



# Targeted deep-intronic sequencing in a cohort of unexplained cases of suspected Lynch syndrome

Anke Marie Arnold<sup>1,2</sup> · Monika Morak<sup>1,3</sup> · Anna Benet-Pagès<sup>1</sup> · Andreas Laner<sup>1</sup> · Dimitrij Frishman<sup>2,4</sup> · Elke Holinski-Feder<sup>1,3</sup>

Received: 26 February 2019 / Revised: 26 September 2019 / Accepted: 27 September 2019 / Published online: 10 December 2019  
© The Author(s), under exclusive licence to European Society of Human Genetics 2019

## Abstract

Lynch syndrome (LS) is caused by germline defects in DNA mismatch repair (MMR) pathway, resulting in microsatellite instability (MSI-H) and loss of immunohistochemical staining (IHC) of the respective protein in tumor tissue. However, not in all clinically suspected LS patients with MSI-H tumors and IHC-loss, causative germline alterations in the MMR genes can be detected. Here, we investigated 128 of these patients to possibly define new pathomechanisms. A search for large genomic rearrangements and deep-intronic regulatory variants was performed via targeted next-generation sequencing (NGS) of exonic, intronic, and chromosomal regions upstream and downstream of *MLH1*, *MSH2*, *MSH6*, *PMS2*, *MLH3*, *MSH3*, *PMS1*, and *EPCAM*. Within this cohort, two different large rearrangements causative for LS were detected in three cases, belonging to two families (2.3%). The sensitivity to detect large rearrangements or copy number variations (CNV) was evaluated to be 50%. In 9 of the 128 patients (7%), previously overlooked pathogenic single-nucleotide variants (SNV) and two variants of uncertain significance (VUS) were identified in *MLH1*, *MSH2*, and *MSH6*. Pathogenic aberrations were not found in *MLH3*, *MSH3*, and *PMS1*. A potential effect on regulation was exerted for 19% of deep-intronic SNVs, predominantly located in chromosomal regions where the modification of histone proteins suggests an enhancer function. In conclusion, conventional variation analysis of coding regions is missing rare genomic rearrangements, nevertheless they should be analyzed. Assessment of deep-intronic SNVs is so far non-conclusive for medical questioning.

---

These authors contributed equally: Dimitrij Frishman, Elke Holinski-Feder

**Supplementary information** The online version of this article (<https://doi.org/10.1038/s41431-019-0536-9>) contains supplementary material, which is available to authorized users.

✉ Elke Holinski-Feder  
elke.holinski-feder@mgz-muenchen.de

<sup>1</sup> Medizinisch Genetisches Zentrum—MGZ, Munich, Germany

<sup>2</sup> Department of Bioinformatics, Wissenschaftszentrum Weihenstephan, Technische Universität München, Freising, Germany

<sup>3</sup> Medizinische Klinik und Poliklinik IV, Campus Innenstadt, Klinikum der Universität München, Munich, Germany

<sup>4</sup> Laboratory of Bioinformatics, RASA Research Center, St. Petersburg State Polytechnic University, St. Petersburg, Russia195251

## Introduction

Colorectal cancer (CRC) is the third most common cancer and cause of cancer-related death worldwide [1]. Although the majority of individuals who develop CRC have sporadic disease, up to 20% may have inherited a genetic predisposition [2]. Lynch syndrome (LS) is the most common form of hereditary CRC and accounts for 1–3% of patients with CRC or hereditary nonpolyposis colorectal cancer (HNPCC)-associated tumors (endometrium, small intestine, urinary system) [3]. This autosomal dominant disorder is caused by germline defects in DNA mismatch repair (MMR) genes. It is characterized by the development of colorectal cancer, as well as endometrial cancer and various other cancers at a young age [4]. MMR genes implicated in LS include *MLH1*, *MSH2*, *MSH6*, and *PMS2* [3, 5]. Abnormalities in the function of these MMR genes lead to errors during DNA replication, in particular microsatellite instability (MSI) [6]. LS tumors typically show high MSI and an immunohistochemical (IHC) loss of the respective MMR protein expression. For ~65% of the patients

documented in the German Consortium of Familial Colon Cancer that fulfill the clinical criteria for LS, a pathogenic germline variant in an MMR gene, classified as “class 4/5” by the InSiGHT Mismatch Repair Gene Variant Classification Criteria in the LOVD [7] confirms the diagnosis [8]. Copy number variations (CNV) and large rearrangements account for 20% of the known disease causing variants in LS [9]. However, the detection of either a germline variant of uncertain significance (VUS; 17%), or no relevant variant at all (18%) in the MMR genes leaves one-third of these cases unsolved [8, 10]. Therefore, for a considerable number of patients with suspected LS, the causative genetic predisposition cannot be identified. With a view to uncover further uncommon pathomechanisms, we set up a cohort of 128 patients that are highly suspicious of having LS but remained unsolved in previous genetic tests. We performed exonic and deep-intronic sequencing of eight genes associated with LS or involved in the MMR pathway (*MLH1*, *MSH2*, *MSH6*, *PMS2*, *MLH3*, *MSH3*, *PMS1*, *EPCAM*) and conducted bioinformatics analysis of both coding and non-coding single-nucleotide variants (SNV) and small insertions/deletions (InDel), as well as of large genomic rearrangements.

## Materials and methods

### Patients and control cohort

#### Patient cohort

All patients provided their informed consent for cancer genetics research compliant with ethical standards. The cohort of 128 patients suspected of hereditary colorectal cancer was defined by the following criteria: (i) at least one of the Bethesda criteria met (of those, 32 had family history of CRC, and 15 were Amsterdam criteria positive), (ii) MMR-deficient tumors with loss of *MLH1* and *PMS2*, *MSH2* and *MSH6*, *PMS2* or *MSH6*, and (iii) no causative variation detected in the respective gene(s) showing deficiency in the tumor. Germline analysis in this cohort was performed by denaturing high performance liquid chromatography (DHPLC) or Sanger sequencing and MLPA based on IHC results [11–14]. For cases with IHC loss of *MLH1*, promoter methylation was excluded in tumors.

#### Control cohort

We sequenced the DNA of 13 individuals that are known to have gross deletions/duplications or complex rearrangements in the four major MMR genes, previously detected by MLPA or long-range PCRs. In two of the control cases, the exact rearrangement breakpoints are already determined,

whereas for the remaining 11 individuals only the deleted or duplicated exons are known. Table 1 lists the rearrangements of the 13 control individuals, which were used as a test set for the establishment of the rearrangement pipeline.

## Deep-intronic sequencing

### Panel design

We used a custom-designed Agilent SureSelectXT enrichment kit (Agilent Technologies, Santa Clara, California, USA), which covers the following genes associated with LS or involved in the MMR pathway: *MLH1*, *MSH2*, *MSH6*, *PMS2*, *MLH3*, *PMS1*, *MSH3*, and *EPCAM*. To be able to detect breakpoints of chromosomal rearrangements and regulatory variations, we included all exonic and intronic regions of the eight genes, as well as large parts of their upstream and downstream regions. In the UCSC Genome Browser [15], we defined the flanking regions of the longest transcripts by assessing DNase sensitivity and conserved transcription factor-binding sites. Thereby, the target region comprises 96 kb upstream of *EPCAM* through *MSH2* up to 36 kb downstream of *MSH6*, 23 kb upstream to 26 kb downstream of *PMS1*, 71 kb upstream to 23 kb downstream of *MLH1*, 23 kb upstream to 88 kb downstream of *MSH3*, 14 kb upstream to 30 kb downstream of *PMS2* and 11 kb upstream to 8 kb downstream of *MLH3*. By adding these large-flanking regions to the design, additional genes located in these regions are also covered. Figure 1 displays the exact regions contained in the kit. In total, the target region covers 1.3 Mb of the human genome.

### Library preparation

DNA was obtained from peripheral blood of all patients. For panel enrichment approximately 85 ng genomic DNA was required. Sample preparation was conducted following the manufacturer’s instructions.

### Next-generation sequencing (NGS)

Sequencing was carried out on an Illumina NextSeq instrument (Illumina, San Diego, California, USA) as 150 bp paired-end runs with V2 chemistry. Fifteen samples were pooled for a sequencing run to ensure a sequencing depth of at least 20-fold for the complete target region.

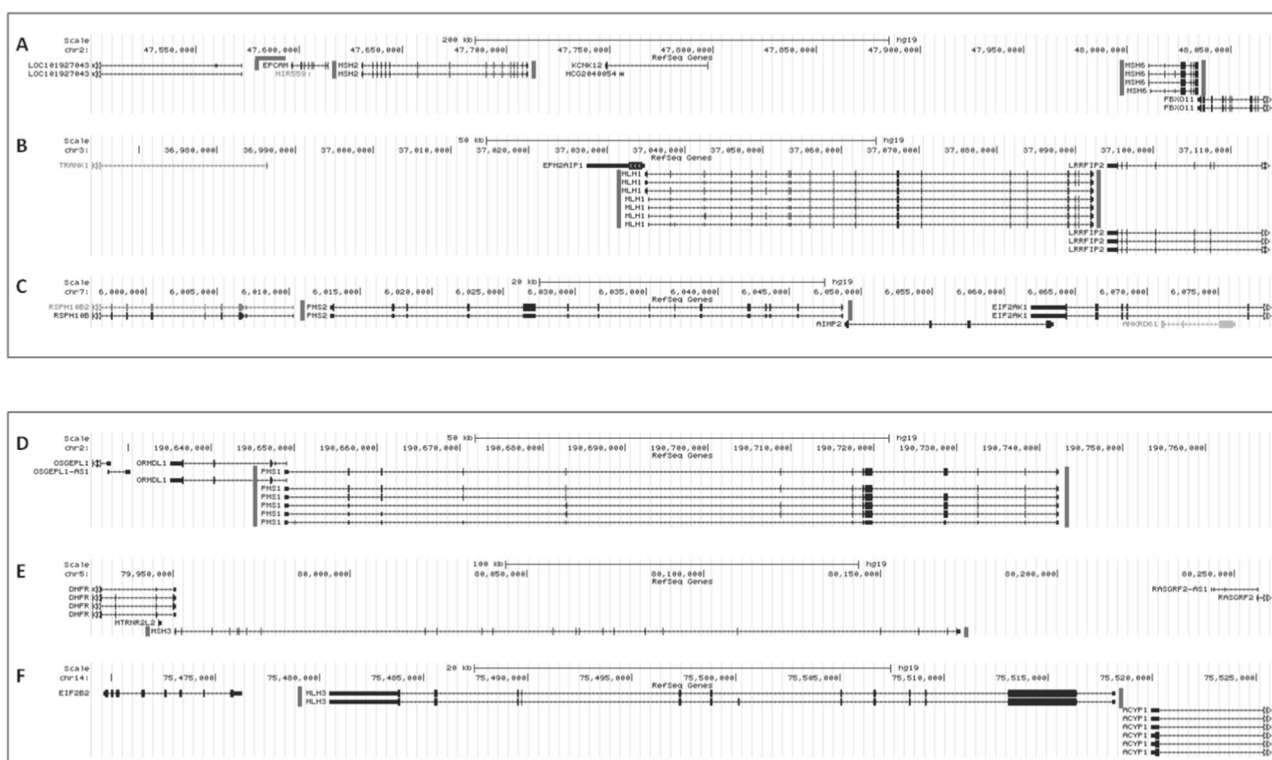
### Alignment, genotype calling, variant annotation, and filtering of NGS data

Demultiplexing of the sequencing reads were performed using the *bcl2fastq2* (v2.15.0.4) software from Illumina (downloaded from <https://support.illumina.com/sequencing/>)

**Table 1** CNV and rearrangement events representing the control cohort.

Sample ID	IHC loss	Event	Breakpoints
50823	<i>MLH1/PMS2</i>	<i>Inversion MLH1 / LRRFIP2</i>	<i>NC_000003.11(NM_000249.3):c.1731 + 2144_(NM_006309.2):c.2055 + 224inv</i> BPs: <i>chr3:37085966 (&gt; 30 reads) &amp; chr3:37095678 (&gt; 30 reads)</i>
54450	<i>MLH1/PMS2</i>	Deletion <i>MLH1 E9-E15</i>	BPs: <i>chr3:37054432 (&gt; 30 reads) &amp; chr3:37087879 (&gt; 30 reads)</i>
75637	<i>MLH1/PMS2</i>	Deletion <i>MLH1 E15</i>	BPs: <i>chr3:37082201 (30 reads) &amp; chr3:37084519 (&gt; 30 reads)</i>
79032	<i>MSH6</i>	Deletion <i>MSH6 E5-E6</i>	BPs: <i>chr2:48030352 (30 reads) &amp; chr2:48032578 (27 reads)</i>
83746	<i>MSH2/MSH6</i>	Duplikation <i>MSH2 E5-E16</i>	BPs: <i>chr2:47640781 (&gt; 30 reads) &amp; chr2:47711090 (19 reads)</i> Not detected
43564	<i>MLH1/PMS2</i>	Duplikation <i>MLH1</i>	Not detected
45524	<i>MSH2/MSH6</i>	Fusion <i>EPCAM E3-E8 &amp; MSH2 E1-E8</i>	Not detected
54695	<i>MSH2/MSH6</i>	Fusion <i>EPCAM E6-E9</i>	BPs: <i>chr2:47611616 (1 read) &amp; chr2:47628425 (3 reads)</i> Not detected (BPs: <i>chr3:37085695 &amp; chr3:37055515</i> )
58915	<i>MLH1/PMS2</i>	Duplikation <i>MLH1 E9-E15</i>	Not detected
62454	<i>PMS2</i>	Deletion <i>PMS2 E9-E12</i>	Not detected
72394	<i>PMS2</i>	Deletion <i>PMS2 E11-E15</i>	Not detected
87183	<i>MSH2/MSH6</i>	Duplikation <i>MSH2 E3-E7</i>	Not detected
49738	<i>PMS2</i>	Deletion <i>PMS2 E1-E10</i>	Not detected

In two samples the exact breakpoints were known from previous analyses (italic). With the deep-intronic sequencing kit we discovered the so far unknown breakpoints five control samples (BPs are listed). The chromosomal positions are with regard to hg19



**Fig. 1 Target region of the deep-intronic sequencing kit.** The six tracks show the complete chromosomal region that is covered by the custom-designed kit. The four genes associated with LS (a–c) are highlighted in red and the four genes involved in the MMR pathway (d–f) are highlighted in green. The target region includes the following

chromosomal positions: chr2:47,500,000–48,070,000; chr2:190,625,643–190,768,067; chr3:36,964,000–37,115,000; chr5:79,927,190–80,260,701; chr7:5,996,249–6,078,750; chr14:75,469,099–75,525,752 (color figure online).

sequencing\_software/bcl2fastq-conversion-software.html). Reads were aligned to the human reference genome (GRCh37/g19) [16] using BWA (v 0.7.8-r455) [17] with standard parameters. Duplicate reads and reads that did not map unambiguously were removed with Picard Tools (downloaded from <http://broadinstitute.github.io/picard>). The percentage of reads overlapping targeted regions and coverage statistics of targeted regions were calculated using Shell scripts. SNVs and small InDels were called using SAMtools (v1.1) [18]. The following parameters were used: a maximum read depth of 10000 (parameter-d), a maximum per sample depth of 10000 for InDel calling (parameter-L), adjustment of mapping quality (parameter-C), and recalculation of per-base alignment quality (parameter-E). Additionally, SNVs were required to fulfill the following criteria: a minimum of 20% of reads showing the variant base and presence in reads originating from different strands. For InDels at least 15% of reads covering this position were required to indicate the InDel. Variant annotation was performed with snpEff (v 4.0e) [19] and Alamut-Batch (v 1.3.1) according to the following RefSeq genomic/transcript sources: NG\_007109.2/NM\_000249.3 (*MLH1*), NG\_007110.2/NM\_000251.2 (*MSH2*), NG\_007111.1/NM\_000179.2 (*MSH6*), NG\_008466.1/NM\_000535.6

(*PMS2*), NG\_008649.1/NM\_001040108.1 (*MLH3*), NG\_016607.2/NM\_002439.4 (*MSH3*), NG\_008648.1/NM\_000534.4 (*PMS1*), and NG\_012352.2/NM\_002354.2 (*EPCAM*). The above mentioned genomic/transcript sources are used for exon numbering. The snpEff nomenclature was adopted for all variant effects except for splice-site variants. According to Cartegni et al. [20], those variants were classified as splice-site defects when located up to 12 bps upstream of an exon or up to 8 bps downstream of an exon, as well as variants that affect coding nucleotides 1–2 at 5'-end of an exon and coding nucleotides 1–3 at 3'-end of an exon. Splice-site prediction was performed with MaxEntScan (MES) [21] and Human Splice-Site Finder (SSF) [22]. Splice-site variants fulfilling both of the following criteria are considered as conspicuous: (i) MES: difference between wildtype and variant >15% and (ii) SSF: difference between wildtype and variant >5%. To predict the damaging potential of aminoacid substitutions we used PolyPhen-2 [23], Mutation Taster 2 [24], SIFT [25], and AGVGD [26, 27]. Missense variants are considered as conspicuous if they are predicted to have damaging effect in three of the four in silico prediction algorithms. The class 3, 4, and 5 variants (SNVs and rearrangements) listed in Tables 1 and 2 were submitted to the LOVD3 database

**Table 2** Detected variants with functional or unknown effects in the LS patient cohort.

Sample ID	IHC loss	Gene	Isoform	MOI	Variant	Genotype	ACMG Class
30669	MSH2/MSH6	<i>MSH2</i>	NM_000251.2	AD	c.416del:p.(Asn139Metfs*35)	Heterozygous	Class 4
87083	MSH2/MSH6	<i>MSH2</i>	NM_000251.2	AD	c.211 + 1 G > T	Heterozygous	Class 4
66012	MSH2/MSH6	<i>MSH2</i>	NM_000251.2	AD	c.942 + 3 A > T	Heterozygous	Class 5
KB10471	MSH6	<i>MSH2</i>	NM_000251.2	AD	c.942 + 3 A > T	Heterozygous	Class 5
77187	MLH1/PMS2	<i>MSH6</i>	NM_000179.2	AD	c.3882del p.(Pro1295Leufs*32)	Heterozygous	Class 4
29781	MLH1/PMS2	<i>MSH6</i>	NM_000179.2	AD	c.1421_1422dup p.(Gln475Cysfs*7)	Heterozygous	Class 5
67408	PMS2	<i>MLH1</i>	NM_000249.3	AD	c.799_800delinsAG p.(Val267Arg)	Heterozygous	Class 3
83122	MSH2/MSH6	<i>MSH6</i>	NM_000179.2	AD	c.3604dup p.(Met1202Asnfs*13)	Heterozygous	Class 4
KB9121	MSH2/MSH6	<i>MSH6</i>	NM_000179.2	AD	c.3488_3489del p.(Glu1163Glyfs*8)	Heterozygous	Class 4
80168	MLH1/PMS2	<i>MLH1</i>	NM_000249.3	AD	c.2162 A > G p.(Tyr721Cys)	Heterozygous	Class 3
48120	MLH1/PMS2	<i>MLH1</i>	NM_000249.3	AD	c.-63_-62insCACGAGCA	Heterozygous	Class 4

The first three variants were not detectable in previous DHPLC variation screenings. In four cases (rows 5, 6, 8, 9), analysis of the deep-intronic sequencing data revealed a misleading IHC result. The locus of the last variant was not covered in previous analyzed

and are accessible via the following URL and submission IDs: URL: <https://databases.lovd.nl/shared/individuals/submissionID>; IDs: 00263999, 00263997, 00197203, 00265379, 00263998, 00188512, 00265377, 00187512, 00187024, 00265378, 00188557, 00263983, 00263984, 00263985, 00263986, 00263987, 00263988, 00263989, 00263990, 00263991, 00263992, 00263993, 00263994, 00263995.

### Analysis of rearrangements and CNVs

Large chromosomal rearrangements and breakpoints of CNVs were analyzed with the public tools Pindel (version v0.2.5b8) [28] and Delly (version v0.7.6) [29]. The methods are based on the evaluation of the span and orientation of paired-end reads and the investigation of split sequence-read signatures breaking alignment with the reference. The rearrangement pipeline requires the complete mapping data of a sample, i.e., reads that were flagged as unmapped, not properly paired or duplicated (PCR or optical) are not removed from the mapping files. The rearrangement pipeline is conducted subsequent to demultiplexing and mapping of the SNV analysis pipeline. The results of Pindel and Delly are combined and annotated within a master Shell script.

### Analysis of non-coding variants

We used the functional annotation tool SNIIPA [30] for effect predictions on both the gene and the regulatory level. SNIIPA provides information about regulatory elements, gene trait annotations, as well as variant associations for many variants and their locus context. As SNIIPA only contains annotations for SNVs in phase 3 version 5 of the 1000 genomes project [31], we additionally intersected all

detected non-coding variant calls with the following regulatory genomic regions tracks obtained from the UCSC Table Browser [32]: (1) “*Dnase Cluster: Integrated Regulation from ENCODE Tracks*”, (2) “*H3K4Me1, H3K4Me3, H3K27Ac: Integrated Regulation from ENCODE Tracks*”, (3) “*TFBS: HMR Conserved Transcription Factor Binding Sites*”, and (4) “*miRNA: TargetScan miRNA Regulatory Sites*”.

## Results

### Quality of targeted deep-intronic sequencing data

The target region of the patients and the control cohort showed a mean coverage of 211.91 reads and a median of 201 reads. Only 2.34% of the target region had zero coverage, 90.82% was covered by at least ten reads and 85.71% was covered by at least 20 reads. Overall, we obtained sufficient coverage in both exons, introns and flanking regions for all analyzed individuals (patients and controls) in order to conduct variant calling and rearrangement detection.

### SNV analysis

SNV and InDel analysis of each LS patient resulted in a mean of 1236 (SD 204) calls for each individual, and, out of this number, a mean of 304 (SD 91) SNVs per patient were considered as rare variants with a frequency of <1% in the general population (ExAC data [33]). There were on average 11.4 (SD 2.2) coding variants per individual, of which 4.9 (SD 1.8) were silent variants, 9.6 (SD 1.7) were missense variants, and 1.8 (SD 2.2) showed a strong truncating effect as defined by MES and SSF. In each patient, a mean

**Table 3** Allocation of non-coding variants into putative regulatory regions.

Type	All	Dnase cluster	miRNA	TFBS	H3K4Me1	H3K4Me3	H3K27Ac
splice site region	28	6	0	0	18	17	9
intron	9069	807	0	41	4159	4997	3655
UTR'3 prime	189	46	0	4	114	128	105
UTR'5 prime	57	33	0	4	32	38	34
downstream	489	25	0	5	181	282	213
upstream	392	108	0	11	205	234	161

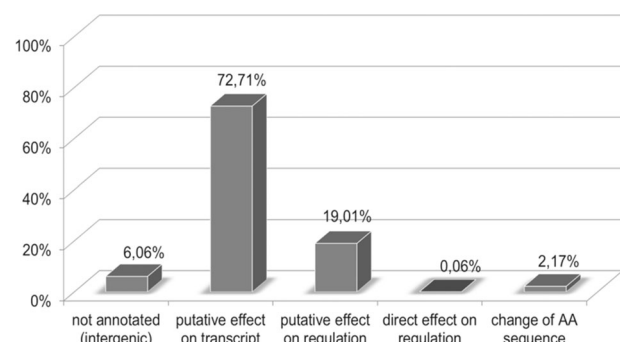
All non-coding variants were assigned to UCSC regulatory region tracks. Most of the variants are located in regions where modification of histone proteins is suggestive of enhancer or other regulatory activity

of 1127 (SD 199) non-coding variants were detected, with 2.5 (SD 1) of these located at splice-sites as defined by Cartegni et al. [20], 1031 (SD 174) in the introns, 34 (SD 9) upstream, and 59 (SD 15) downstream of the respective genes. SNV analysis of the 128 LS samples solved the phenotype of LS in nine patients by detecting class 4 or class 5 variants in the coding regions of the MMR genes and revealed two variants with uncertain significance (VUS). These 11 variants were missed initially by DHPLC or Sanger sequencing or due to misleading IHC results. Table 2 lists the class 3, 4, and 5 variants, which include five frame-shift variants, three SNVs affecting a splice-site, one promoter insertion, and two consecutive missense variants. The nine causative variants and the two VUS only affected three genes: *MSH2*, *MSH6*, and *MLH1*, whereas no clear variants affecting function could be detected in *PMS2* or the other targeted genes, *MLH3*, *MSH3*, and *PMS1*. In four patients, the disease causing variant was overseen, because a misleading IHC result indicated a defect of a specific MMR gene, which was then analyzed selectively, whereas the gene harboring the variant affecting the function was not analyzed (Table 2, rows 5, 6, 8, 9). IHC was repeated for these cases; the results of IHC then matched with the genes where disease causing variants associated with LS were detected. *MSH2* variants c.211 + 1 G > T, c.416delA p.(Asn139Ilefs\*35) and c.942 + 3 A > T were not detectable in previous DHPLC variation screening [14], confirming an incomplete variation detection rate of ~90%. The patient with variant *MLH1* c.-63\_-62insCACGAGCA was externally analyzed by Sanger Sequencing and analysis did not cover the location of the variant. The patients KB10471 and 67408 (Table 2, lines 4, 7) were also analyzed externally and the variants were probably overlooked because they affect the heterodimers and not the respective gene with IHC loss.

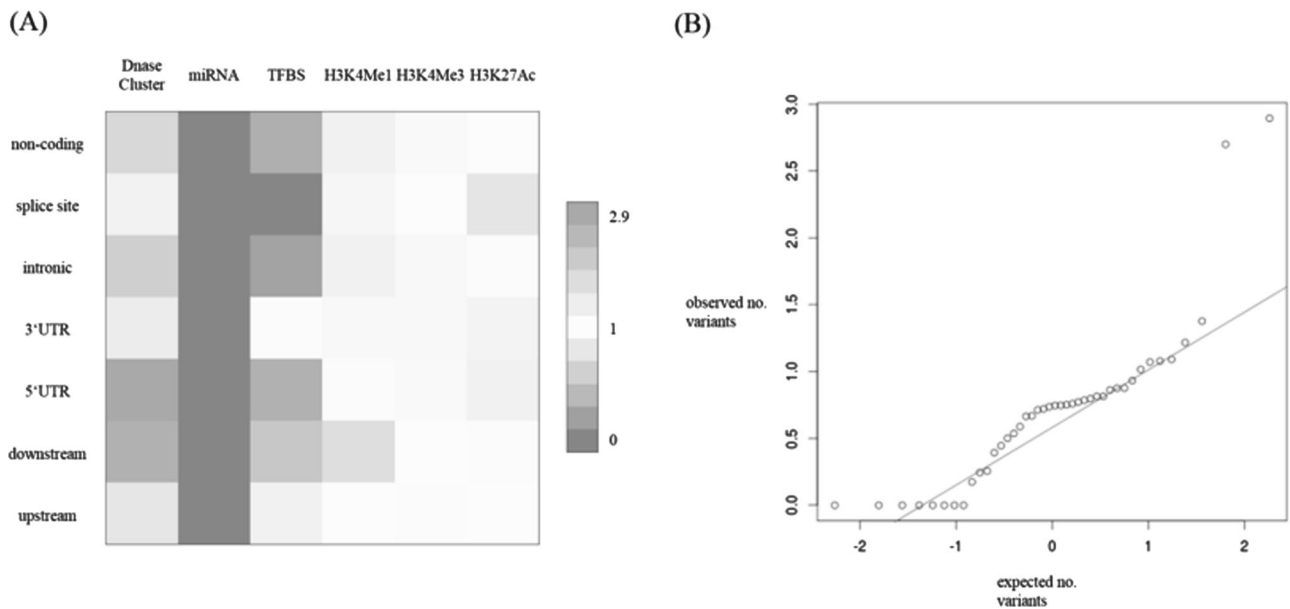
### Regulatory evaluation of deep-intronic variants

Assessing all detected unique variant calls from the whole cohort of patients, a total of 10,224 distinct variants were annotated as non-coding. Of these, 9069 were located in introns, 189 at the 3'UTR, 57 at the 5'UTR, 489

downstream, 392 upstream, and 28 SNVs in splice-site regions of the corresponding coding genes (Table 3). Among the 28 splice-site variants, three SNVs affect intronic nucleotides at +/- 1-2 and 25 were located within regions +/- 3-8. In addition to the 28 non-coding splice-site SNVs, we detected two coding SNVs within the defined splice-site region of two genes previously not associated with LS [*MSH3*:NM\_002439.4:c.235 A > G p.(Ile79Val), *EIF2B2*:NM\_014239.3:c.599 G > T p.(Gly200Val)]. Besides the two splice-site variants that solved the phenotype of LS (Table 2), none of the remaining splice-site variants had a suspicious result in splice-site prediction (MES, SSF). SNVs with a rare (<1%) or no documented population frequency were subjected to comprehensive annotation yielding effect predictions on both the gene and the regulatory level. Figure 2 provides a summary of effect categories as predicted by the SNIIPA functional annotations [30]. The majority (72%; 7361) of the intronic, UTR, or non-coding variants may affect transcript processing; however, there is no further evidence available that would allow for more accurate insights into this matter. For a definite classification of these variants RNA analyses are planned to be performed in a next step. Moreover, according to SNIIPA, 19% (1943) of the variants were predicted to have a putative effect on regulation, as they were mapped into regulatory regions such as promoter/enhancer regions.

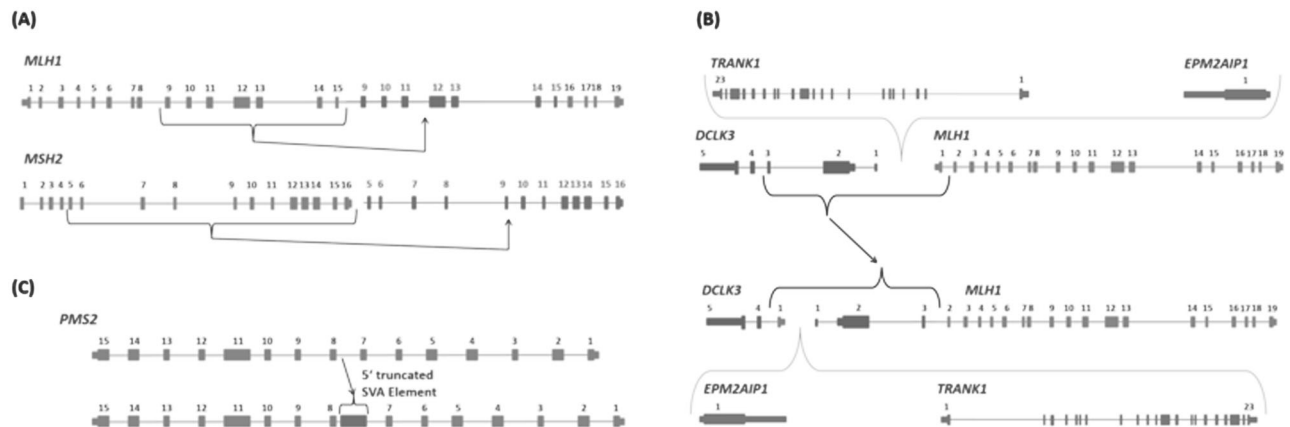


**Fig. 2 Overview of annotations for all coding and deep-intronic SNV calls detected in the whole-patient cohort.** Functional annotations and linkage disequilibrium information for all detected SNVs were received with the SNIIPA tool. The majority of variants are located in intronic regions with a putative effect on transcript.



**Fig. 3 Distribution of non-coding variants in putative regulatory regions.** Ratios of observed versus expected number of variants per regulatory region track are shown in a heatmap (a) and as a QQ-plot (b). It was assumed that the variants are evenly distributed in the target

region and consequently, the expected number of variants was defined by the size of the respective regulatory region track. 5'UTR variants occur more frequently in DNase-hypersensitivity regions and transcription factor-binding sites (b).



**Fig. 4 Results of rearrangement detection.** a Schematic view of two duplication events in the control cohort. In these two cases, rearrangement analysis of the deep-intronic sequencing data revealed the exact breakpoints and, thereby, also the chromosomal locus of the duplicated parts. The green parts are duplicated. b In one sample of the patient cohort, we detected a complex rearrangement on chromosome

2 that causes a paracentric inversion between the *DCLK3* gene and *MLH1*. The same structural variation was identified in two additional family members. c In another patient, we identified a gross insertion (2.2 Kb) of a 5' truncated SVA element in *PMS2* intron 7 (color figure online).

Table 3 gives a detailed description of the mapping of the variants into predicted regulatory regions. A total of 1025 variants were located at DNase I Hypersensitivity sites; the majority of the variants (over 4000) mapped to H3K4Me1, H3K27Ac, or H3K4Me3. A lower number of 65 variants were located at transcription factor-binding sites (TFBS) conserved in the human/mouse/rat alignment. No variants mapped to the TargetScan miRNA Regulatory Sites. Figure 3 shows a heatmap of the ratio of the observed versus expected number of variants in the regulatory region tracks.

The variants were classified based on their consequences for the transcript. We observed an almost three times higher number of 5'UTR variants than expected in regions, where chromatin is hypersensitive to cutting by the DNase enzyme (obs./exp. ratio 2.89) and in regions where transcription factors bind (obs./exp. ratio 2.69). A closer look at these 5'UTR variants revealed no causative effect, as they were either detected in patients with another variant affecting the function or in patients with IHC loss of proteins not corresponding to the genes affected by these 5'UTR variants.

## CNV and rearrangement analysis

### Control cohort

We were able to detect the breakpoints in six of the 13 individuals with known CNVs (Table 1). In one (ID 50823) of these six cases, both detected breakpoints were already known, whereas for the remaining five cases, only the deleted or duplicated exons were known. We thus could determine the exact position of duplicated gene regions in one further patient (ID 83746). In one case (ID 58915), no breakpoints were detected even though breakpoints were known.

### Duplications

In one patient (ID 58915), a gene context disruption was shown before for an *MLH1* duplication, which explains the underlying pathomechanism. The duplicated region affects exon 9 to exon 15, which was inserted in forward orientation in intron 15; the exon sequence in this patient is thus E1-15-E9-15-E16-19 (Fig. 4a). In this case, the breakpoints were known but could not be confirmed with the deep-intronic panel, even though the respective region showed sufficient coverage (>30-fold). In sample ID 83746, with an *MSH2* duplication, IHC loss of *MSH2* in the tumor implies the presence of a defect; however, the pathomechanism is unclear, as the parental *MSH2* transcript is predicted to be unaffected by the duplication. The exact breakpoint could be identified and the affected exons 5 to 16 were duplicated downstream to *MSH2* far after the poly-adenylation signal. Subsequent complementary DNA analyses revealed that aberrant *MSH2* transcripts were generated (Fig. 4a; publication in preparation). In patient ID 43564, a complete *MLH1* duplication had been identified previously. For this sample, the complete target region around *MLH1* showed a one-third/two-thirds allele distribution, which suggests that the breakpoints of the duplication are located outside the designed panel's target region. The duplication breakpoints of *MSH2* exons 3 to 7 in sample ID 87183 could not be detected, despite sufficient coverage of both intronic regions.

### Deletions

Control patient ID 54450 has a deletion of *MLH1* exons 9 to 15. The breakpoints were known and could be confirmed with the deep-intronic panel (Table 1). The same was the case for the *MLH1* and *MSH6* deletions of sample ID 75637 and sample ID 79032: both breakpoints were already known and were confirmed by the rearrangement pipeline with a sufficient coverage of >20 reads. The *EPCAM* deletion of exons 6 to 9 in patient ID 54695 was confirmed with the

deep-intronic panel and, more importantly, the previously unknown breakpoints of the event could be detected and revealed a deletion of the promoter region of *MSH2*. In patient ID 49738, a deletion of exons 1 to 10 of *PMS2* had been detected, but so far it has not been possible to determine the exact breakpoints. The region upstream from *PMS2*, which is targeted by the deep-intronic panel, showed a homozygous variant-allele distribution in this sample, which indicates that the corresponding breakpoint may be located outside the target region. A detailed coverage analysis of *PMS2* intron 10 revealed an uncovered part of 8% in which the second breakpoint may have remained undetected. Similarly, the breakpoints of the *PMS2* deletions in sample ID 72394 and sample ID 62454 may have not been detected because they are located in a region with zero coverage of *PMS2*. Coverage analysis revealed an uncovered proportion of 10%, 5% (*PMS2* downstream, *PMS2* intron 10) and 13%, 10% (*PMS2* intron 8, *PMS2* intron 12) for these two samples. A deletion of *MSH2* exons 1 to 8, which reaches upstream into the *EPCAM* gene, was detected before for sample ID 45524, but the breakpoints could not be identified with the deep-intronic panel. For this sample, the *MSH2* intron 8 region showed four coverage gaps with a size between 300 to 1600 bps. Furthermore, the region upstream from *MSH2* showed only homozygous variants up to the end of the target region.

### Inversions

The control cohort included one patient (ID 50823) with a paracentric inversion of *MLH1* exons 1–15 and *LRRFIP2* exon 29 in-frame, and *LRRFIP2* exons 1–3 fused with *MLH1* exons 16–19 in-frame [34]. The inversion breakpoints in *MLH1* intron 15 and *LRRFIP2* intron 3 with deletion of *LRRFIP2* exons 4–28 were verified successfully with the rearrangement pipeline.

### Patient cohort

In one patient with a combined IHC loss of *MLH1* and *PMS2* in the tumor, we reported a germline heterozygous SNV in *MLH1* exon 8 revealed mono-allelic expression in previous cDNA analyses, indicating an *MLH1* germline defect [13]. In this patient, our rearrangement pipeline detected a complex novel rearrangement on chromosome 2 that results in a paracentric inversion between the genes *DCLK3* and *MLH1* (Fig. 4b). To investigate the recurrence of rearrangements in further unsolved patients suspected of LS we currently perform analysis, which was recently published [35]. *DCLK3* is located about 280 kb upstream from *MLH1* with antisense-orientation. In total, the inversion affects 273,592 bp, including the two antisense genes *TRANK1* and *EPM2AIP1*, which are located between



*DCLK3* and *MLH1*. The breakpoints are located in *MLH1* intron 1 and *DCLK3* intron 3, resulting in two new fusion transcripts: *MLH1* exon 1 and *DCLK3* exons 4–5 in-frame, and *DCLK3* exons 1–3 fused with *MLH1* exons 2–19 in-frame [breakpoints: chr3:36762825 (9 reads) & chr3:37036416 (17 reads)]. The same structural variation was identified in two additional family members.

In another patient with IHC loss of *PMS2*, we detected a 2.2 kb insertion of a 5' truncated SVA element in *PMS2* intron 7 [breakpoints: chr7:6035324 (over 30 reads) & chr7:77138494 (over 30 reads)], which was already described by van der Klift et al. [36] (Fig. 4c). In-silico predictions and cDNA analyses showed that the SVA insertion results in the exonization of a 71 bp fragment, including 14 intronic bp and 57 bp of the 5'-end of the SVA element. The fragment inserted in the transcript causes a premature termination codon that triggers the degradation of the transcript through nonsense-mediated messenger RNA decay.

The two rearrangements we found were consistent with the IHC loss in the tumor and obviously abrogate gene function, therefore we classified these rearrangements as class 5.

## Discussion

The aim of the present study was to evaluate the incidence of large rearrangements and the contribution of sequence variants in the regions of the major and subordinate MMR genes in a cohort of patients who: (i) met the Amsterdam or revised Bethesda guidelines; (ii) showed a loss of gene expression in their tumor tissue for at least one of the MMR proteins; and (iii) had shown no causative germline variant in these MMR genes during state-of-the-art routine diagnostics. In a second step, deep-intronic germline variants were analyzed with regard to their location in regulatory regions.

We were able to solve the phenotype of LS in 11 out of 128 patients that were unsolved in preceding genetic tests. Primarily, the causative DNA alterations were single-nucleotide variants. Of these, 82% were pathogenic SNVs overlooked in the course of routine diagnostics; large chromosomal rearrangements were identified in only 18% of the clarified cases, which is consistent with the general contribution of large rearrangements to known variants associated with LS [9]. All causative variants affected one of the four major MMR genes (*MLH1*, *PMS2*, *MSH2*, *MSH6*), whereas no causative variation was detected in the other genes (*PMS1*, *MLH3*, *MSH3*, *EPCAM*). So far, heterozygous germline variants in additional MMR genes such as *MSH3*, *PMS1*, or *MLH3* alone are not regarded causative for LS [12]. However, biallelic *MSH3* defects have been

identified in two patients with recessive adenomatous polyposis [37]. For *EPCAM*, only large deletions, including the last exon are causative of LS, whereas variants altering an aminoacid in the protein are not associated with LS.

## Sensitivity of rearrangement detection

We were able to determine <50% of the rearrangements breakpoints in the control cohort. One reason for this sensitivity was an uneven coverage in the custom-designed capture kit. As mentioned before, 9.18% of the target region of the deep-intronic panel was covered by ten sequencing reads or less and 2.34% were not covered at all. This uneven coverage might be the consequence of limitations in the bait design due to highly homologous genomic regions where baits could not be placed. Possibly, breakpoints are enriched in these difficult regions and thus could not be detected so far. In the cases of the *PMS2* deletions, the breakpoints may be missed because of ambiguous alignment results due to the many regions (e.g., *PMS2CL*) with high homology to *PMS2*. Additionally, the target region of the custom-designed capture kit is limited and may not cover all the breakpoint genomic regions. If we compare the samples for which we identified the causative variation to those in which we could not detect any breakpoints, there is no discrepancy in overall quality. We were not able to detect breakpoints in some of the best-covered samples, and, conversely, identified clearly disease causing variants in samples with rather poor coverage. One strategy to overcome these issues and improve sensitivity would be whole-genome sequencing (WGS) of the patient cohort. WGS data shows evenly distributed coverage, which facilitates the detection of CNV breakpoints and large chromosomal rearrangements. Moreover, with WGS no limits are imposed on the region to be analyzed and therefore breakpoints cannot be missed due to insufficient target region size [38, 39]. Identification of the exact CNV breakpoints is important, especially for duplications, since the interpretation of a genomic duplication depends heavily on the locus of the duplication event (e.g., outside of the natural genomic context versus disruption of the reading frame of the parental gene). Nevertheless, even if the breakpoints are known, in some cases the underlying pathomechanism can be very difficult to deduce. For sample ID 83746 (see Table 2) we revealed that the duplicated gene (*MSH2*) is intact but the corresponding transcript affected by the duplication; the tumor, showed an IHC loss of *MSH2*. In order to confirm this case, *MSH2* should be sequenced in the tumor tissue for identification of causative somatic variants. As with other studies, targeted capture-based NGS data is a powerful tool for detection of CNV breakpoints and large rearrangements [40], which are indispensable for the identification of the causative variant or mechanism and

should be analyzed with a high and evenly distributed coverage and, if possible, in an unrestricted target region.

### Intronic variants

In a second approach, all rare deep-intronic variants of the patient cohort were analyzed. Most of the variants mapped to H3K4Me1, H3K27Ac and H3K4Me3 regions or DNase I hypersensitivity sites (Table 2). The histone modifications H3K4Me1 and H3K27Ac are quite broad and often indicative of enhancer and other regulatory activity. The H3K4Me3 track shows a histone mark associated with promoters. Regulatory regions, and especially promoters, tend to be DNase-sensitive. We determined an accumulation of observed versus expected 5'UTR variants in DNase-hypersensitivity regions and TFBS; however, these variants were classified as benign or likely benign. Rare intronic variants are of scientific interest, but their contribution to disease is difficult to determine. Montalban et al. [41] identified the first *BRCA1* deep-intronic variant associated with hereditary breast-ovarian cancer syndromes by pseudoexon activation. In another study, Clendenning et al. [42] identified a variant deep within an intron of *MSH2* that causes LS. These studies highlight the importance of investigating non-coding regions by performing additional molecular analyses and utilizing more extensive sequence analysis methods to solve unexplained cases of suspected LS where routine diagnostics fail to determine a causative variation. Intronic variants predicted to create de novo splice-sites or activate cryptic splice-sites should be selected for further characterization, e.g., in patient RNA. RNAseq-based expression data could be useful here in order to reveal new pathomechanisms.

### Unexplained LS suspected cases

In concordance with literature, we failed to identify a causative germline MMR gene variation for the majority of suspected LS cases [43]. Another explanation of the MMR deficiency detected in the tumors may be the presence of biallelic somatic hits in MMR genes [44–46], most probably representing sporadic cases. Also, germline variations in other genes that cause impaired DNA error correction, such as *MUTYH*, *POLE* or *POLD1*, may in some instances cause MMR deficiency in the tumors because of the somatic alteration of MMR genes [47, 48]. For most of the samples, analysis of somatic alterations is not possible as the tumors are not available. Nevertheless, complex rearrangements may have been missed in the unexplained LS cases because they are located outside of the targeted region or may be hidden in regions difficult to analyze, e.g., in common fragile site region, as it was the case in a study of Vogt et al.

[49] where they found a “*PMS2* inactivation by a complex rearrangement involving an HERV retroelement and the inverted 100-kb duplison on 7p22.1”.

### Conclusion

Taken together, complex rearrangements are missed during routine diagnostics covering only exonic and splice junctions, but they occur very seldom. However, their low incidence may be biased by the fact that many events are missed due to the difficulties in detection and the lack of covering the breakpoint regions. Integrating special regions for targeted breakpoint analysis has the advantages it is not as expensive as WGS, and the data can be analyzed much easier and faster. Otherwise, it is challenging to determine all necessary regions to be targeted and, moreover, to achieve uniform coverage on the capture data. SNVs within deep-intronic regions are still non-conclusive for medical questioning. The main reason for this lies in the lack of conclusive annotations about intronic variants, especially their impact on diseases. More extensive sequence analysis methods of intronic regions and additional molecular analyses of their downstream effects are necessary to cope with this lack of information. Therefore, extending routine diagnostics to regions that are known to be prone to breakpoints and well-annotated intronic regions may strike a good balance between not missing causative variation, gain new information about intronic regions and save time, money, and effort.

**Acknowledgements** We want to thank Christina Rapp and Kristina Lenhard for assistance with sample preparation and sequencing.

**Funding** Funded by the German Cancer Aid, project no. 111222.

### Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflict of interest.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

### References

1. Siegel R, Naishadham D, Jemal A. Cancer statistics, 2012. *CA Cancer J Clin.* 2012;62:10–29.
2. Jass JR. What's new in hereditary colorectal cancer? *Arch Pathol Lab Med.* 2005;129:1380–4.
3. Hampel H, Frankel WL, Martin E, Arnold M, Khanduja K, Kuebler P, et al. Screening for the Lynch syndrome (hereditary nonpolyposis colorectal cancer). *N Engl J Med.* 2005;352:1851–60.
4. Lynch HT, de la Chapelle A. Hereditary colorectal cancer. *N Engl J Med.* 2003;348:919–32.

5. Vasen HFA, Möslein G, Alonso A, Bernstein I, Bertario L, Blanco I, et al. Guidelines for the clinical management of Lynch syndrome (hereditary non-polyposis cancer). *J Med Genet.* 2007;44:353–62.
6. Moreira L, Balaguer F, Lindor N, de la Chapelle A, Hampel H, Aaltonen LA, et al. Identification of Lynch syndrome among patients with colorectal cancer. *JAMA.* 2012;308:1555–65.
7. Thompson BA, Spurdle AB, Plazzer J-P, Greenblatt MS, Akagi K, Al-Mulla F, et al. Application of a 5-tiered scheme for standardized classification of 2,360 unique mismatch repair gene variants in the insight locus-specific database. *Nat Genet.* 2014;46:107–15.
8. Mangold E, Pagenstecher C, Friedl W, Mathiak M, Buettner R, Engel C, 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.
9. Hampel H. Genetic testing for hereditary colorectal cancer. *Surg Oncol Clin N Am.* 2009;18:687–703.
10. Müller-Koch Y, Kopp R, Lohse P, Baretton G, Stoetzer A, Aust D, et al. Sixteen rare sequence variants of the hMLH1 and hMSH2 genes found in a cohort of 254 suspected HNPCC (hereditary nonpolyposis colorectal cancer) patients: mutations or polymorphisms? *Eur J Med Res.* 2001;6:473–82.
11. Grabowski M, Müller-Koch Y, Grasbon-Frodl E, Koehler U, Keller G, Vogelsang H, et al. Deletions account for 17% of pathogenic germline alterations in MLH1 and MSH2 in hereditary nonpolyposis colorectal cancer (HNPCC) families. *Genet Test.* 2005;9:138–46.
12. Morak M, Käsbauer S, Kersch M, Laner A, Nissen AM, Benet-Pagès A, et al. Loss of MSH2 and MSH6 due to heterozygous germline defects in MSH3 and MSH6. *Fam Cancer.* 2017;16:491–500.
13. Morak M, Ibsler A, Keller G, Jessen E, Laner A, Gonzales-Fassrainer D, et al. Comprehensive analysis of the *MLH1* promoter region in 480 patients with colorectal cancer and 1150 controls reveals new variants including one with a heritable constitutional *MLH1* epimutation. *J Med Genet.* 2018;55:240–8.
14. Holinski-Feder E, Müller-Koch Y, Friedl W, Moeslein G, Keller G, Plaschke J, et al. DHPLC mutation analysis of the hereditary nonpolyposis colon cancer (HNPCC) genes hMLH1 and hMSH2. *J Biochem Biophys Methods.* 2001;47:21–32.
15. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. *Genome Res.* 2002;12:996–1006.
16. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J. Initial sequencing and analysis of the human genome. *Nature.* 2001;409:860–921.
17. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2010;26:589–95.
18. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. *Bioinformatics.* 2009;25:2078–9.
19. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin).* 2012;6:80–92.
20. Cartegni L, Chew SL, Krainer AR. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet.* 2002;3:285–98.
21. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol.* 2004;11:377–94.
22. Desmet F-O, Hamroun D, Lalande M, Collod-Bérout G, Claustres M, Bérout C. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res.* 2009;37:e67.
23. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P. A method and server for predicting damaging missense mutations. *Nat Methods.* 2010;7:248–9.
24. Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods.* 2014;11:361–2.
25. Sim N-L, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* 2012;40:W452–7.
26. Tavtigian SV, Deffenbaugh AM, Yin L, Judkins T, Scholl T, Samollow PB. Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. *J Med Genet.* 2006;43:295–305.
27. Mathe E, Olivier M, Kato S, Ishioka C, Hainaut P, Tavtigian SV. Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods. *Nucl. Acids Res.* 2006;34:1317–25.
28. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z, Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics.* 2009;25:2865–71.
29. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics.* 2012;28:i333–9.
30. Arnold M, Raffler J, Pfeufer A, Suhre K, Kastenmüller G. SNIpA: an interactive, genetic variant-centered annotation browser. *Bioinformatics.* 2015;31:1334–6.
31. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. *Nature.* 2015;526:68–74.
32. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, et al. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* 2004;32:D493–6.
33. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T. Analysis of protein-coding genetic variation in 60,706 humans. *Nature.* 2016;536:285–91.
34. Morak M, Koehler U, Schackert HK, Steinke V, Royer-Pokora B, Schulmann K, et al. Biallelic MLH1 SNP cDNA expression or constitutional promoter methylation can hide genomic rearrangements causing Lynch syndrome. *J Med Genet.* 2011;48:513–9.
35. Morak M, Schaefer K, Steinke-Lange V, Koehler U, Keinath S, Massdorf T. Full-length transcript amplification and sequencing as universal method to test mRNA integrity and biallelic expression in mismatch repair genes. *Eur J Hum Genet.* 2019;27:1808–20. Epub 2019 Jul 22.
36. van der Klift HM, Tops CM, Hes FJ, Devilee P, Wijnen JT. Insertion of an SVA element, a nonautonomous retrotransposon, in PMS2 intron 7 as a novel cause of Lynch syndrome. *Hum Mutat.* 2012;33:1051–5.
37. Adam R, Spier I, Zhao B, Klothe M, Marquez J, Hinrichsen I, et al. Exome sequencing identifies biallelic MSH3 germline mutations as a recessive subtype of Colorectal Adenomatous Polyposis. *Am J Hum Genet.* 2016;99:337–51.
38. Meienberg J, Bruggmann R, Oexle K, Matyas G. Clinical sequencing: is WGS the better WES? *Hum Genet.* 2016;135:359–62.
39. Tattini L, D'Aurizio R, Magi A. Detection of genomic structural variants from next-generation sequencing data. *Front Bioeng Biotechnol.* 2015;3:92.
40. Mu W, Li B, Wu S, Chen J, Sain D, Xu D, et al. Detection of structural variation using target captured next-generation sequencing data for genetic diagnostic testing. *Genet Med.* 2018;21:1603–10.
41. Montalban G, Bonache S, Moles-Fernández A, Gisbert-Beamud A, Tenés A, Bach V, et al. Screening of *BRCA1/2* deep intronic

- regions by targeted gene sequencing identifies the first germline *BRCA1* variant causing pseudoexon activation in a patient with breast/ovarian cancer. *J Med Genet.* 2019;56:63–74.
42. Clendenning M, Buchanan DD, Walsh MD, Nagler B, Rosty C, Thompson B, et al. Mutation deep within an intron of *MSH2* causes Lynch syndrome. *Fam Cancer.* 2011;10:297–301.
  43. Carethers JM. Differentiating Lynch-like from Lynch Syndrome. *Gastroenterology.* 2014;146:602–4.
  44. Geurts-Giele WR, Leenen CH, Dubbink HJ, Meijssen IC, Post E, Sleddens HF, et al. Somatic aberrations of mismatch repair genes as a cause of microsatellite-unstable cancers. *J Pathol.* 2014;234:548–59.
  45. Haraldsdottir S, Hampel H, Tomsic J, Frankel WL, Pearlman R, de la Chapelle A, et al. Colon and endometrial cancers with mismatch repair deficiency can arise from somatic, rather than germline, mutations. *Gastroenterology.* 2014;147:1308–16.
  46. Mensenkamp AR, Vogelaar IP, van Zelst-Stams WA, Goossens M, Ouchene H, Hendriks-Cornelissen SJ, et al. Somatic mutations in *MLH1* and *MSH2* are a frequent cause of mismatch-repair deficiency in Lynch syndrome-like tumors. *Gastroenterology.* 2014;146:643–6.
  47. Jansen AM, van Wezel T, van den Akker BE, Ventayol Garcia M, Ruano D, Tops CM, et al. Combined mismatch repair and *POLE/POLD1* defects explain unresolved suspected Lynch syndrome cancer. *Eur J Hum Genet.* 2016;24:1089–92.
  48. Morak M, Heidenreich B, Keller G, Hampel H, Laner A, de la Chapelle A, et al. Biallelic *MUTYH* mutations can mimic Lynch syndrome. *Eur J Hum Genet.* 2014;22:1334–7f.
  49. Vogt J, Wernstedt A, Ripperger T, Pabst B, Zschocke J, Kratz C, et al. *PMS2* inactivation by a complex rearrangement involving an *HERV* retroelement and the inverted 100-kb duplison on 7p22.1. *Eur J Hum Genet.* 2016;24:1598–604.