPs within UCEs and risk for CRC

Genotype	Training set				Replication set				Pooled set	
	Control (<i>N</i> = 341)	Case (<i>N</i> = 520)	OR (95% CI) ^a	<i>P</i> value	Control (<i>N</i> = 210)	Case (<i>N</i> = 267)	OR (95% CI)	<i>P</i> value	OR (95% CI) ^a	<i>P</i> value
rs7849										
AA	217	352	Reference		130	190	Reference		Reference	
AG	101	154	0.95 (0.70–1.30)	0.744	68	63	0.58 (0.37–0.90)	0.014	0.82 (0.64–1.06)	0.126
GG	23	14	0.36 (0.18–0.74)	0.005	12	14	0.75 (0.32–1.73)	0.499	0.51 (0.30–0.86)	0.012
AG + GG versus AA			0.84 (0.63–1.13)	0.248			0.61 (0.40–0.91)	0.016	0.77 (0.61–0.97)	0.030
AA + AG versus GG			0.37 (0.18–0.75)	0.006			0.89 (0.39–2.03)	0.780	0.54 (0.32–0.91)	0.021
<i>P</i> for the trend			0.78 (0.63–1.00)	0.005			0.72 (0.52–0.99)	0.044	0.77 (0.63–0.93)	0.008
rs9784100										
GG	115	211	Reference		64	89	Reference		Reference	
GC	165	224	0.74 (0.55–1.02)	0.062	103	136	0.89 (0.58–1.37)	0.600	0.80 (0.62–1.02)	0.074
CC	61	85	0.81 (0.54–1.22)	0.308	43	42	0.77 (0.44–1.34)	0.350	0.78 (0.56–1.08)	0.135
GC + CC versus GG			0.76 (0.57–1.02)	0.067			0.86 (0.57–1.29)	0.450	0.79 (0.63–1.00)	0.052
GG + GC versus CC			0.95 (0.66–1.38)	0.797			0.82 (0.50–1.34)	0.430	0.89 (0.66–1.19)	0.430
<i>P</i> for the trend			0.87 (0.71–1.06)	0.166			0.88 (0.67–1.15)	0.350	0.87 (0.74–1.02)	0.081
rs1399685										
TT	317	465	Reference		194	230	Reference		Reference	
TA	23	53	1.63 (0.97–2.74)	0.067	15	35	2.17 (1.13–4.19)	0.020	1.74 (1.16–2.61)	0.008
AA	0	1			1	2	0.90 (0.07–10.89)	0.936	1.79 (0.18–17.53)	0.616
TA + AA versus TT			1.66 (0.99–2.80)	0.056			2.08 (1.10–3.94)	0.025	1.74 (1.16–2.60)	0.007
rs6124509										
AA	241	392	Reference		141	187	Reference		Reference	
AG	90	118	0.77 (0.56–1.07)	0.117	65	69	0.81 (0.53–1.24)	0.330	0.78 (0.61–1.01)	0.059
GG	10	9	0.52 (0.20–1.32)	0.169	4	9	2.17 (0.60–7.81)	0.240	0.83 (0.40–1.71)	0.611
AG + GG versus AA			0.75 (0.54–1.02)	0.069			0.88 (0.58–1.32)	0.530	0.79 (0.61–1.01)	0.056

Bold represents significant ORs and *P* values.

^aAdjusted by age, gender, ethnicity and smoking status.

Table III. Associations between SNPs within UCEs and risk for RCRC

SNP	Training set				Replication set				Pooled set	
	Control, N = 341	Case, N = 151	OR (95% CI) ^a	P value	Control, N = 210	Case, N = 73	OR (95% CI)	P value	OR (95% CI) ^a	P value
rs6124509										
AA	241	119	Reference		141	51	Reference		Reference	
AG	90	28	0.60 (0.36–0.98)	0.042	65	21	0.94 (0.51–1.74)	0.853	0.69 (0.47–1.01)	0.053
GG	10	4	0.86 (0.25–2.97)	0.816	4	1	1.05 (0.10–10.68)	0.967	0.92 (0.31–2.69)	0.880
AG + GG versus AA			0.62 (0.39–1.00)	0.049			0.95 (0.52–1.73)	0.863	0.70 (0.49–1.02)	0.061
rs4243289										
GG	101	42	Reference		65	23	Reference		Reference	
GC	177	69	0.95 (0.59–1.53)	0.838	100	36	1.00 (0.53–1.89)	0.992	0.97 (0.67–1.42)	0.882
CC	63	40	1.64 (0.94–2.88)	0.083	45	14	0.93 (0.41–2.10)	0.866	1.38 (0.88–2.18)	0.160
GC + CC versus GG			1.13 (0.72–1.76)	0.599			0.98 (0.54–1.78)	0.954	1.08 (0.76–1.54)	0.658
GG + GC versus CC			1.69 (1.05–2.72)	0.030			0.93 (0.45–1.91)	0.845	1.41 (0.96–2.08)	0.084
P for the trend			1.26 (0.95–1.68)	0.109			0.97 (0.65–1.44)	0.883	1.16 (0.92–1.46)	0.201
rs12218935										
AA	120	49	Reference		54	24	Reference		Reference	
AG	159	58	0.92 (0.57–1.47)	0.718	104	32	0.65 (0.34–1.26)	0.203	0.82 (0.56–1.19)	0.290
GG	62	44	1.69 (0.97–2.95)	0.064	52	17	0.66 (0.30–1.46)	0.306	1.23 (0.78–1.92)	0.370
AG + GG versus AA			1.12 (0.72–1.73)	0.624			0.66 (0.36–1.21)	0.176	0.93 (0.65–1.32)	0.678
AA + AG versus GG			1.78 (1.10–2.89)	0.019			0.86 (0.43–1.71)	0.663	1.39 (0.94–2.05)	0.099
P for the trend			1.27 (0.96–1.68)	0.100			0.80 (0.53–1.19)	0.265	1.08 (0.86–1.36)	0.498

Bold represents significant ORs and *P* values.

^aOR Adjusted by age, gender, ethnicity and smoking status.

Table IV. Associations between SNPs within UCEs and risk for LCRC

SNP	Training set				Replication set				Pooled set	
	Control, N = 341	Case, N = 366	OR (95% CI) ^a	P value	Control, N = 210	Case, N = 185	OR (95% CI) ^a	P value	OR (95% CI) ^a	P value
rs1399685										
TT	317	325	Reference		194	160	Reference		Reference	
TA	23	40	1.72 (1.00–2.99)	0.052	15	24	2.13 (1.05–4.31)	0.036	1.76 (1.15–2.71)	0.010
AA	0	1	N/A		1	1	0.91 (0.05–46.87)	0.949	1.72 (0.15–19.29)	0.660
TA + AA versus TT			1.77 (1.02–3.06)	0.041			2.04 (1.02–4.07)	0.043	1.76 (1.15–2.70)	0.009
rs9784100										
GG	115	162	Reference		64	64	Reference		Reference	
GC	165	149	0.66 (0.47–0.91)	0.013	103	88	0.79 (0.49–1.27)	0.321	0.70 (0.53–0.91)	0.009
CC	61	55	0.67 (0.43–1.05)	0.084	43	33	0.83 (0.45–1.51)	0.534	0.71 (0.50–1.01)	0.058
GC + CC versus GG			0.66 (0.48–0.90)	0.009			0.80 (0.51–1.25)	0.322	0.70 (0.54–0.90)	0.006
GG + GC versus CC			0.85 (0.56–1.27)	0.423			0.96 (0.56–1.62)	0.865	0.87 (0.63–1.20)	0.393
P for the trend			0.78 (0.63–0.97)	0.027			0.89 (0.66–1.20)	0.453	0.81 (0.69–0.97)	0.020
rs7849										
AA	217	255	Reference		130	125	Reference		Reference	
AG	101	102	0.86 (0.61–1.20)	0.378	68	49	0.66 (0.41–1.06)	0.082	0.81 (0.62–1.06)	0.129
GG	23	9	0.30 (0.13–0.69)	0.004	12	11	0.87 (0.35–2.12)	0.753	0.51 (0.28–0.91)	0.022
AG + GG versus AA			0.76 (0.55–1.04)	0.089			0.69 (0.44–1.08)	0.101	0.76 (0.59–0.98)	0.036
AA + AG versus GG			0.32 (0.14–0.72)	0.006			1.00 (0.42–2.41)	0.999	0.54 (0.30–0.96)	0.037
P for the trend			0.71 (0.55–0.94)	0.015			0.79 (0.55–1.13)	0.193	0.76 (0.62–0.94)	0.012
rs7033100										
GG	134	171	Reference		98	88	Reference		Reference	
GA	170	141	0.68 (0.49–0.95)	0.022	97	85	0.89 (0.58–1.38)	0.603	0.76 (0.59–0.98)	0.038
AA	37	54	1.24 (0.76–2.02)	0.398	15	12	0.88 (0.38–2.07)	0.773	1.13 (0.75–1.71)	0.558
GA + AA versus GG			0.78 (0.57–1.06)	0.106			0.89 (0.58–1.35)	0.587	0.82 (0.64–1.05)	0.114
GG + GA versus AA			1.51 (0.96–2.39)	0.076			0.93 (0.41–2.13)	0.872	1.31 (0.88–1.93)	0.183
P for the trend			0.97 (0.77–1.21)	0.768			0.92 (0.65–1.29)	0.612	0.95 (0.79–1.14)	0.578
rs1956211										
AA	204	230	Reference		135	106	Reference			
AG	114	126	1.01 (0.73–1.40)	0.939	63	67	1.15 (0.73–1.81)	0.547	1.09 (0.84–1.41)	0.533
GG	19	9	0.36 (0.15–0.85)	0.020	11	11	1.14 (0.46–2.87)	0.777	0.61 (0.33–1.11)	0.108
AG + GG versus AA			0.92 (0.67–1.25)	0.581			1.15 (0.74–1.77)	0.531	1.02 (0.79–1.31)	0.902
AA + AG versus GG			0.36 (0.15–0.84)	0.018			1.08 (0.44–2.68)	0.861	0.59 (0.32–1.07)	0.084
P for the trend			0.84 (0.64–1.09)	0.189			1.11 (0.78–1.58)	0.564	0.94 (0.77–1.17)	0.597

Bold represents significant ORs and *P* values.

^aAdjusted by age, gender, ethnicity and smoking status.

significance. The homozygous variant genotype of rs7849 conferred a decreased CRC risk with an OR of 0.32 (95% CI, 0.14–0.72; $P = 0.006$) and the per-variant allele OR was 0.71 with a dose–response trend (95% CI, 0.55–0.94; P for trend = 0.015). However, no significant association was observed in the replication set. Again, Both ORs were significant in the pooled analysis. A borderline increased risk was observed for the homozygous variant genotype of rs7033100 (OR, 1.51; 95% CI, 0.96–2.39; $P = 0.076$), whereas the homozygous variant genotype of rs1956211 was associated with a decreased CRC risk (OR, 0.36; 95% CI, 0.15–0.84; $P = 0.018$). The associations were not observed in replication set or in the pooled dataset.

Since no SNP was replicated for RCRC (Table III) and only one SNP was replicated in LCRC (Table IV), assessment of cumulative risk of SNPs was not applicable to RCRC or LCRC. We therefore only assessed joint effects of rs7849 and rs1399685, the two SNPs that were replicated in the overall CRC (Table II). The analysis was done in the replication population only. We defined the genotypes associated with increased risk, which included the wild-type genotype of rs7849 and variant genotypes of rs1399685. There was a significant dose–response trend of increased CRC risk with increasing numbers of unfavorable genotypes (P for trend = 0.003; Table V). Compared with the reference group of subjects with 0 unfavorable genotypes, a 1.52-fold increased CRC risk (95% CI, 0.99–2.34; $P = 0.059$) was noted for subjects with 2 unfavorable genotypes and a 3.97-fold increased risk (95% CI, 1.70–9.28; $P = 0.003$) was noted for subjects with 3 unfavorable genotypes (Table V).

We used CART analysis to explore the higher order gene–gene interactions among significant SNPs to explore whether interactions among the SNPs could further modulate CRC risk. CART analysis identified rs1399685 as the initial split. The final tree structure identified five terminal nodes with different subgroups of CRC risk (Figure 1). Individuals with the rs1399685 TT genotype, rs9784100 GC + CC genotype and rs6124509 AG + GG genotype exhibited the lowest CRC risk. Using this group as the reference, we found that other subgroups (as defined by different SNP combinations) showed an increased CRC risk. Note that subjects carrying the rs1399685 TA + AA genotype and the rs9784100 GC + CC genotype, showed the highest CRC risk (OR, 3.72; 95% CI, 2.08–6.64; $P = 8.93E-06$).

Discussion

In this case–control study, we systematically evaluated the individual and joint effects of genetic polymorphisms within UCEs on CRC risk. We found that rs7849 was significantly associated with CRC risk, and the associations were replicated. Moreover, there was a joint effect of SNPs on CRC risk. Our data also suggested that SNPs that conferred risks for RCRC and LCRC were different, indicating that genetic susceptibility to LCRC and RCRC may be determined by different loci within UCEs.

The main effect of SNP rs7849 on CRC risk was replicated, suggesting that the association is unlikely a chance finding. This SNP was located in the 3' untranslated region of stearyl-CoA desaturase 1 gene (*SCD1*). *SCD1* is a key regulator of fatty acid synthesis in mammalian cells. Increased expression of *SCD1* was found in various cancers including CRC (21), which may improve the biosynthesis of saturated fatty acids and monounsaturated fatty acids to fulfill the increasing demand for phospholipids for cell replication (22). In lung cancer cells, reduction of *SCD1* expression has been found to significantly delay the formation of tumors (23), indicating that *SCD1* may be involved in tumorigenesis. This hypothesis is supported by the finding that *SCD1* overexpression was associated with hepatocarcinogenesis in mice and rats (24). The rare allele of rs7849 was shown to have a strong effect on both insulin sensitivity and waist

Table V. Joint effects of unfavorable genotypes on CRC risk

No. of unfavorable genotype ^b	Case, n (%)	Control, n (%)	OR (95% CI) ^a	P value
0	69 (48.59)	73 (51.41)	1 (reference)	
1	8 (53.33)	7 (46.67)	1.24 (0.41–3.78)	0.702
2	161 (57.09)	121 (42.91)	1.52 (0.99–2.34)	0.059
3	29 (76.32)	9 (23.68)	3.97 (1.70–9.28)	0.001
P for trend				0.003

^aAdjusted by age, gender, ethnicity and smoking status.
^bAdverse genotypes are defined as wild-type genotype of rs7849 and variant genotypes of rs1399685.

Fig. 1. Tree structure generated using CART analysis in CRC patients.

circumference (25), suggesting its potential biological significance. Thus, it is plausible that the rare allele of rs7849 may increase the production of the SCD1 protein and promote CRC carcinogenesis. The functional significance of this SNP on CRC requires further study.

The increasing incidence of RCRC and the decreasing incidence of LCRC (26) suggest that different risk factors may be involved in RCRC and LCRC etiologies. Consistent with this hypothesis, our data showed that different SNPs conferred risks for RCRC and LCRC. These results suggest that site differences should be considered in risk assessment of CRC. The SNP rs1399685 showed consistent association with increased risk for overall CRC and LCRC, which may affect the non-exonic element uc.81. The non-exonic UCE elements have been shown to function as long-range enhancers (27,28). The nearest downstream gene of rs1399685 is activin A receptor, type II (ACVR2). ACVR2 is a member of the transforming growth factor- β type II receptor family and is involved in cellular proliferation, angiogenesis, differentiation, cellular migration and apoptosis (29). Jung *et al.* (30) demonstrated that loss of ACVR2 protein expression due to loss of heterozygosity correlated with the chromosomal instability phenotype. Given that CRC with chromosomal instability is predominant in LCRC (31), it is plausible that rs1399685 may alter the production of ACVR2 and thus contribute to LCRC tumorigenesis.

In addition, three SNPs (rs9784100, rs7849 and rs7033100) also exhibited association with LCRC susceptibility in this study. However, none of the associations were replicated although significant associations were observed in the pooled analysis. It should be noted that the significant results in the pooled analysis did not provide further justification of significance and future validation studies are needed to confirm the findings. Although the results were not validated in our study, there are evident that these SNPs have possible functional significance. The most significant association was found for rs9784100 in non-exonic uc.60, which showed a strong inverse association with LCRC ($P = 0.0008$). The nearest downstream gene of rs9784100 is Fanconi anemia, complementation group L (FANCL), 1 of the 13 known Fanconi anemia genes. Cole *et al.* (32) found recently that FANCL was a catalytic subunit required for monoubiquitination of FANCD2 in DNA repair. The role of FANCL in tumorigenesis was revealed by Zhang *et al.* (33), who reported that a novel splice variant of FANCL could cause chromosomal instability and promote tumor development. Further study demonstrated that FANCD2 was associated with breast cancer risk (34). rs7033100 is located in exonic uc.264 overlapping with the heterogeneous nuclear ribonucleoprotein K, which has known functions in transcription, messenger RNA shuttling and RNA editing and translation, indicating that heterogeneous nuclear ribonucleoprotein K may have a direct role in carcinogenesis (35). Heterogeneous nuclear ribonucleoprotein K was found to be a coactivator for p53 target genes and to play a key role in coordinating transcriptional responses to regulate cell cycle progression (36). rs6124509 in uc.455 (1 of 111 exonic elements) was found to confer a decreased risk of RCRC. This exonic UCE has the putative functions of RNA-splicing and protein-binding genes. Ni *et al.* (13) found that uc.455 overlapped the stop codon exon of the RNA splicing-associated gene RNPC2 (RNA-binding region containing 2), strongly suggesting a regulatory function for uc.455. Furthermore, the presence of ultraconserved stop codon-containing exons promotes transcript sensitivity to nonsense-mediated decay (37). Therefore, rs6124509 may regulate RNPC2 expression posttranscriptionally. Given that overexpression of RNPC2 in carcinogenesis has been established (38), rs6124509 could be a biologically plausible candidate SNP for increased CRC susceptibility. However, since the associations of the three SNPs were not replicated in our analysis, the results should not be overinterpreted and future validation studies are needed to confirm the findings.

We further evaluate the cumulative influences on CRC risk of multiple SNPs within the UCEs. Our data demonstrated that SNPs exhibited strong joint effects on CRC risk. Moreover, CART analysis indicated possible higher interactions among the SNPs that would further modify CRC risk, consistent with the polygenic nature of CRC susceptibility.

Our study has several strengths. To our knowledge, this is the first study to systematically evaluate the risk of CRC using a comprehensive panel of polymorphisms within UCEs. Moreover, previous association studies, including genome-wide association studies, did not consider tumor location (left versus right side) in their design. In this study, we specifically explored whether different polymorphisms within UCEs were associated with risks for RCRC and LCRC. Our results support the hypothesis that different genetic variants within UCEs may affect risk of RCRC and LCRC. However, the SNPs found in RCRC require further validation. Moreover, the study included a validation stage, allowing for replication of the results to further increase detection power and reduce false-positive findings. The limitation of this study is that this is a hospital-based case-control study, and selection bias may exist. However, as the study tested a genetic-driven hypothesis rather than an exposure-driven hypothesis, selection bias is less of a concern. Moreover, other CRC risk factors not included in this study cannot be examined for confounding.

In conclusion, to our knowledge, this is the first study to systematically evaluate CRC risk in a comprehensive panel of polymorphisms within UCEs. Our results strongly suggest that genetic variants in the UCEs may contribute to CRC susceptibility, individually and jointly, and that different genetic etiologies may be involved in RCRC and LCRC. Future studies are needed to validate our findings in independent populations to further confirm these associations, and functional experiments are warranted to elucidate the underlying molecular mechanisms.

Supplementary material

Supplementary Table 1 can be found at <http://carcin.oxfordjournals.org/>.

Funding

This work was supported by a Multidisciplinary Research Program grant on colorectal cancer from The University of Texas M. D. Anderson Cancer Center. This research was supported in part by the National Institutes of Health through M. D. Anderson's Cancer Center support grant (CA016672).

Conflict of Interest Statement: None declared.

References

1. Lichtenstein, P. *et al.* (2000) Environmental and heritable factors in the causation of cancer: analyses of cohorts of twins from Sweden, Denmark, and Finland. *N. Engl. J. Med.*, **343**, 78–84.
2. Jass, J.R. *et al.* (2002) Emerging concepts in colorectal neoplasia. *Gastroenterology*, **123**, 862–876.
3. Gervaz, P. *et al.* (2004) Two colons-two cancers: paradigm shift and clinical implications. *J. Surg. Oncol.*, **88**, 261–266.
4. Bufill, J.A. (1990) Colorectal cancer: evidence for distinct genetic categories based on proximal or distal tumor location. *Ann. Intern. Med.*, **113**, 779–788.
5. Iacopetta, B. (2002) Are there two sides to colorectal cancer? *Int. J. Cancer*, **101**, 403–408.
6. Araki, K. *et al.* (1996) Comparison of mucosal microvasculature between the proximal and distal human colon. *J. Electron Microsc. (Tokyo)*, **45**, 202–206.
7. McMichael, A.J. *et al.* (1985) Host factors in carcinogenesis: certain bile-acid metabolic profiles that selectively increase the risk of proximal colon cancer. *J. Natl Cancer Inst.*, **1985**, 185–191.
8. Grady, W.M. (2004) Genomic instability and colon cancer. *Cancer Metastasis Rev.*, **23**, 11–27.
9. Tsigris, C. *et al.* (2007) Genetic association studies in digestive system malignancies. *Anticancer Res.*, **27**, 3577–3587.
10. Bejerano, G. *et al.* (2004) Ultraconserved elements in the human genome. *Science*, **304**, 1321–1325.
11. Baira, E. *et al.* (2008) Ultraconserved elements: genomics, function and disease. *RNA Biol.*, **5**, 132–134.
12. Pennacchio, L.A. *et al.* (2006) *In vivo* enhancer analysis of human conserved non-coding sequences. *Nature*, **444**, 499–502.

13. Ni, J.Z. *et al.* (2007) Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. *Genes Dev.*, **21**, 708–718.
14. Bernstein, B.E. *et al.* (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, **125**, 315–326.
15. Feng, J. *et al.* (2006) The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. *Genes Dev.*, **20**, 1470–1484.
16. Calin, G.A. *et al.* (2007) Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. *Cancer Cell*, **12**, 215–229.
17. Katzman, S. *et al.* (2007) Human genome ultraconserved elements are ultra-selected. *Science*, **317**, 915.
18. Yang, R. *et al.* (2008) SNPs in ultraconserved elements and familial breast cancer risk. *Carcinogenesis*, **29**, 351–355.
19. Hudmon, K.S. *et al.* (1997) Identifying and recruiting healthy control subjects from a managed care organization: a methodology for molecular epidemiological case-control studies of cancer. *Cancer Epidemiol. Biomarkers Prev.*, **6**, 565–571.
20. Taccioli, C. *et al.* (2009) UCbase & miRfunc: a database of ultraconserved sequences and microRNA function. *Nucleic Acids Res.*, **37**, D41–D48.
21. Li, J. *et al.* (1994) Characterization of a cDNA for human stearyl-CoA desaturase and changes in its mRNA expression in some normal and malignant tissues. *Int. J. Cancer*, **1994**, 348–352.
22. Scaglia, N. *et al.* (2009) Inhibition of stearyl-CoA desaturase-1 inactivates acetyl-CoA carboxylase and impairs proliferation in cancer cells: role of AMPK. *PLoS One*, **4**, e6812.
23. Scaglia, N. *et al.* (2008) Inhibition of stearyl-CoA desaturase 1 expression in human lung adenocarcinoma cells impairs tumorigenesis. *Int. J. Oncol.*, **33**, 839–850.
24. Falvella, F.S. *et al.* (2002) Stearyl-CoA desaturase 1 (Scd1) gene overexpression is associated with genetic predisposition to hepatocarcinogenesis in mice and rats. *Carcinogenesis*, **2002**, 1933–1936.
25. Warensjö, E. *et al.* (2007) Polymorphisms in the SCD1 gene: associations with body fat distribution and insulin sensitivity. *Obesity (Silver Spring)*, **15**, 1732–1740.
26. Cucino, C. *et al.* (2002) Continued rightward shift of colorectal cancer. *Dis. Colon Rectum*, **45**, 1035–1040.
27. Nobrega, M.A. *et al.* (2003) Scanning human gene deserts for long-range enhancers. *Science*, **302**, 413.
28. Licastro, D. *et al.* (2010) Promiscuity of enhancer, coding and non-coding transcription functions in ultraconserved elements. *BMC Genomics*, **11**, 151.
29. Chen, Y.G. *et al.* (2002) Regulation of cell proliferation, apoptosis, and carcinogenesis by activin. *Exp. Biol. Med. (Maywood)*, **227**, 75–87.
30. Jung, B. *et al.* (2009) Activin signaling in microsatellite stable colon cancers is disrupted by a combination of genetic and epigenetic mechanisms. *PLoS One*, **4**, e8308.
31. Chou, C.L. *et al.* (2010) Microsatellite instability screening should be done for right-sided colon cancer patients less than 60 years of age. *Int. J. Colorectal Dis.*, **25**, 47–52.
32. Cole, A.R. *et al.* (2010) The structure of the catalytic subunit FANCL of the Fanconi anemia core complex. *Nat. Struct. Mol. Biol.*, **17**, 294–298.
33. Zhang, J. *et al.* (2010) FAVL elevation in human tumors disrupts Fanconi anemia pathway signaling and promotes genomic instability and tumor growth. *J. Clin. Invest.*, **120**, 1524–1534.
34. Barroso, E. *et al.* (2009) The Fanconi anemia family of genes and its correlation with breast cancer susceptibility and breast cancer features. *Breast Cancer Res. Treat.*, **118**, 655–660.
35. Carpenter, B. *et al.* (2006) Heterogeneous nuclear ribonucleoprotein K is over expressed, aberrantly localised and is associated with poor prognosis in colorectal cancer. *Br. J. Cancer*, **95**, 921–927.
36. Moumen, A. *et al.* (2005) hnRNP K: an HDM2 target and transcriptional coactivator of p53 in response to DNA damage. *Cell*, **123**, 1065–1078.
37. Sathirapongsasuti, J.F. *et al.* (2011) Ultraconserved cDNA segments in the human transcriptome exhibit resistance to folding and implicate function in translation and alternative splicing. *Nucleic Acids Res.*, **39**, 1967–1979.
38. Shi, Y.Y. *et al.* (2005) Identification and analysis of tumour-associated antigens in hepatocellular carcinoma. *Br. J. Cancer*, **92**, 929–934.

Received September 1, 2011; revised January 11, 2012;
accepted February 1, 2012