Confirmation of Linkage to and Localization of Familial Colon Cancer Risk Haplotype on Chromosome 9q22

Courtney Gray-McGuire, Kishore Guda, Indra Adrianto, Chee Paul Lin, Leanna Natale, John D. Potter, Polly Newcomb, Robert Jenkins, John Hopper, Mark Jenkins, Joanne Young, Daniel Buchanan, Steve Gallinger, Mark Adams, Susan Lewis, Joseph Willis, Robert Elston, and Georgia L. Wiesner

Abstract
Genetic risk factors are important contributors to the development of colorectal cancer. Following the definition of a linkage signal at 9q22-31, we fine mapped this region in an independent collection of colon cancer families. We used a custom array of single-nucleotide polymorphisms (SNP) densely spaced across the candidate region, performing both single-SNP and moving-window association analyses to identify a colon neoplasia risk haplotype. Through this approach, we isolated the association effect to a five-SNP haplotype centered at 98.15 Mb on chromosome 9q. This haplotype is in strong linkage disequilibrium with the haplotype block containing HABP4 and may be a surrogate for the effect of this CD30 Ki-1 antigen. It is also in close proximity to GALNT12, also recently shown to be altered in colon tumors. We used a predictive modeling algorithm to show the contribution of this risk haplotype and surrounding candidate genes in distinguishing between colon cancer cases and healthy controls. The ability to replicate this finding, the strength of the haplotype association (odds ratio, 3.68), and the accuracy of our prediction model (~60%) all strongly support the presence of a locus for familial colon cancer on chromosome 9q. Cancer Res; 70(13); OF1–10. ©2010 AACR.

Introduction
Colorectal cancer is the second leading cause of cancer mortality in adult Americans, with 135,000 new cases and 57,000 deaths annually (1). Each American has, on average, a 6% lifetime risk of developing colorectal cancer, and although early-stage cancers are highly curable by surgery and adjuvant chemotherapy, late-stage colon cancers remain incurable (2). Both somatic and germline mutations have been associated with the development of colon cancers and their common precursor adenomatous polyps. However, familial colon cancers with known cause, familial adenomatous polyposis (FAP) and Lynch syndrome, also called hereditary nonpolyposis colorectal cancer (HNPPC), respectively account for <1% (3, 4) and ∼5% (5, 6) of all colon cancer cases annually. This leaves a large proportion of the estimated up to 35% heritability of colorectal adenocarcinoma unexplained (7, 8).

Results of linkage and association studies further support the involvement of additional genetic variants in predisposing to colon cancer. Specifically, case-control association studies have identified loci for colon cancer on 8q24 (9–11), 9p24, 18q21, 11q23, 10p14, 8q23.3 and 15q13.3, 19p13, 20p12, 16q22, and 14q22 (10, 12–14). Family linkage studies have additionally reported linkage to 3q21–24, 7q31, 11q23, and 9q22–31 (15–20). Not clear, however, are the penetrance and effect

Authors’ Affiliations: 1Department of Arthritis and Immunology, Oklahoma Medical Research Foundation, Oklahoma City, Oklahoma; Departments of 2Medicine, 3Genetics, 4Epidemiology and Biostatistics, 5Pathology and 6Case Comprehensive Cancer Center, Case Western Reserve University, Cleveland, Ohio; 7Cancer Prevention Program, Fred Hutchinson Cancer Research Center, Seattle, Washington; Departments of 8Medical Genetics, 9Health Science Research, and 10Medical Genetics, Laboratory Medicine and Pathology, Mayo Clinic, Rochester, Minnesota; 11Epidemiology Program, Cancer Research Center of Hawaii, University of Hawaii, Honolulu, Hawaii; 12Keck School of Medicine and 13Department of Preventive Medicine, USC/Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, California; 14Melbourne School of Population Health, University of Melbourne, Melbourne, Victoria, Australia; 15Division of Genetics and Population Health, Queensland Institute of Medical Research, Brisbane, Queensland, Australia; 16Samuel Lunenfeld Research Institute, Toronto General Hospital, Toronto, Ontario, Canada; 17Howard Hughes Medical Institute, Chevy Chase, Maryland; and 18Center for Human Genetics and Departments of 19Medicine, Hematology/Oncology, and 20Pathology, University Hospitals Case Medical Center, Cleveland, Ohio.

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

C. Gray-McGuire and K. Guda contributed equally to this work.

The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the Colon Cancer Family Registries, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government or the Colon Cancer Family Registry.

Corresponding Author: Georgia L. Wiesner, Center for Human Genetics, University Hospitals Case Medical Center, 11100 Euclid Avenue, LKSD 1500, Cleveland, OH 44106. Phone: 216-844-3938; Fax: 216-844-7497; E-mail: ghw2@case.edu.

doi: 10.1158/0008-5472.CAN-10-0188

©2010 American Association for Cancer Research.
size of each of these identified variants, nor why they are found in some studies and not in others.

It was the purpose of this study first to replicate, in an independent sample, the linkage finding at 9q22-31 that we initially identified in a genome scan of 53 kindreds with multiple colon cancer and/or advanced colon adenoma cases (20) and then confirmed by two other studies (17, 18). Indeed, we have narrowed the linkage on 9q22-31 from 13.5 to 7.7 cM and show here that the addition of 69 independent colon neoplasia kindreds increases evidence of linkage to this region. Second, we report further isolation of this effect via a family-based association analysis of >3,000 single-nucleotide polymorphisms (SNP) very densely spaced across the region of interest. Third, we identify a five-SNP haplotype associated with risk of early-onset familial colon neoplasia and its estimated effect size in our study sample. Finally, we offer some reconciliation of the inconsistent results of the many linkage and association studies of colon cancer to date.

Materials and Methods

Replication and localization of linkage signal

Sample. To further validate our initial finding of linkage to D9S1786 for familial colon neoplasia, we repeated our linkage analysis using a cohort of 70 colon neoplasia kindreds (herein referred to as the confirmation collection), each with at least two affected siblings enrolled and DNA available for genotyping. However, 7 of these kindreds came from our own Colon Neoplasia Sibling Study (CNSS) and another 63 were independently recruited by the Colon Cancer Family Registry (CCFR), a National Cancer Institute–supported consortium established in 1997 to create a multinational comprehensive collaborative infrastructure for interdisciplinary studies in the genetic epidemiology of colorectal cancer. Detailed information about the CCFR can be found at http://epi.grants.cancer.gov/CFR/, as well as being described in detail by Newcomb and colleagues (21). The CCFR registries that were used to identify eligible cases are located at The Fred Hutchinson Cancer Research Center (Seattle, WA), University of Hawaii Cancer Research Center, Cancer Care Ontario, University of Melbourne Australasia Colorectal Cancer Family Registry, University of Southern California Consortium, and Mayo Clinic as well as a CCFR data center at the University of California, Irvine. This study included data from CCFR participants recruited from both population-based sources and clinic-based sources.

Additionally, the original collection of 53 kindred from the CNSS reported by Wiesner and colleagues (20) was modified in view of updated affection status in family members (changed two affected to unaffected resulting in the loss of two families); these changes resulted in a data set (herein referred to as the revised original collection) comprising 93 discordant sibling pairs, 67 concordantly affected pairs, and 22 concordantly unaffected pairs, all of whom are informative for linkage. All kindreds who self-reported to be of European descent, but collected from multiple sites, were included. Details of the enrollment procedures were published previously, but, in brief, we classified individuals as affected if and only if they had a confirmed diagnosis of colorectal cancer or adenomatous polyp ≥1 cm by age 65 and no inflammatory bowel disease, FAP, or Lynch syndrome/HNPCC. Persons for whom tumors were available were tested for microsatellite instability and, if positive, removed from the sample. Unique to the CNSS study, individuals were classified as unaffected only if they had undergone an endoscopy and no cancer or adenomatous polyps were found, and classified as unknown if they had not been screened.

The expanded cohort (herein referred to as the combined collection) consisted of 121 families (51 from the original collection, 7 additional CNSS, and 63 additional from the CCFR) and 272 sibling pairs and included the revised original collection and the new independent cohort. However, because linkage analysis and association analyses are susceptible to bias (either toward or away from the null) due to population stratification, 10 families were excluded because of non-Caucasian or incomplete ancestry information and 1 for double enrollment by two different collection sites. The final family sample therefore comprised 256 sibling pairs from 110 families (50 from the original collection, 6 additional CNSS, and 54 additional from the CCFR; Table 1).

Linkage analysis. To confirm our initial linkage finding, we performed linkage analysis on these 256 siblings pairs. Rather than simply genotyping the 5 markers identified in our original sample, we opted to genotype an additional 17 fine-mapping markers across the originally defined 13.5-Mb linkage interval (average spacing of ∼1.6 cM) in both the revised original and combined collection (Supplementary Table S1). This offered us the potential to further localize, as well as replicate, our result.

| Table 1. Sample description for revised original, confirmation, and combined pedigree collection both before and after exclusions |
|-----------------------------|-----------------------------------|-----------------|-------------------|
| **Collection**              | **Source**                        | **Sample before exclusions** | **Sample after exclusions** |
|                            |                                   | **Kindreds/sibpairs**     | **Kindreds/sibpairs** |
| Revised original           | CNSS                              | 51/182                   | 50/179             |
| Confirmation               | CCFR                              | 63/75                    | 54/63              |
|                            | CNSS                              | 7/15                     | 6/14               |
| Combined                   | CCFR and CNSS                     | 121/272                  | 110/256            |
Before linkage analysis, the genotype data were examined for Mendelian inconsistencies using MARKERINFO in the Statistical Analysis for Genetic Epidemiology program (S.A.G.E. 5.4) and for Hardy-Weinberg proportion disequilibrium in family data using FREQ (S.A.G.E. 5.4; ref. 22). Those genotypes believed to be erroneous were removed from subsequent analysis. We then calculated multipoint identity-by-descent (IBD) sharing estimates every 2 cM using GENIBD (S.A.G.E. 5.4) and used those estimates to perform two tests for linkage with SIBPAL (S.A.G.E. 5.4): the weighted Haseman-Elston regression method and the mean tests for concordant affected, discordant, and concordant unaffected siblings. The weighted Haseman-Elston regression method regresses on the multipoint IBD sharing estimates using as the dependent variable a weighted average of the squared sibpair trait difference and the squared mean-corrected sum expressed as follows

\[
\left( w\left( (y_i - \bar{y}) - (y_j - \bar{y}) \right)^2 - (1 - w) \left( (y_i - \bar{y}) + (y_j - \bar{y}) \right)^2 \right) / 4
\]

where \( \bar{y} \) is the best linear unbiased predictor including an adjustment for the population prevalence (23) and the chosen weights are optimal if the sample size is large enough. Note that the Haseman-Elston regression method relies on the presence of, at minimum, two sibling pairs from among concordant affected, discordant, or concordant unaffected pairs. For this reason, we could not perform this test in our confirmation sample alone as it comprised only affected sibling pairs. The mean tests, however, assess the statistical significance of the departure of the observed IBD sharing estimates in each of the respective pair types from what would be expected in the absence of linkage. This test was therefore conducted in the revised original, confirmation, and combined collections.

Finally, using the results of these tests for linkage and the IBD sharing estimates, we ranked the families on their likelihood of being linked to this region via the quantitative linkage score (QLS) proposed by Wang and Elston (24) and expressed as follows:

\[
U_{ij} = (y_i - \bar{y}) (y_j - \bar{y}) (IBD_{ij} - 0.5)
\]

This QLS allowed us to prioritize families within which to pursue the addition of both marker genotypes and family members, where \( i \) and \( j \) are two siblings in a family and \( \bar{y} \) can be fixed at any value. In our case, we conservatively classified families as linked if they had positive mean and sib-specific QLS scores for all of \( \bar{y} = 0.25, 0.5, \) and 0.75.

**Association analysis and risk haplotype identification**

**Sample and genotyping.** To further localize and assess risk attributable to the variant at 9q22,31, we conducted a joint family-based and case-control association study from 106 of our families (4 families were removed from the study because of low DNA concentration) comprising 222 affected and 48 unaffected sibpairs and 201 additional, independent controls. Controls were selected from individuals who presented for a colonoscopy at University Hospitals Case Medical Center in Cleveland, Ohio, were independent of the families studied, had clean colonoscopies, were at least 60 years old, had no personal history of cancer, and had no more than one first-degree relative with reported cancer of any type.

We genotyped 2,699 SNPs across the 13.5 cM region of interest, with an average spacing of 4,000 bp (4 kb). The SNPs were chosen initially based on tagging using the Human HapMap CEU (Caucasian European from Utah) samples, imposing a linkage disequilibrium (LD) threshold of \( r^2 = 0.8 \); we then added additional SNPs to achieve an average spacing of 4 kb and ensure adequate coverage given strong Caucasian LD in this region. Based on the HapMap CEPH samples, tagging alone would have resulted in one SNP every 5.4 kb with several LD blocks spanning >10 kb. All SNPs had a minor allele frequency of at least 5% and were either golden-gate or double-hit validated. All SNP genotypes were collected using the Illumina Bead Station platform via the Case Comprehensive Cancer Center genotyping core.

**Single-SNP and moving-window association analysis.** We performed both single-SNP and moving-window analysis of our SNP genotype data using a regression model, based on Elston and colleagues (25), of the following form:

\[
h(y)_i = h(\alpha + \gamma_i c_1 + \gamma_2 c_2 + \ldots + \gamma_n c_n + \delta z_i) + p_i + f_i
\]

where for any individual \( i \), with liability \( y_i \) and \( n \) covariate values \( c_j \). In this formulation, \( h \) is the logit link function in the context of a generalized linear mixed model under the assumption that the random effects are normally distributed, \( p_i \) is a random polygenic effect, \( f_i \) and \( f_j \) are random common nuclear family effects, \( m_j \) is a random marital effect, \( s_j \) is a random common sibship effect, \( z_i \) is a random individual effect, and \( c_j \) is a genotype indicator for the allele A at a diallelic locus with alleles A and B, such that, when considering only a single locus under an additive model,

\[
z_i = \begin{cases} 
    -1 & \text{for genotype BB} \\
    0 & \text{for genotype AB} \\
    1 & \text{for genotype AA}
\end{cases}
\]

In addition to testing one SNP at a time, we used a multilocus approach to produce more robust results and potentially improve power under conditions in which multiple SNP markers are associated with the disease or in LD with a causal locus. To simultaneously analyze multiple nearby correlated genetic variants, we incorporated a moving-window approach into our family-based association method. For an appropriate window width (e.g., five SNPs), we moved the window from the first marker to the last marker in the candidate genome region. For each window, we fitted the regression model using all markers within the window and calculated the corresponding \( P \) values based on asymptotic theory. Assume \( k \) markers within the window. For any individual \( i \), with trait \( y_i \) and \( j \)-th covariate values \( c_{ij} \), the
regression model shown above now takes the following form:

\[ h(y_i) = h(\alpha + \gamma c_{i1} + \gamma c_{i2} + \ldots + \gamma c_{in} + \delta (z_{i1}, z_{i2}, \ldots, z_{in})) + p_i + f_i + f'_i + m_i + e_i \]

where \( z_i \) is a genotype indicator as indicated above and \( f(g) \) is the smoothing function, which here we chose to be the mean of the additive genetic variants.

To identify the optimal window size while avoiding the penalty of performing additional tests, we blinded ourselves to the SNP names and then plotted, from the results of moving windows of various sizes, the \( P \) values for the likelihood ratio test (LRT) against the \( P \) values for the one-degree-of-freedom Wald \( \chi^2 \) test for each group of SNPs. Agreement in these two test statistics serves as an indication of stability and therefore represents the most appropriate window size. With a linear correlation coefficient of 0.98 between the Wald and LRT \( P \) values, the window size of 5 was deemed most appropriate (compared with 0.85, 0.89, 0.93, and 0.87 for window sizes of 1, 3, 7, and 9, respectively).

**Haplotype identification.** After identifying significant regions based on both our single-SNP and moving-window analyses, using the full sample, we constructed a statistically predicted risk haplotype based on the estimated odds ratios (OR) from the regression analysis outlined above. For example, if the “A” allele was coded as the reference allele and the OR was >1, then A was deemed as the risk allele, and if the OR was <1, then “B” was deemed the risk allele and so forth for each SNP marker in the haplotype.

**Molecular haplotype confirmation.** We were then able to verify these haplotypes in a select sample of 12 individuals from the CNSS family collection for whom we already had molecular haplotypes obtained by genotyping of uniparental monochromosomal somatic cell hybrids (converted clones) as described by Yan and colleagues (26). We then calculated, using the entire sample, the numbers of individuals who were unambiguous carriers of two or zero copies of the risk haplotype, respectively, but included in the counts only a single count per family, even if multiple members were informative. We tested the difference in counts of the risk haplotype between cases and controls using either a Pearson’s \( \chi^2 \) or a Fisher’s exact test, as appropriate.

**Prediction modeling.** The ability of a group of SNPs to predict cancer versus no cancer in our data set was assessed through use of recursive feature elimination in a support vector machine framework (SVM-RFE). Weka data mining software (27) was used to perform the SVM-RFE experiments with a 10-fold cross-validation. Cross-validation estimates how well a given model predicts the same outcome in an unseen data set drawn from the same statistical distribution (28). A 10-fold cross-validation randomly divided the data set into 10 equal subsets. The first 9 subsets were used as training data to construct a predictive model and the 10th was used as test data. Then, the 2nd through 10th subsets were used as training data, and the first subset was used as test data. We stopped this procedure after each subset had been used as test data. The performance measures of our predictive model can be expressed by accuracy, sensitivity, and specificity on the test data. Accuracy is the probability that a subject will be correctly classified as either a case or control by a predictive model. Sensitivity refers to the probability of a positive prediction among patients with disease. Specificity measures the probability of a negative prediction among subjects without disease. A good predictive model should simultaneously optimize accuracy, sensitivity, and specificity. We determined the threshold for these calculations by selecting the point on the probability output of our predictive model that gave the highest score common to these three performance measures. It is important to note that we performed our predictive modeling on a random group of independent individuals (102 cases, one from each family: 201 controls).

**Results**

**Replication and localization of linkage signal**

The strongest signal in the Haseman-Elston linkage analysis of the revised original and combined collection of families was with marker D9S1786 (−log \( P \) values = 3.35, and 3.38, respectively), exactly at the marker location (Fig. 1). The mean test yielded \( P \) values of 0.0005, 0.04, and 0.0001 for the revised original, confirmation, and combined collections, respectively. The equivalent of a 1.1-LOD drop from the linkage maximum defined a 7.5 cM (8.8 Mb) linkage interval bounded by D9S1815-D9S1857 (29). As shown, this expanded colon neoplasia cohort showed increased statistical significance for
linkage of disease to the prespecified marker D9S1786, with the $P$ value for linkage in this expanded cohort = 0.00016, a 3-fold increase in significance from the value of 0.00045 seen in our initial study (20). This increase, as well as the fact that the mean test analysis of the confirmation sample met the $P < 0.05$ criterion touted as necessary for significance, shows confirmation of the same linkage among the newly added kinships. It additionally narrows the linkage region to an 8.8-Mb interval. Also of note, in this combined collection, concordantly affected sibling pairs show an excess over 0.5 IBD sharing of 0.60, which corresponds to 0.60% of colon neoplasia kindreds in the expanded analysis being linked to a potentially autosomal dominant disease gene at 9q22.2-31.2 (with 95% confidence limits for this estimate now being 21–59%). Thus, this combined collection not only provides replication of our initial finding of linkage of familial colon neoplasia to 9q22.2-31.2 but also strengthens our conclusion that this locus accounts for the development of disease in a major subgroup of colon neoplasia kindreds.

Further, in our follow-up linkage analysis, 25 of our best linked CNSS families had been extended by an additional 3 affected and 42 unaffected family members not available for our originally published linkage study (20). Model-based linkage analysis of these extended families led to an increased LOD score for both the recessive and dominant model (from 3.953 to 4.277 and 2.75 to 3.01, respectively). Despite the fact that these families were identified as being linked and targeted for the addition of family members, and therefore these results are subject to the effects of ascertainment, the increase in LOD score supports the presence of a disease variant in this region.

**Association analysis and risk haplotype identification**

Analysis of each SNP individually did not produce any SNPs meeting a very conservative $P$ value of $2.8 \times 10^{-5}$ after Bonferroni correction [based on the effective number of SNPs analyzed (2,699) $\times$ average proportion of LD (0.67) = 1,808 independent tests]. However, each of three regions, centered at approximately 92, 98, and 102 Mb, was suggestive of association with SNPs with $P$ values of $<2.5 \times 10^{-3}$ (Fig. 2).

The moving-window analysis produced, as expected, results that were more robust (i.e., less sensitive to the requirements of asymptotic assumptions, as explained in Materials and Methods), more precise, and, in this case, statistically significant. Although the increased statistical significance could be due to increased type I error, in view of the uniformity of the increases, it is more likely to represent a gain in power due to multiple SNPs in the same region being associated with a disease locus or in LD with a causal locus. The three regions that were statistically significant in the single-SNP analysis were also significant in the moving-window analyses, with significance increasing to $P < 1.0 \times 10^{-5}$, $P < 2.5 \times 10^{-5}$, and $P < 5.0 \times 10^{-5}$, respectively. As can be seen in Fig. 3, the moving-window approach also resulted in a smoothing of the peaks, with more associations clustering at ~98 Mb. We further characterized these regions by examining the LD structure within and among the three regions, as well as calculating the OR and 95% confidence interval for the most significant regional SNPs. For the region at 98 Mb, there was a break in statistical significance (and therefore the risk haplotype) between 98,157,208 and 98,296,272 bp. We therefore characterized separately the two regions: rs7865648 to rs3780442 (centered at 98,146,452 bp) and rs10818948 to rs1953087 (centered at 98,298,339 bp). We therefore characterized separately the two regions: rs7865648 to rs3780442 (centered at 98,146,452 bp) and rs10818948 to rs1953087 (centered at 98,298,339 bp). There is strong LD across all regions (Fig. 3) and specifically between rs10820943 and rs3802477 ($r^2 > 0.9$, LOD > 2) and between rs998952 and rs10818948 ($r^2 > 0.9$, LOD > 2; Fig. 3). Further, when confining the analysis to only those families most linked to this region, the association becomes stronger and spans even more SNPs, ultimately pointing to
the region centered at 98.15 Mb as the most likely to house a causal variant.

This result was further supported by haplotype analysis. In fact, we found 6 cases compared with 4 controls with two copies of the five-SNP risk haplotype at 98.15 Mb, as well as almost twice as many controls (148 compared with 83 cases) with no copies (OR, 3.68, comparing cases with controls with two or zero copies).

**Molecular confirmation**

To verify that the statistically predicted haplotype indeed exists, as described, we additionally genotyped the uniparental monochromosomal somatic cell hybrids and verified that the haplotype of interest centering ~98.15 Mb is indeed present. In fact, 9 of the 12 affected persons from our most-linked families for whom we have converted clones had at least one copy of the risk haplotype, whereas only 3 had none. In contrast, the risk haplotype in the region centered at 98.29 Mb was observed in duplicate (two copies) in almost equal numbers in cases and controls (OR, 0.86). No individuals were homozygous for all risk alleles across the region, and therefore, none could be unambiguously classified as carrying the risk haplotype in the region centering ~92 Mb and similarly for the region ~102 Mb.

**Prediction modeling**

Using SVM in a recursive framework as described above, we were able to predict colon cancer using a specified subset of SNPs with sensitivity, specificity, and accuracy each just under 60%. We arrived at this value by varying,
from 0 to 1, the probability threshold for which we declared someone correctly classified. We then plotted the proportion of correctly classified samples using cases alone (specificity), controls alone (sensitivity), and the full sample (accuracy). The point of intersection of each of the three attributes of interest occurred at a probability threshold of 0.365 (Fig. 4). When including all SNPs contained within the four associated regions outlined above, the subset of SNPs that best predicted colon cancer in these data spans the region 98.145-98.296 (Fig. 5A). This result affirmed our previous conclusion because SVM does not depend on strength of association but rather on predictability. Additional SVM analysis incorporating SNPs within candidate genes in the region (Fig. 5B) further improved these predictive models and is discussed below.

Discussion

The results presented here confirm the validity of our initially published finding of statistically significant linkage of familial colon neoplasia to a chromosome 9q22.2-31.2 disease locus. Despite the fact that some of the recent genome-wide association scans do not report association to SNPs in the 9q22.2-31.2 region, we have shown a gain in statistical significance both by adding independent families and by expanding existing families where possible. These linkage results also show, as is not possible with case-control association analyses, that this signal is not due to population stratification or some other form of ascertainment bias and is therefore unlikely to be a false-positive result.

We have further isolated this signal via combined family and population-based association analysis to a 151,602-bp region centering at 98.15 Mb. This localization was verified via LD characterization and haplotype analysis. Specifically, as can be seen in Fig. 3, the LD structure of this region shows the strength of LD between the other two regions of significance, helping to show that the region at 98.15 most likely represents the causal variant. The haplotype analysis confirms this with an OR of 2.76 for cases compared with controls when scored for two or zero copies of the risk allele. Furthermore, an analysis of all individuals in the family data set (not just the randomly selected independent cases and controls) increased the OR to 3.68. Although a correlation between related individuals could inflate the OR, it may be a more conservative estimate because of the stronger difference between related cases and controls. This result also verifies that the findings are not an artifact of our choice of controls. Finally, the predictive modeling via SVM, which has been successfully used to model other cancers, including breast (28) and esophageal cancer (30), supports this association. It is important, however, to validate any genetic association, and the collection of an independent replication sample is currently under way.

Although well-supported statistically, we recognize that without biological evidence, the causal variant may only be represented by, but not actually contained within, this figure.
Figure 6. LD plot of regions with significant moving-window association results and in LD with candidate genes for all cases (n = 225; A) and controls (n = 248; B). Solid ovals indicate significant regions, the solid box indicates the risk haplotype at 98.15 Mb, brackets indicate genes, and dashed ovals highlight regions of LD between the haplotype block containing HABP4 and the four regions of significance.

Region 1, the R-squared is $0.05 < R^2 < 0.11$
Region 2, the R-squared is $0.01 < R^2 < 0.05$
Region 3-5, the R-squared is $0.03 < R^2 < 0.06$
haplotype block. It is to our good fortune that the candidate region in which we are interested has been fairly well characterized and houses multiple genes (ZNF367, HABP4, GABBR2, and GALNT12). Two of these plausibly have a functional effect on cancer, specifically the hyaluronan binding protein 4 gene (HABP4), also called Ki-1/57, and the GalNac transferase 12 gene (GALNT12). HABP4 encodes a CD30 Ki-1 antigen first discovered as a marker for Reed-Sternberg cells in Hodgkin lymphoma but was later found to be expressed in a variety of cell lines, including normal lymphocytes and monocyte-derived macrophages. The Ki-57 molecule, with which Ki-1 interacts, occurs intracellularly only in the cytoplasm, nuclear pores, and the nucleus (31, 32). This antigen has also been shown to interact with the chromohelicase–DNA-binding domain protein 3, a nuclear protein involved in the regulation of transcription and chromatin remodeling, and the receptor of activated kinase 1, and further coprecipitates protein kinase C (PKC). PKC is a tumor promoter and has been extensively studied and linked to breast, bladder, skin, and other forms of cancer. GALNT12 has been suggested to play a role in the initial step of mucin-type oligosaccharide biosynthesis in digestive organs (33) and has been shown to be highly expressed in digestive organs such as small intestine, stomach, pancreas, and colon and moderately expressed in testis, thyroid gland, and spleen. Recent studies from our group report both rare germline GALNT12 mutations that are present in some individuals who develop colon cancer and rare somatic GALNT12 mutations in certain colon cancer tumors (34).

It is therefore reasonable to ask if our association signal points to either of the above-mentioned genes as the most likely candidate. Five of the SNPs typed as a part of our association study lie within HABP4 (4 intronic and 1 3′ untranslated region) and 11 SNPs lie within GALNT12 (11 intronic and 1 synonymous coding SNP). Although none of these SNPs met our threshold for statistical significance in the association analysis, there are examples in the literature of causal variants identified within genes outside of the region of greatest statistical significance in an association study (35). Further, a comparison of the LD structure in the cases and the controls in our data set suggests that the underlying variants are likely uncommon. All of this suggests that the disease locus housed on 9q is specific to a familial syndrome with a phenotype of younger age of onset (<75, <45, and <35, respectively; ref. 18). Further, by supplementing a set of tag SNPs with additional, more uniformly spaced SNPs, we were able to capture a much larger proportion of the variability in this region. We point toward the studies by the CORGI consortium as support for these differences explaining the various results; they did not find the signal at 9q in 69 families with a mixture of colon cancer, adenomas, and polyps (15), but they did replicate our findings in 57 families with three or more affected persons using strict age-of-onset criteria for the three phenotypes mentioned above (<75, <45, and <35, respectively; ref. 18). All of this suggests that the disease locus housed on 9q is specific to a familial syndrome with a phenotype of younger age of onset and/or severity of colon neoplasia. Finally, whereas the prevalence of the syndrome we describe is unknown, we suggest that the underlying variants are likely uncommon. This and further characterization of the effect at 9q are the subject of ongoing research.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

We thank the individuals and families who participated in this study.

Grant Support

The results of this paper were obtained by using the program package S.A.G.E., which is supported by USPHS Resource Grant RR03655 from the National Center for Research Resources. This work was also supported by...
the Prevent Cancer Foundation; the NIH National Cancer Institute; and National Institute of General Medical Sciences, USPHS awards R01 CA130901, R01 CA104667, P30 CA043703, and R01GM28356 and through cooperative agreements with members of the CCFR. Each CCFR center that provided data for the analysis was supported as follows: Australasian Colorectal Cancer Family Registry (U15 CA997735), Familial Colorectal Neoplasia Collaborative Group (U01 CA074799), Mayo Clinic Cooperative Family Registry for Colon Cancer Studies (U01 CA074800), Ontario Registry for Studies of Familial Colorectal Cancer (U01 CA074783), Seattle Colorectal Cancer Family Registry (U01 CA074794), University of Hawaii Colorectal Cancer Family Registry (U01 CA074806), and University of California, Irvine Informatics Center (U01 CA078296).

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 01/17/2010; revised 03/25/2010; accepted 04/14/2010; published OnlineFirst 06/15/2010.

References

Confirmation of Linkage to and Localization of Familial Colon Cancer Risk Haplotype on Chromosome 9q22


Cancer Res  Published OnlineFirst June 15, 2010.