Computational and Structural Biotechnology Journal 27 (2025) 1732–1740

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

Combining long-read DNA and RNA sequencing to enhance molecular understanding of structural variations leading to copy gains

Jade Fauqueux [a], Jean-Pascal Meneboo [b], Roseline Caumes [a,c], Luc Thomès [a], Emilie Ait Yahya [d], Caroline Thuillier [e], Elise Boudry [e], Celine Villenet [b], Martin Figeac [b], Jamal Ghoumid [a,c], Thomas Smol [a,e,*]

[a] ULR 7364—RADEME, Univ. Lille, FHU-G4 Génomique, Lille F-59000, France
[b] Univ. Lille, CNRS, Inserm, CHU Lille, Institut Pasteur de Lille, US 41-UAR 2014-PLBS, FHU-G4 Génomique, Lille F-59000, France
[c] CHU Lille, Clinique de Génétique, Lille, Lille F-59000, France
[d] CHU Lille, Cellule de Bioinformatique, Plateau Commun de Séquençage, CHU Lille, Lille F-59000, France
[e] CHU Lille, Institut de Génétique Médicale, Lille, Lille F-59000, France

ARTICLE INFO

Keywords:
Long Read Sequencing
Nanopore
Cytogenomics
RERE
ZMYM2

ABSTRACT

Structural variants (SVs) significantly contribute to human disease, but their complexity often makes accurate characterization difficult with conventional methods. Advances in long-read sequencing (LRS) offer potential by spanning kilobases and directly resolving SVs. In this study, we examined two individuals with unresolved SVs. LRS on both DNA and cDNA provided single-base resolution of all breakpoint junctions, revealing detailed rearrangement structures and underlying mechanisms. Transcriptomic analyses identified abnormal fusion transcripts and clarified their functional consequences, including haploinsufficiency and potential dominant-negative effects. In one case, a triplication affecting the *ZMYM2* gene was precisely mapped, revealing a truncated variant that may escape nonsense-mediated decay. In the second case, a highly complex reciprocal translocation involving *RERE* and *FHAD1* disrupted *RERE* expression, with Hi-C data showing minimal impact on enhancer-promoter interactions. Due to their complexity, these SVs were not fully resolved by standard methods. By integrating LRS with transcriptomic and chromosomal conformation analyses, we provided a comprehensive understanding of SV formation and its pathogenic impact. Our findings emphasize the need for advanced genomic approaches to resolve complex SVs, enhance diagnostic accuracy, and inform clinical management.

1. Introduction

Structural variants (SVs) are large-scale genomic rearrangements that affect DNA segments of 50 base pairs (bp) or more. SVs are broadly classified into two main types based on their effect on DNA dosage: unbalanced and balanced variants. Unbalanced variants, also known as copy number variants (CNVs), include duplications and deletions that result in DNA gains and losses, disrupting genetic dosage or the regulatory environment. In contrast, balanced variants, such as inversions and balanced translocations, do not change the overall DNA content but reorganize the genomic structure. These rearrangements can affect gene expression or regulation, particularly when breakpoints disrupt functionally critical genes [1–3].

Accurate characterization of SVs is essential for precise diagnosis. Their detection typically involves conventional and molecular cytogenetic techniques, including karyotyping, fluorescence *in situ* hybridization (FISH), comparative genomics hybridization arrays (CGH), and short read sequencing (SRS) [1]. However, these approaches face significant limitations, particularly in resolving complex genomic rearrangements (CGRs) and SVs within highly repetitive genomic regions. Such challenges complicate the precise characterization of some duplications, which is critical for assessing their potential effects on nearby genes or regulatory elements [4]. As a result, some SVs identified by SRS or CGH may remain classified as variants of uncertain significance (VUS), presenting significant challenges for clinical interpretation and decision-making. [5].

In this context, the use of long read sequencing (LRS) technologies to characterize complex events has significantly advanced the understanding of SVs [6,7]. Reads generated by LRS technologies range from 10 to 150 kbp, enabling the capture of full-length or near full-length SVs.

* Corresponding author at: CHU Lille, Institut de Génétique Médicale, Lille, Lille F-59000, France.
E-mail address: thomas.smol@univ-lille.fr (T. Smol).

https://doi.org/10.1016/j.csbj.2025.04.031
Received 26 February 2025; Received in revised form 22 April 2025; Accepted 23 April 2025
Available online 24 April 2025

By reducing sequencing bias and resolving the organization of complex regions at both the DNA and cDNA levels, LRS strategies are considered a key component of cytogenomics [8].

In this study, we employed LRS on both DNA and cDNA to fully characterize two unresolved structural cytogenetic cases: a two-copy gains of the *ZMYM2* gene and a copy gain of the *RERE* gene. The *ZMYM2* gene encodes a zinc finger transcriptional regulator involved in promoting and maintaining cell identity [9]. Monoallelic *ZMYM2* loss-of-function variants cause an autosomal dominant craniofacial syndrome with renal and cardiac anomalies (NECRC, MIM #619522). Additionally, some evidence suggests a potential dominant-negative mechanism through escape from nonsense-mediated decay (NMD) [10]. The *RERE* gene encodes a nuclear receptor coregulator that regulates the retinoic acid signal [11]. Pathogenic variants within *RERE* are associated with an autosomal dominant neurodevelopmental disorder characterized by brain, ocular, and cardiac anomalies (MIM #605226). Most pathogenic variants in *RERE* lead to loss of function, although severe phenotypes have also been reported with specific missense variants [11,12].

Our goal was to integrate a comprehensive DNA and cDNA approach using LRS to refine the classification of structural variants previously designated as variants of uncertain significance.

2. Materials and methods

Peripheral blood samples were collected from the affected individuals and their parents after obtaining informed consent. This study was approved by the *Comité de Protection des Personnes* (CPP) ethics committee, reference #2023-A00473–42. Informed written consent was obtained from all participants or their legal guardians.

Clinical descriptions Patient 1 is a 2-year-old boy born to unrelated parents. The patient was delivered at 39 weeks' gestation after a pregnancy with suspected congenital heart defect that was not confirmed after birth. At 15 months of age, his weight was 9.7 kg (-1 SD), his height was 75 cm (-2 SD), and his cranial circumference was 44 cm (-3 SD). Clinical features include axial hypotonia that began at 2 months of age, feeding difficulties, moderate developmental delay, postnatal microcephaly and cryptorchidism. CGH analysis identified a two-copy gain at 13q12.11 encompassing the *PSPC1*, *ZMYM5*, and *ZMYM2* genes for Patient 1.

Patient 2 is an 11-year-old girl born to unrelated parents after an uncomplicated full-term pregnancy. At 10 years of age, the patient weighed 41 kg (+1 SD), was 1.46 m tall (+2 SD), and had a cranial circumference of 55 cm. She presented with speech delay, with her first words appearing around 3 years of age. She showed global delay and moderate intellectual disability. A copy gain at 1p36.23 encompassing the *RERE* and *ENO1* genes was initially identified by CGH for Patient 2.

2.1. CGH-array

Patient peripheral blood samples were collected in EDTA tubes. Genomic DNA was extracted using the CHEMAGIC-STAR robotic extraction system (Hamilton, Reno, US-NV). Experiments were performed on genomic DNA. Agilent Human Genome CGH 60 K oligonucleotide arrays (Agilent, Santa Clara, CA, USA) were used according to the manufacturer's protocols.

2.2. Nanopore long-read DNA sequencing

DNA libraries were prepared using the Ligation Sequencing Kit SQK-LSK114 (Oxford Nanopore Technologies ONT, Oxford, UK). Sequencing was then performed according to the standard protocol on an ONT PromethION sequencer, using a PromethION flow cell (R10.4.1) connected to a GridION sequencer as the computing device. After sequencing, basecalling was performed using the Guppy basecaller (version 6.4.6) in High Accuracy (HAC) mode.

The resulting reads were mapped to the human reference genome GRCh38 using MiniMap2 (version 2.24) [13]. These mapped data were then used for downstream analysis. Sniffles2 (version 2.0.7) and CuteSV (version 2.1.1) were used to detect structural variation [14,15]. Sniffles2 calling was optimized with specific parameters to enhance detection of complex variants: –long-dup-length 5000000 and –minsupport 3, while CuteSV was used with the parameters –max_size −1 and –min_support 3. Visual inspection of structural variants within target regions was performed using tools the Integrative Genomics Viewer (IGV version 2.10). Breakpoints were confirmed by Sanger sequencing using specific primers, which are available in Additional Table S3.

2.3. qPCR targeting analysis

Quantitative PCR was performed using genomics DNA extracted from the patient and the PowerUp SYBR Green kit (ThermoFisher Scientific, Waltham, MA, USA) with primers specifically designed to quantify copy number within the variant locus. Copy number was normalized to that of a control sample processed under the same conditions as the patient sample. Relative primer levels were calculated using the $2^{-\Delta\Delta Ct}$ method with normalization to a reference gene, and statistical analysis was performed using appropriate software. All experiments were performed in triplicate.

2.4. Optical genome mapping (OGM)

Ultra-high molecular weight DNA was isolated from patient whole blood using the Bionano Prep SP protocol for frozen human blood (Bionano Genomics, San Diego, CA, USA). Briefly, DNA was labeled with the DLE-1 enzyme at the CTTAAG motif and the backbone was stained using the Direct Label and Stain Kit according to the Bionano Prep DLS protocol (Bionano Genomics). The labeled DNA was quantified using the Promega Quantifluor HS assay and then loaded onto a Saphyr chip for linearization and imaging on the Saphyr instrument.

2.5. Cell culture

Human lymphoblastoid cell lines (LCL) were established by Epstein-Barr virus immortalization of the subject's blood lymphocytes and maintained in RPMI 1640 medium (ThermoFisher Scientific) supplemented with 15 % fetal bovine serum (FBS) (ThermoFisher Scientific) and 1 % penicillin-streptomycin (Life Technologies).

2.6. Nanopore long-read cDNA sequencing

RNA was extracted from LCL using the PureLink RNA Minikit (Invitrogen, Carlsbad, CA, USA). For the ONT sequencing experiment, the Direct cDNA sequencing protocol (SQK-LSK114 kit) was followed according to the manufacturer's instructions. Briefly, for each patient, 2–4 μg of total RNA was reverse transcribed using patient-specific primers, followed by strand switching to produce full-length cDNA. The prepared cDNA libraries were then loaded onto separate PromethION flow cells, one for each patient, for sequencing. Base calling was performed using the Guppy basecaller (version 6.4.6) in High Accuracy (HAC) mode. The resulting reads were aligned to the human reference genome GRCh38 using Minimap2 with splice-aware alignment settings to accommodate transcriptomic data. RT-qPCR was performed to examinate the distribution of the transcripts of candidate gene in both patients.

2.7. Conventional cytogenetic analysis

Karyotyping was performed using RH G-banding on metaphase spreads prepared from peripheral blood of patient 2 according to standard procedures. FISH analysis was performed on metaphase and interphase nuclei from blood lymphocytes using fosmid clones

G248P82408E3, G248P86196F8, G248P82014D9, and G248P89339F11 for patient 2.

2.8. Hi-C sequencing

Hi-C sequencing was performed for patient 2 on LCLs using the Arima HiC kit for mammalian cell lines (Arima Genomics, Carlsbad, CA, USA). Briefly, 10 million cells were cross-linked with 2 % formaldehyde, lysed, and digested with enzymes provided in the Arima kit. The resulting fragments were labeled with biotin, ligated, and cross-linking was reversed. Next, the biotin-labeled DNA fragments were sheared to a size of approximately 400 bp using a Covaris S220 instrument (Covaris LLC, Woburn, MA, USA) and AFA tubes. Biotin-labeled fragments were then selectively enriched with beads from the kit. DNA was further processed through repair steps according to the Arima protocol using the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs, Ipswich, MA, USA). Sequencing libraries were prepared by ligation of adapters and indexes using the NEBNext Multiplex Oligos for Illumina kit (New England Biolabs). PCR purification and size selection were performed using Agencourt AMPure XP beads (Beckman Coulter, A63881). The libraries were subjected to deep sequencing (~300 million reads) in a 150 bp paired end run on a NextSeq550 sequencer (Illumina, San Diego, CA, USA). Paired reads obtained from the sequencing were aligned using the Burrows-Wheeler Aligner (BWA-MEM 0.7.17) [16]. Parameters were

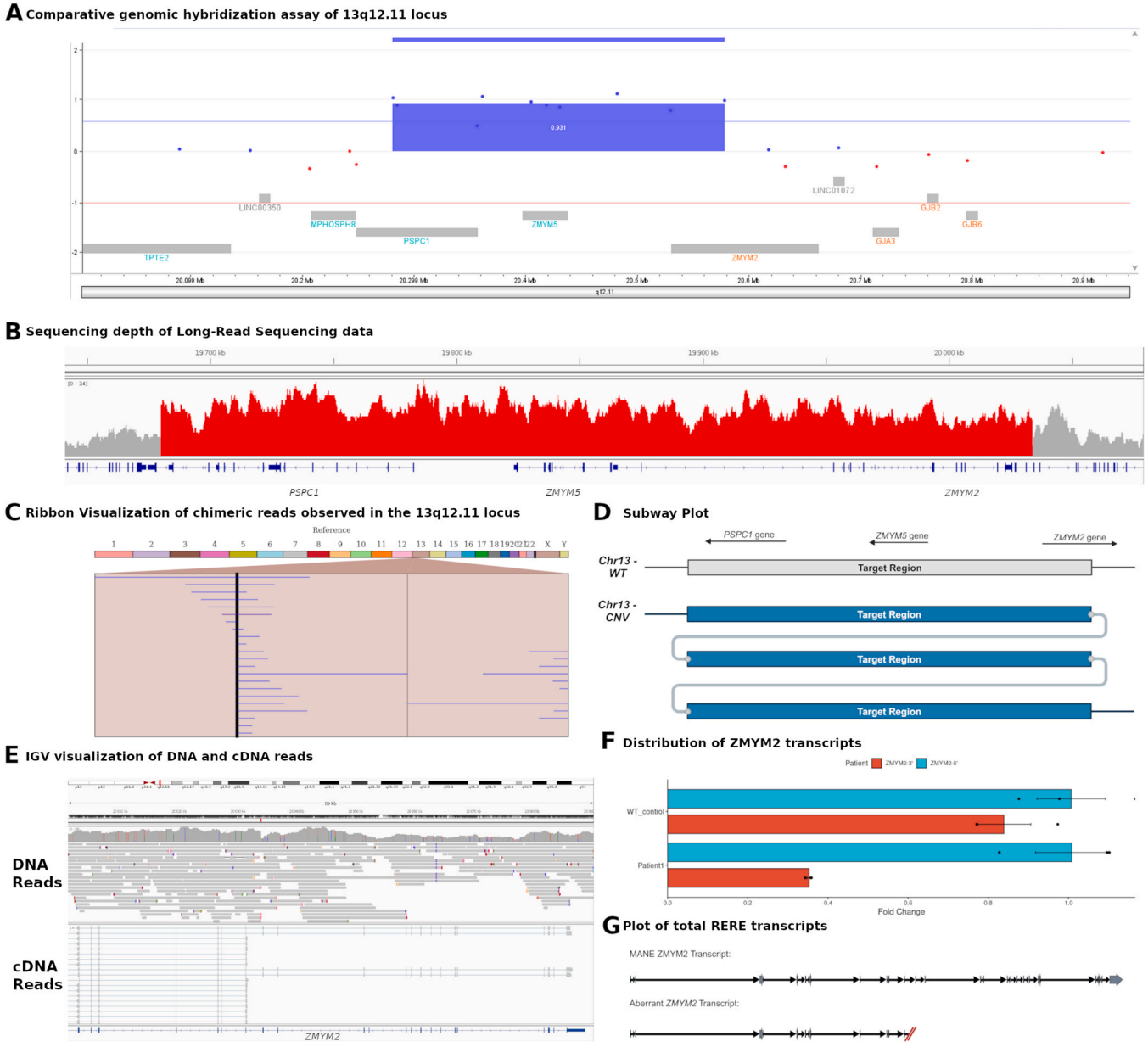

Fig. 1. Molecular analysis of the two-copy gain in Patient 1. (A) Comparative genomic hybridization assay (CGH) analysis reveals a 0.3 Mb gain encompassing the entire *PSPC1* and *ZMYM5* genes, along with a partial duplication of the *ZMYM2* gene. **(B)** Coverage plot from DNA long-read nanopore sequencing shows a two-copy gain encompassing the entire *PSPC1* and *ZMYM5* genes and exons 1–10 of *ZMYM2*, with precise breakpoint identification. **(C)** Ribbon Genome Viewer visualization of long-read nanopore sequencing data identifies a tandem triplication event [43]. **(D)** Schematic representation of wild-type and triplicated alleles, with boxed regions highlighting copy gain loci. **(E)** Integrated Genome Viewer (IGV) visualization of DNA and cDNA long-read nanopore sequencing data showing genome breakpoints and their effect on *ZMYM2* transcripts. **(F)** RT-qPCR analysis quantifying total transcripts (blue bars, targeting the 5′ region of *ZMYM2*) and wild-type transcripts (red bars, targeting the 3′ region of *ZMYM2*). **(G)** Schematic representation of MANE (Matched Annotation from NCBI and EMBL-EBI) and aberrant *ZMYM2* transcripts.

set in accordance to the documentation of HiCExplorer 3.7.3 [17]. Briefly, mates of the Hi-C sequencing read pairs were aligned separately with parameters setting a high gap extension penalty (-E50) and the penalty for 5' and 3' clipping to zero (-L0). As proposed in the documentation, we also set the matching score to 1 (-A1) and the mismatch penalty to 4 (-B4). Alignments were converted from SAM to BAM files and sorted using Sambamba [18]. For further data exploration, Hi-C contact matrices were produced at various resolutions using HiCExplorer. Resolutions of 50 kb and 1 Mb were selected for the visualization of local contact or TADs and inter-chromosomal contacts respectively.

3. Results

3.1. LRS enables base-level characterization of a triplication involving the *ZMYM2* gene

A two-copy gain at 13q12.11 encompassing the *PSPC1*, *ZMYM5*, and *ZMYM2* genes was initially identified by CGH for Patient 1 (Fig. 1A). This finding was subsequently confirmed by targeted qPCR and

Fig. 2. Molecular Analysis of the Highly Complex RERE Rearrangement in Patient 2. **(A)** Comparative genomic hybridization assay (CGH) analysis identifies a 0.5 Mb gain encompassing the entire *ENO1* gene and a partial duplication of the *RERE* gene. **(B)** Coverage plot of DNA long-read nanopore sequencing reveals regions with five-copy and two-copy gains, labeled A to F. **(C)** Subway plot depicting the complete complex rearrangement observed in Patient 2. Gray structures represent wild-type chromosomes, yellow denotes the derivative chromosome 1, and red denotes the derivative chromosome 22. **(E)** IGV visualization of cDNA long-read Nanopore sequencing data, showing full-length wild-type and aberrant *RERE* transcripts terminating at exon 10 and exon 12. Green bar indicate exon 10 of the *RERE* gene and Red bar indicate exon 12 of the *RERE* gene. **(E)** Schematic map of the MANE and Aberrant *RERE* transcripts. **(F)** RT-qPCR analysis quantifying total transcripts (blue bars, targeting the 5′ region of *RERE*) and wild-type transcripts (red bars, targeting the 3′ region of *RERE*).

considered to be *de novo* (Additional Figure S1A).

LRS using Nanopore technology was performed to characterize the rearrangement. All sequencing data are reported in Additional Table S4. Visual read depth analysis and the structural variants callers confirmed the presence of the copy gains (Fig. 1B). Chimeric reads identified the breakpoints at chr13:19,680,196–20,033,950, with each copy gain spanning 353,755 bp and arranged in a direct tandem configuration (Fig. 1C). This duplication encompassed the entire *PSPC1* and *ZMYM5* genes, while the *ZMYM2* gene was disrupted by truncation at intron 10 (Fig. 1D-E). This specific copy number gain has not been previously reported in the gnomAD CNV or DGV population databases. While *ZMYM2* is known to be associated with disease, alterations in *PSPC1* and/or ZMYM5 have not been reported in pathology. In addition, predictive sensitivity scores for copy number gains or losses did not suggest significant pathogenic potential for *ZMYM5* or *PSPC1* [19].

Chimeric sequence analysis revealed junction sequences both upstream and downstream of the copy gains, with a 3-bp microhomology at the breakpoints that was validated by Sanger sequencing (Additional Figure S1B). Furthermore, the absence of similar repeat elements in the region and at the breakpoint suggests that a non-allelic homologous recombination (NAHR) mechanism is unlikely (Additional Tables S1-S2).

The detailed structural characterization provided by LRS confirmed the tandem-direct nature of the rearrangement. This analysis revealed two wild-type alleles of *ZMYM2*, in addition to two truncated alleles containing exons 1–10. Given the predictive sensitivity scores from the pTriplo tool and the absence of previously reported cases with copy number gains affecting these genes, no further investigation into the impact on *PSPC1* and *ZMYM5* was deemed necessary.

cDNA analysis using LRS Nanopore sequencing confirmed the presence of both full-length *ZMYM2* transcripts and aberrant truncated transcripts terminating near exon 10 with polyadenylation signals (Fig. 1E). RT-qPCR was performed to quantify the distribution of full length versus aberrant transcripts and revealed a 50:50 ratio of full length to truncated transcripts in the patient. This finding suggests a frameshift consequence of the triplication SV involving *ZMYM2* (Fig. 1F-G).

3.2. LRS unravels a highly complex structural variant involving the RERE gene

CGH analysis identified a copy gain at 1p36.23 spanning 499 kbp and encompassing the *RERE* and *ENO1* genes in Patient 2 (Fig. 2A). Further examination revealed three consecutive probes within this region with a log ratio of approximately 1, suggesting the presence of a second copy gain within the first, resulting in a DUP-TRP-DUP configuration. The rearrangement was subsequently confirmed by targeted qPCR and considered to be *de novo*. This specific copy number gain has not been previously reported in the gnomAD CNV or DGV population databases [20,21].

LRS using Nanopore technology was performed to further characterize the rearrangement. All sequencing data are reported in Additional Table S4. Depth of coverage analysis confirmed the unbalanced DUP-TRP-DUP, with segment sizes of 87,189 bp, 213,819 bp, and 246,772 bp, named 'B', 'C' and 'D', respectively, corresponding to the CGH result (Additional Table S1). Additionally, three smaller CNVs were identified: a copy gain upstream of the 'B' segment, named 'A', with a size of 1871 bp; and two additional copy gains located 7 Mb downstream at the 1p36.21 locus, with segment sizes of 46,495 bp and 1034 bp, named 'E' and 'F', respectively (Fig. 2B, Additional Figures S2-S3). These variants were validated by qPCR with fold-change values of 1.5, consistent with duplications, and 2.0, consistent with triplications (Additional Figure S4A). For clarity, we will use only the A-B-C-D-E-F designation for each genomic segment (Fig. 2C, Additional Figure S3). Detailed analysis of chimeric read patterns identified multiple organizations involving these different segments, notably an interaction between *locus* 1p36.23 and *locus* 1p36.21, and between these two *loci* and the 22q11.2 region (Fig. 2C). This suggested a more complex rearrangement pattern.

In this configuration, two potential hypotheses of rearrangement were possible: one involving an insertion in chromosome 22 and another involving a reciprocal translocation. Complementary cytogenetic analyses were then performed to differentiate between a translocation from an insertion. Karyotyping and chromosome painting confirmed the presence of 22pter on 1pter and 22qter on 1qter, supporting the hypothesis of a reciprocal translocation (Supplementary Figure S4B-C). OGM and FISH were also performed with probes targeting the 'B', 'C', 'D', and 'E' regions. OGM identified eight breakpoints, while 'A' and 'F' were not detected due to their small size, and confirmed the translocation t(1;22) involving a highly repeated region in the 22q11.2 *locus* (Supplementary Figure S4D-F). Thus, the complete SV, initially identified as a duplication of the *RERE* gene, was found to involve derivative chromosomes 1 and 22 resulting from a t(1;22) translocation, with two unbalanced complex regions at both breakpoints that include genomic segments from the *RERE* region.

The rearrangement configuration indicated the presence of one wild-type allele of the *RERE* gene and two aberrant alleles arising from the der (1) and der(22) chromosomes. LRS transcript analysis revealed the presence of a full-length wild-type transcript along with two distinct aberrant transcript types (Fig. 2D-E). The first aberrant transcript aligned with the *RERE* gene up to exon 10 and continued on chromosome 22, corresponding to der(1). The second aberrant transcript extended to exon 12 of the *RERE* gene, then aligned with exon 3 of the *FHAD1* gene, followed by exons 2–7 of the long non-coding RNA *FAM230G*, corresponding to der(22). Notably, the terminal sequence of exon 12 in *RERE* and the start sequence of exon 3 in *FHAD1* are phased, suggesting the possible formation of a fusion transcript involving *RERE*, *FHAD1*, and *FAM230G* (Fig. 2E). RT-qPCR was performed to quantify the distribution of full length versus aberrant transcripts and revealed a 50:50 ratio of full length to truncated transcripts in the patient. (Fig. 2F). This finding suggests a frameshift equivalent of the CGR involving *RERE* from both der(1) and der(22).

Hi-C sequencing was then used to investigate the effect of the characterized CGR on the 3D organization of chromosomes 1 and 22. Interchromosomal contacts between chromosomes 1 and 22 were confirmed, consistent with the translocation hypothesis (Fig. 3 A). In addition, intra-chromosomal interactions were observed on chromosome 1 between regions 1p36.23 and 1p36.21 (Fig. 3B). Given the disorganization of the 1p36 region, we investigated the potential for novel enhancer-gene interactions. While the translocation appeared to alter chromatin architecture, its effect on enhancer-promoter associations was minimal or negligible (Fig. 3 C). No significant novel enhancer-promoter interactions were observed, suggesting limited regulatory disruption (data not shown).

4. Discussion

The accurate characterization of SVs, especially complex events involving copy gains, remains one of the major challenges in cytogenomics. The advent of new sequencing technologies, particularly long-read sequencing for both DNA and cDNA, has greatly improved the ability to unravel the organization and consequences of SVs [6–8,22]. In this study, the LRS approaches facilitated the characterization of two CGRs, enabling the precise identification of breakpoint locations and a clear definition of the nature of the SVs.

LRS provided a comprehensive analysis for patient 1, precisely characterizing the CNV as a tandem triplication with a breakpoint within intron 10 of the *ZMYM2* gene. This approach localized the triplication and its breakpoints at the nucleotide level, with a 3-bp microhomology at the two breakpoints (Additional Table S2, Additional Figure S1B). This microhomology, in the context of copy gains, provides valuable insight into potential mechanisms and strongly suggests the

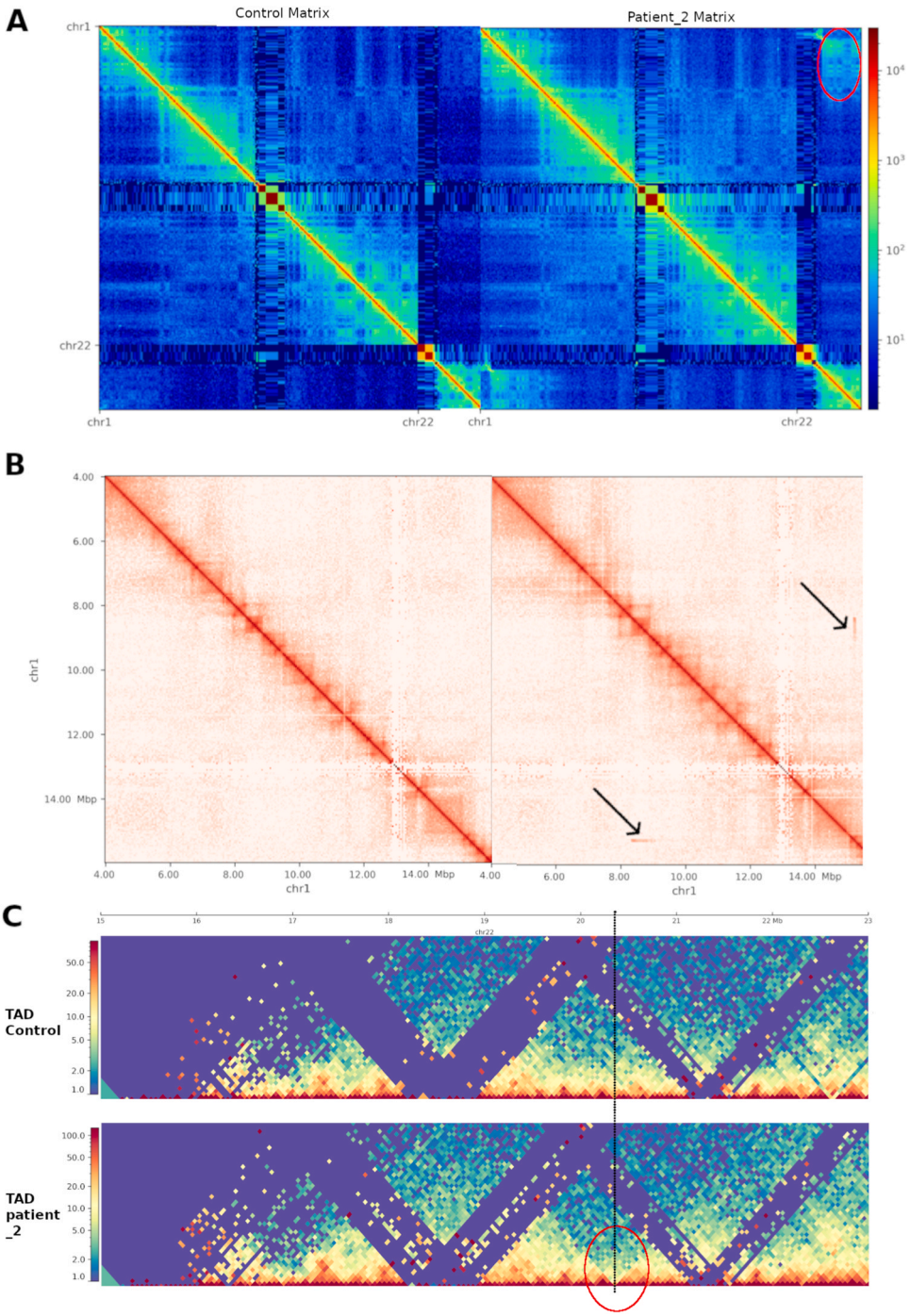

Fig. 3. 3D Genome Analysis Targeting the Structural Rearrangement of Patient 2. **(A)** Hi-C cis contact maps for chromosomes 1 and 22 comparing a wild-type (WT) control and patient 2. Red frames highlight ectopic interchromosomal contacts observed between chromosomes 1 and 22 in patient 2. **(B)** Hi-C cis contact map for chromosome 1 in WT control and patient 2. Black arrows indicate ectopic intra-chromosomal contacts within chromosome 1 in patient 2. **(C)** Topologically associated domain (TAD) map of chromosome 22 (chr22:15,000,000–23,000,000; hg38) derived from Hi-C data for WT control and patient 2. Alterations in the TAD structures in patient 2 are evident compared to the WT control.

involvement of replication-based processes such as microhomology-mediated break-induced replication (MMBIR) or fork stalling and template switching (FoSTeS). These mechanisms likely utilized the 3 bp microhomology as a repair template during the rearrangement event [2,23,24]. Such processes can repeatedly collapse during replication, resulting in structurally more complex variants such as the observed triplication (Fig. 1D). The direct tandem triplication of *ZMYM2* is different from copy gains mediated by the non-allelic homologous recombination (NAHR) mechanism [1], such as type I triplications, which lacks flanking duplications or inversions, or type II triplications, which typically contain inversions and flanking duplications [1,25]. The NAHR mechanism does not apply in this case due to the absence of repeat elements in the region (Additional Tables S1-S2).

In previous reports, microcephaly was documented in 4 of 14 patients, neurodevelopmental disorders in 9 of 14, hypotonia in 3 of 14, feeding difficulties in 2 of 14 and cryptorchidism in 1 patient. Patient 1, who presented with developmental delay, cryptorchidism and microcephaly, was reminiscent of this condition [10]. LRS on cDNA from

Patient 1 provided insight into the effect of the *ZMYM2* rearrangement on transcript profiles. Notably, we observed a 50:50 distribution of wild-type and truncated transcripts in the LCL sample, which is predicted to result in the production of an aberrant ZMYM2 protein. We hypothesize that the truncated *ZMYM2* transcripts may escape nonsense-mediated decay (NMD). As suggested by Connaughton *et al.,* similar truncating pathogenic variants were identified in a family, that displayed a comparable phenotype. Using Bioluminescence Resonance Energy Transfer (BRET) assays, which measure protein-protein interactions by transferring energy from a luciferase to a fluorophore, the authors demonstrated that some truncating pathogenic variants could disrupt the ZMYM2 interactome. If these variants allow the transcript to escape nonsense-mediated decay (NMD), they may impact DNA-binding transcription factors, co-repressors, and proteins involved in chromatin regulation and organization [10]. These alterations may have downstream effects on key factors involved in urinary tract development and may contribute to the observed urinary tract malformations in the patient described in this study [26]. In particular, truncated ZMYM2 variants have been shown to partially interact with FOXP1, a transcription factor implicated in both neurodevelopmental and renal diseases [27]. In line with this hypothesis, these findings suggest that the truncated ZMYM2 proteins could disrupt the native interactome by competing with wild-type proteins, thereby contributing to the patient's phenotype.

LRS of the genome of Patient 2 revealed a CGR with multiple copy number gains, interspersed inversions, and a reciprocal translocation between chromosomes 1 and 22, with breakpoints within the *RERE* and *FHAD1* genes (Fig. 2C). LRS facilitated the precise identification of all copy number gains, their arrangements, and the exact locations of the breakpoints, including base-pair resolution sequencing of the breakpoints themselves [6,28,29]. Despite the challenging nature of the breakpoint locations, which often included repetitive elements, regions of microhomology were identified between interacting genomic regions (Additional Table S1-S2).

We propose a two-step mechanism underlying the rearrangement. In the first step, repair by mechanisms such as MMBIR/FoSTeS likely induced chromoanasynthesis on chromosome 1, leading to the observed CNV. Microhomologies ranging from 3 to 8 base pairs may have served as repair templates. The microhomology between segments 'C' and 'F' was not directly observed, possibly due to the loss of the corresponding chimeric read, suggesting that this interaction was disrupted by subsequent repair events [2,23,24].

In a second step, two double-strand breaks (DSBs) on chromosomes 1 and 22 appear to have triggered the formation of the reciprocal translocation [30]. At the chromosome 22 breakpoint, a region containing multiple copies of a type A translocation breakpoint sequence –characterized by AT-rich, highly variable segments– was identified (Additional Figure S5A) [31]. This region tends to form cruciform secondary structures, which increase genomic instability and make the region more susceptible to DSBs (Additional Figure S5B) [32]. Similarly, the DSB on chromosome 1 may have been driven by an inversion near region 'B' which could have induced secondary structure formation, further promoting genomic instability and leading to a break [33].

Once these DSBs occurred, cellular repair mechanisms were engaged to restore genomic integrity. Based on the sequences identified at the breakpoints, non-homologous end joining (NHEJ) appears to be the most likely mechanism, facilitating the fusion of 1pTER/22pTER and 1qTER/22qTER (Fig. 2C, Additional Figure S5C). This repair process explains the reciprocal translocation and provides a plausible model for the observed complex rearrangement. This proposed two-step mechanism could have occurred during different developmental windows—premeiotic, meiotic, or postzygotic. While these stages are considered critical for the emergence of such genomic alterations, pinpointing the exact timing remains challenging.

Transcriptomic analysis and structural studies provided comprehensive insights into how the observed complex rearrangement affects *RERE* gene expression. Monoallelic pathogenic variants of *RERE* are associated with a neurodevelopmental disorder characterized by brain, eye and heart abnormalities (MIM #605226) [12]. Pathogenic variants in *RERE* are predominantly loss-of-function. However, severe phenotypes have also been associated with specific missense variants within the atrophin-1 domain, suggesting a dominant-negative mechanism may contribute to more severe clinical presentations [11,12]. The cumulative data –showing one wild-type transcript alongside two aberrant truncating *RERE* transcripts (Fig. 2D-E)– strongly support haploinsufficiency as the most likely pathogenic mechanism.

RT-qPCR analysis, which revealed a 50:50 ratio of wild-type to aberrant transcripts, highlights several intriguing findings. The first aberrant transcript, comprising exons 1–10 of the *RERE* gene and aligned to chromosome 22, appears to be degraded *via* the NMD pathway [34]. The second aberrant transcript, spanning exons 1–12 of the *RERE* gene, exon 3 of the *FHAD1* gene, and exons 2–7 of the *FAM230G* lncRNA, may have functional relevance. This fusion transcript suggests the possibility of producing a chimeric protein that may contribute to the observed phenotype. Notably, fusion transcripts involving lncRNAs, as observed in cancer studies, have been shown to generate functional proteins. This raises the possibility that the second aberrant transcript exerts a biological effect that contributes to the phenotype. The observed 50:50 ratio between the second aberrant transcript and the WT transcript further supports its potential functional relevance [35]. Consistent with the literature, patients with *RERE* variants leading to haploinsufficiency often present with milder and less specific phenotypes, aligning with the clinical features observed in Patient 2 [11,12]. Surprisingly, the complex SV involving multiple CNVs, inversions, and a translocation appears to have minimal impact on the regulatory environment and may only lead to *RERE* dysregulation. Although Hi-C data suggested a possible reorganization of topologically associated domains (Fig. 3 C), no evidence of enhancer hijacking events was observed in this case [36–38].

In both reported cases, LRS significantly enhanced our understanding of CNVs and CGRs. Although both SVs occurred *de novo* in the patients, they could potentially be transmitted to future generations. The detailed characterization of these complex rearrangements is therefore essential for genetic counseling, enabling a better understanding of potential reproductive risks. This technology allowed us to identify the likely molecular diagnosis in these patients, who experienced a "diagnostic odyssey". Nevertheless, analysis of LRS data in the context of medical molecular diagnosis remains challenging, and many bioinformatics issue nedd to be addressed. Current structural variant callers often fail to detect all breakpoints and to capture the full complexity of such rearrangments [39]. The two variant calling software tools used, CuteSV and Sniffles2, accurately both identified only the duplication in Patient 1 and a single inversion in Patient 2 [14,15]. Additionally, CuteSV identified an inversion between regions 'A' and 'B', and a deletion between regions 'A' and 'E'. The detection of this deletion is not surprising, as it reflects the contact observed between these two regions. However, the variant is inaccurately named: it is not a true deletion, but rather an insertion of a duplicated segment, which the caller did not take into account. This discrepancy highlights the limitations of current variant calling algorithms, especially when analyzing complex rearrangements. Factors such as structural variant complexity, sequencing quality, and read depth play a key role in variant detection and characterization. For example, low coverage or regions of high sequence complexity can hinder breakpoint resolution, leading to underrepresentation of specific rearrangements and incomplete variant profiles [39]. The development of improved algorithms that effectively integrate depth, read mapping consistency, and sequence context will be key to unlocking the full potential of long-read sequencing in genomic diagnostics. These advances could improve breakpoint detection and enable more accurate characterization of complex rearrangements. However, manual curation and complementary experimental validation remain essential components of comprehensive genomic analysis to

ensure accurate interpretation and clinical relevance of identified structural variants.

LRS has demonstrated its versatility by enabling both genome analysis and transcriptome characterization. Unlike traditional methods, it enables direct sequencing of full-length transcripts without prior PCR amplification, preserving native RNA features and minimizing amplification-induced bias [40,41]. This capability was essential for the accurate identification of truncated and fusion transcripts in both patients, providing critical insight into the functional impact of structural variants on gene expression. SRS techniques lack the ability to directly span large genomic regions, relying instead on statistical inference to piece together sequences. As a result, they often struggle to accurately resolve intricate fusion events and other complex structural variations [42]. While OGM is a powerful tool for detecting large structural variants and provides an intuitive visualization of genome-wide rearrangements, it has resolution limitations that may hinder its ability to capture finer details of complex genomic rearrangements (CGRs). Specifically, OGM has a resolution threshold of approximately 5 kb, which can result in missed detection of smaller events, as observed in our study with regions ‘A’ and ‘F’. In a separate study comparing OGM to long-read sequencing (LRS) using Nanopore technology, OGM detected 74 % of the junctions in complex cases, whereas LRS alone identified 100 % of these junctions [7].

Our findings emphasize the importance of integrating long-read sequencing of DNA and cDNA into routine clinical diagnostics for rare diseases. These approaches provide essential insights into the characteristics of structural variants, facilitate hypotheses about their formation mechanisms and pathogenic effects, and ultimately improve the classification of VUS. By enhancing diagnostic accuracy, this comprehensive strategy supports more personalized and effective clinical management for affected individuals.

CRediT authorship contribution statement

Thuillier Caroline: Validation, Methodology, Investigation. **Yahya Emilie:** Software, Methodology, Investigation. **Meneboo Jean-Pascal:** Software, Methodology, Investigation, Formal analysis. **Fauqueux Jade:** Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis. **Thomès Luc:** Software, Methodology, Investigation. **Caumes Roseline:** Investigation, Data curation. **Figeac Martin:** Software, Methodology, Investigation. **Villenet Celine:** Formal analysis. **Smol Thomas:** Writing – review & editing, Writing – original draft, Validation, Resources, Project administration, Funding acquisition, Conceptualization. **Ghoumid Jamal:** Writing – review & editing, Writing – original draft, Validation, Resources, Project administration, Conceptualization. **Boudry Elise:** Formal analysis, Data curation.

Declaration of Competing Interest

All authors declare that they have no conflicts of interest.

Acknowledgments

We thank the patients and their families for participating in this study. This work was supported by the University Hospital of Lille, France (Grant No. BPI22-07-1531). We thank Delphine Ceraso, Alexis Leurent, Heidi Tampere, and Pauline Grave for their technical assistance with LRS, primer design, and RT-PCR experiments. We thank Julien Merlin for his technical assistance with FISH analysis.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2025.04.031.

Data Availability

All data sets used of analyzed in this study are available from the corresponding author upon request.

References

[1] Weckselblatt B, Rudd MK. Human structural variation: mechanisms of chromosome rearrangements. Trends Genet 2015;31:587–99. https://doi.org/10.1016/j.tig.2015.05.010.
[2] Carvalho CMB, Lupski JR. Mechanisms underlying structural variant formation in genomic disorders. Nat Rev Genet 2016;17:224–38. https://doi.org/10.1038/nrg.2015.25.
[3] Schuy J, Grochowski CM, Carvalho CMB, Lindstrand A. Complex genomic rearrangements: an underestimated cause of rare diseases. Trends Genet TIG 2022;38:1134–46. https://doi.org/10.1016/j.tig.2022.06.003.
[4] Newman S, Hermetz KE, Weckselblatt B, Rudd MK. Next-generation sequencing of duplication CNVs reveals that most are tandem and some create fusion genes at breakpoints. Am J Hum Genet 2015;96:208–20. https://doi.org/10.1016/j.ajhg.2014.12.017.
[5] Boycott KM, Rath A, Chong JX, Hartley T, Alkuraya FS, Baynam G, et al. International cooperation to enable the diagnosis of all rare genetic diseases. Am J Hum Genet 2017;100:695–705. https://doi.org/10.1016/j.ajhg.2017.04.003.
[6] Miller DE, Sulovari A, Wang T, Loucks H, Hoekzema K, Munson KM, et al. Targeted long-read sequencing identifies missing disease-causing variation. Am J Hum Genet 2021;108:1436–49. https://doi.org/10.1016/j.ajhg.2021.06.006.
[7] De Clercq G, Vantomme L, Dewaele B, Callewaert B, Vanakker O, Janssens S, et al. Full characterization of unresolved structural variation through long-read sequencing and optical genome mapping. Sci Rep 2024;14:29142. https://doi.org/10.1038/s41598-024-80068-z.
[8] Damaraju N, Miller AL, Miller DE. Long-read DNA and RNA sequencing to streamline clinical genetic testing and reduce barriers to comprehensive genetic testing. J Appl Lab Med 2024;9:138–50. https://doi.org/10.1093/jalm/jfad107.
[9] Owen DJ, Aguilar-Martinez E, Ji Z, Li Y, Sharrocks AD. ZMYM2 controls human transposable element transcription through distinct co-regulatory complexes. eLife 2023;12:RP86669. https://doi.org/10.7554/eLife.86669.
[10] Connaughton DM, Dai R, Owen DJ, Marquez J, Mann N, Graham-Paquin AL, et al. Mutations of the transcriptional corepressor ZMYM2 cause syndromic urinary tract malformations. Am J Hum Genet 2020;107:727–42. https://doi.org/10.1016/j.ajhg.2020.08.013.
[11] Fregeau B, Kim BJ, Hernández-García A, Jordan VK, Cho MT, Schnur RE, et al. De Novo mutations of RERE cause a genetic syndrome with features that overlap those associated with proximal 1p36 deletions. Am J Hum Genet 2016;98:963–70. https://doi.org/10.1016/j.ajhg.2016.03.002.
[12] Jordan VK, Fregeau B, Ge X, Giordano J, Wapner RJ, Balci TB, et al. Genotype-phenotype correlations in individuals with pathogenic RERE variants. Hum Mutat 2018;39:666–75. https://doi.org/10.1002/humu.23400.
[13] Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinforma Oxf Engl 2018;34:3094–100. https://doi.org/10.1093/bioinformatics/bty191.
[14] Smolka M, Paulin LF, Grochowski CM, Horner DW, Mahmoud M, Behera S, et al. Detection of mosaic and population-level structural variants with Sniffles2. Nat Biotechnol 2024. https://doi.org/10.1038/s41587-023-02024-y.
[15] Jiang T, Liu Y, Jiang Y, Li J, Gao Y, Cui Z, et al. Long-read-based human genomic structural variation detection with cuteSV. Genome Biol 2020;21:189. https://doi.org/10.1186/s13059-020-02107-y.
[16] Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM 2013. https://doi.org/10.48550/arXiv.1303.3997.
[17] Ramírez F, Bhardwaj V, Arrigoni L, Lam KC, Grüning BA, Villaveces J, et al. High-resolution TADs reveal DNA sequences underlying genome organization in flies. Nat Commun 2018;9:189. https://doi.org/10.1038/s41467-017-02525-w.
[18] Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. Bioinforma Oxf Engl 2015;31:2032–4. https://doi.org/10.1093/bioinformatics/btv098.
[19] Riggs ER, Andersen EF, Cherry AM, Kantarci S, Kearney H, Patel A, et al. Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). Genet Med J Am Coll Med Genet 2020;22:245–57. https://doi.org/10.1038/s41436-019-0686-8.
[20] Chen S, Francioli LC, Goodrich JK, Collins RL, Kanai M, Wang Q, et al. A genomic mutational constraint map using variation in 76,156 human genomes. Nature 2023. https://doi.org/10.1038/s41586-023-06045-0.
[21] MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW. The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res 2014;42:D986–92. https://doi.org/10.1093/nar/gkt958.
[22] Fujimoto A, Wong JH, Yoshii Y, Akiyama S, Tanaka A, Yagi H, et al. Whole-genome sequencing with long reads reveals complex structure and origin of structural variation in human genetic variations and somatic mutations in cancer. Genome Med 2021;13:65. https://doi.org/10.1186/s13073-021-00883-1.
[23] Bahrambeigi V, Song X, Sperle K, Beck CR, Hijazi H, Grochowski CM, et al. Distinct patterns of complex rearrangements and a mutational signature of microhomeology are frequently observed in PLP1 copy number gain structural variants. Genome Med 2019;11:80. https://doi.org/10.1186/s13073-019-0676-0.

[24] Currall BB, Talkowski Chiang C, Morton ME. CC. Mechanisms for structural variation in the human genome. Curr Genet Med Rep 2013;1:81–90. https://doi.org/10.1007/s40142-013-0012-8.
[25] Shimojima K, Mano T, Kashiwagi M, Tanabe T, Sugawara M, Okamoto N, et al. Pelizaeus-Merzbacher disease caused by a duplication-inverted triplication-duplication in chromosomal segments including the PLP1 region. Eur J Med Genet 2012;55:400–3. https://doi.org/10.1016/j.ejmg.2012.02.013.
[26] Estruch SB, Graham SA, Quevedo M, Vino A, Dekkers DHW, Deriziotis P, et al. Proteomic analysis of FOXP proteins reveals interactions between cortical transcription factors associated with neurodevelopmental disorders. Hum Mol Genet 2018. https://doi.org/10.1093/hmg/ddy230.
[27] Wu S-T, Feng Y, Song R, Qi Y, Li L, Lu D, et al. Foxp1 is required for renal intercalated cell differentiation and acid-base regulation. J Am Soc Nephrol JASN 2024;35:533–48. https://doi.org/10.1681/ASN.0000000000000319.
[28] Merker JD, Wenger AM, Sneddon T, Grove M, Zappala Z, Fresard L, et al. Long-read genome sequencing identifies causal structural variation in a Mendelian disease. Genet Med J Am Coll Med Genet 2018;20:159–63. https://doi.org/10.1038/gim.2017.86.
[29] Nakamichi K, Huey J, Sangermano R, Place EM, Bujakowska KM, Marra M, et al. Targeted long-read sequencing enriches disease-relevant genomic regions of interest to provide complete Mendelian disease diagnostics. JCI Insight 2024;9:e183902. https://doi.org/10.1172/jci.insight.183902.
[30] Lieber MR, Gu J, Lu H, Shimazaki N, Tsai AG. Nonhomologous DNA end joining (NHEJ) and chromosomal translocations in humans. Subcell Biochem 2010;50:279–96. https://doi.org/10.1007/978-90-481-3471-7_14.
[31] Delihas N. A family of long intergenic non-coding RNA genes in human chromosomal region 22q11.2 carry a DNA translocation breakpoint/AT-rich sequence. PloS One 2018;13:e0195702. https://doi.org/10.1371/journal.pone.0195702.
[32] Correll-Tash S, Lilley B, Salmons Iv H, Mlynarski E, Franconi CP, McNamara M, et al. Double strand breaks (DSBs) as indicators of genomic instability in PATRR-mediated translocations. Hum Mol Genet 2021;29:3872–81. https://doi.org/10.1093/hmg/ddaa251.
[33] Ait Saada A, Guo W, Costa AB, Yang J, Wang J, Lobachev KS. Widely spaced and divergent inverted repeats become a potent source of chromosomal rearrangements in long single-stranded DNA regions. Nucleic Acids Res 2023;51:3722–34. https://doi.org/10.1093/nar/gkad153.
[34] Hug N, Longman D, Cáceres JF. Mechanism and regulation of the nonsense-mediated decay pathway. Nucleic Acids Res 2016;44:1483–95. https://doi.org/10.1093/nar/gkw010.
[35] Sánchez-Marín D, Silva-Cázares MB, Porras-Reyes FI, García-Román R, Campos-Parra AD. Breaking paradigms: Long non-coding RNAs forming gene fusions with potential implications in cancer. Genes Dis 2024;11:101136. https://doi.org/10.1016/j.gendis.2023.101136.
[36] Wang X, Xu J, Zhang B, Hou Y, Song F, Lyu H, et al. Genome-wide detection of enhancer-hijacking events from chromatin interaction data in re-arranged genomes. Nat Methods 2021;18:661–8. https://doi.org/10.1038/s41592-021-01164-w.
[37] Melo US, Jatzlau J, Prada-Medina CA, Flex E, Hartmann S, Ali S, et al. Enhancer hijacking at the ARHGAP36 locus is associated with connective tissue to bone transformation. Nat Commun 2023;14:2034. https://doi.org/10.1038/s41467-023-37585-8.
[38] Ramani V, Shendure J, Duan Z. Understanding spatial genome organization: methods and insights. Genom Proteom Bioinforma 2016;14:7–20. https://doi.org/10.1016/j.gpb.2016.01.002.
[39] Helal AA, Saad BT, Saad MT, Mosaad GS, Aboshanab KM. Benchmarking long-read aligners and SV callers for structural variation detection in Oxford nanopore sequencing data. Sci Rep 2024;14:6160. https://doi.org/10.1038/s41598-024-56604-2.
[40] Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. Nat Rev Genet 2020;21:597–614. https://doi.org/10.1038/s41576-020-0236-x.
[41] Vollmers C, Penland L, Kanbar JN, Quake SR. Novel exons and splice variants in the human antibody heavy chain identified by single cell and single molecule sequencing. PloS One 2015;10:e0117050. https://doi.org/10.1371/journal.pone.0117050.
[42] Nattestad M, Goodwin S, Ng K, Baslan T, Sedlazeck FJ, Rescheneder P, et al. Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. Genome Res 2018;28:1126–35. https://doi.org/10.1101/gr.231100.117.
[43] Nattestad M, Aboukhalil R, Chin C-S, Schatz MC. Ribbon: intuitive visualization for complex genomic variation. Bioinforma Oxf Engl 2021;37:413–5. https://doi.org/10.1093/bioinformatics/btaa680.