IJC
International Journal of Cancer

Using linkage studies combined with whole-exome sequencing to identify novel candidate genes for familial colorectal cancer

Claudio Toma[1,2†], Marcos Díaz-Gay[3†], Sebastià Franch-Expósito[3], Coral Arnau-Collell[3], Bronwyn Overs[1,2], Jenifer Muñoz[3], Laia Bonjoch[3], Yasmin Soares de Lima[3], Teresa Ocaña[3], Miriam Cuatrecasas[4], Antoni Castells[3], Luis Bujanda[5], Francesc Balaguer[3], Joaquín Cubiella[6], Trinidad Caldés[7], Janice M. Fullerton[1,2] and Sergi Castellví-Bel[3]

[1]Neuroscience Research Australia, Sydney, Australia
[2]School of Medical Sciences, University of New South Wales, Sydney, Australia
[3]Gastroenterology Department, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBERehd), Hospital Clínic, University of Barcelona, Barcelona, Spain
[4]Pathology Department, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBERehd) and Tumor Bank-Biobank, Hospital Clínic, Barcelona, Spain
[5]Gastroenterology Department, Hospital Donostia-Instituto Biodonostia, Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBERehd), Basque Country University (UPV/EHU), San Sebastian, Spain
[6]Gastroenterology Department, Complexo Hospitalario Universitario de Ourense, Instituto de Investigación Sanitaria Galicia Sur, Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBERehd), Ourense, Spain
[7]Molecular Oncology Laboratory, Hospital Clinico San Carlos, CIBERONC (Centro de Investigacion Biomedica en Red de Cancer), IdISSC, Madrid, Spain

Colorectal cancer (CRC) is a complex disorder for which the majority of the underlying germline predisposition factors remain still unidentified. Here, we combined whole-exome sequencing (WES) and linkage analysis in families with multiple relatives affected by CRC to identify candidate genes harboring rare variants with potential high-penetrance effects. Forty-seven affected subjects from 18 extended CRC families underwent WES. Genome-wide linkage analysis was performed under linear and exponential models. Suggestive linkage peaks were identified on chromosomes 1q22–q24.2 (maxSNP = rs2134095; LODlinear = 2.38, LODexp = 2.196), 7q31.2–q34 (maxSNP = rs6953296; LODlinear = 2.197, LODexp = 2.149) and 10q21.2–q23.1 (maxSNP = rs1904589; LODlinear = 1.445, LODexp = 2.195). These linkage signals were replicated in 10 independent sets of random markers from each of these regions. To assess the contribution of rare variants predicted to be pathogenic, we performed a family-based segregation test with 89 rare variants predicted to be deleterious from 78 genes under the linkage intervals. This analysis showed significant segregation of rare variants with CRC in 18 genes (weighted *p*-value › 0.0028).

[†]C.T and M.D.G. contributed equally to this work

Additional Supporting Information may be found in the online version of this article.

Key words: colorectal cancer, whole-exome sequencing, linkage analysis, genetic predisposition to disease

Abbreviations: BWA: Burrows–Wheeler Aligner; CRC: colorectal cancer; ExAC: Exome Aggregation Consortium; GATK: Genome Analysis Toolkit; GESE: gene-based segregation test; GWAS: genome-wide association study; IBD: identity-by-descent; IPA: Ingenuity pathway analysis; NGS: next-generation sequencing; NPL: nonparametric linkage; SNP: single nucleotide polymorphism; SNV: single nucleotide variant; WES: whole-exome sequencing

Conflict of interest: The authors declare that they have no conflict of interest.

Grant sponsor: "la Caixa" Foundation; **Grant number:** LCF/BQ/DI18/11660058; **Grant sponsor:** Agència de Gestió d'Ajuts Universitaris i de Recerca; **Grant numbers:** 2018FI_B1_00213, GRPRE 2017SGR21, GRC 2017SGR653; **Grant sponsor:** Australian National Health and Medical Research; **Grant numbers:** 1063960, 1066177; **Grant sponsor:** COST (European Cooperation in Science and Technology); **Grant number:** CA17118; **Grant sponsor:** Departament d'Universitats, Recerca i Societat de la Informació; **Grant sponsor:** Fondo de Investigación Sanitaria/FEDER; **Grant number:** 17/00878; **Grant sponsor:** Fundación Científica de la Asociación Española contra el Cáncer; **Grant number:** GCB13131592CAST; **Grant sponsor:** Generalitat de Catalunya, Salut; **Grant number:** PERIS SLT002/16/00398; **Grant sponsor:** Instituto de Salud Carlos III; **Grant sponsor:** Juan de la Cierva postdoctoral contract; **Grant number:** FJCI-2017-32593
DOI: 10.1002/ijc.32683

History: Received 18 Apr 2019; Accepted 23 Aug 2019; Online 16 Sep 2019

Correspondence to: S. Castellví-Bel, E-mail: sbel@clinic.cat

Protein network analysis and functional evaluation were used to suggest a plausible candidate gene for germline CRC predisposition. Etiologic rare variants implicated in cancer germline predisposition may be identified by combining traditional linkage with WES data. This approach can be used with already available NGS data from families with several sequenced members to further identify candidate genes involved germline predisposition to disease. This approach resulted in one candidate gene associated with increased risk of CRC but needs evidence from further studies.

What's new?

Inherited genetic factors are thought to account for more than one-third of colorectal cancer (CRC) cases. Most predisposing genetic factors, however, remain unidentified. Here, genome-wide linkage analysis using whole-exome sequencing (WES) data was performed in families with marked CRC aggregation. The combined linkage-sequencing approach identified possible linkage peaks on chromosomes 1q22-q24.2, 7q31.2-q34, and 10q21.2-q23.1. Analyses of potentially pathogenic variants revealed significant segregation of rare variants in 18 genes, while functional analyses identified a plausible candidate gene for germline CRC predisposition. The findings underscore the utility of linkage analysis employing WES for the discovery of candidate genes for disease predisposition.

Introduction

Colorectal cancer (CRC), like other complex diseases, is caused by both genetic and environmental factors. Although environmental causes such as smoking and diet are without doubt risk factors for CRC, studies in twins show that 35% of the variability in susceptibility corresponds to inherited genetic factors.[1,2] Approximately 6% of cases present with a strong family aggregation and belong to the well-known forms of hereditary CRC, caused by germinal mutations in *APC*, *MUTYH* or the DNA mismatch repair genes. In addition, 30% of CRC cases show family history outside of these known hereditary CRC genes and are categorized as familial CRC, whereas the remaining 65% are classified as sporadic CRC cases.[3]

In the past decade, several studies attempted to identify new germline genetic risk factors for CRC by using genetic linkage analysis. Indeed, the first studies pointed to loci on chromosomes 9q22 and 3q22 which contained putative susceptibility variants implicated in the disease.[4,5] Additional reports pointed to loci on other chromosomes including 11q, 14q and 22q,[6] 7q31,[7] 10q23,[8] 4q21, 8q13, 12q24 and 15q22,[9] and 4p16.3, 9q31.1, 17p13.2 and Xp22.33.[10] However, previous studies were not able to clearly identify candidate genes that were responsible for those linkage signals. On the other hand, genome-wide association studies have achieved greater success in pinpointing additional germline factors by discovering up to 100 common, low-penetrance genetic variants involved in susceptibility to CRC.[11,12]

Next-generation sequencing (NGS) technologies have facilitated the identification of genes involved in disease predisposition.[13,14] Sequencing directed to genome coding regions (exons) or whole-exome sequencing (WES) has become the most fruitful application of NGS in translational biomedicine.[15] Recently, NGS has discovered germinal mutations in genes that cause hereditary CRC, such as *POLE* and *POLD1* or *NTLH1*.[16,17] Several other studies using WES on familial CRC cohorts have proposed several candidate genes for germline predisposition but have also evidenced that the number of candidate variants remains too high after NGS and should be reduced by other means.[18–21]

Linkage analysis studies performed with common polymorphisms from WES data have been suggested as a cost-effective strategy to simultaneously identify linkage regions and then focus on variant-level gene mapping within these intervals.[22–24] This approach has already been used successfully to identify candidate genes in several diseases,[25,26] but its strength is yet to be determined in complex disorders with genetic heterogeneity.

In the present study, we combined WES and linkage analysis in 18 multiplex or extended families with unaffiliated CRC aggregation. Sequencing data from 47 patients were used to identify rare variants with potential high-penetrance effects from the identified linkage intervals. By doing so, we aim to identify novel candidate genes involved in germline CRC predisposition, adding to the knowledge base for future genetic counseling and prevention protocols.

Materials and Methods

Study participants

Eighteen multiplex and extended families with at least three affected relatives with unaffiliated strong CRC aggregation were selected from a previously described cohort of 38 families (Fig. 1).[19,27,28] Families were selected based on the following criteria: three or more relatives with CRC, two or more consecutive affected generations and at least one CRC diagnosed before the age of 60. The presence of germline alterations in well-known genes related to hereditary CRC syndromes (APC, MUTYH and the DNA MMR genes) were previously discarded for all probands. Tumors from probands were microsatellite stable and negative for *MLH1* methylation. High-risk adenomas were adenomas with villous histology or high-grade dysplasia or ≥10 mm in size). Our study was approved by the institutional

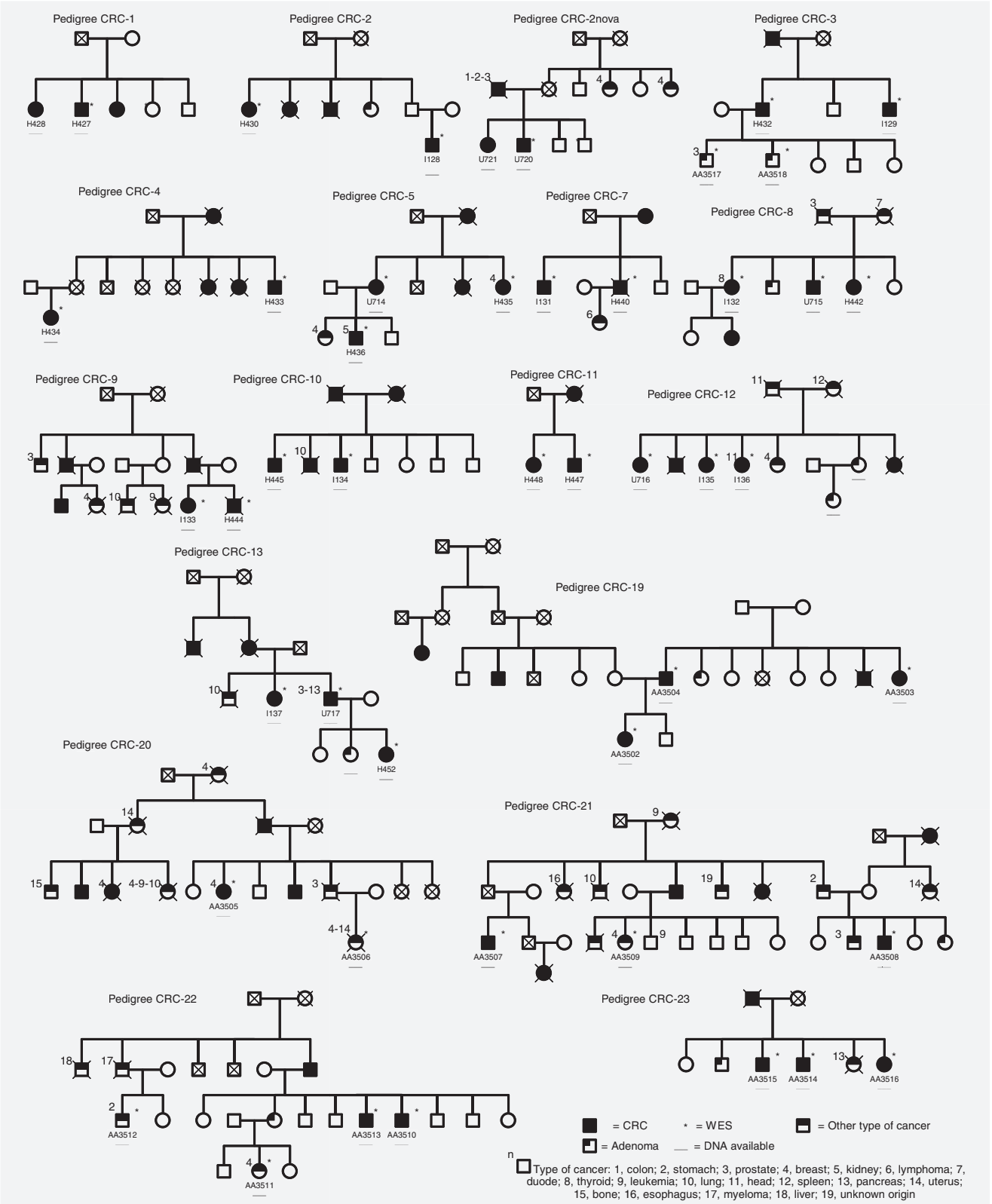

Figure 1. Pedigree structures of the 18 CRC multiplex and extended families examined in our study. Males are indicated with squares, females with circles and diagnosis are shown by dark shading (full, patients diagnosed with colorectal cancer, CRC, left quarter, patients diagnosed with high-risk adenoma; half, patients diagnosed with any other type of cancer detailed in the figure legend; unshaded, unaffected individuals or unknown). Patients analyzed by whole-exome sequencing are indicated by an asterisk, and all subjects with DNA available are underlined.

ethics committee and written informed consent was obtained in all cases.

WES data

The entire cohort had germline WES data available from previous studies.[19,27,28] The group studied herein comprised 47 patients from 18 families, 41 were diagnosed with CRC, two with high-risk colorectal adenoma, and four with other type of cancer. The six selected individuals with high-risk adenoma or other neoplasms were offspring of patients diagnosed with CRC. Briefly, WES was performed using the HiSeq2000 platform (Illumina, San Diego, CA) and SureSelectXT Human All Exon V4 or V5 for exon enrichment (Agilent, Santa Clara, CA). Indexed libraries were pooled and massively parallel sequenced using a paired-end 2 × 75 bp read length protocol. Burrows–Wheeler Aligner (BWA-MEM) was used for read mapping to the human reference genome (build hs37d5, based on NCBI GRCh37).[29] PCR duplicates were discarded using the MarkDuplicates tool from Picard and then indel realignment and base quality score recalibration were performed with the Genome Analysis Toolkit (GATK).[30] The HaplotypeCaller GATK tool was used for variant calling.[30]

Linkage analysis procedures

Genotypes were called from aligned WES reads using SAMtools pileup and filtered to include haplotype-informative markers (HapMap CEU population) using LINKDATAGEN.[31] WES-derived genotypes were used to confirm familial relationships by pair-wise identity-by-descent (IBD) using PLINK,[32] and Z0, Z1 and Z2 values were obtained. All genotyped-derived genetic relationships were consistent with demographic information from the clinical records.

A linkage study was performed using 5,723 WES-derived single-nucleotide polymorphisms (SNPs) from 22 autosomal chromosomes (n = 5,563 SNPs) and the X chromosome (n = 160 SNPs) across all 18 families. Nonparametric linkage (NPL) analyses were performed using the "all" statistic implemented in Merlin,[33] under the Kong and Cox linear (LOD) and exponential (ExLOD) models. All genotyped individuals in our study were affected and, where possible, from the most distant branches of each pedigree. Pedigree structures and diagnoses are detailed in Figure 1. The results of the genome-wide linkage scan under both linear and exponential models were plotted using the "lodplot" R package (https://cran.r-project.org/src/contrib/Archive/lodplot).

Fine mapping of linkage peaks

A LOD score threshold of greater than 2 was considered as suggestive evidence of linkage. A fine-mapping study was performed including additional polymorphic SNPs in flanking regions of a linkage peak to increase allelic informativeness and where inter-marker interval was greater than 1 cM. Twelve additional SNPs with high heterozygosity in Caucasian Europeans (http://www.internationalgenome.org/) were selected in 1q22–q24.2 (rs1002599, rs1509022 and rs2134095), 7q31.2–q34 (rs6968786, rs6651125, rs6953296 and rs447266) and 10q21.2–q23.1 (rs7893379, rs2271698, rs12774070, rs3750736 and rs12263204). The fine-mapping linkage analysis included 513 markers for chromosome 1,303 markers for chromosome 7, and 291 markers for chromosome 10 within 1-LOD-drop interval. The allelic frequencies of markers used to perform the fine-mapping linkage analysis were extracted from a Spanish control population of 629 individuals.[34] After fine mapping, the relative family contribution to overall linkage was computed using the –perFamily option in Merlin. Further examination of the robustness of the three linkage peaks was assessed through a replicative analysis using 10 sets of randomly selected WES-derived markers from chromosome 1 (4,539 SNPs), chromosome 7 (2,117 SNPs) and 10 (2,094 SNPs) that were nonmonomorphic in the HapMap CEU population.

Rare variant selection

Different parameters were considered for variant annotation including population frequency (1000 Genomes, Exome Variant Server, Exome Aggregation Consortium, Collaborative Spanish Variant Server), functional consequences, pathogenicity and position (SnpEff, ANNOVAR, dbNSFP).

Variant filtering was performed with an in-house pipeline written in R language already described in previous studies.[19,27] The parameters taken into account were sequencing quality (coverage ≥10× and genotype quality ≥50), germline allelic frequency (≤0.1% in ExAC database), internal cohort frequency (≤25%) and functional effect (truncating or predicted disrupting missense variants). Missense pathogenicity prediction was assessed with PhyloP (score ≥1.6), SIFT (damaging), PolyPhen2 (probably or possibly damaging), MutationTaster (disease-causing), LRT (deleterious) and CADD (score ≥15). Missense variants predicted to be pathogenic in the least 3 out of 6 predictor tools were selected for further analysis. Variants were also visually inspected with the Integrative Genomics Viewer, and discarded if any sequencing artifact due to strand bias was detected.[35]

Family-based association analysis for rare variants

A total of 91 rare and potentially disruptive variants regardless their segregation status in the families from genes spanning the three detected linkage intervals were included for a family-based association test, using the gene-based segregation test (GESE) package implemented in R (https://cran.r-project.org/web/packages/GESE/).[36] Segregation of rare variants was assessed including all individuals diagnosed with CRC, high-risk adenoma or other type of cancer sequenced in our study. Values of p for statistical significance were calculated after 100,000 simulations. Per-family weights were included in the analysis to assess a relative symptom severity variable based on the number of CRC patients per family, age of onset and the presence of high-risk adenoma and other extracolonic neoplasms.

Selection of novel candidate genes for CRC

The prioritization process was completed with the selection of the putative candidate genes arising from the GESE family-

based association analysis and their interaction partners. This selection process was performed using the Ingenuity pathway analysis (IPA) software program (Ingenuity Systems Inc., Redwood City, CA), which can identify significant networks using a built-in scientific literature-based database.[37] The associated genes with unadjusted p-value < 0.05 from the family-based analysis were used as input for IPA and combined with well-known hereditary CRC genes (*APC*, *MLH1*, *MSH2*, *MSH6*, *PMS2*, *MUTYH*, *BMPR1A*, *BMP4*, *PTPRJ*, *GALNT12*, *EPHB2*, *AXIN2*, *UNC5C*, *GREM1*, *STK11*, *SMAD4*, *PTEN*, *KLLN*, *POLE*, *POLD1*, *BUB3*, *BUB1*, *BUB1B*, *RNF43*, *ATM*, *PALB2*, *SEMA4A*, *RPS20*, *NTHL1*, *FAN1*, *MCM9*, *BLM*, *LRP6*, *SMAD9*, *MSH3*, *EPCAM*, *SETD6* and *BRF1*) plus an additional 115 general cancer predisposition genes.[38] Networks containing any of the GESE genes were considered of interest.

Functional candidate gene presented here was (*i*) expressed in colon tissue, using mRNA expression data from the GTEX dataset (RPKM > 1) and protein expression from the Human Protein Atlas; (*ii*) compatible with cancer predisposition based on reviewing of bibliographic and functional data present in different databases (NCBI, Gene Ontology, KEGG, Reactome). Final candidate variant was confirmed by Sanger sequencing (GATC Biotech, Germany).

Data availability

The data that support the findings of our study are available from the corresponding author upon reasonable request.

Results

Linkage analysis

A genome-wide nonparametric linkage analysis was performed using WES derived-genotype data from 18 multiplex and extended families with unaffiliated strong CRC aggregation (Fig. 1). The highest peak LOD scores were identified on three loci: chromosome 1q22–q24.2 (with LOD_{linear} = 2.11 and LOD_{exp} = 1.872 at rs10753668 or 180.122 cM); chromosome 7q31.2–q34 (with LOD_{linear} = 2.023 and LOD_{exp} = 1.838 at rs2075371 or 142.05 cM); and chromosome 10q21.2–q23.1 (with LOD_{linear} = 1.423 and LOD_{exp} = 2.118 at rs1904589 or 94.855 cM; Fig. 2). When additional markers were added to fine-map each region and reduce intermarker intervals, evidence for linkage at the 1q22–q24.2 locus increased to LOD_{linear} = 2.383 (p-value = 4.6E−04) and LOD_{exp} = 2.196 (p-value = 7.3E−04) with rs2134095 being the peak marker (Fig. 3*a*, Supporting Information Table S1). Likewise, evidence for linkage at the 7q31.2–q34 locus increased to LOD_{linear} = 2.197 (p-value = 7.3E−04) and LOD_{exp} = 2.149 (p-value = 8.2E−04) with rs6953296 being the

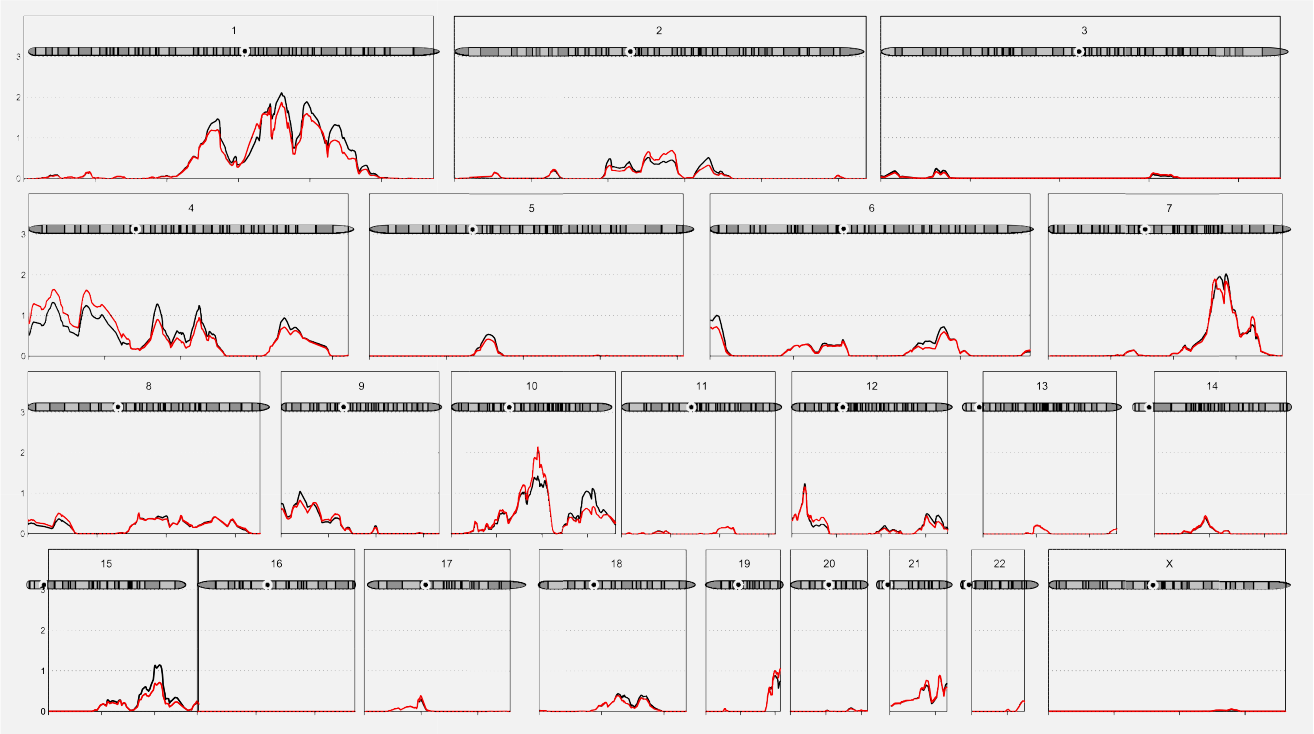

Figure 2. Results of the genome-wide linkage analysis. Nonparametric linkage analysis was performed under the linear (black line) and exponential (red line) models in 18 multiplex/extended CRC families. Each chromosome is represented in a separate plot, including the X chromosome. A linkage signal with LOD › 2 was observed at chromosomes 1q22–q24.2 with a maximum linear LOD score at marker rs10753668 of 2.11 (linear model), 7q31.2–q34 with a maximum linear LOD score at marker rs2075371 of 2.023 (linear model) and 10q21.2–q23.1 with a maximum linear LOD score at marker rs1904589 of 2.118 (exponential model). Additional markers were subsequently added to fine-map these linkage peaks.

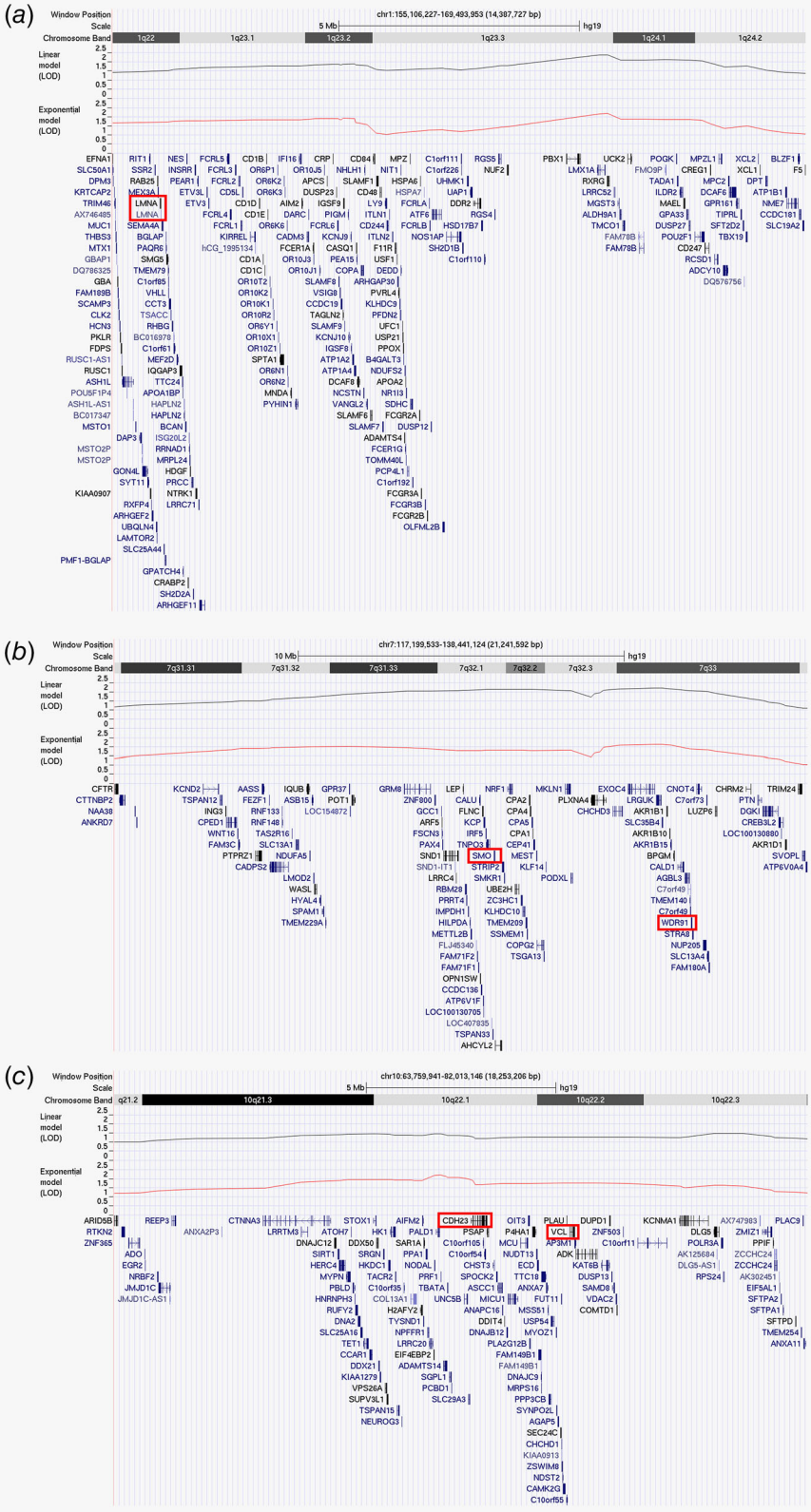

Figure 3. Schematic of the linkage intervals and the gene content between the proximal and distal boundaries on chromosome 1q22–q24.2 (*a*), 7q31.2–q34 (*b*) and 10q21.2–q23.1 (*c*) after fine mapping with 12 additional SNPs. The maximum LOD score under linear and exponential models are shown at each locus. The locations of known protein-coding genes in the linkage interval are provided in the images below which are generated using the UCSC genome browser (https://genome.ucsc.edu). Final candidate genes for CRC, after gene network analysis, colon gene expression evaluation and sequence quality are highlighted using a red box.

Table 1. Family-based association test of rare variants under the linkage peaks

Gene	SNVs/ Seg-SNV	Family (patients with Seg-SNV)	*p*-value	Weighted *p*-value	Specific LOD at locus
NDST2	1/1	CRC-23 (3)	1.28E−08	2.00E−07	0.49
FAM78B	1/1	CRC-20 (2)	1.28E−08	2.00E−07	0.20
GRM8	1/1	CRC-1 (2)	2.56E−08	1.00E−07	0.29
SYNPO2L	1/1	CRC-11 (2)	2.56E−08	4.00E−07	0.30
COL13A1	1/1	CRC-11 (2)	2.56E−08	4.00E−07	0.30
ZSWIM8	1/1	CRC-9 (2)	1.99E−06	5.00E−06	0.47
MYPN	1/1	CRC-19 (3)	3.94E−06	4.90E−05	0.14
LMNA	1/1	CRC-9 (2)	2.18E−05	2.60E−05	0.47
WDR91	1/1	CRC-10 (2)	3.16E−05	6.00E−05	0.27
LY9	1/1	CRC-23 (3)	3.55E−05	3.90E−04	0.29
INSRR	1/1	CRC-7 (2)	4.74E−05	4.20E−04	0.29
TMEM79	1/1	CRC-5 (3)	1.26E−04	4.50E−04	−0.003
SMO	1/1	CRC-13 (3)	1.77E−04	6.90E−04	0.14
C1orf85	1/1	CRC-4 (2)	3.06E−04	3.08E−03	0.20
CDH23	4/1	CRC-8 (3)	2.97E−04	3.00E−03	0.59
TNPO3	1/1	CRC-11 (2)	4.71E−04	4.74E−03	0.30
CFTR	1/1	CRC-20 (2)	7.52E−04	7.47E−03	0.19
VCL	2/1	CRC-10 (2)	9.15E−04	2.88E−03	0.30

Results of 89 rare variants (SNVs) segregating with patients (Seg-SNV) across the 18 CRC families (n = 47 patients), after simulations and weight corrections. Only significant genes (p-value < 0.05) are reported.
Abbreviations: SNVs, number of SNVs regardless segregation in CRC patients; Seg-SNV, number of SNVs segregating in all CRC patients in this family. Specific LOD, linkage contribution from this gene to a specific peak.

peak marker (Fig. 3*b*, Supporting Information Table S2). Finally, evidence for linkage remained stable at the 10q21.2–q23.1 locus (LOD_{linear} = 1.455, p-value = 0.005; LOD_{exp} = 2.195, p-value = 7.3E−03) with rs1904589 remaining as the peak marker (Fig. 3*c*, Supporting Information Table S3).

We examined the robustness of the observed linkage signals using a replicative analysis, performing linkage analysis in 10 replicate SNP sets using random markers from chromosome 1, 7 and 10 (Supporting Information Figs. S1–S3). The linkage peaks previously identified were replicated in all of the 10 data sets with a LOD > 2 under either the linear or the exponential model, or both. These results suggest that the linkage to these regions is not being driven by a particular set of SNPs and is robust to SNP selection. A formal permutation analysis to exclude false-positive signals and to determine empirical significance was not possible, as all subjects were affected and permuting the subjects' phenotypes would be uninformative.

The CRC linkage intervals, as defined by a 1-LOD drop interval, spanned a genetic distance of 20.955 cM (1q22–q24.2), 18.339 cM (7q31.2–q34) and 20.214 cM (10q21.2–q23.1). Per-family linkage analysis showed locus heterogeneity, whereby the number of CRC families contributing positively to the overall LOD score at 1q22–q24.2, 7q31.2–q34 and 10q21.2–q23.1, were 13, 14 and 12 families, respectively (Supporting Information Table S4), and seven families contributed to signals at all three peaks.

Family-based association analysis for rare variants under specific linkage intervals

Next, we explored the possibility that rare alleles with higher penetrance effects, explained the linkage peaks at each locus. We extracted 530 single nucleotide variants (SNVs) from WES data within 271 protein-coding genes under the linkage peak intervals (Fig. 3). After quality control filters and more stringent selection criteria, 89 SNVs from 78 genes were selected for segregation analysis (Supporting Information Table S5).

Then, we performed a family-based segregation test of the 89 rare variants using the GESE package. Allele-frequency weighted segregation analysis revealed significant rare variant segregation with CRC in 18 genes (Table 1).

Novel candidate genes for CRC

Genes with evidence of familial segregation for rare pathogenic variants in CRC patients were further inspected using a protein–protein network analysis (IPA) to identify the most plausible candidate genes from the linkage peaks involved in hereditary CRC. We pooled together the 18 genes with segregating variants from

Table 2. Candidate genes within regions with positive linkage on chromosomes 1, 7 and 10 after considering hereditary cancer networks

Gene	Variant	Family	Chromosomal region	Colon gene expression (RPKM)	IGV	Gene function/OMIM
LMNA	c.1718C>T (p.Ser573Leu)	CRC-9	1q22–q24.2	52	+	Muscular dystrophies
SMO	c.1921C>G (p.Pro641Ala)	CRC-13	7q31.2–q34	5.4	+	Familial or sporadic basal cell carcinoma/Curry–Jones syndrome
WDR91	c.699G>A (Asp239Tyr)	CRC-10	7q31.2–q34	4.6	+	Neuronal development
CDH23	c.4885A>C (p.Ile1629Leu)	CRC-8	10q21.2–q23.1	0.6	+	Deafness
VCL	c.590C>T (p.Thr197Ile)	CRC-10	10q21.2–q23.1	101	+	Cardiomyopathy

Information about the identified genetic variant, the CRC family, gene expression level in colon, sequence quality, gene function and previous involvement in hereditary conditions are listed.
Abbreviations: CRC, colorectal cancer; IGV, integrative genomics viewer: + validated, − not validated; OMIM, online Mendelian inheritance in man, www.omim.org; RPKM, reads per kilobase per million mapped reads (from the Human Protein Atlas, GTEx dataset-colon); Curry-Jones syndrome, craniofacial malformations, polysyndactyly, abnormal skin and gut development.

the GESE analysis with established genes for hereditary CRC (38 genes) and germline predisposition to cancer (115 genes) to investigate specific CRC networks. Six networks were produced and two of them contained GESE genes. Potential CRC candidate genes were prioritized as functionally plausible when a network contained them (Supporting Information Fig. S4). This procedure resulted in the retention of five candidate genes listed in Table 2 to which further exclusion criteria were applied. Candidate genes with negligible expression in colon tissue were excluded (*CDH23*). Only genes with a function compatible with cancer predisposition were retained. Two genes were disregarded due to their involvement in the predisposition to diseases not related to cancer (*LMNA*, muscular dystrophy; *VCL*, cardiomyopathy) or for a gene function that was not tumor-related (*WDR91*, neuronal development). Accordingly, *SMO* remained as a plausible candidate to be involved in germline predisposition to familial CRC although evidence from further studies is needed.

Discussion

Inherited variants are considered to be the underlying cause in an important number of CRC cases,[1,2] with familial aggregation estimated to be present in up to 35% of CRC patients. During the past 40 years, several approaches have been used to identify genetic factors causing this hereditary predisposition. In Mendelian disorders, linkage analysis has been the only approach used for decades and has reported numerous examples of gene discovery.[39] The use of large families in these linkage studies permitted the identification of the main hereditary genes (*APC* and the DNA mismatch repair genes). Linkage approaches have also been applied using nonparametric models in complex disorders, and CRC linkage studies have identified several susceptibility loci at chromosomes 9q22–q31.2,[4] 3q22,[5] 11q, 14q, 22q,[6] 3q29, 4q31.3, 7q31,[7] 10q23,[8] 4q21, 8q13, 12q24, 15q22,[9] 4p16.3, 9q31.1, 17p13.2 and Xp22.33.[10] However, apart from previously commented successful examples, linkage studies in CRC have rarely converged on the same top results, and more generally association studies in linkage regions have failed to identify common variants implicated in the disorder.[40] Subsequently, linkage studies have been superseded by genome-wide association studies (GWAS), which have been able to identify low-risk genetic variants.[11] However, variants identified by GWAS only explain approximately 10% of the variance in genetic liability to CRC,[12] suggesting that rare variants with higher penetrance effects may play a substantial role in disease, particularly in familial forms.

The discovery of high-penetrance germline variants in CRC genes is feasible using NGS technologies. Recently, WES studies performed in approximately 2,000 familial or early-onset CRC cases suggested candidate genes harboring potential pathogenic rare variants.[18–21,41] Despite the encouraging results, the number of identified causative genes has remained limited and poorly replicated. In a previous study, we performed WES in our unaffiliated familial CRC cohort of 71 patients from 38 families that led to the identification of potential candidate genes including *CDKN1B*, *XRCC4*, *EPHX1*, *NFKBIZ*, *SMARCA4* and *BARD1*[19] and a Fanconi anemia pathway enrichment.[27] WES data have also been used to infer rare copy number variants that could act as the germline mutational event in some families.[28]

After more than a decade, linkage analysis has re-emerged as a successful approach to uncover genes implicated in Mendelian diseases when used in combination with NGS.[26,42–44] This approach has also been recently applied in complex disorders where susceptibility loci are examined for rare variants with higher penetrance effects that are expected to segregate among patients in large families. This combined strategy has been employed in large individual families or few combined families in some complex disorders,[24,25,45,46] but has never been applied in multiplex or extended families with cancer. Our study is the first to employ linkage and WES data in cancer families by examining 47 patients from 18 extended CRC families.

We identified three suggestive linkage peaks on chromosomes 1q22–q24.2, 7q31.2–q34 and 10q21.2–q23.1. Two of our CRC susceptibility loci overlap with previously identified linkage regions at 7q31 and 10q23.[7,8] Neklason *et al.* performed a genome-wide scan in 151 DNA samples from 70 families which implicated chromosome 7q31, whereas Nieminen *et al.* performed a linkage scan in a large Finnish CRC type X family that yielded a suggestive signal on 10q23, and *BMPR1A* was suggested as causative gene. Our findings provide additional evidence in support of these previous results implicating 7q31.2–q34 and 10q21.2–q23.1 in the germline predisposition to CRC.

Segregation analysis for potentially pathogenic rare variants inherited by CRC patients followed by a protein network analysis identified *SMO* as the most relevant candidate for germline CRC predisposition, which lies in the center of the 7q31.2–q34 linkage peak. The SMO protein is a G-coupled receptor that interacts with the patched protein, a receptor for hedgehog proteins. Alterations in the *SMO* gene have been related to the familial or sporadic forms of basal cell carcinoma,[47] and Curry–Jones syndrome, a multisystem disorder characterized among other symptoms by skin lesions, polysyndactyly, brain malformations and intestinal malrotation with myofibromas or hamartomas.[48] It is possible that finding a rare variant in SMO is due to chance, given the limited number of available affected subjects in the family with the segregating variant and the absence of segregating SMO variants in other families showing linkage to 7q31.2–q34, although this may also indicate locus heterogeneity.

The interesting findings presented in our study must be considered in light of the limitations that were present. First, we did not account for the potential contribution of common variants associated with CRC. However, it could be argued that common variants may have a reduced impact on multiplex families with segregating illness. Second, while rare variants from noncoding regions are not covered in WES studies (and are potentially more difficult to ascribe functional relevance than those observed in protein-coding regions), they may explain additional contribution to the observed linkage intervals. Third, predictions of pathogenicity from variants at untranslated regions are not as reliable as predictions based

on missense variants, so while we considered only predicted pathogenic rare variants from coding regions, we could not exclude etiologic variants from those regions not examined here. Finally, while we considered multiplex families with high rates of cancer, most families had only 2–3 patients with DNA available and many of those were close relatives, so the power to detect significant linkage was low. Furthermore, the power of segregation analysis of individual rare variants in small families is limited, and presence of phenocopies within a family would impact the interpretation of apparent non-segregation of potentially pathogenic variants, thus rare variants in other genes within the 7q31.2–q34 linkage interval should not be discounted. The identification and analysis of more distally related affected relatives from these and other families may yield more information on these and other risk loci for CRC. While we considered only rare protein-coding variants predicted to be pathogenic with perfect segregation with CRC in these multiplex/extended families in defining the most likely candidate gene, we cannot exclude the contribution of pathogenic variants that partly segregate with the phenotype, given the allelic heterogeneity observed in CRC.

In summary, we performed a genome-wide linkage analysis using WES-derived genotype data from 18 multiplex and extended families with unaffiliated strong CRC aggregation and found suggestive risk loci on chromosomes 1q22–q24.2, 7q31.2–q34 and 10q21.2–q23.1. Rare variant segregation analysis and protein network analyses identified *SMO* as a plausible candidate for germline CRC predisposition. Replication in additional cohorts (including targeted sequencing of large numbers of families with hereditary CRC) and further functional studies are required to confirm this novel potential candidate for CRC germline predisposition. The present approach can be used with already available NGS data from families with several sequenced members to further identify candidate genes involved germline predisposition to the disease.

Acknowledgements

We are sincerely grateful to the patients, CNAG, the Biobank of Hospital Clínic–IDIBAPS and Biobanco Vasco para la Investigación/O + ehun-Hospital Donostia. M.D.-G. was supported by a contract from Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR; Generalitat de Catalunya, 2018FI_B1_00213). S.F.-E., C.A.-C., J.M. and J.J.L. were supported by a contract from CIBEREHD. LB was supported by a Juan de la Cierva postdoctoral contract (FJCI-2017-32593). YSL was supported by a fellowship (LCF/BQ/DI18/11660058) from "la Caixa" Foundation (ID 100010434) funded EU Horizon 2020 programme (Marie Skłodowska-Curie grant agreement no. 713673). CIBEREHD and CIBERONC are funded by the Instituto de Salud Carlos III. C.T., B.J.O. and J.M.F. were supported by Australian National Health and Medical Research (NHMRC) Project Grants 1063960 and 1066177. This research was supported by grants from Fondo de Investigación Sanitaria/FEDER (17/00878), Fundación Científica de la Asociación Española contra el Cáncer (GCB13131592CAST), PERIS (SLT002/16/00398, Generalitat de Catalunya), CERCA Programme (Generalitat de Catalunya) and Agència de Gestió d'Ajuts Universitaris i de Recerca (Generalitat de Catalunya, GRPRE 2017SGR21, GRC 2017SGR653). This article is based upon work from COST Action CA17118, supported by COST (European Cooperation in Science and Technology). www.cost.eu. The work was carried out (in part) at the Esther Koplowitz Centre, Barcelona.

References

1. Lichtenstein P, Holm NV, Verkasalo PK, et al. Environmental and heritable factors in the causation of cancer—analyses of cohorts of twins from Sweden, Denmark, and Finland. *N Engl J Med* 2000;343(2):78–85.
2. Hemminki K, Chen B. Familial risk for colorectal cancers are mainly due to heritable causes. *Cancer Epidemiol Biomarkers Prev* 2004;13(7):1253–6.
3. Ma H, Brosens LAA, Offerhaus GJA, et al. Pathology and genetics of hereditary colorectal cancer. *Pathology* 2018;50(1):49–59.
4. Wiesner GL, Daley D, Lewis S, et al. A subset of familial colorectal neoplasia kindreds linked to chromosome 9q22.2-31.2. *Proc Natl Acad Sci USA* 2003;100(22):12961–5.
5. Kemp Z, Carvajal-Carmona L, Spain S, et al. Evidence for a colorectal cancer susceptibility locus on chromosome 3q21-q24 from a high-density SNP genome-wide linkage scan. *Hum Mol Genet* 2006;15(19):2903–10.
6. Djureinovic T, Skoglund J, Vandrovcova J, 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(3):362–6.
7. Neklason DW, Kerber RA, Nilson DB, et al. Common familial colorectal cancer linked to chromosome 7q31: a genome-wide analysis. *Cancer Res* 2008;68(21):8993–7.
8. Nieminen TT, Abdel-Rahman WM, Ristimäki A, et al. BMPR1A mutations in hereditary non-polyposis colorectal cancer without mismatch repair deficiency. *Gastroenterology* 2011;141(1):e23–6.
9. Cicek MS, Cunningham JM, Fridley BL, et al. Colorectal cancer linkage on chromosomes 4q21, 8q13, 12q24, and 15q22. *PLoS One* 2012;7(5):e38175.
10. Kontham V, von Holst S, Lindblom A. Linkage analysis in familial non-Lynch syndrome colorectal cancer families from Sweden. *PLoS One* 2013;8(12):e83936.
11. Peters U, Bien S, Zubair N. Genetic architecture of colorectal cancer. *Gut* 2015;64(10):1623–36.
12. Huyghe JR, Bien SA, Harrison TA, et al. Discovery of common and rare genetic risk variants for colorectal cancer. *Nat Genet* 2019;51(1):76–87.
13. Metzker ML. Sequencing technologies - the next generation. *Nat Rev Genet* 2010;11(1):31–46.
14. Toma C, Torrico B, Hervás A, et al. Exome sequencing in multiplex autism families suggests a major role for heterozygous truncating mutations. *Mol Psychiatry* 2014;19(7):784–90.
15. Ng SB, Turner EH, Robertson PD, et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 2009;461(7261):272–6.
16. Palles C, Cazier JB, Howarth KM, et al. Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nat Genet* 2013;45(2):136–44.
17. Weren RD, Ligtenberg MJ, Kets CM, et al. A germline homozygous mutation in the base-excision repair gene NTHL1 causes adenomatous polyposis and colorectal cancer. *Nat Genet* 2015;47(6):668–71.
18. Gylfe AE, Katainen R, Kondelin J, et al. Eleven candidate susceptibility genes for common familial colorectal cancer. *PLoS Genet* 2013;9(10):e1003876.
19. Esteban-Jurado C, Vila-Casadesús M, Garre P, et al. Whole-exome sequencing identifies rare pathogenic variants in new predisposition genes for familial colorectal cancer. *Genet Med* 2015;17(2):131–42.
20. de Voer RM, Hahn MM, Weren RD, et al. Identification of Novel Candidate Genes for Early-Onset Colorectal Cancer Susceptibility. *PLoS Genet* 2016;12(2):e1005880.
21. Brea-Fernandez AJ, Fernandez-Rozadilla C, Alvarez-Barona M, et al. Candidate predisposing germline copy number variants in early onset colorectal cancer patients. *Clin Transl Oncol* 2017;19(5):625–32.
22. Ott J, Wang J, Leal SM. Genetic linkage analysis in the age of whole-genome sequencing. *Nat Rev Genet* 2015;16(5):275–84.
23. Gazal S, Gosset S, Verdura E, et al. Can whole-exome sequencing data be used for linkage analysis? *Eur J Hum Genet* 2016;24(4):581–6.

24. Toma C, Shaw AD, Allcock RJN, et al. An examination of multiple classes of rare variants in extended families with bipolar disorder. *Transl Psychiatry* 2018;8(1):65.
25. Norton N, Li D, Rampersaud E, et al. Exome sequencing and genome-wide linkage analysis in 17 families illustrate the complex contribution of TTN truncating variants to dilated cardiomyopathy. *Circ Cardiovasc Genet* 2013;6(2):144–53.
26. Eggers S, Smith KR, Bahlo M, et al. Whole exome sequencing combined with linkage analysis identifies a novel 3 bp deletion in NR5A1. *Eur J Hum Genet* 2015;23(4):486–93.
27. Esteban-Jurado C, Franch-Expósito S, Muñoz J, et al. The Fanconi anemia DNA damage repair pathway in the spotlight for germline predisposition to colorectal cancer. *Eur J Hum Genet* 2016; 24(10):1501–5.
28. Franch-Expósito S, Esteban-Jurado C, Garre P, et al. Rare germline copy number variants in colorectal cancer predisposition characterized by exome sequencing analysis. *J Genet Genomics* 2018;45(1):41–5.
29. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;25(14):1754–60.
30. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20(9):1297–303.
31. Smith KR, Bromhead CJ, Hildebrand MS, et al. Reducing the exome search space for Mendelian diseases using genetic linkage analysis of exome genotypes. *Genome Biol* 2011;12(9):R85.
32. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007;81(3):559–75.
33. Abecasis GR, Cherny SS, Cookson WO, et al. Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* 2002;30(1): 97–101.
34. Fernandez-Rozadilla C, Cazier JB, Tomlinson IP, et al. A colorectal cancer genome-wide association study in a Spanish cohort identifies two variants associated with colorectal cancer risk at 1p33 and 8p12. *BMC Genomics* 2013;14:55.
35. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* 2013;14(2):178–92.
36. Qiao D, Lange C, Laird NM, et al. Gene-based segregation method for identifying rare variants in family-based sequencing studies. *Genet Epidemiol* 2017;41(4):309–19.
37. Krämer A, Green J, Pollard J Jr, et al. Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics* 2014;30(4):523–30.
38. Rahman N. Realizing the promise of cancer predisposition genes. *Nature* 2014;505(7483):302–8.
39. Turnbull C, Sud A, Houlston RS. Cancer genetics, precision prevention and a call to action. *Nat Genet* 2018;50(9):1212–8.
40. Abulí A, Fernández-Rozadilla C, Giráldez MD, et al. A two-phase case-control study for colorectal cancer genetic susceptibility: candidate genes from chromosomal regions 9q22 and 3q22. *Br J Cancer* 2011;105(6):870–5.
41. Valle L. Recent Discoveries in the Genetics of Familial Colorectal Cancer and Polyposis. *Clin Gastroenterol Hepatol* 2017;15(6):809–19.
42. Gal M, Levanon EY, Hujeirat Y, et al. Novel mutation in TSPAN12 leads to autosomal recessive inheritance of congenital vitreoretinal disease with intra-familial phenotypic variability. *Am J Med Genet A* 2014;164A(12):2996–3002.
43. Hildebrand MS, Tankard R, Gazina EV, et al. PRIMA1 mutation: a new cause of nocturnal frontal lobe epilepsy. *Ann Clin Transl Neurol* 2015;2(8):821–30.
44. Marsh AP, Heron D, Edwards TJ, et al. Mutations in DCC cause isolated agenesis of the corpus callosum with incomplete penetrance. *Nat Genet* 2017;49(4):511–4.
45. Georgi B, Craig D, Kember RL, et al. Genomic view of bipolar disorder revealed by whole genome sequencing in a genetic isolate. *PLoS Genet* 2014;10(3):e1004229.
46. Corominas J, Klein M, Zayats T, et al. Identification of ADHD risk genes in extended pedigrees by combining linkage analysis and whole-exome sequencing. *Mol Psychiatry* 2018. https://doi.org/10.1038/s41380-018-0210-6.
47. Stone DM, Hynes M, Armanini M, et al. The tumour-suppressor gene patched encodes a candidate receptor for Sonic hedgehog. *Nature* 1996; 384(6605):129–34.
48. Twigg SRF, Hufnagel RB, Miller KA, et al. A recurrent mosaic mutation in SMO, encoding the hedgehog signal transducer smoothened, is the major cause of Curry-Jones syndrome. *Am J Hum Genet* 2016;98(6):1256–65.