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

Claudio Toma1,2†, Marcos Díaz-Gay3†, Sebastià Franch-Expósito5, Coral Arnau-Collell3, Bronwyn Overs1,2, Jenifer Muñoz3, Laia Bonjoch7, Yasmin Soares de Lima3, Teresa Ocaña5, Miriam Cuatrecasas4, Antoni Castells5, Luis Bujanda5, Francesc Balaguer7, Joaquín Cubiella6†, Trinidad Caldés7, Janice M. Fullerton1,2 and Sergi Castellví-Bel6†

1Neuroscience Research Australia, Sydney, Australia
2School of Medical Sciences, University of New South Wales, Sydney, Australia
3Gastroenterology Department, Institut d’Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBERehd), Hospital Clinic, University of Barcelona, Barcelona, Spain
4Pathology 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 Clinic, Barcelona, Spain
5Gastroenterology Department, Hospital Donostia-Instituto Biobank, Hospital Clinic, Barcelona, Spain
6Gastroenterology 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
7Gastroenterology 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

Molecular Oncology Laboratory, Hospital Clínico 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).

1C.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/11660058; Grant sponsor: Agència de Gestió d’Ajuts Universitaris i de Recerca; Grant numbers: 2018FI_B1_00213, GRPRES2017SR, GRC 2017SR653; 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: GCB1313592CAST; Grant sponsor: Generalitat de Catalunya, Salut; Grant number: PERIS SLT002/1600398; 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
This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.
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 germline 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 Xp22.3.10 However, previous studies were not able to clearly identify candidate genes 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 germline mutations in genes 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 consecutively 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. 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). 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). The HaplotypeCaller GATK tool was used for variant calling.

**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. WES-derived genotypes were used to confirm familial relationships by pair-wise identity-by-descent (IBD) using PLINK, 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, 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 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. 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. The parameters taken into account were sequencing quality (coverage ≥10x 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.

**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/). 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. 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, PTEN, GALNT12, EPHB2, AXIN2, UNC5C, GREM1, STK11, SMAD4, PTEF1, KLLN, POLE, POLED1, BUB3, BUB1B, RNF43, ATM, PALB2, SMAD9, MSH3, EPCAM, SETD6 and BRF1) plus an additional 115 general cancer predisposition genes. 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 LODlinear = 2.11 and LODexp = 1.872 at rs10753668 or 180.122 cM); chromosome 7q31.2–q34 (with LODlinear = 2.023 and LODexp = 1.838 at rs2075371 or 142.05 cM); and chromosome 10q21.2–q23.1 (with LODlinear = 1.423 and LODexp = 2.149 with rs2134095 being the peak marker (Fig. 3a, Supporting Information Table S1). Likewise, evidence for linkage at the 1q22–q24.2 locus increased to LODlinear = 2.383 (p-value = 4.6E−04) and LODexp = 2.196 (p-value = 7.3E−04) with rs2134095 being the peak marker (Fig. 3a, Supporting Information Table S1). Likewise, evidence for linkage at the 7q31.2–q34 locus increased to LODlinear = 2.197 (p-value = 7.3E−04) and LODexp = 2.149 (p-value = 8.2E−04) with rs6953296 being the peak marker (Fig. 3a, Supporting Information Table S1). Additional markers were subsequently added to fine-map each region and reduce intermarker intervals, evidence for linkage at the 1q22–q24.2 locus increased to LODlinear = 2.383 (p-value = 4.6E−04) and LODexp = 2.196 (p-value = 7.3E−04) with rs2134095 being the peak marker (Fig. 3a, Supporting Information Table S1). Likewise, evidence for linkage at the 7q31.2–q34 locus increased to LODlinear = 2.197 (p-value = 7.3E−04) and LODexp = 2.149 (p-value = 8.2E−04) with rs6953296 being the peak marker (Fig. 3a, Supporting Information Table S1).
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

<table>
<thead>
<tr>
<th>Gene</th>
<th>SNVs/Seg-SNV</th>
<th>Family (patients with Seg-SNV)</th>
<th>p-value</th>
<th>Weighted p-value</th>
<th>Specific LOD at locus</th>
</tr>
</thead>
<tbody>
<tr>
<td>NDST2</td>
<td>1/1</td>
<td>CRC-23 (3)</td>
<td>1.28E-08</td>
<td>2.00E-07</td>
<td>0.49</td>
</tr>
<tr>
<td>FAM78B</td>
<td>1/1</td>
<td>CRC-20 (2)</td>
<td>1.28E-08</td>
<td>2.00E-07</td>
<td>0.20</td>
</tr>
<tr>
<td>GRM8</td>
<td>1/1</td>
<td>CRC-1 (2)</td>
<td>2.56E-08</td>
<td>1.00E-07</td>
<td>0.29</td>
</tr>
<tr>
<td>SYNPO2L</td>
<td>1/1</td>
<td>CRC-11 (2)</td>
<td>2.56E-08</td>
<td>4.00E-07</td>
<td>0.30</td>
</tr>
<tr>
<td>COL13A1</td>
<td>1/1</td>
<td>CRC-11 (2)</td>
<td>2.56E-08</td>
<td>4.00E-07</td>
<td>0.30</td>
</tr>
<tr>
<td>ZSWIM8</td>
<td>1/1</td>
<td>CRC-9 (2)</td>
<td>1.99E-06</td>
<td>5.00E-06</td>
<td>0.47</td>
</tr>
<tr>
<td>MYPN</td>
<td>1/1</td>
<td>CRC-19 (3)</td>
<td>3.94E-06</td>
<td>4.90E-05</td>
<td>0.14</td>
</tr>
<tr>
<td>LMNA</td>
<td>1/1</td>
<td>CRC-9 (2)</td>
<td>2.18E-05</td>
<td>2.60E-05</td>
<td>0.47</td>
</tr>
<tr>
<td>WDR91</td>
<td>1/1</td>
<td>CRC-10 (2)</td>
<td>3.16E-05</td>
<td>6.00E-05</td>
<td>0.27</td>
</tr>
<tr>
<td>LYY</td>
<td>1/1</td>
<td>CRC-23 (3)</td>
<td>3.55E-05</td>
<td>3.90E-04</td>
<td>0.29</td>
</tr>
<tr>
<td>INSRR</td>
<td>1/1</td>
<td>CRC-7 (2)</td>
<td>4.74E-05</td>
<td>4.20E-04</td>
<td>0.29</td>
</tr>
<tr>
<td>TEMEM79</td>
<td>1/1</td>
<td>CRC-5 (3)</td>
<td>1.26E-04</td>
<td>4.50E-04</td>
<td>-0.003</td>
</tr>
<tr>
<td>SMO</td>
<td>1/1</td>
<td>CRC-13 (3)</td>
<td>1.77E-04</td>
<td>6.90E-04</td>
<td>0.14</td>
</tr>
<tr>
<td>C1orf5</td>
<td>1/1</td>
<td>CRC-4 (2)</td>
<td>3.06E-04</td>
<td>3.08E-03</td>
<td>0.20</td>
</tr>
<tr>
<td>CDH23</td>
<td>4/1</td>
<td>CRC-8 (3)</td>
<td>2.97E-04</td>
<td>3.00E-03</td>
<td>0.59</td>
</tr>
<tr>
<td>TNPO3</td>
<td>1/1</td>
<td>CRC-11 (2)</td>
<td>4.71E-04</td>
<td>4.74E-03</td>
<td>0.30</td>
</tr>
<tr>
<td>CFTR</td>
<td>1/1</td>
<td>CRC-20 (2)</td>
<td>7.52E-04</td>
<td>7.47E-03</td>
<td>0.19</td>
</tr>
<tr>
<td>VCL</td>
<td>2/1</td>
<td>CRC-10 (2)</td>
<td>9.15E-04</td>
<td>2.88E-03</td>
<td>0.30</td>
</tr>
</tbody>
</table>

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

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

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.

WES and linkage in familial colorectal cancer

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, 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. 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, 11q, 14q, 22q, 3q29, 4q31.3, 7q31, 10q23, 4q21, 8q13, 12q24, 15q22, 4p16.3, 9q31.1, 17p13.2 and Xp22.33. 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. Subsequently, linkage studies have been superseded by genome-wide association studies (GWAS), which have been able to identify low-risk genetic variants. However, variants identified by GWAS only explain approximately 10% of the variance in genetic liability to CRC, 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. 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 and a Fanci anemia pathway enrichment. WES data have also been used to infer rare copy number variants that could act as the germline mutational event in some families.

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. 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, 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. 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 BMPRIA 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, and Curry-Jones syndrome, a multisystem disorder characterized among other symptoms by skin lesions, polysyndactyly, brain malformations and intestinal malrotation with myofibromas or hamartomas. 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 Clinic-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, 2018 FI_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 (1CF/BQ/ D18/11660058) from “la Caixa” Foundation (ID 100001043) 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 1063980 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 (GCIB1313192CAST), 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, GRPRRE 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
