MicroRNA Polymorphisms and Risk of Colorectal Cancer

Stephanie L. Schmit1,2, Jeremy Gollub3, Michael H. Shapero3, Shu-Chen Huang2, Hedy S. Rennert4, Andrea Finn3, Gad Rennert4,5,6, and Stephen B. Gruber1,2

Abstract

Background: miRNAs act as post-transcriptional regulators of gene expression. Genetic variation in miRNA-encoding sequences or their corresponding binding sites may affect the fidelity of the miRNA–mRNA interaction and subsequently alter the risk of cancer development.

Methods: This study expanded the search for miRNA-related polymorphisms contributing to the etiology of colorectal cancer across the genome using a novel platform, the Axiom miRNA Target Site Genotyping Array (237,858 markers). After quality control, the study included 596 cases and 429 controls from the Molecular Epidemiology of Colorectal Cancer study, a population-based case–control study of colorectal cancer in northern Israel. The association between each marker and colorectal cancer status was examined assuming a log-additive genetic model using logistic regression adjusted for sex, age, and two principal components.

Results: Twenty-three markers had P values less than 5.0E–04, and the most statistically significant association involved rs2985 (chr6:34845648; intronic of UHRF1BP1; OR = 0.66; P = 3.7E–05). Furthermore, this study replicated a previously published risk locus, rs1051690, in the 3′-untranslated region of the insulin receptor gene INSR (OR = 1.38; P = 0.03), with strong evidence of differences in INSR gene expression by genotype.

Conclusions: This study is the first to examine associations between genetic variation in miRNA target sites and colorectal cancer using a genome-wide approach. Functional studies to identify allele-specific effects on miRNA binding are needed to confirm the regulatory capacity of genetic variation to influence risk of colorectal cancer.

Impact: This study demonstrates the potential for an miRNA-targeted genome-wide association study to identify candidate susceptibility loci and prioritize them for functional characterization. Cancer Epidemiol Biomarkers Prev; 24(1); 65–72. ©2014 AACR.

Introduction

In addition to protein-coding messenger RNAs (mRNAs), other classes of small RNA molecules exist with specialized regulatory and processing functions. Among these types of regulatory RNAs are microRNAs (miRNAs), short (18–24 nucleotide) non–protein-coding molecules that act as post-transcriptional regulators of gene expression (1). The biogenesis of a miRNA begins with transcription from a small, stand-alone gene or an intron or exon of a known protein-coding gene and transitions through a series of conversion steps from hairpin precursors to duplexed pre-miRNAs intermediates to mature, single-stranded miRNAs (2, 3). miRNAs exert their regulatory effects via binding to complementary approximately 6–8 nucleotide target seed sites in the 3′ untranslated regions (3′-UTR) of one or more mRNAs. Depending on the fidelity and context of the interaction, this binding acts to repress translation of the messenger into protein or to signal for degradation of the targeted mRNA (1, 4, 5). Each miRNA typically binds multiple, even thousands, of messenger targets, offering the potential for widespread downstream effects (6, 7).

Deregulated miRNA profiles have been described across a range of cancers including colorectal cancer (8, 9). Furthermore, some suggest that miRNA biology can be integrated into the molecular subtyping of colorectal tumors and into the traditional model of genetic alterations accompanying progression from normal mucosa to carcinoma, particularly among tumors that develop through the chromosomal instability pathway (7, 10–13). As an extension of this work, several miRNAs have been proposed as biomarkers for colorectal cancer early detection, prognosis, and progression (14–16).

Despite extensive miRNA profiling in colorectal tumors, the factors driving aberrations in miRNA expression and their impact on colorectal cancer development and progression are not yet well defined. One hypothesis proposes that SNPs in genes encoding the miRNA sequence or 3′-UTR regions of the corresponding binding sites affect miRNA transcription, miRNA processing, and/or the fidelity of the miRNA–mRNA interaction. Any of these alterations could plausibly impact target miRNA translation into proteins critical for cellular differentiation and proliferation. Evidence from studies of candidate miRNA-related genetic

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

Corresponding Author: Stephen B. Gruber, USC Norris Comprehensive Cancer Center, 1441 Eastlake Avenue, B302L, 3003 Eastlake Road, Los Angeles, CA 90089.

Phone: 323-865-0816; Fax: 323-865-0102; E-mail: sgruber@usc.edu

doi: 10.1158/1055-9965.EPI-14-0219

©2014 American Association for Cancer Research.

www.aacrjournals.org
alterations supports this hypothesis and suggests that such SNPs may alter expression of some miRNAs in colorectal cancer (17) and increase or decrease the risk of tumor development (18). Target site polymorphisms that confer risk in specific populations have been identified in INSR (18), CD86 (18), IL16 (19), RPA2 (20), and GTF2HH1 (20); however, replication of these findings has been limited with the exception of rs1051690 in INSR and rs17281995 in CD86 (21). To date, published studies have not comprehensively investigated polymorphisms implicated in the miRNA regulatory pathway across the genome.

In this study, we expanded the search for miRNA-related genetic variants important in the etiology of colorectal cancer across the genome and investigated the association between thousands of genetic variants in miRNA target sites in 3′-UTR regions and miRNA-encoding genes and colorectal cancer risk using a novel genotyping platform. In contrast to a classical genome-wide association study (GWAS) approach that relies on haplotype-tagging SNPs, we leveraged genotyping of SNPs bioinformatically predicted to have functional implications specific to the miRNA regulatory pathway. We then characterized the predicted miRNA-binding consequences of our most significantly associated SNPs and further explored these associations with expression quantitative trait loci (eQTL) analyses. This study was designed to evaluate the feasibility of a targeted GWAS approach for identifying lead candidates and prioritizing them for functional characterization based on biologically relevant hypotheses. The genetically homogenous founder population of Ashkenazi Jewish individuals experiences a high burden of colorectal cancer and served as the focus of this study (22).

**Materials and Methods**

**Study population: Molecular Epidemiology of Colorectal Cancer study**

The Molecular Epidemiology of Colorectal Cancer (MECC) study is a population-based, case–control study of pathologically confirmed, incident cases of colorectal cancer recruited from a geographically defined region of northern Israel (23). Recruitment began in 1998 and remains ongoing. Individually matched controls with no prior history of colorectal cancer were selected from the same source population that gave rise to cases matched controls with no prior history of colorectal cancer were selected from the same source population that gave rise to cases and matched across the genotyping platforms. IMPUTE2 v2.2.2 was used to impute missing genotypes based on reference haplotypes from the 1000 Genomes Project (March 2012 release to identify markers. In addition, the array included a panel of n = 4,470 ancestry informative markers (AIM) and loci with known complex trait associations from the August 16, 2011 National Human Genome Research Institute (NHGRI) GWAS Catalog (29). In an ancillary study, MECC samples were also genotyped on a custom Affymetrix Axiom platform with approximately 1.3 million SNPs and indels as part of the ColoRectal Transdisciplinary (CORECT) Study, and concordance was compared across the genotyping platforms.

**Genotyping and quality control**

Germline DNA was extracted from peripheral blood samples, purified, quantified by nanodrop and Qubit fluorometric quantitation, and genotyped by Affymetrix on a novel Axiom miRNA Target Site Genotyping array with 237,858 SNPs and indels (Supplementary Table S1). Markers were selected for the microarray from four online bioinformatic databases: PolymiRTS (86,340), dPORE (10,400), Patrocles (1,200), and microRNA.org (158,400). These databases were leveraged to select polymorphic loci that overlap genes encoding miRNAs, miRNA gene regulatory regions, proteins important for miRNA processing, and/or target seed sites (24–28). For microRNA.org, Affymetrix used the database’s high-quality predictions of miRNA-binding sites (both conserved and non-conserved) and intersected microRNA.org’s predicted sites with the 1000 Genomes Phase 1 (March 2012) release to identify markers. In addition, the array included a panel of n = 4,470 ancestry informative markers (AIM) and loci with known complex trait associations from the August 16, 2011 National Human Genome Research Institute (NHGRI) GWAS Catalog (29). In an ancillary study, MECC samples were also genotyped on a custom Affymetrix Axiom platform with approximately 1.3 million SNPs and indels as part of the ColoRectal Transdisciplinary (CORECT) Study, and concordance was compared across the genotyping platforms. IMPUTE2 v2.2.2 was used to impute missing genotypes based on reference haplotypes from phase I of the 1000 Genomes Project (March 2012 release; n = 1,092; refs. 30, 31).

MECC genotype data was filtered on the basis of quality control metrics at the individual subject and SNP levels (Supplementary Fig. S1). Samples with >5% missing genotypes, sex mismatches (between self-reported and genotypic predicted sex), and duplicate samples were identified and subsequently removed. Monomorphic markers and markers with >5% call rate were excluded. Furthermore, SNPs that were not consistent with Hardy–Weinberg Equilibrium in controls were excluded. Principal components analysis (PCA) was conducted using a panel of AIMs and the pcaMethods Biocoreductor package (32) R to identify ethnic outliers for removal and to later adjust for population stratification. Principal component definitions of ancestry were used to exclude ethnic outliers and non-Ashkenazi Jewish individuals from the analysis. Thus, all individuals in the final analysis dataset (described in Table 1) were genetically of Ashkenazi Jewish descent, regardless of their self-reported ethnicity.

**Gene expression quantification**

Gene expression levels from 419,473 probe sets derived from two Affymetrix expression arrays were quantified on RNA isolated from snap frozen tumors of 331 MECC colorectal cancer cases. Of these 331 cases, 135 also had high-throughput germline genotype data available (63 on the Affymetrix Axiom CORECT custom array
and 72 on the Illumina HumanOmni 2.5S-1 BeadChip). Methods for gene expression quantification via hybridization to GeneChip Human Genome U133A 2.0 and Human Genome U133 Plus 2.0 Arrays have been described elsewhere (33). Briefly, expression was measured in two batches (one for each array) followed by quantile normalization and log_{2} transformation of MAS 5.0-calculated signal intensities. Data from the two batches were aligned after individual batch pre-processing and quality control. These microarray data have been deposited in the Gene Expression Omnibus database (accession number GSE26682) to comply with Minimum Information About a Microarray Gene Experiment guidelines.

Statistical analysis

Logistic regression was employed to examine the marginal association between each marker on the miRNA target site array with minor allele frequency (MAF) ≥1% (\textit{f}_{\text{marker}} = 55,208) and colorectal cancer risk assuming a log-9additive genetic model. Here, each additional copy of the minor allele was assumed to confer the same magnitude of risk or protection. Each model was run both unadjusted and adjusted for sex, age, and the first two principal components. We calculated beta coefficients, SEs, ORs with associated 95% confidence intervals (95% CI), and \textit{P} values from unconditional logistic regression. The Bonferroni-corrected alpha level was set at 9.0 \times 10^{-7} (0.05/55,406 SNPs). After taking this genome-wide approach, we then examined previously published SNPs from three studies in the candidate miRNA-related polymorphism literature to assess our ability to replicate purported risk loci (18–20).

To begin the bioinformatic characterization of functional consequences of our most significantly associated SNPs, we investigated predicted changes in miRNA binding using a combination of databases: miRcorna.org, miRBase, PolymiRTS, and dPORE (24, 25, 28, 34–37). MicroSniper was also used to identify the potential disruption or creation of miRNA-binding sites for the following 3′-UTR SNPs in Table 2: rs1380466, rs1972820, and rs2985 (38). A seed site of a minimum of either 7 or 8 bases was specified for each of these SNPs. In addition, we conducted ANOVA to compare differences in gene expression by genotype for all SNPs with association \textit{P} values less than 5 \times 10^{-4} as well as for a previously published risk locus, where expression and genotype data permitted. Expression of the gene nearest to each SNP was considered.

Results

Targeted GWAS

Plots of the first 3 eigenvalues from PCA indicated that the original samples selected for analysis included some non-Ashkenazi Jewish individuals (almost exclusively among subjects without colorectal cancer) that inhibited our ability to control for confounding due to population stratification through principal component adjustment (data not shown). However, following the removal of 5 ethnic outliers and 211 non-Ashkenazi identified on the basis of genetic definitions, the first two principal components were sufficient to control for the remaining population stratification, as indicated by genomic control lambda \textit{GC} \textit{λ} values shown below. The distributions of demographic and clinical characteristics of the final analysis dataset were comparable across case and control groups (Table 1).

Quantile–quantile (Q–Q) and Manhattan plots visually display −\log_{10}(\textit{P} values) resulting from the logistic regression models adjusted for age, sex, and two principal components (Fig. 1). The Q–Q plot in the left panel plots the rank-ordered observed −\log_{10}(\textit{P} value) against the rank-ordered expected −\log_{10}(\textit{P} value). It demonstrates that, on average, we did not observe SNPs with associations more statistically significant than expected under a uniform distribution of \textit{P} values. The GC \textit{λ} value of 1 suggests that principal components 1 and 2 were sufficient to control for population stratification in our ethnically homogenous study sample. The Manhattan plot displays the summary results by ordered chromosomal position and shows that our smallest \textit{P} values are in the 10^{-5} range with none reaching genome-wide statistical significance after correction for multiple testing.

Although none of the individual SNPs achieved genome-wide significance, our top findings are detailed in Table 2. Interestingly, 7 of our 9 most statistically significant SNPs yield a predicted change in miRNA binding in an allele-specific manner. Each of these seven most statistically significant variants predict either a change from no miRNA binding to one or more miRNAs binding or from one set of miRNAs to a different set. None of the most significant miRNA-related SNPs has been previously reported as significantly associated with risk of colorectal cancer.

Gene expression analysis for top association findings

Of the top 25 SNPs that met our \textit{P} value threshold of 5 \times 10^{-4} from the association analysis, 11 corresponding nearest genes had at least 1 matching probe in our gene expression dataset. Among the 21 total probes quantifying gene expression from these 11 genes (some genes had multiple probes), 13 probes for 6 genes had a corresponding genotype measured in MECG cases from the cohort GWAS array. The Bonferroni-corrected alpha level was set at 9.0 \times 10^{-7} (0.05/55,406 SNPs). After taking this genome-wide approach, we then examined previously published SNPs from three studies in the candidate miRNA-related polymorphism literature to assess our ability to replicate purported risk loci (18–20).

To begin the bioinformatic characterization of functional consequences of our most significantly associated SNPs, we investigated predicted changes in miRNA binding using a combination of databases: miRcorna.org, miRBase, PolymiRTS, and dPORE (24, 25, 28, 34–37). MicroSniper was also used to identify the potential disruption or creation of miRNA-binding sites for the following 3′-UTR SNPs in Table 2: rs1380466, rs1972820, and rs2985 (38). A seed site of a minimum of either 7 or 8 bases was specified for each of these SNPs. In addition, we conducted ANOVA to compare differences in gene expression by genotype for all SNPs with association \textit{P} values less than 5 \times 10^{-4} as well as for a previously published risk locus, where expression and genotype data permitted. Expression of the gene nearest to each SNP was considered.

Results

Targeted GWAS

Plots of the first 3 eigenvalues from PCA indicated that the original samples selected for analysis included some non-Ashkenazi Jewish individuals (almost exclusively among subjects without colorectal cancer) that inhibited our ability to control for confounding due to population stratification through principal component adjustment (data not shown). However, following the removal of 5 ethnic outliers and 211 non-Ashkenazi identified on the basis of genetic definitions, the first two principal components were sufficient to control for the remaining population stratification, as indicated by genomic control lambda \textit{GC} \textit{λ} values shown below. The distributions of demographic and clinical characteristics of the final analysis dataset were comparable across case and control groups (Table 1).

Quantile–quantile (Q–Q) and Manhattan plots visually display −\log_{10}(\textit{P} values) resulting from the logistic regression models adjusted for age, sex, and two principal components (Fig. 1). The Q–Q plot in the left panel plots the rank-ordered observed −\log_{10}(\textit{P} value) against the rank-ordered expected −\log_{10}(\textit{P} value). It demonstrates that, on average, we did not observe SNPs with associations more statistically significant than expected under a uniform distribution of \textit{P} values. The GC \textit{λ} value of 1 suggests that principal components 1 and 2 were sufficient to control for population stratification in our ethnically homogenous study sample. The Manhattan plot displays the summary results by ordered chromosomal position and shows that our smallest \textit{P} values are in the 10^{-5} range with none reaching genome-wide statistical significance after correction for multiple testing.

Although none of the individual SNPs achieved genome-wide significance, our top findings are detailed in Table 2. Interestingly, 7 of our 9 most statistically significant SNPs yield a predicted change in miRNA binding in an allele-specific manner. Each of these seven most statistically significant variants predict either a change from no miRNA binding to one or more miRNAs binding or from one set of miRNAs to a different set. None of the most significant miRNA-related SNPs has been previously reported as significantly associated with risk of colorectal cancer.

Gene expression analysis for top association findings

Of the top 25 SNPs that met our \textit{P} value threshold of 5 \times 10^{-4} from the association analysis, 11 corresponding nearest genes had at least 1 matching probe in our gene expression dataset. Among the 21 total probes quantifying gene expression from these 11 genes (some genes had multiple probes), 13 probes for 6 genes had a corresponding genotype measured in MECG cases from the custom Affymetrix CORECT Axiom and/or Illumina Omni platforms. ANOVA results for gene expression [\textit{log}_{2} (normalized intensity)] by genotype for the 6 represented nearest genes with appropriate data availability revealed only one statistically significant SNP, intergenic rs6827968 that falls downstream of the RAP guanine nucleotide exchange factor 2 (\textit{RAPGEF2}) gene (\textit{F} = 5.71, \textit{P} = 0.02). \textit{RAPGEF2} expression levels for two probes plotted against the number of copies of the minor allele at this SNP locus in our study sample can be visualized in Fig. 2, which provides evidence of an eQTL (probe 215992_s_at: \textit{F} = 3.6, \textit{P} = 0.06; probe 203096_s_at: \textit{F} = 5.7, \textit{P} = 0.02).

Replication of previously published risk SNPs

We also examined the colorectal cancer association with 19 candidate miRNA SNPs previously presented in the literature [8 from Landi and colleagues (18), 5 from Azimzadeh and colleagues (19), and 6 from Naccarati and colleagues (20)], of which 6 were statistically significant in the original report. In our dataset, we replicated only one of the previously reported findings (Table 3). The replicated variant (rs1051690), originally reported in Landi and colleagues (18), is located in the 3′-UTR region of the insulin receptor gene \textit{INSR} (Table 3; OR = 1.38; \textit{P} = 0.03) and has predicted miRNA-binding consequences. Our eQTL analysis demonstrated a statistically significant association between the rs1051690 variant and expression of the \textit{INSR} gene, with increasing expression tracking with each additional copy of the minor allele (Supplementary Fig. S2: \textit{F} = 21.3; \textit{P} = 8.98 \times 10^{-4}).

Genotype concordance: miRNA targeted array versus custom GWAS array

Only 14,436 markers were directly measured on both the Affymetrix miRNA Target Site Genotyping Array and the CORECT
Axiom 1.3M custom array in the same set of samples. For the 14,436 overlapping markers, there was a 99.89% overall genotype concordance across arrays. However, because the number of directly measured markers shared by both arrays was low, we then compared genotypes for directly measured markers on the miRNA targeted array with 1000 Genomes imputed genotypes.

Table 2. Summary of variants from the Axiom miRNA Target Site Genotyping Array with the most statistically significant associations from age-, sex-, and principal component 1–2–adjusted logistic regression models.

<table>
<thead>
<tr>
<th>rsID</th>
<th>CHR</th>
<th>BP</th>
<th>A1*</th>
<th>MAF</th>
<th>OR</th>
<th>SE</th>
<th>P</th>
<th>Gene</th>
<th>A1 miRNA</th>
<th>A2 miRNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs1288559</td>
<td>10</td>
<td>32856746</td>
<td>C</td>
<td>0.07</td>
<td>0.54</td>
<td>0.18</td>
<td>4.3E–04</td>
<td>CCDC7 (coding)</td>
<td>miR2837</td>
<td>miR4270</td>
</tr>
<tr>
<td>rs853158</td>
<td>19</td>
<td>32902386</td>
<td>G</td>
<td>0.03</td>
<td>0.38</td>
<td>0.28</td>
<td>4.5E–04</td>
<td>H56577 (3’-UTR)</td>
<td>miR3162-5p</td>
<td>miR5441</td>
</tr>
<tr>
<td>rs6379924</td>
<td>12</td>
<td>32745248</td>
<td>C</td>
<td>0.07</td>
<td>0.29</td>
<td>0.37</td>
<td>4.9E–04</td>
<td>FLJ50403 (exon, 3’-UTR)</td>
<td>miR4758-5p</td>
<td>miR3659</td>
</tr>
<tr>
<td>rs31072275</td>
<td>20</td>
<td>32974395</td>
<td>A</td>
<td>0.15</td>
<td>1.59</td>
<td>0.13</td>
<td>4.9E–04</td>
<td>FLJ14914 (exon)</td>
<td>miR1284</td>
<td>miR337-3p</td>
</tr>
<tr>
<td>rs117299563</td>
<td>19</td>
<td>52095600</td>
<td>C</td>
<td>0.12</td>
<td>0.28</td>
<td>0.37</td>
<td>4.9E–04</td>
<td>FLJ50403 (exon, 3’-UTR)</td>
<td>miR4758-5p</td>
<td>miR3659</td>
</tr>
<tr>
<td>rs117299563</td>
<td>19</td>
<td>52095600</td>
<td>C</td>
<td>0.12</td>
<td>0.28</td>
<td>0.37</td>
<td>4.9E–04</td>
<td>FLJ50403 (exon, 3’-UTR)</td>
<td>miR4758-5p</td>
<td>miR3659</td>
</tr>
<tr>
<td>rs79029362</td>
<td>1</td>
<td>178516408</td>
<td>G</td>
<td>0.03</td>
<td>2.91</td>
<td>0.30</td>
<td>4.6E–04</td>
<td>Chor220 (exon)</td>
<td>miR3162-5p</td>
<td>miR5441</td>
</tr>
<tr>
<td>rs47665992</td>
<td>12</td>
<td>31373531</td>
<td>A</td>
<td>0.26</td>
<td>1.45</td>
<td>0.11</td>
<td>4.7E–04</td>
<td>RHISA (upstream)</td>
<td>miR4758-5p</td>
<td>miR3659</td>
</tr>
<tr>
<td>rs56391924</td>
<td>10</td>
<td>32745248</td>
<td>C</td>
<td>0.07</td>
<td>0.54</td>
<td>0.18</td>
<td>4.7E–04</td>
<td>CCDC7 (missense)</td>
<td>miR4758-5p</td>
<td>miR3659</td>
</tr>
<tr>
<td>rs31072275</td>
<td>20</td>
<td>32974395</td>
<td>A</td>
<td>0.15</td>
<td>1.59</td>
<td>0.13</td>
<td>4.9E–04</td>
<td>FLJ14914 (exon)</td>
<td>miR4758-5p</td>
<td>miR3659</td>
</tr>
<tr>
<td>rs1972820</td>
<td>2</td>
<td>22243422</td>
<td>G</td>
<td>0.36</td>
<td>1.14</td>
<td>0.10</td>
<td>4.9E–04</td>
<td>ERBB4 (3’-UTR)</td>
<td>miR4758-5p</td>
<td>miR3659</td>
</tr>
<tr>
<td>rs12247495</td>
<td>10</td>
<td>32802829</td>
<td>C</td>
<td>0.07</td>
<td>0.54</td>
<td>0.18</td>
<td>4.9E–04</td>
<td>CCDC7 (intron)</td>
<td>miR4758-5p</td>
<td>miR3659</td>
</tr>
</tbody>
</table>

NOTE: A1 miRNA and A2 miRNA refer to miRNAs predicted to bind in the presence of each alternative allele. Gray indicates the availability of high-throughput genotype data and gene expression data for the nearest gene. Abbreviation: NA, no prediction found. A*, effect allele for the corresponding odds ratios. SE, standard error of the beta estimate.
from the Axiom 1.3M custom array. Of the 88,205 directly genotyped, post–quality control markers on the miRNA-targeted array, 63,407 were imputed with high quality from the Axiom 1.3M custom array. A comparison of the 63,407 miRNA-targeted array genotypes and the corresponding best call genotypes derived from the Axiom 1.3M custom array imputation showed that heterozygote genotype concordance was severely depressed for SNPs with MAF ≤ 5%. That is, consistent with prior GWAS studies using imputed genotypes, imputation did not perform well for the less common alleles important in the miRNA pathway and did not accurately reflect the directly measured genotypes.

Discussion

To our knowledge, this is the first study to examine associations between genetic variation in miRNA-encoding genes or target seed sites and risk of colorectal cancer using a genome-wide approach informed by bioinformatic miRNA prediction algorithms. While we did not identify any genome-wide significant associations meeting the traditional threshold of $5 \times 10^{-8}$, this study did highlight suggestive variants with predicted miRNA-binding implications. These findings led us to replicate a previously reported association between rs1051690 in INSR and colorectal cancer risk and to demonstrate variability in INSR gene expression by genotype at this locus. While limited with respect to power, our initial study of only 596 cases and 429 controls demonstrated the potential for a targeted miRNA GWAS approach to identify candidate susceptibility loci and to prioritize them based on biologic insights for further functional characterization.

Alterations of expression from miRNA targets may be mediated by seed site polymorphisms that strengthen or weaken the miRNA–mRNA interaction. We illustrated a relevant example in this study from an association finding through eQTL analysis, and more generally, demonstrated our novel approach (Fig. 3). The INSR 3'-UTR variant (rs1051690) association with colorectal cancer had previously been detected in both Czech Republic and Spanish case–control studies assuming a codominant model (18, 21). We were able to replicate this risk locus based on a log-additive genetic model assumption. To date, few studies have examined the functional consequences miRNA-related SNPs. However, INSR is a notable exception. The same group that originally identified the INSR association later conducted in vitro luciferase reporter assays to show that the minor allele differentially regulates reporter gene expression (21). Evidence from our eQTL analysis corroborates this finding and provides an example of how such a target site polymorphism could influence that same gene's expression in a

![Figure 1](image1.png)

**Figure 1.** Q-Q and Manhattan plots of $P$ values for association between each genetic marker and colorectal cancer from logistic regression models adjusted for sex, age, and two principal components for ancestry. MAF ≥ 1%; $n_{	ext{marker}} = 55,406, \text{GC} \chi = 1.$

![Figure 2](image2.png)

**Figure 2.** $R\text{APGEF2}$ gene expression [log$_2$(normalized intensity)] measured by two separate probes in 72 MECC colorectal cancer cases by rs6827968 genotype.
Table 3. MECC association results from logistic regression adjusted for age, sex, and principal components 1-2 for Axiom miRNA Target Site Genotyping Array markers previously studied as miRNA-related variants.

<table>
<thead>
<tr>
<th>SNP</th>
<th>CHR</th>
<th>Position</th>
<th>Gene</th>
<th>A1</th>
<th>Published OR</th>
<th>Published P</th>
<th>MECC OR</th>
<th>MECC SE*</th>
<th>MECC P</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs1051690*</td>
<td>19</td>
<td>7116963</td>
<td>INSR</td>
<td>T</td>
<td>1.86</td>
<td>0.05</td>
<td>1.38</td>
<td>0.15</td>
<td>0.03</td>
</tr>
<tr>
<td>rs1568439*</td>
<td>5</td>
<td>158742014</td>
<td>IL2B</td>
<td>G</td>
<td>0.65</td>
<td>0.80</td>
<td>1.14</td>
<td>0.12</td>
<td>0.06</td>
</tr>
<tr>
<td>rs11515*</td>
<td>9</td>
<td>21968199</td>
<td>CDKN2A</td>
<td>G</td>
<td>0.67</td>
<td>1.04</td>
<td>1.05</td>
<td>0.10</td>
<td>0.32</td>
</tr>
<tr>
<td>rs17256547*</td>
<td>3</td>
<td>1486757</td>
<td>XPC</td>
<td>C</td>
<td>1.13</td>
<td>0.73</td>
<td>1.14</td>
<td>0.13</td>
<td>0.27</td>
</tr>
<tr>
<td>rs3155500*</td>
<td>16</td>
<td>50766886</td>
<td>NOD2</td>
<td>A</td>
<td>0.72</td>
<td>1.07</td>
<td>0.99</td>
<td>0.09</td>
<td>0.37</td>
</tr>
<tr>
<td>rs1131445*</td>
<td>15</td>
<td>81601782</td>
<td>IL16</td>
<td>C</td>
<td>0.99</td>
<td>0.99</td>
<td>0.93</td>
<td>0.09</td>
<td>0.41</td>
</tr>
<tr>
<td>rs1131445*</td>
<td>15</td>
<td>81601782</td>
<td>IL16</td>
<td>C</td>
<td>2.21</td>
<td>0.004</td>
<td>0.93</td>
<td>0.09</td>
<td>0.41</td>
</tr>
<tr>
<td>rs1051690*</td>
<td>3</td>
<td>12625747</td>
<td>RAF1</td>
<td>T</td>
<td>1.11</td>
<td>0.85</td>
<td>0.90</td>
<td>0.14</td>
<td>0.44</td>
</tr>
<tr>
<td>rs4596*</td>
<td>11</td>
<td>18388128</td>
<td>GTF2H1</td>
<td>C</td>
<td>0.79</td>
<td>0.03</td>
<td>0.93</td>
<td>0.10</td>
<td>0.44</td>
</tr>
<tr>
<td>rs2229090*</td>
<td>3</td>
<td>1487345</td>
<td>XPC</td>
<td>C</td>
<td>0.91</td>
<td>0.38</td>
<td>1.08</td>
<td>0.12</td>
<td>0.54</td>
</tr>
<tr>
<td>rs17281995*</td>
<td>3</td>
<td>12183964</td>
<td>CD6B</td>
<td>C</td>
<td>2.93</td>
<td>0.01</td>
<td>1.05</td>
<td>0.12</td>
<td>0.66</td>
</tr>
<tr>
<td>rs11677*</td>
<td>1</td>
<td>20301964</td>
<td>PLAG2</td>
<td>T</td>
<td>1.02</td>
<td>0.97</td>
<td>1.06</td>
<td>0.14</td>
<td>0.70</td>
</tr>
<tr>
<td>rs7356*</td>
<td>1</td>
<td>2282100</td>
<td>RPA2</td>
<td>C</td>
<td>1.33</td>
<td>0.04</td>
<td>1.02</td>
<td>0.09</td>
<td>0.81</td>
</tr>
<tr>
<td>rs1803541*</td>
<td>2</td>
<td>12801493</td>
<td>ERCC5</td>
<td>T</td>
<td>0.96</td>
<td>0.70</td>
<td>0.97</td>
<td>0.14</td>
<td>0.86</td>
</tr>
<tr>
<td>rs2760224*</td>
<td>5</td>
<td>40692940</td>
<td>PTGER4</td>
<td>A</td>
<td>0.21</td>
<td>0.14</td>
<td>0.99</td>
<td>0.12</td>
<td>0.94</td>
</tr>
<tr>
<td>rs1670224*</td>
<td>5</td>
<td>40692940</td>
<td>PTGER4</td>
<td>A</td>
<td>0.29</td>
<td>0.11</td>
<td>0.99</td>
<td>0.12</td>
<td>0.94</td>
</tr>
<tr>
<td>rs4781563*</td>
<td>16</td>
<td>14045599</td>
<td>ERCC4</td>
<td>A</td>
<td>0.68</td>
<td>0.09</td>
<td>1.00</td>
<td>0.10</td>
<td>0.97</td>
</tr>
<tr>
<td>rs960555*</td>
<td>17</td>
<td>4534834</td>
<td>ALOX5</td>
<td>C</td>
<td>0.98</td>
<td>0.01</td>
<td>1.00</td>
<td>0.10</td>
<td>0.98</td>
</tr>
<tr>
<td>rs743554*</td>
<td>17</td>
<td>73754248</td>
<td>ITGB4</td>
<td>A</td>
<td>0.76</td>
<td>0.36</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

NOTE: Data row 1 shows replication of a variant in INSR that was suggestively associated with colorectal cancer risk in prior publications by Landi et al. in 2008 and 2011 (OR (95% CI), 1.86 (0.99–3.50); P = 0.052).

*SE, standard error of the beta estimate.

This study has limitations with respect to power and modeling assumptions. Our sample size is limited to 1,025 samples. However, this analysis, which was able to replicate a previously identified miRNA risk locus and characterize preliminary functionality, provides justification for study in a larger sample. Our lack of genome-wide significant findings is likely attributable to a lack of power, and our sample size did not permit the investigation of effects for rare variants with MAF < 1%. Also, not all SNPs exert their effects according to the assumed log-additive genetic model, and this choice made to restrict multiple testing could inhibit our ability to identify risk loci that are consistent with a recessive, dominant, or codominant model. Furthermore, we did not consider interactions between these potentially risk-confering variants or variant effects in the context of environmental risk factors. Finally, our ability to examine gene expression was limited by data availability and restriction to studying the SNP's nearest gene.

dose-response manner (Supplementary Fig. S2). A link between insulin resistance and colorectal cancer has long been recognized (39). It is possible that each additional copy of the minor/risk allele reduces miRNA-mRNA binding to the point of inhibiting miRNA degradation, which is what may lead to the increased INSR gene expression observation. It is also possible that the SNP exerts an effect analogous to haploinsufficiency, such that one copy of the major allele is not sufficient to appropriately repress INSR protein expression. Further functional work is necessary to elucidate this particular SNP’s mechanism of action.

Another illustrative example for the success of our approach lies with rs6827968, our third most statistically significant finding from the targeted GWAS that also showed evidence as a cis-eQTL. Although rs6827968 is highly unlikely to exert a direct regulatory influence via the miRNA pathway on the nearest gene as it is an AIM, RAPGEF2 encodes a protein that could plausibly be linked to colorectal cancer etiology. RAPGEF2 activates RAS through promotion of the active GTP-bound state in a GTP/GDP-regulated signal transduction switch (40).

Sethupathy and Collins suggested in 2008 that studies elucidating the role of miRNA-related polymorphisms in complex diseases such as colorectal cancer should focus on three domains: genetic, functional (testing altered miRNA targeting mediated by genetic variation), and mechanistic (testing the mechanism by which altered miRNA leads to tumor development; 41). We provided evidence with respect to genetic and functional studies. The next step is to expand our genotyped dataset to increase power for detecting novel risk loci. With respect to genotyping platform for future studies, this study highlighted the advantage of this novel genotyping array over a traditional GWAS array with imputation based on the 1000 Genomes Project haplotypes, particularly when rare variants are of interest. A comparison of concordance between the Affymetrix miRNA Target Site Genotyping Array genotypes and imputed best call genotypes from the CORECT Axiom 1.3M custom array showed that the targeted miRNA array has added value over the GWAS array for the thousands of markers from this regulatory pathway with MAF ≤ 5%. Also, functional studies are underway to identify SNP effects on miRNA-binding fidelity (for rs1051690 as well as other top association findings) and to find the best in vitro model for allele-specific effects. Furthermore, replication and fine-mapping will strengthen our confidence in both novel and previously published findings. Finally, this study suggests the benefit of reexamining previously published colorectal cancer susceptibility regions identified through GWAS for potential functional SNPs in miRNA-binding sites or other miRNA pathway-related sequences. Given that most GWAS risk loci identified to date have MAF ≥ 5%, reanalysis of existing, imputed GWAS datasets using the bioinformatic approaches described here has a high probability of yielding insights into the functional relevance of these regions.

Published OnlineFirst October 23, 2014; DOI: 10.1158/1055-9965.EPI-14-0219

Downloaded from cebp.aacrjournals.org on August 18, 2017. © 2015 American Association for Cancer Research.
Identify SNPs from targeted association study

Evaluate statistically significant SNPs for predicted impact using online miRNA database of origin

Investigate whether alternate allele leads to a change in miRNA binding site

eQTL analysis

Figure 3. Bioinformatic pipeline for the predicted impact of statistically significant SNPs on miRNA target seed sites and cis gene expression. SNPs identified from the targeted association study were evaluated using PolymiRTS, Patrocles, and microRNA.org (in conjunction with markers identified by the 1000 Genomes Project) for indications of target seed sites that would be disrupted by the presence of the alternate allele for each genetic variant. Markers from the PolymiRTS database were further annotated for potential creation of new seed sites. Statistically significant markers were further investigated with MicroSNPER to identify additional target seed sites both disrupted and potentially created by the presence of the alternate allele. The final step was cis-eQTL analysis where gene expression data were available.

Despite these limitations, we provide evidence that a targeted genome-wide approach for studying germline susceptibility can be extended beyond known or purported cancer biology pathways to the exploration of a regulatory pathway with widespread post-transcriptional effects. A better understanding of the mechanisms by which aberrations in miRNA expression and binding impact colorectal cancer development and progression may offer insights for prevention and targeted therapeutics. Specifically, the JNSR variant warrants further investigation in a functional setting to elucidate its role in the alteration of colorectal cancer risk.

Disclosure of Potential Conflicts of Interest
A. Finn is a genotyping specialist at Affymetrix Inc. No potential conflicts of interest were disclosed by the other authors.

Disclaimer
The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Authors’ Contributions
Conception and design: S.I. Schmit, A. Finn, S.B. Gruber
Development of methodology: S.I. Schmit, J. Gollub, H.S. Rennert, S.B. Gruber
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): J. Gollub, H.S. Rennert, G. Rennert, S.B. Gruber
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): S.I. Schmit, H.S. Rennert, G. Rennert, S.B. Gruber
Writing, review, and/or revision of the manuscript: S.I. Schmit, M.H. Shapero, H.S. Rennert, A. Finn
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): J. Gollub, S.-C. Huang, S.B. Gruber
Study supervision: G. Rennert, S.B. Gruber

Acknowledgments
The authors acknowledge the ColoRectal Transdisciplinary (CORECT) Study and its investigators for contributing genotypes for a comparative analysis of the miRNA-targeted array versus a custom GWAS array.

Grant Support
This work was supported by the National Cancer Institute at the NIH (U19 CA148107 to S.B. Gruber; P30 CA041089 to S.B. Gruber; and RO1 CA14848 to S.B. Gruber), the National Human Genome Research Institute at the NIH (T32 HG000040 to S.L. Schmit), the National Institute of Environmental Health Sciences at the NIH (T32 ES013678 to S.I. Schmit), and the Rackham Graduate School at the University of Michigan (Rackham Predoctoral Fellowship to S.L. Schmit).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received March 17, 2014; revised September 23, 2014; accepted October 6, 2014; published OnlineFirst October 23, 2014.

miRNA-Related Genetic Variants and Risk of Colorectal Cancer

References


