The following full text is a publisher's version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/154700

Please be advised that this information was generated on 2021-01-03 and may be subject to change.
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

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 \leq 40 years of age, or from multiple affected CRC families with at least 1 first-degree relative diagnosed with CRC at \leq 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.
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

© The Author(s) 2015. Published by Baishideng Publishing Group Inc. All rights reserved.

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 (MLHI, 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 600,000 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, BMPRIA, 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 ≤ 40 years of age\(^2\) or from multiple affected CRC families with at least one first-degree relative diagnosed with CRC at ≤ 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 guidelines. 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 ≥ 10 reads, ≥ 5 variant reads and ≥ 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 (NHBLI) 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 ≥ 3.0) nucleotide positions, were included in our analyses. Alamut v.2.0 software (Interactive Biosoftware) and integrated mutation prediction software (align GVD, 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/LOVDv2.0/) 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/).\(^{30}\)

**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...
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 non-synonymous 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
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 downstream is utilized in its place. Furthermore, a frameshifting in-frame deletion spanning 19 amino acids was found in the MLH1 protein sequence of patient 106-2A (colon cancer at age of 31). This genotype likely results in a protein that cannot form active mismatch repair complexes.

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 downstream is utilized in its place. Furthermore, a frameshifting in-frame deletion spanning 19 amino acids was found in the MLH1 protein sequence of patient 106-2A (colon cancer at age of 31). This genotype likely results in a protein that cannot form active mismatch repair complexes.
Table 3  Prioritization scheme for exome data analysis of all 23 patients

<table>
<thead>
<tr>
<th>Type of prioritization filter</th>
<th>Remaining variants (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All variants</td>
<td>1106642</td>
</tr>
<tr>
<td>Coding region and canonical splice site variants after quality filtering (total ≥ 10 reads, ≥ 5 variant reads and ≥ 20% variant reads)</td>
<td>13819</td>
</tr>
<tr>
<td>Non-synonymous variants, canonical splice site variants</td>
<td>9833</td>
</tr>
<tr>
<td>Variants that result in alterations in protein function (protein truncation, splice site defects and missense mutations at highly conserved (phyloP ≥ 3.0) nucleotide positions.)</td>
<td>4432</td>
</tr>
<tr>
<td>Not in in-house database and MAF ≤ 0.001 in dbSNPv138</td>
<td>2883</td>
</tr>
<tr>
<td>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)</td>
<td>61</td>
</tr>
<tr>
<td>Variants/genes validated by Sanger sequencing</td>
<td>39 (32 different variants in 23 genes)</td>
</tr>
</tbody>
</table>

1Including 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.
Family 49

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, II:2) carried the MLH1 mutation and both did not develop cancer. A third brother (II:4) carried the MSH6 mutation, but not the MLH1 mutation. 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 49-5A (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 4  Identification of germline mutations in known colorectal cancer predisposing genes

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Gene name</th>
<th>Gene ID</th>
<th>Genomic change</th>
<th>cDNA change</th>
<th>Protein change</th>
<th>Pathogenicity</th>
</tr>
</thead>
<tbody>
<tr>
<td>43-1A</td>
<td>MLH1</td>
<td>NM_000249</td>
<td>g.chr3:37033590G&gt;A</td>
<td>c.677G&gt;A</td>
<td>p.Arg226Gln</td>
<td>__________________</td>
</tr>
<tr>
<td>106-2A</td>
<td>MLH1</td>
<td>NM_000249</td>
<td>chr3:37405555G&gt;T</td>
<td>c.453+1G&gt;T</td>
<td>SSM</td>
<td>__________________</td>
</tr>
<tr>
<td>116-1A</td>
<td>MLH1</td>
<td>NM_000249</td>
<td>g.chr3:37402445G&gt;A</td>
<td>c.208-1G&gt;A</td>
<td>SSM</td>
<td>__________________</td>
</tr>
<tr>
<td>180-1</td>
<td>MUTYH</td>
<td>NM_001128245</td>
<td>g.chr1:45797997G&gt;A</td>
<td>c.799C&gt;T</td>
<td>p.Gln267*</td>
<td>Yes[4,5]</td>
</tr>
<tr>
<td>180-1</td>
<td>MUTYH</td>
<td>NM_001128245</td>
<td>g.chr1:45797994G&gt;A</td>
<td>c.857G&gt;A</td>
<td>p.Gly268Glu</td>
<td>Yes[4,5]</td>
</tr>
<tr>
<td>49-4A</td>
<td>MLH1</td>
<td>NM_000249</td>
<td>g.chr3:37067252_37067255insT</td>
<td>c.1163_1164insT</td>
<td>p.Arg389Profs*6</td>
<td>NR</td>
</tr>
<tr>
<td>49-5A</td>
<td>MLH1</td>
<td>NM_000249</td>
<td>g.chr3:37067252_37067255insT</td>
<td>c.1163_1164insT</td>
<td>p.Arg389Profs*6</td>
<td>NR</td>
</tr>
<tr>
<td>49-4A</td>
<td>MSH6</td>
<td>NM_000179</td>
<td>g.chr2:48807423G&gt;T</td>
<td>c.2308C&gt;G</td>
<td>p.Thr767Ser</td>
<td>NR</td>
</tr>
<tr>
<td>50-11A</td>
<td>MSH2</td>
<td>NM_000251</td>
<td>g.chr2:47641406A&gt;T</td>
<td>c.793-2A&gt;T</td>
<td>SSM</td>
<td>NR</td>
</tr>
</tbody>
</table>

1This substitution results in a complete loss of exon 8 of MLH1 by RNA analysis[44]. NR: Not reported; SSM: Splice site mutation.

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 49-5A (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.
(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 signalosomes. 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, larger and more hydrophobic residue, which may lead to loss of interactions with other molecules or residues. Through variant p.Arg1307Trp, a positively charged residue is replaced by a neutral smaller residue, which again may lead to loss of interactions and hydrogen bonds and/or disturb correct protein folding giving rise to collisions 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
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 ($\chi^2$ 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 $\leq 40$ years of age or from multiple affected CRC families with at least one first-degree relative diagnosed with CRC at $\leq 55$ 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

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

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Gene name</th>
<th>Gene/pathway involved</th>
<th>cDNA change</th>
<th>Protein change</th>
<th>Domain</th>
<th>PhyloP score</th>
<th>Grantham score</th>
<th>Align GVGD</th>
<th>SIFT score</th>
<th>SIFT prediction</th>
<th>Polyphen2 score</th>
<th>Polyphen2 prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>77-1A</td>
<td>LRP5</td>
<td>WNT</td>
<td>c.2156A&gt;G</td>
<td>p.Tyr719Cys</td>
<td>LDLR class B repeat</td>
<td>4.751</td>
<td>194</td>
<td>C65</td>
<td>0.000</td>
<td>D</td>
<td>0.999</td>
<td>PrD</td>
</tr>
<tr>
<td>43-1A</td>
<td>LRP5</td>
<td>WNT</td>
<td>c.3536G&gt;A</td>
<td>p.Arg1179His</td>
<td>LDLR class B repeat</td>
<td>3.712</td>
<td>29</td>
<td>C25</td>
<td>0.000</td>
<td>D</td>
<td>0.953</td>
<td>PrD</td>
</tr>
<tr>
<td>54-2A</td>
<td>LRP5</td>
<td>WNT</td>
<td>c.3919C&gt;T</td>
<td>p.Arg1307Trp</td>
<td>LDLR class A repeat</td>
<td>3.172</td>
<td>101</td>
<td>C35</td>
<td>0.000</td>
<td>D</td>
<td>0.948</td>
<td>PrD</td>
</tr>
<tr>
<td>110-1A</td>
<td>RPS6KB2</td>
<td>PI3K/AKT</td>
<td>c.331A&gt;G</td>
<td>p.Lys111Glu</td>
<td>Protein kinase, catalytic domain</td>
<td>4.639</td>
<td>56</td>
<td>C55</td>
<td>0.000</td>
<td>D</td>
<td>0.535</td>
<td>PoD</td>
</tr>
<tr>
<td>43-1A</td>
<td>RPS6KB2</td>
<td>PI3K/AKT</td>
<td>c.683C&gt;A</td>
<td>p.Thr228Asn</td>
<td>Protein kinase, catalytic domain</td>
<td>5.062</td>
<td>65</td>
<td>C55</td>
<td>0.001</td>
<td>D</td>
<td>0.994</td>
<td>PrD</td>
</tr>
<tr>
<td>43-1A</td>
<td>RYR2</td>
<td>Somatic mutation gene</td>
<td>c.2701G&gt;A</td>
<td>p.Gly901Ser</td>
<td>Ryanodine receptor</td>
<td>6.081</td>
<td>56</td>
<td>C55</td>
<td>0.010</td>
<td>D</td>
<td>1.000</td>
<td>PrD</td>
</tr>
<tr>
<td>103-1A</td>
<td>RYR2</td>
<td>Somatic mutation gene</td>
<td>c.6457A&gt;G</td>
<td>p.Lys2153Glu</td>
<td>Intraacellular calcium-release channel</td>
<td>6.067</td>
<td>56</td>
<td>C0</td>
<td>0.020</td>
<td>D</td>
<td>0.615</td>
<td>PoD</td>
</tr>
<tr>
<td>102-1A</td>
<td>RYR3</td>
<td>Somatic mutation gene</td>
<td>c.13507G&gt;A</td>
<td>p.Val4503Met</td>
<td>Ryanodine Receptor TM 4-6</td>
<td>6.012</td>
<td>21</td>
<td>C15</td>
<td>0.000</td>
<td>D</td>
<td>1.000</td>
<td>PrD</td>
</tr>
<tr>
<td>71A</td>
<td>ETV4</td>
<td>Cancer gene</td>
<td>c.991G&gt;A</td>
<td>p.Glu331Lys</td>
<td>PEA3-type ETS-domain transcription factor, N-terminal</td>
<td>6.424</td>
<td>56</td>
<td>C55</td>
<td>0.001</td>
<td>D</td>
<td>0.862</td>
<td>PoD</td>
</tr>
<tr>
<td>103-1A</td>
<td>PRDM1</td>
<td>Cancer gene</td>
<td>c.1499G&gt;A</td>
<td>p.Gln500Arg</td>
<td>Zinc finger, C2H2</td>
<td>4.875</td>
<td>43</td>
<td>C0</td>
<td>0.050</td>
<td>D</td>
<td>0.570</td>
<td>PoD</td>
</tr>
<tr>
<td>164-1A</td>
<td>TSC2</td>
<td>Cancer gene, PI3K/AKT</td>
<td>c.5200G&gt;A</td>
<td>p.Asp1734Asn</td>
<td>Rap/ran-GAP</td>
<td>5.538</td>
<td>23</td>
<td>C0</td>
<td>0.000</td>
<td>D</td>
<td>0.998</td>
<td>PrD</td>
</tr>
<tr>
<td>71A</td>
<td>MTOR</td>
<td>Cancer gene</td>
<td>c.5857G&gt;T</td>
<td>p.Val1953Leu</td>
<td>PIK-related kinase</td>
<td>5.634</td>
<td>32</td>
<td>C0</td>
<td>0.001</td>
<td>D</td>
<td>0.827</td>
<td>PoD</td>
</tr>
<tr>
<td>154-1A</td>
<td>DAAM1</td>
<td>WNT</td>
<td>c.667G&gt;A</td>
<td>p.Val223Met</td>
<td>Diaphanous GTPase-binding Frizzled protein</td>
<td>6.347</td>
<td>21</td>
<td>C0</td>
<td>0.000</td>
<td>D</td>
<td>0.998</td>
<td>PrD</td>
</tr>
<tr>
<td>71A</td>
<td>FZD10</td>
<td>WNT</td>
<td>c.1341C&gt;G</td>
<td>p.Phe447Leu</td>
<td>High mobility group, HMG1/HMG2</td>
<td>4.229</td>
<td>22</td>
<td>C15</td>
<td>0.000</td>
<td>D</td>
<td>0.984</td>
<td>PrD</td>
</tr>
<tr>
<td>164-1A</td>
<td>TCF7</td>
<td>WNT</td>
<td>c.572G&gt;T</td>
<td>p.Arg191Met</td>
<td>High mobility group, HMG1/HMG2/PI3K</td>
<td>4.202</td>
<td>91</td>
<td>C65</td>
<td>0.000</td>
<td>D</td>
<td>0.999</td>
<td>PrD</td>
</tr>
<tr>
<td>71A</td>
<td>MAST2</td>
<td>Transposon studies</td>
<td>c.3482A&gt;G</td>
<td>p.Asn1161Ser</td>
<td>PDZ/DHR/GLGF</td>
<td>4.854</td>
<td>46</td>
<td>C0</td>
<td>0.000</td>
<td>D</td>
<td>0.999</td>
<td>PrD</td>
</tr>
</tbody>
</table>

D: Damaging; PoD: Possibly damaging; PrD: Probably damaging.
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/P4-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[54]. Mutations in the MAX gene have been reported to be associated with 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.Gln164* 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[57]. Patient 106-2A was found to carry both BUB1 p.Gln164* 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[58]. 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[59]. 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[60].
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
We appreciate Dr. Ying Han, Dr. Hai-Hong Wang, Dr. Xin Wang, Dr. Ai-Qin Li, Dr. Xiao-Wei Wang and Dr. Hui Su from Department of Gastroenterology, General Hospital of Beijing Military Region, Beijing, China for their kindly help in sample collection and we thank the patients and families for participating and their cooperation in this study.

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


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


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]


