Published in final edited form as:
Cancer Epidemiol Biomarkers Prev. 2013 July ; 22(7): 1239–1251. doi:10.1158/1055-9965.EPI-12-1226.

Identification of Novel Variants in Colorectal Cancer Families by High-Throughput Exome Sequencing

Melissa S. DeRycke[1], **Shanaka R. Gunawardena**[1], **Sumit Middha**[1], **Yan W Asmann**[1], **Daniel J. Schaid**[1], **Shannon K. McDonnell**[1], **Shaun M. Riska**[1], **Bruce W Eckloff**[1], **Julie M. Cunningham**[1], **Brooke L. Fridley**[2], **Daniel J. Serie**[1], **William R. Bamlet**[1], **Mine S. Cicek**[1], **Mark A. Jenkins**[3], **David J. Duggan**[4], **Daniel Buchanan**[5], **Mark Clendenning**[5], **Robert W. Haile**[6], **Michael O. Woods**[7], **Steven N. Gallinger**[8], **Graham Casey**[6], **John D. Potter**[9], **Polly A. Newcomb**[9], **Loic Le Marchand**[10], **Noralane M. Lindor**[11], **Stephen N. Thibodeau**[1], and **Ellen L. Goode**[1,*] **for the Colon Cancer Family Registry**

[1]Departments of Health Sciences Research, Biomedical Statistics and Informatics, Laboratory Medicine and Pathology, Medical Genetics, Medical Genomics Technology and Advanced Genomics Technology Center, Mayo Clinic College of Medicine, Rochester, MN, 55905, USA

[2]Department of Biostatistics, University of Kansas Medical Center, Kansas City, KS 66160, USA

[3]Centre for Molecular, Environmental, Genetic and Analytic Epidemiology, University of Melbourne, Victoria 3010, Australia

[4]Translational Genomics Research Institute, Phoenix, AZ, 85004, USA

[5]Cancer and Population Studies Group, Queensland Institute of Medical Research, Queensland, Australia

[6]Department of Preventive Medicine, University of Southern California, Los Angeles, CA, 90033, USA

[7]Discipline of Genetics, Faculty of Medicine, Memorial University of Newfoundland, St. Johns, NL, Canada

[8]Department of Surgery, University of Toronto, Toronto, ON M5G 2C4, Canada

[9]Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, 98109, USA

[10]Department of Epidemiology, University of Hawaii, Honolulu, HI, USA

[11]Department of Health Sciences Research, Mayo Clinic, Scottsdale, AZ 85259, USA

Abstract

Background—Colorectal cancer (CRC) in densely affected families without Lynch Syndrome may be due to mutations in undiscovered genetic loci. Familial linkage analyses have yielded disparate results; the use of exome sequencing in coding regions may identify novel segregating variants.

Methods—We completed exome sequencing on 40 affected cases from 16 multi-case pedigrees to identify novel loci. Variants shared among all sequenced cases within each family were identified and filtered to exclude common variants and single nucleotide variants (SNVs) predicted to be benign.

*Correspondence to, Mayo Clinic College of Medicine, 200 First Street SW, Rochester, MN, 55905, USA, Phone 507/266-7997; Fax 507/266-2478; egoode@mayo.edu.

Conflict of Interest: The authors have no potential conflicts of interest to disclose.

Results—We identified 32 nonsense or splice-site SNVs, 375 missense SNVs, 1,394 synonymous or non-coding SNVs, and 50 indels in the 16 families. Of particular interest are two validated and replicated missense variants in *CENPE* and *KIF23*, which are both located within previously reported CRC linkage regions, on chromosomes 1 and 15, respectively.

Conclusions—Whole-exome sequencing identified DNA variants in multiple genes. Additional sequencing of these genes in additional samples will further elucidate the role of variants in these regions in colorectal cancer susceptibility.

Impact—Exome sequencing of familial CRC cases can identify novel rare variants that may influence disease risk.

Keywords

colorectal cancer; familial and hereditary cancers; exome sequencing; rare variants; family study design

INTRODUCTION

Colorectal cancer (CRC) is the third most common cancer and the third leading cause of cancer death in the United States for both men and women (1). Family history is a consistent risk factor (2); without CRC family history, the lifetime risk for an individual is 5% to 6%, but, 10% to 15% if a first-degree relative has CRC (3-5) and 30% to 100% in familial genetic syndromes (6). Lynch Syndrome represents up to 5% of CRCs and results from germline mutations that affect DNA mismatch repair (MMR) genes *MLH1*, *MSH2*, *MSH6*, and *PMS2*.Tumors from these patients demonstrate a defective MMR (dMMR) phenotype manifested by DNA microsatellite instability (MSI) and absence of MMR protein expression (7, 8).

Beyond the known familial genetic syndromes, linkage studies have implicated several additional regions in CRC susceptibility, including 3q21-24, 4q21, 7q31, 8q13, 8q23, 8q24, 9q22-31, 10p14, 11q23, 12q24, 15q22, and 18q21 (9-21). Genome-wide association studies (GWAS) of CRC have reported evidence of many common risk variants in several genetic regions, including chromosomes 1q41, 3q26, 6p21, 6q25, 8q23, 8q24, 9p24, 10p14, 11q13, 12q13, 12q24, 14q22, 15q13, 16q22, 18q21, 19q13, 20p12, 20q13, and Xp22 (14, 15, 19, 21-30). However, results from linkage studies have not been consistent, and GWAS are not ideal for the identification of rare variants. Hypothesizing that coding regions may harbor rare variants segregating with susceptibility, we sequenced the exomes of 40 affected individuals from 16 CRC families. To our knowledge, this represents the first family-based application of massively-parallel sequencing in this disease (31-36).

MATERIALS AND METHODS

Study Participants

We utilized The Colon Cancer Family Registry (Colon CFR), an NCI–supported consortium established to create an infrastructure for interdisciplinary studies of the genetic and molecular epidemiology of CRC (37-39). Families were enrolled via the Mayo Clinic, Memorial University of Newfoundland, the University of Southern California, or the University of Melbourne (37). Risk-factor data, blood samples, and pathology reports were collected on participants, using standardized core protocols, and germline DNA was isolated from blood. Sixty-six pedigrees were reviewed with ≥2 invasive CRC cases and no evidence of Lynch syndrome, *MUTYH* mutations (37), or familial adenomatous polyposis. Sixteen families were selected based on the presumption of a genetic predisposition to disease due to

1) large numbers of affected relatives and 2) younger ages at diagnosis (Table 1). Forty affected individuals were chosen for sequencing based on genetic relatedness (preferring distant relatives), including three cases per family where possible (Figure S1). All aspects of this work received institutional review board approval under the policies of the Colon CFR.

Library Preparation, Target Capture, and Sequencing

Due to the rapid pace of technological advances during the course of our experiments, library capture and sequencing conditions varied by family (Table 1). Exome capture was completed using Agilent's 36 Mb (n=37 individuals) or 50 Mb (n=3) All Human Exon chip. Libraries were sequenced once on Illumina's GAIIx (n=19), twice on a GAIIx (n=8), once on a HiSeq 2000 (n=11) or once each on a GAIIx and HiSeq 2000 (n=2). All samples were run in a single lane of a flow cell; samples run twice were sequenced in separate flow cells and BAM files from separate runs were merged for analysis.

Libraries were prepared following manufacturers' protocols (Illumina and Agilent). Briefly, 3 μg of genomic DNA was fragmented to 150-200 bp using the Covaris E210 sonicator. The ends were repaired, and an A base was added to the 3′ ends. Paired end DNA adaptors (Illumina) with a single T base overhang at the 3′ end were ligated and the resulting constructs were purified using AMPure SPRI beads from Agencourt. Adapter-modified DNA fragments were enriched by four cycles of PCR using PE 1.0 forward and PE 2.0 reverse primers (Illumina). Concentration and size distributions were determined on an Agilent Bioanalyzer DNA 1000 chip. Whole-exon capture used the protocol for Agilent's SureSelect Human All Exon kit (36 Mb or 50 Mb). Five hundred ng of the prepared library was incubated with whole-exon biotinylated RNA capture baits supplied in the kit for 24 hours at 65°C. Captured DNA:RNA hybrids were recovered using Dynabeads MyOne Streptavidin T1 from Dynal and DNA was eluted from the beads and purified using Ampure XP (Beckman). Purified capture products were amplified using the SureSelect GA PCR primers (Agilent) for 12 cycles. Libraries were loaded onto paired-end flow cells at concentrations of 6 pM to 8 pM (GAIIx) or 4 pM to 5 pM (HiSeq 2000) to generate cluster densities of 250,000-350,000 per tile (GAIIx) or 300,000-500,000 per mm^2 (HiSeq 2000) following Illumina's standard protocol using the Illumina cluster station and paired end cluster kit version 4 (GAIIx) or the Illumina cBot and HiSeq Paired-end cluster kit version 1 (HiSeq 2000).

Illumina GAIIx flow cells were sequenced as 101X2 paired-end indexed reads using SBS sequencing kit version 4 and SCS version 2.5 data collection software; base-calling used Illumina's Pipeline version 1.5.1. Illumina HiSeq 2000 flow cells were sequenced as 101X2 paired-end reads using TruSeq SBS sequencing kit version 1 and HiSeq 2000 data collection version 1.1.37.0; base-calling used Illumina's RTA version 1.7.45.0. Results from samples run in duplicate or triplicate were pooled for analysis.

Bioinformatics

Sequences were analyzed using TREAT (Targeted RE-sequencing and Annotation Tool) for sequence alignment, variant calling, functional prediction, and variant annotation (40). Reads were aligned to the human reference genome using BWA and duplicated read levels were evaluated using SAMtools's rmdup method (41-43). The BWA alignment was improved using the Genome Analysis ToolKit (GATK) (44) local re-alignments. SNVs were called using SNVMix (45), and indels were called by GATK with default parameter settings. A 0.8 SNVMix posterior probability threshold was chosen for filtering based on analysis of a CEU sample sequenced by the 1000 Genomes Project (46). Variants located within the target regions were retained. SIFT (47) and SeattleSeq (http://gvs.gs.washington.edu/SeattleSeqAnnotation131) provided functional annotation. Read depths at each variant

position and the average mapping quality score were generated by curating the BAM pile-up files using SAMtools (42). Potential splice variants were defined as those within two bp of exon-intron boundaries; eSplices were those within coding regions. We excluded reads and variants with poor mapping or quality scores (Qphred score <20 and probability score (0.8) and required a minimum number of high-quality reads supporting alternative alleles (10 for SNVs and three for indels). During read alignment, we identified several reads that aligned to off-target coding regions with high mapping scores and expanded the target region to 80 Mb to include high-quality reads in the Agilent capture definition.

Variant Filtering and Analysis

Shared variants (shared among all sequenced cases within each family) with MAF <0.05 (dbSNP Build 130) were identified and categorized as either a nonsense or splice SNV, missense SNV, other SNV (synonymous variants and non-coding), or a frame-shifting indel variant. We took two analysis approaches based on the following two questions: what novel genes harbor variants that may cause predisposition to CRC and what variants can be found in genes and regions previously associated with CRC? And what variants can be found in genes and regions previously associated with CRC? First, as an agnostic approach, we excluded variants not likely to be disease-causing, based on the following criteria: 1) shared in ≥4 pedigrees (likely representing artifacts or reference sequence annotation errors or newly identified common variants); 2) MAF ≥0.01 in CEU populations (HapMap, 1000Genomes, and Beijing Genome Institute); 3) annotation errors of nonsense and splice-site variants (variants incorrectly identified as nonsense or splice-site were correctly categorized then subjected to the standard exclusion criteria); 4) prediction of pathogenicity (missense and indel variants predicted to be benign by PolyPhen or tolerated by SIFT); 5) indel variants in splice sites that did not alter the splice site. Second, we examined variants in *a priori* candidate genomic regions using less stringent filtering criteria. All variants present that were shared among family members, even those with a low probability of causing disease, were included. These included: 1) 27 known CRC susceptibility genes (*AKT1*, *APC, AXIN1*, *AXIN2, BLM*, *BMPR1A*, *BRCA1, BRCA2, CHEK2*, *GALNT12, MCC, MLH1*, *MLH3*, *MSH2*, *MSH3*, *MSH6*, *MUTYH, MYH11, PMS1*, *PMS2, PTEN*, *SMAD4*, *SMAD7, STK11*, *TGFB1*, *TGFBR2*, and *TP53*); *2*) previously identified linkage regions (3q21-24, 4q21, 7q31, 8q13, 8q23, 8q24, 9q22-31, 10p14, 11q23, 12q24, 15q22, and 18q21), and 3) GWAS regions (1q41, 3q26.2, 8q24, 9p24, 10p14, 11q23, 12q13, 14q22, 16q22, 18q21, 19q13, and 20p12) (9-27). In 15 families, we examined regions with family-specific dominant or recessive LOD scores > 1.3 found in multipoint linkage analysis using MERLIN version 1.1.2 (48) and genotype data from Affymetrix Linkage 2.0 or Illumina Linkage Panel 12 arrays, as described previously (12). Expected sharing was calculated as in described in Feng et al. and compared to actual sharing for each family.(49)

Technical Validation and Replication of Select Variants

Variants prioritized from the whole exome sequencing were validated using Sanger sequencing. Primers were designed using GRCh37/hg19 reference assembly for selected nonsense, splice-site, and missense variants identified in families 1, 2, 3, 8, 9, 12, and 16. Sequencing was performed on 16 individuals within the families that had been WES to validate the variants and on 31 available relatives with DNA to look at segregation of the variant with additional CRC- and polyp-affected and unaffected relatives. Briefly, 25 ng of leukocyte DNA was amplified in a 15 μl PCR containing 7.5 μl of GoTaq master mix (Promega, Madison, WI, USA) and 5 pmoles of each primer (available on request). Reactions were cycled on a Biorad iCycler (Biorad, Hercules, CA, USA) using the following profile: 94°C for 2 minutes, followed by 45 cycles of 94°C for 15 seconds, 60°C for 15 seconds and 72°C for 15 seconds, cycling was finalized at 72°C for 5 minutes. PCR reactions were subsequently cleaned up using Montage PCR96 Cleanup plates (Millipore,

Bedford, MA, USA) according to the manufacturer's guidelines. PCR product (0.5μl) was then used in an 8μl sequencing reaction comprising 0.4μl BigDye Terminator v3.1, 1.4μl 5x reaction buffer and 1.5pmoles of either primer. Reactions were cycled for 96°C for 1 minute, followed by 25 cycles of 96°C for 10 seconds, 50°C for 5 seconds and 60°C for 90 seconds. Prior to running on an ABI3100 genetic analyzer (Applied Biosystems, Foster City, CA, USA), sequencing reactions were cleaned up using Xterminator reagent (Applied Biosystems, Foster City, CA, USA) according to the manufacturer's instructions. Resultant sequences were analyzed using SeqMan software (DNASTAR, Madison, WI, USA). Nonparametric linkage analyses used MERLIN version 1.1.2 and non-parametric Kong & Cox LOD (NPL) scores were computed for validated SNVs. (48)

RESULTS

Comparison of Exome Capture and Sequencing Platforms

We completed germline exome sequencing of forty cases from 16 familial CRC families (Table 1). Cases were selected to be distant relatives to decrease the number of shared, non-susceptibility variants. Sequencing technologies advanced rapidly during our work; both capture and sequencing technologies were updated, providing an opportunity for comparison across platforms. As expected, more variants were identified in samples captured with the 50 Mb chip than the 36 Mb chip. Most variant types were increased modestly, with three notable exceptions: intergenic indels increased by 21.6-fold and indels near the 3′ and 5′ ends of a gene were increased by 15-fold and 4.3-fold, respectively, when using the 50 Mb compared to the 36 Mb chip (Table S1). These increases likely represent the expanded target of the 50 Mb chip; similar increases are expected for future versions targeting more UTR regions. Samples run twice on a GAIIx, showed an approximately two-fold increase in the number of reads and variants identified compared to those run once (Figure S2a, Table 1). Samples sequenced twice on a GAIIx also increased the coverage similar to that of samples run once on a HiSeq 2000 (Figure S2b).

Agnostic Search for Novel Loci Identifies Candidate Genes

As described in the Methods, variant filtering was applied to the full whole-exome sequence dataset. We found that, on average, affected cases within families shared 33 variants (Table 1). There was great disparity in the number of shared variants by family and platform, with as few as four shared variants in Family 7 and up to 70 in Family 12. The majority of shared variants (75.8%) were synonymous or non-coding (intronic or intergenic). Missense variants represented 18.5% of all variants, while indels and nonsense or splice-site variants represented 3.6% and 2.1% of variants, respectively. On average, related cases shared approximately three nonsense or splice-site variants within a family, of which two were private.

Thirty-five nonsense or splice-site variants shared among affected family members were identified, each in a unique gene (Table 2). Although most of these variants were found in only one family, multiple families shared the one variant observed in each of *SHROOM3*, *CDC27*, *ARSD*, *H2BFM*, and *TMC2*.

There were 375 missense variants and 70 indels shared among affected family members after filtering. Three hundred fifty-eight of the missense SNVs (95%) and 62 of the indels (89%) were private, 10 missense SNVs and 8 indels were present in two families, and seven missense SNVs were found in three families (Table 3). Two genes had two variants each (*CTBP2* and *MUC6*); the variants were shared between the same families. In both genes, the variants were <50 bp apart and likely due to an inherited haplotype in the families.

Seventeen genes had >1 missense SNV, including six variants in *CDC27* (Table S2). Private missense and indel variants are shown in Tables S3 and S4, respectively.

Synonymous, intronic, and intergenic variants were the most abundant, with 1,394 shared among affected family members after filtering. Most of these were private (86%), while 191 were detected in ≥2 families. Over half of the variants were in genes without any other variant present (n=837); the remaining 557 variants were found in 152 genes (range 2-31 variants per gene). Summarizing across variant types, 46 genes had at least 4 variants (Table 4).

Pedigree structures of five families suggested recessive inheritance; these families were separately investigated to identify genes with homozygous variant alleles or compound heterozygosity (Table S5). In Family 2, five genes were identified with multiple variants. Three had only non-coding variants, while one (*CTBP2*) harbored two missense variants and a 5′ UTR variant and another gene (*PDE4DIP)* harbored two indels. In Family 3, three genes had multiple variants. One gene contained only non-coding variants; the remaining genes had a missense and a non-coding variant (*PYROXD1*) or a missense variant and an indel (*PTPN9*). In Family 6, three genes harbored multiple variants; however, all variants were non-coding. In Family 11, 19 genes had multiple variants. Fourteen of these genes had only non-coding variants, one had two indels (*HLA-DQA1*), two had missense variants (*DDX12*, *MUC2*), and the remaining two had a combination of variants (*NBPF10*, *ZNF717*). In Family 13, 11 genes harbored multiple variants; variants in ten of the genes were non-coding, while *GGT3P* had one missense and one non-coding variant.

Technical Validation of Select Variants

Thirty-one variants identified in seven families were selected for technical validation and segregation studies by Sanger sequencing. Additional variants in the families were not tested due to the presence of homologous sequences, because the variant had been identified as a common sequencing error, or because the gene was hypervariable. Of the 31 variants tested, 27 were validated in the previously exome sequenced individuals and four were found to be a false positives, including two nonsense variants (*SCN1A* and *SHROOM3*) and two indels (*B3GNT6* and *RBMX*) (Table 5).

Segregation Analysis of Validated Variants

For all variants validated, additional affected and non-affected family members, and family members with polyps were Sanger sequenced to determine co-segregation (Table 6). Only one variant (*PTPN9*) was not replicated in any of the new samples; others were replicated in one to six additional family members. Several of the variants appeared to segregate with affection status, such as *TMC2*, *ADH6*, *CENPE*, *AASDH*, *C6orf170*, *AHSG*, *SF1*, *RPGRIP1*, and *KIF23*. Particularly interesting are the variants in *CENPE* and *KIF2.* Both are very rare; the *KIF23* variant is seen only once in the ESP database of European Americans, while the variant in *CENPE* is not present in any public database. Non-parametric LOD scores were calculated for validated SNVs. The maximum possible NPL score was <2.5 for all families (Table S6). No variants had an observed NPL LOD score >1, possibly due to the few individuals and families with available data for analysis.

Search of the Known Susceptibility Genes and Regions also Identifies *CENPE* and *KIF23*

In addition to the agnostic search for novel loci, we investigated 27 known or suspected high-risk and familial CRC genes and several candidate regions. We required that all affected family members shared the variants as in our previous analyses. However, we did not exclude any variants beyond that, as the genes and regions we were targeting are well documented risk regions and we didn't want to overlook any potential candidate variants.

Our selection of non-MMR families was effective - no shared variants were observed in *MLH1*, *MLH3*, *MSH2*, *MSH3*, *MSH6*, or *PMS1*. Two SNPs in *PMS2* were identified in Families 5, 6, and 9 (Table S7). However, both SNPs were common and expected to be tolerated. *BRCA2* had missense (n=3) and synonymous or intronic (n=3) variants with a MAF between 1 - 5%; five of the SNVs were found only in Family 12, increasing the likelihood that the region containing the variants was inherited as a haplotype block. *MCC* harbored two SNVs, a missense variant resulting in a glycine-to-arginine substitution in Family 13, and a non-coding variant found in five families. *BRCA1* harbored a GLN to ARG substitution (rs1799950) predicted to be damaging; the same rare allele has been associated with a decreased risk of developing breast cancer, however it has not been described in colon cancer previously (50). *APC*, *AXIN2*, *GALNT12*, *MYH11*, and *TP53* each had one non-coding variant. No SNVs or indels were shared among affected family members in the remaining 12 HCC genes (Table S7).

Previously, we reported four regions linked to CRC with HLOD scores greater than 3.0 in 356 families, including 15 of the currently studied families (12). Two regions, 4q21.1 and 15q22.31, harbored variants (Table S8). The 4q21.1 region contained five shared variants, including the *SHROOM3* nonsense variant, which found to be a false positive, and four non-coding variants. The 15q22.31 region contained two missense variants (*CGNL1* and *KIF23)* and five non-coding variants. No variants were found in the other linkage regions examined. Family-specific linkage analysis yielded LOD scores >1.3 in 10 regions in five of the families sequenced (Table S9). None of the regions contained a gene with a shared nonsense or splice variants and six of the regions harbored only non-coding variants. The linkage peak on chromosome 4 harbored 2 missense SNVs in Family 1, one each in *ADH6* and *CENPE*. In Family 2, two linkage peaks on chromosomes 1 and 3 harbored missense variants in *WDR47*, *AHSG*, and *MASP1*.The linkage peaks in Family 5 contained two missense variants (*CFTR* and *ZC3HC1)* and two non-coding variants. In addition, although it is expected that variants responsible for disease in densely affected families differ from modest penetrance variants, we investigated SNVs within the +/-500 Kb regions surrounding SNPs shown to be associated in GWAS with CRC risk (14, 15, 19, 21-27). One indel variant was found in *TPD52L3* within the 9p24 region in Family 5; however, this family does not carry the identified risk allele at rs719725. Twenty-one missense and 45 synonymous or non-coding variants were also identified in the GWAS regions, however, many were common (MAF>5%) and not likely to contribute to CRC genesis (Table S10). Thus, we were able to identify two variants of interest (*CENPE* and *KIF23*) in the regions previously implicated in CRC risk that were also identified by our earlier agnostic search. These two variants were validated and replicated in the affected families, strengthening the results.

DISCUSSION

This analysis of whole-exome sequence data in 16 high-risk CRC families demonstrates the utility of massively-parallel exome sequencing to identify novel candidate genes for complex diseases. We have enumerated potential novel variants as well as those in prior candidate genes and regions. After excluding variants not shared among affected family members, common variants (MAF≥0.01), and those expected to be benign or tolerated, several remained, including protein-truncating mutations in genes involved in: cell shape and motility (*ZRANB1*); mitosis (*CDC27*, *CENPE*, *DDX12*, *HAUS6/FAM29A*, *HIST1H2BE*, *KIF23*, *TACC2*, and *ZC3HC1*); transcription regulation (*CTBP2*, *IRF5*, *MED12*, *RNF111*, *SF1*, *TLE1*, *TLE4*, *TRIP4*); and the immune response (*BTNL2*, *BAGE*, *CARD8*, *FANK1*, *KIR2DL1*, *KIR2DS4*, *KIR3DL3*, *MASP1*, and *NLRP8*) (51-77), as well as numerous missense and indel variants.

It is likely that some of the identified genes are causal. We divided the variants into three categories, based on the likelihood of causing a loss of protein or protein function: the most likely to be causal (nonsense and splice site), those with an elevated risk of being deleterious (missense) and those with the lowest likelihood of being damaging (synonymous and non-coding). Given the lengthy list of candidate genes, the possibility of false positive results, and the paucity of functional information, additional targeted sequencing studies in a large set of independent cases and controls is warranted. Targeted sequencing of novel candidate genes (e.g., those with nonsense or splice site variants) in upwards of 1,000 familial CRC cases would be an informative next step.

This study represents only one point in the journey to identifying genetic predisposition to CRC. CRC is highly heterogeneous and polygenic; unlike the very distinctive Mendelian diseases for which whole-exome sequencing has been successful. Studies directed at identifying candidate susceptibility genes for familial CRC are not readily yielding causal variants (78-81). We debated family selection strategies. Without defined criteria about the optimal selection methods, identification of families and individuals best suited for exome sequencing proved more challenging than expected. We based selection on what we believed would maximize the chance of including families with a high risk genetic predisposition based upon widely held tenets: multiple, closely related affected relatives and younger ages at diagnosis (49, 82).To investigate this further, we compared the observed proportion of shared variants to the expected proportion of shared variants. Significant increases in nonsense, missense, and indel variants (Table S11) strengthened our belief that the methods for choosing families was suitable.

For our families, we chose to sequence two individuals when distantly related cases were available and three individuals when only closely related individuals (siblings) were available, following the recommendations of Feng and colleagues for studying complex diseases by sequencing (49). Model-based approaches, such as estimation of expected lod scores, could also have been used (83). Missing information on earlier generations meant that sequenced samples may be connected through unaffected relatives (in one avuncular pair, additional data became available confirming this to be the case), demonstrating the challenges of incomplete penetrance and phenocopies in studying cancer and complex traits. Sequencing of non-affected family members to help distinguish between causal and benign variants was also discussed; however, the penetrance for CRC in families that met Amsterdam criteria (84), but do not have MMR defects (Type X), is lower than for Lynch Syndrome (85). This makes sequencing unaffected relatives less useful, compared to disorders with complete or very high penetrance.

Our study has weaknesses. First, it involves only a small number of families, chosen to include those with no evidence of Lynch syndrome, *MUTYH* mutations, or Familial Adenomatous Polyposis. Several other candidate families were considered, however, funding was only available for a small number of families. As sequencing costs decline, combining with other collections will be more feasible and needed to identify additional genes of interest. Second, for each family, only a limited number of affected individuals were available for sequencing. The relationships of those selected were preferentially chosen to be cousins or avuncular. However, for several families, the only individuals with DNA available were siblings, reducing power to detect causal variants. Third, we had a false discovery rate of ~13%, which is higher than expected based on previous studies. This may in be due to the fact that rare variants, such as the ones we choose to validate, have a lower rate of validation than more common variants (86), highlighting the critical need for Sanger validation. It is interesting to note that the three false positive variants identified were all in the same family (Family 12), which was the only one utilizing the 50 Mb capture system. Multiple factors may have contributed to the false positives identified in this family,

including degraded DNA for the individuals tested, increased target size, resulting in localized areas of decreased coverage, or misalignment due to poor probe design. Fourth, samples were not all subject to the same capture or sequencing conditions, resulting in increased coverage for samples sequenced toward the study's end. It is possible that some variants detected in later families were present in the earlier families, but went undetected, skewing our perceptions of the allele frequencies. Differences in capture and sequencing technologies also likely affect public databases; numerous variants identified had little available frequency information. Lastly, the most appropriate method to filter for causal variants in complex diseases is unknown. We first narrowed the number of variants by filtering those not shared among the affected family members, which may have excluded causal variants that don't perfectly co-segregate with disease. We used several strategies, including examining candidate genes and regions, looking for genes with multiple variants, and agnostic searching for novel loci. We restricted our search to rare variants, hypothesizing that genes important for the development of CRC will harbor several private variants

In summary, we have completed exome sequencing of 40 familial CRC cases from 16 families and identified and technically validated several candidate CRC variants. Follow-up studies to determine the frequency of variants in many of the identified genes are currently underway. Further sequencing and functional studies will be needed to confirm the identified genes and determine their role in the genesis of CRC.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Financial Support: This work was supported by the National Cancer Institute, National Institutes of Health under RFA # CA-95-011 and through cooperative agreements with members of the Colon Cancer Family Registry and P.I.s. 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 CFRs, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government or the CFR. Collaborating centers include the Australian Colorectal Cancer Family Registry (UO1 CA097735), the Familial Colorectal Neoplasia Collaborative Group (UO1 CA074799), Mayo Clinic Cooperative Family Registry for Colon Cancer Studies (UO1 CA074800), Ontario Registry for Studies of Familial Colorectal Cancer (UO1 CA074783), and University of California, Irvine Informatics Center (UO1 CA078296).

REFERENCES

1. Siegel, R.; Naishadham, D.; Jemal, A. Cancer statistics, 2012. Vol. 62. A Cancer Journal for Clinicians; CA: 2012. p. 10-29.
2. Lynch HT, de la Chapelle A. Hereditary colorectal cancer. New England Journal of Medicine. 2003; 348:919–32. [PubMed: 12621137]
3. Hemminki K, Vaittinen P, Dong C, Easton D. Sibling risks in cancer: clues to recessive or X-linked genes? British Journal of Cancer. 2001; 84:388–91. [PubMed: 11161404]
4. Johns LE, Houlston RS. A systematic review and meta-analysis of familial colorectal cancer risk. American Journal of Gastroenterology. 2001; 96:2992–3003. [PubMed: 11693338]
5. Jenkins MA, Baglietto L, Dite GS, Jolley DJ, Southey MC, Whitty J, et al. After hMSH2 and hMLH1—what next? Analysis of three-generational, population-based, early-onset colorectal cancer families. International Journal of Cancer. 2002; 102:166–71.
6. Rustgi AK. The genetics of hereditary colon cancer. Genes & Development. 2007; 21:2525–38. [PubMed: 17938238]
7. Boland CR, Thibodeau SN, Hamilton SR, Sidransky D, Eshleman JR, Burt RW, et al. A National Cancer Institute Workshop on microsatellite instability for cancer detection and familial

predisposition: Development of international criteria for the determination of microsatellite instability in colorectal cancer. Cancer Research. 1998; 58:5248–57. [PubMed: 9823339]

8. Umar A, Boland CR, Terdiman JP, Syngal S, Chapelle Adl, Rüschoff J, et al. Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch Syndrome) and microsatellite instability. Journal of the National Cancer Institute. 2004; 96:261–8. [PubMed: 14970275]
9. Kemp Z, Carvajal-Carmona L, Spain S, Barclay E, Gorman M, Martin L, et al. Evidence for a colorectal cancer susceptibility locus on chromosome 3q21–q24 from a high-density SNP genome-wide linkage scan. Human Molecular Genetics. 2006; 15:2903–10. [PubMed: 16923799]
10. Picelli S, Vandrovcova J, Jones S, Djureinovic T, Skoglund J, Zhou X-L, et al. Genome-wide linkage scan for colorectal cancer susceptibility genes supports linkage to chromosome 3q. BMC Cancer. 2008; 8:87. [PubMed: 18380902]
11. Middeldorp A, Jagmohan-Changur SC, van der Klift HM, van Puijenbroek M, Houwing-Duistermaat JJ, Webb E, et al. Comprehensive genetic analysis of seven large families with mismatch repair proficient colorectal cancer. Genes, Chromosomes and Cancer. 2010; 49:539–48. [PubMed: 20222047]
12. Cicek MS, Cunningham JM, Fridley BL, Serie DJ, Bamlet WR, Diergaarde B, et al. Colorectal cancer linkage on chromosomes 4q21, 8q13, 12q24, and 15q22. PLoS ONE. 2012; 7:e38175. [PubMed: 22675446]
13. Neklason DW, Kerber RA, Nilson DB, Anton-Culver H, Schwartz AG, Griffin CA, et al. Common familial colorectal cancer linked to chromosome 7q31: A genome-wide analysis. Cancer Research. 2008; 68:8993–7. [PubMed: 18974144]
14. Tomlinson I, Webb E, Carvajal-Carmona L, Broderick P, Howarth K, Pittman AM, et al. A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. Nature Genetics. 2008; 40:623–30. [PubMed: 18372905]
15. Wiesner GL, Daley D, Lewis S, Ticknor C, Platzer P, Lutterbaugh J, et al. A subset of familial colorectal neoplasia kindreds linked to chromosome 9q22.2-31.2. Proceedings of the National Academy of Sciences. 2003; 100:12961–5.
16. Kemp ZE, Carvajal-Carmona LG, Barclay E, Gorman M, Martin L, Wood W, et al. Evidence of linkage to chromosome 9q22.33 in colorectal cancer kindreds from the United Kingdom. Cancer Research. 2006; 66:5003–6. [PubMed: 16707420]
17. Skoglund J, Djureinovic T, Zhou X-L, Vandrovcova J, Renkonen E, Iselius L, et al. Linkage analysis in a large Swedish family supports the presence of a susceptibility locus for adenoma and colorectal cancer on chromosome 9q22.32–31.1. Journal of Medical Genetics. 2006; 43:e07.
18. Gray-McGuire C, Guda K, Adrianto I, Lin CP, Natale L, Potter JD, et al. Confirmation of linkage to and localization of familial colon cancer risk haplotype on chromosome 9q22. Cancer Research. 2010; 70:5409–18. [PubMed: 20551049]
19. Djureinovic T, Skoglund J, Vandrovcova J, Zhou X-L, Kalushkova A, Iselius L, et al. A genome wide linkage analysis in Swedish families with hereditary non-familial adenomatous polyposis/non-hereditary non-polyposis colorectal cancer. Gut. 2006; 55:362–6. [PubMed: 16150854]
20. Pittman AM, Naranjo S, Webb E, Broderick P, Lips EH, van Wezel T, et al. The colorectal cancer risk at 18q21 is caused by a novel variant altering SMAD7 expression. Genome Research. 2009; 19:987–93. [PubMed: 19395656]
21. Tenesa A, Farrington SM, Prendergast JGD, Porteous ME, Walker M, Haq N, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. Nature Genetics. 2008; 40:631–7. [PubMed: 18372901]
22. Houlston RS, Cheadle J, Dobbins SE, Tenesa A, Jones AM, Howarth K, et al. Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. Nature Genetics. 2010; 42:973–7. [PubMed: 20972440]
23. Tomlinson I, Webb E, Carvajal-Carmona L, Broderick P, Kemp Z, Spain S, et al. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. Nature Genetics. 2007; 39:984–8. [PubMed: 17618284]
24. Zanke BW, Greenwood CMT, Rangrej J, Kustra R, Tenesa A, Farrington SM, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. Nature Genetics. 2007; 39:989–94. [PubMed: 17618283]

25. Poynter JN, Figueiredo JC, Conti DV, Kennedy K, Gallinger S, Siegmund KD, et al. Variants on 9p24 and 8q24 are associated with risk of colorectal cancer: Results from the Colon Cancer Family Registry. Cancer Research. 2007; 67:11128–32. [PubMed: 18056436]
26. (COGENT) CCG. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. Nature Genetics. 2008; 40:1426–35. [PubMed: 19011631]
27. Broderick P, Carvajal-Carmona L, Pittman AM, Webb E, Howarth K, Rowan A, et al. A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. Nature Genetics. 2007; 39:1315–7. [PubMed: 17934461]
28. Cui R, Okada Y, Jang SG, Ku JL, Park JG, Kamatani Y, et al. Common variant in 6q26-q27 is associated with distal colon cancer in an Asian population. Gut. 2011; 60:799–805. [PubMed: 21242260]
29. Peters U, Hutter C, Hsu L, Schumacher F, Conti D, Carlson C, et al. Meta-analysis of new genome-wide association studies of colorectal cancer risk. Human Genetics. 2012; 131:217–34. [PubMed: 21761138]
30. Dunlop MG, Dobbins SE, Farrington SM, Jones AM, Palles C, Whiffin N, et al. Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk. Nat Genet. 2012; 44:770–6. [PubMed: 22634755]
31. Zhu X, Feng T, Li Y, Lu Q, Elston RC. Detecting rare variants for complex traits using family and unrelated data. Genetic Epidemiology. 2010; 34:171–87. [PubMed: 19847924]
32. Shi G, Rao DC. Optimum designs for next-generation sequencing to discover rare variants for common complex disease. Genetic Epidemiology. 2011; 35:572–9. [PubMed: 21618604]
33. Ionita-Laza I, Ottman R. Study designs for identification of rare disease variants in complex diseases: The utility of family-based designs. Genetics. 2011; 189:1061–8. [PubMed: 21840850]
34. Zhu Y, Xiong M. Family-based association studies for next-generation sequencing. The American Journal of Human Genetics. 2012; 90:1028–45.
35. Ku C-S, Cooper DN, Wu M, Roukos DH, Pawitan Y, Soong R, et al. Gene discovery in familial cancer syndromes by exome sequencing: prospects for the elucidation of familial colorectal cancer type X. Mod Pathol. 2012; 25:1055–68. [PubMed: 22522846]
36. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012; 487:330–7. [PubMed: 22810696]
37. Newcomb PA, Baron J, Cotterchio M, Gallinger S, Grove J, Haile R, et al. Colon Cancer Family Registry: An international resource for studies of the genetic epidemiology of colon cancer. Cancer Epidemiology Biomarkers & Prevention. 2007; 16:2331–43.
38. Green R, Green J, Buehler S, Robb J, Daftary D, Gallinger S, et al. Very high incidence of familial colorectal cancer in Newfoundland: a comparison with Ontario and 13 other population-based studies. Familial Cancer. 2007; 6:53–62. [PubMed: 17039269]
39. Stuckless S, Parfrey P, Woods M, Cox J, Fitzgerald G, Green J, et al. The phenotypic expression of three MSH2 mutations in large Newfoundland families with Lynch syndrome. Familial Cancer. 2007; 6:1–12. [PubMed: 17039271]
40. Asmann YW, Middha S, Hossain A, Baheti S, Li Y, Chai H-S, et al. TREAT: a bioinformatics tool for variant annotations and visualizations in targeted and exome sequencing data. Bioinformatics. 2012; 28:277–8. [PubMed: 22088845]
41. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics. 2010; 26:589–95. [PubMed: 20080505]
42. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25:2078–9. [PubMed: 19505943]
43. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009; 25:1754–60. [PubMed: 19451168]
44. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research. 2010; 20:1297–303. [PubMed: 20644199]
45. Goya R, Sun MGF, Morin RD, Leung G, Ha G, Wiegand KC, et al. SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. Bioinformatics. 2010; 26:730–6. [PubMed: 20130035]

46. A map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061–73. [PubMed: 20981092]
47. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Research. 2003; 31:e15. [PubMed: 12582260]
48. Abecasis G, Cherny S, Cookson W, Cardon L. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. Nature Genetics. 2002; 30:97–101. [PubMed: 11731797]
49. Feng B-J, Tavtigian SV, Southey MC, Goldgar DE. Design considerations for massively parallel sequencing studies of complex human disease. PLoS ONE. 2011; 6:e23221. [PubMed: 21850262]
50. Dunning AM, Chiano M, Smith NR, Dearden J, Gore M, Oakes S, et al. Common BRCA1 variants and susceptibility to breast and ovarian cancer in the general population. Human Molecular Genetics. 1997; 6:285–9. [PubMed: 9063749]
51. Plageman TF Jr, Zacharias AL, Gage PJ, Lang RA. Shroom3 and a Pitx2-N-cadherin pathway function cooperatively to generate asymmetric cell shape changes during gut morphogenesis. Developmental Biology. 2011; 357:227–34. [PubMed: 21726547]
52. Plageman TF, Chauhan BK, Yang C, Jaudon F, Shang X, Zheng Y, et al. A Trio-RhoA-Shroom3 pathway is required for apical constriction and epithelial invagination. Development. 2011; 138:5177–88. [PubMed: 22031541]
53. Bai S, Herrera-Abreu M, Rohn J, Racine V, Tajadura V, Suryavanshi N, et al. Identification and characterization of a set of conserved and new regulators of cytoskeletal organization, cell morphology and migration. BMC Biology. 2011; 9:54. [PubMed: 21834987]
54. Tugendreich S, Tomkiel J, Earnshaw W, Hieter P. CDC27Hs colocalizes with CDC16Hs to the centrosome and mitotic spindle and is essential for the metaphase to anaphase transition. Cell. 1995; 81:261–8. [PubMed: 7736578]
55. Putkey FR, Cramer T, Morphew MK, Silk AD, Johnson RS, McIntosh JR, et al. Unstable kinetochore-microtubule capture and chromosomal instability following deletion of CENP-E. Developmental Cell. 2002; 3:351–65. [PubMed: 12361599]
56. Parish JL, Rosa J, Wang X, Lahti JM, Doxsey SJ, Androphy EJ. The DNA helicase ChlR1 is required for sister chromatid cohesion in mammalian cells. Journal of Cell Science. 2006; 119:4857–65. [PubMed: 17105772]
57. Zhu H, Coppinger JA, Jang C-Y, Yates JR, Fang G. FAM29A promotes microtubule amplification via recruitment of the NEDD1–γ-tubulin complex to the mitotic spindle. The Journal of Cell Biology. 2008; 183:835–48. [PubMed: 19029337]
58. Marzluff WF, Gongidi P, Woods KR, Jin J, Maltais LJ. The human and mouse replication-dependent histone genes. Genomics. 2002; 80:487–98. [PubMed: 12408966]
59. Liu X, Zhou T, Kuriyama R, Erikson RL. Molecular interactions of Polo-like-kinase 1 with the mitotic kinesin-like protein CHO1/MKLP-1. Journal of Cell Science. 2004; 117:3233–46. [PubMed: 15199097]
60. Dou Z, Ding X, Zereshki A, Zhang Y, Zhang J, Wang F, et al. TTK kinase is essential for the centrosomal localization of TACC2. FEBS Letters. 2004; 572:51–6. [PubMed: 15304323]
61. Bassermann F, von Klitzing C, Münch S, Bai R-Y, Kawaguchi H, Morris SW, et al. NIPA defines an SCF-type mammalian E3 ligase that regulates mitotic entry. Cell. 2005; 122:45–57. [PubMed: 16009132]
62. Razmara M, Srinivasula SM, Wang L, Poyet J-L, Geddes BJ, DiStefano PS, et al. CARD-8 rrotein, a new CARD family member that regulates Caspase-1 activation and apoptosis. Journal of Biological Chemistry. 2002; 277:13952–8. [PubMed: 11821383]
63. Furusawa T, Moribe H, Kondoh H, Higashi Y. Identification of CtBP1 and CtBP2 as corepressors of zinc finger-homeodomain factor δEF1. Molecular and Cellular Biology. 1999; 19:8581–90. [PubMed: 10567582]
64. Shi Y, Sawada J-i, Sui G, Affar EB, Whetstine JR, Lan F, et al. Coordinated histone modifications mediated by a CtBP co-repressor complex. Nature. 2003; 422:735–8. [PubMed: 12700765]
65. Barnes BJ, Kellum MJ, Field AE, Pitha PM. Multiple regulatory domains of IRF-5 control activation, cellular localization, and induction of chemokines that mediate recruitment of T lymphocytes. Molecular and Cellular Biology. 2002; 22:5721–40. [PubMed: 12138184]

66. Philibert RA, Madan A. Role of MED12 in transcription and human behavior. Pharmacogenomics. 2007; 8:909–16. [PubMed: 17716226]

67. Levy L, Howell M, Das D, Harkin S, Episkopou V, Hill CS. Arkadia activates Smad3/Smad4-dependent transcription by triggering signal-induced SnoN degradation. Molecular and Cellular Biology. 2007; 27:6068–83. [PubMed: 17591695]

68. Parker KL, Schimmer BP. Steroidogenic factor 1: A key determinant of endocrine development and function. Endocrine Reviews. 1997; 18:361–77. [PubMed: 9183568]

69. Ali SA, Zaidi SK, Dobson JR, Shakoori AR, Lian JB, Stein JL, et al. Transcriptional corepressor TLE1 functions with Runx2 in epigenetic repression of ribosomal RNA genes. Proceedings of the National Academy of Sciences. 2010; 107:4165–9.

70. Milili M, Gauthier L, Veran J, Mattei M-G, Schiff C. A new Groucho TLE4 protein may regulate the repressive activity of Pax5 in human B lymphocytes. Immunology. 2002; 106:447–55. [PubMed: 12153506]

71. Lee JW, Choi HS, Gyuris J, Brent R, Moore DD. Two classes of proteins dependent on either the presence or absence of thyroid hormone for interaction with the thyroid hormone receptor. Molecular Endocrinology. 1995; 9:243–54. [PubMed: 7776974]

72. Arnett HA, Escobar SS, Gonzalez-Suarez E, Budelsky AL, Steffen LA, Boiani N, et al. BTNL2, a butyrophilin/B7-like molecule, is a negative costimulatory molecule modulated in intestinal inflammation. The Journal of Immunology. 2007; 178:1523–33. [PubMed: 17237401]

73. Boël P, Wildmann C, Sensi ML, Brasseur R, Renauld J-C, Coulie P, et al. BAGE: a new gene encoding an antigen recognized on human melanomas by cytolytic T lymphocytes. Immunity. 1995; 2:167–75. [PubMed: 7895173]

74. Wang H, Song W, Hu T, Zhang N, Miao S, Zong S, et al. Fank1 interacts with Jab1 and regulates cell apoptosis via the AP-1 pathway. Cellular and Molecular Life Sciences. 2011; 68:2129–39. [PubMed: 20978819]

75. Natarajan K, Dimasi N, Wang J, Mariuzza RA, Margulies DH. Structure and function of natural killer cell receptors: Multiple molecular solutions to self, nonself discrimination. Annual Review of Immunology. 2002; 20:853–85.

76. Takahashi K. Mannose-binding lectin and the balance between immune protection and complication. Expert Review of Anti-infective Therapy. 2011; 9:1179–90. [PubMed: 22114968]

77. Tschopp J, Martinon F, Burns K. NALPs: a novel protein family involved in inflammation. Nature Reviews Molecular Cell Biology. 2003; 4:95–104.

78. Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, et al. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat Genet. 2010; 42:790–3. [PubMed: 20711175]

79. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, et al. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet. 2010; 42:30–5. [PubMed: 19915526]

80. Polvi A, Linnankivi T, Kivelä T, Herva R, Keating James P, Mäkitie O, et al. Mutations in CTC1, encoding the CTS Telomere Maintenance Complex Component 1, cause cerebroretinal microangiopathy with calcifications and cysts. The American Journal of Human Genetics. 2012; 90:540–9.

81. Sobreira NLM, Cirulli ET, Avramopoulos D, Wohler E, Oswald GL, Stevens EL, et al. Whole-genome sequencing of a single proband together with linkage analysis identifies a Mendelian disease gene. PLoS Genet. 2010; 6:e1000991. [PubMed: 20577567]

82. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet. 2010; 11:415–25. [PubMed: 20479773]

83. Ploughman LM, Boehnke M. Estimating the power of a proposed linkage study for a complex genetic trait. Am J Hum Genet. 1989; 44:543–51. [PubMed: 2929597]

84. Vasen HFA, Mecklin JP, Khan PM, Lynch HT. The International Collaborative Group on Hereditary Non-Polyposis Colorectal Cancer (ICG-HNPCC). Diseases of the Colon & Rectum. 1991; 34:424–5. [PubMed: 2022152]

85. Lindor NM, Rabe K, G.M. P, Haile R, Casey G, Baron J, et al. Lower cancer incidence in amsterdam-i criteria families without mismatch repair deficiency: Familial colorectal cancer type x. JAMA. 2005; 293:1979–85. [PubMed: 15855431]

86. Marth G, Yu F, Indap A, Garimella K, Gravel S, Leong W, et al. The functional spectrum of low-frequency coding variation. Genome Biology. 2011; 12:R84. [PubMed: 21917140]

Table 1

Family Characteristics, Sequencing Conditions, and Number of Shared Variants Identified

Family	N Affected	Mean Age at Diagnosis (Range)	N Sequenced (Relation)	Library Capture	Sequencing Platform	N Variants (N Private)				
						Nonsense, splice site	Indel	Missense	Other	Total
1	8	48.8 (25-72)	2 (first cousins)	36 Mb	GAIIx	1 (0)	4 (4)	15 (14)	39 (32)	59 (50)
2	4	49.0 (31-68)	3 (siblings)	36 Mb	GAIIx	3 (3)	5 (5)	22 (20)	74 (57)	104 (85)
3	6	64.2 (51-69)	3 (siblings)	36 Mb	GAIIx	1 (0)	2 (2)	16 (16)	47(36)	66 (54)
4	6	58.0 (44-89) *	2 (first cousins once removed)	36 Mb	GAIIx	2 (2)	8 (7)	15 (11)	51 (38)	76 (58)
5	6	61.0 (50-79) *	3 (siblings)	36 Mb	GAIIx	1 (1)	4 (2)	28 (28)	61 (54)	94 (85)
6	3	50.8 (39-71) *	3 (siblings)	36 Mb	GAIIx	3 (3)	3 (3)	17 (15)	52 (45)	75 (66)
7	4	66.0 (48-73)	3 (2 siblings, 1 first cousin)	36 Mb	GAIIx	1 (1)	1 (1)	2 (1)	8 (7)	12 (10)
8	6	64.2 (50-79)	3 (avuncular pair, first cousin)	36 Mb	GAIIx †	1 (0)	2 (1)	8 (6)	42 (22)	53 (29)
9	5	51.0 (40-68) *	3 (2 siblings, 1 first cousin)	36 Mb	GAIIx †	3 (1)	3 (1)	7 (4)	51 (27)	64 (33)
10	5	50.2 (28-66)	2 (avuncular pair)	36 Mb	GAIIx †	1 (0)	8 (6)	37 (33)	173 (115)	219 (154)
11	6	66.3 (54-81) *	2 (first cousins)	36 Mb	HiSeq 2000	4 (2)	6 (4)	30 (26)	120 (74)	160 (106)
12	3	49.7 (42-56)	2 (avuncular pair)	36 Mb	HiSeq 2000	2 (2)	9 (8)	59 (57)	169 (112)	239 (179)
13	5	61.4 (53-72)	2 (first cousins)	36 Mb	HiSeq 2000	2 (0)	4 (3)	19 (16)	125 (79)	150 (98)
14	3	49.7 (34-63)	2 (avuncular pair)	36 Mb	HiSeq 2000	9 (7)	7 (5)	28 (25)	161 (111)	205 (148)
15	5	57.4 (45-67)	2 (avuncular pair)	36 Mb	GAIIx, HiSeq 2000 †	5 (3)	4 (2)	43 (35)	173 (140)	225 (180)
16	5	59.2 (56-67)	3 (siblings)	50 Mb	HiSeq 2000	5 (5)	8 (8)	53 (51)	291 (240)	357 (304)

* Age of diagnosis unknown for one or two individual(s); these individuals were excluded from calculation

† Samples in these families were sequenced twice and the results were pooled for analysis

Table 2

Shared Nonsense and Splice Site SNVs

Chr	Position	rsID	Gene	DNA Change	AA Change	Families
1	85,546,961	-	*WDR63*	G/T	GLU>stop	6
	148,932,885	rs1048214	*LOC645166*	C/T	GLN>stop	5
	152,277,622	-	*FLG*	G/T	SER>stop	9
2	166,904,221	-	*SCN1A*	G/T	TYR>stop	16
3	40,231,748	-	*MYRIP*	A/T	LYS>stop	15
	75,787,516	-	*ZNF717*	C/G	Splice Site	4
4	77,660,829	rs73826426	*SHROOM3*	C/A	TYR>stop	8, 10
6	49,425,475	-	*MUT*	G/A	ARG>stop	6
	121,560,230	-	*C6orf170*	G/A	ARG>stop	2
	168,226,602	-	*C6orf124*	C/T	TRP>stop	7
7	76,751,068	rs71555938	*CCDC146*	G/C	TYR>stop	16
10	123,846,924	-	*TACC2*	C/T	GLN>stop	4
	135,490,903	rs36130162	*DUX1*	C/T	ARG>stop	16
11	1,857,515	-	*SYT8*	G/A	Splice Site	14
12	4,870,307	rs61758971	*GALNT8*	C/T	GLN>stop	14
	109,690,964	-	*ACACB*	G/A	Splice Site	14
13	24,243,246	-	*TNFRSF19*	C/T	ARG>stop	12
14	21,779,981	-	*RPGRIP1*	A/G	Splice Site	16
	58,832,019	rs62621193	*ARID4A*	G/A	Splice Site	14
15	75,562,499	-	*GOLGA6C*	C/T	GLN>stop	14
	78,807,407	-	*AGPHD1*	T/A	TYR>stop	14
16	24,873,990	-	*SLC5A11*	G/A	TRP>stop	2
	84,495,318	rs4782970	*ATP2C2*	A/C	Splice Site	15
17	45,234,277	-	*CDC27*	A/C	Splice Site	13, 14
18	14,513,786	rs8095431	*POTEC*	T/C	Splice Site	4
19	11,943,225	-	*ZNF440*	C/T	ARG>stop	12
	40,195,184	-	*LGALS14*	G/A	Splice Site	16
	43,699,204	-	*PSG4*	C/A	GLU>stop	4
	56,459,551	-	*NLRP8*	C/T	ARG>stop	6
20	2,597,716	-	*TMC2*	A/T	Splice Site	1, 9
	30,226,904	-	*COX4I2*	T/G	Splice Site	14

Chr	Position	rsID	Gene	DNA Change	AA Change	Families
22	44,287,073	-	*PNPLA5*	G/A	GLN>stop	2
	2,832,668	-	*ARSD*	A/G	Stop>GLN	9, 15
X	55,185,663	-	*FAM104B*	T/G	Splice Site	15
	103,294,760	rs2301384	*H2BFM*	C/T	GLN>stop	11, 13, 14

Table 3

Missense and Indel Variants Shared in Multiple Families

Chr	Position	rsID	Gene	DNA Change	AA Change	Families
1	117,142,736	-	*IGSF3*	A/G	ILE>THR	1, 4, 12
	145,293,515	rs12565078	*NBPF10*	A/G	ASN>SER	11, 15
	148,023,040	-	*NBPF14*	G/C	SER>CYS	9, 7, 4
	154,171,908	-	*C1orf189*	TC/-	FS	9, 15
6	31,324,603	rs66519358	*HLA-B*	T/-	FS	9, 10
7	76,619,625	rs2302541	*PMS2P11*	C/T	ARG>CYS	12, 14, 15
	99,434,077	rs61469810	*CYP3A43*	A/-	FS	8, 11
10	118,215,310	-	*PNLIPRP3*	-/A	FS	4, 14
	126,673,560	-	*ZRANB1*	-/A	FS	5, 15
	126,678,112	-	*CTBP2*	T/C	ASN>SER	2, 15
	126,678,148	-	*CTBP2*	G/C	ALA>GLY	2, 15
11	1,017,337	-	*MUC6*	T/C	THR>ALA	4, 16
	1,017,338	-	*MUC6*	C/A	GLN>HIS	4, 16
	5,172,795	-	*OR52A1*	-/C	FS	5, 11
	71,529,890	-	*FAM86C & DEFB108B* *	A/T	ILE>LYS	6, 15
12	9,581,791	rs4763566	*DDX12*	T/C	LYS>GLU	11, 13, 15
14	19,378,312	rs61969158	*OR11H12*	T/G	VAL>GLY	11, 13, 14
16	85,132,883	-	*FAM92B*	A/C	PHE>VAL	6, 10
19	41,622,107	rs11399890	*CYP2F1*	-/C	FS	10, 14
	44,778,796	-	*ZNF233*	T/-	FS	12, 13
	58,385,748	-	*ZNF814*	G/A	ALA>VAL	9, 10
22	18,846,088	rs9605845	*GGT3P & DGCR6* *	A/G	MET>THR	8, 13
X	2,832,715	rs73632953	*ARSD*	T/C	LYS>ARG	9, 15
	55,185,656	rs5003001	*FAM104B*	C/A	ARG>ILE	8, 10, 15
	68,725,640	rs1171942	*FAM155B*	T/C	LEU>PRO	10, 11, 14

* Intergenic variants, identified are the two closest genes

Table 4

Genes with Multiple Variants Shared in Affected Family Members: Number of Variants

Gene(s)	Nonsense and Splice	Missense	Indel	Other	Total
ZNF717	1	-	-	30	31
ANKRD30BL	-	-	-	25	25
FRG1B, NCAPG2	-	-	-	18	18
CDC27	1	6	-	8	15
CTBP2	-	4	-	9	13
KIR2DS4	-	-	-	10	10
ROCK1P1, TTTY23/GYG2P1 *	-	-	-	9	9
HLA-DRB1	-	1	1	6	8
MST1P2, MUC12	-	-	-	8	8
ARSD	1	1	-	5	7
ACHE, BAGE/BAGE4, KCNJ12, MST1P9	-	-	-	7	7
FAM104B	1	2	-	4	7
BCL8, KIR3DL3, LOC642846, MUC3A, NBPF10, RACGAP1P	-	-	-	6	6
AQP7P1, SIGLEC16, CROCCP2, FANK1, KIR2DL,1 KRT16P2/TNFRSF13B, KRTAP5-4*	-	-	-	5	5
C6orf10	-	2	-	3	5
NBPF12	-	3	-	2	5
CFTR	-	3	-	1	4
ADAM6, HLA-DRB5, HLA-DRB6, HSP90AB2P, MED12, NBPF1, NBPF9, PCDHB17, POLA1, RPGR, TBC1D3P2, WASH2P	-	-	-	4	4

* Intergenic variants, identified are the two closest genes

Table 5

Validation of Identified Variants

GRCh37 Chr:Position	Gene Name	dbSNP130	Variant Type	Family	Technical validation of exome sequenced CRC-affected carriers
20:2,597,716	*TMC2*	-	Splice-site	1	2/2
4:100,130,075	*ADH6*	rs149932401	Missense		2/2
4:104,030,143	*CENPE*	-	Missense		2/2
19:48,735,017	*CARD8*	rs146319637	Frameshift		2/2
4:57,204,689	*AASDH*	-	Frameshift		2/2
6:121,560,230	*C6orf170*	-	Nonsense	2	3/3
16:24,873,990	*SLC5A11*	-	Nonsense		3/3
22:44,287,073	*PNPLA5*	-	Nonsense		3/3
3:187,003,786	*MASP1*	-	Missense		3/3
3:186,331,094	*AHSG*	-	Missense		3/3
12:70,088,219	*BEST3*	-	Frameshift		3/3
11:64,543,927	*SF1*	rs34514973	Frameshift		3/3
19:55,327,891	*KIR3DL1*	rs71367103	Upstream	3	3/3
15:75,798,025	*PTPN9*	-	Frameshift		3/3
4:77,660,829	***SHROOM3***	**rs73826426**	**Nonsense**	**8, 9**	**0/3**
13:24,243,246	*TNFRSF19*	-	Nonsense	12	2/2
19:11,943,225	*ZNF440*	-	Nonsense		2/2
19:44,778,796	*ZNF233*	-	Frameshift		2/2
16:336,700	*PDIA2*	rs201624048	Frameshift		2/2
2:196,661,361	*DNAH7*	-	Frameshift		2/2
7:128,587,351	*IRF5*	rs60344245	Deletion		2/2
7:15,601,409	*AGMO*	-	Frameshift		2/2
20:34,215,234	*CPNE1*	rs76294482	Frameshift		2/2
19:40,195,184	*LGALS14*	-	Nonsense	16	3/3
2:166,904,221	***SCN1A*** *	**-**	**Nonsense**		**0/3**
14:21,779,981	*RPGRIP1*	-	Splice-site		3/3
15:69,732,770	*KIF23*	-	Missense		3/3
3:178,960,766	*KCNMB3*	rs143962239	Frameshift		3/3
9:43,844,264	*CNTNAP3B*	-	Frameshift		3/3
11:76751,603	***B3GNT6*** *	**-**	**Frameshift**		**0/3**
X:135,960,146	***RBMX*** *	**-**	**Frameshift**		**0/3**

*Variants in bold were not validated and considered false positives.

Table 6

Replication and Segregation of Validated Variants

GRCh37 Chr:Position	Gene Name	dbSNP130	Variant Type	Family	Segregation		
					Additional CRC-affected carriers	unaffected carriers	Polyp-affected carriers
20:2,597,716	*TMC2*	-	Splice-site		2/3*	0/1	3/4
4:100,130,075	*ADH6*	rs149932401	Missense		3/3*	0/1	3/4
4:104,030,143	*CENPE*	-	Missense	1	3/3*	0/1	3/4†
19:48,735,017	*CARD8*	rs146319637	Frameshift		2/3*	1/1¥	2/4
4:57,204,689	*AASDH*	-	Frameshift		2/3*	0/1	4/4
6:121,560,230	*C6orf170*	-	Nonsense		0/0	0/2	4/7
16:24,873,990	*SLC5A11*	-	Nonsense		1/1*	1/2	5/7
22:44,287,073	*PNPLA5*	-	Nonsense		1/1*	1/2	3/7
3:187,003,786	*MASP1*	-	Missense	2	0/0	1/2	2/7
3:186,331,094	*AHSG*	-	Missense		1/1*	1/2	3/7
12:70,088,219	*BEST3*	-	Frameshift		1/1*	0/2	5/7
11:64,543,927	*SF1*	rs34514973	Frameshift		1/1*	1/2	4/7
19:55,327,891	*KIR3DL1*	rs71367103	Upstream	3	2/2	1/1	n/a
15:75,798,025	*PTPN9*	-	Frameshift		0/2	0/1	n/a
13:24,243,246	*TNFRSF19*	-	Nonsense		1/1	1/4	1/1
19:11,943,225	*ZNF440*	-	Nonsense		1/1	0/4	0/1
19:44,778,796	*ZNF233*	-	Frameshift		1/1†	3/4	1/1
16:336,700	*PDIA2*	rs201624048	Frameshift	12	1/1	1/4	0/1
2:196,661,361	*DNAH7*	-	Frameshift		1/1	1/4	1/1
7:128,587,351	*IRF5*	rs60344245	Deletion		1/1	4/4	1/1
7:15,601,409	*AGMO*	-	Frameshift		1/1	1/4	0/1
20:34,215,234	*CPNE1*	rs76294482	Frameshift		0/1	3/4	0/1
19:40,195,184	*LGALS14*	-	Nonsense		0/1	1/3	n/a
14:21,779,981	*RPGRIP1*	-	Splice-site		1/1	0/3	n/a
15:69,732,770	*KIF23*	-	Missense	16	1/1	0/3	n/a
3:178,960,766	*KCNMB3*	rs143962239	Frameshift		0/1	2/3	n/a
9:43,844,264	*CNTNAP3B*	-	Frameshift		1/1	3/3	n/a

*Includes one obligate carrier

†at least 1 individual is homozygous for the variant

¥has stomach cancer

All individuals with available DNA in each family (excluding the original WES samples) were tested for each variant (Family 1, n=8; Family 2, n=10; Family 3, n=3; Family 12, n=6; Family 16, n=5). Only results of successful sequencing are reported in the table.