Received: 20 January 2023 | Revised: 1 May 2023 | Accepted: 21 May 2023

DOI: 10.1002/kjm2.12715

ORIGINAL ARTICLE

Kaohsiung Journal of Medical Sciences KJMS WILEY

Whole exome sequencing identifies *MAP3K1*, *MSH2*, and *MLH1* as potential cancer-predisposing genes in familial early-onset colorectal cancer

Nayeralsadat Fatemi[1] | Siang-Jyun Tu[2] | Chin-Chun Chung[2] | Pardis Ketabi Moghadam[3] | Ehsan Nazemalhosseini Mojarad[3] | Amir Sadeghi[3] | Mehdi Totonchi[1,4] | Hamid Asadzadeh Aghdaei[1] | Jan-Gowth Chang[2,5]

[1]Basic and Molecular Epidemiology of Gastrointestinal Disorders Research Center, Research Institute for Gastroenterology and Liver Diseases, Shahid Beheshti University of Medical Sciences, Tehran, Iran

[2]Center for Precision Medicine, China Medical University Hospital, Taichung, Taiwan

[3]Gastroenterology and Liver Diseases Research Center, Research Institute for Gastroenterology and Liver Diseases, Shahid Beheshti University of Medical Sciences, Tehran, Iran

[4]Department of Genetics, Reproductive Biomedicine Research Center, Royan Institute for Reproductive Biomedicine, ACECR, Tehran, Iran

[5]Department of Bioinformatics and Medical Engineering, Asia University, Taichung, Taiwan

Correspondence
Hamid Asadzadeh Aghdaei, Basic and Molecular Epidemiology of Gastrointestinal Disorders Research Center, Research Institute for Gastroenterology and Liver Diseases, Shahid Beheshti University of Medical Sciences, Tehran, Iran.
Email: hamid.assadzadeh@gmail.com

Jan-Gowth Chang, Center for Precision Medicine, China Medical University Hospital, Taichung, Taiwan, P.O. Box: 404327, Taichung, Taiwan.
Email: d6781@mail.cmuh.org.tw

Funding information
China Medical University Hospital, Grant/Award Number: DMR-112-233

Abstract

The incidence of early-onset colorectal cancer (CRC), which affects people under 50, is increasing for unknown reasons. Additionally, no underlying genetic cause is found in 20%–30% of patients suspected of having familial CRC syndrome. Whole exome sequencing (WES) has generated evidence for new genes associated with CRC susceptibility, but many patients remain undiagnosed. This study applied WES in five early-onset CRC patients from three unrelated families to identify novel genetic variants that could be linked to rapid disease development. Furthermore, the candidate variants were validated using Sanger sequencing. Two heterozygote variations, c.1077-2A>G and c.199G>A, were found in the *MSH2* and the *MLH1* genes, respectively. Sanger sequencing analysis confirmed that these (likely) pathogenic mutations segregated in all the affected families' members. In addition, we identified a rare heterozygote variant (c.175C>T) with suspected pathogenic potential in the *MAP3K1* gene; formally the variant is of uncertain significance (VUS). Our findings support the hypothesis that CRC onset may be oligogenic and molecularly heterogeneous. Larger and more robust studies are needed to understand the genetic basis of early-onset CRC development, combined with novel functional analyses and omics approaches.

KEYWORDS
colorectal cancer, early-onset, likely pathogenic, whole exome sequencing

1 | INTRODUCTION

Colorectal cancer (CRC) is one of the most common cancers worldwide, and its incidence was estimated to be around 1.9 million new cases in 2020.[1] The highest incidence rates of CRC are observed in developed countries, particularly North America, Europe, and Oceania, where rates exceed 40 per 100,000 population. By contrast, the lowest incidence rates are observed in Africa and South-Central Asia, where rates are less than 10 per 100,000 population.[2]

Evidence shows the incidence of CRC is 14.6 per 100,000 males and 11.1 per 100,000 females per year in Iran.[3] Approximately 20%–30% of CRC cases have a heritable component. This is either a pathogenic variant in CRC-susceptibility genes such as mismatch repair (MMR) genes in Lynch syndrome (LS), *APC*, *MUTYH*, and *NTHL1* in adenomatous polyposis syndromes, or a positive family history with an unknown genetic cause.[4–7] These patients might carry rare variants that are dominantly inherited or with low-moderate penetrance that act synergistically.[5] The lack of a genetic diagnosis is a global problem in familial and early onset CRC (EOCRC; cases with an age of onset <50 years) and has led to studies using larger gene panels to explain the genetic causes. Overall, variable genes have been examined, but often a monogenic cause has been identified in up to 22% of the patients analyzed, and the highest diagnostic yield was in younger patients.[5,8,9] Identifying underlying genetic causes in early-onset and/or familial CRC patients facilitates proper clinical management of the patient and their families and surveillance of affected and unaffected carriers. Given the broad spectrum of potential causative variants in these patients, multi-gene panels or whole exome sequencing (WES) testing is recommended.[8,9] Here, we used WES of three unrelated families diagnosed with CRC at a young age. We searched for novel or rare pathogenic variants that could explain the premature occurrence of CRC. In addition, to analyze the pathogenic potential of one of the candidate variants, we performed in silico protein structure modeling.

2 | MATERIALS AND METHODS

2.1 | Study patients and clinical evaluation

Figure 1A and Table 1 show the families' pedigrees and patient characteristics. We obtained blood samples from the available family members (affected and healthy) and reviewed the subjects' medical records. The selected patients were all diagnosed with CRC before/at 40. No previous mismatch repair (MMR) analysis was conducted for the index person or a relative. The study protocol was approved by the ethics committee at the Research Center for Gastroenterology and Liver Diseases (RCGLD), and the study participants gave informed consent.

2.2 | Whole exome sequencing data analysis

A total of 300 ng DNA (based on Qubit quantification) was used for library preparation. The MGIEasy FS DNA Library Prep Set was used for fragmentation, end repair, adaptor ligation, and enrichment of DNA fragments, and the MGIEasy Exome Capture V5 Probe Set was used for exome capture. Before sequencing, circular DNA was formed using the Circularization Kit of MGIEasy FS DNA Library Prep Set, after which we constructed DNA nanoballs (DNBs) using a high-throughput Sequencing Set (FCL PE150). All quality control was performed using the Qubit 4.0 Fluorometer (Invitrogen) and an Agilent 4200 Bioanalyzer (Agilent Technologies). The library was sequenced on the MGI DNBSEQ-G400 with 100× mean target coverage.

Bioinformatics analysis was performed using the MegaBOLT bioinformatics analysis accelerator (MegaBOLT v2.2.2.0). In detail, fastq files were first trimmed for adaptors ("AAGTCGGAGGCCAAGCGGTCTTAGGAAGACAA" and "AAGTCGGATCGTAGCCATGTCGTTCTGTGAGCCAAGGAGTTG"), and low-quality reads were filtered out. The minimap2[10] tool was used to align the reads against the human genome (hg19). Variant identification followed GATK Best Practices,[11] including marking duplicates, performing base quality score recalibration (BQSR), and calling variants using HaplotypeCaller with known variant sites from the dbSNP (v151)[12] and the 1000 Genomes Project.[13]

2.3 | Mutation filtering and annotation

To identify cancer-related variants, only 724 cancer-related genes were selected.[14] Functional annotation was performed using ANNOVAR,[15] and population frequency was analyzed using gnomAD (v2.1.1).[16] The clinical interpretation was assessed by ClinVar (accessed November 07, 2021)[17] and HGMD,[18] and pathogenic prediction scores were obtained using CADD (v1.6)[19] and REVEL.[20] Potentially pathogenic variants were then defined as falling into one of two categories: (1) pathogenic/likely pathogenic (P/LP) in ClinVar or (2) having an allele frequency of <1% and a CADD score >20 (CADD20).

2.4 | Mutation validation and co-segregation analysis

The mutations identified by WES in the *MSH2*, *MLH1*, and *MAP3K1* genes were confirmed in the probands of each family. Moreover, the mutations were segregated with CRC in each family member with CRC manifestation and none of the healthy individuals, except the *MAP3K1* variant. The primers were designed by the Primer3 software (v0.4.0; 29). Sequences of the primers used in this study are listed in Table 2. Conventional polymerase chain reaction (PCR) was performed using Taq DNA Polymerase Master Mix (Ampliqon, Odense, Denmark). PCR was performed at 94°C for 4 min, followed by 32 cycles at 94°C for 30 s, 65°C for *MSH2* and *MAP3K1*, 54°C for *MLH1* for 30 s, and 72°C for 30 s. A final extension step was performed for 5 min at 72°C. The PCR products were assayed using the 3130xl genetic analyzer (Applied Biosystems). We used Sequencing Analysis v5.2 (Applied Biosystems) and FinchTV v1.5.0 (Geospiza Inc.)

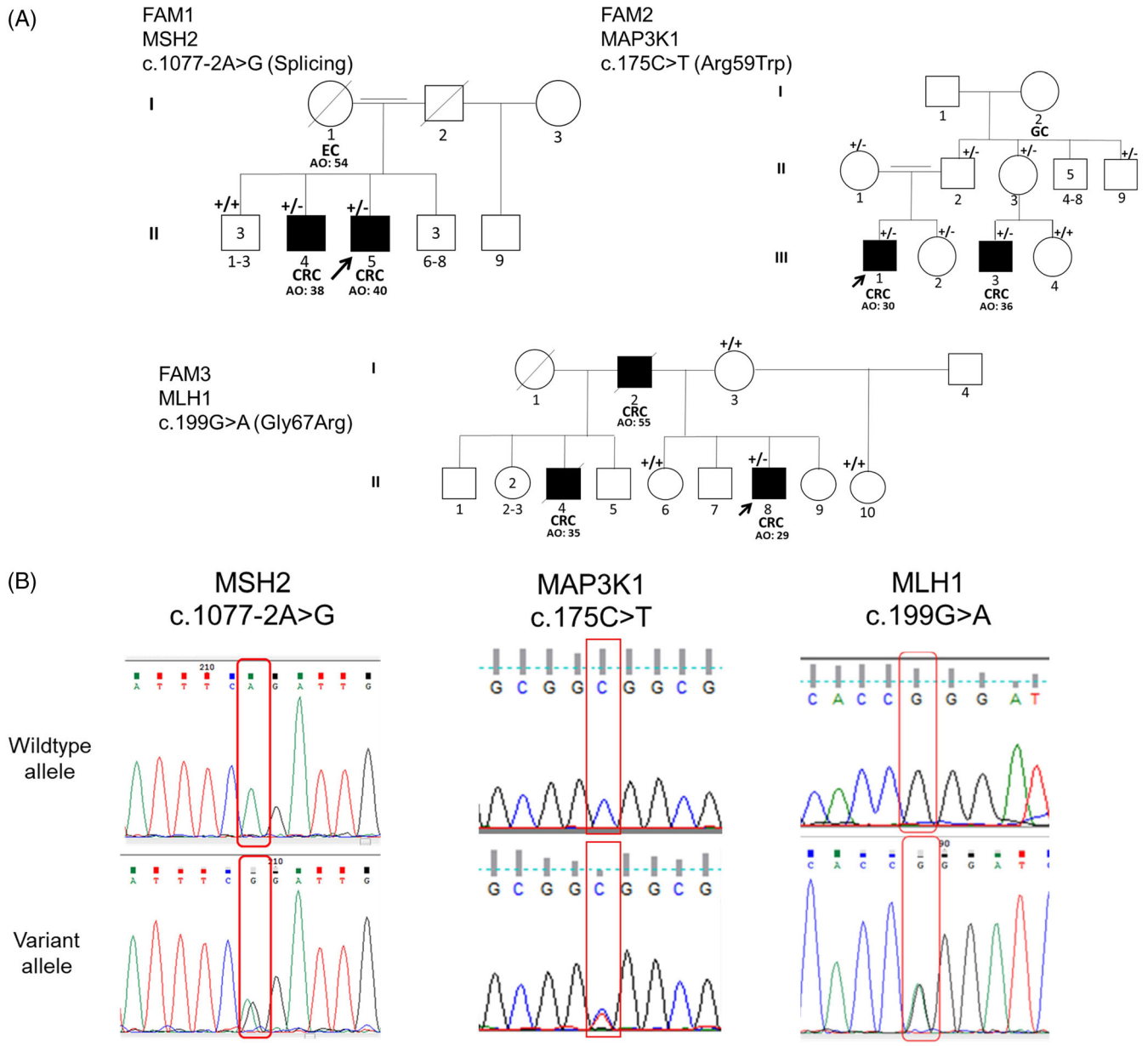

FIGURE 1 *MSH2* (c.1077-2A>G), *MAP3K1* (c.175C>T), and *MLH1* (c.199G>A) variants co-segregating in the FAM1, FAM2, and FAM3 families. (A) The FAM1, FAM2, and FAM3 pedigrees and co-segregation of the three identified variants. (B) Electropherograms showing the variant and wild-type alleles for the three identified variants. Filled squares indicate those affected by CRC. The arrow indicates the proband. (+/−), mutation carrier; (+/+), wild type; AO, age of onset; CRC, colorectal cancer; EC, endometrium cancer; GC, gastric cancer.

to determine the PCR product sequences and validate the candidate mutations.

2.5 | Pathogenic potential prediction for the *MAP3K1* c.175 C>T variant

Since *MAP3K1* c.175 C>T (rs749840532) has yet to be established to have clinical significance in ClinVar (https://www.ncbi.nlm.nih.gov/snp/rs749840532), we used several methods such as the VarSome platform and the AlphaFlod database for protein structure prediction and protein functional domain assessment to predict its pathogenic potential.

3 | RESULTS

3.1 | Clinical data

We enrolled three Iranian families (FAM1, FAM2, and FAM3) with CRC. In FAM1, two brothers with CRC (II-4 and II-5), born from a consanguineous marriage between first cousins (I-1 and I-2), were 40 and 42, respectively. Their medical history revealed that II-4 and II-5 had undergone partial colon resection for CRC in the cecum diagnosed at 38 and 40 years, respectively. The other family members did not report any history of CRC. In another family (FAM2), CRC was diagnosed at the age of 30 in the proband (III-1), who was born from a consanguineous marriage between first cousins (II-1 and II-2).

A total colectomy was performed for cancer of the descending colon in the proband and his first cousin (III-3) at 30 and 36 years, respectively. In FAM3, the proband (II-8) was a 30-year-old man from an inbred family who was subjected to a partial colectomy due to CRC. The affected individuals existed in two generations; the proband's brother (II-4) and his father (I-2) passed away due to CRC at 35 and 55, respectively.

3.2 | Pathogenic variant identification and validation

3.2.1 | *MSH2*: c.1077-2A>G

Genetic sequence analysis detected a heterozygote splice acceptor variation at c.1077-2A>G (NM_000251, rs267607943) in the *MSH2* gene in FAM1 (II-4 and II-5). The c.1077-2A>G in hereditary nonpolyposis colorectal neoplasms has been previously reported as likely pathogenic in ClinVar. The variant was not listed in the gnomAD and the Iranome in-house databases and was predicted to be deleterious by SIFT, probably damaging by PolyPhen, and disease-causing using Mutation Taster prediction tools (Table 3). The genetic counselor's recommendations for the affected individuals (II-4 and II-5), under the National Comprehensive Cancer Network guidelines for LS and specifically for MSH2-LS, were a high-quality colonoscopy every 1–2 years (English version available at: https://www2.tri-kobe.org/nccn/guideline/colorectal/english/genetics_colon.pdf). Genetic testing

TABLE 1 Patient characteristics.

Characteristics	Value
Families	3 (5 patients)
Patient characteristics	
Male	5 patients
Female	-
Age of onset (years)	34.6 (range 29–40)
Phenotype per patient	
CRC	5
Pattern of inheritance	
Dominant	1
Recessive	2

TABLE 2 Primer sequences.

Primer	Sequence (5′-3′)
MSH2 (F)	TGTACAGACGGGGTTTCTCC
MSH2 (R)	CGGTTAAGATCTGGGAATCG
MAP3K1 (F)	AGCCCGCGAGAGAAAATGGC
MAP3K1 (R)	CACCGCCACAGGCTGGAAG
MLH1 (F)	GCCAGTTTAGATGCAAAATCC
MLH1 (R)	ACACATCCCCTGAACAGTGC

TABLE 3 Overview of three variants identified by WES and in silico pathogenicity analysis.

GENE	ClinVar	HGMD	HGVS	Variant type	Protein change	gnomAD	CADD	REVEL	PolyPhen	SIFT	Mutation taster
MSH2	Likely_pathogenic	Colorectal cancer, non-polyposis	NM_000251: c.1077-2A>G	Splice acceptor	Splicing	No data	35	–	Probably damaging	Deleterious	Disease-causing
MAP3K1	NA	–	NM_005921:c.175C>T	Missense	Arg59Trp	0.0002915	28.1	0.22	Possibly damaging	Deleterious	Disease-causing
MLH1	Pathogenic	Colorectal cancer, non-polyposis	NM_000249:c.199G>A	Missense	Gly67Arg	No data	31	0.988	Probably damaging	Deleterious	Disease-causing

Abbreviations: CADD, combined annotation dependent depletion; ClinVar, clinical significance of the variation; gnomAD, the Genome Aggregation Database; HGMD, the human gene mutation database; HGVS, the human genome variation society; MutationTaster, a free web-based application to evaluate DNA sequence variants for their disease-causing potential; NA, not available; PolyPhen, polymorphism phenotyping; REVEL, rare exome variant ensemble learner; SIFT, sorting intolerant from tolerant; predicting amino acid changes that affect protein function.

for the likely pathogenic variant c.1077-2A>G in the *MSH2* gene was recommended for the I and II-degree relatives of the proband.

3.2.2 | *MAP3K1*: c.175C>T

A rare heterozygote variant at c.175C>T (NM_005921, Arg59Trp, rs749840532) in the *MAP3K1* gene was identified in III-1 and III-3 of FAM2. The mutation is predicted to be deleterious by SIFT, possibly damaging by PolyPhen, and disease-causing using the Mutation Taster prediction tools. An association between the c.175C>T variation and CRC has yet to be reported in clinical databases such as OMIM and ClinVar (Table 3). In addition, we validated the *MAP3K1* mutation status of formalin-fixed paraffin-embedded (FFPE) tissue samples for patient III-1 with the c.175C>T heterozygote mutation.

3.2.3 | *MLH1*: c.199G>A

We detected a heterozygote variant at c.199G>A (NM_000249, Gly67Arg, rs63750206) in the *MLH1* gene in the proband (II-8) of FAM3. No frequency for this variant was reported in the genomeAD-

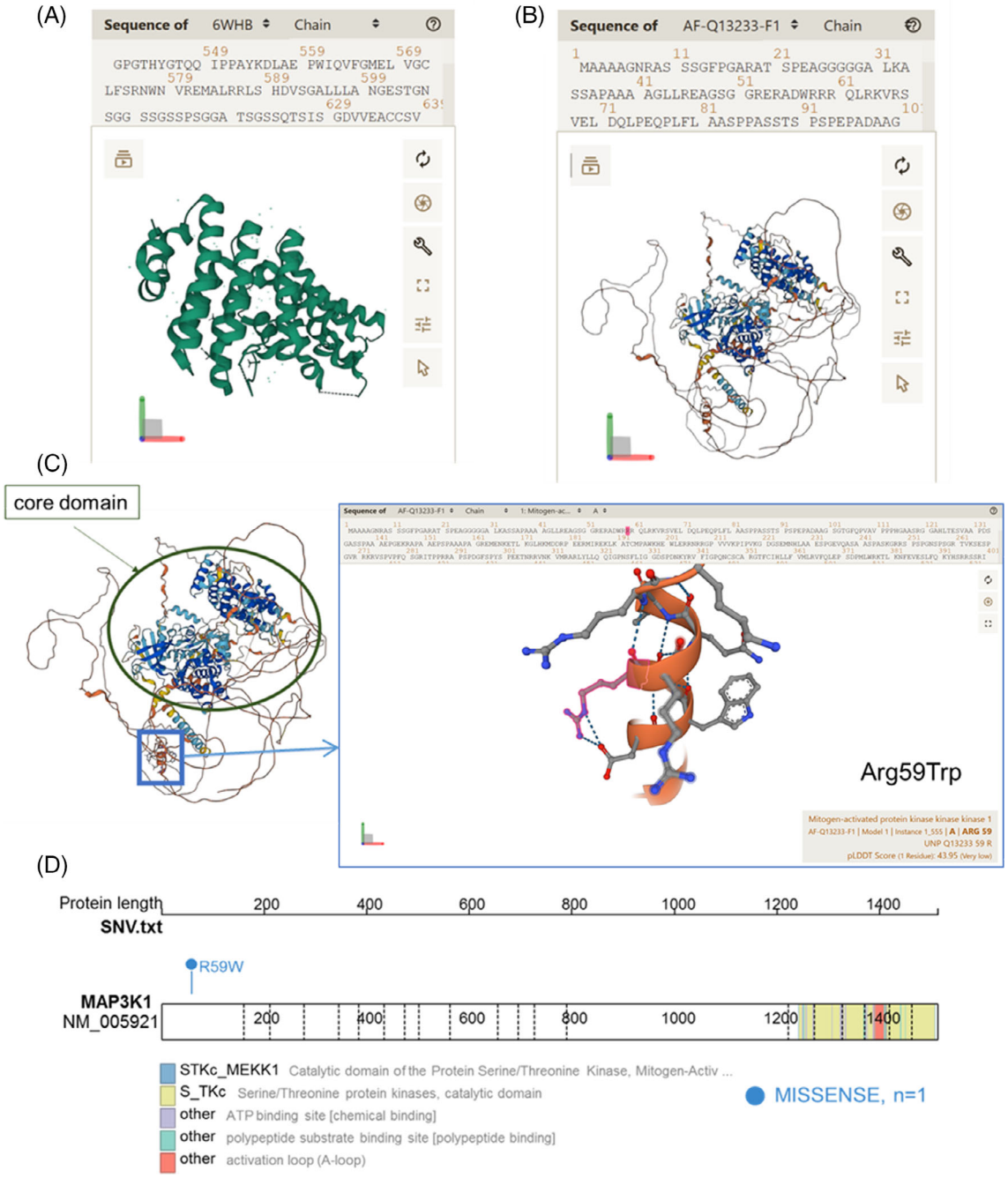

FIGURE 2 Three-dimensional structure from PDB (A: representative) and AlphaFold (B: predicted) for the *MAP3K1* gene. (C) AlphaFold prediction of the position of variant (Arg59Trp) in MAP3K1 protein. (D) The core functional domain is local at the tail of MAP3K1, and the blue pin shows the mentioned variant (R59W).

genome East Asian and Iranome in-house databases. Moreover, this variant is predicted to be deleterious by SIFT, probably damaging by PolyPhen, and disease-causing using the Mutation Taster prediction tools. In addition, the c.199G>A, Gly67Arg in hereditary nonpolyposis colorectal neoplasms has previously been reported as pathogenic in ClinVar (Table 3). The I and II-degree relatives of the proband were recommended genetic counseling and testing for the pathogenic variant c.199G>A in the *MLH1* gene, like *MSH2*.

Further, genetic analysis of members of the three families showed segregation of the mutations with disease status (Figure 1B). These observations suggest that NM_000251 (*MSH2*): c.1077-2A>G, NM_005921 (*MAP3K1*): c.175C>T, and NM_000249 (*MLH1*): c.199G>A mutations could be the possible causal factor of EOCRC in these families.

3.3 | Structural modeling of the c.175C>T variant of *MAP3K1*

The varsome.com website performs ACMG classification based on several prediction tools, variant frequency data, and clinical data. For the c.175C>T variant (rs749840532), the prediction is "Likely Benign." However, only one experimental structure is available for MAP3K1 in the PDB database (Access: 6WHB). However, this variant (NP_005912: p. Arg59Trp) is not present in the core domain structure of 6WHB (the amino acid of the 6WHB structure starts from 539 and extends to 888) (Figure 2A). We used the predicted structure from AlphaFold (Figure 2B) and found that the variant (Arg59Trp) was distant from the core domain (Figure 2C). The core functional domain is local at the tail of MAP3K1, and the variant (R59W) is not in a functional domain (Figure 2D).

4 | DISCUSSION

As CRC rarely develops before age 40, genetic predisposition is suspected. The most common familial CRC syndrome is LS, which is genetically heterogeneous autosomal dominant and is characterized by early onset of CRC and neoplasia in other organs.[21] This study conducted a gene panel analysis of 724 cancer-related genes in patients ($n = 5$) with EOCRC from three different families.

We identified three high-risk pathogenic variants in *MSH2*, *MAP3K1*, and *MLH1*. The likely pathogenic *MSH2* variant (NM_000251: c.1077-2A>G), and the pathogenic *MLH1* variant (NM_000249: c.199G>A, p. Gly67Arg), had previously been identified; however, we report these variants as possible causative agents of EOCRC in the Iranian population for the first time.

The *MSH2* gene causes LS (MIM #120435). The incidence of LS is estimated to be 1/226 in the general population,[22] and approximately 40% of cases are caused by germline mutations in *MSH2*.[23] The *MSH2* variant correlates with the second highest risk of CRC and is only slightly lower than that of *MLH1*. Among individuals over 75, 46.6% of women and 51.4% of men are affected by this variation.[24]

Bogomilova et al. used next-generation sequencing to identify a splice site mutation in *MSH2* (c.1386+1G>A) in subjects with very early onset CRC.[25] In another study,[26] a novel pathogenic *MSH2* mutation (c.1552C>T; p.Q518X) was detected in a 33-year-old Tunisian patient with CRC. The *MLH1* variant correlates with the highest risk of developing CRC,[27] with rates ranging from 0% (age 30) to 48.3% (age 75) in females and from 4.5% (age 30) to 57.1% (age 75) in males.[24] Hesson et al. showed two MLH1 5′UTR variants (c.-28A>G and c.-7C>T) associated with EOCRC.[28] Another study identified a heterozygous novel mutation (c.206delG) in the *MLH1* gene in an extended family with LS. The proband of this study was a 40-year-old male with early onset CRC.[29] Furthermore, in one cohort study, *MLH1* was reported as the most frequently mutated gene in early-onset sporadic CRC patients.[30] As a result of this study, four pathogenic variations were identified in *MLH1*: c.C793T (p.R265C), c.C1029A (p.Y343X), c.C793T (p.R265C), and c.C1029A (p.Y343X).[30]

Mitogen-activated protein kinase (MAP3K1/MEKK1) can induce anti-apoptotic signaling. As a result of caspase cleavage, a 91-kDa fragment generated by MAP3K1 induces apoptosis, suggesting that MAP3K1 acts as a molecular switch to regulate apoptosis.[31] Due to its dual roles in cell survival and apoptosis, the significant number of *MAP3K1* mutations in various types of cancer, including CRC, remains poorly understood.[32] This study classified the identified variation of *MAP3K1* (NM_005921: c.175C>T) into the CADD20up group. In addition, various prediction tools predict this variant to be deleterious, possibly damaging, and disease-causing. After further reviewing the pathogenic potential of variant *MAP3K1* (NM_005921: exon1: c.175C>T, NP_005912: p. Arg59Trp), we suggest reclassifying the variant as "likely pathogenic" rather than a "variant of uncertain significance" (VUS). The reason for this reclassification is that the CADD prediction score is 28.1, suggesting a moderate level of pathogenic potential, while the REVEL score is 0.22, indicating a weaker level of pathogenic potential. Given these discrepant results, further studies are warranted to clarify the role of this variant in the development of CRC. Furthermore, the incomplete co-segregation result for FAM2 suggests that the c.175C>T variant does not lead to the CRC phenotype in other carriers. However, most individuals in generation II (II-2 and his brothers) show digestive problems, including rectal bleeding, severe constipation, and even fissure surgery. Reduced penetrance likely underlies this issue. Reduced penetrance results from the modulation of differential allele expression, copy number variation, or the effect of additional genetic variants in cis or trans. Age and/or sex influence the penetrance of some pathogenic genotypes. Other factors may also affect penetrance, such as unlinked modifier genes, epigenetic changes, and environmental factors.[33] The clinical importance of the *MAP3K1* variant in CRC is currently debated; however, it can become an influential factor in CRC. Moreover, we analyzed beyond the 724 genes but found no pathogenic or likely pathogenic mutations in other known cancer-related genes. Therefore, we speculate that this family likely carried other polygenic factors, but the *MAPK3K1* c.175C>T mutation played a major role. Su et al. conducted a comprehensive analysis of the genomic profile and polygenic risk score (PRS) in Taiwanese patients with CRC. The study found a significant

difference in driver gene mutation rates between Taiwanese individuals and those of white ethnicity.[34] *MAP3K1* mutation found in the Iranian population was not found in Taiwanese patients with CRC. To date, there have been limited genome-wide association studies on CRC in Iran, and the analysis of PRS may be insufficient for our study.

A VUS with a high potential of being reclassified to likely pathogenic is the *MSH2* variant, c.2168C>T, p. (Ser723Phe). The variant has already been detected in a family[35] but was considered to be a VUS. Meanwhile, a suggested functional effect[36–39] combined with the co-segregation results in two individuals with early onset CRC[40] indicates that *MSH2* c.2168C>T interrupts normal MMR functions and is considered likely pathogenic.

Further, more extensive studies are needed to elucidate the genetic background of hereditary and/or early-onset CRC and identify novel cancer genes. Therefore, coordinated efforts must be made to obtain robust results in the era of Open Data Science, particularly for rare and molecularly heterogeneous cases within the CRC spectrum. Additionally, it is hoped that functional data, such as CRISPR assays or in vitro organoid models, will complement data interpretation, increasing the functional screening throughput of candidate genes and providing invaluable data to validate novel susceptibility loci for CRC.

Identifying novel susceptibility variants associated with early-onset familial CRC may be possible by sequencing these patients. This study identified only two known likely pathogenic and pathogenic variants, *MSH2* c.1077-2A>G and *MLH1* c.199G>A. In the remaining family, the most promising candidates are *MAP3K1* c.175C>T, although this does not fully co-segregate with the family phenotypes. Early development of this disease may be caused by multiple and variable rare, moderately-penetrant variants, illustrating that CRC susceptibility is oligogenic and heterogeneous. Further, more extensive, and appropriately powered studies are needed to clarify the underlying genetic causes. For example, gene prioritization algorithms could be refined using complementary approaches such as interconnected omics sources and high-throughput functional studies.

CONFLICT OF INTEREST STATEMENT

All authors declare no conflict of interest.

ORCID

Jan-Gowth Chang https://orcid.org/0000-0003-0375-1427

REFERENCE

1. de la Chapelle A. The incidence of Lynch syndrome. Fam Cancer. 2005;4(3):233–7.
2. Rawla P, Sunkara T, Barsouk A. Epidemiology of colorectal cancer: incidence, mortality, survival, and risk factors. Prz Gastroenterol. 2019;14(2):89–103.
3. Erratum: Global cancer statistics 2018. GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2020;70(4):313.
4. Valle L. Recent discoveries in the genetics of familial colorectal cancer and polyposis. Clin Gastroenterol Hepatol. 2017;15(6):809–19.
5. Pearlman R, Frankel WL, Swanson B, Zhao W, Yilmaz A, Miller K, et al. Prevalence and Spectrum of germline cancer susceptibility gene mutations among patients with early-onset colorectal cancer. JAMA Oncol. 2017;3(4):464–71.
6. Valle L, Vilar E, Tavtigian SV, Stoffel EM. Genetic predisposition to colorectal cancer: syndromes, genes, classification of genetic variants and implications for precision medicine. J Pathol. 2019;247(5):574–88.
7. Babaei K, Khaksar R, Zeinali T, Hemmati H, Bandegi A, Samidoust P, et al. Epigenetic profiling of MUTYH, KLF6, WNT1 and KLF4 genes in carcinogenesis and tumorigenesis of colorectal cancer. Biomedicine (Taipei). 2019;9(4):22.
8. Dominguez-Valentin M, Nakken S, Tubeuf H, Vodak D, Ekstrom PO, Nissen AM, et al. Identification of genetic variants for clinical management of familial colorectal tumors. BMC Med Genet. 2018;19(1):26.
9. Stoffel EM, Koeppe E, Everett J, Ulintz P, Kiel M, Osborne J, et al. Germline genetic features of young individuals with colorectal cancer. Gastroenterology. 2018;154(4):897–905.e1.
10. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100.
11. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491–8.
12. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29(1):308–11.
13. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68–74.
14. Chang YS, Chao DS, Chung CC, Chou YP, Chang CM, Lin CL, et al. Cancer carrier screening in the general population using whole-genome sequencing. Cancer Med. 2023;12(2):1972–83.
15. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164.
16. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434–43.
17. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46(D1):D1062–7.
18. Stenson PD, Mort M, Ball EV, Chapman M, Evans K, Azevedo L, et al. The human gene mutation database (HGMD((R))): optimizing its use in a clinical diagnostic or research setting. Hum Genet. 2020;139(10):1197–207.
19. Rentzsch P, Schubach M, Shendure J, Kircher M. CADD-splice-improving genome-wide variant effect prediction using deep learning-derived splice scores. Genome Med. 2021;13(1):31.
20. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. Am J Hum Genet. 2016;99(4):877–85.
21. Kastrinos F, Mukherjee B, Tayob N, Wang F, Sparr J, Raymond VM, et al. Risk of pancreatic cancer in families with Lynch syndrome. JAMA. 2009;302(16):1790–5.
22. Haraldsdottir S, Rafnar T, Frankel WL, Einarsdottir S, Sigurdsson A, Hampel H, et al. Comprehensive population-wide analysis of Lynch syndrome in Iceland reveals founder mutations in MSH6 and PMS2. Nat Commun. 2017;8:14755.
23. Moller P, Seppala T, Bernstein I, Holinski-Feder E, Sala P, Evans DG, et al. Incidence of and survival after subsequent cancers in carriers of pathogenic MMR variants with previous cancer: a report from the prospective Lynch syndrome database. Gut. 2017;66(9):1657–64.
24. Dominguez-Valentin M, Sampson JR, Seppala TT, Ten Broeke SW, Plazzer JP, Nakken S, et al. Cancer risks by gene, age, and gender in 6350 carriers of pathogenic mismatch repair variants: findings from

the prospective Lynch syndrome database. Genet Med. 2020;22(1):15–25.
25. Bogomilova Kamburova Z, Lubenova Popovska S, Stefanova Kovacheva K, Todorov Petrov K, Enkova Nikolova S. Familial Lynch syndrome with early age of onset and confirmed splice site mutation in MSH2: a case report. Biomed Rep. 2022;16(5):39.
26. Jaballah-Gabteni A, Tounsi H, Kabbage M, Hamdi Y, Elouej S, Ben Ayed I, et al. Identification of novel pathogenic MSH2 mutation and new DNA repair genes variants: investigation of a Tunisian Lynch syndrome family with discordant twins. J Transl Med. 2019;17(1):212.
27. Win AK, Jenkins MA, Dowty JG, Antoniou AC, Lee A, Giles GG, et al. Prevalence and penetrance of major genes and polygenes for colorectal cancer. Cancer Epidemiol Biomarkers Prev. 2017;26(3):404–12.
28. Hesson LB, Packham D, Kwok CT, Nunez AC, Ng B, Schmidt C, et al. Lynch syndrome associated with two MLH1 promoter variants and allelic imbalance of MLH1 expression. Hum Mutat. 2015;36(6):622–30.
29. Ghaedi H, Ramsheh SM, Omidvar ME, Labbaf A, Alehabib E, Akbari S, et al. Whole-exome sequencing identified a novel mutation of MLH1 in an extended family with lynch syndrome. Genes Dis. 2019;7(4):614–9.
30. Chang PY, Chang SC, Wang MC, Chen JS, Tsai WS, You JF, et al. Pathogenic germline mutations of DNA repair pathway components in early-onset sporadic colorectal polyp and cancer patients. Cancers (Basel). 2020;12(12):3560.
31. Pham TT, Angus SP, Johnson GL. MAP3K1: genomic alterations in cancer and function in promoting cell survival or apoptosis. Genes Cancer. 2013;4(11–12):419–26.
32. Xue Z, Vis DJ, Bruna A, Sustic T, van Wageningen S, Batra AS, et al. MAP3K1 and MAP2K4 mutations are associated with sensitivity to MEK inhibitors in multiple cancer models. Cell Res. 2018;28(7):719–29.
33. Cooper DN, Krawczak M, Polychronakos C, Tyler-Smith C, Kehrer-Sawatzki H. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. Hum Genet. 2013;132(10):1077–130.
34. Su MW, Chang CK, Lin CW, Chu HW, Tsai TN, Su WC, et al. Genomic and metabolomic landscape of right-sided and left-sided colorectal cancer: potential preventive biomarkers. Cell. 2022;11(3):527.
35. Nilbert M, Wikman FP, Hansen TV, Krarup HB, Orntoft TF, Nielsen FC, et al. Major contribution from recurrent alterations and MSH6 mutations in the Danish Lynch syndrome population. Fam Cancer. 2009;8(1):75–83.
36. Gammie AE, Erdeniz N, Beaver J, Devlin B, Nanji A, Rose MD. Functional characterization of pathogenic human MSH2 missense mutations in Saccharomyces cerevisiae. Genetics. 2007;177(2):707–21.
37. Drost M, Zonneveld JB, van Hees S, Rasmussen LJ, Hofstra RM, de Wind N. A rapid and cell-free assay to test the activity of lynch syndrome-associated MSH2 and MSH6 missense variants. Hum Mutat. 2012;33(3):488–94.
38. Houlleberghs H, Dekker M, Lantermans H, Kleinendorst R, Dubbink HJ, Hofstra RM, et al. Oligonucleotide-directed mutagenesis screen to identify pathogenic Lynch syndrome-associated MSH2 DNA mismatch repair gene variants. Proc Natl Acad Sci U S A. 2016;113(15):4128–33.
39. Rath A, Mishra A, Ferreira VD, Hu C, Omerza G, Kelly K, et al. Functional interrogation of Lynch syndrome-associated MSH2 missense variants via CRISPR-Cas9 gene editing in human embryonic stem cells. Hum Mutat. 2019;40(11):2044–56.
40. Djursby M, Madsen MB, Frederiksen JH, Berchtold LA, Therkildsen C, Willemoe GL, et al. New pathogenic germline variants in very early onset and familial colorectal cancer patients. Front Genet. 2020;11:566266.

How to cite this article: Fatemi N, Tu S-J, Chung C-C, Moghadam PK, Mojarad EN, Sadeghi A, et al. Whole exome sequencing identifies *MAP3K1*, *MSH2*, and *MLH1* as potential cancer-predisposing genes in familial early-onset colorectal cancer. Kaohsiung J Med Sci. 2023;39(9):896–903. https://doi.org/10.1002/kjm2.12715