Noncoding Aberrations in Mismatch Repair Genes Underlie a Substantial Part of the Missing Heritability in Lynch Syndrome

ERN-GENTURIS Lynch-Like Working Group; Te Paske, Iris B.A.W.; Mensenkamp, Arjen R.; Neveling, Kornelia; Hoogerbrugge, Nicoline; Ligtenberg, Marjolijn J.L.; De Voer, Richarda M.

Published in:
Gastroenterology

DOI:
10.1053/j.gastro.2022.08.041

Publication date:
2022

Noncoding Aberrations in Mismatch Repair Genes Underlie a Substantial Part of the Missing Heritability in Lynch Syndrome

Lynch syndrome (LS) is characterized by the development of mismatch repair–deficient (dMMR) colorectal and endometrial cancers at a young age in life. LS is caused by germline pathogenic variants (PVs) in 1 of the MMR genes *MLH1*, *MSH2*, *MSH6*, or *PMS2* or deletions affecting the 3′ region of *EPCAM*.[1] Current germline diagnostics for LS include targeted short-read sequencing and multiplex ligation-dependent probe amplification of the coding regions of the MMR genes, after the exclusion of somatic *MLH1*-promoter hypermethylation. In the absence of a germline PV in an MMR gene, the presence of somatic dMMR is investigated. However, a proportion of individuals with dMMR tumors remain genetically unresolved after germline and somatic analyses. These individuals have an unexplained dMMR tumor and are known as individuals with Lynch-like syndrome (LLS) (Figure 1*A*).[2,3] For individuals with LLS and their relatives, treatment options and surveillance are yet unclear. Here, by applying targeted long-read sequencing of the MMR genes, we show that a substantial proportion of individuals with LLS actually can be diagnosed with LS, because they carry deep intronic MMR gene aberrations that result in aberrant splicing (for further reading on aberrant splicing, see Lord et al.[4]).

The study cohort consisted of 32 individuals diagnosed with an unexplained dMMR cancer (Figure 1*A*) aged ≤70 years with (n = 18) or without (n = 14) a familial history of colorectal or endometrial cancer. For each individual, the MMR gene(s) of interest (Supplementary Table 1) was amplified using long-range polymerase chain reaction amplicons of 7–18 kb in size (Supplementary Figure 1*A*) on germline DNA and sequenced using Single Molecule Real-Time Sequencing (Pacific Biosciences). For details on sequencing and annotation see Supplementary Methods. This study (CMO-2018-4922) was approved by the Radboudumc Ethical Committee.

Nine different noncoding aberrations in 9 of 32 individuals (28.1%) were identified (Figure 1*B*). Five different deep intronic single nucleotide variants in *MSH2* (n = 3), *MLH1* (n = 1), and *PMS2* (n = 1) that are likely to introduce novel splice sites based on in silico predictions were identified in 6 individuals. Additionally, 2 intronic *Alu* element insertions (>96% similarity to *Alu* elements) in introns 1 and 8 of *MSH2* (of which the latter was in *cis* with a deep-intronic single nucleotide variant), 1 1704-bp intronic deletion in intron 3 of *MLH1*, and 1 tandem duplication in the *MLH1* promoter region were identified. None of the variants was previously reported in population databases (GnomAD-G, 1000G). We performed co-segregation analyses where possible and analyzed the effects of 7 potentially pathogenic noncoding variants in a mini-gene assay to assess altered splicing and the effect of the duplication in the *MLH1* promoter region using a luciferase reporter assay.

Individual LLP004 carried 2 deep intronic *MSH2* variants: c.793-603C>T and c.2458+976A>G. Co-segregation analysis showed that c.793-603C>T did not co-segregate with cancer phenotypes within the family and therefore was excluded for functional analysis, whereas c.2458+976A>G did segregate in 2 affected cousins and an affected aunt (Figure 1*C*, family A). Quantitative analysis of *MSH2* mRNA from a lymphoblastoid cell line from an affected cousin showed nonsense-mediated decay and decreased *MSH2* expression compared with healthy control samples (Supplementary Figure 1*D*). The same *MSH2* c.2458+976A>G variant was also found in individual LLP009, who is as far as we could determine not related to LLP004. Mini-gene analysis of c.2458+976A>G showed alternative splicing because of the generation of a new splice donor site and activation of an intronic splice acceptor site leading to a premature stop codon at p.(Gly820Glu*fs**47).

Individual LLP031 carried an *Alu* element insertion (c.1387-3546_1387-3545ins351) and a deep intronic variant (c.2459-954A>G) *in cis* in *MSH2*. Both variants co-segregated in the sister and the mother of the index (Figure 1*C*, family B). Mini-gene analysis of *MSH2* c.1387-3546_1387-3545ins351 showed predominant expression of the wild-type transcript but also some transcripts of alternative size (Supplementary Figure 1*B*). However, because the mini-gene analysis of the c.2459-954A>G variant showed the inclusion of a pseudoexon (68 bp) between exons 14 and 15 that led to a premature stop codon at p.(Gly820Glu*fs**44), we considered the latter as the PV in this individual.

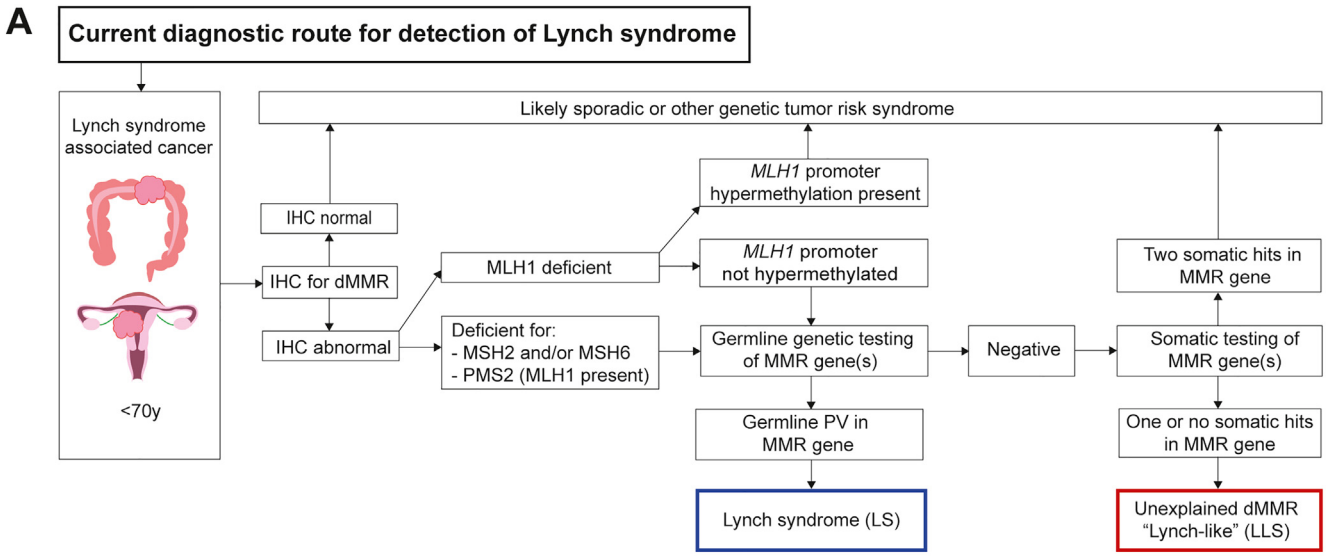
Co-segregation analysis was not possible for the other LLS individuals. Therefore, the remaining potential PVs were only analyzed by mini-gene analysis or a luciferase reporter assay. Mini-gene analysis of the *Alu* element insertion affecting the splice acceptor consensus sequence of *MSH2* exon 2 (c.212-4_213-3ins366; individual LLP024) and the deep intronic variants *MLH1* c.306+1001_307-642delinsTA (individual LLP032) and *MLH1* c.306+1070C>G (individual LLP002) induced altered splicing compared with wild-type and all resulted in predicted premature stop codons (Figure 1*D*, Supplementary Figure 1*B* and *C*).

Abbreviations used in this paper: **dMMR, mismatch repair deficient; LS, Lynch syndrome; LLS, Lynch-like syndrome; MMR, mismatch repair; PV, pathogenic variant.**

Most current article

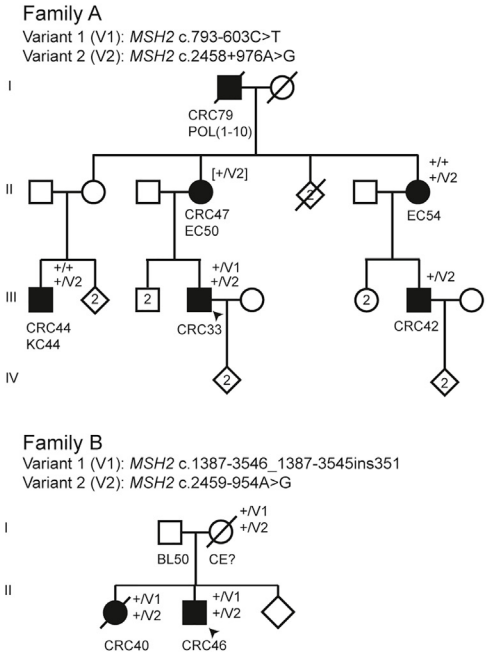
0016-5085
https://doi.org/10.1053/j.gastro.2022.08.041

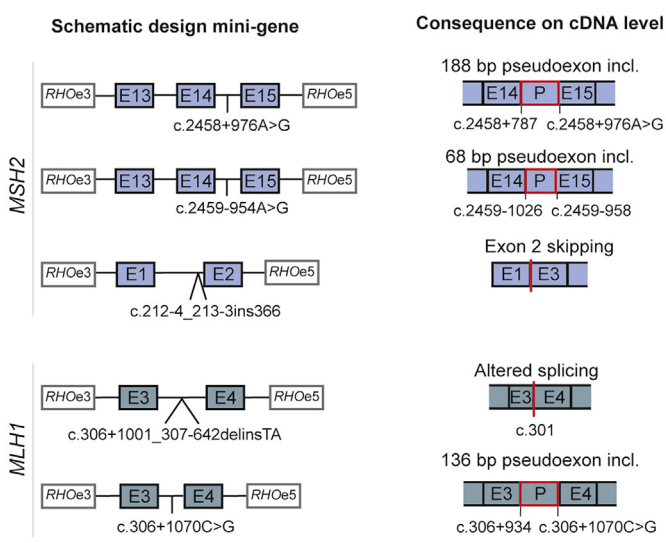
The effect of the 48-bp tandem duplication in the *MLH1* promoter in individual LLP025 (*MLH1* c.-404_-357dup) was assessed using a luciferase reporter assay (Supplementary Methods).[5] However, the *MLH1* c.[-404_-357dup;-93G>A] haplotype of the individual showed a similar level of promoter activity compared with the wild-type *MLH1* promoter (Supplementary Figure 1*E*). In addition, germline *MLH1* promoter hypermethylation was not observed. Therefore,



B

Gene	Individual	Gender	Diagnosis and age	FH CRC/EC	Germline variant found in this study	Disease mechanism based on study
MSH2	LLP004	M	CRC33	Y	c.793-603C>T	Likely not pathogenic
					c.2458+976A>G	Pseudoexon inclusion and premature stop
	LLP009	M	UC38, CRC44	N	c.2458+976A>G	Pseudoexon inclusion and premature stop
	LLP031	M	CRC46	Y	c.1387-3546_1387-3545ins351	Variant of unknown significance
					c.2459-954A>G	Pseudoexon inclusion and premature stop
	LLP024	F	CRC32	U	c.212-4_212-3ins366	Exon skipping and premature stop
MLH1	LLP032	M	CRC43	U	c.306+1001_307-642delinsTA	Altered splicing and premature stop
	LLP002	F	KC53, CRC58	U	c.306+1070C>G	Pseudoexon inclusion and premature stop
	LLP025	M	CRC50	Y	c.-404_-357dup	Variant of unknown significance
PMS2	LLP014	F	CRC31	Y	c.2276-400G>C	Likely not pathogenic
	LLP029	M	BDC61	Y	c.2276-400G>C	Likely not pathogenic





RESEARCH LETTERS

this duplication is currently considered as a variant of unknown significance. Mini-gene analysis of the deep intronic *PMS2* c.2276-400G>C variant, identified in individuals LLP014 and LLP029, did not indicate altered splicing compared with wild-type.

Until now, deep intronic PVs, *MLH1* promoter variants, and *Alu*-mediated structural variants have only been reported in isolated cases or in very small proportions of LLS cohorts.[5–9] However, our analyses indicate that a substantial proportion of individuals with LLS (18.8%; 6/32 individuals) can be diagnosed with LS because of a germline noncoding pathogenic aberration in an MMR gene. These findings warrant the expansion of current diagnostic short-read–sequencing panels for known pathogenic intronic variants to increase the diagnostic yield for LS. Moreover, for individuals with LLS who remain without a genetic diagnosis after both germline and somatic routine diagnostic analyses, it is of interest to perform germline sequencing of the complete MMR gene loci. In such an approach, long-read sequencing may facilitate the detection of deep intronic variants not covered by current diagnostic panels and allow the detection of *Alu* or tandem repeats, which are very difficult to detect using short-read–sequencing technologies. Together, analyses of the intronic regions of the MMR genes further optimizes LS diagnostics and consequently improves treatment and cancer surveillance in patients and relatives.

Supplementary Material

Note: To access the supplementary material accompanying this article, visit the online version of *Gastroenterology* at www.gastrojournal.org and at https://doi.org/10.1053/j.gastro.2022.08.041.

IRIS B.A.W. TE PASKE
Department of Human Genetics
Radboud University Medical Center
Radboud Institute for Molecular Life Sciences
Nijmegen, the Netherlands

ARJEN R. MENSENKAMP
Department of Human Genetics
Radboud University Medical Center
Radboud Institute for Molecular Life Sciences
Nijmegen, the Netherlands

KORNELIA NEVELING
Department of Human Genetics
Radboud University Medical Center
Donders Institute for Brain, Cognition and Behaviour
Nijmegen, the Netherlands

ERN-GENTURIS LYNCH-LIKE WORKING GROUP

NICOLINE HOOGERBRUGGE
Department of Human Genetics
Radboud University Medical Center
Radboud Institute for Molecular Life Sciences
Nijmegen, the Netherlands

MARJOLIJN J.L. LIGTENBERG
Department of Human Genetics
Department of Pathology
Radboud University Medical Center
Radboud Institute for Molecular Life Sciences
Nijmegen, the Netherlands

RICHARDA M. DE VOER
Department of Human Genetics
Radboud University Medical Center
Radboud Institute for Molecular Life Sciences
Nijmegen, the Netherlands

References

1. Lynch HT, et al. Clin Genet 2009;76:1–18.
2. Elze L, et al. Gastroenterology 2021;160:1414–1416.

Figure 1. Analysis of noncoding aberrations in individuals with Lynch-like syndrome. (*A*) Flowchart of the current diagnostic route for the detection of LS. Individuals with Lynch-associated MMR deficient cancer (CRC and/or endometrium cancer [EC] in individuals aged < 70 years) without somatic *MLH1* promoter methylation are selected for germline MMR testing. Individuals without a detectable pathogenic germline variant or 2 somatic variants in the MMR genes are considered to have LLS. IHC, Immunohistochemistry. (*B*) Characteristics of individuals in which a potential pathogenic noncoding variant was found that is likely to lead to aberrant splicing based on in silico predictions or in which a structural variant is found. Variant positions are based on reference transcripts NM_000249.4 (*MLH1*), NM_000251.3 (*MSH2*), NM_000535.7 (*PMS2*) on GRCH37/hg19. BDC, bile duct cancer; F, female; FH, family history; KC, kidney cancer; M, male; N, no; UC, urothelial cancer; U, unknown; Y, yes. (*C*) Pedigrees of family A and family B. In family A *MSH2* c.2458+976A>G co-segregated with the affected disease status. Index individuals are indicated with *black arrowheads*. Individuals affected with EC and CRC are visualized by *filled figures.* BL, bladder cancer; CE, cervical cancer; POL, polyposis; ?, unknown age at diagnosis. (*D*) Schematic representation of the molecular consequence per noncoding variant based on cDNA sequence analysis. Overview of mini-gene constructs (*left*) and consequence of variant on complementary DNA level (*right*). Only variants that resulted in aberrant splicing are displayed. *MSH2* c.2458+976A>G, *MSH2* c.2459-954A>G, and *MLH1* c.306+1070C>G result in inclusion of a pseudoexon. In *MLH1* c.306+1001_307-642delinsTA aberrant splicing was observed compared with wild-type because of the use of an alternative splice site at position c.301. The *MSH2* c.212-4_213-3ins366 mutant mini-gene showed skipping of exon 2 compared to wild-type. *Red* highlights the change. For details see also Supplementary Figure 1. E, exon; P, pseudoexon; *RHOe*3, *RHO* exon 3; *RHOe*5, *RHO* exon 5.

3. **Mensenkamp AR**, **Vogelaar IP**, et al. Gastroenterology 2014;146:643–646.
4. Lord J, et al. Front Genet 2021;12:689892.
5. Hitchins MP, et al. Cancer Cell 2011;20:200–213.
6. Clendenning M, et al. Fam Cancer 2011;10:297–301.
7. Li L, et al. Hum Mutat 2006;27:388.
8. Ward RL, et al. Genet Med 2013;15:25–35.
9. Arnold AM, et al. Eur J Hum Genet 2020;28:597–608.

Author names in bold designate shared co-first authorship.

Received June 29, 2022. Accepted August 22, 2022.

Correspondence
Address correspondence to: Richarda M. de Voer, PhD, Department of Human Genetics, Radboud University Medical Center, Geert Grooteplein Zuid 10, 6525GA Nijmegen, the Netherlands. e-mail: Richarda.devoer@radboudumc.nl.

Acknowledgments
The authors thank all study participants. The authors thank Neeltje Arts, Edris Askar, Anouk Bertram, Linske de Bruijn, Ronny Derks, Michael Kwint, Mieke Lutje-Berenbroek, Luke O'Gorman, Bruce Poppe, Robin de Putter, Hanneke Volleberg-Gorissen, Marcel Nelen, and Lisenka Vissers for their contributions to this project. Furthermore, the authors want to thank the Radboudumc Genome Technology Center for infrastructural and computational support.

The ERN-GENTURIS Lynch-like Working Group (alphabetical order, grouped per hospital) includes Stéphanie Baert-Desurmont,[1] Kathleen B. M. Claes,[2] Kim de Leeneer,[2] Lisa Elze,[3] Simone van den Heuvel,[3] Rachel S. van der Post,[4] Yvonne van Twuijver,[3] Tjakko J. van Ham,[5] Anja Wagner,[5] Mirjam M. de Jong,[6] Edward M. Leter,[7] Maartje Nielsen,[8] from the [1]Department of Genetics, Normandy Center for Genomic and Personalized Medicine, UNIROUEN, Inserm U1245 and Rouen University Hospital, Rouen, France; [2]Centre for Medical Genetics, Ghent University Hospital, Department of Biomolecular Medicine, Ghent University, Medical Genetics, Ghent, Belgium; [3]Department of Human Genetics, Radboud University Medical Center, Nijmegen, the Netherlands; [4]Department of Pathology, Radboud University Medical Center, Radboud Institute for Molecular Life Sciences, Nijmegen, the Netherlands; [5]Department of Clinical Genetics, Erasmus University Medical Center, Erasmus MC Cancer Institute, Rotterdam, the Netherlands; [6]Department of Clinical Genetics, University Medical Center Groningen, Groningen, the Netherlands; [7]Department of Clinical Genetics, Maastricht University Medical Center +, Maastricht, the Netherlands; and [8]Center for Human and Clinical Genetics, Leiden University Medical Center, Leiden, the Netherlands.

CRediT Authorship Contributions
Iris B. A. W. te Paske, MSc (Data curation: Equal; Formal analysis: Equal; Investigation: Equal; Validation: Equal; Writing – original draft: Equal; Writing – review & editing: Equal).

Arjen R. Mensenkamp, PhD (Data curation: Equal; Formal analysis: Equal; Investigation: Equal; Validation: Equal; Writing – review & editing: Equal).

Kornelia Neveling, PhD (Data curation: Equal; Formal analysis: Equal; Writing – review & editing: Equal).

Nicoline Hoogerbrugge, MD, PhD (Conceptualization: Equal; Formal analysis: Equal; Investigation: Equal; Supervision: Equal; Validation: Equal; Writing – review & editing: Equal).

Marjolijn J. L. Ligtenberg, PhD (Conceptualization: Equal; Formal analysis: Equal; Investigation: Equal; Supervision: Equal; Validation: Equal; Writing – review & editing: Equal).

Richarda M. de Voer, PhD (Conceptualization: Equal; Formal analysis: Equal; Investigation: Equal; Supervision: Equal; Writing – original draft: Equal; Writing – review & editing: Equal).

Conflicts of interest
The authors disclose no conflicts.

Funding
This project received funding from the Sacha Swarttouw-Hijmans Foundation. The Solve-RD project received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 779257. This research is supported (not financially) by the European Reference Network on Genetic Tumor Risk Syndromes (ERN GENTURIS)—Project ID No. 739547 and the Dutch X-omics Initiative, which is partly funded by NWO (184.034.019). ERN GENTURIS is partly co-funded by the European Union within the framework of the Third Health Program ERN-2016 with Framework Partnership Agreement 2017–2021.

Supplementary Methods

Study Cohort

The study cohort included 32 individuals without a germline PV and without somatic dMMR in 1 of the MMR genes, without *MLH1* promoter methylation, and with at most 1 somatically inactivated allele (Figure 1*A*). Germline and somatic variants of unknown significance (class 3 variants) were considered as not explanatory. Inclusion was prioritized based on the presence of a LS-associated cancer in persons aged $\leq$ 70 years (Supplementary Table 1). For every individual, immunohistochemistry results and somatic sequencing results were taken into consideration to decide which genes should be amplified.

Long-read Sequencing

Region selection and long-range polymerase chain reaction amplification. For each MMR gene (transcripts used for design: *MLH1*:NM_000249.4; *EPCAM-MSH2*:NM_002354.3-NM_000251.3; *MSH6*:NM_000179.3; *PMS2:*NM_000535.7) long-range polymerase chain reaction (PCR) amplicons were designed ranging from 7 to 18 kb, with a 1-kb overlap between amplicons (Supplementary Figure 1*A*). For every amplicon, leukocyte-derived genomic DNA (gDNA) was amplified on a ProFlex PCR system (Thermo Fisher Scientific) using LongAmp Hot Start Taq 2x master mix (New England Biolabs) (protocol and primer sequences are available on request). PCR products were checked on 0.8% agarose gel, and concentrations were measured by Qubit HS dsDNA Assay (Thermo Fisher) before equimolar pooling per individual.

Library preparation and single molecule real-time sequencing. Library preparation was performed according to the protocol Procedure and Checklist—Preparing Single Molecule Real-time (SMRT)bell Libraries using PacBio Barcoded Adapters for Multiplex SMRT Sequencing (Pacific Biosciences). The generation of polymerase-bound SMRTbell complexes was performed using the Sample Setup option in SMRTLink (Pacific Biosciences), and the SMRTbell complex was loaded onto an SMRTcell and sequenced either on a Sequel I or IIe system. Circular consensus long-reads meeting quality control (QC) metrics $\geq$ 20 ($\leq$1% error rate) were mapped against GRCH37/hg19 within SMRTlink.

Variant Calling and Annotation

For single nucleotide variant and small indel detection, mapped bam files were loaded into JSI SeqNext Software v5.1.0 Build 503 (JSI Medical Systems GmbH) to perform quality filtering and variant calling. Variant calling format (VCF) files containing distinct variants ($\geq$20% coverage per direction) were annotated using an in-house annotation pipeline and were filtered for GnomAD-G allele frequency $<$ 0.1% and in-house database frequency ($>$26,000 alleles) $<$ 0.1%. Coding variants and variants with a SpliceAI delta score $>$ 0.1 were included for (re)evaluation. Coding variants located upstream of *EPCAM* exon 8 were excluded. Noncoding variants were filtered for variant allele frequency $\geq$ 35% (correction for mapping difficulties in mononucleotide repeats) and selected if there was a SpliceAI delta score $>$ 0.1.[1] All bam files were also visually assessed using the Integrative Genomics Viewer v.2.10/12 (Broad Institute), with settings to flag supplementary aligned reads and flag indels ($>$10 bp) to identify structural variants.

Validations

Co-segregation analysis. Co-segregation analysis was performed by Sanger sequencing of gDNA isolated from leukocytes or formalin-fixed paraffin-embedded tissue of the index individual and available family members.

Mini-genes for in vitro splice assays and transcript analysis. For each in silico predicted likely pathogenic variant, a mini-gene construct was generated as previously described by Sangermano et al.[2] In short, gDNA or the corresponding long-range PCR product from the affected individual or reference gDNA was amplified using specific primers that were located in the genomic regions upstream and downstream of the exons that were flanking the specific variant. The product was cloned into a pDONR201 vector by means of Gateway Cloning (Thermo Fisher). When reference gDNA was used, site-directed mutagenesis was performed to introduce the variant of interest. Subsequently, wild-type and mutant constructs were cloned into the pCI-NEO-RHOexon3,5/DEST. Expression vectors were transfected using Fugene6 (Promega) in HEK293T and HCT116 cell lines. Cells were harvested 48 hours after transfection for total mRNA isolation using the RNeasy Mini kit (Qiagen). Complementary DNA (cDNA) was generated of mRNA using the iScript cDNA Synthesis kit (Bio-Rad), followed by reverse transcriptase (RT) PCR amplification of the region of interest using primers located in *RHO* exon 3 and exon 5. RT-PCR products were run on agarose gel. Bands were purified from gel and analyzed by Sanger sequencing.

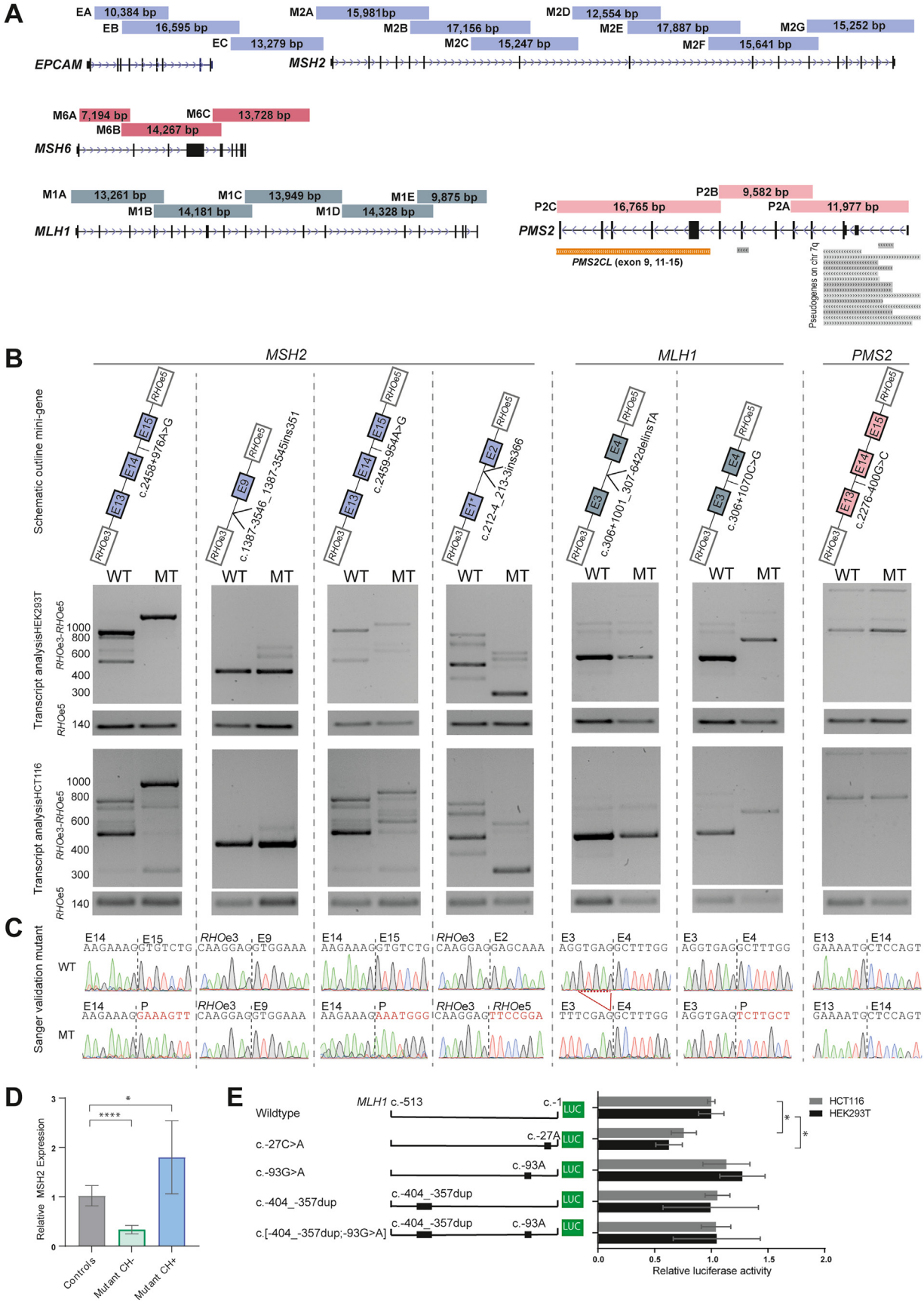
***MSH2* expression analysis.** mRNA was isolated using the QIAamp RNA Blood-Kit (Qiagen) according to the manufacturer's protocol from cultured peripheral blood lymphocytes that were cultured in the presence and absence of cycloheximide as previously described by Weren et al.[3] Total cDNA was generated as mentioned above. For transcript quantification, 5 μL of cDNA (concentration, 0.2 ng/μL) was mixed, with GoTaq qPCR Master Mix (Promega), according to manufacturer's protocol. To determine *MSH2* expression, primers targeting *MSH2* exons 2–3, 6–7, and 10–11 were used. Real-time quantitative RT-PCR was performed on a 7500 Fast Real-Time PCR system (Applied Biosystems) with *HPRT1* as control. Data represent the mean $\pm$ SD of 3 replicates. Statistical significance was determined using the 2-tailed unpaired Welch's *t*-test.

Luciferase assay. The *MLH1* promoter region c.$-$513 to c.-1 was amplified from gDNA of the individual affected with the *MLH1* c.-404_-357dup variant and from reference gDNA and cloned into the pGL3-Basic-GW vector containing a *firefly* luciferase cassette by means of Gateway Cloning. Per transfection, 50 ng *firefly* reporter plasmids containing wild-type, positive control (c.-27C>A; associated with decreased luciferase activity[4]), LLS individual wild-type

allele (c.-93G>A), LLS individual mutant allele (c.[-404_-357dup;-93G>A]), and duplication only (c.-404_-357dup), an empty vector, and pGL3-SV40 promoter (data not shown) were cotransfected with 5 ng pRL-SV40 *Renilla* luciferase reporter vector (Promega) into HEK293T and HCT116 cells using FuGene6. After 48 hours cells were harvested and lysed and luciferase activity was measured by Dual-Luciferase Reporter Assay kit (Promega), according to manufacturer's protocol. *Firefly* luciferase units were normalized with *Renilla* luciferase units. Grubbs test was applied (alpha 0.1) to remove outliers. Data represent the mean ± SD of 3 replicates. Statistical significance was determined using the 2-tailed unpaired *t*-test.

Supplementary References

1. Jaganathan K, et al. Cell 2019;176:535–548.
2. Sangermano R, et al. Ophthalmology 2016; 123:1375–1385.
3. Weren RD, et al. Nat Genet 2015;47:668–671.
4. Hitchins MP, et al. Cancer Cell 2011;20:200–213.


A
EA 10,384 bp
EB 16,595 bp
EC 13,279 bp
M2A 15,981bp
M2B 17,156 bp
M2C 15,247 bp
M2D 12,554 bp
M2E 17,887 bp
M2F 15,641 bp
M2G 15,252 bp
EPCAM
MSH2
M6A 7,194 bp
M6B 14,267 bp
M6C 13,728 bp
MSH6
M1A 13,261 bp
M1B 14,181 bp
M1C 13,949 bp
M1D 14,328 bp
M1E 9,875 bp
MLH1
P2B 9,582 bp
P2C 16,765 bp
P2A 11,977 bp
PMS2
PMS2CL (exon 9, 11-15)
Pseudogenes on chr 7q
B
MSH2
MLH1
PMS2
Schematic outline mini-gene
RHOe3
E13
E14
E15
RHOe5
c.2458+976A>G
E9
c.1387-3546_1387-3545ins351
c.2459-954A>G
E2
E1*
c.212-4_213-3ins366
E3
E4
c.306+1001_307-642delinsTA
c.306+1070C>G
c.2276-400G>C
WT
MT
Transcript analysisHEK293T
Transcript analysisHCT116
RHOe3-RHOe5
RHOe5
1000
800
600
400
300
140
C
Sanger validation mutant
WT
MT
E14 E15
AAGAAAGGTGTCTG
RHOe3 E9
CAAGGAGGTGGAAA
AAGAAAGGAAAGTT
AAGAAAGAAATGGG
RHOe3 E2
CAAGGAGGAGCAAA
CAAGGAGTTCCGGA
E3 E4
AGGTGAGGCTTTGG
TTTCGAGGCTTTGG
E3 P
AGGTGAGTCTTGCT
E13 E14
GAAAATGCTCCAGT
D
Relative MSH2 Expression
Controls
Mutant CH-
Mutant CH+
E
MLH1 c.-513
Wildtype
c.-27C>A
c.-93G>A
c.-404_-357dup
c.[-404_-357dup;-93G>A]
c.-1
c.-27A
c.-93A
LUC
HCT116
HEK293T
Relative luciferase activity

Supplementary Figure 1. Analyses of noncoding variants found in individuals with a LLS phenotype. (*A*) Design of 7- to 18-kb amplicons for LS-associated gene regions. Amplicons, including size in bp, are displayed above the gene (gene assembly derived from UCSC genome browser on GRCh37/hg19). For *PMS2*, highly homologous regions also present on chromosome 7 are visualized below the gene region (*orange,* >99% similarity; *gray*, 90%–98% similarity). For *MLH1*, *MSH2*, and *PMS2* 100% of the coding regions are covered. For *MSH6*, 5% of the locus that encompasses the 5′UTR and the first 201 bp of exon 1 could not be amplified, likely due to GC-richness of the exon 1 and promoter region. (*B*) Overview of mini-gene constructs and cDNA analysis by RT-PCR after transfection of wild-type and mutant mini-genes in HEK293T and HCT116 cells. *RHO* exon 5 RT-PCR is used as transfection control. *MSH2* c.2458+976A>G, c.2459-954A>G, c.212-4_213-3ins366, and *MLH1* c.306+1001_307-642delinsTA and c.306+1070C>G showed altered splicing compared with wild-type. *MSH2* c.1387-3546_1387-3545ins351 shows low levels of alternatively spliced transcripts. E, exon; MT, mini-gene with a mutant sequence; P, pseudoexon; *RHO*e3, *RHO* exon 3; *RHO*e5, *RHO* exon 5; WT, mini-gene with a wildtype sequence. (*C*) Sanger sequencing results of dominantly expressed transcripts from the mini-gene analyses. Sanger sequencing was performed on the most dominantly expressed wild-type and mutant transcripts from the HEK293T transfections. The wild-type and mutant cDNA sequence are included above the chromatograms. The *dotted line* represents the exon–exon border. Molecular changes due to the presence of the variant are displayed in *red*. Bands > 1000 bp in the *PMS2* mini-gene construct represent use of alternative polyadenylation sites. **MSH2* exon 1 is not spliced into the transcript because it does not have a splice acceptor site. (*D*) Quantitative analysis of *MSH2* expression of Epstein Barr virus (EBV)-transformed cell line carrying *MSH2* c.2458+976G>A (Mutant) compared with EBV-transformed controls (Controls) normalized to *HPRT1*. Data represent the mean ± SD. **P* < .05; *****P* < .0001. CH, cycloheximide. (*E*) Relative luciferase activity of different *MLH1* promoter sequences. Sequences are schematically represented at the left. *Firefly* luciferase activity normalized to *Renilla* luciferase activity and per transfection normalized to wild-type and measured HEK293T cells and HCT116 cells. Data represent mean ± SD. **P* < .05.

Supplementary Table 1. Characteristics of the Study Cohort

Individual	gender	Diagnosis and age	Family history of CRC/EC	Immunohistochemistry result	Germline variant found in diagnostics	Genes tested	Somatic variant (class 4/5)	Somatic variant (class 3)	Somatic variants	Germline variant found in this study
LLP001	F	CRC23, POL32	Y	MLH1/PMS2 negative	N	*MLH1*	*MLH1* LOH	N	One	N
LLP002	F	KC53, CRC58	U	MLH1 negative, PMS2 positive	N	*MLH1*	*MLH1* LOH	N	One	SNV (deep intronic): *MLH1* c.306+1070C>G
LLP003	M	CRC66	N:POL	MLH1/PMS2 negative	N	*MLH1*	*MLH1* LOH	N	One	N
LLP004	M	CRC33	Y	MSH2/MSH6 negative	N	*MSH2*	*MSH2* c.942+3A>T; *no indication for LOH*	N	One	SNVs (deep intronic): *MSH2* c.793-603C>T; *MSH2* c.2458+976A>G
LLP005	F	EC50	Y	MSH2/MSH6 negative	N	*MSH2*	*MSH2* c.211+2delT; *no indication for LOH*	N	One	N
LLP006	F	EC52	U	MSH6 negative	N	*MSH6*	*MSH6* c.3261dup: p.(Phe1088fs); *no indication for LOH*	N	One	N
LLP007	M	CRC59	Y	MSH6 negative	N	*MSH6*	*MSH6* LOH	N	One	N
LLP008	M	CRC51	N	PMS2 weak positive	c.2533C>G; p.(His845Asp) (Class 3)	*PMS2*	*PMS2* c.780del: p.(Asp261*); *no indication for LOH*	N	One	N
LLP009	M	UC38, CRC44	N	MSH2/MSH6 negative	N	*MSH2*	*No hits in somatic analysis, no indication for LOH*	N	None	SNV (deep intronic): *MSH2* c.2458+976A>G
LLP010	M	CRC25	Y	MLH1/PMS2 negative	N	*MLH1*	*MLH1* LOH	*MLH1* c.193G>T: p.(Gly65Cys)	One	N
LLP011	M	CRC37	N	MLH1/PMS2 negative	N	*MLH1*	*MLH1* c.884+4A>G; *no indication for LOH*	*MLH1* c.922C>T: p.(His308Tyr)	One	N
LLP012	F	CRC41	N:POL	MLH1/PMS2 negative	N	*MLH1*	*SNPs do not show LOH*	*MLH1* c.1585T>C: p.(Ser529Pro), *MLH1* c.1919C>T: p.(Pro640Leu)	None	N
LLP013	M	CRC43	N	MSH2/MSH6 weak positive	N	*MSH2* and *MSH6*	*MSH6* c.3119_3120del: p.(Phe1040*); *no indication for LOH*	*MSH6* c.1153_1155del: p.(Arg385del)	One	N

Supplementary Table 1. Continued

Individual	gender	Diagnosis and age	Family history of CRC/EC	Immunohistochemistry result	Germline variant found in diagnostics	Genes tested	Somatic variant (class 4/5)	Somatic variant (class 3)	Somatic variants	Germline variant found in this study
LLP014	F	CRC31	Y	PMS2 negative, MSH6 partly positive	N	*PMS2*	*PMS2* c.486del: p.(Leu162fs); *no indication for LOH*	N	One	SNV (deep intronic): *PMS2* c.2276-400G>C
LLP015	M	CRC35	N	MSH2/MSH6 negative	N	*MSH2*	*MSH2* LOH	N	One	N
LLP016	F	CRC46	Y	MLH1/PMS2 negative	N	*MLH1*	*MLH1* LOH	*MLH1* c.977T>A: p.(Val326Glu)	One	N
LLP017	M	CRC43	N	MSH2/MSH6 negative	N	*MSH2*	Chr2(GRCh37):g.(?_47630331)_(47710089_?)	N	One	N
LLP018	F	CRC48	Y	MSH2/MSH6 negative	N	*MSH2* and *MSH6*	*MSH6* c.3613_3615del: p.(Thr1205del); *no indication for LOH*	MSH6 c.3261dup: p.(Phe1088fs) (Low variant allele frequency, likely due to MSI)	None	N
LLP019	M	CRC43	N	MSH2/MSH6 negative	N	*MSH2*	*MSH2* c.1901T>G: p.(Leu634*)	N	One	N
LLP020	M	CRC30	N	MSH2 negative	N	*MSH2*	*MSH2* LOH	*MSH2* c.2459-11A>G	One	N
LLP021	F	CRC43	Y	Not performed	N	*MLH1*	*MLH1* LOH	*MLH1* c.2059C>T: p.(Arg687Trp)	One	N
LLP022	F	IC43	Y	MSH6 negative	N	*MSH6*	*MSH6* LOH	N	One	N
LLP023	F	CRC38	Y	MSH2/MSH6 negative	N	*MSH2*	*MSH2* LOH	N	One	N
LLP024	F	CRC32	U	MSH2/MSH6 negative	N	*MSH2*	*No indication for LOH*	N	None	Alu insertion (intronic): *MSH2* c.212-4_212-3ins366
LLP025	M	CRC50	Y	MLH1/PMS2 negative	N	*MLH1*	*Too little informative SNPs*	N	None	Tandem duplication (promoter region): *MLH1* c.-404_-357dup
LLP026	F	CRC47	Y	MLH1/PMS2 negative	N	*MLH1* and *MSH2*[a]	*MSH2* c.802del: p.(Ser268HisfsTer6); *no indication for LOH in MLH1*	N	One	N

Supplementary Table 1. Continued

Individual	gender	Diagnosis and age	Family history of CRC/EC	Immunohistochemistry result	Germline variant found in diagnostics	Genes tested	Somatic variant (class 4/5)	Somatic variant (class 3)	Somatic variants	Germline variant found in this study
LLP027	M	CRC40	Y	MSH6 negative	N	*MSH6*	*No indication for LOH*	*MSH6* c.2295C>G: p.(Cys765Trp) (variant allele frequency 50%)	None	N
LLP028	M	CRC57	Y	MSH2/MSH6 negative	N	*MSH2*	*MSH2* LOH	N	One	N
LLP029	M	BDC61	Y	MLH1/PMS2 negative	N	*MLH1* and *PMS2*[b]	*MLH1* c.(?_-1)_(1731+1_1732-1)del	N	One	SNV (deep intronic): *PMS2* c.2276-400G>C
LLP030	M	CRC62	Y	MLH1/PMS2 negative	N	*MLH1*	*MLH1* c.1838_1854del	N	One	N
LLP031	M	CRC46	Y	MSH2/MSH6 negative	N	*MSH2*	*No indication for LOH*	N	None	Alu insertion (intronic): *MSH2* c.1387-3546_1387-3545ins351; SNV (deep intronic): *MSH2* c.2459-954A>G
LLP032	M	CRC43	U	MLH1/PMS2 negative	N	*MLH1*	*No indication for LOH*	N	None	Deletion (intronic): *MLH1* c.306+1001_307-642delinsTA

Immunohistochemistry and somatic variants are of the tumor are underlined. BDC, bile duct cancer; EC, endometrium cancer; LOH, loss of heterozygosity; IC, ileocecal cancer; KC, kidney cancer; MSI, microsatellite instability; N, no; POL, polyps; SNPs, single nucleotide polymorphisms; SNV, single nucleotide variant; UC, urothelial cancer; U, unknown; Y, yes; ?, unknown age at diagnosis.

[a]*MSH2* c.802del; p.(Ser268HisfsTer6) found in tumor with allele frequency of 63% and 90% tumor cell percentage.

[b]Suggestive LOH in *PMS2* found.