IJC
International Journal of Cancer

Copy number variation analysis and targeted NGS in 77 families with suspected Lynch syndrome reveals novel potential causative genes

Katrin Kayser[1], Franziska Degenhardt[1,2], Stefanie Holzapfel[1,3], Sukanya Horpaopan[1,4], Sophia Peters[1,2], Isabel Spier[1,3], Monika Morak[5,6], Deepak Vangala[7], Nils Rahner[8], Magnus von Knebel-Doeberitz[9,10], Hans K. Schackert[11], Christoph Engel[12], Reinhard Büttner[13], Juul Wijnen[14], Tobias Doerks[15,†], Peer Bork[15], Susanne Moebus[16], Stefan Herms[2,17,18], Sascha Fischer[17], Per Hoffmann[1,2,17,18], Stefan Aretz[1,3] and Verena Steinke-Lange[1,5,6]

[1]Institute of Human Genetics, University of Bonn, Bonn, Germany
[2]Department of Genomics, Life and Brain Center, University of Bonn, Bonn, Germany
[3]Center for Hereditary Tumor Syndromes, University of Bonn, Bonn, Germany
[4]Department of Anatomy, Faculty of Medical Science, Naresuan University, Phitsanulok, Thailand
[5]Medizinische Klinik und Poliklinik IV, Campus Innenstadt, Klinikum der Universität München, Munich, Germany
[6]Medical Genetics Center (MGZ), Munich, Germany
[7]Department of Internal Medicine, Knappschaftskrankenhaus, Ruhr-University Bochum, Bochum, Germany
[8]Institute of Human Genetics, University of Düsseldorf, Düsseldorf, Germany
[9]Department of Applied Tumor Biology, Institute of Pathology, University Hospital of Heidelberg, Heidelberg, Germany
[10]Cooperation Unit Applied Tumor Biology, German Cancer Research Center (DKFZ), Heidelberg, Germany
[11]Department of Surgical Research, Technische Universität Dresden, Dresden, Germany
[12]Institute of Medical Informatics, Statistics, and Epidemiology, University of Leipzig, Leipzig, Germany
[13]Institute of Pathology, University of Cologne, Cologne, Germany
[14]Department of Clinical Genetics, Leiden University Medical Center, Leiden, The Netherlands
[15]Structural and Computational Biology Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany
[16]Centre for Urban Epidemiology, University Hospital of Duisburg-Essen, University of Duisburg-Essen, Essen, Germany
[17]Human Genomics Research Group, Department of Biomedicine, University of Basel, Basel, Switzerland
[18]Insitute of Medical Genetics and Pathology, University Hospital of Basel, Basel, Switzerland

In many families with suspected Lynch syndrome (LS), no germline mutation in the causative mismatch repair (MMR) genes is detected during routine diagnostics. To identify novel causative genes for LS, the present study investigated 77 unrelated, mutation-negative patients with clinically suspected LS and a loss of MSH2 in tumor tissue. An analysis for genomic copy number variants (CNV) was performed, with subsequent next generation sequencing (NGS) of selected candidate genes in a subgroup of the cohort. Genomic DNA was genotyped using Illumina's HumanOmniExpress Bead Array. After quality control and filtering, 25 deletions and 16 duplications encompassing 73 genes were identified in 28 patients. No recurrent CNV was

Key words: Lynch syndrome, HNPCC, MSH2, CNV analysis, NGS

Abbreviations: CADD: combined annotation-dependent depletion; CNV: copy number variant; CRC: colorectal cancer; EC: endometrial cancer; ExAC: exome aggregation consortium; HI: haploinsufficiency score; HNPCC: hereditary nonpolyposis colorectal cancer; HNR: Heinz Nixdorf RECALL; IBS: identity-by-state; IHC: immunohistochemical staining; kb: kilobase; LBF: log Bayes factor; LS: Lynch syndrome; MLPA: multiplex ligation-dependent probe amplification; MMR: mismatch repair; MSI: microsatellite instability; NGS: next generation sequencing; PPAP: polymerase proofreading-associated polyposis; QC: quality control; RVIS: residual variant intolerance score; SCID: severe combined immunodeficiency; SNP: single-nucleotide polymorphism; STRING: search tool for the retrieval of interacting genes/proteins; TGCA: the Cancer Genome Atlas; VUS: variants of unknown significance.

Additional Supporting Information may be found in the online version of this article.

K.K., F.D., S.A., and V.S.-L. contributed equally to this work.

[†]deceased.

Grant sponsor: Deutsche Krebshilfe; **Grant numbers:** 190370; **Grant sponsor:** German Ministry of Education and Science and the German Research Council; **Grant sponsor:** Heinz Nixdorf Foundation; **Grant sponsor:** German Cancer Aid; **Grant numbers:** 190370
DOI: 10.1002/ijc.31725
History: Received 10 July 2017; Accepted 26 Mar 2018; Online 10 July 2018

Correspondence to: Verena Steinke-Lange, Medical Genetics Center (MGZ), Bayerstrasse 3-5, D-80335 Munich, Germany, E-mail: verena.steinke-lange@mgz-muenchen.de; Tel.: +49-89/30-90-886-0, Fax: +49-89/30-90-886-66

detected, and none of the CNVs affected the regulatory regions of *MSH2*. A total of 49 candidate genes from genomic regions implicated by the present CNV analysis and 30 known or assumed risk genes for colorectal cancer (CRC) were then sequenced in a subset of 38 patients using a customized NGS gene panel and Sanger sequencing. Single nucleotide variants were identified in 14 candidate genes from the CNV analysis. The most promising of these candidate genes were: (*i*) *PRKCA*, *PRKDC*, and *MCM4*, as a functional relation to *MSH2* is predicted by network analysis, and (*ii*) *CSMD1*, as this is commonly mutated in CRC. Furthermore, six patients harbored *POLE* variants outside the exonuclease domain, suggesting that these might be implicated in hereditary CRC. Analyses in larger cohorts of suspected LS patients recruited via international collaborations are warranted to verify the present findings.

What's new?

While several causal genetic factors for Lynch syndrome (LS), or hereditary non-polyposis colorectal cancer, have been identified, in many families with suspected LS no germline mutation is detected during routine diagnostics. Here, a genome-wide copy number variant (CNV) analysis in a large cohort of patients with suspected LS and MSH2 loss identified rare alterations in candidate genes (*PRKCA*, *PRKDC*, *MCM4*, and *CSMD1*) that may predispose to colorectal tumorigenesis. The study demonstrates that rare germline CNVs and point mutations are likely to contribute to the hereditary risk for colorectal tumors, and the underlying genetic factors are likely to be very heterogeneous.

Introduction

Lynch syndrome (LS) or hereditary nonpolyposis colorectal cancer (HNPCC) is a tumor predisposition syndrome characterized by a high risk of colorectal cancer (CRC); endometrial cancer (EC); and a variety of additional malignancies.[1] Research has identified several causal genetic factors. These comprise germline mutations in four mismatch repair (MMR) genes (*MLH1*, *MSH2*, *MSH6*, and *PMS2*), and deletions in the 3' region of the *EPCAM* gene upstream of *MSH2*.[1] In LS-associated cancers, tumor tissue usually displays both a loss of the respective DNA repair protein, and high microsatellite instability (MSI) as a sign of the DNA repair defect.[1] The pattern of loss detected via immunohistochemical staining (IHC) indicates the underlying genetic defect, or epigenetic alterations in case of MLH1 loss.

However, a germline mutation in one of the aforementioned MMR genes is only detectable in around 53% of patients who fulfill the clinical criteria for suspected LS (revised Bethesda guidelines[2]) and show MSI in their tumor tissue.[3] In patients whose tumor tissue shows immunohistochemical loss of MSH2/MSH6, the mutation detection rate is slightly higher (70%). Nonetheless, in 30% of these patients, no causative germline mutation is detected during routine diagnostics.[3] Mutation-negative patients include individuals with a young age of onset and/or a positive family history of LS-associated cancers, factors which are strong indicators that mutations in as yet undiscovered genes might be causative for LS.

Since its establishment in 1999, the German HNPCC Consortium has collected one of the largest cohorts of patientswith clinically suspected LS worldwide. The cohort includes a large number of individuals who fulfill the revised Bethesda criteria, show signs of a DNA repair defect in their tumor tissue, and have no detectable MMR germline mutation. In the majority of these patients, a loss of MLH1 and/or PMS2 has been detected in the respective tumor tissue. Research has demonstrated that this tumor phenotype is often caused by somatic changes, in particular methylation of the *MLH1* promoter. Loss of MSH2/MSH6 in tumor tissue is identified less frequently, and is more often attributable to an underlying germline alteration. Therefore, a focus on patients with a loss of MSH2 is a promising approach to the identification of novel causal genes for LS.

The aim of the present work was to identify novel causative genes in unrelated mutation-negative patients with clinically suspected LS and a loss of MSH2/MSH6 protein expression in their tumor tissue. Therefore, we chose a two-step study design. First, a genome-wide copy number variant (CNV) analysis was performed as large heterozygous deletions and duplications contribute significantly to the germline mutation spectrum of the MMR genes and therefore CNVs in yet undiscovered LS genes might also be disease causing. Second, promising candidate genes of the CNV analysis were sequenced in a subgroup of the cohort to identify pathogenic point mutations.

Materials and Methods

Patients

Patients were recruited at six German university hospitals participating in the German HNPCC Consortium as described elsewhere,[4] and at the Leiden University Medical Center. No patient had intellectual disability or any other severe mental disease given the clinical impression and personal history. The study was approved by the ethics committee of each participating institution, and all study procedures complied with

the Declaration of Helsinki. All patients provided written informed consent prior to inclusion.

The initial cohort comprised 137 index cases from the German HNPCC Consortium and 12 index patients from the Leiden University Medical Center. All of the respective families met the Amsterdam II criteria[5] or the revised Bethesda Guidelines.[2] For each patient, IHC of tumor tissue had been performed for a minimum of MLH1 and MSH2. With the exception of two patients, the result indicated MSH2 deficiency (and additional MSH6 deficiency where applicable). The families of the two patients in whom no MSH2 or MSH6 deficiency was detected met the Amsterdam II criteria. In the first patient, only adenoma tissue was available for IHC. This displayed no MMR protein loss. In the second patient, tumor tissue showed MSI and no MMR protein loss in IHC.

If not previously performed, the following were conducted in the leucocyte DNA of all patients: (*i*) Sanger or next generation sequencing (NGS) of all coding exons and adjacent intronic regions of *MSH2;* and (*ii*) multiplex ligation-dependent probe amplification (MLPA) analysis of all *MSH2* exons, and *EPCAM* exons 8 and 9. These analyses were also performed for the *MSH6* gene. However, in some cases, the *MSH6* analysis had been performed by DHPLC prescreening and subsequent sequencing of the conspicuous exons only, and the remaining amount of DNA was insufficient for additional *MSH6* analysis. We also excluded two large recurrent *MSH2* inversions described in the literature[6,7] in most of our patients (where enough DNA was available).

Variants of unknown significance (VUS) in *MSH2* or *MSH6* were reevaluated prior to study inclusion using the *InSiGHT* database (http://insight-group.org/variants/classifications). The initial patient cohort comprised 149 patients with loss of MSH2 expression in their tumor tissue and no pathogenic *MSH2* germline mutation. Of these, 53 patients harbored VUS in *MSH2* or *MSH6*. Reevaluation of these variants revealed that many of these have since been classified as pathogenic (Class 5) or probably pathogenic (Class 4). As a result, 39 patients were excluded from the analysis. Prior to the CNV analysis, 15 additional patients were found to harbor a pathogenic variant in *MSH2* or *MSH6* through the application of up-dated routine diagnostics. In one patient, an unexpected *MLH1* mutation was detected by NGS, and one patient was diagnosed with *MUTYH*-associated polyposis after a clinical review of his medical data prompted a mutation analysis of the *MUTYH* gene. All of these patients were excluded from the present analyses, in addition to 16 patients whose DNA amount or quality was insufficient for array analysis.

The final genome-wide CNV analysis was conducted in 77 unrelated mutation-negative patients. A subset of 39 patients with sufficient DNA amount and quality was selected for NGS of candidate genes.

Microsatellite analysis

MSI analysis was performed on matched pairs of tumor and normal DNA samples using the National Cancer Institute/International Collaborative Group on HNPCC (NCI/ICG-HNPCC) reference-marker panel for the evaluation of MSI in CRC, as described elsewhere.[3]

IHC of MMR proteins

IHC for the MMR proteins MLH1, MSH2, MSH6, and PMS2 was performed as described elsewhere.[4] The level of protein staining in tumor cells was compared to the protein level in normal tissue. MMR protein expression was considered deficient if the nuclei showed no immunostaining, or only very weak immunostaining, relative to normal tissue.

Genotyping and quality control prior to CNV detection

For all individuals, DNA from venous blood samples was genotyped on the Illumina HumanOmniExpress-12 BeadChip. To minimize technical artifacts in CNV calling, stringent quality control (QC) criteria were applied. Single-nucleotide polymorphisms (SNPs) with a call rate of <98% were excluded. Individuals with the following characteristics were removed from the dataset: (*i*) DNA call rate <98%; and/or (*ii*) difference between X-chromosomally inferred and phenotypic sex.

CNV detection

The SNP-chip data of each participant were analyzed with QuantiSNP, as described elsewhere (v2.2, http://www.well.ox.ac.uk/QuantiSNP).[8–10] Participants were excluded if the standard deviation from the log R ratio calculated over all SNPs exceeded 0.30. X-chromosome data were excluded from the analysis as we did not expect X-chromosomal inheritance.

Filtering against an in-house control dataset

The in-house control dataset comprises 1,320 population-based controls, as described elsewhere.[10] These individuals were drawn from the population-based Heinz Nixdorf RECALL (HNR) study (Risk Factors, Evaluation of Coronary Calcium and Lifestyle).[11] All individuals were genotyped on the Illumina HumanOmniExpress-12 BeadChip. SNPs with a call rate of <98% were excluded. Participants with the following characteristics were removed from the dataset: (*i*) DNA call rate <98%; (*ii*) difference between X-chromosomally inferred and phenotypic sex; (*iii*) DNA sample doublets identified by identity-by-state (IBS) estimates (defined as IBS = 2); (*iv*) cryptic relatedness (IBS ≥ 1.6); and/or (*v*) population outlier status according to multidimensional scaling with HapMap phase 2.2 data. The CNV detection protocol was equivalent to that used for patients.

CNV filter criteria

To be considered for downstream analysis, each CNV was required to: (*i*) span ≥10 kb; (*ii*) encompass ≥5 consecutive markers; (*iii*) have a max. log Bayes factor (LBF) of ≥10

(deletions) or ≥20 (duplications); and (*iv*) lie within 50 kb upstream or downstream of a RefSeq gene boundary (according to GRch37/hg19) or the putative regulatory regions 1.5 Mb upstream and downstream of *MSH2*.

As the focus of the present study was the identification of rare and highly penetrant CNVs, all deletions and duplications fulfilling the above mentioned criteria were filtered against the in-house control dataset. LS has a prevalence of around 0.2% in the general population.[12] Thus all CNVs with a frequency of ≥0.2% in the in-house control dataset were excluded from the downstream analyses. Furthermore, CNVs that showed partial or complete overlap with known segmental duplications were excluded from further analysis, as these regions are naturally prone to copy number changes, and are unlikely to cause monogenic disease.

Experimental verification of predicted CNVs

CNVs that surpassed the present filter criteria were visually inspected in Genome-Studio (v2011.1, http://www.illumina.com/software/genomestudio_software.ilmn). Experimental verification of predicted CNVs was performed in triplicate using qPCR and Fast SYBR®Green (Life Technologies, Carlsbad, CA), as described previously.[13,14] Briefly, three to four primer pairs were designed for each region (sequences available upon request). Relative copy numbers were measured in comparison to three housekeeping genes (*BCN1*, *CFTR*, and *RnaseP*). CNVs that implicated a gene that was followed-up in the sequencing analyses were either experimentally verified, or assumed to be a true CNV finding on the basis of a LBF of >60.

Compilation of an *MSH2* interaction partner list

A genomic context analysis was performed to compile a list of the 100 closest interaction partners of *MSH2*. The genomic context analysis approach is described elsewhere.[15] All genes larger than 10 kb were covered with ≥3 SNPs on the Illumina HumanOmniExpress-12 BeadChip. Three X-chromosomal genes were excluded from further analysis as the X-chromosome was not included in CNV analysis.

Selection of candidate genes for NGS

To validate the etiological relevance of the candidate genes, promising candidate genes were further investigated by NGS analysis of leukocyte DNA using a customized panel in order to identify additional point mutations (i.e., truncating variants and putative missense variants, as well as in-frame deletions/duplications and silent mutations predicted to be deleterious on the basis of a high CADD [combined annotation-dependent depletion] score). Genes affected by a CNV in the present study were prioritized according to function. A subset of 47 genes was thereby selected for NGS analysis, two further genes for Sanger sequencing. Candidate gene selection was restricted to genes expressed in normal colon mucosa, as indicated by two publicly available databases, EST profiles reported in the UniGene database, and RNA expression data from normal human tissues reported in GeneCards. Genes were considered to be expressed if the value of transcripts per million was above zero. X-chromosomal genes were excluded from the analysis. Genes for long noncoding RNAs were excluded, as their function remains unclear. In addition, 30 established or putative CRC risk genes from the literature were included to exclude other hereditary cancer syndromes in the present cohort mimicking the LS phenotype.

Targeted high-throughput sequencing

Mutation screening of the 77 selected candidate genes was performed using targeted NGS and TruSeq enrichment protocols (Illumina). Oligonucleotide probes were designed using the Illumina DesignStudio software. One microgram of genomic DNA was extracted from leukocytes using standard protocols and fragmented using sonication technology (Bioruptor, Diagenode Liège, Belgium). The fragments were end repaired and adapter ligated using Illumina's TruSeq® DNA HT Sample Preparation Kit. Custom capture of targeted regions was performed on pools of 12 index libraries using Illumina's TruSeq enrichment protocol. The captured DNA was sequenced using Illumina's MiSeq2000 sequencer with 2x100bp paired-end reads. Coverage of 30x was achieved for at least 96% of targeted bases. Data were filtered using Illumina Realtime Analysis® software.

To identify somatic *MSH2* mutations, *MSH2* and *MSH6* were analyzed in the tumor samples of 11 patients of whom tumor DNA was available using NGS and the TruSight Cancer Panel (Illumina) and Illumina's MiSeq2000 sequencer.

Alignment, genotype calling, variant annotation, and filtering of NGS data

Reads were aligned to the hg19 human reference genome using the *in silico* PCR und BLAST tools of the UCSC Genome Browser.[16] Variant call quality was assessed with VariantStudio v2.2 (Illumina). A minimum quality score of 30, a minimum coverage of 10x, and an alternative variant frequency of >10% were required. Allele frequencies were obtained from the Exome Aggregation Consortium Browser (release 0.3.1), after the exclusion of samples from patients with known cancer. Common variants with an allele frequency of >0.005 were excluded from the analysis. All variants withstanding these filter criteria were confirmed by Sanger sequencing.

In silico prediction of NGS variants

Pathogenic effects of missense variants were predicted using four *in silico* analysis tools (SIFT, Polyphen-2, MutationTaster, and PROVEAN).[17–20] Amino acid insertions and deletions were analyzed by MutationTaster and PROVEAN only. Exonic silent variants and intronic variants (+/− 10 bp of RefSeq exon boundaries) were analyzed using MutationTaster and Human Splicing Finder.[21]

CADD scores were obtained for all variants, as described by Kircher *et al.*[22] A PHRED-like scaled C-score (CADD score) of ≥10 indicates the 10% most probable deleterious substitutions in the human genome, a score of ≥20 indicates the 1% most deleterious. A CADD threshold score of ≥20 was applied to the present data.

The pathogenic relevance of the variants was further explored by evaluating: (*i*) the genetic intolerance to functional variation of the respective gene, as measured by the residual variation intolerance score (RVIS) which ranges from 0% (most intolerant genes) to 100% (most tolerant genes) (RVIS v4 constructed on the ExAC v2 data release)[23]; and (*ii*) the likelihood of haploinsufficiency of the respective gene, as measured by the haploinsufficiency score, where a low ranking (e.g., 0–10%) indicates that a gene is more likely to exhibit haploinsufficiency (from Supporting Information dataset S2, including imputed values).[24]

Data on the frequency of somatic mutations in colorectal tumors were obtained from the exome database of *The Cancer Genome Atlas* (TCGA), after the exclusion of hypermutated tumors, as described elsewhere.[10]

Network analysis, STRING

The database STRING (*Search Tool for the Retrieval of Interacting Genes/Proteins*) v.10.5 (*https://string-db.org*) was used to detect functional associations between MSH2 and the proteins of candidate CNV genes.[15] The following settings were used: organism, *Homo sapiens*; meaning of network edges, evidence; active interaction sources, Textmining, Experiments, Databases, Coexpression, Neighborhood, Gene Fusion, Cooccurrence; and minimum required interaction score, medium confidence (0.400). For the network analysis, all CNV genes were uploaded with *MSH2* and possible interactions were analyzed.

Ingenuity pathway analysis, Qiagen

In silico pathway analysis was performed using the web-based software Ingenuity Pathway Analysis Qiagen (www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis/). For the *in silico* pathway analysis, all CNV genes were uploaded together with known CRC genes, and analyses were performed to identify shared canonical cancer pathways.

Calculation of somatic mutation frequencies in nonhypermutated tumor samples

Data concerning the frequency (percentage) of colorectal tumors with somatic mutations in candidate genes were obtained from the exome database of TCGA (https://tcga-data.nci.nih.gov/tcga/). Somatic variants identified in exome data from colonic (n = 400) and rectal (n = 137) adenocarcinomas were downloaded from the TCGA data portal. To correct the data for the presence of passenger mutations, we excluded hypermutated tumors from the dataset.[10] Therefore, the distribution of somatic variants in the TCGA exomes was analyzed, and all tumors with >200 variants (41% of the tumors) were excluded. We used the remaining 315 exomes (59% of tumors) to calculate the frequency of somatic mutations in candidate genes.

Results

CNV analysis

For the genome-wide CNV analysis, 77 apparently unrelated patients were included (Table 1; for more detailed information, see Supporting Information Table S1). After stringent QC and filtering, a total of 71 patients and 41 rare germline CNVs remained for analysis. These 41 CNVs were identified in 28 patients (39%), and comprised 25 unique heterozygous deletions and 16 unique duplications. The majority of *in silico* predicted CNVs were subjected to experimental verification by qPCR (Supporting Information Table S2). The majority of patients (68%) carried one CNV only. The remaining patients carried a maximum of four CNVs. No homozygous deletions were found.

The 25 deletions had an average size of around 67 kb (13–387 kb) and affected 36 RefSeq genes. The 16 duplications had an average size of around 135 kb (16–501 kb) and affected 37 RefSeq genes. Thus, a total of 73 genes were affected (Supporting Information Table S2). None of the CNVs were recurrent or overlapping. No gene was implicated in both a deletion and a duplication, or was affected in more than one patient. Ten CNVs (eight deletions and two duplications) affected the intronic region of a gene only. Segregation analysis was not possible in any of the families.

Table 1. Patient characteristics

Patients	Count
Total	77
Gender	
Female	34
Male	43
Inclusion criteria	
Revised Bethesda criteria[1]	67
Amsterdam criteria	10
Results microsatellite analysis	
MSI-H	64
MSI-L	3
MSS	1[2]
Results immunohistochemistry	
Loss of MSH2 expression	75
Intact MSH2 expression	2[3]
Medial age of onset first LS tumor (years)	43

[1]Amsterdam Criteria not fulfilled; two patients not certain because of the lack of information.
[2]Result in an adenoma from an Amsterdam positive patient.
[3]One result in an adenoma from an Amsterdam positive patient, one patient with MSI-H but normal expression of all MMR genes.

After the exclusion of three X-chromosomal genes, the compiled list of 100 possible interaction partners of the MSH2 protein (Supporting Information Table S3) was compared to the CNV data, and a partial deletion (exons 1–10) of *MCM4* was identified in patient 41. This deletion in patient 41 also affected the *PRKDC* gene (exons 1–40). None of the other putative MSH2 interaction partners was affected by a CNV in the present cohort.

As loss of MSH2 in tumor tissue was a central focus of the study, the putative regulatory regions 1.5 Mb up- and downstream of *MSH2* were analyzed for rare CNVs. No CNV was detected in these regions in the present cohort.

Functional evaluation of CNV genes

A network analysis (STRING) and a pathway analysis (Ingenuity Pathway Analysis, Qiagen) were performed to determine a functional connection between candidate genes from the CNV analysis, and between the present candidate genes and *MSH2*. A common network was detected for eight of the CNV genes (11%) (Supporting Information Fig. S1). The closest functional connections to *MSH2* were found for *MCM4* and *PRKDC*. Both interactions have been determined experimentally in previous studies.[25–28]

Three CNV genes (*PRKCA, PRKDC,* and *MCM4)* were involved in canonical cancer pathways (Supporting Information Table S4). These genes were also of interest from a functional perspective. *PRKCA* is involved in a wide variety of signaling pathways. These include pathways that regulate cell proliferation, apoptosis, differentiation, migration, adhesion, and tumorigenesis. *PRKDC* is involved in DNA damage checkpoint regulation. *MCM4* is involved in the control of chromosomal replication.

Independent of the network and pathway analysis, *TSPAN5* and *CSMD1* were considered interesting candidate genes. The TSPAN5 protein plays a role in the regulation of cell adhesion, migration, and proliferation. The function of the protein encoded by *CSMD1* remains unclear. However, *CSMD1* is a candidate gene for oral and oropharyngeal squamous cell carcinoma.[29,30]

Comparing our results to literature data, we found that four more germline CNVs in NRG3, which was affected in patient 38 by a partial deletion, were recently reported in another study.[31]

Sequencing of candidate genes

Of the 73 genes implicated by CNVs in the present analysis, the most promising 47 genes, which affected 22 patients, were selected for NGS analysis, as well as 30 established or putative CRC risk genes (Supporting Information Table S5). NGS sequencing was conducted in 39/77 patients. For 38 of these patients, high-quality sequencing data were obtained. For each coding exon of the 77 genes, high-quality data were obtained. In total, 98 rare (allele frequency < 0.005) heterozygous germline variants were detected (Supporting Information Table S6). Of these, 36 had a CADD score of >20 and 20 were located within a total of 12 candidate genes identified in the CNV analysis (Table 2). Mutations with a CADD score of >20 included 33 missense variants, one in-frame deletion, one synonymous variant, one splice acceptor variant, and no nonsense or frameshift variants. Segregation analysis was not feasible for any of these mutations as no DNA from additional family members was available. No variants were detected via the Sanger sequencing of two additional candidate genes from the CNV analysis (miRNA genes *MIR208A* and *MIR4506*).

Germline variants in CNV genes

Regarding only variants with a CADD score >20, NGS revealed that 14 genes affected by a CNV in one of the present patients harbored point variants in at least one further patient (Table 2). For three CNV genes (*GRIK4*, *NOSIP,* and *TRIM41*), additional NGS variants were found in more than one patient. The *GRIK4* missense variants c.514C>T;p.Leu172Phe, c.2219T>C;p.Ile740Thr, c.896C>G;p.Thr299Ser, and c.569C>T;p.Ser190Phe were found in one patient, respectively. The silent *NOSIP* variant c.72G>T;p.= and the missense *NOSIP* variant c.277G>A;p.Gly93Ser were detected in one patient, respectively. The *TRIM41* missense variants c.1084C>T;p.Arg362Cys and c.1561G>T;p.Gly521Cys were identified in one patient, respectively.

In patient 41, the CNV analysis revealed a heterozygous deletion of *PRKDC,* which also spanned the gene *MCM4.* This patient had been diagnosed with EC and CRC at the age of 48 and 72 years, respectively. Her family history was suggestive of LS: sister, EC and CRC; mother, liver cancer; and father, leukemia. Interestingly, in patient 31, NGS analysis revealed two rare missense variants in *PRKDC* (c.8809G>A;p.Val2937Ile and c.7201C>T;p.Leu2401Phe; allelic phase unknown). These variants had CADD scores of 22 and 20, respectively. Patient 31 had been diagnosed with CRC at the age of 37 years. His uncle had been diagnosed with CRC at the age of 73 years. No other patient carried two variants (CNV and / or point variant) in a single CNV gene.

Among the other 10 CNV genes in which only one point variant in addition to the CNV was found, *PRKCA* and *CSMD1* were considered interesting candidate genes, as described above. Patient 42 carried a heterozygous deletion affecting intron 3 of *PRKCA*. This patient had been diagnosed with CRC at the age of 43 years, and his father had been diagnosed with CRC at the age of 55 years. A *PRKCA* in-frame variant (c.1658_1660delACA;p.Asn554del) was detected via NGS in patient 45. She had been diagnosed with CRC at the age of 50 years. Her mother died of cancer of unknown primary at the age of 45 years, and her grandfather had been diagnosed with CRC at the age of 83 years.

Patient 23 who was diagnosed with small bowel cancer at the age of 81 and later with another cancer (probably an upper tract urothelial carcinoma) harbored a partial duplication of the *CSMD1* gene. A missense variant (c.8789C>G) in

Table 2. Mutations in rare CNV genes and CRC genes from literature with CADD score >20

Gene	Origin of gene	ExAC_ RVIS v4	HI score dataset 2	cDNA change	Predicted AA change	Patient no.	ExAC freq[1]	CADD-score	Mutation taster	PolyPhen	Sift	Provean prediction[2]	Human Splicing Finder v3.0
AGMO	CNV	90.8%	NA	c.712G>T	p.Gly238Cys	36	0.000448666	34.0	Disease causing	Probably damaging	Deleterious	Deleterious	Potential alteration of splicing
CSMD1	CNV	0.3%	49.3%	c.8789C>G	p.Ser2930Cys	65	9.47454E-06	22.6	Disease causing	Benign	Tolerated	Neutral	Potential alteration of splicing
DDX24	CNV	74.7%	49.9%	c.1805G>A	p.Arg602His	21	0.001422301	35.0	Disease causing	Probably damaging	Deleterious	Deleterious	Potential alteration of splicing
GRIK4	CNV	13.0%	41.1%	c.514C>T	p.Leu172Phe	8	9.52599E-06	29.1	Disease causing	Probably damaging	Deleterious	Deleterious	Probably no impact on splicing
GRIK4	CNV	13.0%	41.1%	c.2219T>C	p.Ile740Thr	39	1.88377E-05	25.9	Disease causing	Possibly damaging	Deleterious	Deleterious	Potential alteration of splicing
GRIK4	CNV	13.0%	41.1%	c.896C>G	p.Thr299Ser	46	4.74347E-05	20.2	Disease causing	Benign	Tolerated	Neutral	Potential alteration of splicing
GRIK4	CNV	13.0%	41.1%	c.569C>T	p.Ser190Phe	36	0	23.9	Disease causing	Probably damaging	Tolerated	Deleterious	Potential alteration of splicing
KATNAL1	CNV	26.2%	28.5%	c.1082C>T	p.Pro361Leu	22	8.48288E-05	28.6	Disease causing	Probably damaging	Deleterious	Deleterious	Potential alteration of splicing
KLHDC4	CNV	98.1%	36.0%	c.83G>A	p.Arg28His	56	0.000629053	29.7	Disease causing	Unknown	Deleterious	Deleterious	Potential alteration of splicing
NOSIP	CNV	55.7%	51.4%	c.72G>T	p.=	46	0.000591366	21.6	Disease causing	Na	Tolerated	Neutral	Potential alteration of splicing
NOSIP	CNV	55.7%	51.4%	c.277G>A	p.Gly93Ser	44	0	23.1	Polymorphism	Benign	Tolerated	Neutral	Probably no impact on splicing
OTUB2	CNV	62.4%	8.5%	c.146G>A	p.Gly49Glu	38	0	27.5	Disease causing	Probably damaging	Deleterious	Deleterious	Potential alteration of splicing
PRKCA	CNV	10.3%	5.1%	c.1658_ 1660delACA	p.Asn554del	45	0	23.1	Disease causing	Na	Na	Deleterious	Potential alteration of splicing
PRKDC	CNV	2.9%	8.7%	c.8809G>A	p. Val2937Ile	31	9.53925E-06	22.2	Disease causing	Benign	Tolerated	Neutral	Probably no impact on splicing
PRKDC	CNV	2.9%	8.7%	c.7201C>T	p.Leu2401Phe	31	0.000596727	20.2	Disease causing	Benign	Tolerated	Neutral	Potential alteration of splicing

(Continues)

Table 2. Mutations in rare CNV genes and CRC genes from literature with CADD score >20 (Continued)

Gene	Origin of gene	ExAC_ RVIS v4	HI score dataset 2	cDNA change	Predicted AA change	Patient no.	ExAC freq[1]	CADD-score	Mutation taster	PolyPhen	Sift	Provean prediction[2]	Human Splicing Finder v3.0
PRRG2	CNV	2.9%	8.7%	c.254C>A	p.Thr85Asn	55	0	22.2	Polymorphism	Possibly damaging	Tolerated	Neutral	Probably no impact on splicing
TRIM41	CNV	43.6%	82.8%	c.1084C>T	p.Arg362Cys	38	4.22833E-05	24.3	Disease causing	Benign	Tolerated	Deleterious	Potential alteration of splicing
TRIM41	CNV	43.6%	82.8%	c.1561G>T	p.Gly521Cys	65	0	23.6	Polymorphism	Probably damaging	Deleterious	Neutral	Potential alteration of splicing
TRMT44	CNV	95.3%	NA	c.1928-2_1929delAGAG		42	5.75805E-05	35.0	Disease causing	Na	Na	Na	Most probably affecting splicing
VWDE	CNV	NA	NA	c.4430G>A	p.Arg1477His	47	0.000631	23.3	Polymorphism	Possibly damaging	Tolerated	Neutral	Potential alteration of splicing
AMER1	Literature	7.84%	NA	c.401A>C	p.His134Pro	11	0.00118574	22.4	Polymorphism	Possibly damaging	Deleterious	Neutral	NA
APC	Literature	0.19%	0.467	c.1703G>A	p.Ser568Asn	22	0	28.2	Disease causing	Possibly damaging	Deleterious	Deleterious	Probably no impact on splicing
APC	Literature	0.19%	0.467	c.8425G>A	p.Val2809Met	24	0	25.4	Disease causing	Possibly damaging	Deleterious	Neutral	Potential alteration of splicing
CHEK2	Literature	70.34%	0.987	c.1441G>T	p.Asp481Tyr	47	0.000246	33.0	Disease causing	Probably damaging	Deleterious	Deleterious	Potential alteration of splicing
MSH2	Literature	8.66%	0.998	c.1864C>A	p.Pro622Thr	22	0	27.9	Disease causing	Probably damaging	Deleterious	Deleterious	Potential alteration of splicing
MSH6	Literature	1.54%	0.974	c.722G>T	p.Ser241Ile	46	0	24.3	Polymorphism	Benign	Tolerated	Neutral	Potential alteration of splicing
MSH6	Literature	1.54%	0.974	c.1372C>T	p.His458Tyr	65	0	26.2	Disease causing	Probably damaging	Deleterious	Deleterious	Probably no impact on splicing
PDGFRA	Literature	15.7%	0.948	c.899A>C	p.Lys300Thr	32	1.88477E-05	22.5	Disease causing	Possibly damaging	Tolerated	Neutral	Potential alteration of splicing
POLD1	Literature	23.37%	0.961	c.2052G>C	p.Gln684His	25	0.000388876	26	Disease causing	Possibly damaging	Deleterious	Neutral	Potential alteration of splicing
POLE	Literature	2.77%	0.628	c.2770C>T	p.Arg924Cys	23	1.88306E-05	34	Disease causing	Probably damaging	Deleterious	Deleterious	Probably no impact on splicing

(Continues)

Table 2. Mutations in rare CNV genes and CRC genes from literature with CADD score >20 (Continued)

Gene	Origin of gene	ExAC_ RVIS v4	HI score dataset 2	cDNA change	Predicted AA change	Patient no.	ExAC freq[1]	CADD-score	Mutation taster	PolyPhen	Sift	Provean prediction[2]	Human Splicing Finder v3.0
POLE	Literature	2.77%	0.628	c.2683G>A	p.Ala895Thr	25	0.000376712	35	Disease causing	Probably damaging	Deleterious	Deleterious	Potential alteration of splicing
POLE	Literature	2.77%	0.628	c.3245G>A	p.Arg1082His	32	0.000151875	24.7	Disease causing	Benign	Tolerated	Deleterious	Potential alteration of splicing
POLE	Literature	2.77%	0.628	c.274A>C	p.Ser92Arg	37	6.59183E-05	27.8	Disease causing	Possibly damaging	Deleterious	Deleterious	Potential alteration of splicing
POLE	Literature	2.77%	0.628	c.139C>T	p.Arg47Trp	52	0.000856825	33	Disease causing	Possibly damaging	Deleterious	Deleterious	Potential alteration of splicing
POLE	Literature	2.77%	0.628	c.4709G>A	p.Arg1570Gln	52	0	35	Disease causing	Possibly damaging	Deleterious	Deleterious	Probably no impact on splicing
POLE	Literature	2.77%	0.628	c.5492T>C	p.Leu1831Pro	65	2.82598E-05	25.9	Disease causing	Probably damaging	Deleterious	Deleterious	Probably no impact on splicing

[1]Samples selected for tumor patients have been excluded.
[2]Based on the longest gene transcript; NA, not analyzed.

this gene was detected in patient 65, who had been diagnosed with CRC at the age of 38 years and had no family history of cancer. In addition to the *CSMD1* variant, patient 65 carried a VUS in *MSH6* (c.1372C>T;p.His458Tyr) affecting a highly conserved amino acid and a POLE variant (see below).

Germline variants in CRC genes

In total, 16 variants with a CADD score >20 were found in genes with a known association to CRC (all rare variants in CRC and CNV genes—independent of their CADD score—are shown in Supporting Information Table S6). Five patients harbored more than one of these variants.

The known pathogenic *CHEK2* variant (c.1441G>T;p.Asp481Tyr) was found in one patient (no. 47). Patient 47 had been diagnosed with Hodgkin's disease at the age of 40 years, and developed CRC four years later. His father had been diagnosed with prostate cancer. The family history was otherwise unremarkable.

Two patients (no. 22 and no. 24, both without an adenomatous polyposis phenotype) carried each one heterozygous *APC* missense variant. Both variants were predicted to be disease causing by two of three *in silico* programs. Patient 22 was diagnosed with CRC at the age of 32 years. She had no known family history of cancer. In addition to the identified *APC* variant c.1703G>A;p.Ser568Asn in coding exon 13, she was found to carry a VUS in *MSH2* (c.1864C>A;p.Pro622Thr) affecting a highly conserved amino acid. Patient 24 with the *APC* variant c.8425G>A;p.Val2809Met had been diagnosed with EC at the age of 30 years, and with CRC at the age of 63 years. Her family history was strongly suggestive of hereditary CRC, as three maternal relatives had been diagnosed with CRC at a young age (i.e., <50 years). Neither *APC* variant is listed in the LOVD database.

Missense variants were detected in *POLE* (n = 7) and *POLD1* (n = 1) in a total of six patients. Five of these patients had been diagnosed with CRC at <45 years of age. Research has shown that specific missense variants in the exonuclease domains of *POLE* and *POLD1* cause polymerase proofreading-associated polyposis (PPAP). None of the variants found in these six patients lie within this region, or cluster in any other specific region. Patient 52 carried two different *POLE* variants (c.139C>T;p.Arg47Trp, and c.4709G>A;p.Arg1570Gln). He has been diagnosed with CRC at the age of 38 years, and had a maternal family history of breast cancer and CRC. Patient 25 harbored a *POLD1* missense variant (c.2052G>C;p.Gln684His) in addition to the *POLE* variant c.2683G>A;p.Ala895Thr. He had been diagnosed with CRC at the age of 29 years. His grandfather had gastric cancer. The family history was otherwise unremarkable. Patient 65, who had been diagnosed with CRC at the age of 38 years and had no family history of cancer, was found to carry the *POLE* variant (c.5492T>C;p.Leu1831Pro), as well as a possibly pathogenic *VUS* in *MSH6*

(c.1372C>T;p.His458Tyr) and a *CSMD1* missense variant (see above).

In addition to the VUS in *MSH2* (c.1864C>A) and *MSH6* (c.1372C>T) mentioned above, the *MSH6* VUS c.722G>T which had been detected in routine diagnostics was confirmed in patient 46 who had been diagnosed with CRC at the age of 56 years. No family history was available.

MSH2 sequencing in tumor DNA

As a loss of MSH2 in tumor cells can be caused by biallelic somatic variants, *MSH2* and *MSH6* analyses were performed in tumor samples. For 11 patients (for details see Supporting Information Table S1), tumor tissue of sufficient quality was available for *MSH2* and *MSH6* mutation analysis using the TruSight Cancer Panel (Illumina).

In the tumor tissue of three patients, potentially pathogenic mutations were found: Patient 23 harbored two heterozygous *MSH2* missense variants (c.1993C>G;p.His665Asp and c.1835C>T;p.Ser612Leu) in his small bowel cancer tissue. Both variants had a CADD score of >25, indicating a possible pathogenic effect. Patient 6 had a heterozygous *MSH2* frameshift variant (c.1100delT;p.Val367Glufs*6) in the CRC. No second point variant was found on the other allele. To look for a large deletion on the other allele as a second hit, we checked the genotypes of the known SNPs in MSH2 and found three SNPs (c.211+9C>G in exon 1, c.1511-9A>T in exon 10, and c.1661 +12G>A in exon 10) heterozygous in both leukocyte and tumor DNA excluding a large deletion. Patient 1 harbored the heterozygous *MSH6* missense variant c.122C>T;p.Ser41Phe (CADD score 17 in the CRC, however, again no second variant was identified. With the exception of patient 1, whose father and paternal grandmother had also been diagnosed with CRC, the family histories of the 11 patients were unremarkable in terms of LS-associated cancers.

Discussion

The aim of the present study was to identify novel causative genes for LS in a cohort of patients who: (*i*) met the revised Bethesda guidelines; (*ii*) showed a loss of MSH2 expression in their tumor tissue; and (*iii*) had shown no pathogenic germline variant in *MSH2*, *MSH6*, or *EPCAM* during state-of-the-art routine diagnostics (the two large recurrent *MSH2* inversions described in the literature[6,7] were also excluded in most patients). This phenotype, in combination with MSI in tumor tissue, is strongly suggestive of an underlying genetic basis. In a very small number of patients, the phenotype might also be explained by MMR mosaicism, deep intronic germline variants, or pathogenic variants classified to date as VUS.

A two-step study design was selected. In the first step, a genome-wide CNV analysis was performed in 77 unrelated patients. In a second step, 49 selected genes implicated in the CNV analysis, and 30 known or assumed CRC risk genes, were sequenced in a subset of 38 patients (one patient was excluded) using a customized NGS gene panel and Sanger sequencing.

Assuming a dominant or recessive mode of inheritance with high penetrance, the expected frequency of causative CNVs in the general population is much lower than 1%. By applying an established stringent filter and validation workflow, which included comparisons with large control cohorts, 41 unique or rare (i.e., frequency <0.2% in the in-house control dataset) germline CNVs were identified in 39% of the 77 index patients. These comprised 25 different heterozygous deletions and 16 different duplications, all of which were nonrecurrent. Most of the 73 protein coding genes that were affected by CNVs in the present cohort have not been reported to harbor germline alterations in familial cancer patients in previous research.

The majority of patients carried a single CNV. All of the CNVs were ≤500 kb in size, which is consistent with the absence of intellectual disability or other syndromic features in these patients. No rare or unique CNV was identified in the 30 established or putative CRC risk genes, or in the 1.5 Mb region upstream and downstream of *MSH2*. However, one out of 97 putative interaction partners of *MSH2* (*MCM4*) was affected by a partial deletion. These findings are consistent with the few systematic genome-wide CNV screens performed in cohorts with unexplained familial tumor syndromes to date. These investigations identified varying numbers of nonrecurrent rare CNVs, and very little overlap is apparent between studies in terms of the affected genes.[32]

The main approach used to detect novel causative genes for LS in the present study was to detect genes affected by recurrent germline variants, that is, CNVs present in more than one patient or point variants in additional patients. In total, 20 rare point variants with a CADD score of >20 were identified in 12 candidate genes from the CNV analysis (Table 2).

From a functional perspective, *PRKDC*, *PRKCA*, and *MCM4* were the most promising candidate genes identified in the present study. According to the network analysis, PRKDC and MCM4 interact directly with MSH2, although the nature of the interaction remains unclear. Besides their role in the correction of replication errors, MMR genes are also involved in the earliest steps of checkpoint regulation.[31,33] As MCM4 and PRKDC are both involved in cell cycle control, their interaction with MSH2 may be linked to the cell cycle control system. Although previous research has demonstrated that PRKCA is implicated in cell cycle control, the present *in silico* pathway analysis did not demonstrate a significant involvement of PRKCA in any of the cell cycle control pathways. All three genes have low haploinsufficiency and intolerance scores, comparable to those predicted in known monogenic disease genes.

PRKDC encodes a serine/threonine-protein kinase that is involved in the repair of DNA double-strand breaks and nonhomologous end joining. According to the TCGA database,

PRKDC is mutated in 72 out of 537 (13.4%) colorectal carcinomas, indicating a possible causal connection. If all hypermutated tumors are excluded, the number drops to 3.8%. As *PRKDC* is involved in DNA repair and somatic *PRKDC* mutations might be as well jointly responsible for the hypermutation phenotype, it seems not reasonable to exclude these tumors. Biallelic pathogenic germline variants in *PRKDC* have been associated with immunodeficiency 26, which is a severe combined immunodeficiency (SCID) syndrome. In the present study, two *PRKDC* missense variants were detected in patient 31. Polyphen, SIFT, and PROVEAN classified both variants as benign. As the patient had no clinical signs of SCID, biallelic pathogenic *PRKDC* variants are unlikely. However, the possibility that hypomorphic *PRDKC* variants in a biallelic state cause an increased cancer risk without the SCID phenotype cannot be excluded.

PRKDC maps next to *MCM4* on chromosome 8. The detected heterozygous 91 kb deletion in patient 41 spans a part of both genes. Therefore, the additional deletion of *MCM4* in this patient might result in an increased cancer risk. *MCM4* encodes a highly conserved minichromosome maintenance protein, which is essential for eukaryotic genome replication. No additional variants were detected in *MCM4* in the present cohort. However, another gene family member, *MCM3AP*, was affected by germline variants (a partial duplication, a frameshift variant, and two missense variants predicted to be deleterious) in four unrelated patients with a colorectal adenomatosis phenotype in a recent publication by Horpaopan *et al.*[10] Unfortunately, the analysis of samples from the family of patient 41 was not possible.

The functions of the protein encoded by *PRKCA* include roles in cell adhesion and cell cycle checkpoint control. An intronic deletion of *PRKCA* was detected in patient 42, who was diagnosed with CRC at the age of 43 years. Results of the tumor tissue analysis in this patient were not completely reliable, as the respective tissue fragments were very small. Subsequent analysis of a liver metastasis sample revealed no loss of *MSH2*, and so this patient may not in fact have fulfilled the present study inclusion criteria. One additional *PRKCA* variant (an in-frame deletion of one amino acid) was found in patient 45, who had been diagnosed with CRC at the age of 50 years. Masson *et al.*[34] found a deletion of *PRKCA* in a mutation-negative HNPCC patient. No information regarding the phenotype of this patient was provided by the authors.

The most frequently mutated CNV gene in the present cohort was *GRIK4* with heterozygous missense variants in four patients. In addition to the deletion found in the CNV analysis, rare missense variants were detected in four patients during the sequencing step. Three of the variants lie within the ligand binding region of the protein. Given its function as a subunit of a glutamate-gated ionic channel, it is unlikely to be a candidate gene for hereditary CRC, particularly as variants in this gene are only found in around 2.5% of all nonhypermutated CRC. However, the high number of variants detected in this gene is noteworthy.

Two missense variants were found in the CNV genes *NOSIP* (deleted in one patient) and *TRIM41* (partially duplicated in one patient), respectively. The function of the encoded proteins remains unclear. NOSIP has a probable role in neurodevelopment, and TRIM41 may be implicated in protein kinase C signaling. However, neither protein has any obvious relation to cancer development. As the HI score and RVIS also indicate that *NOSIP* and *TRIM41* are prone to variation, they are unlikely to have been disease causing in the present cohort.

The function of the protein encoded by *CSMD1*, which was duplicated in one patient and found to carry a mismatch variant in a second patient, is not fully understood, but is of potential interest. Previous authors have proposed *CSMD1* as a candidate gene for oral and oropharyngeal squamous cell carcinoma.[29,30] Interestingly, somatic variants in *CSMD1* are found in around 11% of nonhypermutated CRC (TCGA data, see Materials and Methods section), suggesting a possible causative role in tumorigenesis. As these are mainly missense variants, a gain rather than a loss of function might contribute to cancer development.

To exclude other hereditary cancer syndromes in the present cohort, which might mimic the LS phenotype, 30 genes with a known or assumed association to CRC were analyzed using NGS. A total of 41 rare variants (16 with a CADD score of >20) were identified in these genes. The most interesting of these genes were *APC* and *POLE/POLD1*.

The causal relevance of the *APC* missense variants that were found in two patients is difficult to assess. In both cases, the clinical history was not indicative of a colorectal adenomatous polyposis. However, the clinical presentation may still represent a very low penetrance phenotype. As very few pathogenic missense variants in *APC* have been reported to date,[35] these variants must be considered to be of unclear significance until appropriate evidence is generated via segregation or functional studies.

Specific somatic or germline missense variants in *POLE* or *POLD1* lead to an impaired exonuclease activity, resulting in ultramutation in tumors and PPAP. When other repair genes, such as *POLE/POLD1*, *MUTYH,* or *NTHL1,* are affected, the tumor might also show MSI and loss of the respective MMR proteins, as shown recently for MSH2/MSH6.[36–38] This phenotypic overlap between PPAP and LS renders clinical discrimination difficult. For the six patients with *POLE* missense variants in the present cohort, the medical records contained no indication of adenomatous polyposis. Five of the six patients had been diagnosed with CRC at <45 years of age. Although none of the variants were located in the exonuclease domain of the respective protein, the identification of rare variants in six patients in a gene that is relatively intolerant to variation (RVIS 9%) and which has a low haploinsufficiency score (11%) is remarkable. Three of the variants affect

evolutionary highly conserved amino acids and three moderately conserved amino acids, which do not cluster in a specific protein domain. To date, only one of the two *POLE* variants detected in patient 52 (i.e., c.139C>T) has been classified as probably benign (ClinVar). The issue of whether missense variants outside the exonuclease domains can contribute to hereditary tumor predisposition remains unclear, and requires further exploration in larger patient cohorts.

The *CHEK2* variant c.1441G>T;p.Asp481Tyr, which was found in patient 47, is a known moderate risk factor for breast cancer, and previous authors have postulated that it is also associated with an increased risk for CRC. Therefore, the *CHEK2* variant may have contributed to the phenotype of patient 47, even if the loss of MSH2 in his CRC remains unexplained.

MSI and lost MMR protein expression in the tumor tissue of LS patients is caused by the combination of a MMR germline variant and a second hit, i. e. a somatic variant that inactivates the second allele of the respective gene in the tumor cells. Nevertheless, two pathogenic somatic variants affecting the two alleles of one MMR gene generally have the same effect. Several recent publications have shown that this is not uncommon in MSH2 and MSH6 negative tumors, particularly in the presence of an additional *POLE* or *POLD1* variant or biallelic *MUTYH* variants.[37,39–42] No two clearly pathogenic somatic MMR variants were found in any of the 11 tumors available for analysis in the present study. However, the small bowel cancer of patient 23 may be attributable to the two heterozygous somatic *MSH2* missense variants in his tumor tissue as the late age at diagnosis (81 years) and his family history, which is atypical for LS, argue against the presence of an underlying LS. Similarly, it is possible that the tumor of patient 6—who was diagnosed with CRC at the age of 36 years, had no family history of LS associated tumors, and carried a pathogenic somatic *MSH2* variant—may carry an additional somatic *MSH2* variant or epimutation that could not be detected by the presently applied method. The heterozygous *MSH6* missense variant c.122C>T;p.Ser41Phe (CADD score 17) in the tumor tissue of patient 1 must be considered a VUS at the present time, and is unlikely to have caused the loss of MSH2/MSH6 in the tumor. Therefore, 2 out the 11 analyzed tumors may be attributable to somatic *MSH2* variants. Somatic changes are insufficient to explain the remaining familial cases, where no germline variant could be found. In addition, patients with a very young age of onset are unlikely to have gathered biallelic somatic variants in their tumor cells. Nonetheless, familiarity and young age of onset suggest strong underlying genetic effects, and argue for the existence of as yet undiscovered novel genetic causes for LS.

The present study had two main limitations. First, as a result of the very specific phenotype of interest, the sample size was relatively small despite being drawn from the largest published cohorts of its type to date. Therefore, the CNVs detected in the initial CNV analysis probably account for only a fraction of all rare CNVs present in the genome of patients with suspected LS and MSH2 loss. Second, the possibility that small CNVs affecting single exons may have been overlooked due to the resolution of the bead array cannot be excluded.

To our knowledge, the present study is the first genome-wide CNV analysis to be performed in a specific subgroup of mutation-negative HNPCC patients. Although Talseth-Palmer *et al.*[43] conducted a CNV analysis in mutation-positive LS cases and HNPCC patients with no known pathogenic variant, they did not distinguish between these two groups, or consider different MMR expression profiles in the tumor tissue. The present study is therefore not comparable to this analysis, which may partly explain the absence of an overlap in terms of the identification of CNVs. Masson *et al.*[34] also analyzed 125 mutation-negative HNPCC patients without distinguishing between MSI positive and negative samples. As mentioned above, the authors found a *PRKCA* deletion in one patient as well as a partial duplication of the functionally related *PRKCI* gene in two other patients.

The causal relevance of *NRG3* is difficult to judge. It is a ligand to the ERBB4 receptor, is implicated in the proliferation and differentiation of neuroblasts, and has been discussed as a possible susceptibility gene for schizophrenia. Although we found a germline alteration (partial deletion) in one patient only, Masson *et al.*[34] identified a CNV affecting the *NRG3* gene in four of 165 patients including one partial deletion. However, the fact that (*i*) so far no pathogenic *NRG3* germline point mutations have been published, (*ii*) three of the four CNVs detected by Masson *et al.* are partial duplications, lying in a region commonly affected by partial duplications, (*iii*) the evidence for haploinsufficiency is low (high HI score), and (*iv*) *NRG3* has not been implicated to date in cancer development, argues against a relevant contribution to the phenotype.

In a case–control investigation conducted without reference to MSI and immunohistochemistry status, Yang *et al.*[44] investigated CNVs in a cohort of familial CRC cases and found a rare structural variation at 12p12.3 in cases. No CNVs were detected in this region in the present analyses.

In conclusion, CNV analysis and subsequent NGS analysis of selected candidate genes in mutation-negative HNPCC patients with a loss of MSH2 in tumor tissue revealed novel candidate genes for LS—in particular *PRKCA*, *PRKDC*, *MCM4*, and *CSMD1*—and suggested that *POLE* variants outside the exonuclease domain might act as etiological factors for hereditary CRC. Analyses in larger cohorts of patients with clinically suspected LS recruited via international collaborations are warranted to verify the present findings.

Acknowledgements

Since its foundation in 1999, the German HNPCC Consortium has received continuous funding from the German Cancer Aid. The funding institution had no influence on the study design, data collection, data analysis, data interpretation, the writing of the report, or the decision to

submit the paper for publication. The present study was supported by German Cancer Aid grant no. 190370. The HNR study is supported by the Heinz Nixdorf Foundation (Germany) and additionally by the German Ministry of Education and Science and the German Research Council (DFG; Project SI 236/8-1, SI236/9-1, ER 155/6-1).

Web resources

Combined Annotation Dependent Depletion (CADD), http://cadd.gs.washington.edu/home

ExAC Browser, http://exac.broadinstitute.org

Ensembl (release 54), http://may2009.archive.ensembl.org/index.html

GeneCards, http://www.genecards.org

HGMD, http://www.hgmd.cf.ac.uk

Human Splicing Finder 3.0, http://www.umd.be/HSF3/

LOVD, http://chromium.lovd.nl/LOVD2/colon_cancer/home.php?select_db=APC

MutationTaster, http://www.mutationtaster.org

NCBI, http://www.ncbi.nlm.nih.gov/

PolyPhen-2, http://genetics.bwh.harvard.edu/pph2/

PROVEAN, provean.jcvi.org/index.php

RefSeq, http://www.ncbi.nlm.nih.gov/RefSeq

Residual Variation Intolerance Score (RVIS), http://genic-intolerance.org/

The Cancer Genome Atlas, https://cancergenome.nih.gov/abouttcga

UCSC Genome Browser, http://genome.ucsc.edu/cgi-bin/hgGateway

Unigene, http://www.ncbi.nlm.nih.gov/unigene

REFERENCES

1. Lynch HT, Snyder CL, Shaw TG, et al. Milestones of lynch syndrome: 1895-2015. *Nat Rev Cancer* 2015;15:181–94.
2. Umar A, Boland CR, Terdiman JP, et al. Revised Bethesda guidelines for hereditary nonpolyposis colorectal cancer (lynch syndrome) and microsatellite instability. *J Natl Cancer Inst* 2004;96:261–8.
3. Steinke V, Holzapfel S, Loeffler M, et al. Evaluating the performance of clinical criteria for predicting mismatch repair gene mutations in lynch syndrome: a comprehensive analysis of 3,671 families. *Int J Cancer J Int Cancer* 2014;135:69–77.
4. Mangold E, Pagenstecher C, Friedl W, et al. Spectrum and frequencies of mutations in MSH2 and MLH1 identified in 1,721 German families suspected of hereditary nonpolyposis colorectal cancer. *Int J Cancer* 2005;116:692–702.
5. Vasen HF, Watson P, Mecklin JP, et al. New clinical criteria for hereditary nonpolyposis colorectal cancer (HNPCC, lynch syndrome) proposed by the international collaborative group on HNPCC. *Gastroenterology* 1999;116:1453–6.
6. Wagner A, van der Klift H, Franken P, et al. A 10-Mb paracentric inversion of chromosome arm 2p inactivates MSH2 and is responsible for hereditary nonpolyposis colorectal cancer in a north-American kindred. *Genes Chromosomes Cancer* 2002;35:49–57.
7. Rhees J, Arnold M, Boland CR. Inversion of exons 1-7 of the MSH2 gene is a frequent cause of unexplained lynch syndrome in one local population. *Fam Cancer* 2014;13:219–25.
8. Colella S, Yau C, Taylor JM, et al. QuantiSNP: an objective Bayes hidden-Markov model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res* 2007;35:2013–25.
9. Priebe L, Degenhardt F, Strohmaier J, et al. Copy number variants in German patients with schizophrenia. *PloS One* 2013;8:e64035.
10. Horpaopan S, Spier I, Zink AM, et al. Genome-wide CNV analysis in 221 unrelated patients and targeted high-throughput sequencing reveal novel causative candidate genes for colorectal adenomatous polyposis. *Int J Cancer J Int Cancer* 2015;136:E578–89.
11. Schmermund A, Möhlenkamp S, Stang A, et al. Assessment of clinically silent atherosclerotic disease and established and novel risk factors for predicting myocardial infarction and cardiac death in healthy middle-aged subjects: rationale and design of the Heinz Nixdorf RECALL study. Risk factors, evaluation of coronary calcium and lifestyle. *Am Heart J* 2002;144:212–8.
12. Chen S, Wang W, Lee S, et al. Prediction of germline mutations and cancer risk in the lynch syndrome. *JAMA* 2006;296:1479–87.
13. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(−Delta Delta C(T)) method. *Methods San Diego Calif* 2001;25:402–8.
14. Engels H, Wohlleber E, Zink A, et al. A novel microdeletion syndrome involving 5q14.3-q15: clinical and molecular cytogenetic characterization of three patients. *Eur J Hum Genet EJHG* 2009;17:1592–9.
15. Franceschini A, Szklarczyk D, Frankild S, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* 2013;41:D808–15.
16. Kent WJ, Sugnet CW, Furey TS, et al. The human genome browser at UCSC. *Genome Res* 2002;12:996–1006.
17. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 2009;4:1073–81.
18. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* 2013; Chapter 7:Unit7.20.
19. Schwarz JM, Cooper DN, Schuelke M, et al. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods* 2014;11:361–2.
20. Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinforma Oxf Engl* 2015;31:2745–7.
21. Desmet F-O, Hamroun D, Lalande M, et al. Human splicing finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res* 2009;37:gkp215.
22. Kircher M, Witten DM, Jain P, et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014;46:310–5.
23. Petrovski S, Wang Q, Heinzen EL, et al. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet* 2013;9:e1003709.
24. Huang N, Lee I, Marcotte EM, et al. Characterising and predicting Haploinsufficiency in the human genome. *PLoS Genet* 2010;6:e1001154.
25. Byron A, Humphries JD, Craig SE, et al. Proteomic analysis of α4β1 integrin adhesion complexes reveals α-subunit-dependent protein recruitment. *Proteomics* 2012;12:2107–14.
26. von Eyss B, Maaskola J, Memczak S, et al. The SNF2-like helicase HELLS mediates E2F3-dependent transcription and cellular transformation. *EMBO J* 2012;31:972–85.
27. Karmakar S, Mahajan MC, Schulz V, et al. A multiprotein complex necessary for both transcription and DNA replication at the β-globin locus. *EMBO J* 2010;29:3260–71.
28. Li T, Zhang J, Zhu H, et al. Proteomic analysis of differentially expressed proteins involved in peel senescence in harvested mandarin fruit. *Front Plant Sci* 2016;7:725.
29. Scholnick SB, Richter TM. The role of CSMD1 in head and neck carcinogenesis. *Genes Chromosomes Cancer* 2003;38:281–3.
30. Toomes C, Jackson A, Maguire K, et al. The presence of multiple regions of homozygous deletion at the CSMD1 locus in oral squamous cell carcinoma question the role of CSMD1 in head and neck carcinogenesis. *Genes Chromosomes Cancer* 2003;37:132–40.
31. Jiricny J. The multifaceted mismatch-repair system. *Nat Rev Mol Cell Biol* 2006;7:335–46.
32. Krepischi ACV, Pearson PL, Rosenberg C. Germline copy number variations and cancer predisposition. *Future Oncol Lond Engl* 2012;8:441–50.
33. Bak ST, Sakellariou D, Pena-Diaz J. The dual nature of mismatch repair as antimutator and mutator: for better or for worse. *Front Genet* [Internet] 2014; 5, 287. Available from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4139959/
34. Masson AL, Talseth-Palmer BA, Evans T-J, et al. Copy number variation in hereditary non-polyposis colorectal cancer. *Genes* 2013;4:536–55.

35. Azzopardi D, Dallosso AR, Eliason K, et al. Multiple rare nonsynonymous variants in the adenomatous polyposis coli gene predispose to colorectal adenomas. *Cancer Res* 2008;68: 358–63.
36. Elsayed FA, Kets CM, Ruano D, et al. Germline variants in POLE are associated with early onset mismatch repair deficient colorectal cancer. *Eur J Hum Genet EJHG* 2015;23: 1080–1084.
37. Jansen AM, van Wezel T, van den Akker BE, et al. Combined mismatch repair and POLE/POLD1 defects explain unresolved suspected lynch syndrome cancers. *Eur J Hum Genet EJHG* 2016;24:1089–92.
38. Morak M, Heidenreich B, Keller G, et al. Biallelic MUTYH mutations can mimic lynch syndrome. *Eur J Hum Genet EJHG* 2014; 22:1334–7.
39. Haraldsdottir S, Hampel H, Tomsic J, et al. Colon and Endometrial cancers with mismatch repair deficiency can Arise from somatic, rather than Germline, mutations. *Gastroenterology* 2014;147:1308–1316.e1.
40. Lefevre JH, Colas C, Coulet F, et al. MYH biallelic mutation can inactivate the two genetic pathways of colorectal cancer by APC or MLH1 transversions. *Fam Cancer* 2010;9:589–94.
41. Vargas-Parra GM, González-Acosta M, Thompson BA, et al. Elucidating the molecular basis of MSH2-deficient tumors by combined germline and somatic analysis. *Int J Cancer* 2017;141:1365–80.
42. Jansen AML, Geilenkirchen MA, van Wezel T, Jagmohan-Changur SC, Ruano D, van der Klift HM, van den Akker BEWM, Laros JFJ, van Galen M, Wagner A, Letteboer TGW, Gómez-García EB, et al. Whole gene capture analysis of 15 CRC susceptibility genes in suspected lynch syndrome patients. *PLoS One* [Internet] 2016; 11: e0157381. Available from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4907507/
43. Talseth-Palmer BA, Holliday EG, Evans T-J, et al. Continuing difficulties in interpreting CNV data: lessons from a genome-wide CNV association study of Australian HNPCC/lynch syndrome patients. *BMC Med Genomics* 2013;6:10.
44. Yang R, Chen B, Pfütze K, et al. Genome-wide analysis associates familial colorectal cancer with increases in copy number variations and a rare structural variation at 12p12.3. *Carcinogenesis* 2014;35:315–23.

THE DIFFERENCE OF BREAKTHROUGH MOMENTS

WITH COMPLETE SOLUTIONS FOR GROUNDBREAKING DISCOVERIES FROM A TRUSTED PARTNER.

Your next breakthrough could be closer than you imagine, especially with the right resources to help you advance your research. At BD, we are dedicated to helping you get the data you need, when, where and how you need it. Our integrated solutions in instrumentation, software and reagents are optimized to work together to bring you closer to your next breakthrough moment. And you can depend on us for world-class training, service and support to help you get the most from the results your research depends on. Discover a range of optimized solutions that open endless possibilities for your future research. **Discover the new BD.**

Learn how you can advance your research >

BD