npj | genomic medicine

Case report

Published in partnership with CEGMR, King Abdulaziz University

https://doi.org/10.1038/s41525-025-00485-5

Long-read genome and RNA sequencing resolve a pathogenic intronic germline LINE-1 insertion in *APC*

Alexandra A. Baumann[1,2,3,16], Lisanne I. Knol[2,4,5,16], Marie Arlt[1,3], Tim Hutschenreiter[1,3], Anja Richter[1,2,3], Thomas J. Widmann[2,6], Marcus Franke[1,3], Karl Hackmann[1,3], Sylke Winkler[7], Daniela Richter[2,4,5,8,9], Isabel Spier[10,11], Stefan Aretz[10,11], Daniela Aust[12,13], Joseph Porrmann[1,3], Doreen William[1,3], Evelin Schröck[1,2,3,7,8,9,17], Hanno Glimm[2,4,5,8,9,14,15,17] & Arne Jahn[1,2,3,8,9,17] ✉

Familial adenomatous polyposis (FAP) is caused by pathogenic germline variants in the tumor suppressor gene *APC*. Confirmation of diagnosis was not achieved by cancer gene panel and exome sequencing or custom array-CGH in a family with suspected FAP across five generations. Long-read genome sequencing (PacBio), short-read genome sequencing (Illumina), short-read RNA sequencing, and further validations were performed in different tissues of multiple family members. Long-read genome sequencing resolved a 6 kb full-length intronic insertion of a heterozygous LINE-1 element between exons 7 and 8 of *APC* that could be detected but not fully resolved by short-read genome sequencing. Targeted RNA analysis revealed aberrant splicing resulting in the formation of a pseudo-exon with a premature stop codon. The variant segregated with the phenotype in several family members allowing its evaluation as likely pathogenic. This study supports the utility of long-read DNA sequencing and complementary RNA approaches to tackle unsolved cases of hereditary disease.

Pathogenic germline variants in hereditary cancer genes cause genetic tumor risk syndromes[1]. Identification of germline variants can have an impact on cancer screening, preventive interventions, (targeted) therapy, and reproductive choices. Loss-of-function germline variants in the tumor suppressor gene *APC* cause autosomal dominant familial adenomatous polyposis (FAP)[2,3]. FAP is characterized by multiple colonic adenomas (> 100 in classic FAP) and a near complete penetrance for early-onset colorectal cancer (mean age 39 years), as well as extracolonic manifestations[4]. Upon diagnosis, intensified tumor surveillance and colectomy are recommended[5,6].

Genetic testing can allow to differentiate between FAP and differential diagnoses[7]. Currently, genetic testing is typically limited to exonic and flanking intronic regions, resolving approximately 34-58% of pathogenic variants in *APC* for individuals with at least 100 adenomas[8–10]. Further genetic analyses, such as whole-genome sequencing, should be offered for unresolved cases of expected rare diseases. It allows for a more comprehensive variant detection, including deep intronic, promoter[11], and structural variants (SVs), and reaches a higher diagnostic yield than exome sequencing[12].

[1]Institute for Clinical Genetics, University Hospital Carl Gustav Carus at TUD Dresden University of Technology and Faculty of Medicine of TUD Dresden University of Technology, Dresden, Germany. [2]National Center for Tumor Diseases (NCT), NCT/UCC Dresden,, a partnership between DKFZ, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, and Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Dresden, Germany. [3]ERN GENTURIS, Hereditary Cancer Syndrome Center Dresden, Dresden, Germany. [4]Department of Translational Medical Oncology, NCT Dresden and DKFZ, Dresden, Germany. [5]Translational Medical Oncology, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Germany. [6]Pfizer-University of Granada-Junta de Andalucía Centre for Genomics and Oncological Research (GENYO), PTS Granada, managed by Fundación Pública Andaluza Progreso y Salud (FPS), Granada, Spain. [7]Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany. [8]German Cancer Consortium (DKTK), Dresden, Germany. [9]German Cancer Research Center (DKFZ), Heidelberg, Germany. [10]Institute of Human Genetics, Medical Faculty, University of Bonn, Bonn, Germany. [11]National Center for Hereditary Tumor Syndromes, University Hospital Bonn, Bonn, Germany. [12]Institute of Pathology, University Hospital Carl Gustav Carus at TUD Dresden University, Dresden, Germany. [13]Tumor- and Normal Tissue Bank of the University Cancer Center (UCC), University Hospital Carl Gustav Carus, Medical Faculty, TUD Dresden University of Technology, Dresden, Germany. [14]Center for Personalized Oncology, NCT Dresden and University Hospital Carl Gustav Carus, Faculty of Medicine and TUD Dresden University of Technology, Dresden, Germany. [15]Translational Functional Cancer Genomics, NCT Heidelberg and DKFZ, Heidelberg, Germany. [16]These authors contributed equally: Alexandra A. Baumann, Lisanne I. Knol.[17]These authors jointly supervised this work: Evelin Schröck, Hanno Glimm, Arne Jahn. ✉e-mail: Arne.Jahn@ukdd.de

SVs are larger than 50 bp and include deletions, insertions, duplications, inversions, translocations, and complex rearrangements[13,14]. One source of SVs are retrotransposons that use a "copy-and-paste" mechanism to multiply throughout the genome, such as Long INterspersed Element-1 (LINE-1) and LINE-1 dependent Alu-elements, comprising approx. 17% and 11% of the human genome, respectively[15,16]. Full-length LINE-1 elements are approx. 6 kb long[17] and contain a 5' UTR with internal promoter activity, two expressed open reading frames (ORF1, coding for an RNA binding protein, and ORF2, coding for a protein with endonuclease and reverse transcriptase activity), and a 3' UTR with a poly(A) tail[18]. Retrotransposition can lead to genomic alterations due to the full-length or partial LINE-1/Alu insertion itself, as well as genomic rearrangements like deletions (up to 1 Mb)[15,19–21]. These genomic alterations can result in aberrant splicing, aberrant expression, and/or epigenetic alterations of affected genes or regions[15,22].

Detecting and studying SVs, including those derived from retrotransposition events, has been challenging with short-read sequencing[13,14,23,24], especially as SVs often occur in genomic regions of low complexity or around repetitive elements[25], and specialised callers[26] or orthogonal validations might be required[27]. Therefore, the impact of LINE-1 disease-causing SVs in monogenic diseases has likely been underestimated[28–30]. Long-read sequencing (LRS, reads >10 kb) has largely improved the detection of SVs[31] and has resolved complex regions of the human genome[32], because longer reads can span the full length of the SV and reduce the fraction of multi-mapping reads[30].

In this study, we describe the detection and the resolution of the, to our knowledge first, germline intronic full-length LINE-1 insertion in genomic DNA (gDNA) in a family with suspected FAP using long-read genome sequencing. A co-inserted flanking DNA sequence (3' DNA transduction) allowed the identification of the donor LINE-1 element. Segregation and cDNA analysis revealed partial exonization of the LINE-1 element and allowed variant evaluation as likely pathogenic.

Results

Family with a suspected familial adenomatous polyposis

Classic familial adenomatous polyposis (FAP) was expected in a family presenting with five generations of patients with adenomatous colonic polyposis (II:3, III:4, IV:3, V:1) or colorectal cancer (I:3) (Fig. 1). Family members underwent colectomy (individuals II:3, III:4 and IV:3) and/or *APC*-specific cancer surveillance (individuals IV:3, III:4, V:1). One individual of the family had an unremarkable colonoscopy at the age of 36 (IV:1) and one individual did not receive a colonoscopy at the age of nine years (V:2).

Long-read genome sequencing resolved a 6.1 kb insertion in *APC*

Extensive routine genetic diagnostics for the index patient (IV:3), including cancer gene panel sequencing, exome sequencing, and copy number analyses, were inconclusive ("Methods" section) and did not reveal a pathogenic variant in *APC*. Only a heterozygous, likely benign, synonymous *APC* variant (NM_000038.6:c.1959G>A (p.Arg653 =))[33,34] was found in individuals IV:3 and III:4 (Fig. 1).

Short-read whole-genome sequencing (WGS) (Illumina) followed by SV calling detected an insertion in intron 7 (chr5:g.112800732ins) of *APC* (Supplementary Fig. 1A, B). However, the sequence of the insertion could not be determined, since the span of the insertion surpassed that of the reads.

Therefore, PacBio long-read WGS was carried out. The median read length was 13,757 bp, exceeding the 150 bp length of the short-read sequencing. Three out of four SV callers (SVIM, PBSV and Sniffles2) detected a heterozygous ≈ 6.1 kb insertion (chr5:g.112800732_112800733insN[6116 ± 1]) at the same location in *APC* as the short-read WGS (Fig. 2 and Supplementary Fig. 1C). Of 29 reads mapped to the region, eight reads spanned the entire insertion and five reads spanned one SV breakpoint. A consensus sequence of the insertion was determined, with an uncertainty of one A in the poly-A_{16} region at the 3'-end (Supplementary Fig. 1D, whole consensus sequence in Supplementary Material 1). The sequence of the insertion was confirmed by long-range PCR followed by amplicon sequencing but yielded two NGS and mapping errors (poly-C artifact and low coverage in the poly-A stretch, Supplementary Fig. 1E). Furthermore, segregation analysis (long-range PCR) revealed the *APC* insertion in both the affected mother (III:4) of the index patient (IV:3) and her affected son (V:1), but not in the unaffected sister (IV:1) (Supplementary Fig. 2A). An additional rare 10 bp deletion in intron 2 of *APC* (chr5:g.112765678_112765687del) detected in the index (IV:3) was not considered pathogenic as it was also observed in her unaffected sister (IV:1), and no other pathogenic variants were found in genes associated with colorectal polyposis.

Characterization of the SV reveals an insertion of an intact LINE-1 retrotransposon

A BLAST (NCBI, BLAST + 2.15.0[35]) search of the insertion consensus sequence matched with a complete human LINE-1 element inserted in the sense direction of *APC* (5' to 3' integration) and could be further manually subgrouped as LINE-1 Ta1d element based on Boissinot et al.[36], the

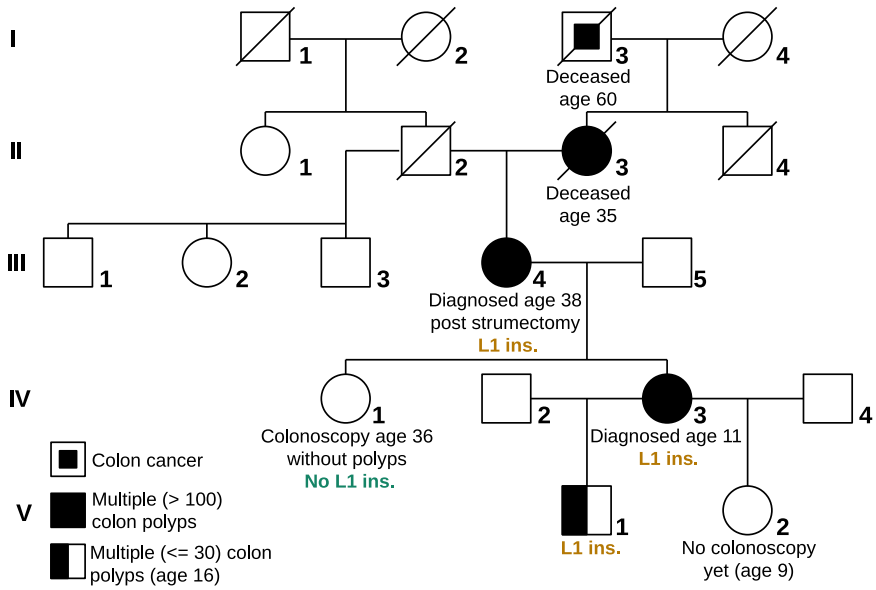

Fig. 1 | Family pedigree with suspected familial adenomatous polyposis. Phenotype of individuals and genotype of LINE-1 insertion in APC (L1 ins.) = chr5:g.112800732_112800733insN[6116 ± 1] depicted. Individuals III:4, IV:1, IV:3, and V:1 were tested and in all but IV:1, the variant was found.

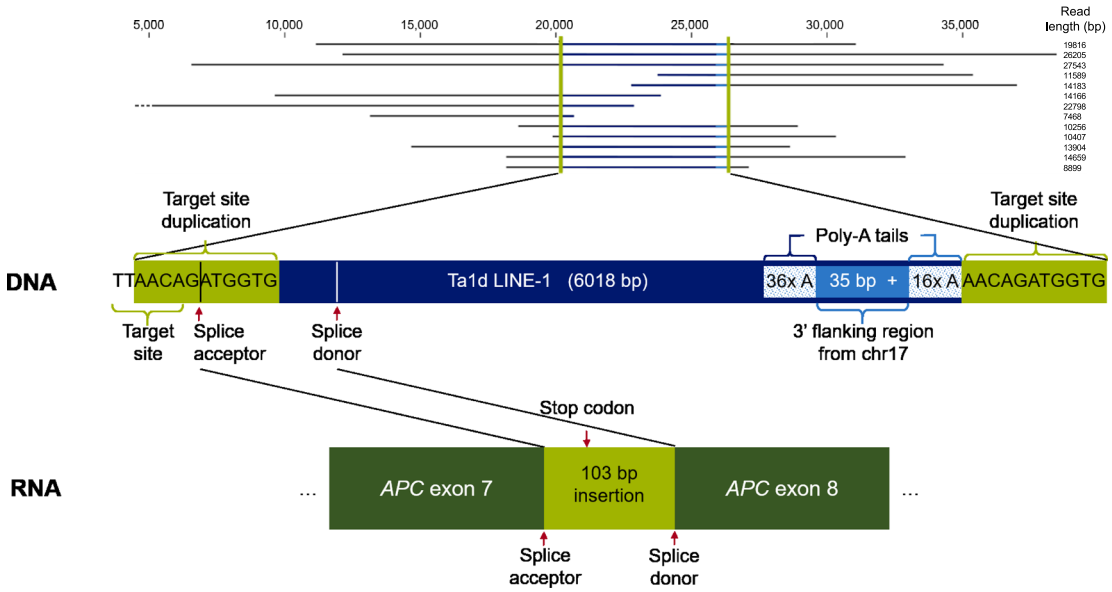

Fig. 2 | Long-read mapping of the genomic region of intron 7 of APC spanning the LINE-1 insertion and enlarged sequence indicating the specific disease-causing genotype (screenshot of CLC Genomics Workbench v21.0.3, QIAGEN). Below, the transcript model of the aberrant transcript with a partial exonization of the LINE-1 insertion in APC (NM_000038.6) is shown.

youngest and potentially active human LINE-1 subfamily. BLAST, as well as the UCSC-genome browser BLAT[37] of the full-length insertion against the human reference T2T-CHM13v2.0 (Jan. 2022), revealed a match to a LINE-1 element on chromosome 17 (start at chr17:67510707), based on a 35 bp flanking region at its 3'-end (chr17:67516763-67516797). No other LINE-1 insertion with the same characteristics and flanking region was found in the patient. The AA | TTGT sequence on the minus strand (chr5:112800731-112800736) was the expected target site of the LINE-1 element due to its similarity to the typical AA | TTTT sequence[38] (Fig. 2). An 11 bp homology region (AACAGATGGTG) between the insertion and the reference sequence was found. This goes in line with typical LINE-1 retrotransposition by target-site primed reverse transcription at an asymmetric AT-rich sequence in the genome, typically leading to a duplication of the target site (TSD)[16,18]. To investigate methylation of the element and the surrounding region, the 5-methylcytosine pattern was studied based on long-read sequencing data and suggested heavy methylation of CpG sites within the LINE-1 element in the blood (Supplementary Fig. 2B, C).

Targeted RNA sequencing revealed aberrant splicing resulting in nonsense-mediated decay

Even though the genomic variant was fully resolved, the impact on RNA level was still unknown. Thus, in silico splice prediction tools were tested on the full-length insertion. The command line tool from SpliceAI (v1.3.1[39]) could process the full LINE-1 insertion sequence and predicted aberrant splicing, but no (delta) positions for acceptor gain (delta score 0.98) and donor gain (delta score 0.64).

Hence, RNA from lymphoblastoid cell lines (LCLs) and stabilized whole blood was analyzed to investigate the transcriptional impact of the intronic *APC* LINE-1 insertion. Short-read mRNA sequencing yielded insufficient coverage of the region of interest. Therefore, a custom target enrichment panel encompassing exonic regions of 303 cancer-related genes was used for enrichment of cDNA and resulted in a coverage of ~300x of the region of interest in the *APC* gene (total coverage 35 Mio reads, Supplementary Fig. 3A, B). The mapping indicated partial exonization of the LINE-1 element. Targeted PCRs on blood-derived cDNA revealed a splice acceptor gain in the target-site duplication and a splice donor gain in the LINE-1 element, leading to a 103 bp pseudo-exon between exons 7 and 8 of *APC* (Fig. 2, Supplementary Fig. 3C, Supplementary Material 1). Validations on cDNA derived from FFPE tissue from colon epithelium as well as a colon polyp of the index patient (individual IV:3) yielded both splice junctions separately (Supplementary Fig. 3D).

The partial exonization of the 5'-end of the LINE-1 insertion included a stop codon (after 10 amino acids), expected to lead to nonsense-mediated decay (NMD) (Fig. 2). To study this further, LCLs of the index patient (IV:3) and of an unaffected family member (IV:1) were treated with the NMD inhibitor cycloheximide. In the index patient (IV:3), but not in the healthy control (IV:1), a shift in the variant allele frequency ratio of a shared benign exonic *APC* variant (chr5:g.112827157 T > C) was observed, supporting that exonization of the LINE-1 element led to NMD (Supplementary Fig. 3E). Additional cDNA analysis of this heterozygous variant with an exon-exon junction spanning primer revealed that less than 10% of all *APC* wildtype transcript was derived from the LINE-1 allele. This suggested a near complete aberrant splicing from this allele (Supplementary Fig. 3F). Based on the evidence from the gDNA and mRNA analysis, *APC*-specific ACMG/AMP criteria[40] PS3_strong, PM2_supporting (absent in gnomAD SVs v4.0[41]), and PS4_supporting were given and the full-length LINE-1 insertion in *APC* was evaluated as likely pathogenic.

Discussion

We applied short- and long-read genome sequencing of gDNA as well as targeted RNA sequencing in a family with suggested FAP that identified a heterozygous likely pathogenic intronic full-length LINE-1 insertion in *APC*, leading to a near complete pseudo-exonization and NMD.

A somatic LINE-1 insertion in the last exon of *APC* in a colon cancer was reported as one of the first disease-causing retrotransposition events[42], and other studies similarly described the relevance of somatic retrotransposition events in *APC* during cancerogenesis in colorectal cancer[43]. While a likely benign full-length LINE-1 germline insertion was detected in a population database (INS_CHR5_0FF534E3 in 95 of 126,092 alleles of gnomAD SV v.4.0[41]), to our knowledge, this report describes the first pathogenic LINE-1 germline insertion in *APC*. Besides LINE-1 elements, five Alu germline insertions[44–47] and two SVA insertions in *APC*[48] have previously been reported.

Only 80–100 evolutionary young full-length LINE-1 elements (HS or Ta1d[49]) are potentially active. Some of them may escape host defenses such as methylation and histone modifications[50] and show retrotransposition activity. The *APC* LINE-1 insertion described here is an excellent example of the mutagenic potential of mobile elements, even though being inserted into

an intron. While an increased de novo mutation rate for the element in *APC* is not expected, as experimental data does not support insertion-prone regions[51], it remains to be investigated whether it is a founder variant. The co-retrotransposition of a short 3' flanking sequence allowed the identification of a element on chr17 in the human T2T[32]. Most likely, the same element on chr17 from a different source has been described to be highly active in a vector-based retrotransposition assay in HeLa cells (identification number 2–12, 130% of control L1.3 LINE-1 element activity)[52]. Therefore, we assume that the chromosome 17 LINE-1 element is a recurrent element and the element of origin for the LINE-1 insertion in *APC* described here. Since these LINE-1 elements are both full-length, they likely retain their retrotransposition capacity.

Despite their prevalence, estimated to be one in 250 to 1000 pathogenic variants[53], LINE-1 retrotransposition is rarely detected as a cause of monogenetic diseases. However, like in this family, their frequency and that of other SVs has likely been underestimated. Considering the size of the *APC* gene (108 kb), these and other retrotransposition events could partially solve unexplained cases of suspected FAP and other monogenetic diseases. Short-read genome sequencing improves the detection of non-coding variants and SVs, and could serve as first-tier diagnostics for patients with rare diseases[54]. Short-read genome sequencing had detected the breakpoints of the LINE-1 insertion described here but did not resolve it. Genotyping with targeted short-read sequencing might be challenging due to poly-A stretches and GC-rich regions[27], as exemplified in this report. Additionally, for diseases with a larger genetic heterogeneity, SV detection is challenging due to many false positives[55]. Long-read sequencing has a higher sensitivity to detect SVs, especially in repetitive areas[56,57], and fully resolves them as shown for this LINE-1 insertion and other complex genomic rearrangements in the *APC* locus, likely increasing the diagnostic yield[58]. Thereby, in principle, it obviates the need for validations (e.g., NGS panel sequencing, PCR, and Sanger sequencing). For other large and complex SVs, other diagnostic tools, such as Optical Genome Mapping might prove valuable[59]. As an additional advantage, long-read sequencing provides information on cytosine methylation. As expected, and likely explained by the presence of the YY1-sequence in the LINE-1 promoter[60], the CpG sites within the LINE-1 element were methylated (5mC). This should lead to transcriptional silencing, as expected in blood and nearly all other differentiated cells. We infer that the unmethylated parental LINE-1 element integrated a copy of itself into the *APC* gene during the early embryogenesis of an ancestor[50,52].

Dependent on the specific genotype of the retrotransposition-induced SV and the insertion site, various effects on RNA level are possible[15,22] and can impact clinically relevant genes[19,61–64]. In this case, the LINE-1 insertion in *APC* was partially exonized, as it provided both a splice acceptor (surprisingly in the target-site duplication) and a splice donor (in the LINE-1 promoter region), leading to truncation, NMD, and near complete functional loss of this allele. Yet, the transcriptional impact of full-length LINE-1 insertions remains to be systematically investigated.

Only the command line version of the in silico splice prediction tool SpliceAI[39] could process the full LINE-1 insertion and predicted aberrant splicing without providing details. Thus, experimental tests on RNA level had to be conducted to assess the functional impact. The low *APC* expression in blood (median TPM of *APC* in whole blood = 1.81, GTEx[65]) limited the usage of total poly-A enriched RNA-seq. Amplicon-based targeted cDNA analysis has been described[47] and proved helpful for quantification of aberrant splicing here. Alternatively, a targeted enrichment of cDNA using oligonucleotides of a routine diagnostic cancer panel, short-read sequencing, and bioinformatic tools to identify aberrant expression (DROP)[66] could be used as a screening method. This complementary approach can increase the diagnostic yield for unresolved cases or prioritized variants[67,68].

However, an accurate quantification of *APC* expression with this method is dependent on the used tissue, gene panel and multiple control samples. Therefore, a low-cost targeted long-read RNA sequencing[69] is an interesting alternative for accurate isoform detection and quantification in genetic diagnostics.

In conclusion, multi-platform genomics, including long-read sequencing and further omics beyond routine diagnostics allowed to molecularly confirm a FAP in this family and are promising for patients with rare diseases[70–72] and in precision oncology[73,74]. Such a strategy should improve the diagnostic yield of non-coding variants and SVs, including transposable elements, and shorten the time to diagnosis[56].

Methods

Patients, samples, and ethics

The family was initially seen at the Institute for Clinical Genetics, University Hospital Dresden, in 2016. Clinical data, specimens, and other biological materials were collected, used, and stored after obtaining signed, informed consent from the participating individuals. Peripheral blood samples (EDTA) were available for individuals III:4, IV:1, IV:3, and V:1 (Fig. 1). EBV-immortalized lymphoblastoid cell lines (LCL) were generated[75] from individuals III:4 and IV:3. Additionally, PAXgene Blood RNA tubes (Becton Dickinson) were collected according to manufacturer's instructions. For index patient IV:3, FFPE tissue of healthy colon epithelium, as well as from a colon polyp, were macrodissected. Clinical investigations were conducted according to the German Gene-Diagnostic Act, and the Declaration of Helsinki principles and approved by the local institutional ethical committees (BO-EK-497112022).

Genomic DNA sequencing and validation

Previous analyses. Previous genetic diagnostics for the index patient (IV:3) included targeted sequencing of specific intronic *APC* variants[47], short-read panel sequencing (TruSightCancer94 (Illumina), CEU HuGx EBM panel including *APC*, *POLD1* and *POLE* (Illumina)), exome sequencing (TWIST Bioscience, comprehensive), copy number analyses (customized high-resolution array for CGH including the *APC*-gene[76] and *MLPA* (Salsa MLPA Kit P043-E1 APC, MRC-Holland)).

Short-read DNA sequencing. Genomic DNA from whole blood for short-read sequencing was extracted via the NucleoSpin Blood L Kit (MACHEREY-NAGEL) according to the manufacturer's instructions.

Genomic DNA isolated from K2 EDTA blood of individual IV:3 according to the manufacturer's instructions with 500 ng final DNA input was used for the Illumina DNA PCR-Free Library Prep kit (Illumina). Sequencing was performed using an S2 flow cell on a NovaSeq 6000 Sequencing System yielding an average coverage of ~41x (2×150 bp).

The reads were mapped on a DRAGEN v3.9.5 platform (Illumina) with default parameters. The TruSight Software Suite (Illumina) was used for variant filtering and IGV for variant visualization[77]. Independently, FASTQ reads they were mapped with bwa-mem (v0.7.17[78]) and further processed with samtools fixmate, sort, and markdup (v1.11[79]). Structural variants were called with the DRAGEN pipeline (including Manta[80]) and GRIDSS (v2.11.1)[81] using standard parameters.

Long-read whole-genome DNA sequencing. High-molecular weight (HMW) genomic DNA (gDNA) of individual IV:3 was extracted from human K2 EDTA whole blood sample with the Nanobind HMW DNA Extraction—Mammalian Whole Blood Protocol (Circulomics, Document ID: EXT-BLH-001, Release Date: 03/24/2021) according to the manufacturer's instructions.

In brief, 600 µl of whole human blood stored in K2 EDTA was digested with Proteinase K and RNase A in the dedicated blood lysis buffer (BL3) on a Thermomixer with 900 rpm at 55 °C for 10 min. After incubation, the released gDNA was bound to the Circulomics Nanobind disk upon the addition of isopropanol. After several washing steps, the HMW gDNA was eluted from the Nanobind disk and kept in buffer EB. The quality and length of the extracted gDNA were analyzed by Pulse field gel electrophoresis using the Femtopulse device (Agilent). The fragment length of the extracted HMW gDNA was about 40 to 120 kb.

Three HiFi libraries of Circulomics-extracted genomic DNA (HMW gDNA) of *human blood* were prepared as recommended by Pacific

Biosciences according to the 'Guidelines for preparing HiFi SMRTbell libraries using the 'SMRTbell Express Template Prep Kit 2.0'.

In summary, HMW gDNA was post-purified with 1x pretreated AMPure beads (Beckman). HMW gDNA were sheared twice to 14 kb fragments with the setting 25 kb and 20 kb on a MegaRuptor™ device (Diagenode).

5 µg of sheared gDNA has been used for each library preparation according to the PacBio guidelines. All three PacBio SMRTbell™ libraries were size selected for fragments larger than 7 kb with the BluePippin™ device according to the manufacturer's instructions. The size selected libraries were loaded on the Sequel II with 65 and 70 pM on plate, and ran on three Sequel II SMRT cells (8 M) with the Sequel II polymerase 2.2, the sequencing primer v5, and the Sequel II sequencing kit 2.0 for 30 h on a Sequel II, yielding to an average coverage of 22x.

Circular consensus sequences (CCS) and 5mC marks were called, making use of the default SMRTLink tools (SMRTlink v11.0.0.146107). 5mC CpG sites were defined by kinetic analysis of the raw PacBio subreads.

CCS reads were generated with the PacBio CCS tool (https://github.com/nlhepler/pbccs), and DeepConsensus was applied to improve yield and accuracy[82]. The remaining PacBio Adapters were identified with a blast and removed.

DNA extraction, PacBio library preparation, sequencing, consensus calling, and read polishing were carried out by the DRESDEN-concept Genome Center.

Subsequently, reads were aligned to the NCBI GRCh38 genome without alternative contigs and including the Epstein-Bar genome as a decoy sequence (https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.15_GRCh38/seqs_for_alignment_pipelines.ucsc_ids/) using pbmm2 (version 1.9.0, https://github.com/PacificBiosciences/pbmm2) with –preset CCS. Mapping statistics were reported and visualized with Nanoplot (version 1.40.2)[83]. Subsequently, four different SV callers were used: pbsv (version 2.8.0, https://github.com/PacificBiosciences/pbsv), sniffles (version 2.0.7)[84,85], SVIM (version 1.4.2)[86]and cuteSV (version 1.0.8)[87].

CuteSV was run with settings –max_cluster_bias_INS 1000, –diff_ratio_merging_INS 0.9, –max_cluster_bias_DEL 1000, and –diff_ratio_merging_DEL 0.5, while the other SV callers were run with standard settings. For both pbsv and sniffles, annotated tandem repeats were provided as recommended to increase sensitivity and recall. Subsequently, the detected SVs surrounding the *APC* gene were studied for each caller and compared with the tool SURVIVOR[88].

CLC Genomics Workbench (version 21.0.3, QIAGEN) was used to call a consensus sequence of the LINE-1 insertion based on the reads overlapping the insertion site.

Validations. For PCR validations, 2 µl of 10 ng/µl genomic DNA or cDNA were used in a 20 µl reaction mix with 2 µl of 10 pmol forward and reverse in-house designed primers (Supplementary Table 1) and 16 µl of a mastermix with reagents of the Qiagen HotStarTaq DNA Polymerase (250 U) #203203 Kit (except the dNTP Mix). The mastermix consisted of 0.2 µl HotStart Taq Polymerase (250 U), 0.6 µl $MgCl_2$ (25 mM), 1 µl dNTP Mix (10 mM each, NEB #N0447S), 2 µl 10x PCR buffer, 4 µl 5x Q-Solution and 8.2 µl water. PCR steps were initial denaturation (15 min at 95 °C), 34 PCR amplification cycles (30 s at 95 °C, 30 s annealing according to Supplementary Table 1 and 45 s of 72 °C) and final extension (10 min at 72 °C). The resulting PCR products were purified with the QIAquick PCR & Gel Cleanup Kit (QIAGEN) or, when multiple PCR products were present, with the QIAquick Gel Extraction Kit (QIAGEN) according to the manufacturer's instructions. Sanger sequencing with the respective primers was performed at Seqlab. For NGS of the PCR product, a library was prepared with the Illumina DNA Prep (M) Tagmentation kit (Illumina, 20060059) according to the standard protocol and sequenced on an Illumina NextSeq500 or NextSeq550 system (2 × 150 bp reads). The sequencing analysis workflow is described in the section "Targeted short-read RNA sequencing with a cancer gene panel enrichment".

Long-range PCR (> 1 kb) of genomic DNA was carried out with the Long Range PCR Master Mix (2x) (Biotechrabbit) according to the manufacturer's indications and with in-house designed primers (Supplementary Table 1).

RNA sequencing and validation

RNA extraction. RNA from PAXgene Blood RNA tubes (Becton Dickinson) was extracted with the PAXgene Blood RNA Kit (PreAnalytiX). Cell pellets of lymphoblastoid cell lines were stored in RNAprotect solution before RNA was extracted using the RNeasy Mini Kit (QIAGEN), including on-column DNase digestion with the RNAse-Free DNase Set (QIAGEN). First-strand cDNA for PCR validations on mRNA level was synthesized with the SuperScript VILO cDNA Synthesis Kit (Invitrogen) according to the manufacturer's specifications.

RNA of normal colon tissue (FFPE) and adenomatous colon tissue (FFPE) was extracted via the QIAGEN RNeasy FFPE Kit (cat. 73504). All protocols were carried out according to the manufacturer's instructions.

Short-read total mRNA sequencing. RNA extracted from blood and LCLs were sequenced on an Illumina NextSeq550 or NextSeq500 (10 M reads per sample, 2×75 bp reads) after library preparation (Illumina TruSeq RNA Sample Preparation Kit v2, and Illumina Stranded mRNA Prep Kit with an RNA poly-A enrichment step).

The quality of the sequencing data was assessed with FastQC (v0.11.9, https://github.com/s-andrews/FastQC) and reads were clipped using trimmomatic (v0.39)[89] (ILLUMINACLIP: TruSeq3-PE.fa:2:30:10:2:true LEADING:15 TRAILING:15 SLIDINGWINDOW:4:15 MINLEN:36 CROP:75).

The trimmed reads were mapped with primary (v2.7.9a)[90] on the GRCh37 and GRCh38 reference genome (–readFilesCommand zcat –alignIntronMax 500000 –alignMatesGapMax 500000 –outBAMcompression 0 –outSAMtype BAM SortedByCoordinate –outSAMprimaryFlag OneBestScore –outFilterMultimapNmax 100 –outFilterMismatchNoverLmax 0.05 –chimSegmentMin 15 –chimOutType WithinBAM –chimScoreMin 1 –chimScoreJunctionNonGTAG 0 –chimJunctionOverhangMin 15 –chimSegmentReadGapMax 3 –alignSJstitchMismatchNmax 5 -1 5 5). The targeted RNA-seq samples were mapped to GRCh37. The resulting alignment file (BAM file) was indexed with samtools (version 1.9)[91].

Targeted short-read RNA sequencing with a cancer gene panel enrichment. To obtain long cDNA strands with oligo dT primers, RNA from LCLs (1000 ng) was incubated with 2 µl of oligo dT_{12-18} Primers (500 µg/ml, Invitrogen, cat. 18418012) and 1 µl of a dNTP mix (10 mM of each nucleotide, NEB, cat. N0447S) at 65 °C for 5 min, then spun briefly and put on ice. 1 µl of the Induro Reverse Transcriptase (200 U/µl, NEB, cat. M0681S), 4 µl of the corresponding 5x Induro RT Reaction Buffer (NEB, cat. B0681A), 0.2 µl of an RNasin Plus Ribonuclease Inhibitor (40 U/µl, Promega, cat. N261A), and nuclease-free water were added to a final volume of 20 µl. After an incubation time of 30 min at 45 °C, the reverse transcriptase was inactivated at 95 °C for 1 min and put on ice. For second-strand cDNA synthesis, 45 µl nuclease-free water, 15 µl of 5x Second-Strand Reaction Buffer (Invitrogen, cat. 10812014), and 1.5 µl of dNTP mix (10 mM of each nucleotide, NEB, cat. N0447S) were added to the single-stranded cDNA sample on ice. The reaction mix was incubated with 0.5 µl *E. coli* DNA ligase (10 U/µl, Invitrogen, cat. 18052019), 2 µl *E. coli* DNA polymerase I (10 U/µl, Invitrogen, cat. 18010017), and 1 µl *E. coli* RNase H (2 U/µl, Invitrogen, cat. 18021014) at 16 °C for 2 h. 2 µl T4 DNA Polymerase (5 U/µl, Invitrogen, cat. 1368437) was supplemented to the sample and further incubated at 16 °C for 5 min. After the addition of 5 µl EDTA (0.5 M, Invitrogen, cat. AM9261), the double-stranded cDNA was purified with the High Pure PCR Product Purification Kit (Sigma-Aldrich, cat. 11732668001) following the manufacturer's instructions.

The length of the product was assessed via a Fragment Analyzer System with a HS Genomic DNA 50 kb Kit (Agilent, cat. DNF-468-0500).

To enrich the cDNA for regions of interest, a Custom Cancer Panel with exonic regions of 303 cancer-related genes (Twist Biosciences) was used according to standard procedure for DNA. The Library Preparation EF 2.0 with Enzymatic Fragmentation and Twist Universal Adapter System (Rev2.0) (Twist Biosciences) was followed according to the manufacturer's instructions for NGS library preparation. This included use of the Enzymatic Fragmentation Library kit (EF, Twist Biosciences) and Universal adapter system (UDI, Twist Biosciences). The Twist Target Enrichment Standard Hybridization v2 (Rev3.0) for NGS workflow protocol (Twist Biosciences) was followed for target enrichment with specific custom cancer capture probes. The resulting cDNA was inspected via a Fragment Analyzer System with an HS NGS Fragment Kit (1–6000 bp) (Agilent) prior to sequencing on an Illumina NextSeq500 or NextSeq550 system (35 M reads, 2 × 150 bp reads).

After sequencing and quality control with FastQC (v0.11.9, https://github.com/s-andrews/FastQC), reads were trimmed with trimmomatic (v0.39)[89] (ILLUMINACLIP: data/TruSeq3-PE.fa:2:30:10:2:true LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36). The trimmed reads were indexed the same way as the short-read total mRNA sequencing data (section "Short-read total mRNA sequencing").

Nonsense-mediated decay inhibition. Two sets of 3×10^6 LCLs of individuals IV:1 and IV:3 were seeded in 10 ml RPMI 1640 (Gibco, 21875034) with 15% FCS (Gibco, 10270106) and 1% Penicillin/Streptomycin (Gibco, 11548876). LCLs were incubated either with or without 100 µg/ml Cycloheximide (CHX, Sigma-Aldrich, 01810-1 G) as a nonsense-mediated decay (NMD) inhibitor for 6 h at 37 °C. The cells were washed with 1xPBS (Gibco, 14190136), and the pellet was resuspended in 500 µl RNAprotect solution (Qiagen, 76526) for storage at −80 °C. RNA isolation and (short) first-strand cDNA synthesis were done as described above (see DNA and RNA extraction, cDNA synthesis). PCRs of *APC* segments containing the heterozygous exonic variant chr5:g.112827157 T > C was carried out as described above (see subsection "Validations in section Genomic DNA sequencing and validation").

Data availability

The consensus sequence of the LINE-1 insertion in *APC* can be found in Supplementary Material 1. The structural variant identified in this study was submitted to ClinVar under the organization number 505632 with accession SCV005685034. Sequencing data may be accessed by qualified researchers through direct contact with the authors. All variant positions refer to the GRCh38/hg38 reference genome except if stated otherwise.

Received: 17 September 2024; Accepted: 28 February 2025;
Published online: 04 April 2025

References

1. Rahman, N. Realising the promise of cancer predisposition genes. *Nature* **505**, 302–308 (2014).
2. Groden, J. et al. Identification and characterization of the familial adenomatous polyposis coli gene. *Cell* **66**, 589–600 (1991).
3. Nishisho, I. et al. Mutations of chromosome 5q21 genes in FAP and colorectal cancer patients. *Science* **253**, 665–669 (1991).
4. Yen, T. et al. APC-Associated Polyposis Conditions. 1998 [Updated 2022]. In: Adam, M. P. et al. (eds) GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993–2025.
5. Zaffaroni, G. et al. Updated European guidelines for clinical management of familial adenomatous polyposis (FAP), MUTYH-associated polyposis (MAP), gastric adenocarcinoma, proximal polyposis of the stomach (GAPPS) and other rare adenomatous polyposis syndromes: a joint EHTG-ESCP revision. *Br. J. Surg.* **111**, znae070 (2024).
6. National Comprehensive Cancer Network. *Genetic/Familial High-Risk Assessment: Colorectal Version 1.2023*. (National Comprehensive Cancer Network, 2023).
7. Valle, L. & Monahan, K. J. Genetic predisposition to gastrointestinal polyposis: syndromes, tumour features, genetic testing, and clinical management. *Lancet Gastroenterol. Hepatol.* **9**, 68–82 (2024).
8. Grover, S. et al. Prevalence and phenotypes of APC and MUTYH mutations in patients with multiple colorectal adenomas. *JAMA* **308**, 485–492 (2012).
9. Stanich, P. P. et al. Prevalence of germline mutations in polyposis and colorectal cancer–associated genes in patients with multiple colorectal polyps. *Clin. Gastroenterol. Hepatol.* **17**, 2008–2015.e3 (2019).
10. Aretz, S. et al. Large submicroscopic genomic APC deletions are a common cause of typical familial adenomatous polyposis. *J. Med. Genet.* **42**, 185–192 (2005).
11. Rohlin, A. et al. Inactivation of promoter 1B of APC causes partial gene silencing: evidence for a significant role of the promoter in regulation and causative of familial adenomatous polyposis. *Oncogene* **30**, 4977–4989 (2011).
12. Chung, C. C. Y. et al. Meta-analysis of the diagnostic and clinical utility of exome and genome sequencing in pediatric and adult patients with rare diseases across diverse populations. *Genet. Med.* **25**, 100896 (2023).
13. Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
14. Huddleston, J. et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.* **27**, 677–685 (2017).
15. Beck, C. R., Garcia-Perez, J. L., Badge, R. M. & Moran, J. V. LINE-1 elements in structural variation and disease. *Annu. Rev. Genomics Hum. Genet.* **12**, 187–215 (2011).
16. Deininger, P. Alu elements: know the SINEs. *Genome Biol.* **12**, 236 (2011).
17. Dombroski, B. A., Mathias, S. L., Nanthakumar, E., Scott, A. F. & Kazazian, H. H. Isolation of an active human transposable element. *Science* **254**, 1805–1808 (1991).
18. Ostertag, E. M. & Kazazian, H. H. Jr Biology of mammalian L1 retrotransposons. *Annu. Rev. Genet.* **35**, 501–538 (2001).
19. Hancks, D. C. & Kazazian, H. H. Roles for retrotransposon insertions in human disease. *Mob. DNA* **7**, 9 (2016).
20. Vogt, J. et al. SVA retrotransposon insertion-associated deletion represents a novel mutational mechanism underlying large genomic copy number changes with non-recurrent breakpoints. *Genome Biol.* **15**, R80 (2014).
21. Zhou, D. et al. Insertion of LINE-1 Retrotransposon inducing exon inversion causes a rotor syndrome phenotype. *Front. Genet.* **10**, 1–6 (2020).
22. Belancio, V. P., Roy-Engel, A. M. & Deininger, P. The impact of multiple splice sites in human L1 elements. *Gene* **411**, 38–45 (2008).
23. Van Belzen, I. A. E. M., Schönhuth, A., Kemmeren, P. & Hehir-Kwa, J. Y. Structural variant detection in cancer genomes: computational challenges and perspectives for precision oncology. npj Precis. *Onc* **5**, 1–11 (2021).
24. Sedlazeck, F. J., Lee, H., Darby, C. A. & Schatz, M. C. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat. Rev. Genet*. **19**, 329–346 (2018).
25. Carvalho, C. M. B. & Lupski, J. R. Mechanisms underlying structural variant formation in genomic disorders. *Nat. Rev. Genet* **17**, 224–238 (2016).
26. Joe, S. et al. Comparison of structural variant callers for massive whole-genome sequence data. *BMC Genomics* **25**, https://doi.org/10.1186/s12864-024-10239-9 (2024).
27. Yang, L. A practical guide for structural variation detection in the human genome. *CP Hum. Genet.* **107**, e103 (2020).

28. Huddleston, J. & Eichler, E. E. An incomplete understanding of human genetic variation. *Genetics* **202**, 1251–1254 (2016).
29. Korbel, J. O. et al. Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420–426 (2007).
30. De Coster, W. & Van Broeckhoven, C. Newest methods for detecting structural variations. *Trends Biotechnol.* **37**, 973–982 (2019).
31. Mitsuhashi, S. & Matsumoto, N. Long-read sequencing for rare human genetic diseases. *J. Hum. Genet.* **65**, 11–19 (2020).
32. Nurk, S. et al. The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
33. Aretz, S. et al. Familial adenomatous polyposis: aberrant splicing due to missense or silent mutations in the APC gene. *Hum. Mutat.* **24**, 370–380 (2004).
34. Grandval, P. et al. The UMD-APC Database, a model of nation-wide knowledge base: update with data from 3,581 variations. *Hum. Mutat.* **35**, 532–536 (2014).
35. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
36. Boissinot, S., Chevret, P. & Furano, A. V. L1 (LINE-1) retrotransposon evolution and amplification in recent human history. *Mol. Biol. Evol.* **17**, 915–928 (2000).
37. Kent, W. J. et al. The Human Genome Browser at UCSC. *Genome Res* **12**, 996–1006 (2002).
38. Cost, G. J. & Boeke, J. D. Targeting of human retrotransposon integration is directed by the specificity of the L1 endonuclease for regions of unusual DNA structure. *Biochemistry* **37**, 18081–18093 (1998).
39. Jaganathan, K. et al. Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535–548.e24 (2019).
40. Spier, I. et al. Gene-specific ACMG/AMP classification criteria for germline APC variants: Recommendations from the ClinGen InSiGHT Hereditary Colorectal Cancer/Polyposis Variant Curation Expert Panel. *Genet. Med.* **26**, 100992 (2024).
41. Collins, R. L. et al. A structural variation reference for medical and population genetics. *Nature* **581**, 444–451 (2020).
42. Miki, Y. et al. Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. *Cancer Res*. **52**, 643–645 (1992).
43. Scott, E. C. et al. A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. *Genome Res.* **26**, 745–755 (2016).
44. Qian, Y. et al. Identification of pathogenic retrotransposon insertions in cancer predisposition genes. *Cancer Genet.* **216–217**, 159–169 (2017).
45. Halling, K. C. et al. Hereditary desmoid disease in a family with a germline alu i repeat mutation of the APC gene. *Hum. Heredity* **49**, 97–102 (1999).
46. Su, L.-K. et al. Genomic rearrangements of the APC tumor-suppressor gene in familial adenomatous polyposis. *Hum. Genet.* **106**, 101–107 (2000).
47. Spier, I. et al. Deep intronic APC mutations explain a substantial proportion of patients with familial or early-onset adenomatous polyposis. *Hum. Mutat.* **33**, 1045–1050 (2012).
48. Elias, M. et al. Identification and characterization of SVA retroelement insertions through NGS hereditary cancer panel testing and RNA analysis. *Am. Soc. Hum. Genet. Meet.* Abstract **2343** (2020).
49. Boissinot, S., Entezam, A., Young, L., Munson, P. J. & Furano, A. V. The insertional history of an active family of L1 retrotransposons in humans. *Genome Res*. **14**, 1221–1231 (2004).
50. Garcia-Perez, J. L., Widmann, T. J. & Adams, I. R. The impact of transposable elements on mammalian development. *Development* **143**, 4101–4114 (2016).
51. Flasch, D. A. et al. Genome-wide de novo L1 retrotransposition connects endonuclease activity with replication. *Cell* **177**, 837–851.e28 (2019).
52. Beck, C. R. et al. LINE-1 retrotransposition activity in human genomes. *Cell* **141**, 1159–1170 (2010).
53. Kazazian, H. H. in *Retrotransposons and Human Disease.* 115–127 (WORLD SCIENTIFIC, 2022).
54. van der Sanden, B. P. G. H. et al. The performance of genome sequencing as a first-tier test for neurodevelopmental disorders. *Eur. J. Hum. Genet*. **31**, 81–88 (2023).
55. Liu, Z. et al. Towards accurate and reliable resolution of structural variants for clinical diagnosis. *Genome Biol.* **23**, 68 (2022).
56. Chaisson, M. J. P. et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019).
57. Audano, P. A. et al. Characterizing the major structural variant alleles of the human genome. *Cell* **176**, 663–675.e19 (2019).
58. Scharf, F. et al. Constitutional chromothripsis of the APC locus as a cause of genetic predisposition to colon cancer. *J. Med. Genet.* **59**, 976–983 (2021).
59. Alesi, V. et al. Deep intronic LINE-1 insertions in NF1: expanding the spectrum of neurofibromatosis type 1-associated rearrangements. *Biomolecules* **13**, 725 (2023).
60. Sanchez-Luque, F. J. et al. LINE-1 evasion of epigenetic repression in humans. *Mol. Cell* **75**, 590–604.e12 (2019).
61. Meischl, C., de Boer, M., Åhlin, A. & Roos, D. A new exon created by intronic insertion of a rearranged LINE-1 element as the cause of chronic granulomatous disease. *Eur. J. Hum. Genet.* **8**, 697–703 (2000).
62. Samuelov, L., Fuchs-Telem, D., Sarig, O. & Sprecher, E. An exceptional mutational event leading to Chanarin-Dorfman syndrome in a large consanguineous family. *Br. J. Dermatol*. **164**, 1390–1392 (2011).
63. Kagawa, T. et al. Recessive inheritance of population-specific intronic LINE-1 insertion causes a rotor syndrome phenotype. *Hum. Mutat.* **36**, 327–332 (2015).
64. Rodriguez-Martin, B. et al. Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. *Nat. Genet.* **52**, 306–319 (2020).
65. Lonsdale, J. et al. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
66. Wai, H. A. et al. Blood RNA analysis can increase clinical diagnostic rate and resolve variants of uncertain significance. *Genet. Med.* **22**, 1005–1014 (2020).
67. Landrith, T. et al. Splicing profile by capture RNA-seq identifies pathogenic germline variants in tumor suppressor genes. *NPJ Precis. Oncol.* **4**, 4 (2020).
68. Yépez, V. A. et al. Detection of aberrant gene expression events in RNA sequencing data. *Nat. Protoc.* **16**, 1276–1296 (2021).
69. Wang, F. et al. TEQUILA-seq: a versatile and low-cost method for targeted long-read RNA sequencing. *Nat. Commun.* **14**, 4760 (2023).
70. Marwaha, S., Knowles, J. W. & Ashley, E. A. A guide for the diagnosis of rare and undiagnosed disease: beyond the exome. *Genome Med.* **14**, 23 (2022).
71. Hartley, T. et al. Exome and genome sequencing for rare genetic disease diagnosis: a scoping review and critical appraisal of clinical guidance documents produced by genetics professional organizations. *Genet. Med.* **25**, 100948 (2023).
72. Lunke, S. et al. Integrated multi-omics for rapid rare disease diagnosis on a national scale. *Nat. Med.* **29**, 1681–1691 (2023).
73. Horak, P. et al. Comprehensive genomic and transcriptomic analysis for guiding therapeutic decisions in patients with rare cancers. *Cancer Discov.* **11**, 2780–2795 (2021).
74. Jahn, A. et al. Comprehensive cancer predisposition testing within the prospective MASTER trial identifies hereditary cancer patients and supports treatment decisions for rare cancers. *Ann. Oncol.* **33**, 1186–1199 (2022).

75. Wall, F. E., Henkel, R. D., Stern, M. P., Jenson, H. B. & Moyer, M. P. An efficient method for routine epstein-barr virus immortalization of human B lymphocytes. *Vitr. Cell Dev. Biol. Anim.* **31**, 156–159 (1995).
76. Hackmann, K. et al. Ready to clone: CNV detection and breakpoint fine-mapping in breast and ovarian cancer susceptibility genes by high-resolution array CGH. *Breast Cancer Res. Treat.* **159**, 585–590 (2016).
77. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
78. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* [Preprint] (2013). https://arxiv.org/abs/1303.3997.
79. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
80. Chen, X. et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).
81. Cameron, D. L. et al. GRIDSS2: comprehensive characterisation of somatic structural variation using single breakend variants and structural variant phasing. *Genome Biol.* **22**, 202 (2021).
82. Baid, G. et al. DeepConsensus improves the accuracy of sequences with a gap-aware sequence transformer. *Nat. Biotechnol.* **41**, 232–238 (2023).
83. De Coster, W. & Rademakers, R. NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinformatics* **39**, btad311 (2023).
84. Sedlazeck, F. J. et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
85. Smolka, M. et al. Detection of mosaic and population-level structural variants with Sniffles2. *Nat. Biotechnol.* **42**, 1571–1580 (2024).
86. Heller, D. & Vingron, M. SVIM: structural variant identification using mapped long reads. *Bioinformatics* **35**, 2907–2915 (2019).
87. Jiang, T. et al. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol.* **21**, 189 (2020).
88. Jeffares, D. C. et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 14061 (2017).
89. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
90. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
91. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

Acknowledgements

We would like to thank the patients and the members of the family for allowing us to perform this study. In addition, we would like to thank the DRESDEN-concept Genome Center, Center for Molecular and Cellular Bioengineering, TUD Dresden University of Technology, Dresden, Germany for short- and long-read genome sequencing. Long-read sequencing was performed by the LongRead Team of the Max-Planck Institute of Molecular Cell Biology and Genetics (MPI-CBG) – as part of the DcGC Dresden-concept Genome Center, a core facility of the CMCB and Technology Platform of the TUD Dresden University of Technology. This study was funded by the NCT Dresden. This work was supported (not financially) by the European Reference Network on Genetic Tumor Risk Syndromes (ERN GENTURIS) - Project ID No 739547. ERN GENTURIS is partly co-funded by the European Union within the framework of the Third Health Program "ERN-2016 — Framework Partnership Agreement 2017-2021."

Author contributions

Concept and design: A.A.B., L.K., E.S., H.G., A.J. Drafting of the manuscript: A.A.B., L.K., T.W., E.S., A.J. Bioinformatics: A.A.B., L.K., T.H., D.W. Administrative, technical, or material support: A.R., T.W., M.F., K.H., S.W., D.R., I.S., S.A., D.A., J.P., D.W. Supervision: E.S., H.G., A.J. All the authors contributed for critical revision of the manuscript for important intellectual content, acquisition, analysis, or interpretation of data, accountability for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved and final approval of completed version of manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

E.S.: Honoraria: Illumina. A.J.: Honoraria: AstraZeneca. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41525-025-00485-5.

Correspondence and requests for materials should be addressed to Arne Jahn.