ORIGINAL ARTICLE

Basic Study

Candidate colorectal cancer predisposing gene variants in Chinese early-onset and familial cases

Jun-Xiao Zhang, Lei Fu, Richarda M de Voer, Marc-Manuel Hahn, Peng Jin, Chen-Xi Lv, Eugène TP Verwiel, Marjolijn JL Ligtenberg, Nicoline Hoogerbrugge, Roland P Kuiper, Jian-Qiu Sheng, Ad Geurts van Kessel

Jun-Xiao Zhang, Richarda M de Voer, Marc-Manuel Hahn, Eugène TP Verwiel, Marjolijn JL Ligtenberg, Nicoline Hoogerbrugge, Roland P Kuiper, Ad Geurts van Kessel, Department of Human Genetics, Radboud University Medical Center, Radboud Institute for Molecular Life Sciences, 6500 HB Nijmegen, The Netherlands

Lei Fu, Peng Jin, Chen-Xi Lv, Jian-Qiu Sheng, Department of Gastroenterology, General Hospital of Beijing Military Region, Beijing 100700, China

Lei Fu, Third Military Medical University, Chongqing 400038, China

Marjolijn JL Ligtenberg, Department of Human Genetics and Department of Pathology, Radboud University Medical Center, 6500 HB Nijmegen, The Netherlands

Correspondence to: Jian-Qiu Sheng, Professor, Department of Gastroenterology, General Hospital of Beijing Military Region, 5 Nanmenchang, Dongcheng, Beijing 100700, China. jianqiu@263.net
Telephone: +86-10-66721299
Fax: +86-10-66721299

Abstract

AIM: To investigate whether whole-exome sequencing may serve as an efficient method to identify known or novel colorectal cancer (CRC) predisposing genes in early-onset or familial CRC cases.

METHODS: We performed whole-exome sequencing in 23 Chinese patients from 21 families with non-polyposis CRC diagnosed at ≤ 40 years of age, or from multiple affected CRC families with at least 1 first-degree relative diagnosed with CRC at ≤ 55 years of age. Genomic DNA from blood was enriched for exome sequences using the SureSelect Human All Exon Kit, version 2 (Agilent Technologies) and sequencing was performed on an Illumina HiSeq 2000 platform. Data were processed through an analytical pipeline to search for rare germline variants in known or novel CRC predisposing genes.

RESULTS: In total, 32 germline variants in 23 genes were identified and confirmed by Sanger sequencing. In 6 of the 21 families (29%), we identified 7 mutations in 3 known CRC predisposing genes including *MLH1* (5 patients), *MSH2* (1 patient), and *MUTYH* (biallelic, 1 patient), five of which were reported as pathogenic. In

the remaining 15 families, we identified 20 rare and novel potentially deleterious variants in 19 genes, six of which were truncating mutations. One previously unreported variant identified in a conserved region of EIF2AK4 (p.Glu738_Asp739insArgArg) was found to represent a local Chinese variant, which was significantly enriched in our early-onset CRC patient cohort compared to a control cohort of 100 healthy Chinese individuals scored negative by colonoscopy (33.3% *vs* 7%, *P* < 0.001).

CONCLUSION: Whole-exome sequencing of early-onset or familial CRC cases serves as an efficient method to identify known and potential pathogenic variants in established and novel candidate CRC predisposing genes.

Key words: Colorectal cancer; Cancer predisposition; Early-onset; Germline variants; Exome sequencing

Core tip: Mendelian colorectal cancer (CRC) predisposition syndromes underlie about 5% of all CRCs, and are caused by germline mutations in a limited set of genes. The overall heritability of CRC, however, is estimated to be approximately 30% and as yet many families at risk remain unexplained. This research identifies seven mutations of known CRC predisposing genes (*MLH1*, *MSH2* and *MUTYH*) in 6 of the 21 families (29%), five of which were previously reported as pathogenic. One unreported variant EIF2AK4 (p.Glu738_Asp739insArgArg) located at conserved region was found to represent a local Chinese variant and significantly enriched in our early-onset CRC patient cohort.

INTRODUCTION

Colorectal cancer (CRC; MIM 114500) is the third most common cancer worldwide and the fourth leading cause of cancer-related death, with over one million new cases diagnosed and approximately 600000 deaths each year[1]. In China, it is the third most common cancer and the fifth leading cause of death from cancer. Moreover, the incidence of CRC in China has been increasing in recent years[2]. Genetic factors are estimated to account for the development of approximately 30% of all CRC cases[3]. However, Mendelian colorectal cancer predisposition syndromes, such as Lynch syndrome (LS), familial adenomatous

polyposis (FAP), MUTYH-associated polyposis (MAP), juvenile polyposis syndrome (JPS) and polymerase proofreading-associated polyposis (PPAP), account for only approximately 5%-10% of all CRC cases and are associated with high-penetrance germline mutations in various mismatch repair (MMR) genes or the *APC*, *MUTYH*, *SMAD4*, *BMPR1A*, *POLE* and *POLD1* genes, respectively[4,5]. The remaining approximately 20%-25% of the cases are thought to be due to moderate- to low-penetrance variants, most of which remain to be identified.

CRC patients with a family history of CRC or an early age at diagnosis are especially suggestive of a hereditary contribution and may be used in genetic association studies to increase the likelihood of identifying susceptibility variants[6-10]. Whereas CRC families with multiple affected individuals may be employed to search for high penetrance genetic susceptibility variants using linkage-based approaches, moderate- to low-penetrance variants cannot be identified through linkage-based studies in large families. In more recent years, multiple low-penetrance genetic loci associated with CRC susceptibility have been identified by genome-wide association studies (GWAS)[11,12]. However, not all results from linkage studies turned out to be consistent, and GWAS are not ideal for the identification of rare variants. Recent advances in next-generation sequencing (NGS) technologies, in particular whole-exome sequencing, have provided efficient means to identify germline variants in individuals with familial or inherited cancer syndromes[5,13-15]. We hypothesized that the majority of the yet unidentified CRC predisposing variants can be identified using whole-exome sequencing when applied to a strictly selected cohort of CRC patients and families. Several cellular signaling pathways appear to be involved in the development of CRC, including the WNT, DNA repair, BMP/TGF-β, apoptosis, MMIF/GIF, and PI3K/AKT pathways[16]. In addition, "sleeping beauty" transposon tagging has recently been employed as an effective forward genetic screening tool for the discovery of novel cancer initiating genes in the mouse intestinal tract, resulting in the identification of hundreds of novel candidate cancer driver genes[17-19].

In this study, we aimed to identify rare and novel germline variants in known and novel candidate CRC predisposing genes by performing whole-exome sequencing of germline DNA of 23 Chinese patients from 21 families diagnosed with non-polyposis CRC at a young age. We initially focused on genes that, based on genetic and functional data, are likely to play a role in CRC development, and on candidate genes that have been identified through GWAS studies.

MATERIALS AND METHODS

Recruitment of patient and control cohorts
Twenty-three patients from 21 families included in this study were recruited through the Department of

Gastroenterology of the General Hospital of Beijing Military Region, Beijing, China. All patients were diagnosed with CRC without polyposis at $\leq$ 40 years of age[20] or from multiple affected CRC families with at least one first-degree relative diagnosed with CRC at $\leq$ 55 years of age. Additionally, 100 colonoscopy test-negative, unrelated controls with Chinese Han ancestry without inflammatory bowel disease or any family history of CRC were collected from a subject pool who participated in health check-up programs, including colonoscopy, at the department of Gastroenterology of the General Hospital of Beijing Military Region, Beijing, China. This study was approved by the Institutional Review Board of the General Hospital of Beijing Military Region (No. 2014-035), and all patients have provided written informed consent.

Whole-exome sequencing
Genomic DNA was extracted from peripheral blood cells using a QIAamp DNA Kit (QIAGEN, Hilden, Germany) according to the protocol provided by the manufacturer and whole-exome sequencing was performed at the Beijing Genome Institute (BGI, Shenzhen, China) according to manufacturer's guide-lines. Briefly, genomic DNA was fragmented and enriched for exome sequences using the SureSelect Human All Exon Kit, version 2 (Agilent Technologies, Santa Clara, CA, United States) and sequencing was performed at a minimal average coverage of 50 × on an Illumina HiSeq 2000 platform (Illumina, Inc., San Diego, CA).

Bioinformatics analyses
After removing sequence adaptors and low-quality reads, Burrows-Wheeler Aligner (BWA)[21] was used to align the reads to the NCBI human reference genome (hg19). Single nucleotide variants (SNVs) were called using SOAPsnp[22] and small insertion/deletions (indels) were detected using the SAMtools software package[23]. All variants were annotated using an in-house annotation pipeline, as described previously[24]. High-confidence variants (total $\geq$ 10 reads, $\geq$ 5 variant reads and $\geq$ 20% variant reads) were subsequently prioritized for variants that were non-synonymous and not found in our in-house database (1302 in-house analyzed exomes, mostly from European ancestry). In addition, dbSNPv138, the National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project database (ESP, 6503 exomes, http://evs.gs.washington.edu/EVS/), and 700 control exome data sets from Chinese subjects with Han ancestry (Juan Tian and Zhimin Feng, BGI, personal communication) were used to exclude recurrent variants with a minor allele frequency (MAF) > 0.001.

Functional impact of variant analyses
Non-synonymous variants that result in alterations in protein function, including protein truncation, splice site

defects and missense mutations at highly conserved (phyloP $\geq$ 3.0) nucleotide positions, were included in our analyses. Alamut v.2.0 software (Interactive Biosoftware) and integrated mutation prediction software (align GVDV, SIFT and PolyPhen-2)[25-27] packages were used for analyses of the identified variants. The prediction of splicing effects was evaluated based on five different algorithms (SpliceSiteFinder, MaxEntScan, NNSPLICE, GeneSplicer, Human Splicing Finder) through the bioinformatics tools of the Alamut v.2.0 software. The online tool "Project HOPE"[28] (http://www.cmbi.ru.nl/hope/) was used for revealing the structural consequences of missense mutations.

Candidate gene selection
We initially selected germline variants in CRC predisposing genes known to be associated with hereditary CRC syndromes and searched for evidence of pathogenicity in relevant databases, *i.e.*, InSiGHT (http://www.insight-group.org/), LOVD (https://atlas.cmm.ki.se/LOVDv.2.0/) and the Mismatch Repair Genes Variant Database (http://www.med.mun.ca/mmrvariants/).

Next to the identification of variants in known CRC predisposing genes, we searched for potential pathogenic variants in novel candidate genes using the remaining exome data of our CRC patient cohort. For the selection of these variants, we focused on genes that meet the following criteria: (1) genes exhibiting recurrent variants; (2) 582 known cancer genes, including somatically mutated cancer genes (Cancer Gene Census, http://www.sanger.ac.uk/genetics/CGP/Census/)[29,30], cancer predisposing genes of which rare germline variants are known to confer a highly or moderately increased risk of cancer and for which at least 5% of individuals with the relevant variants develop cancer[31], and genes that are included in the Radboud university medical center hereditary cancer gene list[32]; (3) 286 genes that have been identified as candidate CRC driver genes by the "sleeping beauty" transposon tagging system in mice[18,19]; (4) 588 genes included in the following KEGG pathways: WNT signaling pathway (hsa04310), TGF-β signaling pathway (hsa04350), base excision repair (BER, hsa03410), nucleotide excision repair (NER, hsa03420), mismatch repair (MMR, hsa03430), non-homologous end-joining (NHEJ, hsa03450), Fanconi anemia pathway (hsa03460) and pathways involved in cancer (hsa05200); and (5) 268 genes likely to play a role in CRC susceptibility identified by GWAS studies[11,12,33,34] and included in the NHGRI GWAS Catalog (http://www.genome.gov/gwastudies/)[35].

Variant validation by Sanger sequencing
Identified germline variants were validated by Sanger sequencing after PCR amplification. The PCR primers were designed *in silico* using the Primer3 software package[36]. PCR reactions were performed on a Dual

Table 1 Clinical characteristics and family histories of 23 early-onset and familial colorectal cancer patients			
Patient ID	**Gender**	**Patient's history**	**Family history**
43-1A	Female	RC at 37 yr	Brother RC at 53 yr
43-2A	Male	RC at 53 yr	Sister RC at 37 yr
49-4A	Male	CC at 30 yr	Brother CC at 43 yr; sister CC at 23 yr
49-5A	Female	CC at 23 yr	Brother CC at 43 yr; brother CC at 30 yr
50-11A	Male	CC at 34 yr and relapse at 36	Father CRC at 35 yr and death at 52 yr; brother CC at 34 yr and death at 36 yr
54-2A	Female	RC at 44 yr	Sister CRC; Brother CC at 76 yr and death
66-1-1A	Female	CC at 47 yr	Sister CC at 51 yr
71A	Female	RC at 57 yr	Sister RC at 53 yr
77-1A	Female	CRC at 38 yr	Father EC at 64 yr and death; uncle CRC at 68 yr and death
102-1A	Male	RC at 25 yr	
103-1A	Male	CC at 53 yr	Brother CC at 36 yr and death at 48 yr; mother IO at 63 yr and death
106-2A	Male	JC at 34 yr, CC at 39 yr, KC at 44 yr and PC at 45 yr	Father EC and death; mother RC at 42 yr and death; Sister CP
108-1A	Male	RC at 33 yr	
110-1A	Male	CC at 36 yr	
116-1A	Female	CC at 31 yr and HC at 57 yr	Brother intussusception and death at 40 yr; Brother CC at 50 yr, RC and SMT at 58 yr; brother IC at 50 yr, CC at 53 yr and RC at 61 yr; sister GC at 56 yr
120-1A	Female	RC at 36 yr	
142-1A	Male	RC at 34 yr	
149-1A	Male	CRC at 31 yr	Father EC and death, mother GC at 56 yr
154-1A	Female	CRC at 40 yr	Father HC, RC and death at 57 yr
156-1A	Female	CRC at 54 yr	Sister CP at 54 yr; sister CP; mother CC at 48 yr; grandfather EC and death.
164-1A	Male	CC at 30 yr	Uncle colonitls at 42 yr
165-1A	Male	CRC at 43 yr	Sister RC at 31 yr and death; grandmother RC at 65 yr and death.
180-1	Male	CRC at 40 yr	Sister CP at 46 yr

CRC: Colorectal cancer; CC: Colon cancer; RC: Rectal cancer; IC: Ileocecus carcinoma; IO: Intestinal obstruction; JC: Jejunum cancer; KC: Kidney cancer; PC: Pulmonary carcinoma; CP: Colonic polyps; HC: Hepatic carcinoma; GC: Gastric cancer; EC: Esophageal cancer; SMT: Splenic metastatic tumors.

96-Well GeneAmp PCR System 9700 (Applied Biosystems) using standard protocols (primer sequences available upon request). Mutation analyses were performed using the Vector NTI software package (Invitrogen, Paisley, United Kingdom).

RESULTS

Patient cohort characteristics

In order to identify known and potential pathogenic variants in established and novel candidate CRC predisposing genes, we performed whole-exome sequencing on germline DNA of 23 CRC patients from 21 families with non-polyposis CRC diagnosed at ⩽ 40 years of age (n = 16), or from multiple affected CRC families with at least one first-degree relative diagnosed with CRC at ⩽ 55 years of age (n = 7). The mean age at diagnosis was 38.6 years, and 43% (n = 10) of the patients were female (Table 1).

Exome sequencing performance

Overall, we generated a mean of 68 M raw reads per sample, of which 77.6% to 89.5% were aligned to the human reference genome (hg19; Table 2). The mean coverage of the exome for the 23 samples was 58.5 × (range: 53.0-64.7×). On average, 87.03% of the reads was covered at least 10 times and 76.35% of the reads was covered at least 20 times.

We identified on average 46437 SNVs (range: 44353-48114) and 1678 indels (range: 1630-1719)

per exome. Over 95.3% of these substitutions and 73.1% of indels represented known variants listed in private and public databases (Figure 1). A prioritization scheme was applied to identify candidate variants (Table 3). Initial quality filtering (total ⩾ 10 reads, ⩾ 5 variant reads and ⩾ 20% variant reads) resulted in the identification of 13819 genetic variants in coding regions or canonical splice sites, including 9833 nonsynonymous changes. A total of 4432 variants that result in alterations in protein function, including 172 nonsense variants, 188 frame shift variants, 943 canonical splice site variants, 237 in-frame deletions, 191 in-frame insertions and 2701 missense variants with high conservation scores (phyloP ⩾ 3.0), were identified. Subsequently, we excluded known variants present in our in-house database and variants with MAF scores > 0.001 in dbSNPv138, reducing the number of variants to 2883. Subsequently, we prioritized variants in known CRC predisposing genes and in genes likely to play a role in CRC development, and excluded variants with MAF scores > 0.1 in the ESP database or in the 700 control exomes from Chinese subjects with Han ancestry, thereby reducing the number of candidate variants to 61. Of these 61, 39 (32 different variants in 23 genes) were validated by Sanger sequencing (Figure 2).

Identification of germline variants in known CRC predisposing genes

A total of seven CRC patients from six families (30%) were identified with germline variants in known CRC

Table 2 Alignment and coverage statistics for 23 early-onset and familial colorectal cancer patients

| Sample ID	Total reads	Total mapped	Reads mapped to genome	Covered ≥ 4×	Covered ≥10×	Covered ≥ 20×	Average target coverage
43-1A	62997602	52130593	45528212	93.30%	85.80%	74.30%	55.88×
43-2A	57099664	50367772	43924906	93.60%	86.40%	75.00%	54.94×
49-4A	67025248	51978393	45418701	93.30%	85.70%	74.10%	55.30×
49-5A	60632336	51598017	45450431	92.60%	85.00%	73.40%	55.49×
50-11A	68991044	58507454	51033221	94.00%	87.20%	76.80%	60.71×
54-2A	68459832	57860336	50820626	93.40%	86.50%	76.10%	61.71×
66-1-1A	69759994	58838035	51472112	94.10%	87.60%	77.50%	61.82×
71A	68055130	58181783	51012277	94.00%	87.60%	77.70%	61.50×
77-1A	65956248	56894265	49817369	93.80%	87.30%	77.10%	61.06×
102-1A	64702600	57086284	49672873	94.40%	87.90%	77.70%	59.92×
103-1A	66004146	55109962	48218769	93.80%	86.90%	76.40%	59.28×
106-2A	61956558	54367359	47567033	93.80%	86.80%	76.00%	57.97×
108-1A	64764180	56469665	49473520	94.00%	87.20%	76.50%	57.52×
110-1A	68883264	56962439	49975545	94.10%	87.30%	76.90%	59.31×
116-1A	68975484	60681318	53305706	93.70%	86.70%	75.60%	55.86×
120-1A	64307066	56593051	49900259	94.20%	87.30%	75.90%	53.02×
142-1A	72999930	65321752	57822754	94.70%	87.60%	76.20%	53.19×
149-1A	69636008	59789305	52641740	93.90%	87.20%	76.90%	61.19×
154-1A	80632788	63934448	56196297	94.40%	88.30%	78.90%	64.69×
156-1A	94340904	82125696	73199086	94.20%	87.10%	76.10%	56.48×
164-1A	67813680	58837471	51779826	93.60%	86.50%	75.80%	58.83×
165-1A	68657326	59845292	52561646	94.30%	87.80%	77.70%	60.99×
180-1	65727112	57908057	50742938	94.40%	87.90%	77.50%	59.01×
Average	68190354	58321250	51197211	93.90%	87.03%	76.35%	58.51×
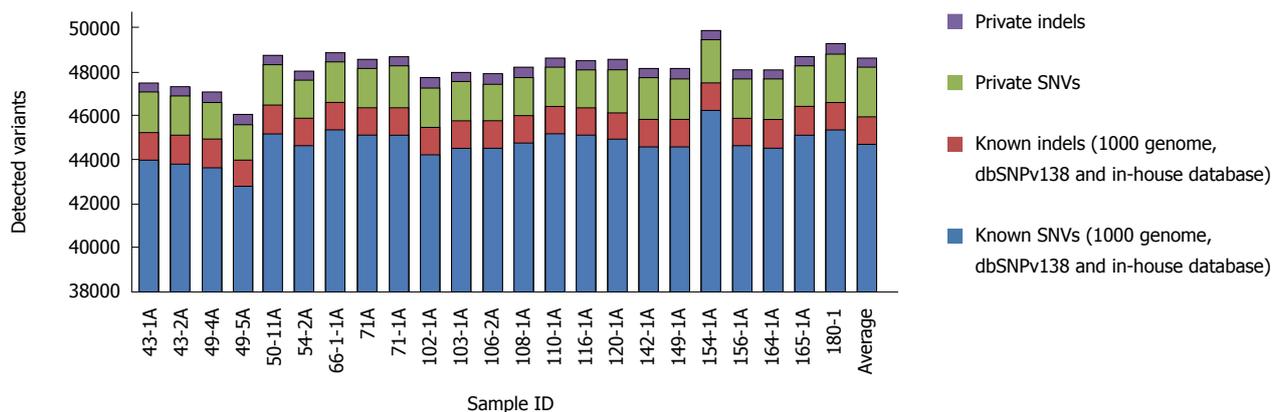
Figure 1 Variant statistics (marked in colors) for 23 early-onset colorectal cancer patient samples. The numbers of detected variants are listed on the left hand side and the patient samples ID on the bottom. The different colors represent different types of variants, *i.e.*, purple represents "private indels", green represents "private SNVs", red and blue represent known indels and known SNVs listed in the 1000 genome, dbSNPv138 and in-house databases, respectively. SNV: Single nucleotide variant.

predisposing genes. Of these, five variants (in four patients) were reported as being pathogenic in public databases, three of which were located in *MLH1*[37] (Table 4), including a canonical splice site mutation (c.453+1G>T) in patient 106-2A (colon cancer at age of 39), a canonical splice site mutation (c.208-1G>A) in patient 116-1A (colon cancer at age of 31), and a missense mutation (c.677G>A, p.Arg226Gln) in patient 43-1A (rectal cancer at age of 37). This latter mutation has been reported to result in a complete skipping of exon 8 at the mRNA level[38]. The brother of patient 43-1A was also subjected to exome sequencing (patient 43-2A, rectal cancer at age of 53), but the *MLH1* mutation c.677G>A was not encountered in

this patient, and subsequent Sanger sequencing confirmed this finding. Compound heterozygous MUTYH mutations (p.Gln267* and p.Gly286Glu) were found in patient 180-1 (CRC at age of 40). The sister of patient 180-1 (colonic polyps at age of 46) also carried both MUTYH mutations (p.Gln267* and p.Gly286Glu). Both mutations have been reported to be causative for MUTYH-associated polyposis (MAP)[39,40].

Three mismatch repair gene mutations, observed in three unrelated patients, were not previously reported in public databases. A novel splice site mutation in *MSH2* (c.793-2A>T) was identified in patient 50-11A (colon cancer at age of 34). This canonical splice site is inactivated and a splice site seven nucleotides

Table 3 Prioritization scheme for exome data analysis of all 23 patients

Type of prioritization filter	Remaining variants (*n*)
All variants	1106642
Coding region and canonical splice site variants after quality filtering (total ≥ 10 reads, ≥ 5 variant reads and ≥ 20% variant reads)	13819
Non-synonymous variants, canonical splice site variants	9833
Variants that result in alterations in protein function (protein truncation, splice site defects and missense mutations at highly conserved (phyloP ≥ 3.0) nucleotide positions.	4432[1]
Not in in-house database and MAF ≤ 0.001 in dbSNPv138	2883
Variants in known CRC predisposing genes and genes likely to play a role in CRC development (MAF ≤ 0.001 in ESP and 700 control Chinese exome data sets)	61
Variants/genes validated by Sanger sequencing	39 (32 different variants in 23 genes)

[1]Including 172 nonsense variants, 188 frame shift variants, 943 canonical splice site variants, 237 in-frame deletions, 191 in-frame insertions and 2701 missense variants with highly conserved (phyloP ≥ 3.0); In-house database: 1302 in-house analyzed exomes, mostly from European ancestry. MAF: Minor allele frequency; ESP: Exome Sequencing Project database (6503 exomes, http://evs.gs.washington.edu/EVS/); 700 control Chinese exome data sets: Chinese subjects with Han ancestry (Juan Tian and Zhi-Min Feng, BGI, personal communication).
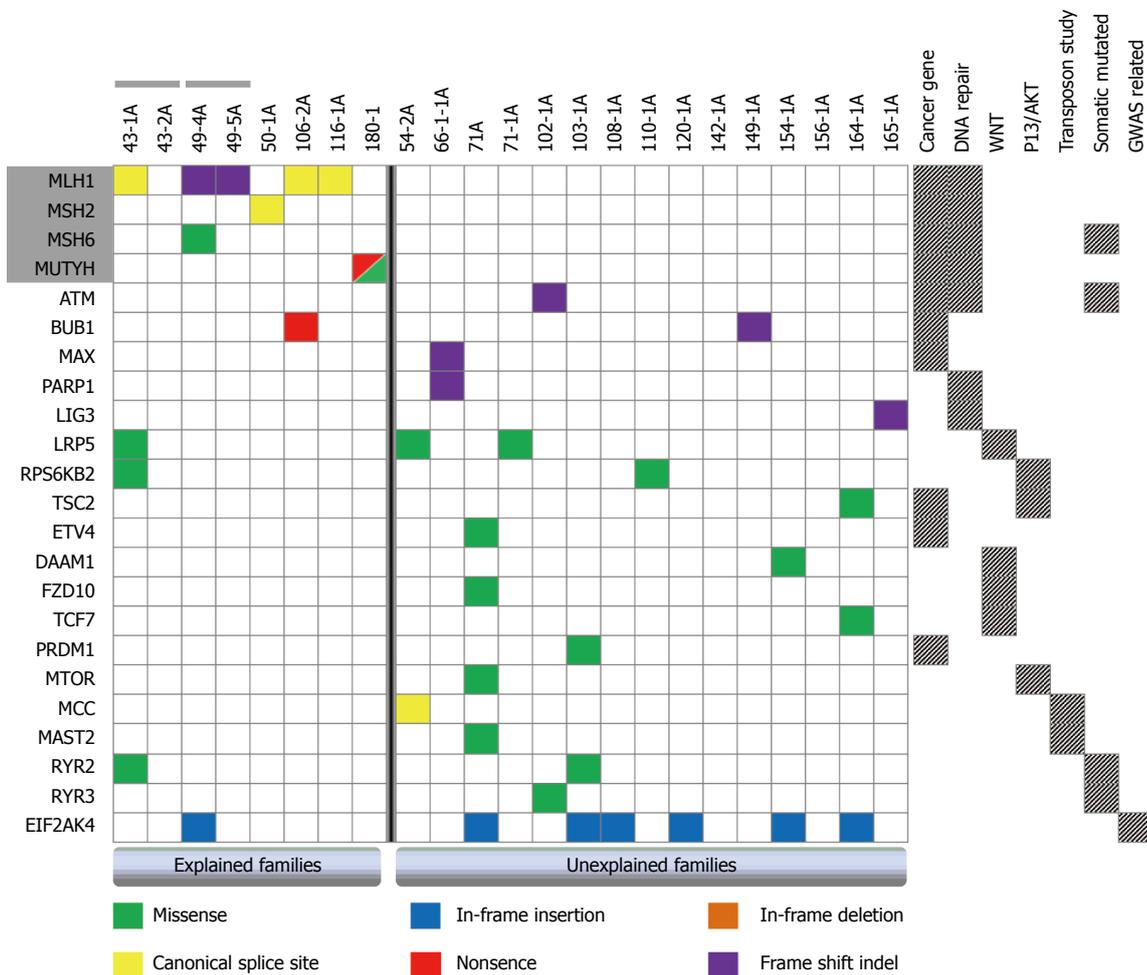
Figure 2 Germline variants identified in known colorectal cancer predisposing genes and genes likely to play a role in colorectal cancer development. The genes are listed on the left hand side and the patient samples on top. Patient samples from the same families are marked (bars). Known colorectal cancer (CRC) predisposing genes are marked by shading (left). The shades at the right hand side of the figure indicate functional (groups of) genes considered to play a role in CRC development. The different variant types are indicated in colors (right). The red-triangle/green-triangle square in sample 180-1 indicates the presence of one *MUTYH* nonsense and one *MUTYH* missense mutation.

Table 4	**Identification of germline mutations in known colorectal cancer predisposing genes**					
Sample ID	**Gene name**	**Gene ID**	**Genomic change**	**cDNA change**	**Protein change**	**Pathogenicity**
43-1A	*MLH1*	NM_000249	g.chr3:37053590G>A	c.677G>A[1]	p.Arg226Gln[a]	Yes[38,44]
106-2A	*MLH1*	NM_000249	chr3:g.37048555G>T	c.453+1G>T	SSM	Yes[42]
116-1A	*MLH1*	NM_000249	g.chr3:37042445G>A	c.208-1G>A	SSM	Yes[43]
180-1	*MUTYH*	NM_001128425	g.chr1:45797972G>A	c.799C>T	p.Gln267*	Yes[39]
180-1	*MUTYH*	NM_001128425	g.chr1:45797914C>T	c.857G>A	p.Gly286Glu	Yes[40]
49-4A	*MLH1*	NM_000249	g.chr3:37067252_37067253insT	c.1163_1164insT	p.Arg389Profs*6	NR
49-5A	*MLH1*	NM_000249	g.chr3:37067252_37067253insT	c.1163_1164insT	p.Arg389Profs*6	NR
49-4A	*MSH6*	NM_000179	g.chr2:48027422C>G	c.2300C>G	p.Thr767Ser	NR
50-11A	*MSH2*	NM_000251	g.chr2:47641406A>T	c.793-2A>T	SSM	NR

[1]This substitution results in a complete loss of exon 8 of *MLH1* by RNA analysis[38]. NR: Not reported; SSM: Splice site mutation.
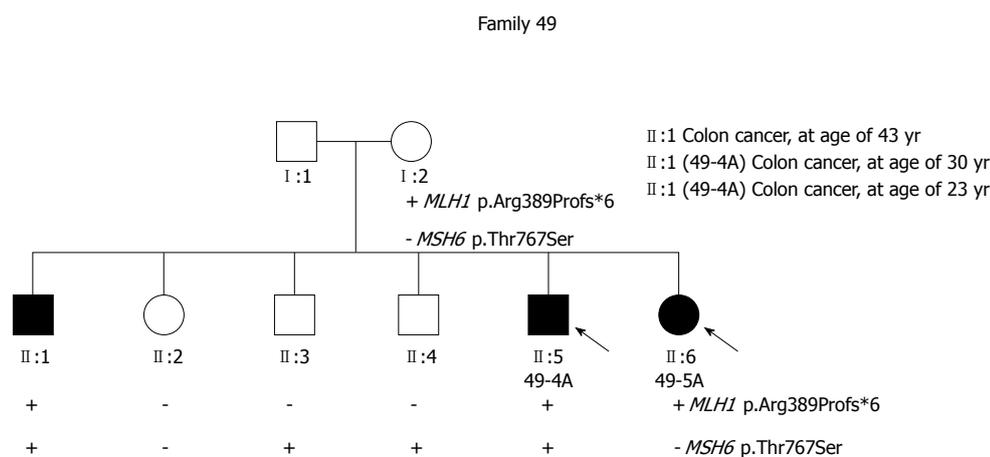
Figure 3 Pedigree and segregation analysis in family members of index patients for *MLH1* and *MSH6* mutations. Index patients are indicated by arrows. Both index patients Ⅱ:5 (sample 49-4A) and Ⅱ:6 (sample 49-5A) carried *MLH1* frame shift mutation (c.1163_1164insT, p.Arg389Profs*6) and Ⅱ:5 also carried *MSH6* missense mutation (c.2300C>G, p.Thr767Ser). Two brothers Ⅱ:1 (colon cancer at age of 43) and Ⅱ:3 (no cancer) carried both mutations. A sister (Ⅱ:2, no cancer) carried neither the *MLH1* nor the *MSH6* mutation. A third brother (Ⅱ:4) carried the *MSH6* mutation, but not the *MLH1* mutation. And the mother of index patients carried *MLH1* mutation, but not the *MSH6* mutation. Both did not develop cancer.

downstream is used according to Alamut prediction. Both a frame shift mutation in MLH1 (p.Arg389Profs*6) and a missense variant in MSH6 (p.Thr767Ser) were found in patient 49-4A (colon cancer at age of 30). The MLH1 mutation p.Arg389Profs*6 was also found in his sister, patient 49-5A (colon cancer at age of 23), whereas this sister was found to be negative for the MSH6 variant p.Thr767Ser. Segregation analysis of four siblings and the mother in this family (Figure 3) showed that the brothers of index patient 49-4A, *i.e.*, family members Ⅱ:1 (colon cancer at age of 43 years) and Ⅱ:3 (no cancer), carried both mutations. The MLH1 p.Arg389Profs*6 mutation-positive, MSH6 wild-type mother Ⅰ:2 and the MLH1 wild-type, MSH6 p.Thr767Ser variant-positive brother Ⅱ:4 both did not develop cancer. We, therefore, conclude that the MLH1 frame shift mutation (p.Arg389Profs*6) acts as the main contributor to the development of CRC in this family.

Rare germline variants of novel candidate CRC predisposing genes

After extrusion of variants in known CRC predisposing genes, a set of 24 rare candidate germline variants remained (Table 5). Of these, seven represent truncating mutations (five frame-shift indels, one nonsense and one canonical splice site). In addition, one in-frame insertion and 16 highly conserved non-synonymous missense variants are present in this set. For these latter variants, SIFT and Polyphen2 algorithms were used to estimate their functional effects on the respective encoded proteins. In all cases, both SIFT and Polyphen2 predicted the variants to be functionally impaired or possibly/probably functionally impaired (Table 6). Four rare or novel variants were found in cancer predisposing genes that are not directly linked to an increased CRC risk, including ATM p.Lys468Glufs*18 in patient 102-1A (rectal cancer at age of 25 years), MAX p.Leu61Serfs*15 in patient 66-1-1A (colon cancer at age of 47 years), TSC2 p.Asp1734Asn in patient 164-1A (colon cancer at age of 30 years) and ETV4 p.Glu331Lys in patient 71A (rectal cancer at age of 57 years). ATM and MAX are involved in DNA repair pathways, and TSC2 plays a role in the PI3K/AKT pathway. These pathways are also active in CRC. Interestingly, in patient 66-1-1A we also observed a potentially deleterious variant in PARP1

Table 5 Characteristics of 24 variants identified in 19 novel genes likely to play a role in colorectal cancer development

Sample ID	Gene name	Gene/pathway involved	cDNA change	Protein change	rs ID in dbSNP138	MAF (700 Chinese exomes)	MAF (NHLBI ESP)	MAF (1000 genome)
102-1A	*ATM*	Cancer gene, DNArep	c.1402_1403del	p.Lys468Glufs*18	NR	NR	NR	NR
66-1-1A	*PARP1*	DNArep	c.758dup	p.Lys254Glufs*6	NR	NR	0.000077	NR
66-1-1A	*MAX*	Cancer gene	c.181del	p.Leu61Serfs*15	NR	NR	NR	NR
106-2A	*BUB1*	Cancer gene	c.46C>T	p.Gln16*	NR	NR	NR	NR
149-1A	*BUB1*	Cancer gene	c.2844del	p.Gln949Argfs*3	NR	NR	NR	NR
165-1A	*LIG3*	DNArep	c.218del	p.Phe73Serfs*41	NR	NR	NR	NR
54-2A	*MCC*	Transposon studies	c.1355+1_1355+2ins14	SMM	NR	NR	NR	NR
49-4A	*EIF2AK4*	GWAS related	c.2214_2215insCGACGA	p.Glu738_Asp739insArgArg	NR	NR	NR	NR
71A	*EIF2AK4*	GWAS related	c.2214_2215insCGACGA	p.Glu738_Asp739insArgArg	NR	NR	NR	NR
103-1A	*EIF2AK4*	GWAS related	c.2214_2215insCGACGA	p.Glu738_Asp739insArgArg	NR	NR	NR	NR
108-1A	*EIF2AK4*	GWAS related	c.2214_2215insCGACGA	p.Glu738_Asp739insArgArg	NR	NR	NR	NR
120-1A	*EIF2AK4*	GWAS related	c.2214_2215insCGACGA	p.Glu738_Asp739insArgArg	NR	NR	NR	NR
154-1A	*EIF2AK4*	GWAS related	c.2214_2215insCGACGA	p.Glu738_Asp739insArgArg	NR	NR	NR	NR
164-1A	*EIF2AK4*	GWAS related	c.2214_2215insCGACGA	p.Glu738_Asp739insArgArg	NR	NR	NR	NR
77-1A	*LRP5*	WNT	c.2156A>G	p.Tyr719Cys	NR	NR	NR	NR
43-1A	*LRP5*	WNT	c.3536G>A	p.Arg1179His	NR	NR	0.000077	NR
54-2A	*LRP5*	WNT	c.3919C>T	p.Arg1307Trp	NR	NR	0.000077	NR
110-1A	*RPS6KB2*	PI3K/AKT	c.331A>G	p.Lys111Glu	NR	0.00075	NR	NR
43-1A	*RPS6KB2*	PI3K/AKT	c.683C>A	p.Thr228Asn	rs183360785	NR	NR	0.001
43-1A	*RYR2*	Somatic mutation gene	c.2701G>A	p.Gly901Ser	NR	NR	NR	NR
103-1A	*RYR2*	Somatic mutation gene	c.6457A>G	p.Lys2153Glu	NR	NR	NR	NR
102-1A	*RYR3*	Somatic mutation gene	c.13507G>A	p.Val4503Met	NR	NR	NR	NR
71A	*ETV4*	Cancer gene	c.991G>A	p.Glu331Lys	NR	NR	NR	NR
103-1A	*PRDM1*	Cancer gene	c.1499A>G	p.Gln500Arg	rs201512476	NR	NR	0.001
164-1A	*TSC2*	Cancer gene, PI3K/AKT	c.5200G>A	p.Asp1734Asn	NR	NR	NR	NR
71A	*MTOR*	PI3K/AKT	c.5857G>T	p.Val1953Leu	NR	0.000714	NR	NR
154-1A	*DAAM1*	WNT	c.667G>A	p.Val223Met	NR	NR	NR	NR
71A	*FZD10*	WNT	c.1341C>G	p.Phe447Leu	NR	NR	NR	NR
164-1A	*TCF7*	WNT	c.572G>T	p.Arg191Met	NR	NR	NR	NR
71A	*MAST2*	Transposon studies	c.3482A>G	p.Asn1161Ser	NR	NR	0.000077	NR

MAF: Minor allele frequency; NR: Not reported; DNArep: DNA repair pathway; WNT: WNT signaling pathway; SSM: Splice site mutation.

(p.Lys254Glufs*6), another gene involved in DNA repair.

Genes recurrently affected by potentially deleterious variants

Despite the limited size of our cohort, the recurrent detection of rare potentially deleterious variants is another way to select candidates from the list of rare variants. Four genes were found to be recurrently affected by different rare variants, and two of them (*BUB1* and *LRP5*) were encountered in patients that also carried pathogenic *MLH1* mutations (patients 106-2A and 43-1A, respectively; Figure 2). In total, two truncating BUB1 variants were found (p.Gln16* and p.Gln949Argfs*3). As reported previously, these BUB1 variants may be associated with an increased risk for aneuploidy and, in patient 106-2A, this may have contributed to somatic loss of the wild-type MLH1 allele in the tumor[15]. The other recurrently affected genes were *LRP5*, *RPS6KB2* and *RYR2*. LRP5 may be of particular interest since it is a component of the WNT-FZD-LRP5-LRP6 complex that triggers β-catenin signaling through the induction of aggregation of receptor-ligand complexes into ribosome-sized signalsomes. We identified three highly conserved *LRP5* missense variants in three unrelated patients (Figure 2). Two of these, p.Tyr719Cys and p.Arg1179His,

were found to be located in the conserved low-density lipoprotein (LDLR) class B repeat region. To investigate the functional consequences of these three mutations on the LRP5 protein structure, the online tool "Project HOPE" was used. By doing so, we found that variant p.Tyr719Cys gives rise to a mutant residue that is smaller and more hydrophobic than the wild-type residue, which may lead to loss of protein-protein interactions and hydrogen bonds and/or disturb correct protein folding. Through variant p.Arg1179His, a positively charged residue is replaced by a neutral and smaller residue, which again may lead to loss of interactions with other molecules or residues. Through variant p.Arg1307Trp, a positively charged residue is replaced by a neutral, larger and more hydrophobic residue, which may lead to loss of interactions with other molecules or residues, loss of hydrogen bonds and/or disturbance of correct protein folding giving rise to collisions with other molecules or residues.

We also identified a recurrent insertion in EIF2AK4 (p.Glu738_Asp739insArgArg) in seven (33.3%) unrelated patients, which was absent in local in-house and public databases. EIF2AK4 is located in a region previously found to be associated with CRC susceptibility in GWAS studies[11,35]. Since this variant could be common in the Han Chinese population, we screened a cohort of 100

Table 6 *In silico* functional prediction of 16 missense variants

Sample ID	Gene name	Gene/ pathway involved	cDNA change	Protein change	Domain	PhyloP score	Grantham score	Align GVGD	SIFT score	SIFT prediction	Polyphen2 score	Polyphen2 prediction
77-1A	LRP5	WNT	c.2156A>G	p.Tyr719Cys	LDLR class B repeat	4.751	194	C65	0.000	D	0.999	PrD
43-1A	LRP5	WNT	c.3536G>A	p.Arg1179His	LDLR class B repeat	3.712	29	C25	0.000	D	0.953	PrD
54-2A	LRP5	WNT	c.3919C>T	p.Arg1307Trp	LDLR class A repeat	3.172	101	C35	0.000	D	0.948	PrD
110-1A	RPS6KB2	PI3K/ AKT	c.331A>G	p.Lys111Glu	Protein kinase, catalytic domain	4.639	56	C55	0.000	D	0.535	PoD
43-1A	RPS6KB2	PI3K/ AKT	c.683C>A	p.Thr228Asn	Protein kinase, catalytic domain	5.062	65	C55	0.001	D	0.994	PrD
43-1A	RYR2	Somatic mutation gene	c.2701G>A	p.Gly901Ser	Ryanodine receptor	6.081	56	C55	0.010	D	1.000	PrD
103-1A	RYR2	Somatic mutation gene	c.6457A>G	p.Lys2153Glu	Intracellular calcium-release channel	5.067	56	C0	0.020	D	0.615	PoD
102-1A	RYR3	Somatic mutation gene	c.13507G>A	p.Val4503Met	Ryanodine Receptor TM 4-6	6.012	21	C15	0.000	D	1.000	PrD
71A	ETV4	Cancer gene	c.991G>A	p.Glu331Lys	PEA3-type ETS-domain transcription factor, N-terminal	6.424	56	C55	0.001	D	0.862	PoD
103-1A	PRDM1	Cancer gene	c.1499A>G	p.Gln500Arg	Zinc finger, C2H2	4.875	43	C0	0.050	D	0.570	PoD
164-1A	TSC2	Cancer gene, PI3K/ AKT	c.5200G>A	p.Asp1734Asn	Rap/ran-GAP	5.538	23	C0	0.000	D	0.998	PrD
71A	MTOR	PI3K/ AKT	c.5857G>T	p.Val1953Leu	PIK-related kinase	5.634	32	C0	0.001	D	0.827	PoD
154-1A	DAAM1	WNT	c.667G>A	p.Val223Met	Diaphanous GTPase-binding	6.347	21	C0	0.000	D	0.998	PrD
71A	FZD10	WNT	c.1341C>G	p.Phe447Leu	Frizzled protein	4.229	22	C15	0.000	D	0.984	PrD
164-1A	TCF7	WNT	c.572G>T	p.Arg191Met	High mobility group, HMG1/ HMG2	4.202	91	C65	0.000	D	0.999	PrD
71A	MAST2	Transposon studies	c.3482A>G	p.Asn1161Ser	PDZ/ DHR/ GLGF	4.854	46	C0	0.000	D	0.999	PrD

D: Damaging; PoD: Possibly damaging; PrD: Probably damaging.

colonoscopy test-negative, unrelated local Han Chinese individuals using Sanger sequencing. We found that 7 (7%) of them carried this variant, revealing a significant enrichment in the early-onset/familial CRC cohort as compared to the ethnicity matched control cohort (χ^2 test, P = 0.000604).

DISCUSSION

In order to identify rare and novel germline variants that may predispose to CRC, we applied whole-exome sequencing to 23 Chinese patients from 21 families with non-polyposis CRC diagnosed at ≤ 40 years of age or from multiple affected CRC families with at least one first-degree relative diagnosed with CRC at ≤ 55 years of age. Initially we selected variants in genes that are known to be associated with hereditary CRC syndromes, and we assessed their pathogenicity as reported in public databases such as InSiGHT, LOVD and the Mismatch Repair Genes Variant database. Among

the 23 patients included, we identified seven patients (from six families; approximately 30%) with variants in known CRC predisposing genes. This percentage is lower than that previously reported by Tanskanen *et al*[41], (42%, 16/38) in a cohort of early-onset CRC patients (< 40 years) using exome sequencing. In a study by Tanskanen *et al*[41], of 38 patients, four were clinically diagnosed with gastrointestinal polyposis (three FAP and one JPS), and 12 were identified with germline MMR mutations and enriched in patients with MSI tumors (86%, 12/14). This discrepancy may be due to the fact that our cohort is a non-polyposis cohort and also includes patients from multiple affected CRC families with at least one first-degree relative diagnosed with CRC at ≤ 55 years of age. In our cohort, six patients were identified with variants in the high-penetrance genes *MLH1*, *MSH2* and *MSH6* underlying Lynch syndrome. In addition, we identified biallelic *MUTYH* mutations, underlying MAP, in one index patient (patient 180-1, CRC at age of 40) and the sister of the patient (colonic polyps at age of 46). Of the eight variants that we identified in known high-penetrance CRC predisposing genes, *MLH1* c.453+1G>T, *MLH1* c.208-1G>A, *MLH1* c.677G>A, MUTYH p.Gln267* and MUTYH p.Gly286Glu were reported as being pathogenic in public databases[39,40,42-44]. In addition, we identified novel rare variants of which two, MLH1 p.Arg389Profs*6 and *MSH2* c.793-2A>T, are most likely pathogenic based on both familial segregation and *in silico* prediction analyses.

In our search for novel germline predisposing variants, we focused on known cancer-associated genes, CRC pathway-associated genes, mouse CRC susceptibility genes identified by transposon ('sleeping beauty') tagging, GWAS-associated genes and genes with reported somatic mutations that are considered likely to be involved in CRC predisposition and/or development. Using these criteria, we identified a total of 19 novel candidate CRC susceptibility genes carrying rare, likely deleterious, variants.

One ATM truncating variant (p.Lys468Glufs*18) identified in patient 102-1A (rectal cancer at age of 25) may be particularly relevant. *ATM* is a gene encoding a protein that belongs to the PI3/PI4-kinase family[45]. The ATM protein represents an important cell cycle checkpoint kinase that is required for a cell's response to DNA damage and for ensuring genomic integrity[46]. Diseases associated with ATM mutations include ataxia telangiectasia (AT), an autosomal recessive disorder[47]. Because of its role in maintaining genomic integrity, ATM may, when mutated, increase the risk for tumor development[48]. Indeed, germline mutations in *ATM* have been shown to increase the risk of breast cancer development through the (de)regulation of BRCA1[49]. In addition, loss of heterozygosity at the *ATM* locus has been found in CRC[50]. Taken together, it appears plausible to assume that germline *ATM* mutations may increase the risk for CRC development. However, considering the high frequency of truncating mutation

in ESP database and in-house database, it is crucial for targeted screening of *ATM* in a large early-onset and/or familial CRC cohort. Another interesting candidate is the truncating MAX variant (p.Leu61Serfs*15) identified in patient 66-1-1A. The protein encoded by the *MAX* gene represents the most conserved dimerization component of the MYC-MAX-MXD1 network of basic helix-loop-helix leucine zipper (bHLHZ) transcription factors that regulate cellular proliferation, differentiation and apoptosis[51,52]. It has been shown that the MAX protein interacts with MSH2[53], and that mutant MAX is able to alter the growth and morphology of CRC cells through inactivation of c-MYC[32]. Mutations in the MAX gene have been reported to be associated with the occurrence of hereditary pheochromocytomas and paragangliomas[54]. Interestingly, an additional truncating variant in PARP1 (p.Lys254Glufs*6) was identified in this patient (66-1-1A). PARP1 is activated in response to DNA damage and plays an important role in DNA repair processes, apoptosis and cell cycle control[55]. Since MAX and PARP1 are both involved in DNA repair, and since it has been shown that PARP1 is essential for c-MYC-induced transactivation and retardation of the G2-M transition in cancer cells[56], the combination of these two variants may have a synergistic effect. Therefore, we anticipate that both truncating variants most likely play a role in CRC development in this family.

Other interesting candidate genes recurrently affected by potentially deleterious variants include *BUB1*, *LRP5* and *EIF2AK4*. Two truncating variants in BUB1 (p.Gln16* and p.Gln949Argfs*3) were found to be present in patient 106-2A and patient 149-1A, respectively. The BUB1 protein is an integral component of the spindle assembly checkpoint (SAC), and we have previously shown that germline variants in the corresponding gene may serve as risk factors for CRC[15]. Patient 106-2A was found to carry both BUB1 p.Gln16* and *MLH1* c.453+1G>T variants. We suggest that BUB1 may have contributed to loss of the wild-type *MLH1* allele in this patient[15]. Obviously, this latter scenario requires validation in larger CRC cohorts.

Three missense LRP5 variants (p.Tyr719Cys, p.Arg1179His and p.Arg1307Trp), found in three CRC cases, were predicted to be deleterious. LRP5 p.Tyr719Cys and LRP5 p.Arg1307Trp were observed in patient 54-2A and patient 77-1A, respectively. In both cases no other putative pathogenic germline variants were detected. Variant LRP5 p.Arg1179His was found in patient 43-1A, who also carried a pathogenic *MLH1* c.677G>A splice site mutation. The LRP5 protein is a component of the WNT-FZD-LRP5-LRP6 complex and, as such, represents an important partner in the WNT signal transduction pathway[57]. Variants LRP5 p.Tyr719Cys and p.Arg1179His are both located in the conserved low-density lipoprotein receptor (LDLR) class B repeat region of LRP5, which is the binding region of Dickkopf-1, a developmental protein antagonist of the canonical WNT-β-catenin pathway[58].

Further assessment of both LRP5 variants using the "Project HOPE" tool indicated that these variants may also result in loss of interactions with other proteins or residues. It has previously been shown that truncated LRP5 proteins are frequently expressed in breast tumors of different developmental stages[59] and that these proteins are strongly implicated in the deregulation of the WNT-β-catenin signaling pathway in hyperparathyroid tumors[60].

One EIF2AK4 variant (p.Glu738_Asp739insArgArg) was recurrently found in seven (33.3%) unrelated patients within our cohort. After comparison of our cohort to an ethnicity matched control cohort, this variant was found to be significantly enriched ($P = 0.000604$). We, therefore, conclude that also this latter gene may be considered a candidate CRC predisposing gene.

A major challenge of using whole-exome sequencing is the identification of predisposing pathogenic variants within the vast background of non-pathogenic variants. Targeted screening of those genes and variants in replicate large early-onset and/or familial CRC cohorts will be instrumental in gaining more robust evidence for pathogenicity. Our current results, however, already vividly illustrate that whole-exome sequencing in carefully selected cases at risk for hereditary cancer may serve as an attractive approach to identify rare and novel variants in known and novel candidate CRC predisposing genes.

ACKNOWLEDGMENTS

COMMENTS

Background
Mendelian colorectal cancer (CRC) predisposition syndromes underlie about 5% of all CRC cases, and are caused by germline mutations in a limited set of genes. The current selection of causative genes to be screened in high-risk families is based on several phenotypic characteristics, including polyposis (*e.g., APC* and *MUTYH*) and microsatellite instability (*MLH1, MSH2, MSH6* and *PMS2*). The overall heritability of CRC, however, is estimated to be approximately 30%. Excluding hereditary forms, there is an important fraction of CRC cases that present familial aggregation for the disease with an unknown germline genetic cause.

Research frontiers
CRC patients with a family history of CRC or an early age at diagnosis are especially suggestive of a hereditary contribution and may be used in genetic association studies to increase the likelihood of identifying susceptibility variants. Whereas CRC families with multiple affected individuals may be employed to search for high penetrance genetic susceptibility variants using linkage-based approaches, moderate- to low-penetrance variants cannot be identified through linkage-based studies in large families. In more recent years, multiple low-penetrance genetic loci associated with CRC susceptibility have been identified by genome-wide association studies (GWAS). However, not all results from linkage studies turned out to be consistent, and GWAS are not ideal for the identification of rare variants. Recently, advances in next-generation sequencing technologies, in particular whole-exome sequencing, have provided efficient means to identify germline variants in individuals with familial or inherited cancer syndromes.

Innovations and breakthroughs
A major challenge of using whole-exome sequencing is the identification of predisposing pathogenic variants within the vast background of non-pathogenic variants. In this study, we performed whole-exome sequencing in a strictly selected cohort of CRC patients and families that are very young CRC patients (diagnosed at ≤ 40 years of age) or familial CRC cases. And data were processed through a tailored analytical pipeline to search for rare germline variants in known or novel CRC predisposing genes.

Applications
The study show that whole-exome sequencing of early-onset or familial CRC cases serves as an efficient method to identify known and potential pathogenic variants in established and novel candidate CRC predisposing genes. The findings also provide insight into the role of these variants in CRC development. Targeted screening of those genes and variants in replicate large early-onset and/or familial CRC cohorts will be instrumental in gaining more robust evidence for pathogenicity.

Terminology
"Early-onset" CRC: CRC is traditionally thought to be a disease of older patients with most being diagnosed after the age of 50 years; however, a significant proportion of young patients present with this disease. Early age of onset is a central characteristic of hereditary predisposition to cancer. Familiar aggregation of tumors and hereditary cases are constantly more frequent under the age of 40 years.

Peer-review
This study investigated the efficiency of whole-exome sequencing in identifying known or novel CRC predisposing genes in early-onset or familial CRC cases. This is a well written paper that has been performed stringently. Although the number of included patients is very low, the authors present very interesting results with a straight forward conclusion.

REFERENCES

1　**Jemal A**, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. *CA Cancer J Clin* 2011; **61**: 69-90 [PMID: 21296855 DOI: 10.3322/caac.20107]

2　**Chen W**, Zheng R, Zhang S, Zhao P, Li G, Wu L, He J. Report of incidence and mortality in China cancer registries, 2009. *Chin J Cancer Res* 2013; **25**: 10-21 [PMID: 23372337 DOI: 10.3978/j.issn.1000-9604.2012.12.04]

3　**Lichtenstein P**, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, Pukkala E, Skytthe A, Hemminki K. 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**: 78-85 [PMID: 10891514 DOI: 10.1056/NEJM200007133430201]

4　**de la Chapelle A**. Genetic predisposition to colorectal cancer. *Nat Rev Cancer* 2004; **4**: 769-780 [PMID: 15510158 DOI: 10.1038/nrc1453]

5　**Palles C**, Cazier JB, Howarth KM, Domingo E, Jones AM, Broderick P, Kemp Z, Spain SL, Guarino E, Salguero I, Sherborne A, Chubb D, Carvajal-Carmona LG, Ma Y, Kaur K, Dobbins S, Barclay E, Gorman M, Martin L, Kovac MB, Humphray S, Lucassen A, Holmes CC, Bentley D, Donnelly P, Taylor J, Petridis C, Roylance R, Sawyer EJ, Kerr DJ, Clark S, Grimes J, Kearsey SE, Thomas HJ, McVean G, Houlston RS, Tomlinson I. Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nat Genet* 2013; **45**: 136-144 [PMID: 23263490 DOI: 10.1038/ng.2503]

6　**Lynch HT**, Smyrk T. Hereditary nonpolyposis colorectal cancer (Lynch syndrome). An updated review. *Cancer* 1996; **78**: 1149-1167 [PMID: 8826936]

7　**Schoen RE**. Families at risk for colorectal cancer: risk assessment and genetic testing. *J Clin Gastroenterol* 2000; **31**: 114-120 [PMID: 10993425]

8　**Gryfe R**, Kim H, Hsieh ET, Aronson MD, Holowaty EJ, Bull

SB, Redston M, Gallinger S. Tumor microsatellite instability and clinical outcome in young patients with colorectal cancer. *N Engl J Med* 2000; **342**: 69-77 [PMID: 10631274 DOI: 10.1056/NEJM200001133420201]

9   **Giráldez MD**, Balaguer F, Bujanda L, Cuatrecasas M, Muñoz J, Alonso-Espinaco V, Larzabal M, Petit A, Gonzalo V, Ocaña T, Moreira L, Enríquez-Navascués JM, Boland CR, Goel A, Castells A, Castellví-Bel S. MSH6 and MUTYH deficiency is a frequent event in early-onset colorectal cancer. *Clin Cancer Res* 2010; **16**: 5402-5413 [PMID: 20924129 DOI: 10.1158/1078-0432.CCR-10-1491]

10   **Chang DT**, Pai RK, Rybicki LA, Dimaio MA, Limaye M, Jayachandran P, Koong AC, Kunz PA, Fisher GA, Ford JM, Welton M, Shelton A, Ma L, Arber DA, Pai RK. Clinicopathologic and molecular features of sporadic early-onset colorectal adenocarcinoma: an adenocarcinoma with frequent signet ring cell differentiation, rectal and sigmoid involvement, and adverse morphologic features. *Mod Pathol* 2012; **25**: 1128-1139 [PMID: 22481281 DOI: 10.1038/modpathol.2012.61]

11   **Tenesa A**, Dunlop MG. New insights into the aetiology of colorectal cancer from genome-wide association studies. *Nat Rev Genet* 2009; **10**: 353-358 [PMID: 19434079 DOI: 10.1038/nrg2574]

12   **Houlston RS**, Cheadle J, Dobbins SE, Tenesa A, Jones AM, Howarth K, Spain SL, Broderick P, Domingo E, Farrington S, Prendergast JG, Pittman AM, Theodoratou E, Smith CG, Olver B, Walther A, Barnetson RA, Churchman M, Jaeger EE, Penegar S, Barclay E, Martin L, Gorman M, Mager R, Johnstone E, Midgley R, Niittymäki I, Tuupanen S, Colley J, Idziaszczyk S, Thomas HJ, Lucassen AM, Evans DG, Maher ER, Maughan T, Dimas A, Dermitzakis E, Cazier JB, Aaltonen LA, Pharoah P, Kerr DJ, Carvajal-Carmona LG, Campbell H, Dunlop MG, Tomlinson IP. Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. *Nat Genet* 2010; **42**: 973-977 [PMID: 20972440 DOI: 10.1038/ng.670]

13   **Jones S**, Hruban RH, Kamiyama M, Borges M, Zhang X, Parsons DW, Lin JC, Palmisano E, Brune K, Jaffee EM, Iacobuzio-Donahue CA, Maitra A, Parmigiani G, Kern SE, Velculescu VE, Kinzler KW, Vogelstein B, Eshleman JR, Goggins M, Klein AP. Exomic sequencing identifies PALB2 as a pancreatic cancer susceptibility gene. *Science* 2009; **324**: 217 [PMID: 19264984 DOI: 10.1126/science.1171202]

14   **Comino-Méndez I**, Gracia-Aznárez FJ, Schiavi F, Landa I, Leandro-García LJ, Letón R, Honrado E, Ramos-Medina R, Caronia D, Pita G, Gómez-Graña A, de Cubas AA, Inglada-Pérez L, Maliszewska A, Taschin E, Bobisse S, Pica G, Loli P, Hernández-Lavado R, Díaz JA, Gómez-Morales M, González-Neira A, Roncador G, Rodríguez-Antona C, Benítez J, Mannelli M, Opocher G, Robledo M, Cascón A. Exome sequencing identifies MAX mutations as a cause of hereditary pheochromocytoma. *Nat Genet* 2011; **43**: 663-667 [PMID: 21685915 DOI: 10.1038/ng.861]

15   **de Voer RM**, Geurts van Kessel A, Weren RD, Ligtenberg MJ, Smeets D, Fu L, Vreede L, Kamping EJ, Verwiel ET, Hahn MM, Ariaans M, Spruijt L, van Essen T, Houge G, Schackert HK, Sheng JQ, Venselaar H, van Ravenswaaij-Arts CM, van Krieken JH, Hoogerbrugge N, Kuiper RP. Germline mutations in the spindle assembly checkpoint genes BUB1 and BUB3 are risk factors for colorectal cancer. *Gastroenterology* 2013; **145**: 544-547 [PMID: 23747338 DOI: 10.1053/j.gastro.2013.06.001]

16   **Wood LD**, Parsons DW, Jones S, Lin J, Sjöblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, Silliman N, Szabo S, Dezso Z, Ustyanksky V, Nikolskaya T, Nikolsky Y, Karchin R, Wilson PA, Kaminker JS, Zhang Z, Croshaw R, Willis J, Dawson D, Shipitsin M, Willson JK, Sukumar S, Polyak K, Park BH, Pethiyagoda CL, Pant PV, Ballinger DG, Sparks AB, Hartigan J, Smith DR, Suh E, Papadopoulos N, Buckhaults P, Markowitz SD, Parmigiani G, Kinzler KW, Velculescu VE, Vogelstein B. The genomic landscapes of human breast and colorectal cancers. *Science* 2007; **318**: 1108-1113 [PMID: 17932254 DOI: 10.1126/science.1145720]

17   **Starr TK**, Largaespada DA. Cancer gene discovery using the

Sleeping Beauty transposon. *Cell Cycle* 2005; **4**: 1744-1748 [PMID: 16294016]

18   **Starr TK**, Allaei R, Silverstein KA, Staggs RA, Sarver AL, Bergemann TL, Gupta M, O'Sullivan MG, Matise I, Dupuy AJ, Collier LS, Powers S, Oberg AL, Asmann YW, Thibodeau SN, Tessarollo L, Copeland NG, Jenkins NA, Cormier RT, Largaespada DA. A transposon-based genetic screen in mice identifies genes altered in colorectal cancer. *Science* 2009; **323**: 1747-1750 [PMID: 19251594 DOI: 10.1126/science.1163040]

19   **March HN**, Rust AG, Wright NA, ten Hoeve J, de Ridder J, Eldridge M, van der Weyden L, Berns A, Gadiot J, Uren A, Kemp R, Arends MJ, Wessels LF, Winton DJ, Adams DJ. Insertional mutagenesis identifies multiple networks of cooperating genes driving intestinal tumorigenesis. *Nat Genet* 2011; **43**: 1202-1209 [PMID: 22057237 DOI: 10.1038/ng.990]

20   **Domati F**, Maffei S, Kaleci S, Di Gregorio C, Pedroni M, Roncucci L, Benatti P, Magnani G, Marcheselli L, Bonetti LR, Mariani F, Alberti AM, Rossi V, Ponz de Leon M. Incidence, clinical features and possible etiology of early onset (≤40 years) colorectal neoplasms. *Intern Emerg Med* 2014; **9**: 623-631 [PMID: 23929387 DOI: 10.1007/s11739-013-0981-3]

21   **Li H**, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009; **25**: 1754-1760 [PMID: 19451168 DOI: 10.1093/bioinformatics/btp324]

22   **Li R**, Li Y, Kristiansen K, Wang J. SOAP: short oligonucleotide alignment program. *Bioinformatics* 2008; **24**: 713-714 [PMID: 18227114 DOI: 10.1093/bioinformatics/btn025]

23   **Li H**, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009; **25**: 2078-2079 [PMID: 19505943 DOI: 10.1093/bioinformatics/btp352]

24   **Vissers LE**, de Ligt J, Gilissen C, Janssen I, Steehouwer M, de Vries P, van Lier B, Arts P, Wieskamp N, del Rosario M, van Bon BW, Hoischen A, de Vries BB, Brunner HG, Veltman JA. A de novo paradigm for mental retardation. *Nat Genet* 2010; **42**: 1109-1112 [PMID: 21076407 DOI: 10.1038/ng.712]

25   **Tavtigian SV**, Deffenbaugh AM, Yin L, Judkins T, Scholl T, Samollow PB, de Silva D, Zharkikh A, Thomas A. Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. *J Med Genet* 2006; **43**: 295-305 [PMID: 16014699 DOI: 10.1136/jmg.2005.033878]

26   **Kumar P**, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 2009; **4**: 1073-1081 [PMID: 19561590 DOI: 10.1038/nprot.2009.86]

27   **Adzhubei IA**, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. *Nat Methods* 2010; **7**: 248-249 [PMID: 20354512 DOI: 10.1038/nmeth0410-248]

28   **Venselaar H**, Te Beek TA, Kuipers RK, Hekkelman ML, Vriend G. Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. *BMC Bioinformatics* 2010; **11**: 548 [PMID: 21059217 DOI: 10.1186/1471-2105-11-548]

29   **Cancer Genome Atlas Network**. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 2012; **487**: 330-337 [PMID: 22810696 DOI: 10.1038/nature11252]

30   **Futreal PA**, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. A census of human cancer genes. *Nat Rev Cancer* 2004; **4**: 177-183 [PMID: 14993899 DOI: 10.1038/nrc1299]

31   **Rahman N**. Realizing the promise of cancer predisposition genes. *Nature* 2014; **505**: 302-308 [PMID: 24429628 DOI: 10.1038/nature12981]

32   **Neveling K**, Feenstra I, Gilissen C, Hoefsloot LH, Kamsteeg EJ, Mensenkamp AR, Rodenburg RJ, Yntema HG, Spruijt L, Vermeer S, Rinne T, van Gassen KL, Bodmer D, Lugtenberg D, de Reuver R, Buijsman W, Derks RC, Wieskamp N, van den Heuvel B, Ligtenberg MJ, Kremer H, Koolen DA, van de Warrenburg BP,

Cremers FP, Marcelis CL, Smeitink JA, Wortmann SB, van Zelst-Stams WA, Veltman JA, Brunner HG, Scheffer H, Nelen MR. A post-hoc comparison of the utility of sanger sequencing and exome sequencing for the diagnosis of heterogeneous diseases. *Hum Mutat* 2013; **34**: 1721-1726 [PMID: 24123792 DOI: 10.1002/humu.22450]

33 **Tomlinson IP**, Carvajal-Carmona LG, Dobbins SE, Tenesa A, Jones AM, Howarth K, Palles C, Broderick P, Jaeger EE, Farrington S, Lewis A, Prendergast JG, Pittman AM, Theodoratou E, Olver B, Walker M, Penegar S, Barclay E, Whiffin N, Martin L, Ballereau S, Lloyd A, Gorman M, Lubbe S, Howie B, Marchini J, Ruiz-Ponte C, Fernandez-Rozadilla C, Castells A, Carracedo A, Castellvi-Bel S, Duggan D, Conti D, Cazier JB, Campbell H, Sieber O, Lipton L, Gibbs P, Martin NG, Montgomery GW, Young J, Baird PN, Gallinger S, Newcomb P, Hopper J, Jenkins MA, Aaltonen LA, Kerr DJ, Cheadle J, Pharoah P, Casey G, Houlston RS, Dunlop MG. Multiple common susceptibility variants near BMP pathway loci GREM1, BMP4, and BMP2 explain part of the missing heritability of colorectal cancer. *PLoS Genet* 2011; **7**: e1002105 [PMID: 21655089 DOI: 10.1371/journal.pgen.1002105]

34 **Smith CG**, Naven M, Harris R, Colley J, West H, Li N, Liu Y, Adams R, Maughan TS, Nichols L, Kaplan R, Wagner MJ, McLeod HL, Cheadle JP. Exome resequencing identifies potential tumor-suppressor genes that predispose to colorectal cancer. *Hum Mutat* 2013; **34**: 1026-1034 [PMID: 23585368 DOI: 10.1002/humu.22333]

35 **Hindorff LA**, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* 2009; **106**: 9362-9367 [PMID: 19474294 DOI: 10.1073/pnas.0903103106]

36 **Untergasser A**, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. Primer3--new capabilities and interfaces. *Nucleic Acids Res* 2012; **40**: e115 [PMID: 22730293 DOI: 10.1093/nar/gks596]

37 **Thompson BA**, Spurdle AB, Plazzer JP, Greenblatt MS, Akagi K, Al-Mulla F, Bapat B, Bernstein I, Capellá G, den Dunnen JT, du Sart D, Fabre A, Farrell MP, Farrington SM, Frayling IM, Frebourg T, Goldgar DE, Heinen CD, Holinski-Feder E, Kohonen-Corish M, Robinson KL, Leung SY, Martins A, Moller P, Morak M, Nystrom M, Peltomaki P, Pineda M, Qi M, Ramesar R, Rasmussen LJ, Royer-Pokora B, Scott RJ, Sijmons R, Tavtigian SV, Tops CM, Weber T, Wijnen J, Woods MO, Macrae F, Genuardi M. Application of a 5-tiered scheme for standardized classification of 2,360 unique mismatch repair gene variants in the InSiGHT locus-specific database. *Nat Genet* 2014; **46**: 107-115 [PMID: 24362816 DOI: 10.1038/ng.2854]

38 **Pagenstecher C**, Wehner M, Friedl W, Rahner N, Aretz S, Friedrichs N, Sengteller M, Henn W, Buettner R, Propping P, Mangold E. Aberrant splicing in MLH1 and MSH2 due to exonic and intronic variants. *Hum Genet* 2006; **119**: 9-22 [PMID: 16341550 DOI: 10.1007/s00439-005-0107-8]

39 **Kim DW**, Kim IJ, Kang HC, Jang SG, Kim K, Yoon HJ, Ahn SA, Han SY, Hong SH, Hwang JA, Sohn DK, Jeong SY, Choi HS, Hong CW, Lim SB, Park JG. Germline mutations of the MYH gene in Korean patients with multiple colorectal adenomas. *Int J Colorectal Dis* 2007; **22**: 1173-1178 [PMID: 17703316 DOI: 10.1007/s00384-007-0289-8]

40 **Yanaru-Fujisawa R**, Matsumoto T, Ushijima Y, Esaki M, Hirahashi M, Gushima M, Yao T, Nakabeppu Y, Iida M. Genomic and functional analyses of MUTYH in Japanese patients with adenomatous polyposis. *Clin Genet* 2008; **73**: 545-553 [PMID: 18422726 DOI: 10.1111/j.1399-0004.2008.00998.x]

41 **Tanskanen T**, Gylfe AE, Katainen R, Taipale M, Renkonen-Sinisalo L, Mecklin JP, Järvinen H, Tuupanen S, Kilpivaara O, Vahteristo P, Aaltonen LA. Exome sequencing in diagnostic evaluation of colorectal cancer predisposition in young patients. *Scand J Gastroenterol* 2013; **48**: 672-678 [PMID: 23544471 DOI: 10.3109/00365521.2013.783102]

42 **Sheng JQ**, Fu L, Sun ZQ, Huang JS, Han M, Mu H, Zhang H, Zhang YZ, Zhang MZ, Li AQ, Wu ZT, Han Y, Li SR. Mismatch repair gene mutations in Chinese HNPCC patients. *Cytogenet Genome Res* 2008; **122**: 22-27 [PMID: 18931482]

43 **Goldberg Y**, Porat RM, Kedar I, Shochat C, Sagi M, Eilat A, Mendelson S, Hamburger T, Nissan A, Hubert A, Kadouri L, Pikarski E, Lerer I, Abeliovich D, Bercovich D, Peretz T. Mutation spectrum in HNPCC in the Israeli population. *Fam Cancer* 2008; **7**: 309-317 [PMID: 18389388 DOI: 10.1007/s10689-008-9191-y]

44 **Arnold S**, Buchanan DD, Barker M, Jaskowski L, Walsh MD, Birney G, Woods MO, Hopper JL, Jenkins MA, Brown MA, Tavtigian SV, Goldgar DE, Young JP, Spurdle AB. Classifying MLH1 and MSH2 variants using bioinformatic prediction, splicing assays, segregation, and tumor characteristics. *Hum Mutat* 2009; **30**: 757-770 [PMID: 19267393 DOI: 10.1002/humu.20936]

45 **Savitsky K**, Sfez S, Tagle DA, Ziv Y, Sartiel A, Collins FS, Shiloh Y, Rotman G. The complete sequence of the coding region of the ATM gene reveals similarity to cell cycle regulators in different species. *Hum Mol Genet* 1995; **4**: 2025-2032 [PMID: 8589678]

46 **Abraham RT**. Cell cycle checkpoint signaling through the ATM and ATR kinases. *Genes Dev* 2001; **15**: 2177-2196 [PMID: 11544175 DOI: 10.1101/gad.914401]

47 **McKinnon PJ**. ATM and the molecular pathogenesis of ataxia telangiectasia. *Annu Rev Pathol* 2012; **7**: 303-321 [PMID: 22035194 DOI: 10.1146/annurev-pathol-011811-132509]

48 **Pusapati RV**, Rounbehler RJ, Hong S, Powers JT, Yan M, Kiguchi K, McArthur MJ, Wong PK, Johnson DG. ATM promotes apoptosis and suppresses tumorigenesis in response to Myc. *Proc Natl Acad Sci USA* 2006; **103**: 1446-1451 [PMID: 16432227 DOI: 10.1073/pnas.0507367103]

49 **Broeks A**, Urbanus JH, Floore AN, Dahler EC, Klijn JG, Rutgers EJ, Devilee P, Russell NS, van Leeuwen FE, van 't Veer LJ. ATM-heterozygous germline mutations contribute to breast cancer-susceptibility. *Am J Hum Genet* 2000; **66**: 494-500 [PMID: 10677309]

50 **Uhrhammer N**, Bay J, Pernin D, Rio P, Grancho M, Kwiatkowski F, Gosse-Brun S, Daver A, Bignon Y. Loss of heterozygosity at the ATM locus in colorectal carcinoma. *Oncol Rep* 1999; **6**: 655-658 [PMID: 10203610]

51 **Blackwood EM**, Eisenman RN. Max: a helix-loop-helix zipper protein that forms a sequence-specific DNA-binding complex with Myc. *Science* 1991; **251**: 1211-1217 [PMID: 2006410]

52 **Blackwood EM**, Lüscher B, Eisenman RN. Myc and Max associate in vivo. *Genes Dev* 1992; **6**: 71-80 [PMID: 1730411]

53 **Mac Partlin M**, Homer E, Robinson H, McCormick CJ, Crouch DH, Durant ST, Matheson EC, Hall AG, Gillespie DA, Brown R. Interactions of the DNA mismatch repair proteins MLH1 and MSH2 with c-MYC and MAX. *Oncogene* 2003; **22**: 819-825 [PMID: 12584560 DOI: 10.1038/sj.onc.1206252]

54 **Burnichon N**, Cascón A, Schiavi F, Morales NP, Comino-Méndez I, Abermil N, Inglada-Pérez L, de Cubas AA, Amar L, Barontini M, de Quirós SB, Bertherat J, Bignon YJ, Blok MJ, Bobisse S, Borrego S, Castellano M, Chanson P, Chiara MD, Corssmit EP, Giacchè M, de Krijger RR, Ercolino T, Girerd X, Gómez-García EB, Gómez-Graña A, Guilhem I, Hes FJ, Honrado E, Korpershoek E, Lenders JW, Letón R, Mensenkamp AR, Merlo A, Mori L, Murat A, Pierre P, Plouin PF, Prodanov T, Quesada-Charneco M, Qin N, Rapizzi E, Raymond V, Reisch N, Roncador G, Ruiz-Ferrer M, Schillo F, Stegmann AP, Suarez C, Taschin E, Timmers HJ, Tops CM, Urioste M, Beuschlein F, Pacak K, Mannelli M, Dahia PL, Opocher G, Eisenhofer G, Gimenez-Roqueplo AP, Robledo M. MAX mutations cause hereditary and sporadic pheochromocytoma and paraganglioma. *Clin Cancer Res* 2012; **18**: 2828-2837 [PMID: 22452945 DOI: 10.1158/1078-0432.CCR-12-0160]

55 **Schreiber V**, Dantzer F, Ame JC, de Murcia G. Poly(ADP-ribose): novel functions for an old molecule. *Nat Rev Mol Cell Biol* 2006; **7**: 517-528 [PMID: 16829982 DOI: 10.1038/nrm1963]

56 **Pyndiah S**, Tanida S, Ahmed KM, Cassimere EK, Choe C, Sakamuro D. c-MYC suppresses BIN1 to release poly(ADP-ribose) polymerase 1: a mechanism by which cancer cells acquire cisplatin resistance. *Sci Signal* 2011; **4**: ra19 [PMID: 21447800

DOI: 10.1126/scisignal.2001556]

57   **MacDonald BT**, He X. Frizzled and LRP5/6 receptors for Wnt/ β-catenin signaling. *Cold Spring Harb Perspect Biol* 2012; **4**: [PMID: 23209147 DOI: 10.1101/cshperspect.a007880]

58   **Zorn AM**. Wnt signalling: antagonistic Dickkopfs. *Curr Biol* 2001; **11**: R592-R595 [PMID: 11516963]

59   **Björklund P**, Svedlund J, Olsson AK, Akerström G, Westin G. The internally truncated LRP5 receptor presents a therapeutic target in breast cancer. *PLoS One* 2009; **4**: e4243 [PMID: 19158955 DOI: 10.1371/journal.pone.0004243]

60   **Björklund P**, Akerström G, Westin G. An LRP5 receptor with internal deletion in hyperparathyroid tumors with implications for deregulated WNT/beta-catenin signaling. *PLoS Med* 2007; **4**: e328 [PMID: 18044981 DOI: 10.1371/journal.pmed.0040328]

9 771007 932045