Cell cycle–related genes as modifiers of age of onset of colorectal cancer in Lynch syndrome: a large-scale study in non-Hispanic white patients

Jinyun Chen1,2,6,†, Mala Pande1,2,†, Yu-Jing Huang1, Jonghwan Wei1, Christopher Lamos2, Bente A. Talseth-Palmer1,2, Cliff J. Meldrum3, Wei V. Chen2, Ivan P. Gorlov4, Patrick M. Lynch1, Rodney J. Scott1,2 and Marsha L. Frazier1

1Department of Epidemiology and 2Department of Genetics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA, 3School of Biomedical Sciences and Pharmacy, University of Newcastle, 4Hunter Medical Research Institute and 5Hunter Area Pathology Service, Newcastle, NSW, Australia, 6Department of Genitourinary Medical Oncology and 7Department of Gastroenterology, Hepatology and Nutrition, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

†These authors contributed equally to this work.

Heterogeneity in age of onset of colorectal cancer in individuals with mutations in DNA mismatch repair genes (Lynch syndrome) suggests the influence of other lifestyle and genetic modifiers. We hypothesized that genes regulating the cell cycle influence the observed heterogeneity as cell cycle–related genes respond to DNA damage by arresting the cell cycle to provide time for repair and induce transcription of genes that facilitate repair. We examined the association of 1456 single nucleotide polymorphisms (SNPs) in 128 cell cycle–related genes and 31 DNA repair–related genes in 485 non-Hispanic white participants with Lynch syndrome to determine whether there are SNPs associated with age of onset of colorectal cancer. Genotyping was performed on an Illumina GoldenGate platform, and data were analyzed using Kaplan–Meier survival analysis, Cox regression analysis and classification and regression tree (CART) methods. Ten SNPs were independently significant in a multivariable Cox proportional hazards regression model after correcting for multiple comparisons (P < 5 × 10−5). Furthermore, risk modeling using CART analysis defined combinations of genotypes for these SNPs with which subjects could be classified into low-risk, moderate-risk and high-risk groups that had median ages of colorectal cancer onset of 63, 50 and 42 years, respectively. The age-associated risk of colorectal cancer in the high-risk group was more than four times the risk in the low-risk group (hazard ratio = 4.67, 95% CI = 3.16–6.92). The additional genetic markers identified may help in refining risk groups for more tailored screening and follow-up of non-Hispanic white patients with Lynch syndrome.

Introduction

Lynch syndrome (also called hereditary non-polyposis colorectal cancer) is an autosomal dominant inherited cancer predisposition disorder. It is caused by defects in DNA mismatch repair (MMR) due to mutations in DNA MMR genes—MLH1, MSH2, MSH6 and PMS2 (1–4) and more recently also due to mutations in FACSTD1 (or EPCAM) (5,6). The cancers most commonly seen in affected individuals are early onset colorectal cancer (CRC) and endometrial cancer, but cancers of other sites are also observed, such as cancers of the stomach, biliary tract, pancreas, kidneys, brain and skin (7,8). Lynch syndrome tumors are characterized by microsatellite instability resulting from deficient DNA MMR and also demonstrate loss of staining for one or more of the MMR proteins (9–11). These tumor characteristics are key in suspecting Lynch syndrome in an individual and form the basis for genetic testing for Lynch syndrome.

Although mutations in DNA MMR genes are the underlying cause of Lynch syndrome, there is heterogeneity in expression of the cancer phenotype, suggesting that other genetic and lifestyle factors may influence cancer risk. The heterogeneity is particularly evident in the variability in age of onset of CRC seen in these patients. Carriers of MSH6 mutations have a later age of CRC onset than MLH1 and MSH2 mutation carriers, and CRC is less frequent in MSH6 mutation carriers (12,13) but mutations in the different MMR genes only account for some of the variability observed in age of onset of CRC.

Cell cycle checkpoints respond to DNA damage by arresting the cell cycle to provide time for repair and by inducing transcription of genes that facilitate repair (14). Checkpoint loss and perturbation of cell cycle control results in genomic instability and is a hallmark of cancer. More subtle genetic changes due to functional polymorphisms in cell cycle–related genes can act as genetic risk modifiers for the development of cancer. Our previous studies indicate that polymorphisms in the cell cycle–related genes cyclin D1, p53 and AURKA are associated with earlier age of onset of CRC in MMR gene mutation carriers (15–17). Other cell cycle–related genes have also been implicated in modifying cancer risk, including p16 (18), p15 and Rb1 (19), p21 (19), p27 (20) and CHEK2 (21). We hypothesized that in addition to genes regulating MMR, genes regulating the cell cycle influence the heterogeneity in CRC age of onset in patients with Lynch syndrome. To test our hypothesis, we examined the association of 1456 single nucleotide polymorphisms (SNPs) in 128 cell cycle–related genes and 31 DNA repair–related genes in 485 non-Hispanic white subjects with Lynch syndrome to determine whether one or more of the SNPs modified the age-associated risk of CRC. The overarching goal of our study was to provide a better understanding of the role of multiple genetic variants in cell cycle–related genes as risk factors responsible for variation in onset age of Lynch syndrome.

To capture the combined effect of multiple SNPs in the cell cycle pathway, we used a pathways-based genotyping approach, which may amplify the effects of individual polymorphisms that interact in the same pathway and enhance the predictive power. In addition, we utilized a tree-based statistical approach to identify genetic risk factors influencing age-associated risk for Lynch syndrome. We selected a tree-based analysis because it is often able to uncover complex interactions between predictors that may be difficult or impossible to uncover using traditional multivariate techniques. Furthermore, tree-based modeling is adept in uncovering predictors that may be largely operative within specific patient subgroups, but may have minimal effect or none in other patient subgroups.

Materials and methods

Study population

Patients and family members with a confirmed MMR mutation in MLH1, MSH2 or MSH6 were included in the study. To avoid heterogeneity attributable to racial differences in allele frequencies, the analysis was limited to self-reported non-Hispanic white subjects. There were 566 study participants from The University of Texas MD Anderson Cancer Center, USA, and 216 from the Hunter Medical Research Institute, Australia. All participants provided written informed consent for use of their DNA for this research, and the study was approved by the Institutional Review Board of MD Anderson Cancer Center and the Institutional Ethics Review Board of the Hunter New England Health Service.

Gene and SNP selection

To select the cell cycle–related genes included in this study, we used the KnowledgeNet algorithm (22), which is an effective tool to identify genes...
associated with specific function. It combines literature mining with data on functional classification of genes by the Gene Ontology database. First, a list of key words specifying the specific gene function needs to be identified. We used cell cycle, cell cycle progression, cell cycle arrest, cell cycle progression, cell cycle arrest, cycle regulation, cell cycle, cell cycle regulation, cell cycle, cell cycle checkpoint, cycle checkpoint, cell cycle checkpoint control, checkpoint and checkpoint control as key words to identify cell cycle-related genes. KnowledgeNet provides a ranking of genes with confidence scores, which we used to prioritize. We next used Ingenuity Pathways Analysis software (Ingenuity Systems, www.ingenuity.com) to search the cell cycle pathway in the Ingenuity Pathways Analysis library of canonical pathways. Genes that appeared on the list identified by searching KnowledgeNet but did not appear within the Ingenuity Pathways Analysis cell cycle pathway were excluded from the list. From the remaining genes, we chose to study the top 124 genes ranked by KnowledgeNet using the cutoff confidence score >0.04.

To further refine the list, we used the KnowledgeNet algorithm to select genes that are considered to be important for colorectal cancer. We searched using the terms cancer, colorectal cancer, cancer related, cancer genes, sporadic, colorectal cancer, primary cancer, colorectal carcinoma, cancers, cancer risk, familial, colorectal cancer risk, sporadic cancer, familial colorectal, sporadic colorectal cancer, colorectal neoplasms, cancer susceptibility, cancers, sporadic colorectal, colorectal cancer and colorectal carcinoma. Four cell cycle-related genes including TGFBR2, Mlh1, Msh2 and Msh6 that were ranked highly for CRC but were not on our top-124 list (described in the preceding paragraph) were added to our list.

We used SNPbrowser version 4.0 (25) to select tagging SNPs. This software was designed for selection of SNPs based on observed linkage disequilibrium (LD), through construction of metric LD maps and selection of haplotype tagging SNPs. The application provides easy and intuitive selection of SNPs, including visualization of SNPs by showing gene structure, linkage disequilibrium map and haplotype block information. The tagging SNP wizard easily enables the selection of maximally informative tagging SNPs based on user-selected parameters. SNP selection is based on the ethnic-specific LD pattern identified by the HapMap Project (http://hapmap.ncbi.nlm.nih.gov/). The tagging SNPs were chosen with an r² of 0.80 or more and a minor allele frequency (MAF) of 0.05 or more in Caucasian population. SNPs from the adjacent 10 kb regions on either side of the gene were also included. All validated non-synonymous SNPs were included regardless of MAF. We also include some functional SNPs from 31 DNA repair-related genes that have been reported to influence risk for cancer.

Genotyping and data cleaning
A GoldenGate assay (Illumina, San Diego, CA) was developed to examine 1536 SNPs that were assayable (design score >0.60) according to the GoldenGate genotyping platform criteria. Genotypes were called using Beadstudio software (Illumina). Plates were constructed with duplicate and quality control samples. There were 24 duplicate and DNA samples for genotyping quality control. The average discordance rate of duplicates is 0.07%. We removed 12 SNPs with an MAF of 0.01 or less, 26 SNPs with call rate of <98% and 13 SNPs with discordance >0.2. The final data set consisted of 485 white patients with Lynch syndrome with genotyped results for 1456 SNPs.

In silico tools for examining functional relevance of SNPs
We explored the functional consequences of the SNPs using two searchable databases F-SNP (http://compbio.cs.queensu.ca/F-SNP/) (24) and the UCSC Genome browser (http://genome.ucsc.edu/cgi-bin/hgTracks) (25,26). The UCSC Genome browser incorporates visualization of some of the Encyclopedia of DNA elements (ENCODE) functional elements, such as regions of transcription, transcription factor association, chromatin structure and histone modification (27).

Statistical methods
The outcome variable for the analysis was time to CRC onset. The data were explored for differences in CRC age of onset by sex and MMR mutation type using the log-rank test. Hazard ratios (HRs) and 95% CIs were generated using Cox proportional hazards regression analysis to test the association of each of the 1456 SNPs with risk of CRC. All association analyses were adjusted for sex and MMR mutation type, and to allow for correlation between CRC onset age between multiple family members, we applied the Huber–White robust variance correction and clustered on the family ID (28). STATA software (version 10, StataCorp LP, College Station, TX) was used to perform the analyses.

Principal components analysis was conducted to evaluate the potential effects of population structure between Australian and US samples. There was no significant difference in eigenvector loadings for the first five factors showing that Australian–USA differences in structure were a minor source of population variability. Therefore, we did not condition the analysis on study site.

The analyses were performed in five stages, as follows:

Stage 1: Single SNP association analysis with genotypes coded as 0 = wild-type, 1 = heterozygous and 2 = homozygous variant. The models were additive (continuous effect of increasing number of variant alleles 0 versus 1 versus 2), dominant (0 versus 1 and 2), recessive (0, 1 versus 2) and genotypic (0 versus 1, 0 versus 2). All stage 1 Cox regression analyses were adjusted for sex, MMR mutation type and familial correlation.

Stage 2: SNPs that were significant in one or more of the genetic models were ranked by their smallest P value. To limit the probability of false-positives due to multiple testing, a false discovery rate method of Benjamini and Hochberg (29) was used to calculate q-value. A false discovery rate cutoff of 0.05 was applied to select the top SNPs, which limited the probability of false-positives due to multiple tests that were carried out. P values for 14 SNPs exceeded the false discovery rate cutoff.

Stage 3: The top 14 SNPs and the covariates sex and MMR mutation type were run in a Cox forward-selection regression model to select the most parsimonious model.

Stage 4: The top SNPs retained in the multivariable model were classified into favorable and unfavorable (risk) genotypes and survival analysis methods were used to determine the effect of having none versus one or more unfavorable genotypes.

Stage 5: CART analysis was used to construct survival trees to identify subgroups of patients with different risks. We applied the RPART function written in S-PLUS software (version 8.0, Insightful Corporation, Seattle, WA) to construct the risk groups.

The multistage approach allowed us to thoroughly interrogate the association signals and identify clinically meaningful risk groups.

Results
The study sample consisted of 205 men and 280 women from 272 families. All participants were self-reported non-Hispanic Whites. A majority of the subjects had mutations in MLH1 (44.9%) or MSH2 (49.5%), but 27 subjects (5.6%) had MSH6 mutations (Table I). A majority of the families (66.2%) had only one family member in the study, but the study also included families with two family members (accounting for 14.0% of families), three family members (10.3%) or four or more family members (9.5%). The median age at CRC diagnosis was older in women (58 years) than in men (48 years; log-rank test P = 0.0002), and women had a lower rate of CRC diagnosis (Table I). Similarly, the median age at CRC diagnosis was older in MSH6 mutation carriers (66 years) than in carriers of mutations in MLH1 (48 years) or MSH2 (52 years), and MSH6 mutation carriers had a lower rate of CRC diagnosis (Table I).

There were 191 SNPs associated with age at diagnosis of CRC at P < 0.05 in the adjusted Cox regression analysis, adjusting for sex, mutation type and familial correlation due to presence of multiple family members in the sample. Fourteen SNPs in 13 genes remained significant after correction for multiple comparisons (results in Table II). None of the 14 SNPs violated Hardy–Weinberg equilibrium. In the Cox forward-selection regression model, four SNPs on chromosome 5 (CDC25C: rs17171794; KDM3B/FAM53C: rs3734168; CDC25C: rs6874130 and SKP2: rs3804439) were no longer significant at P < 0.05; these SNPs were not included in the multivariable model. Three of the four SNPs that dropped out of the model, rs17171794, rs3734168 and rs6874130, were in high LD with rs3734166 (r² ≥ 60) as seen in a LD plot generated using Haploviz (30) (Supplementary Figure 1, available at Carcinogenesis Online) and were likely dropped from the model because of being correlated with rs3734166. The remaining 10 SNPs in 10 genes were significant in the multivariable model, suggesting independent effects on age of CRC onset.

As 10 genes independently influenced age-associated CRC risk, we performed combined analysis according to the number of unfavorable (risk increasing) genotypes carried by each individual with the underlying hypothesis that people with a larger number of unfavorable genotypes would be at higher risk for developing CRC at a younger age.
Table I. Subject characteristics (n = 485)* and HRs and 95% confidence intervals (CI) for colorectal cancer risk

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Australia</th>
<th>US</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of subjects (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of subjects with CRC (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HR (95% CI)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table II. Association between genetic variants in the cell cycle pathway and age of onset of CRC in non-Hispanic whites with Lynch syndrome

<table>
<thead>
<tr>
<th>Gene</th>
<th>SNP/genotype</th>
<th>MAF</th>
<th>No. of subjects with CRC/total</th>
<th>HR (95% CI)</th>
<th>P</th>
<th>HR (95% CI)</th>
<th>P (multivariate)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPP2R2B</td>
<td>rs10477307</td>
<td></td>
<td>168/352</td>
<td>1 (Reference)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GG</td>
<td>1.1</td>
<td>1.1 (0.83–1.45)</td>
<td>0.49</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GA</td>
<td>0.14</td>
<td>5.45 (3.70–8.05)</td>
<td>1.20E–17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>AA</td>
<td>0.46</td>
<td>1.51 (0.88–1.51)</td>
<td>0.29</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Recessive</td>
<td></td>
<td>5.31 (3.64–7.75)</td>
<td>4.51E–18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Additive</td>
<td></td>
<td>1.22 (0.94–1.58)</td>
<td>0.13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KIF20A</td>
<td>rs10038448</td>
<td></td>
<td>140/303</td>
<td>1 (Reference)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CC</td>
<td>0.87</td>
<td>1.20 (0.84–1.51)</td>
<td>0.13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CG</td>
<td>0.97</td>
<td>2.65 (1.79–3.91)</td>
<td>1.00E–06</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GA</td>
<td>0.21</td>
<td>3.71 (2.2–6.25)</td>
<td>8.50E–07</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Recessive</td>
<td></td>
<td>3.77 (2.25–6.31)</td>
<td>4.54E–07</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Additive</td>
<td></td>
<td>1.18 (0.91–1.53)</td>
<td>0.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TGFB1/CCDC97</td>
<td>rs12980942</td>
<td></td>
<td>178/365</td>
<td>1 (Reference)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GG</td>
<td>0.93</td>
<td>2.55 (1.72–3.78)</td>
<td>0.00001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GA</td>
<td>0.25</td>
<td>3.57 (2.90–4.38)</td>
<td>1.78 (1.03–3.08)</td>
<td>0.038</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Recessive</td>
<td></td>
<td>2.51 (1.70–3.7)</td>
<td>3.83E–06</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Additive</td>
<td></td>
<td>1.22 (0.98–1.51)</td>
<td>0.07</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDC25C</td>
<td>rs3734166</td>
<td></td>
<td>128/274</td>
<td>1 (Reference)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GG</td>
<td>0.95</td>
<td>2.5 (1.67–3.75)</td>
<td>0.00001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GA</td>
<td>0.25</td>
<td>2.5 (1.72–3.78)</td>
<td>0.00001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Recessive</td>
<td></td>
<td>2.51 (1.70–3.7)</td>
<td>3.83E–06</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Additive</td>
<td></td>
<td>1.22 (0.98–1.51)</td>
<td>0.07</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XRCC5</td>
<td>rs1051685</td>
<td></td>
<td>185/377</td>
<td>1 (Reference)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>AA</td>
<td>1.07</td>
<td>5.52 (2.37–11.5)</td>
<td>2.00E–04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>AG</td>
<td>1.12</td>
<td>5.13 (2.35–11.21)</td>
<td>0.00004</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GG</td>
<td>1.18</td>
<td>5.13 (2.35–11.21)</td>
<td>0.00004</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Recessive</td>
<td></td>
<td>5.13 (2.35–11.21)</td>
<td>7.32 (3.40–15.78)</td>
<td>3.7E–07</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Additive</td>
<td></td>
<td>1.40 (0.87–2.26)</td>
<td>0.28</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*All subjects were self-reported non-Hispanic whites.
considered unfavorable. Using this classification method, we found that subjects carried between 0 and 7 adverse genotypes. Compared with people carrying no unfavorable genotype (of any of the 10 SNPs) as the reference group, people carrying one or two unfavorable genotypes had more than twice the risk and those carrying three or more unfavorable genotypes had more than four times the risk of CRC (Table III) after adjusted for sex, type of MMR mutation and familial correlation. The median age at onset differed significantly between the three groups: it was 58 years for people with no unfavorable genotypes, 48 years for those with 1–2 unfavorable genotypes and 40 years for those with three or more unfavorable genotypes (Figure 1).

CART analysis was performed using genotypes of the 10 SNPs, sex and MMR mutation type. The final resulting tree is shown in Figure 2. There was an initial split on KIF20A: rs10038448. The subgroup with oldest age of onset of CRC (node 1) had the following characteristics: KIF20A: rs10038448 wild-type genotype (WW) and heterozygous variant genotype (WM); female; TGFB1: rs12980942 WM/MM; BCL2: rs1531697 WM/MM; and CHFR: rs11610954 WW. The median onset age in these patients was 63 years. The subgroup with the youngest age of onset of CRC, 35 years (node 5) had the following characteristics: KIF20A: rs10038448 WW/WM; female; and TGFB1: rs12980942 homozygous variant genotype (MM). Furthermore, we

The significant results after multiple test correction (FDR q-value < 0.05) were shown in bold.
used the Cox proportional hazards model to estimate HRs for all the groups and used the subgroup with the latest median age of CRC onset (node 1) as the referent (Figure 2). Because there may be a correlation of time to cancer onset in individuals from the same family due to genetic or familial factors, we applied a robust variance correction in the Cox regression analysis to adjust for the differences (34). We grouped the terminal seven nodes into three categories based on the estimated HRs for each node: low risk (node 1), moderate risk (nodes 2–4) and high risk (nodes 5–7). Compared with the low-risk group, we found HRs of 2.19 (95% CI, 1.50–3.19) and 4.67 (95% CI, 3.16–6.92) for the moderate-risk and high-risk groups, respectively (Table IV). The log-rank test (P = 4.81 × 10^{-14}) demonstrated a statistically significant difference among the time-to-onset curves of these three groups (Figure 2). The median age at onset was 42 years for the high-risk group, 50 years for the moderate-risk group and 63 years for the low-risk group.

**Discussion**

In this hypothesis-generating study, we identified 10 SNPs significantly associated with age-associated risk of CRC after correction for multiple comparisons. We used a pathways-based multigene approach to capture the combined effect of multiple SNPs in the cell cycle pathway. The analysis showed evidence of a significant gene-dosage effect. People with a larger number of unfavorable genotypes were at higher risk. Furthermore, CART analysis identified a subgroup with high probability of cancer occurrence at younger ages, a median CRC onset age of 42 years as well as a subgroup with the later age of onset, a median CRC onset age of 63 years.

The genes associated with the 10 significant SNPs were XRCC5, TTC28, TNF, TGFBI, PPP2R2B, KIF20A, CHFR, CDC25C and ATM. These genes are all directly or indirectly involved in cell cycle checkpoint control based on Ingenuity canonical pathways (Ingenuity® Systems, www.ingenuity.com) (Supplementary Figure 2, available at Carcinogenesis Online).

Many of the significant SNPs in these genes are potentially functional as they are in the 5' or 3' untranslated region (UTR) or in the coding region. The SNP rs1051685 is located in the 3' UTR region of XRCC5 and may be of functional relevance because it is located in an exonic splice enhancer sequence as determined by PupaSNP (35). The SNP rs12980942 is located in the upstream 5' UTR of TGFBI and may affect mRNA stability and translation. TGFBI plays a critical role in regulation of cell proliferation, differentiation and apoptosis and serves as a tumor suppressor in normal intestinal epithelium (36). The SNPs rs3734166 and rs1800057 are located in the coding regions of CDC25C and ATM, respectively. CDC25C is a phosphatase that serves as a regulator of G2/M transition and mediates this checkpoint in response to DNA damage. The non-synonymous SNP rs3734166 (CDC25C R70C) was significantly associated with early age of onset of CRC in our study. ATM- and Chk1/2-mediated phosphorylation of CDC25C plays a major role in G2/M arrest. Our study also found that SNP rs1800057 in ATM, resulting in the amino acid change P1045R, was associated with early age of onset of CRC. The SNP was predicted to be deleterious. Heterozygosity for P1054R is associated with decreased ATM expression in tumors (37). The SNP is in complete LD with SNP rs1800056: F858L. The two SNPs were reported to be associated with risk of CRC (38). ATM is critical for regulation of cell cycle checkpoints. Activation of ATM by DNA damage leads to ATM-dependent phosphorylation of CHEK2.

---

**Table III.** Cumulative analysis of unfavorable genotypes

<table>
<thead>
<tr>
<th>No. of unfavorable genotypes</th>
<th>No. of subjects with CRC/total</th>
<th>HR (95% CI)</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>121/292</td>
<td>1 (reference)</td>
<td>7.60E-09</td>
</tr>
<tr>
<td>1–2</td>
<td>86/157</td>
<td>2.29 (1.73–3.04)</td>
<td>1.30E-17</td>
</tr>
<tr>
<td>3 or more</td>
<td>30/36</td>
<td>4.37 (2.99–6.38)</td>
<td>2.60E-14</td>
</tr>
</tbody>
</table>

Unfavorable genotypes: PPP2R2B: rs10477307 AA, KIF20A: rs10938448 GG, TGFβ1/CDC97: rs12980942 AA, CDC25C: rs3734166 AA, XRCC5: rs1051685 GG, TNF: rs3093662 AG+GG, BCL2: rs1531697 AA, TTC28: rs9608696 GG, CHFR: rs11610954 CT+TT, ATM: rs1800057 CG.

---

**Table IV.** Results according to risk groups generated by classification and regression tree analysis

<table>
<thead>
<tr>
<th>Risk group*</th>
<th>No. of subjects with CRC/total</th>
<th>Median age at CRC onset, years</th>
<th>HR (95% CI)</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low risk (node 1)</td>
<td>75/210</td>
<td>63</td>
<td>1.00 (reference)</td>
<td>1.02E-14</td>
</tr>
<tr>
<td>Moderate risk (nodes 2–4)</td>
<td>123/227</td>
<td>50</td>
<td>2.19 (1.50–3.19)</td>
<td>0.00005</td>
</tr>
<tr>
<td>High risk (nodes 5–7)</td>
<td>39/48</td>
<td>42</td>
<td>4.67 (3.16–6.92)</td>
<td>1.02E-14</td>
</tr>
</tbody>
</table>

*See Figure 2 for nodes.
The remaining SNPs that we identified were located in the intronic regions of genes. Recently, the Encyclopedia of DNA elements (ENCODE) project has reported that 80% of the genome is related to some biochemical function after systematically mapping regions of transcription factor association, chromatin accessibility, and histone modification (27). We found that many of the intronic SNPs mapped to areas of histone modification (modification of histone proteins present in chromatin influence gene expression by changing how accessible the chromatin is to transcription), DNAsel hypersensitivity clusters (DNase hypersensitivity sites map to regions believed to contain regulatory elements, including CpG islands, and highly conserved sequences) (39), and altered transcription factor binding sites. SNPs that were predicted to alter a transcription factor binding site using the in silico tool TFSearch (http://www.crcb.jp/research/db/TFSSEARCH.html) included rs12980942 (5′ UTR of TGFB1), rs1051685 (3′ UTR of XRCC5), rs3093662 (intronic of TNF and 3′ of LTA) and rs68741430 (CDCC25). In particular, the A-to-G base change of rs3093662 was predicted to gain a cytidine diphosphate binding site; cytidine diphosphate is a transcription regulator known to be involved in cellular proliferation and cell cycle progression (40,41), which are key processes in carcinogenesis. Using the UCSC Genome browser (http://genome.ucsc.edu/cgi-bin/hgTracks), intronic SNPs related to areas of histone modification included rs10477307 (PP2R2B), rs12980942 (5′ UTR of TGFB1), rs1051685 (3′ UTR of XRCC5) and rs1531697 (BCL2) and some of these SNPs (rs10477307, rs12980942, rs1051685, rs3093662, rs1531697 and rs68741430) were also related to DNaseI hypersensitivity sites. Although intronic, many of these SNPs may therefore influence gene function. Alternatively, the significant SNPs located in intron regions of genes may be linked to other causal SNPs to affect gene activity. In addition, some of the SNPs that we found to be significantly associated with age-associated risk of CRC were associated with susceptibility to other cancers or diseases. For example, SNP rs1051685 in XRCC5 has been reported to be associated with susceptibility to myeloma (35). SNP rs12980942, in the upstream 5′ UTR of TGFB1, has been reported to be associated with increased susceptibility to asthma (42). SNP rs180057 in ATM has been found to be associated with increased risk of prostate cancer and breast cancer and to modify the effect of radiotherapy (43-47).

When we performed combined analysis for all 10 significant SNPs, we found that people with a larger number of unfavorable genotypes were at higher risk. Our findings suggest a cumulative effect of SNPs that interact in the same pathway on age-associated risk of CRC.

The CART analysis identified a few subgroups with higher risk of early onset CRC. There was an initial split on KIF20A: rs10038448, suggesting that this variation was one of the most important risk factors for CRC. SNP rs10038448 was selected as a tagging SNP for cell cycle gene CDC23 and is in complete LD with SNP rs2864, which is located in the 3′ UTR of CDC23. CDC23 is a protein essential for cell cycle progression through the G1/M transition. CDC23 is a component of the anaphase-promoting complex required for degrading mitotic cyclins and other cell cycle regulators (48). Wang et al. (49) reported that mutant human CDC23 protein decreased cell cycle progression of colon epithelial cells and was involved in expression of human Cyclin b1 protein. The tree shown in Figure 2 may provide some clues regarding how genes act together on the age-associated risk of CRC in patients with Lynch syndrome. As shown in the figure, TGFB1: rs12980942, BCL2: rs1531697 and CHFR: rs1610954 act together. TGFB1 is a well-known cell cycle inhibitor. BCL2 family proteins regulate and contribute to programmed cell death or apoptosis. It has been reported that TGFB1 elevates the protein content of the apopto-sis-preventing BCL2 (50), and BCL2 plays a crucial role in regulating the G1/S transition of hematopoietic cells induced by TGFB1 (51). Checkpoint with forkhead and ring finger domains (CHFR) functions as an important checkpoint protein early in the G1/M transition, and its activation delays entry into metaphase in response to mitotic stress (52). The tree in Figure 2 showed that the TNF gene was also a risk factor for CRC. The TNF gene encodes a multifunctional proinflammatory cytokine that belongs to the tumor necrosis factor (TNF) superfamily. Human TNF protein was reported to increase the arrest in G1 phase of
In conclusion, we identified SNPs in 10 genes that were associated with earlier age of onset of CRC in non-Hispanic white patients with Lynch syndrome, and we applied a risk modeling approach to classify individuals into different risk groups on the basis of their genotypes. Ongoing, larger and pooled GWAS analyses as well as studies in other ethnic populations may help identify additional susceptibility alleles and together these may better classify CRC risk. Such classification may help refine the frequency and intensity of screening required for these at-risk subjects. Although our results were based on a large sample size, further validation of these findings is warranted in non-Hispanic whites and in other ethnic groups.

Supplementary material

Supplementary Figures 1 and 2 can be found at http://carcin.oxfordjournals.org

Funding

The University of Texas MD Anderson Cancer Center is supported in part by the National Institutes of Health through Cancer Center Support (CA016672) and National Cancer Institute (CA 70759 to M.L.F. and K07CA160753 to M.P.).

Acknowledgements

We thank Di Zhang, Joshua D. Rother and Haidee Chancoco for technical support with sample preparation and genotyping assays, Laura Lucio for assistance with patient recruitment and Stephanie Deming for scientific editing.

Conflict of Interest Statement: None declared.

References


Received May 8, 2012; revised September 28, 2012; accepted October 25, 2012