Colorectal Cancer Linkage on Chromosomes 4q21, 8q13, 12q24, and 15q22

Mine S. Cicek¹, Julie M. Cunningham¹, Brooke L. Fridley¹, Daniel J. Serie¹, William R. Bamlet¹, Brenda Diergaard², Robert W. Haile³, Loic Le Marchand⁴, Theodore G. Krontiris⁵, H. Banfield Younghusband⁶, Steven Gallinger⁷, Polly A. Newcomb⁸, John L. Hopper⁹, Mark A. Jenkins⁹, Graham Casey⁹, Fredrick Schumacher⁹, Zhu Chen⁹, Melissa S. DeRycke¹, Allyson S. Templeton⁸, Ingrid Winship⁹, Roger C. Green⁶, Jane S. Green⁶, Finlay A. Macrae¹⁰, Susan Parry¹¹,¹⁶, Graeme P. Young¹², Joanne P. Young¹³, Daniel Buchanan¹³, Duncan C. Thomas³, D. Timothy Bishop¹⁴, Noralane M. Lindor¹, Stephen N. Thibodeau¹, John D. Potter¹⁵,¹⁶, Ellen L. Goode¹, for the Colon CFR

1 Departments of Health Sciences Research, Laboratory Medicine and Pathology, and Medical Genetics, Mayo Clinic College of Medicine, Rochester, Minnesota, United States of America, 2 Department of Epidemiology, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America, 3 Department of Preventive Medicine, University of Southern California, Los Angeles, California, United States of America, 4 University of Hawaii Cancer Center, Honolulu, Hawaii, United States of America, 5 Department of Molecular Medicine, Beckman Research Institute of the City of Hope, Duarte, California, United States of America, 6 Faculty of Medicine, Memorial University of Newfoundland, St. Johns, Newfoundland, Canada, 7 Department of Surgery, University of Toronto, Toronto, Ontario, Canada, 8 Cancer Prevention Program, Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America, 9 Departments of Public Health and Medicine, University of Melbourne, Victoria, Australia, 10 Colorectal Medicine and Genetics and Department of Medicine, University of Melbourne, The Royal Melbourne Hospital, Victoria, Australia, 11 New Zealand Familial GI Cancer Registry, Auckland City Hospital, Auckland, New Zealand, 12 Flinders Centre for Cancer Prevention and Control, Flinders University, Adelaide, Australia, 13 Familial Cancer Laboratory, Queensland Institute of Medical Research, Queensland, Australia, 14 University of Leeds, Leeds Institute of Molecular Medicine, Leeds, United Kingdom, 15 Centre for Public Health Research, Massey University, Wellington, New Zealand, 16 Department of Gastroenterology, Middlemore Hospital, Auckland, New Zealand

Abstract

A substantial proportion of familial colorectal cancer (CRC) is not a consequence of known susceptibility loci, such as mismatch repair (MMR) genes, supporting the existence of additional loci. To identify novel CRC loci, we conducted a genome-wide linkage scan in 356 white families with no evidence of defective MMR (i.e., no loss of tumor expression of MMR proteins, no microsatellite instability (MSI)-high tumors, or no evidence of linkage to MMR genes). Families were ascertained via the Colon Cancer Family Registry multi-site NCI-supported consortium (Colon CFR), the City of Hope Comprehensive Cancer Center, and Memorial University of Newfoundland. A total of 1,612 individuals (average 5.0 per family including 2.2 affected) were genotyped using genome-wide single nucleotide polymorphism linkage arrays; parametric and non-parametric linkage analysis used MERLIN in a priori-defined family groups. Five lod scores greater than 3.0 were observed assuming heterogeneity. The greatest were among families with mean age of diagnosis less than 50 years at 4q21.1 (dominant HLOD = 4.51, α = 0.84, 145.40 cM, rs10518142) and among all families at 12q24.32 (dominant HLOD = 4.30, α = 0.32, 132.52 cM, rs10518142). Among families with only two affected individuals yielded a peak at 8q13.2 (recessive HLOD = 3.60, α = 0.48, 285.15 cM, rs952093). Among families with four or more affected individuals and among clinic-based families, a common peak was observed at 15q22.31 (101.40 cM, rs1477798; dominant HLOD = 3.60, α = 0.32, 132.52 cM, rs10518142). These previously unreported linkage peaks demonstrate the continued utility of family-based data in complex traits and suggest that new CRC risk alleles remain to be elucidated.

Citation: Cicek MS, Cunningham JM, Fridley BL, Serie DJ, Bamlet WR, et al. (2012) Colorectal Cancer Linkage on Chromosomes 4q21, 8q13, 12q24, and 15q22. PLoS ONE 7(5): e38175. doi:10.1371/journal.pone.0038175

Editor: Anthony W. I. Lo, The Chinese University of Hong Kong, Hong Kong

Received March 14, 2012; Accepted May 1, 2012; Published May 31, 2012

Copyright: © 2012 Cicek et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by R01-CA104667 through cooperative agreements with the members of the Colon Cancer Family Registry and PIs. Collaborating centers include the Australian Colorectal Cancer Family Registry (U01 CA097735), the USC 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 funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: egoode@mayo.edu

Introduction

Colorectal cancer (CRC) is the third most common cancer and the third leading cause of cancer death in the United States. Approximately 141,210 new cases and 49,380 deaths from CRC were expected in the United States in 2011 [1]. Family history is a consistent risk factor [2]; without CRC family history, the lifetime risk for an individual above the age of 50 years is 5% to 6%, yet this can be as high as 20% when there are first- or second-degree relatives with CRC [3–5], and reaches 80% to 100% in familial...
syndromes [6]. Lynch syndrome represents up to 5% of CRCs and results from germline mutations in one of several DNA mismatch repair (MMR) genes (MLH1, MSH2, MSH6 and PMS2) [7]. MMR mutations result in a defective mismatch repair (dMMR) tumor phenotype manifested by absence of MMR protein expression [8,9] and DNA microsatellite instability (MSI-H). Segregation analyses excluding Lynch syndrome families suggest that additional loci for CRC susceptibility exist [5].

To identify novel loci, case-control association studies and family-based linkage analyses serve as complementary approaches. At least fifteen, common low-penetrance risk alleles have emerged from genome-wide association studies (GWAS) including at 1q41 (rs6691170, DUSP10) [10], 3q26.2 (rs10936599, MLYN) [10], 8q23.3 (rs16892766, ELF3HI) [11], 8q24 (rs6938267) [12,13], 9q24 (rs719725) [14], 10p14 (rs1075668) [11], 11q23 (rs3802842) [15], 12q13.13 (rs11695552) [10], 14q22.2 (rs4444235, BAP1) [16], 15q13 (rs4779584) [15], 16q22.1 (rs9929218, CDH1) [16], 18q21 (rs4939827, SMAD7) [17], 19q13.1 (rs10411210, RHPSY2, 20p12.3 (rs961253) [16], and 20q13.33 (rs4925386, LAMA5) [10]. Studies of CRC linkage in multi-case families or affected sibling pairs have reported evidence of rare, high-risk variants in several genetic regions including 3q21-24, 7q31, 9q22-31, and 11q23 [18–26]. Most linkage studies to date have utilized fewer than 100 families, and only two studies excluded dMMR families [18,26].

Here, we describe a genome-wide linkage scan of 556 white families without evidence of dMMR using family groups defined by age at diagnosis, ascertainment method, and number of families. This represents the largest linkage study of proficient MMR (pMMR) CRC families to date and suggests novel regions with evidence of high-penetrance loci.

Materials and Methods

Ethics Statement


Ascertainment and Collection of Families

A total of 578 linkage-informative families were identified having at least two affected individuals diagnosed with invasive CRC in sibling, half-sibling, cousin, grand-parental, or avuncular pairs [27], absence of sequence-confirmed Lynch Syndrome and MYH-associated polyposis [28], and absence of medical-record-confirmed familial adenomatous polyposis.

The majority of families (N = 480) were from the Colon Cancer Family Registry multi-site NCI-supported consortium (Colon CFR) ascertained between 1997 and 2007 by Cancer Care Ontario (Toronto, Canada), a University of Southern California Consortium (Los Angeles, CA), a University of Melbourne Consortium (Victoria, Australia), the University of Hawaii (Honolulu, HI), the Mayo Clinic (Rochester, MN), and the Fred Hutchinson Cancer Research Center (Seattle, WA) [28]. All study sites ascertained population-based families, although varying sampling schemes based upon age and/or family history were used. Clinic-based families were ascertained by the University of Melbourne Consortium (through family cancer clinics in Adelaide, Perth, Sydney, Brisbane, and Melbourne, Australia and Auckland, New Zealand), the University of Southern California Consortium (through the Cleveland Clinic), and the Mayo Clinic. Epidemiologic data, blood samples, tumor blocks, and pathology reports were collected on all participants with CRC at each site, using standardized core protocols.

Clinic-based families from a City of Hope consortium (N = 59) were recruited between 1998 and 2005 at the City of Hope (Duarte, CA), Tufts University (Medford, MA), the University of Pittsburgh (Pittsburgh, PA), Northwestern University (Chicago, IL), the University of Wisconsin (Madison, WI), Vanderbilt University (Nashville, TN), the University of South Florida/ Moffitt Cancer Center (Tampa, FL), Maine Medical Center (Portland, ME), and Rose Medical Center (Denver, CO). White CRC cases older than 18 years of age, who had at least one living sibling diagnosed with CRC, were enrolled. Blood samples, pathology reports, and a brief questionnaire focused on ethnicity and family history were collected on all cases.

Population-based and clinic-based families (N = 39) from Newfoundland and Labrador, Canada were obtained at Memorial University of Newfoundland as previously described [29,30]. Briefly, pathologically confirmed cases diagnosed under the age of 75 years were enrolled via the provincial tumor registry between 1997 and 2003. Epidemiologic data including family history and risk factors, blood samples, tumor tissue, and pathology reports were collected. Clinic-based families were contacted following referrals to the high-risk cancer clinic of the provincial Medical Genetics Program.

SNP Genotyping and Quality Control

We genotyped all available affected individuals within each family, as well as key unaffected individuals, including siblings, children, and spouses of deceased affected individuals; parents of affected siblings; grandparents of affected cousins; and other individuals useful for estimation of phase [31]. Single nucleotide polymorphism (SNP) genotyping was conducted using the Affymetrix 10K 2.0 array (Affymetrix, Santa Clara, CA) for 327 families (1,753 individuals) and the Illumina Infinium Linkage 12 bead array (Illumina, San Diego, CA) for 251 families (1,001 individuals) following manufacturers’ protocols [32,33]. A CEU trio (Coriell Institute for Medical Research, Camden, NJ, USA) was included in each 96-well plate.

Hardy-Weinberg equilibrium (HWE) testing relied on mean p-values from exact testing of 100 random samples of one individual per pedigree. SNPs were excluded with unknown genetic position (n = 147) (build 36.3), call rate <95% (n = 1,076), minor allele frequency (MAF) <1% (n = 377), HWE p-value<0.001 (n = 17), duplicate concordance <95% (n = 10), or Mendelian error in >2% of families (n = 4). We also excluded SNPs to reduce LD (r²>0.10) (n = 4,512) in order to minimize false-positive linkage findings (Table S1) [34]. For 10,091 unique SNPs remaining combined across arrays, genetic maps were created using the Rutgers linkage-physical map v.2 [35] and converted from Kosambi to Haldane distance.

Family Exclusions

We aimed to analyze white families without relationship errors and without evidence of MMR deficiency. Self-reported family structures were confirmed via evaluation of Mendelian inheritance using PREST [36] and Pedcheck [37] based on SNP data. Where probable sample switches or non-paternity were found, family structures were altered (nine sibships changed to half-sibships) or excluded (34 families excluded). We used EIGENSTRAT [38] to estimate ethnicity for individuals with missing self-reported ethnicity, verify ethnic similarity among related individuals, and
MMR proficiency was evaluated using MSI testing, immunohistochemical (IHC) analysis, and LOD scores at known MMR loci. MSI testing of Colon CFR and Newfoundland families was performed on multiple family members using paired normal and tumor DNA isolated from formalin-fixed, paraffin-embedded (FFPE) material [39]. Ten markers were tested (mono-nucleotide markers BAT25, BAT26, BAT34C4, and BAT40; di-nucleotide markers ACTC, D5S346, D10S197, D17S250, and D18S53; and complex repeat MYCL), and four unequivocal results were required. Eighty-nine families with at least one MSH-H tumor were excluded. IHC analysis of Colon CFR and Newfoundland families for MLH1, MSH2, MSH6, and PMS2 expression was performed on FFPE samples, as previously described [39]. IHC staining across all sites was done at three centers, and pathologist interpretation was conducted blind to MSI status. Forty-one families in which at least one tumor showed protein loss were excluded. Finally, we excluded an additional 15 families with dominant LOD scores >0.4 within 20 kb surrounding MSH2, MLH1, MSH6, PMS2, PMS1, MSH3, or MLH3 (linkage methods described below). Thus, 356 white families with no evidence of MMR deficiency were included in the analysis.

**Linkage Analysis**

Multipoint parametric and nonparametric linkage analyses used MERLIN version 1.1.2 [40]; dominant and recessive models were based on a priori segregation analysis (Table S2) [5]. Parametric linkage in the presence of heterogeneity was assessed using heterogeneity LOD (HLOD) scores, and the proportion of families linked to each locus (\( \alpha \)) was estimated using HOMOG [41]. Non-parametric Kong & Cox LOD (NPL) scores from the linear model were computed along with \( S_d \) statistics [42,43]. As has been useful for other cancers [44,45], we sought to improve power by increasing genetic homogeneity using family sub-groups defined a priori based on presumed genetically relevant characteristics. Thus, family groups were based on mean age at diagnosis (<50 years, \( \geq 50 \) years), ascertainment scheme (population-based, clinic-based, or unknown), and number of affected individuals (2, 3, 4 or more). Likelihood ratio testing evaluated heterogeneity of linkage across the independent subsets of each subgroup factor (i.e., age at diagnosis, ascertainment scheme, and number of affected individuals).

**Association Analysis**

In key regions identified by linkage analysis, we also performed association testing among an additional 1,136 cases (343 family history positive and 793 family history negative cases with and without 1st degree relative with CRC, respectively) and 997 controls from population-based collections of the Colon CFR who were genotyped using the Illumina 1M/1M Duo SNP array, as described previously [46]. Logistic regression estimated association between genotype and CRC risk adjusted for age, gender, study site, and four principal components representing ancestry [46]. A quantile-quantile (Q-Q) plot of genome-wide observed versus expected test statistics indicated no evidence of inflation (\( \lambda = 0.938 \)) [46].

**Results**

This collection of white CRC families with no evidence of dMMR consisted of 277 families from the Colon CFR, 48 families from the City of Hope consortium, and 31 families from Newfoundland. A total of 1,612 individuals were successfully genotyped including, on average, five individuals per family (range, 2–10, mean 2.2 affected and 2.8 unaffected individuals). The mean age at diagnosis was 59.7 years (range, 36–79) and 56.2 years (range, 31–74) among population- and clinic-based families, respectively. The majority of families had two affected members (56%) and an older (>50 years) mean age at diagnosis (84%). MSI data were available on 224 families (209 MSS and 15 MSI-L), and IHC data were available on 253 families and showed no evidence of MMR deficiency (Table 1). Both MSI and IHC data were available on 190 families. Sixty-seven families were not tested but had a LOD<0.04 within 20 kb surrounding MSH2, MLH1, MSH6, PMS2, PMS1, MSH3, or MLH3.

Genome-wide linkage scans of nine family groups were conducted including analysis of all families and of subsets of families defined by age, ascertainment scheme, and number of affected individuals. Four regions in five family groups were observed with HLOD scores greater than 3.0 (Figure 1). The strongest result was based on analysis of 58 families with a mean age at diagnosis <50 years. In this group, we observed a dominant HLOD of 4.51 on chromosome 4q21.1 (145.40 cM, NPL = 2.52) with an estimated 84% of families linked (Table 2). The peak occurred at rs10518142 which is in intron 5 of M444 encoding N-acycloethanolamine acid amidase. The linkage region, defined as a 1-HLOD support interval, spanned 16.0 cM (3.7 Mb). This peak was not seen in older mean age at diagnosis families (Figure S2), although significant heterogeneity by mean age at diagnosis was not observed (LRT p = 0.35). Other regions of interest in families defined by age at diagnosis (HLOD>2.0) are provided in Table 3.

The second strongest linkage peak occurred in analysis of all families (N = 356) at 12q24.32 with a maximum dominant HLOD of 3.60 (285.15 cM, NPL = 2.88) and an estimated 48% of families linked (Table 2). The peak SNP, rs952093, resides in intron 1 of TMEM132C encoding transmembrane protein 132C; the equivalent of a 1-HLOD interval defined a 14 cM (1.3 Mb) region. Three suggestive regions in analysis of all families (HLOD>2.0) were seen on chromosomes 4, 15, and 17 (Figure 1; Table 3), including a region near to the 4q21.1 peak seen in younger age at diagnosis families.

Additional linkage peaks with HLODs just over 3.0 were observed on chromosome 15q22.31 (101.40 cM, rs1477798) among 67 families with four or more affected individuals and on 88 clinic-based families (Figure 1). Among families with at least four affected members, a dominant HLOD of 3.07 was observed (\( \alpha = 0.29, \text{NPL} = 1.03 \)) and among clinic-based families a dominant HLOD of 3.03 was seen (\( \alpha = 0.32, \text{NPL} = 1.03 \)). Thirty-five families contributed to both analyses (i.e., clinic-based families with four or more affected individuals) (Table 4); analysis of these revealed a dominant HLOD of 3.15 (\( \alpha = 0.35, \text{NPL} = 1.88 \)). Of note, this region was also suggested by analysis of all families (HLOD = 2.51, \( \alpha = 0.20, \text{Table 3} \)). This peak was not seen in analysis of smaller families, population-based families, or families with unknown ascertainment, although significant heterogeneity by family size or ascertainment scheme was not observed (all LRT p’s>0.10). rs1477798 is in intron of MEGF11 which encodes multiple EGF-like-domains 11.

An additional HLOD over 3.0 was observed in recessive analysis of 200 families with only two affected family members (Figure 1, Table 2). On 8q13.2, a recessive HLOD of 3.02 was seen at rs1319036 (intron in pseudo-gene LOC30129096, \( \alpha = 0.51, \text{NPL} = 0.08 \)). Linkage assuming a recessive mode of inheritance is consistent with an affected sibling pair family structure. This region was not highlighted in analysis of larger families (Table 3), although significant heterogeneity by family size was not observed (LRT p = 0.94). Another region of note is...
Many of the genes upstream and downstream of \textit{NL44} are members of the chemokine family that are clustered in 4q12-21 region. The CXC chemokines modulate tumor behavior by regulation of angiogenesis, activation of a tumor-specific immune response, and direct stimulation of tumor proliferation in an autocrine or paracrine fashion [48].

Among all families, evidence for linkage was seen at 12q24.32 (HLOD = 3.60) with an estimated 48% of families linked to this locus (1-HLOD-support interval of 14 cM [276 cM-290 cM] spanning 1.3 Mb). This region contains four known genes (\textit{TMEM132C}, \textit{SLC15A4}, \textit{GLT1D1}, and \textit{TMEM132D}), four hypothetical genes (\textit{LOC100128554}, \textit{LOC440117}, and \textit{FLJ37505}), and one microRNA (\textit{MIR3612}). The four known genes in this region are conserved in dog, mouse, and chicken and, in some cases, zebrafish and Arabidopsis. One of these transmembrane proteins (\textit{TMEM132D}) is known to be expressed in mature oligodendrocytes [49], but little else is known about either function or pathology, as is also true of \textit{GLT1D1} (glycosyltransferase 1 domain containing 1) in humans. Members of the \textit{SLC15} (solute carrier family 15) family are electrogenic transporters of short-chain peptides into a variety of cells [50]. Evidence for linkage at 15q22.31, with a 1-HLOD-support interval of 38 cM (78 cM–116 cM) spanning 12.9 Mb, was particularly evident among families enrolled at high-risk clinics or with four or more affected individuals (dominant HLOD = 3.15). This is a large gene-rich region and contains many known genes including \textit{MEGF11} and \textit{RAB11A}. Very little is known about \textit{MEGF11} [51]. \textit{RAB11A} is a RAS oncogene family member expressed in tumor cell lines and suggested to be involved in membrane trafficking [52]. Finally, among families with only two affected individuals, the 1-HLOD-support interval of 12 cM (126 cM–138 cM) spans 5 Mb (8q13.2, recessive HLOD = 3.02) and contains mostly pseudogenes. Notably, \textit{SULF1} in this region has been suggested to modulate signaling by heparin-binding growth factors, and downregulation represents a novel mechanism by which cancer cells can enhance growth factor signaling [53].

### Discussion

Results of this genome-wide linkage scan provide strong evidence for four previously-unreported CRC susceptibility loci. Notably, we identified a region at 4q21.1 among families with younger mean age at diagnosis (dominant HLOD = 4.51) and estimated that 84% of these families were linked. The 1-HLOD-support interval of this region, 16 cM (139 cM–155 cM) spanning 8.7 Mb, contains multiple known genes including \textit{NL44}. \textit{NL44} encodes an N-acylethanolamine-hydrolyzing enzyme and is shown to be expressed in variety of human tissues including colon [47].

**Table 1.** Characteristics of 356 White Colorectal Cancer Families with No Evidence of Defective Mismatch Repair, N (%).

<table>
<thead>
<tr>
<th>Ascertainment Method</th>
<th>Population-based N = 189</th>
<th>Clinic-based N = 88</th>
<th>Unknown* N = 79</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Age of Diagnosis</td>
<td>&lt;50 years 24 (13%)</td>
<td>19 (22%)</td>
<td>15 (19%)</td>
</tr>
<tr>
<td>≥50 years 165 (87%)</td>
<td>51 (77%)</td>
<td>64 (81%)</td>
<td></td>
</tr>
<tr>
<td>Number of Affected Individuals</td>
<td>2</td>
<td>115 (61%)</td>
<td>26 (30%)</td>
</tr>
<tr>
<td>3</td>
<td>51 (27%)</td>
<td>27 (31%)</td>
<td></td>
</tr>
<tr>
<td>4–10</td>
<td>23 (12%)</td>
<td>35 (40%)</td>
<td></td>
</tr>
<tr>
<td>MSL*</td>
<td>MSS</td>
<td>119 (63%)</td>
<td>61 (69%)</td>
</tr>
<tr>
<td></td>
<td>MSL-L</td>
<td>9 (5%)</td>
<td>6 (7%)</td>
</tr>
<tr>
<td>IHC Expression of MMR Genes*</td>
<td>No Loss</td>
<td>147 (78%)</td>
<td>77 (88%)</td>
</tr>
<tr>
<td></td>
<td>Unknown*</td>
<td>42 (22%)</td>
<td>11 (13%)</td>
</tr>
</tbody>
</table>

*Ascertainment method not reported.

17q23.2 which revealed a dominant HLOD of 2.91 among all families (143.49 cM,  \( \alpha = 0.37 \)) and dominant HLOD of 2.87 (143.50 cM,  \( \alpha = 0.42 \)) among 298 families with mean age of diagnosis ≥50 years (Table 3; the peak SNP rs1477798 is in intron 4 of \textit{MS2} which encodes musashi homolog 2 (Drosophila). Additional linkage results are provided in Figure S2. A second recessive model linkage peak downstream of 8q13.2 was observed in the same families with two affected members (HLOD = 2.0, \( \alpha = 0.51 \)) on 8q12.2. These two nearby peaks were 11.2 cM (8.1 Mb) apart.

Finally, we analyzed association within the 1-HLOD-support intervals surrounding each linkage peak with HLOD>3.0 using additional Colon CFR cases (N = 1,176) and controls (N = 997). In 4q21.21, which showed evidence of linkage in younger age at diagnosis families, the linkage SNP rs10518142 showed no evidence of association; however, rs12643573, which is 2 cM downstream, showed some evidence of association (OR 1.64, \( p = 5.4 \times 10^{-3} \); family history positive OR 1.82, \( p = 1.0 \times 10^{-8} \)) (Figure S3). At rs1477798 in 15q22.31 which showed evidence of linkage in clinic-based, larger families, a nominally significant case-control association was observed (OR 1.16, \( p = 0.04 \)) which was modestly strengthened for cases with CRC family history (OR 1.24, \( p = 0.03 \); however, no significant difference in risk by family history was observed and associations were far from genome-wide significant. No other associations of note were observed.

---

17q23.2 which revealed a dominant HLOD of 2.91 among all families (143.49 cM, \( \alpha = 0.37 \)) and dominant HLOD of 2.87 (143.50 cM, \( \alpha = 0.42 \)) among 298 families with mean age of diagnosis ≥50 years (Table 3; the peak SNP rs1477798 is in intron 4 of \textit{MS2} which encodes musashi homolog 2 (Drosophila). Additional linkage results are provided in Figure S2. A second recessive model linkage peak downstream of 8q13.2 was observed in the same families with two affected members (HLOD = 2.0, \( \alpha = 0.51 \)) on 8q12.2. These two nearby peaks were 11.2 cM (8.1 Mb) apart.

Finally, we analyzed association within the 1-HLOD-support intervals surrounding each linkage peak with HLOD>3.0 using additional Colon CFR cases (N = 1,176) and controls (N = 997). In 4q21.21, which showed evidence of linkage in younger age at diagnosis families, the linkage SNP rs10518142 showed no evidence of association; however, rs12643573, which is 2 cM downstream, showed some evidence of association (OR 1.64, \( p = 5.4 \times 10^{-3} \); family history positive OR 1.82, \( p = 1.0 \times 10^{-8} \)) (Figure S3). At rs1477798 in 15q22.31 which showed evidence of linkage in clinic-based, larger families, a nominally significant case-control association was observed (OR 1.16, \( p = 0.04 \)) which was modestly strengthened for cases with CRC family history (OR 1.24, \( p = 0.03 \); however, no significant difference in risk by family history was observed and associations were far from genome-wide significant. No other associations of note were observed.
Figure 1. Genome-wide Linkage Scans of White pMMR Family Groups with HLOD Score > 3.0. HLOD scores from genome-wide linkage scan of five white pMMR family subgroups. The blue line represents HLODs under the dominant model and the red line represents the HLODs under the recessive model. Maximum observed HLODs > 3.0 (in parenthesis) are labeled with the nearest SNP in four regions. (A) Family mean age at diagnosis < 50 years.

B. All Families

C. Families with 4 or More Affected Members

D. Clinic-based Families

E. Families with 2 Affected Members
Like all complex diseases, CRC is heterogeneous and most likely due to multiple partially penetrant susceptibility alleles as well as non-genetic factors. In order to maximize power to detect linkage, we sought to increase genetic homogeneity by grouping families with similar, potentially genetically driven features, such as age at diagnosis, clinic-based ascertainment, and number of affected family members [5]. A number of other groups have taken a similar predefined subset approach, reporting evidence of CRC linkage in specific regions among family subsets [54]. Here, linked regions on 4q21.1 and 8q13.2 become apparent only in the families with younger mean age at diagnosis and only two affected members, respectively, and the 15q22.31 peak suggested by analysis of all families strengthened considering clinic-based or large families only. Two observations provide particular reassurance of the use of this subset approach; first, the subsets predicted by segregation analysis to be more likely to be genetic (younger age at diagnosis, clinic-based) showed greater evidence for linkage; and second, the peak among smaller families (sibling pairs) was identified using a recessive model.

Other CRC linkage scans have reported evidence of linkage at 3q21-24 and 9q22.2-31.2 in more than one study (Table S3) [18–20,22–26,54]. Evidence of linkage on 3q was first reported in 12 large families with an HLOD score of 3.10 (NPL = 3.40) [20], followed by an independent study of 30 Swedish families at a 65 cM region flanked by markers D3S1558 and D3S592 on chromosome 3q13.31-27.1 overlapping with the earlier report [24]. Another study that focused on MSS families specifically showed evidence of linkage at this 3q region with an HLOD score of 1.49 [26]. Wiesner et al [18] identified a linkage peak on chromosome 9q22.2-31.2 region (p = 0.00045) in 53 MSS kindreds in which at least two siblings were diagnosed with colon cancer by age 64 or younger. Subsequently, the linkage peak, flanked by markers D9S283 (80 cM) and D9S938 (104 cM), was narrowed to 7.7 cM by three other studies [21,22,25]. In the current study, we detected no linkage in this 9q22.2 region under either dominant or recessive models. None of the other previously published linkage regions (1p31.1, 3q21.3, 7q31.1, 15q14-22 and 17p13.3 [23,34]) showed evidence of linkage with HLOD of 2.0 or higher in our study, although some regions harbored HLODs close to 1.0 (Table S3).

A number of factors about this study are unique among CRC genome-wide linkage-scans. First, ours is the largest study, thus had higher power for detection. Second, our population included only families with no evidence of MMR deficiency. Only two smaller studies focused on pMMR families [18,26]. In this respect, our approach of studying a large number of pMMR families allowed us to identify specific linkage regions for this subgroup of families who are known to differ clinically from dMMR families and do not arise from MMR mutations [55–57]. Unlike some prior studies, we included MSI-L families (N = 15) in our analysis, because the relatedness of this phenotype to dMMR disease is unknown; in all regions, results did not differ when analyses were repeated exclusion of these families. Finally, the two most significant regions reported here showed similar NPL scores in these regions.

Several GWAS have reported highly replicated low-penetrance loci [10–17], including a meta-analysis of ten independent studies (11,067 cases and 12,517 controls) which replicated eight previously-reported associations [46]. In relation to the four linkage regions reported here, the closest reported GWAS association is on chromosome 12q24 (rs7315438) [46] 3 Mb away for our peak HLOD. It is not surprising that GWAS and linkage analyses may identify different loci due to the complementary strengths of each approach and the evidence, for many cancers, that the familial and non-familial forms of the disease do not often show affected pathways in common. This is largely supported by our analysis of association within the linkage regions reported here. In fact, despite the attractiveness of the two-hit hypothesis, colorectal cancer is an important exception to the pattern among adult cancers, rather than the rule: APC is central to a dominant familial syndrome and frequently mutated somatically in the non-familial disease [58]. There is a similar pattern involving the MMR genes: they are mutated in the germline among those with Lynch syndrome, and MLH1, at least, is frequently hypermethylated in the non-familial cancer.

In conclusion, these results suggest novel CRC susceptibility loci on chromosomes 4q21, 8q13, 12q24, and 15q22. Further confirmatory studies are needed, including targeted sequencing and dense mapping of the identified linkage regions. Targeted sequencing of these regions will facilitate identification of novel variants that may be missed with linkage analysis, while fine-mapping studies will narrow the region of interest to be examined. In addition, pooling of linkage data across multiple genome-wide scans should allow for fine-level analysis of discrepant results across

### Table 2. Summary of Colorectal Cancer Linkage Results with HLOD Scores $>3.0$.

<table>
<thead>
<tr>
<th>Family Group</th>
<th>N Families</th>
<th>Region</th>
<th>cM (Nearest SNP*, Mbp)</th>
<th>Nearest Gene (SNP location)</th>
<th>Model</th>
<th>HLOD (α)</th>
<th>NPL*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Age at Diagnosis &lt;50</td>
<td>58</td>
<td>4q21.1</td>
<td>145.40 (rs10518142, 77.1)</td>
<td>NAAA (intron)</td>
<td>Dominant</td>
<td>4.51 (0.84)</td>
<td>2.52</td>
</tr>
<tr>
<td>All Families</td>
<td>356</td>
<td>12q24.32</td>
<td>285.15 (rs952093, 127.4)</td>
<td>TEMEM132C (intron)</td>
<td>Dominant</td>
<td>3.60 (0.48)</td>
<td>2.88</td>
</tr>
<tr>
<td>4+ Affected</td>
<td>67</td>
<td>15q22.31</td>
<td>101.40 (rs1477798, 64.2)</td>
<td>MEGF11 (intron)</td>
<td>Dominant</td>
<td>3.07 (0.29)</td>
<td>1.03</td>
</tr>
<tr>
<td>Clinic-based</td>
<td>88</td>
<td>15q22.31</td>
<td>101.40 (rs1477798, 64.2)</td>
<td>MEGF11 (intron)</td>
<td>Dominant</td>
<td>3.03 (0.32)</td>
<td>0.98</td>
</tr>
<tr>
<td>2 Affected</td>
<td>200</td>
<td>8q13.2</td>
<td>132.52 (rs1319036, 70.3)</td>
<td>LDOC100129096* (intron)</td>
<td>Recessive</td>
<td>3.02 (0.51)</td>
<td>0.08</td>
</tr>
</tbody>
</table>

*Genotyped SNP nearest to the peak of the linkage region; distances are reported in bp based on the NCBI build 36.2 and Haldane cM.

**Pseudo-gene.**

doi:10.1371/journal.pone.0038175.t002
### Table 3. Summary of Colorectal Cancer Linkage Results with Maximum Observed HLOD Scores between 2.0 and 3.0.

<table>
<thead>
<tr>
<th>Family Group</th>
<th>Linkage Region</th>
<th>cM (Nearest SNP*, Mbp)</th>
<th>Nearest Gene</th>
<th>SNP Locationb (distance, bp)</th>
<th>Model</th>
<th>HLOD (α)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Families (N = 356)</td>
<td>4q21.21</td>
<td>154.41 (rs725826, 82.4)</td>
<td>RASGEF1B</td>
<td>3’ downstream (70,563)</td>
<td>Dominant</td>
<td>2.29 (0.29)</td>
</tr>
<tr>
<td></td>
<td>15q22.31</td>
<td>101.40 (rs1477798, 64.2)</td>
<td>MEGF11</td>
<td>intron</td>
<td>Dominant</td>
<td>2.51 (0.20)</td>
</tr>
<tr>
<td></td>
<td>17q23.2</td>
<td>143.49 (rs888115, 53.0)</td>
<td>MSI2</td>
<td>intron</td>
<td>Dominant</td>
<td>2.91 (0.37)</td>
</tr>
<tr>
<td>Mean Age at Diagnosis &lt;50 (N = 58)</td>
<td>12q24.32</td>
<td>277.56 (rs2013160, 126.3)</td>
<td>LOC644489d</td>
<td>3’ downstream (40,332)</td>
<td>Dominant</td>
<td>2.86 (0.86)</td>
</tr>
<tr>
<td></td>
<td>14q32.1</td>
<td>161.58 (rs1956716, 94.0)</td>
<td>SERPINA12</td>
<td>5’ upstream (18,174)</td>
<td>Recessive</td>
<td>2.68 (0.41)</td>
</tr>
<tr>
<td></td>
<td>22q11.21</td>
<td>7.92 (rs4269007, 16.7)</td>
<td>MICAL3</td>
<td>intron</td>
<td>Recessive</td>
<td>2.07 (0.24)</td>
</tr>
<tr>
<td>Mean Age at Diagnosis ≥50 (N = 298)</td>
<td>15q22.33</td>
<td>106.55 (rs1822829, 65.3)</td>
<td>AAGAB</td>
<td>intron</td>
<td>Dominant</td>
<td>2.66 (0.27)</td>
</tr>
<tr>
<td></td>
<td>12q24.32</td>
<td>277.56 (rs2013160, 126.3)</td>
<td>LOC644489d</td>
<td>3’ downstream (40,332)</td>
<td>Dominant</td>
<td>2.86 (0.86)</td>
</tr>
<tr>
<td></td>
<td>17q23.2</td>
<td>143.49 (rs888115, 53.0)</td>
<td>MSI2</td>
<td>intron</td>
<td>Dominant</td>
<td>2.91 (0.37)</td>
</tr>
<tr>
<td>Population-based (N = 189)</td>
<td>7p14.1</td>
<td>89.98 (rs1949880, 39.0)</td>
<td>POU6F2</td>
<td>intron</td>
<td>Recessive</td>
<td>2.39 (0.38)</td>
</tr>
<tr>
<td></td>
<td>17q23.2</td>
<td>143.49 (rs888115, 53.0)</td>
<td>MSI2</td>
<td>intron</td>
<td>Dominant</td>
<td>2.87 (0.42)</td>
</tr>
<tr>
<td>Clinic-based (N = 88)</td>
<td>1q42.3</td>
<td>467.95 (rs1405633, 234.5)</td>
<td>LOC343508</td>
<td>5’ upstream (231,963)</td>
<td>Dominant</td>
<td>2.13 (0.23)</td>
</tr>
<tr>
<td></td>
<td>6p25.3</td>
<td>2.63 (rs1055368, 852.0)</td>
<td>EXOC2</td>
<td>5’ upstream (214,441)</td>
<td>Recessive</td>
<td>2.95 (0.52)</td>
</tr>
<tr>
<td></td>
<td>8q12.2</td>
<td>121.32 (rs1367972, 62.2)</td>
<td>NPM1P6</td>
<td>5 upstream (111,467)</td>
<td>Recessive</td>
<td>2.00 (0.51)</td>
</tr>
<tr>
<td></td>
<td>13q12.1</td>
<td>3.88 (rs264729, 21.1)</td>
<td>LOC10012806d</td>
<td>3’ downstream (4,619)</td>
<td>Recessive</td>
<td>2.35 (0.40)</td>
</tr>
<tr>
<td></td>
<td>15q22.1</td>
<td>106.55 (rs1822829, 65.3)</td>
<td>AAGAB</td>
<td>intron</td>
<td>Dominant</td>
<td>2.66 (0.27)</td>
</tr>
<tr>
<td>Unknown Ascertainmentc (N = 79)</td>
<td>13q13.3</td>
<td>46.30 (rs1170994, 35.5)</td>
<td>DCLK1</td>
<td>intron</td>
<td>Dominant</td>
<td>2.16 (1.00)</td>
</tr>
<tr>
<td>Mean Age at Diagnosis &lt;50</td>
<td>1q21.2</td>
<td>143.50 (rs888115, 53.0)</td>
<td>MSI2</td>
<td>intron</td>
<td>Dominant</td>
<td>2.87 (0.42)</td>
</tr>
<tr>
<td>Population-based (N = 189)</td>
<td>12q24.32</td>
<td>285.15 (rs952093, 127.4)</td>
<td>LOC10013238d</td>
<td>3’ downstream (77,728)</td>
<td>Dominant</td>
<td>2.01 (0.38)</td>
</tr>
<tr>
<td>Clinic-based (N = 88)</td>
<td>1p21.3</td>
<td>217.34 (rs1144305, 94.3)</td>
<td>LOC644489d</td>
<td>3’ downstream (40,332)</td>
<td>Dominant</td>
<td>2.86 (0.86)</td>
</tr>
<tr>
<td>Unknown Ascertainmentc (N = 79)</td>
<td>13q13.3</td>
<td>46.30 (rs1170994, 35.5)</td>
<td>DCLK1</td>
<td>intron</td>
<td>Dominant</td>
<td>2.16 (1.00)</td>
</tr>
</tbody>
</table>

---

*Genotyped SNP nearest to the peak of the linkage region; distances are reported based on the NCBI build 36.2 and Haldane cM.

bSNP location and the distance in bp is given in regards to the nearest gene.

cAscertainment method not reported.

dPseudo-gene.

doi:10.1371/journal.pone.0038175.t003

### Table 4. The overlap in the distribution of family groups.

<table>
<thead>
<tr>
<th>Family Group</th>
<th>Mean Age at Diagnosis &lt;50</th>
<th>Mean Age at Diagnosis ≥50</th>
<th>Population-based</th>
<th>Clinic-based</th>
<th>2 Affected</th>
<th>3 Affected</th>
<th>4+ Affected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Age at Diagnosis &lt;50</td>
<td>63</td>
<td>0</td>
<td>34</td>
<td>20</td>
<td>9</td>
<td>28</td>
<td>20</td>
</tr>
<tr>
<td>Mean Age at Diagnosis ≥50</td>
<td>0</td>
<td>293</td>
<td>166</td>
<td>69</td>
<td>58</td>
<td>161</td>
<td>68</td>
</tr>
<tr>
<td>Population-based</td>
<td>34</td>
<td>166</td>
<td>200</td>
<td>0</td>
<td>0</td>
<td>115</td>
<td>26</td>
</tr>
<tr>
<td>Clinic-based</td>
<td>20</td>
<td>69</td>
<td>0</td>
<td>89</td>
<td>0</td>
<td>51</td>
<td>27</td>
</tr>
<tr>
<td>2 Affected</td>
<td>9</td>
<td>58</td>
<td>0</td>
<td>0</td>
<td>67</td>
<td>23</td>
<td>35</td>
</tr>
<tr>
<td>3 Affected</td>
<td>28</td>
<td>161</td>
<td>115</td>
<td>51</td>
<td>23</td>
<td>189</td>
<td>0</td>
</tr>
<tr>
<td>4+ Affected</td>
<td>20</td>
<td>68</td>
<td>26</td>
<td>27</td>
<td>35</td>
<td>0</td>
<td>88</td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pone.0038175.t004
family collections. It is clear from this work and the work of others that multiple loci are involved in increasing susceptibility to CRC in families and that family-based studies remain critical to the identification and characterization of these loci.

Supporting Information
Figure S1 Ethnicity Estimation using Eigen Analysis. EIGENSTRAT was used to verify ethnic similarity among related individuals from 544 families based on self-report and to estimate ethnicity for individuals with missing ethnicity. The first two principal components are plotted by (A) Self-reported ethnicity and (B) Genetically-inferred ethnicity which shows a circle surrounding the samples analyzed for linkage. (DOCX)

Figure S2 Genome-wide Linkage Scans of White pMMR Family Groups with HLOD<3.0. Genome-wide linkage scans of three white pMMR family groups with HLODS<3.0. The blue line represents HLODS under the dominant model and the red line represents the HLODs under the recessive model. (A) Family mean age at diagnosis ≥50 years (N = 298). (B) Families with 3 affected members (N = 89). (C) Population-based families (N = 189). (D) Families with unknown ascertainment (N = 79). (DOCX)

Figure S3 Regional Association Plot from Population-based Colorectal Cancer Case Control Analysis in 4q21.1. Plot shows the 1-HLOD interval surrounding rs10518142, the peak linkage SNP among families with younger mean age at diagnosis. The x-axis indicates genomic position. The y-axis indicates −log10 association p-values for genotyped SNPs (solid circles) adjusted for age, gender, study site, and four principal components representing ancestry. The most significantly associated SNP is indicated by a purple diamond. Other than rs10518142 which is indicated by a yellow circle, the colored points indicate the strength of LD with the SNP most associated with CRC risk (purple diamond). Also shown are the SNP build 36 coordinates in kilobases (kb) and a subset of the known genes in the region (below x-axis). (DOCX)

Table S1 SNP Exclusions and Number of Analyzed SNPs. (DOCX)

Table S2 Genetic Models Assumed for Parametric Analyses. (DOCX)

Table S3 Comparison to Prior Colorectal Cancer Linkage Studies. (DOCX)

Author Contributions
Conceived and designed the experiments: MSC JMC DCT SNT NML JDP ELG. Performed the experiments: JMC GC FS ZC. Analyzed the data: BLF DJS WRB FS ZC ELG. Contributed reagents/materials/analysis tools: BD RWH LLM TGK HBY SG PAN JLH MAJ AST IW. Wrote the paper: MSC JMC BLF DJS MSD JDP ELG.

References


