Research Article

Genome-Wide Association Study Identifies Possible Genetic Risk Factors for Colorectal Adenomas

Todd L. Edwards1,2, Martha J. Shrubsole1,3,6, Qiuyin Cai1,3, Guoliang Li1, Qi Dai1,6, Douglas K. Rex7, Thomas M. Ulbright8, Zhenming Fu1, Ryan H. Delahanty1, Harvey J. Murff1,3,6, Walter Smalley4,5,6, Reid M. Ness4, and Wei Zheng1,3,6

University; 6VA-TVHS GRECC, Nashville, Tennessee; 7Department of Gastroenterology, Biomarkers & Prevention Online (http://cebp.aacrjournals.org/).

Abstract

Background: Colorectal cancer is the second leading cause of cancer-related death, and most colorectal cancer usually arises from colorectal adenomas. Removal of polyps reduces mortality from colorectal cancer. Colorectal adenomas are known to aggregate in families; however, the genetic determinants for risk of polyps are largely unknown.

Methods: In this study, we used data from the Tennessee Colorectal Polyp Study and the Tennessee-Indiana Adenoma Recurrence Study to conduct a GWAS of adenoma cases and controls. Our design consisted of discovery and replication phases for a total of 2,551 Caucasian adenoma cases and 3,285 Caucasian controls. We carried out logistic regression to test for association in both the discovery and replication phase and further examined the results with meta-analysis.

Results: No single nucleotide polymorphism (SNP) achieved a genome-wide significant P value; however, the most significantly associated SNPs were either previously associated with colorectal cancer in GWAS, such as rs10505477 in the gene POU5F1 (odds ratio (OR) = 0.87; 95% confidence interval (CI) 0.81–0.94; \( P = 4.4 \times 10^{-5} \)), or have been biologically linked to benign growths in other tissues, such as rs1913914 in the gene histone deacetylase 9 (OR = 1.1; 95% CI, 1.1–1.47; \( P = 1.1 \times 10^{-5} \)).

Conclusions: This study suggests that several SNPs may be related to adenoma risk and provides clues for future studies.

Impact: These results suggest that some known genetic risk factors of colorectal cancer are necessary but not sufficient for carcinogenesis. Cancer Epidemiol Biomarkers Prev; 22(7); 1219–26. ©2013 AACR.

Introduction

Colorectal cancer is the second most common cause of cancer-related death in North America and the fourth most diagnosed cancer (1). The vast majority of colorectal cancers are derived from colorectal adenomas (2, 3), which are well-recognized colorectal cancer precursors (2). Colorectal cancer risk has been shown to be modulated by environmental and genetic factors, in addition to epigenetic phenomena that associate with tumors. In colorectal cancer pathways, normal colonic epithelium is transformed as the result of the progressive accumulation of genetic and epigenetic alterations such as somatic mutations through gain-of-function, loss-of-function, and subsequent genomic instability (4–6). Although this process is understood to some extent with regard to colorectal cancer endpoints, less is known about the genetic determinants for the intermediate steps of pathogenesis through the adenoma–carcinoma progression before the onset of colorectal cancer.

Many successful investigations have been conducted to discover germ-line genetic risk factors for colorectal cancer (7–12), and these associations generalize for the most part to diverse human populations (12, 13). These studies have provided important insight into the biology and pathogenesis of colorectal cancer; however, these studies have not explored earlier stages of colorectal cancer pathogenesis. Studies of colorectal adenoma have the potential to both characterize the role of known colorectal cancer risk factors at benign stages of aberrant tissue growth, to discover novel factors that are specific to risk of adenoma, and to identify the etiologic factors related to the initiation of colorectal cancer. A small number of studies have investigated adenoma risk as a secondary objective, with a recent notable nonstatistically significant discovery in

Authors’ Affiliations: 1Division of Epidemiology, Department of Medicine; 2Center for Human Genetics Research, Vanderbilt University; 3Vanderbilt–Ingram Cancer Center; 4Division of Gastroenterology, Department of Medicine, Vanderbilt University School of Medicine; 5Section of Gastroenterology, Tennessee Valley Veterans Affairs Medical Center and Vanderbilt University; 6VA-TVHS GRECC, Nashville, Tennessee; 7Department of Medicine, Division of Gastroenterology; and 8Department of Pathology, Clarian Pathology Laboratory, Indiana University School of Medicine, Indianapolis, Indiana

Note: Supplementary data for this article are available at Cancer Epidemiology, Biomarkers & Prevention Online (http://cebp.aacrjournals.org/).

Corresponding Author: Todd L. Edwards, Vanderbilt University Medical Center, 2525 West End Ave., Suite 600, Nashville, TN 37203. Phone: 615-322-3652; Fax: 615-322-8291; E-mail: todd.lewards@vanderbilt.edu
doi: 10.1158/1055-9965.EPI-12-1437
©2013 American Association for Cancer Research.
the TP53 polyadenylation signal (14). Other studies have investigated a small number of candidate single nucleotide polymorphisms (SNP) for adenoma risk, usually on the basis of evidence from colorectal cancer studies (15). Here, we present the first genome-wide association study (GWAS) to investigate the genetic determinants of adenoma in a population of patients with European ancestry from the United States.

In this study, we investigated whether common inter-individual genetic variation is a determinant for adenoma formation in participants from the Tennessee Colorectal Polyp Study (TCPS) and the Tennessee–Indiana Adenoma Recurrence Study (TIARS). These studies were designed and implemented to investigate lifestyle, genetic, and other environmental factors for influence on risk of adenoma.

Materials and Methods

Study population and data collection

TCPS was a colonoscopy-based case-control study conducted in Nashville, Tennessee from 2003 to 2010. Eligible participants, between 45 and 70 years of age, were identified from patients scheduled for colonoscopy at the Vanderbilt Gastroenterology Clinic and the Veteran’s Affairs Tennessee Valley Health System Nashville Campus. Demographic characteristics of all participants are described in Table 1. For the purposes of the association analyses, we only included participants of Caucasian race, although original recruitment for TCPS was from a multi-ethnic population.

Exclusions from the study included participants who had hereditary colorectal cancer syndromes, a prior history of inflammatory bowel disease, previous adenomatous polyps, or any cancer other than nonmelanoma skin cancer. Among eligible participants, 65% provided informed consent, and subsequently 84% completed telephonic interviews and 75% completed a food frequency questionnaire (FFQ) specifically designed for the southern phonic interviews and 75% completed a food frequency questionnaire (FFQ) specifically designed for the southern population. In both study populations, colonoscopic procedures were carried out and reported using standard clinical protocols by the patient’s gastroenterologist. Any identified polyps were removed using biopsy forceps or snare techniques. All pathology diagnoses were determined by patient's gastroenterologist. Any identified polyps were removed using biopsy forceps or snare techniques. All pathology diagnoses were determined by hospital pathologists and reported as part of routine care.

Data were abstracted from these reports to classify study participants into the following groups: adenomas only, hyperplastic polyps, and polyp-free controls. In order to be classified as polyp free, the participant had to have a complete colonoscopy reaching the cecum without the observation of polyps. Participants with at least 2 adenomas were further classified as having multiple adenomas. An advanced adenoma was defined as meeting 1 of the following criteria: (i) size 1 centimeter or more; (ii) tubulovillous or villous; or (iii) high-grade dysplasia.

Genotyping

Initial genotyping was done using the Affymetrix Genome-Wide Human SNP Array 5.0 (Affymetrix, Inc.) to agnostically detect associations with adenoma risk throughout the genome. TCPS patients were selected for genotyping with preference for advanced and multiple adenoma cases, with the remaining cases in the discovery
Quality control

Quality control (QC) procedures were carried out on CEL files using the Dynamic Model (DM) algorithm in the Affymetrix Power Tools software package. Genotypes were called in the remaining samples using the BRLMM-P algorithm. The PLINK–sex-check option did not discover any participants whose X-chromosome heterozygosity was inconsistent with their reported sex. Sixteen participants who were first- or second-degree relatives with other study participants were removed from further analysis. One-hundred and sixty-five participants who were missing greater than 5% of their autosomal genotypes were removed from further analysis (Supplementary Fig. S1). The average concordance of genotypes assessed using the PLINK software package within duplicate QC participants was 99.9% (ref. 18; Table 1).
Supplementary Fig. S2). For SNP QC, SNPs were removed if they were missing in greater than 5% of participants, or if the minor allele frequency (MAF) in the samples that passed sample QC was less than 1%. After related and admixed participants were removed, SNPs were removed for major deviations from Hardy–Weinberg equilibrium (HWE), \( P < 1 \times 10^{-6} \). In addition, we attempted to re-genotype 121 case samples for which the GWAS assay failed. Batch effects were observed in association results with a genomic inflation factor of 1.2 when these cases were included, and more stringent QC did not alleviate the issue; therefore, we removed them from further analysis. After sample and SNP QC procedures, 402,326 SNPs remained in 958 adenoma cases and 909 adenoma-free controls (Supplementary Fig. S3). Population stratification was assessed by comparing the study participants to reference panels from the HapMap Phase 3 participants using EIGENSTRAT (19), resulting in the removal of 22 participants with apparent ancestral differences from the rest of the sample (Supplementary Fig. S4).

In the replication phase of the study, SNPs were required to have an imputation quality information score from SNPTEST of at least 0.8, an allele frequency of at least 10%, and an association with adenoma risk with a \( P \) value less than \( 10^{-4} \). Within SNPs passing these criteria, 27 candidate SNPs were selected on the basis of statistical significance. These SNPs were genotyped using the Sequenom genotyping system in 2,028 adenoma and/or polyp cases and 3,087 controls. Samples were checked for duplications, and 41 pairs were removed. In addition, participants who did not self-report as Caucasian were removed from this analysis (394 cases, 618 controls). These SNPs were evaluated for concordance among replicate QC participants (99%), missing data greater than 5%, HWE \( P < 0.001 \), and minor allele frequency agreement with Phase 1. All SNPs passed QC checks. The final data for association analysis consisted of 1,593 cases with adenoma and/or polyps and 2,376 controls, including 1,059 adenoma-only cases.

### Statistical analysis

Two independent samples of participants from TCPS and TIARS were evaluated for associations between genetic variation and adenoma risk in a 2-stage design. In the discovery phase, genotypes from a GWAS were imputed to the 1000 Genomes and HapMap reference panels. In the replication phase, selected SNPs were genotyped in an independent sample of participants, and results from both phases were combined using meta-analysis. In data from the GWAS, we assessed the relationship between genetic variation in candidate genes and the risk of colorectal adenoma using the software package SNPTESTv2.2.0 with the “–method score” option, using logistic regression with frequentist tests, and assuming an additive effect of SNP alleles on risk, adjusted for age and sex (20). In genes where there were multiple nominally significant SNPs, we conducted conditional tests of association for the remaining SNPs, adjusting for the most significant SNP, age, and sex. This procedure mitigates the effect of LD-induced significance and helps identify associations at SNPs that are potentially due to LD with independent mutations on distinct haplotypic backgrounds. The SNPs genotyped for replications were evaluated for association with adenoma and/or polyp risk using PLINK with logistic regression, adjusting for age and sex. We also conducted a secondary analysis by removing all cases with only hyperplastic polyps and conducting the replication analysis using the 1,059 adenoma-only cases.

A meta-analysis was conducted combining the results from both phases of the investigation using the software METAL for the combined sample size of 2,551 cases and 3,285 controls (21). In addition, we conducted a secondary meta-analysis using the 2,017 adenoma-only cases by excluding all hyperplastic polyp cases. All reported \( P \) values are 2-sided.

### Copy number variation analysis

For associated genes with previous evidence of CNVs in tumor tissue [gene for histone deacetylase 9 (HDAC9) and ERICH1] we checked a previous file for the Birdsuite program to determine if common copy number variation (CNV) were present and could be detected in these regions, to evaluate whether such CNVs might also be common standing germ-line variation in human populations. In addition, we downloaded genotypes for 1,092 samples in 2 Mb regions (1 Mb flanking peak SNPs in the regions) from the 1000 Genomes database, extracted any observed CNVs, labeled structural variants in variant call format files, and used PLINK to evaluate LD between peak SNPs and CNVs. We did this for the 4 major world populations (African, American, Asian, and European).

### Results

#### Demographic data

In the discovery phase, cases were significantly older than controls, less educated, were more likely to have a family history of colorectal cancer, were more likely to regularly use tobacco and alcohol, and more likely to be current users of nonsteroidal anti-inflammatory drugs (NSAID; Table 1). In the replication phase, cases were older, less likely to be female, less educated, were more likely to use tobacco and alcohol, and had significantly higher body mass index (BMI), and were less likely to exercise. Differences in these associations between phases are mostly attributable to sample size and the presence of more veterans in the discovery phase, as the direction of effects for family history, exercise, and BMI are consistent between phases.

#### Genetic effects

Results for the most statistically significant regions, identified by SNP \( P \) values after meta-analysis of both the discovery and replication phases are presented in Table 2. No \( P \) value from these analyses was small.
Table 2. Results from tests of association at selected SNPs in the discovery, replication, and combined evidence from both phases using meta-analysis for SNPs with meta-analysis P values lesser than 0.001

<table>
<thead>
<tr>
<th>Chr</th>
<th>SNP</th>
<th>Position</th>
<th>Gene</th>
<th>RA/EA</th>
<th>OR</th>
<th>SE</th>
<th>Discovery</th>
<th>Replication</th>
<th>Direction</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>T/A</td>
<td>0.850</td>
<td>0.073</td>
<td>0.025</td>
<td>A/T</td>
<td>1.315</td>
<td>0.056</td>
<td>1.05</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>C0</td>
<td>1.05</td>
<td>1.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>rs1919314</td>
<td>18106886</td>
<td>HDAC9</td>
<td>A/T</td>
<td>1.551</td>
<td>0.088</td>
<td>1.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>/C2</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>rs1528827</td>
<td>a 68974838</td>
<td>LRRC49</td>
<td>G/A</td>
<td>1.352</td>
<td>0.059</td>
<td>1.21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>–</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>/C2</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>rs1669625</td>
<td>a 647581</td>
<td>ERICH1</td>
<td>T/C</td>
<td>0.710</td>
<td>0.081</td>
<td>2.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>/C0</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>–</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

GWAS of Adenoma Risk

Discussion

In this investigation, we have conducted a GWAS of candidate SNPs, nominated on the basis of statistical significance, minor allele frequency, and imputation quality. Although none of our results for particular SNPs are statistically significant at the canonical GWAS threshold for significance, the most significant results are related to or are already known risk factors in cancer pathways. Here, we present the first GWAS of colorectal adenoma in the literature, and propose that many of the risk factors detected thus far in colorectal cancer genetics research are also risk factors for benign aberrant tissue growth, which is then at increased risk for colorectal cancer. In addition, in secondary analyses, we observed very similar effects and evidence for association when excluding the hyperplastic polyp-only cases, suggesting that the genetic etiologies may be similar for these classes of aberrant colorectal growths.

The SNP rs1919314 near the gene HDAC9 was associated with adenoma risk. HDAC inhibitors have shown promise in the treatment of carcinomas (22, 23). This gene has been previously noted for gain of copy number in uterine leiomyosarcoma using array-based comparative genomic hybridization (24). Leiomyosarcoma are believed to arise from uterine leiomyoma, or fibroids, just as colorectal cancer is believed to arise from adenoma. Exploring CNVs from the 1000 Genomes data within 1 Mb of this SNP indicated that there are no known CNVs with significant LD with the target SNP, although rare copy number gains are difficult to detect through sequencing. Elevated HDAC enzyme activity is associated with cancer risk, including leukemias and solid tumors (25). High HDAC expression may indicate poor prognosis in chronic
lymphocytic leukemia (CLL) patients (26). CLL is also often diagnosed in asymptomatic patients and follows a prolonged, benign latency course.

SNPs rs1528827 and rs2955036 in the gene leucine-rich repeat containing 49 (LRRC49) was associated with adenoma risk. Silencing of LRRC49 by promoter hypermethylation has been observed to be common in breast cancer (27). Family-based studies have also found genome-wide significant evidence for linkage of this region with glioma risk (28). This gene lies nearby the SNP rs1549318 that has been significantly associated with fasting proinsulin levels near the gene LARP6 in a region of linkage disequilibrium in European HapMap participants that include LRRC49 and LARP6 (data not shown; ref. 29). The proinsulin-raising allele in LARP6 was associated with lower expression of the gene in adipose tissue. Thereby, this association with adenoma risk may indicate a pathogenic mechanism related to energy availability. In vitro and in vitro models of cancer suggest that insulin resistance resulting from obesity is a significant risk factor for tumor growth. Obesity leads to increases in IGF-I, insulin, and leptin (involved in appetite suppression) and decreases in adiponectin (regulates sugar uptake and fat breakdown; refs. 30, 31). This leads to decreased action by insulin on glucose transport and metabolism and, eventually, insulin resistance. The ratio of leptin to adiponectin is linked to tumor growth, with the highest growth coming from high leptin to adiponectin ratios (32). Insulin, leptin, and adiponectin are directly involved in AKT–mTOR signaling pathways. AKT–mTOR is activated in obese animals and increases risk for and progression of cancer (30, 31).

The SNP rs478859 in the region nearby the gene cell division cycle 14, S (CDC14A) was associated with risk of adenoma. This gene is essential for cell-cycle progression, and this gene may play a role in preparation for DNA replication during the subsequent cell cycle. This gene shares 64% sequence identity with the yeast gene, cdc14, and cdc14 is also a homolog of phosphatase and tensin homolog (PTEN; ref. 33). PTEN negatively regulates the MAPK pathway and has been shown to have tumor suppressor activity (34). In addition, CDC14A is an essential component in a positive feedback loop in which the anaphase-promoting complex (APC) initiates progression into anaphase, by CDC14A dephosphorylation of securin, which allows securin ubiquitination and sister chromatid separation (35). APC function has been implicated in studies of RAS oncogene virulence and it was further observed that RAS mutant cancer cells were sensitive to treatment with APC knockdown snRNAs (36). This relationship suggests that CDC14A could be an upstream regulator of the APC mechanism, which is an important factor for the pathogenesis of Ras-positive cancers.

The SNP rs1669625 in the gene glutamate-rich 1 (ERIC1H) was associated with adenoma risk. ERIC1H was observed to have copy number and expression changes in a comparative genomic hybridization study of pancreatic cancer tumors (37). Again, CNV data from the 1000 Genomes project does not indicate there are any common CNVs in LD with this SNP.

The SNP rs10505477 near the gene putative POU domain, class 5, transcription factor 1B (POU5F1B) was also associated with adenoma risk. In GWAS studies, this region at 8q24 has been associated with prostate cancer (38–41), breast cancer (42–44), and this and other SNPs in the region have been associated with colorectal cancer (7–11). Previous reports have also shown nominal associations between rs10505477 and the nearby SNP rs6983267 and adenoma risk (9) and, furthermore, that this SNP disrupts a binding site for transcription factor 7-like 2 (TCF7L2) in an enhancer of the oncogene MYC (45, 46). Subsequent experiments in mice showed that deletion of the MYC enhancer element conferred protection from intestinal tumors in APCmin mouse lines (47), suggesting that this sequence is critical for tumorigenesis. Linkage disequilibrium between rs10505477 and rs6983267 is strong ($r^2 = 0.93$), and rs10505477 was also significantly associated with colorectal cancer (9). TCF7L2 is also a known risk factor for type 2 diabetes (29, 48–51).

Our study had at least 80% statistical power to detect effects in the range between an additive OR of 1.7 for a SNP with a MAFF of 0.1 to an OR of 1.37 for a SNP with a MAF of 0.5 in the discovery phase, assuming an effective threshold for significance of $3 \times 10^{-5}$ (27/1,000,000). For the combined samples in the meta-analysis, we had at least 80% power to detect effects in the range between an OR of 1.47 for a SNP with a MAF of 0.1 to an OR of 1.27 for a SNP with a MAF of 0.5 with a threshold for significance of $5 \times 10^{-5}$. It is, therefore, unlikely that effects that are very much stronger than these exist for adenoma risk variants, as we would have most likely observed them unless they were not adequately covered by the GWAS and imputation. Our strongest OR was 1.3 for the most significant SNP rs1919314, with MAF of 0.14, which is barely within our detectable range. The other notable effects were in the OR range from 1.1 to 1.2, which are difficult to conclusively associate with the sample size we employed. To our knowledge, TCPS/TIARS is the largest colonoscopy-based case-control study conducted to study adenomas.

This investigation suggests that several genes are related to adenoma risk. Notable in these results are the implication of genes that are closely related to diabetes risk. The relationship between diabetes and cancer is well-known and has a strong biological basis and, here, we note genetic evidence that supports this relationship between diabetes and colorectal adenoma. We also investigated the TP53 region for association with adenoma and found little evidence among our genotyped and imputed SNPs (data not shown), although this does not cast much doubt on those results due to the differences in fidelity from genotyping versus sequencing and the rare frequency of the causal SNP in that study (14). Furthermore, these findings show that many colorectal cancer susceptibility loci influence the risk of benign colorectal adenoma, which suggests that they function in early stages of transformation and may not be sufficient for cells to become tumors.
Larger studies of adenoma genetic risk are needed to improve the quality of inferences with regard to the putative risk factors we describe here, and to discover other potentially rare risk factors using next-generation sequencing.

Disclosure of Potential Conflicts of Interest
No potential conflicts of interest were disclosed.

Authors’ Contributions
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): M.J. Shrubsole, Q. Cai, G. Li, D.K. Rex, T.M. Ullbright, H.J. Murff, W.E. Smalley, R.M. Ness, W. Zheng
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): T.L. Edwards, M.J. Shrubsole, Q. Dai, Z. Fu, R. Delahanty, W. Zheng

References
Genome-Wide Association Study Identifies Possible Genetic Risk Factors for Colorectal Adenomas

Todd L. Edwards, Martha J. Shrubsole, Qiuyin Cai, et al.


Updated version Access the most recent version of this article at: doi:10.1158/1055-9965.EPI-12-1437

Supplementary Material Access the most recent supplemental material at: http://cebp.aacrjournals.org/content/suppl/2013/05/15/1055-9965.EPI-12-1437.DC1

Cited articles This article cites 51 articles, 15 of which you can access for free at: http://cebp.aacrjournals.org/content/22/7/1219.full.html#ref-list-1

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, contact the AACR Publications Department at permissions@aacr.org.